

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KINH TẾ-LUẬT**



**BÀI BÁO CÁO GIỮA KỲ**

**Môn: PHÂN TÍCH DỮ LIỆU VỚI PYTHON**

**CHỦ ĐỀ: PHÂN TÍCH DỮ LIỆU VÀ DỰ ĐOÁN TỶ LỆ RỜI BỎ CỦA KHÁCH HÀNG CỦA CÔNG TY VIỄN THÔNG BẰNG CÁC MÔ HÌNH HỌC MÁY**

**GIẢNG VIÊN HƯỚNG DẪN: HỒ THỊ LINH**

**THÀNH VIÊN NHÓM:**

- |                                |                   |
|--------------------------------|-------------------|
| <b>1. Phạm Tấn Tùng</b>        | <b>K214060421</b> |
| <b>2. Nguyễn Như Chính Tâm</b> | <b>K214061261</b> |
| <b>3. Lê Dương Thảo</b>        | <b>K214061745</b> |

*Thành phố Hồ Chí Minh, ngày tháng 9 năm 2024*

## LỜI CẢM ƠN

---

Sau một học kỳ nỗ lực học tập môn học "Phân tích dữ liệu với R/python", dự án cuối kỳ của nhóm chúng em là một cột mốc quan trọng, nơi chúng em tổng hợp kiến thức đã học và áp dụng vào các tình huống thực tiễn.

Chúng em cũng xin bày tỏ lòng biết ơn sâu sắc đến cô Hồ Thị Linh, giảng viên của bộ môn "Phân tích dữ liệu với R/python" vì sự tận tâm và truyền đạt những kiến thức vô cùng quý báu trong suốt quá trình học. Sự hướng dẫn nhiệt tình của các cô đã giúp chúng em có nền tảng vững chắc để hoàn thiện dự án này.

Do hạn chế về thời gian và kiến thức còn chưa sâu rộng, dự án của chúng em có thể còn nhiều thiếu sót. Vì vậy, chúng em rất mong nhận được sự thông cảm cũng như những góp ý quý báu từ cô để chúng em có thể hoàn thiện hơn trong tương lai.

Nhóm chúng em xin chân thành cảm ơn.

## MỤC LỤC

---

### Mục lục

<b>1.1 Lý do chọn đề tài.....</b>	<b>2</b>
<b>1.2 Mục tiêu đề tài.....</b>	<b>2</b>
<b>1.3 Đối tượng và phạm vi nghiên cứu .....</b>	<b>3</b>
<b>1.4 Công cụ .....</b>	<b>3</b>
<b>1.5 Ý nghĩa nghiên cứu .....</b>	<b>3</b>
<b>1.6 Cấu trúc báo cáo .....</b>	<b>4</b>
<b>2.1 Tổng quan về Phân tích dữ liệu.....</b>	<b>5</b>
<i>2.1.1. Lợi ích của Phân tích dữ liệu trong kinh doanh .....</i>	<i>5</i>
Phân tích dữ liệu giúp cải thiện trải nghiệm khách hàng, tối ưu hóa chi phí và nâng cao hiệu quả kinh doanh. Cụ thể, các doanh nghiệp có thể:.....	5
<i>2.1.2. Quy trình thực hiện dự án phân tích dữ liệu trong kinh doanh .....</i>	<i>5</i>
<b>2.2. Lý thuyết và phương pháp trong phân tích dữ liệu .....</b>	<b>6</b>
2.2.1. Lý thuyết về sự rời bỏ của khách hàng .....	6
2.2.2. Học máy và mô hình học máy. ....	7
2.2.3. Các thuật toán máy học .....	9
2.2.3.1. <i>K-Nearest Neighbors .....</i>	<i>9</i>
2.2.3.2. <i>Logistic Regression - Hồi quy logistic .....</i>	<i>10</i>
2.2.3.3. <i>Random Forest - Rừng ngẫu nhiên.....</i>	<i>11</i>
2.2.4. Phương pháp phân tích dữ liệu .....	12
<b>3.1. Xác định và phân tích các yêu cầu của người dùng .....</b>	<b>13</b>
<b>3.2. Tổng quan về cơ sở dữ liệu nguồn .....</b>	<b>13</b>
3.2.1. Mô tả dữ liệu: .....	13
3.2.2 Lựa chọn và trình bày dữ liệu để phân tích yêu cầu của người dùng: .....	16
<b>4.1. Giới thiệu về các công cụ và giải pháp phân tích dữ liệu: .....</b>	<b>17</b>
4.1.1. Python .....	17
4.2.2. Google Colab .....	17
<b>4.2. Phân tích dữ liệu, khám phá và trực quan hóa: .....</b>	<b>17</b>

4.2.1. Khai phá và tiền xử lí dữ liệu:.....	17
4.2.2. Phân tích dữ liệu: .....	21
a. Phân tích mô tả: .....	21
b. Phân tích chẩn đoán:.....	25
4.2.3. Phân tích dự đoán và đánh giá mô hình: .....	28
<b>4.3. Thảo luận và đánh giá kết quả để hỗ trợ ra quyết định kinh doanh .....</b>	<b>32</b>
<b>5.1. Kết quả và hạn chế: .....</b>	<b>34</b>
<b>5.2. Hướng phát triển chủ đề .....</b>	<b>34</b>

## DANH MỤC HÌNH ẢNH

---

Hình 1: Sự rời bỏ của khách hàng - Customer churn .....	<b>Error! Bookmark not defined.</b>
Hình 2: Công thức tính tỷ lệ rời bỏ của khách hàng.....	<b>Error! Bookmark not defined.</b>
Hình 3: Công thức khoảng cách Euclidean .....	9
Hình 4: Công thức khoảng cách Manhattan .....	9
Hình 5: Công thức khoảng cách Minkowski .....	10
Hình 6: Định nghĩa Hàm sigmoid .....	10
Hình 7: Phương trình Hồi quy logistic .....	10
Hình 8: Mô tả thuật toán Random Forest .....	<b>Error! Bookmark not defined.</b>
Hình 9: Thông tin chung về bộ dữ liệu.....	19
Hình 10: Boxplot xác định giá trị ngoại lai .....	20
Hình 11 Thay đổi kiểu dữ liệu của thuộc tính .....	<b>Error! Bookmark not defined.</b>
Hình 12: Mã hóa dữ liệu.....	21
Hình 13: Thống kê mô tả của thuộc tính .....	22
Hình 14: Tỷ lệ rời bỏ của khách hàng .....	22
Hình 15: Số lượng khách hàng rời bỏ và giới tính .....	<b>Error! Bookmark not defined.</b>
Hình 16: Bảng tương quan các thuộc tính của bộ dữ liệu .....	23
Hình 17: Phân phối dữ liệu các thuộc tính .....	24
Hình 18: Mối quan hệ giữa sự rời bỏ và số tiền thanh toán hàng tháng.....	25
Hình 19: Mối quan hệ giữa sự rời bỏ và dịch vụ.....	26
Hình 20: Sự phụ thuộc của thời gian gắn bó và chi phí .....	27
Hình 21: Mối quan hệ giữa sự rời bỏ và tổng số tiền sử dụng dịch vụ .....	27
Hình 22: Mô hình dự đoán Logistic Regression .....	30
Hình 23: Mô hình dự đoán K-Nearest Neighbors và Random Forest.....	30
Hình 24: Ma trận nhầm lẫn của mô hình dự đoán Logistic Regression.....	31
Hình 25: Ma trận nhầm lẫn của mô hình dự đoán K-Nearest Neighbors.....	31
Hình 26: Ma trận nhầm lẫn của mô hình dự đoán Random Forest .....	32

## **DANH MỤC BẢNG**

---

Bảng 3.2.1 Bảng mô tả thuộc tính của bộ dữ liệu .....	14
---	----

## GIỚI THIỆU

---

Ngành viễn thông là một trong những lĩnh vực phát triển nhanh chóng và có tính cạnh tranh cao. Việc duy trì khách hàng hiện tại là yếu tố quan trọng để đảm bảo sự phát triển bền vững của doanh nghiệp. Tuy nhiên, tình trạng khách hàng rời bỏ dịch vụ (churn) luôn là mối lo ngại lớn, gây ra tổn thất đáng kể cho doanh thu và lợi nhuận của các công ty viễn thông. Để giảm thiểu tình trạng này, cần thực hiện phân tích dữ liệu để xác định các yếu tố ảnh hưởng và dự đoán xu hướng rời bỏ khách hàng, từ đó giúp doanh nghiệp xây dựng các chiến lược giữ chân khách hàng hiệu quả.

Những khách hàng trung thành không chỉ có xu hướng mua sắm thường xuyên hơn mà còn có khả năng giới thiệu sản phẩm hoặc dịch vụ của công ty cho người khác, tạo ra một nguồn khách hàng tiềm năng mới mà không tốn thêm chi phí quảng cáo. Do đó, việc dự đoán tình trạng ra đi của khách hàng là rất quan trọng, giúp các doanh nghiệp kịp thời phát triển các chiến lược và biện pháp để giữ chân khách hàng.

Công nghệ và phân tích dữ liệu đóng vai trò quan trọng trong việc dự đoán tình trạng ra đi của khách hàng. Bằng cách sử dụng các kỹ thuật máy học và dự đoán chuỗi, các doanh nghiệp có thể phân tích hành vi của khách hàng và xác định các dấu hiệu tiềm ẩn của sự ra đi. Với mục tiêu ứng dụng học máy vào phân tích và dự đoán tỷ lệ rời bỏ, nhóm đã chọn đề tài “Phân tích dữ liệu và dự đoán tỷ lệ rời bỏ của khách hàng của công ty viễn thông bằng các mô hình học máy”. Nhóm sẽ sử dụng các mô hình K-Nearest Neighbors, Random Forest, và Logistic Regression để so sánh độ chính xác của từng mô hình, từ đó đưa ra các giải pháp tối ưu nhằm giúp doanh nghiệp giảm thiểu tỷ lệ rời bỏ khách hàng. Trong báo cáo này, nhóm sẽ giới thiệu bối cảnh dự án và các nền tảng lý thuyết, cũng như trình bày quá trình thực hiện và kết quả của việc xây dựng mô hình dự đoán tình trạng ra đi của khách hàng.

## CHƯƠNG 1: TỔNG QUAN

---

### 1.1 Lý do chọn đề tài

Ngành viễn thông đang phát triển nhanh chóng và có tính cạnh tranh cao, với việc duy trì khách hàng đóng vai trò quan trọng trong sự phát triển bền vững của doanh nghiệp. Tuy nhiên, tình trạng khách hàng rời bỏ dịch vụ (churn) gây tổn thất lớn cho doanh thu. Phân tích dữ liệu để xác định và dự đoán các yếu tố rời bỏ khách hàng là cần thiết, giúp doanh nghiệp phát triển chiến lược giữ chân hiệu quả. Nhiều nghiên cứu đã chỉ ra rằng việc giữ chân khách hàng mang lại lợi nhuận cao hơn so với việc thu hút khách hàng mới (Wouter, 2003; Nor, 2008; Alan, 1994). Theo Freshworks (Aysha Shereen, 2021), chi phí để thu hút khách hàng mới cao hơn từ 7-13% so với việc giữ lại những khách hàng hiện có. Khách hàng trung thành không chỉ mua sắm thường xuyên mà còn giới thiệu sản phẩm đến người khác, tạo thêm khách hàng mới mà không tốn chi phí quảng cáo.

Nhận thấy sự cần thiết của việc giảm tỷ lệ rời bỏ khách hàng trong ngành viễn thông, nhóm đã quyết định chọn đề tài “Phân tích dữ liệu và dự đoán tỷ lệ rời bỏ khách hàng trong ngành viễn thông bằng các mô hình học máy.” Mục tiêu của đề tài là ứng dụng các kỹ thuật học máy, bao gồm K-Nearest Neighbors, Random Forest và Logistic Regression, để phân tích và dự đoán tỷ lệ rời bỏ của khách hàng. Bằng cách so sánh độ chính xác của các mô hình này, nhóm hy vọng có thể đề xuất các giải pháp tối ưu giúp doanh nghiệp không chỉ giữ chân khách hàng mà còn nâng cao hiệu quả kinh doanh. Đây cũng là cơ hội để nhóm phát triển kỹ năng phân tích dữ liệu và áp dụng kiến thức vào thực tiễn.

### 1.2 Mục tiêu đề tài

Phân tích và dự đoán tỷ lệ rời bỏ của khách hàng trong ngành viễn thông bằng cách ứng dụng các kỹ thuật học máy, nhằm hiểu rõ hơn về các yếu tố ảnh hưởng và xu hướng rời bỏ của khách hàng, từ đó đánh giá độ hiệu quả của các mô hình dự đoán khác nhau. Các mục tiêu cụ thể của nghiên cứu được trình bày dưới đây :

- Phân tích các yếu tố ảnh hưởng đến việc rời bỏ của khách hàng trong lĩnh vực viễn thông.
- Ứng dụng các mô hình học máy để dự đoán tỷ lệ rời bỏ của khách hàng.
- So sánh độ chính xác của các mô hình và tìm ra mô hình hiệu quả nhất.
- Trực quan hóa kết quả phân tích và dự đoán bằng cách sử dụng các biểu đồ



- Đưa ra các giải pháp và kiến nghị giúp doanh nghiệp giảm thiểu tỷ lệ rời bỏ khách hàng.

### 1.3 Đối tượng và phạm vi nghiên cứu

- **Đối tượng nghiên cứu:** Các yếu tố liên quan đến hành vi và quyết định rời bỏ dịch vụ của khách hàng trong ngành viễn thông. Các thông tin này bao gồm đặc điểm nhân khẩu học, loại dịch vụ sử dụng, thời gian sử dụng, chi phí hàng tháng, và các yếu tố khác.
- **Phạm vi nghiên cứu:** Dữ liệu được sử dụng trong nghiên cứu là bộ dữ liệu "Telco Customer Churn" từ nền tảng Kaggle, bao gồm khoảng 7.000 mẫu dữ liệu về khách hàng của một công ty viễn thông. Nghiên cứu sẽ tập trung vào việc phân tích và dự đoán tỷ lệ rời bỏ dựa trên các thông tin này.

### 1.4 Công cụ

- **Phần mềm và thư viện**
  - + Ngôn ngữ lập trình: Python (cùng với các thư viện hỗ trợ phân tích và học máy như Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn,...)
  - + Môi trường phát triển: Google Colab
  - + Công cụ trực quan hóa và trình bày kết quả: Canva và
- **Thuật toán và mô hình học máy:**
  - + **K-Nearest Neighbors:** Dự đoán tỷ lệ rời bỏ bằng cách tìm kiếm mối quan hệ giữa các đặc điểm của khách hàng.
  - + **Logistic Regression:** Dự đoán tỷ lệ rời bỏ bằng cách tính xác suất dựa trên các yếu tố như giới tính, độ tuổi, mức độ sử dụng dịch vụ, và các dịch vụ kèm theo,...
  - + **Random Forest:** Dự đoán tỷ lệ rời bỏ dựa trên việc xây dựng nhiều cây quyết định từ các đặc điểm khác nhau của khách hàng.

### 1.5 Ý nghĩa nghiên cứu

Về mặt kinh doanh, nghiên cứu này sẽ cung cấp một cách tiếp cận rõ ràng và có hệ thống để dự đoán tỷ lệ rời bỏ của khách hàng, giúp doanh nghiệp hiểu rõ hơn về các yếu tố tác động và từ đó đưa ra các chiến lược giữ chân khách hàng hiệu quả. Điều này giúp giảm chi phí tiếp thị, tăng doanh thu và cải thiện sự hài lòng của khách hàng.

Về mặt học thuật, đề tài cung cấp một ví dụ thực tiễn về việc ứng dụng các mô hình học máy vào phân tích dữ liệu, giúp nhóm nắm rõ quy trình triển khai một dự án phân tích dữ liệu từ giai đoạn thu thập dữ liệu đến phân tích và trình bày kết quả. Hơn nữa, nghiên cứu này còn giúp nhóm rèn luyện kỹ năng phân tích dữ liệu và áp dụng các mô hình học máy trong thực tiễn.

## 1.6 Cấu trúc báo cáo

**Chương 1 - Tổng quan:** Giới thiệu về dự án, lý do chọn đề tài, mục tiêu nghiên cứu, đối tượng và phạm vi, các công cụ sử dụng, ý nghĩa của nghiên cứu, và cấu trúc tổng quan của báo cáo.

**Chương 2 - Cơ sở lý thuyết :** Tập trung vào lý thuyết liên quan đến churn (rời bỏ khách hàng), những yếu tố ảnh hưởng đến churn trong ngành viễn thông, các lý thuyết liên quan đến các phương pháp phân tích và mô hình dự đoán áp dụng trong dự án, bao gồm mô hình liên quan đến phân tích dữ liệu và dự đoán churn.

**Chương 3 - Phân tích yêu cầu và mô tả dữ liệu:** Xác định và phân tích các yêu cầu của người dùng, mô tả các thuộc tính và cấu trúc dữ liệu từ tập dữ liệu được sử dụng đảm bảo dữ liệu phù hợp với các mục tiêu nghiên cứu đã đặt ra.

**Chương 4 - Phân tích dữ liệu và trình bày kết quả:** Thực hiện phân tích khám phá dữ liệu (EDA), làm sạch và xử lý dữ liệu, xây dựng và đánh giá các mô hình học máy, so sánh kết quả sau đó trình bày kết quả phân tích.

**Chương 5 - Kết luận:** Tổng kết các kết quả phân tích, đưa ra nhận xét đưa ra những kết luận dựa trên phân tích và dự đoán. Đề xuất các biện pháp cho doanh nghiệp viễn thông để giảm tỷ lệ rời bỏ khách hàng dựa trên các yếu tố chính ảnh hưởng đến churn, và gợi ý hướng phát triển tiếp theo cho nghiên cứu.

## CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

---

### 2.1 Tổng quan về Phân tích dữ liệu

#### 2.1.1. Lợi ích của Phân tích dữ liệu trong kinh doanh

Phân tích dữ liệu giúp cải thiện trải nghiệm khách hàng, tối ưu hóa chi phí và nâng cao hiệu quả kinh doanh. Cụ thể, các doanh nghiệp có thể:

- Cải thiện quyết định kinh doanh: Phân tích dữ liệu trong kinh doanh giúp cải thiện quyết định thông qua việc cung cấp thông tin chính xác và kịp thời.
- Tăng cường hiệu suất hoạt động: Cho phép doanh nghiệp tối ưu hóa quy trình làm việc.
- Nâng cao trải nghiệm khách hàng: Việc hiểu rõ nhu cầu và hành vi của khách hàng giúp nâng cao trải nghiệm, từ đó tạo ra sự trung thành và tăng trưởng doanh thu.
- Dự đoán xu hướng thị trường: Phân tích dữ liệu giúp doanh nghiệp dự đoán xu hướng thị trường, từ đó có các chiến lược phù hợp.
- Giảm chi phí: Giúp nhận diện và loại bỏ các yếu tố không cần thiết trong quy trình kinh doanh.

Ngoài ra, việc phân tích dữ liệu còn giúp doanh nghiệp xác định được các nhóm khách hàng có nguy cơ rời bỏ cao và đưa ra chiến lược giữ chân khách hàng kịp thời, cũng như tối ưu hóa chiến lược tiếp thị và tăng cường sự hài lòng của khách hàng.

#### 2.1.2. Quy trình thực hiện dự án phân tích dữ liệu trong kinh doanh

Quy trình phân tích dữ liệu trong kinh doanh là chuỗi các bước được thực hiện nhằm biến dữ liệu thô thành thông tin hữu ích, hỗ trợ doanh nghiệp đưa ra các quyết định chiến lược. Với một quy trình phân tích chặt chẽ, doanh nghiệp có thể khám phá xu hướng, tối ưu hóa hoạt động, và dự báo kết quả kinh doanh.

- 1. Xác định vấn đề kinh doanh:** Bước đầu tiên là xác định rõ vấn đề hoặc mục tiêu mà doanh nghiệp muốn giải quyết. Điều này giúp định hướng dự án và đảm bảo rằng các phân tích sẽ mang lại giá trị cho việc ra quyết định kinh doanh.
- 2. Thu thập dữ liệu:** Sau khi xác định vấn đề, cần thu thập dữ liệu phù hợp từ các nguồn khác nhau, chẳng hạn như cơ sở dữ liệu nội bộ, nguồn dữ liệu bên ngoài, hoặc dữ liệu từ các hệ thống tự động. Dữ liệu có thể bao gồm thông tin khách hàng, dữ liệu giao dịch, dữ liệu thị trường, v.v.

- 3. Phân tích dữ liệu khám phá (EDA):** Trong bước này, các công cụ thống kê và trực quan hóa dữ liệu được sử dụng để tìm hiểu cấu trúc và mối quan hệ giữa các biến trong dữ liệu. EDA giúp xác định các mẫu hình và xu hướng quan trọng.
- 4. Xử lý và làm sạch dữ liệu:** Dữ liệu thô thường chứa nhiều lỗi hoặc giá trị thiếu, vì vậy việc làm sạch, định dạng và chuẩn hóa dữ liệu là cần thiết để đảm bảo tính chính xác và nhất quán cho quá trình phân tích.
- 5. Xây dựng mô hình phân tích:** Dựa trên các mục tiêu và đặc điểm dữ liệu, các mô hình phân tích (chẳng hạn như hồi quy, cây quyết định, hoặc mạng nơ-ron) được xây dựng để giải quyết vấn đề. Các mô hình này sẽ được huấn luyện trên tập dữ liệu để đưa ra các dự đoán hoặc phân tích sâu hơn.
- 6. Đánh giá và tối ưu hóa mô hình:** Sau khi xây dựng mô hình, cần đánh giá độ chính xác và hiệu quả của nó bằng các chỉ số đánh giá mô hình. Nếu kết quả chưa đạt yêu cầu, mô hình sẽ được tối ưu hóa để cải thiện hiệu suất.
- 7. Triển khai và trình bày kết quả:** Sau khi mô hình được hoàn thiện, các kết quả và kết luận sẽ được trình bày một cách rõ ràng, dễ hiểu cho các bên liên quan. Trong một số trường hợp, mô hình sẽ được tích hợp vào hệ thống kinh doanh để hỗ trợ các quyết định hàng ngày.
- 8. Giám sát và cập nhật mô hình:** Môi trường kinh doanh thì sẽ luôn thay đổi, do đó, mô hình phân tích cần được giám sát thường xuyên và cập nhật khi cần thiết để duy trì tính chính xác và giá trị.

## 2.2. Lý thuyết và phương pháp trong phân tích dữ liệu

### 2.2.1. Lý thuyết về sự rời bỏ của khách hàng

Sự rời bỏ của khách hàng (Customer Churn) là hiện tượng khi khách hàng ngừng sử dụng sản phẩm hoặc dịch vụ của một doanh nghiệp trong một khoảng thời gian nhất định. Tỷ lệ rời bỏ của khách hàng cho biết số lượng khách hàng hiện tại không có khả năng tiếp tục mua sản phẩm hoặc dịch vụ từ doanh nghiệp. Đây là một chỉ số quan trọng để đánh giá mức độ hài lòng và sự trung thành của khách hàng.



Hình 1: Sự rời bỏ của khách hàng - Customer churn

(nguồn: FPT IS, 2024)

Công thức tính toán liên quan đến Churn đó là **Tỷ lệ rời bỏ khách hàng** được tính bằng cách lấy **Số lượng khách hàng mất đi** chia cho **Tổng số khách hàng tại thời điểm bắt đầu của kỳ quan sát**, sau đó nhân với 100 để ra kết quả phần trăm.

$$\text{Churn Rate} = \frac{\text{Customers lost in a period}}{\text{Total customers at the beginning of the period}} * 100\%$$

Hình 2: Công thức tính tỷ lệ rời bỏ của khách hàng

(nguồn: Alex Danchenko, 2022 )

Các yếu tố có thể ảnh hưởng tới sự rời bỏ của khách hàng với sản phẩm, dịch vụ:

- Trải nghiệm khách hàng: Khách hàng thường rời bỏ nếu không hài lòng với dịch vụ, trải nghiệm không thuận tiện, hoặc cảm thấy sản phẩm không đáp ứng nhu cầu của họ.
- Quy trình onboard: Nếu khách hàng không được hướng dẫn sử dụng sản phẩm hoặc dịch vụ một cách hiệu quả, họ có thể bỏ cuộc sớm.
- Thiếu tương tác: Khi công ty không duy trì liên lạc và tương tác thường xuyên với khách hàng, điều này có thể tạo cảm giác khách hàng không được trân trọng.
- Thời gian sử dụng ngắn: Khách hàng có thể không thấy giá trị trong sản phẩm nếu họ không sử dụng lâu dài.
- Thay đổi không lường trước: Những thay đổi trong dịch vụ hoặc chính sách có thể gây bối rối và khiến khách hàng rời bỏ.

## 2.2.2. Học máy và mô hình học máy.

Học máy (Machine Learning) là một nhánh của trí tuệ nhân tạo (Artificial Intelligence - AI), tập trung vào việc phát triển các thuật toán và mô hình giúp máy tính

có thể học hỏi từ dữ liệu và cải thiện hiệu suất của chúng mà không cần lập trình cụ thể cho từng tác vụ. Học máy cho phép máy tính tự động cải thiện khả năng dự đoán hoặc quyết định dựa trên dữ liệu mà nó đã được huấn luyện.

- Các thành phần chính của học máy
  - + Dữ liệu: Là nguồn thông tin chính để máy tính học. Dữ liệu có thể là hình ảnh, văn bản, âm thanh, số liệu, v.v.
  - + Mô hình: Là một hàm toán học hoặc một cấu trúc dữ liệu mà máy học sử dụng để đưa ra dự đoán hoặc quyết định.
  - + Thuật toán: Là quy trình hoặc bộ quy tắc để tối ưu hóa mô hình từ dữ liệu. Thuật toán học máy sẽ giúp xác định các tham số của mô hình để tối ưu hóa hiệu suất.
- Các loại học máy
  - + Học có giám sát (Supervised Learning): Dữ liệu được gán nhãn, và mục tiêu là học từ dữ liệu đó để dự đoán nhãn cho dữ liệu mới. Ví dụ: hồi quy, phân loại.
  - + Học không có giám sát (Unsupervised Learning): Dữ liệu không được gán nhãn, và mục tiêu là tìm cấu trúc hoặc mẫu trong dữ liệu. Ví dụ: phân cụm, giảm chiều dữ liệu.
  - + Học tăng cường (Reinforcement Learning): Máy học bằng cách tương tác với môi trường và nhận thưởng hoặc hình phạt, nhằm tối ưu hóa hành động của mình theo thời gian.

Mô hình học máy là chương trình máy tính được sử dụng để nhận diện các mẫu trong dữ liệu hoặc đưa ra dự đoán. Các mô hình học máy được tạo ra từ các thuật toán học máy, trải qua quá trình huấn luyện với dữ liệu có nhãn, không nhãn, hoặc dữ liệu hỗn hợp. Các thuật toán học máy khác nhau sẽ phù hợp với các mục tiêu khác nhau, chẳng hạn như phân loại hoặc dự đoán, vì vậy các nhà khoa học dữ liệu sử dụng các thuật toán khác nhau làm cơ sở cho các mô hình khác nhau. Khi dữ liệu được đưa vào một thuật toán cụ thể, nó sẽ được điều chỉnh để xử lý tốt hơn một nhiệm vụ nhất định và trở thành mô hình học máy.

Mô hình học máy hoạt động thông qua các bước chính sau:

1. **Thu thập và tiền xử lý dữ liệu:** Dữ liệu được thu thập, làm sạch và chuyển đổi để đảm bảo chất lượng tốt trước khi đưa vào mô hình.

2. **Huấn luyện mô hình (Training):** Thuật toán học máy được cung cấp một tập dữ liệu huấn luyện để "học" các mẫu và quan hệ từ dữ liệu đó. Các tham số của mô hình được điều chỉnh để tối ưu hóa kết quả dự đoán.
3. **Đánh giá mô hình (Evaluation):** Mô hình được kiểm tra trên một tập dữ liệu mới (không nằm trong tập huấn luyện) để đánh giá khả năng dự đoán của nó.
4. **Dự đoán và triển khai (Prediction & Deployment):** Sau khi được đánh giá, mô hình có thể được triển khai để dự đoán kết quả từ các dữ liệu chưa biết hoặc mới

### 2.2.3. Các thuật toán máy học

#### 2.2.3.1. K-Nearest Neighbors

K-Nearest Neighbors (KNN) là một thuật toán học máy đơn giản và hiệu quả, được sử dụng chủ yếu cho các bài toán phân loại và hồi quy. KNN dựa trên khoảng cách giữa các điểm dữ liệu để đưa ra dự đoán cho một mẫu mới. Cụ thể, KNN xác định một giá trị K, tức là số lượng “hàng xóm” gần nhất sẽ được xem xét. Khi một mẫu mới cần được phân loại, thuật toán sẽ tính khoảng cách từ mẫu đó đến tất cả các điểm trong tập huấn luyện bằng cách sử dụng các phương pháp như khoảng cách Euclidean. Sau đó, KNN sẽ chọn K điểm gần nhất và dự đoán nhãn cho mẫu mới dựa trên đa số nhãn của nhóm K hàng xóm đó. Nếu là bài toán hồi quy, giá trị dự đoán sẽ được tính bằng trung bình của các giá trị của K “hàng xóm” gần nhất.

Trong K-Nearest Neighbors (KNN), công thức chính thường được dùng để tính khoảng cách giữa các điểm dữ liệu là khoảng cách Euclidean. Công thức này được áp dụng để đo khoảng cách giữa một điểm mới  $X = (x_1, x_2, \dots, x_n)$  và một điểm dữ liệu trong tập huấn luyện  $Y = (y_1, y_2, \dots, y_n)$ :

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^d (x_j - X_{i_j})^2}$$

Hình 3: Công thức khoảng cách Euclidean

Ngoài Euclidean, khoảng cách Manhattan cũng thường được sử dụng, đặc biệt là khi làm việc với dữ liệu có không gian có thể được mô tả dưới dạng lưới (grid-like):

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Hình 4: Công thức khoảng cách Manhattan

Khoảng cách Minkowski là một phiên bản tổng quát của cả khoảng cách Euclidean và Manhattan.

$$d(x, y) = (\sum_{i=1}^n (x_i - y_i)^p)^{\frac{1}{p}}$$

Hình 5: Công thức khoảng cách Minkowski

(nguồn các công thức: *GeeksforGeeks*, updated 2024)

### 2.2.3.2. Logistic Regression - Hồi quy logistic

Hồi quy Logistic là một thuật toán học máy có giám sát, dùng chủ yếu trong các bài toán phân loại nhị phân (ví dụ: xác định một email là spam hay không). Thuật toán này không trực tiếp dự đoán nhãn lớp mà thay vào đó, nó ước tính xác suất rằng một mẫu thuộc về một lớp cụ thể, với giá trị xác suất nằm trong khoảng từ 0 đến 1.

Trọng tâm của hồi quy logistic nằm ở hàm logistic (còn gọi là hàm sigmoid), hàm này chuyển bất kỳ đầu vào nào thành một đầu ra nằm trong khoảng từ 0 đến 1. Hàm logistic có dạng đường cong hình chữ S và được định nghĩa như sau:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Hình 6: Định nghĩa Hàm sigmoid (nguồn: *Viswa*, updated 2024)

Trong đó:

- e: Cơ số của lôgarit tự nhiên (xấp xỉ 2,71828)
- x: Giá trị đầu vào  $\sigma(x)$
- $\sigma(x)$ : Giá trị đầu ra nằm trong khoảng từ 0 đến 1

Trong hồi quy logistic, hàm logistic được kết hợp với một phương trình tuyến tính để mô hình hóa xác suất một mẫu thuộc về lớp dương. Phương trình là:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

$$\Rightarrow P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Hình 7: Phương trình Hồi quy logistic (nguồn: *Viswa*, updated 2024)



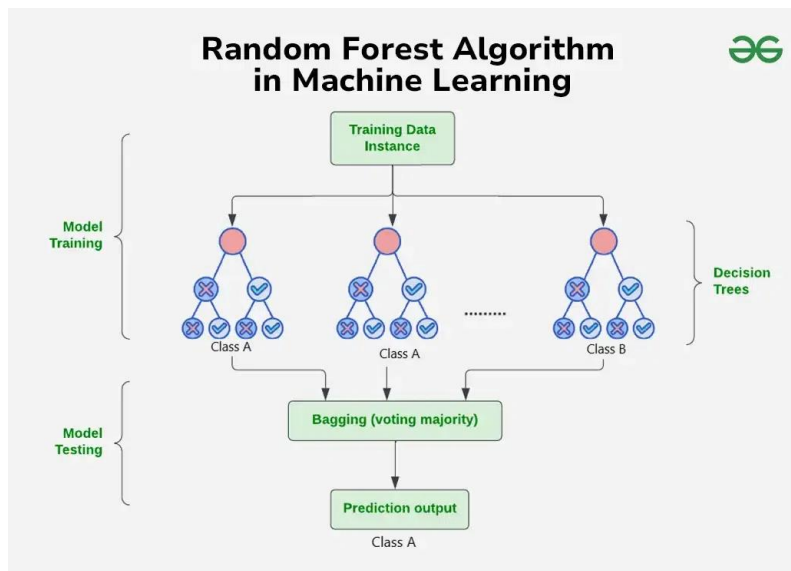
Trong đó:

- $P(Y=1|X)$ : Xác suất của biến mục tiêu bằng 1 với điều kiện các đặc trưng đầu vào  $X$ .
- $\beta_0, \beta_1$  : Các hệ số cần được học từ dữ liệu huấn luyện.
- $X_1, X_2$  : Các đặc trưng đầu vào.

Thuật toán này sử dụng hàm sigmoid để biến đầu ra thành một giá trị xác suất. Nếu xác suất này cao hơn một ngưỡng nhất định (thường là 0,5), mẫu sẽ được phân loại vào một nhóm; nếu thấp hơn ngưỡng, mẫu sẽ thuộc nhóm còn lại. Mặc dù được gọi là "hồi quy," thuật toán này chủ yếu được áp dụng cho các vấn đề phân loại như dự đoán khách hàng có rời bỏ hay không, phân loại email thành spam hoặc không spam, và các ứng dụng tương tự.

#### 2.2.3.3. Random Forest - Rừng ngẫu nhiên

Rừng ngẫu nhiên (Random Forests) là một thuật toán học có giám sát (supervised learning) được xây dựng từ nhiều cây quyết định kết hợp lại nhằm cải thiện độ chính xác trong phân loại và hồi quy. Thuật toán này hoạt động bằng cách tạo ra nhiều cây quyết định từ các mẫu ngẫu nhiên của dữ liệu (bootstrap sampling) và sử dụng một tập hợp ngẫu nhiên các thuộc tính để phân nhánh.



Hình 8: Mô tả thuật toán Random Forest trong Machine Learning

(nguồn: GeeksforGeeks, updated 2024 )

Trong giai đoạn huấn luyện, mỗi cây được xây dựng bằng cách phân chia dữ liệu dựa trên các thuộc tính đã chọn cho đến khi đạt điều kiện dừng. Khi dự đoán cho một mẫu mới, mỗi cây sẽ đưa ra một dự đoán, và kết quả cuối cùng được xác định bằng cách bỏ phiếu (đối với phân loại) hoặc tính giá trị trung bình (đối với hồi quy). Quá trình này giúp cải thiện độ chính xác và ổn định của mô hình, đồng thời giảm nguy cơ overfitting.

#### 2.2.4. Phương pháp phân tích dữ liệu

Trong bối cảnh nghiên cứu, việc lựa chọn phương pháp phân tích dữ liệu phù hợp là rất quan trọng để đảm bảo rằng các kết quả thu được là chính xác và có giá trị. Các phương pháp phân tích dữ liệu có thể được chia thành bốn loại, bao gồm mô tả, chẩn đoán, dự đoán và đề xuất. Trong nghiên cứu này, nhóm sử dụng 3 phương pháp phân tích:

##### **1. Phân Tích Mô Tả (Descriptive Analytics):**

Đây là phương pháp cơ bản nhất trong phân tích dữ liệu, tập trung vào việc tóm tắt và mô tả dữ liệu hiện có để hiểu rõ hơn về tình hình hiện tại. Các kỹ thuật phổ biến bao gồm thống kê mô tả (mean, median, mode), phân tích phân phối tần suất, và trực quan hóa dữ liệu (biểu đồ, đồ thị). Phân tích mô tả giúp nhận diện các mẫu (patterns) cơ bản và xu hướng trong dữ liệu.

##### **2. Phân Tích Chẩn Đoán (Diagnostic Analysis)**

Phân tích chẩn đoán nhằm tìm ra nguyên nhân của các vấn đề hoặc hiện tượng đã xảy ra. Phương pháp này sử dụng các kỹ thuật so sánh, tương quan và hồi quy để xác định mối quan hệ giữa các biến và hiểu rõ lý do dẫn đến các kết quả cụ thể. Phân tích chẩn đoán giúp làm rõ các yếu tố ảnh hưởng đến hiệu suất và các vấn đề trong tổ chức.

##### **3. Phân Tích Dự Đoán (Predictive Analysis):**

Phân tích dự đoán sử dụng các mô hình thống kê và thuật toán học máy để dự đoán các xu hướng hoặc sự kiện trong tương lai dựa trên dữ liệu lịch sử. Mục tiêu của phân tích này là cung cấp thông tin để giúp doanh nghiệp đưa ra quyết định sáng suốt. Các kỹ thuật phổ biến trong phân tích dự đoán bao gồm hồi quy, cây quyết định, và rừng ngẫu nhiên.

## CHƯƠNG 3: PHÂN TÍCH YÊU CẦU CỦA NGƯỜI DÙNG VÀ MÔ TẢ DỮ LIỆU

---

### 3.1. Xác định và phân tích các yêu cầu của người dùng

- **Xác định tỷ lệ rời bỏ hiện tại của khách hàng:** Phân tích dữ liệu lịch sử để xác định tỷ lệ khách hàng rời bỏ trong các khoảng thời gian nhất định. Việc này giúp làm rõ xu hướng và tốc độ mất khách hàng theo thời gian, cung cấp cái nhìn tổng quan về tình hình hiện tại.
- **Phân tích các yếu tố ảnh hưởng đến churn:** Xem xét các yếu tố cụ thể trong dữ liệu, như giá cả dịch vụ, chất lượng dịch vụ, thời gian sử dụng và các đặc điểm cá nhân khác của khách hàng, để tìm ra những yếu tố có thể ảnh hưởng đến quyết định rời bỏ dịch vụ. Điều này cung cấp thông tin hữu ích để xác định và cải thiện các yếu tố có ảnh hưởng tiêu cực đến tỷ lệ giữ chân khách hàng.
- **Dự đoán churn cho từng khách hàng:** Áp dụng các mô hình học máy, bao gồm K-Nearest Neighbors, Random Forest và Logistic Regression, để xây dựng một hệ thống dự đoán churn cho từng khách hàng trong tương lai. Mục tiêu là dự đoán được khả năng rời bỏ của từng khách hàng dựa trên các đặc điểm trong dữ liệu hiện tại, giúp doanh nghiệp đưa ra các biện pháp giữ chân kịp thời.
- **So sánh hiệu suất các mô hình dự đoán:** Đánh giá hiệu quả dự đoán của các mô hình K-Nearest Neighbors, Random Forest và Logistic Regression bằng cách so sánh các chỉ số độ chính xác, độ nhạy và độ đặc hiệu. Việc này giúp xác định mô hình nào là phù hợp và đáng tin cậy nhất trong việc dự đoán churn, từ đó đưa ra lựa chọn tối ưu cho doanh nghiệp.
- **Đưa ra khuyến nghị chiến lược giữ chân khách hàng:** Từ các kết quả phân tích và dự đoán, đề xuất các chiến lược giữ chân khách hàng phù hợp nhằm giảm thiểu tỷ lệ rời bỏ. Các khuyến nghị sẽ được xây dựng dựa trên việc hiểu rõ các yếu tố ảnh hưởng đến churn và đánh giá tính khả thi của các phương pháp khác nhau.

### 3.2. Tổng quan về cơ sở dữ liệu nguồn

#### 3.2.1. Mô tả dữ liệu:

Bộ dữ liệu này được lấy từ Kaggle và chứa thông tin về 7043 khách hàng, với 21 thuộc tính (cột) mô tả đặc điểm của từng khách hàng của một công ty viễn thông. Mỗi hàng trong bộ dữ liệu đại diện cho một khách hàng cụ thể. Bộ dữ liệu cung cấp một cái nhìn sâu sắc về đặc điểm của khách hàng và có thể được sử dụng

cho các phân tích định lượng nhằm phát hiện những yếu tố ảnh hưởng đến khả năng rời bỏ dịch vụ của khách hàng.

Với 21 thuộc tính được mô tả:

*Bảng 3.2.1 Bảng mô tả thuộc tính của bộ dữ liệu*

<b>Thuộc tính</b>	<b>Mô tả</b>	<b>Giá trị tượng trưng</b>	<b>Kiểu dữ liệu</b>
customerID	Mã khách hàng	7795-CFOCW, 5575-GNVDE	Object
gender	Giới tính của khách hàng	male, female	Object
SeniorCitizen	Khách hàng có phải người cao tuổi hay không	0,1	int64
Partner	Khách hàng có phải là đối tác hay không	Yes, No	object
Dependents	Khách hàng có độc lập về tài chính hay không	Yes, No	object
tenure	Thời gian khách hàng đã gắn bó	1, 10 , 45	int64
PhoneService	Khách hàng có sử dụng dịch vụ điện thoại của công ty hay không	Yes, No	object
MultipleLines	Khách hàng có sử dụng nhiều đường dây điện thoại không	Yes, No, No phone service	object
InternetService	Dịch vụ internet mà khách hàng sử dụng	DSL, Fiber optic, No	object
OnlineSecurity	Khách hàng có sử dụng	Yes, No,	object

	dịch vụ bảo mật do công ty cung cấp hay không	No internet service	
OnlineBackup	Khách hàng có đang sử dụng dịch vụ sao lưu trực tuyến không	Yes, No, No internet service	object
DeviceProtection	Khách hàng có sử dụng dịch vụ bảo vệ thiết bị không	Yes, No, No internet service	object
TechSupport	Khách hàng có sử dụng dịch vụ hỗ trợ kỹ thuật không	Yes, No, No internet service	object
StreamingTV	Khách hàng có sử dụng truyền hình trực tuyến không	Yes, No, No internet service	object
StreamingMovies	Khách hàng có sử dụng dịch vụ xem phim trực tuyến không	Yes, No, No internet service	object
Contract	Thời gian khách hàng đã ký kết với công ty	Month-to-month, One year, Two year	object
PaperlessBilling	Khách hàng có sử dụng hóa đơn giấy hay không	Yes, No	object
PaymentMethod	Phương thức thanh toán dịch vụ của khách hàng	Electronic check, Mailed check, Bank transfer (automatic), Credit card	object
MonthlyCharges	Số tiền khách hàng phải trả hàng tháng	19.95 , 84	float64

TotalCharges	Tổng số tiền khách hàng đã thanh toán cho công ty	79.35 , 496.9	object
Churn	Khách hàng đã ngừng sử dụng dịch vụ chưa	Yes, No	object

### 3.2.2 Lựa chọn và trình bày dữ liệu để phân tích yêu cầu của người dùng:

Đề tập trung vào lựa chọn và trình bày dữ liệu trong phân tích rời bỏ của khách hàng, nhóm sẽ lựa chọn các thuộc tính chủ chốt có tác động đáng kể đến hành vi của khách hàng và trình bày theo cách giúp dễ dàng xác định xu hướng.

**Lựa chọn dữ liệu:** Dữ liệu được chia thành các nhóm thuộc tính mang ý nghĩa về đặc điểm cá nhân, dịch vụ sử dụng, hợp đồng và thanh toán, và thông tin tài chính.

**Đặc điểm cá nhân:** Bao gồm các thuộc tính như giới tính (gender), cao tuổi hay không (SeniorCitizen), có đối tác (Partner), và có người phụ thuộc tài chính (Dependents). Các yếu tố này giúp nhận diện các nhóm khách hàng có khả năng rời bỏ cao, chẳng hạn như người cao tuổi hoặc có gia đình có thể có hành vi sử dụng khác biệt.

**Thông tin dịch vụ sử dụng:** Nhóm này gồm các thuộc tính như PhoneService và MultipleLines, giúp xác định khách hàng có sử dụng dịch vụ điện thoại hay nhiều đường dây hay không. Các thuộc tính dịch vụ internet như InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, và StreamingMovies giúp đánh giá mức độ hài lòng và khả năng rời bỏ của khách hàng khi sử dụng các dịch vụ bổ sung.

**Thông tin hợp đồng và thanh toán:** Bao gồm Contract (thời hạn hợp đồng), PaperlessBilling (hóa đơn điện tử), và PaymentMethod (phương thức thanh toán). Những thuộc tính này phản ánh cam kết của khách hàng với công ty. Khách hàng ký hợp đồng dài hạn hoặc chọn tự động thanh toán thường có xu hướng gắn bó hơn.

**Thông tin tài chính:** MonthlyCharges (chi phí hàng tháng) và TotalCharges (tổng chi phí) giúp xác định khả năng duy trì dịch vụ của khách hàng. Chi phí hàng tháng cao có thể là yếu tố ảnh hưởng đến quyết định rời bỏ.

**Trình bày dữ liệu:** Để trực quan hóa, các thuộc tính định tính như gender, Partner, Dependents, InternetService, Contract, và Churn có thể được biểu diễn bằng biểu đồ cột để so sánh tần suất các nhóm khách hàng. Các thuộc tính định lượng như tenure, MonthlyCharges, và TotalCharges có thể dùng biểu đồ histogram hoặc box plot để quan sát phân phối và phát hiện sự khác biệt.

## CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU VÀ KẾT QUẢ

---

### 4.1. Giới thiệu về các công cụ và giải pháp phân tích dữ liệu:

#### 4.1.1. Python

- Mô Tả: Python là một ngôn ngữ lập trình phổ biến trong lĩnh vực phân tích dữ liệu nhờ vào cú pháp dễ hiểu và tính linh hoạt.
- Thư Viện Chính:
  - + Pandas: Thư viện chính để xử lý và phân tích dữ liệu, giúp dễ dàng thao tác với các bảng dữ liệu và thực hiện các phép toán thống kê.
  - + NumPy: Thư viện hỗ trợ tính toán số học với các mảng đa chiều, rất hữu ích trong các phép toán số học phức tạp.
  - + Matplotlib & Seaborn: Các thư viện để tạo biểu đồ và trực quan hóa dữ liệu, giúp dễ dàng phân tích và trình bày kết quả.
  - + Scikit-learn: Thư viện cho học máy, cung cấp nhiều thuật toán và công cụ cho phân tích dự đoán và mô hình hóa dữ liệu.

#### 4.2.2. Google Colab

- Mô Tả: Google Colab là một nền tảng trực tuyến cho phép viết và thực thi mã Python trong trình duyệt, đặc biệt hữu ích cho phân tích dữ liệu và học máy.
- Tính Năng:
  - + Miễn Phí: Cung cấp tài nguyên tính toán miễn phí, bao gồm GPU, giúp thực hiện các tác vụ phân tích dữ liệu lớn một cách hiệu quả.
  - + Tích Hợp với Google Drive: Dễ dàng lưu trữ và chia sẻ các tệp dữ liệu cũng như kết quả phân tích.
  - + Môi Trường Thân Thiện: Hỗ trợ Markdown cho ghi chú và tài liệu, giúp dễ dàng trình bày báo cáo và kết quả phân tích.


### 4.2. Phân tích dữ liệu, khám phá và trực quan hóa:

#### 4.2.1. Khai phá và tiền xử lý dữ liệu:

Trong quá trình khai phá dữ liệu, nhóm đã tiến hành phân tích sơ bộ bộ dữ liệu nhằm xác định cấu trúc và tính toàn vẹn của nó. Nhóm tiến hành trả lời những câu hỏi quan trọng trong việc khai phá và tiền xử lý dữ liệu.

- Dữ liệu có bao nhiêu cột và bao nhiêu hàng?


```
[5] n_rows, n_cols = Telco_data.shape
     print(f'Dataset has {n_rows} rows and {n_cols} columns!')
```

 Dataset has 7043 rows and 21 columns!

Bộ dữ liệu chứa 7043 hàng và 20 cột. Kết quả cho thấy kích thước của bộ dữ liệu không quá lớn, đủ để thực hiện các phân tích và ứng dụng trong phạm vi của môn học. Điều này cho phép nhóm dễ dàng quản lý và thao tác với dữ liệu, đồng thời đảm bảo rằng các phép phân tích có thể được thực hiện một cách hiệu quả và nhanh chóng.

- Có cột nào bị trùng lặp không?


```
[6] any(Telco_data.duplicated())
```

 False

Kết quả phân tích cho thấy không có dữ liệu trùng lặp trong bộ dữ liệu. Tuy nhiên, để hiểu rõ hơn về hành vi của khách hàng, chúng tôi đã tiến hành kiểm tra xem liệu một khách hàng có thể sử dụng các dịch vụ khác nhau vào các thời điểm khác nhau hay không.

Thực tế cho thấy, một customer\_id có thể xuất hiện nhiều lần trong bộ dữ liệu, điều này là hoàn toàn bình thường và hợp lý, vì khách hàng có thể thay đổi dịch vụ hoặc gia hạn hợp đồng trong suốt thời gian sử dụng.

```
[7] dup_id = Telco_data['customerID'].value_counts()
     dup_id = dup_id[dup_id > 1]
     dup_id
```



	count
customerID	

dtype: int64

Sau khi kiểm tra giả thuyết này, nhóm đã xác nhận rằng giả thuyết này không xảy ra, cho thấy rằng bộ dữ liệu của nhóm không chỉ chính xác mà còn phản ánh đúng thực tế hành vi của khách hàng.



- Kiểu dữ liệu các thuộc tính có đang đúng như mô tả không? Cần chỉnh sửa và thay đổi gì?

```
Telco_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   customerID            7043 non-null   object
1   gender                7043 non-null   object
2   SeniorCitizen         7043 non-null   int64
3   Partner               7043 non-null   object
4   Dependents            7043 non-null   object
5   tenure                7043 non-null   int64
6   PhoneService          7043 non-null   object
7   MultipleLines         7043 non-null   object
8   InternetService       7043 non-null   object
9   OnlineSecurity        7043 non-null   object
10  OnlineBackup          7043 non-null   object
11  DeviceProtection      7043 non-null   object
12  TechSupport           7043 non-null   object
13  StreamingTV           7043 non-null   object
14  StreamingMovies       7043 non-null   object
15  Contract              7043 non-null   object
16  PaperlessBilling      7043 non-null   object
17  PaymentMethod         7043 non-null   object
18  MonthlyCharges        7043 non-null   float64
19  TotalCharges          7043 non-null   object
20  Churn                 7043 non-null   object
```

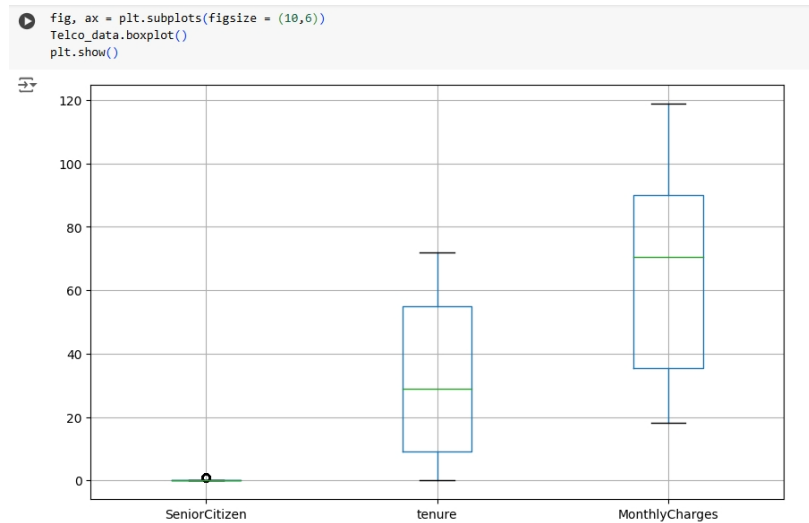
Hình 9: Thông tin chung về bộ dữ liệu

Kết quả cho thấy dữ liệu đang được lưu trữ đúng như mô tả, tuy nhiên chúng ta cần xử lý ở thuộc tính TotalCharges, chuyển kiểu dữ liệu từ object sang float64, vì đây không phải là dữ liệu phân loại mà là một giá trị số thực đại diện cho tổng số tiền khách hàng đã chi tiêu. Việc chuyển đổi này sẽ giúp chúng ta thực hiện các phép toán số học và phân tích thống kê một cách chính xác hơn.

Đối với thuộc tính SeniorCitizen, cần chuyển kiểu dữ liệu từ int64 sang object. Thuộc tính này chỉ đơn giản là một chỉ báo cho việc khách hàng có phải là công dân cao tuổi hay không, và không cần thiết phải lưu trữ dưới dạng số nguyên. Việc chuyển đổi này cũng giúp làm cho việc phân tích và trực quan hóa dữ liệu trở nên rõ ràng hơn, vì các thuộc tính nhị phân thường được thể hiện dưới dạng nhãn (label) thay vì giá trị số.

Sau khi thực hiện những điều chỉnh này, chúng ta sẽ có một bộ dữ liệu đồng nhất và chính xác hơn, giúp cho việc phân tích dữ liệu và xây dựng mô hình trở nên hiệu quả và đáng tin cậy.

- Trả lời cho câu hỏi các giá trị ngoại lai của bộ dữ liệu đang ở đâu?



Hình 10: Boxplot xác định giá trị ngoại lai

Qua kết quả, nhóm nhận thấy rằng thuộc tính TotalCharges có độ phân tán dữ liệu khá lớn. Điều này có thể ảnh hưởng đến tính chính xác của các mô hình học máy nếu không được xử lý đúng cách.

Tuy nhiên, TotalCharges có ý nghĩa quan trọng trong bối cảnh kinh doanh, nên nhóm sẽ thực hiện các bước chuẩn hóa dữ liệu để giảm thiểu ảnh hưởng của các giá trị ngoại lai này. Đặc biệt, nhóm sẽ xem xét sử dụng mô hình như Random Forest, một phương pháp mạnh mẽ có khả năng xử lý các giá trị ngoại lai tốt hơn so với nhiều mô hình khác. Các chi tiết và kết quả cụ thể sẽ được nhóm trình bày trong phần tiếp theo của đồ án

Bắt đầu quá trình tiền xử lý dữ liệu, tiến hành thay đổi kiểu dữ liệu cho phù hợp:

```
Telco_data['TotalCharges'] = pd.to_numeric(Telco_data['TotalCharges'], errors='coerce')
Telco_data['SeniorCitizen'] = Telco_data['SeniorCitizen'].astype('object')
Telco_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
#   Column              Non-Null Count  Dtype
---  -
0   customerID          7043 non-null   object
1   gender               7043 non-null   object
2   SeniorCitizen        7043 non-null   object
3   Partner              7043 non-null   object
4   Dependents           7043 non-null   object
5   tenure               7043 non-null   int64
6   PhoneService         7043 non-null   object
7   MultipleLines        7043 non-null   object
8   InternetService      7043 non-null   object
9   OnlineSecurity       7043 non-null   object
10  OnlineBackup         7043 non-null   object
11  DeviceProtection     7043 non-null   object
12  TechSupport          7043 non-null   object
13  StreamingTV          7043 non-null   object
14  StreamingMovies      7043 non-null   object
15  Contract             7043 non-null   object
16  PaperlessBilling     7043 non-null   object
17  PaymentMethod        7043 non-null   object
18  MonthlyCharges       7043 non-null   float64
19  TotalCharges         7032 non-null   float64
20  Churn                7043 non-null   object
```

Hình 11 Thay đổi kiểu dữ liệu của thuộc tính

- Kiểm tra dữ liệu trống và tiến hành xử lý:

```
Telco_data['TotalCharges'] = pd.to_numeric(Telco_data['TotalCharges'], errors='coerce')
Telco_data['SeniorCitizen'] = Telco_data['SeniorCitizen'].astype('object')
Telco_data.info()
Telco_data.isnull().sum()
```


Có thể thấy TotalCharges đang trống 11 giá trị, điều này có thể lí giải vì khi chuyển từ giá trị Object sang Float64 ( ở trên) rất có thể những giá trị như “abc” sẽ được chuyển thành “NaN”.

Tiến hành điền những giá trị này bằng giá trị trung vị

```
Telco_data['TotalCharges'].fillna(Telco_data['TotalCharges'].median(), inplace=True)
Telco_data.isnull().sum()
```

- Tiến hành mã hóa các dữ liệu phân loại( categorical data) với các nhãn (label encoding):

```
[13] from sklearn.preprocessing import LabelEncoder
cate_data = Telco_data.select_dtypes(['object']).columns
le = LabelEncoder()
for col in cate_data:
    Telco_data[col] = le.fit_transform(Telco_data[col])
Telco_data.head()
```



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	l
0	5375	0	0	1	0	1	0	1	0	0	...	
1	3962	1	0	0	0	34	1	0	0	2	...	
2	2564	1	0	0	0	2	1	0	0	2	...	
3	5535	1	0	0	0	45	0	1	0	2	...	
4	6511	0	0	0	0	2	1	0	1	0	...	

5 rows x 21 columns

Hình 12: Mã hóa dữ liệu

Bây giờ, bộ dữ liệu đã có thể dùng các phương pháp để phân tích dữ liệu và đưa vào mô hình học máy. Điều này sẽ được nhóm trình bày bên dưới.

#### 4.2.2. Phân tích dữ liệu:

##### a. Phân tích mô tả:

Số liệu thống kê ban đầu:

```
Telco_data[['tenure', 'MonthlyCharges', 'TotalCharges']].describe().round(2)
```

	tenure	MonthlyCharges	TotalCharges
count	7043.00	7043.00	7043.00
mean	32.37	64.76	2281.92
std	24.56	30.09	2265.27
min	0.00	18.25	18.80
25%	9.00	35.50	402.22
50%	29.00	70.35	1397.48
75%	55.00	89.85	3786.60
max	72.00	118.75	8684.80

Hình 13: Thống kê mô tả của thuộc tính

Từ số liệu này, ta có thể thấy:

Có sự khác biệt lớn về thời gian gắn bó giữa các khách hàng, một số khách hàng chỉ mới gia nhập và cũng có những người đã gắn bó nhiều năm. Chi phí hàng tháng biến động từ thấp đến cao, có thể thấy các gói dịch vụ hoặc mức độ sử dụng khác nhau. Tổng số tiền chi trả cho thấy mức chi tiêu khá lớn từ một số khách hàng, dự đoán có khả năng là những khách hàng lâu năm hoặc sử dụng dịch vụ ở mức độ cao.

- Tỷ lệ rời bỏ của khách hàng đang như thế nào?

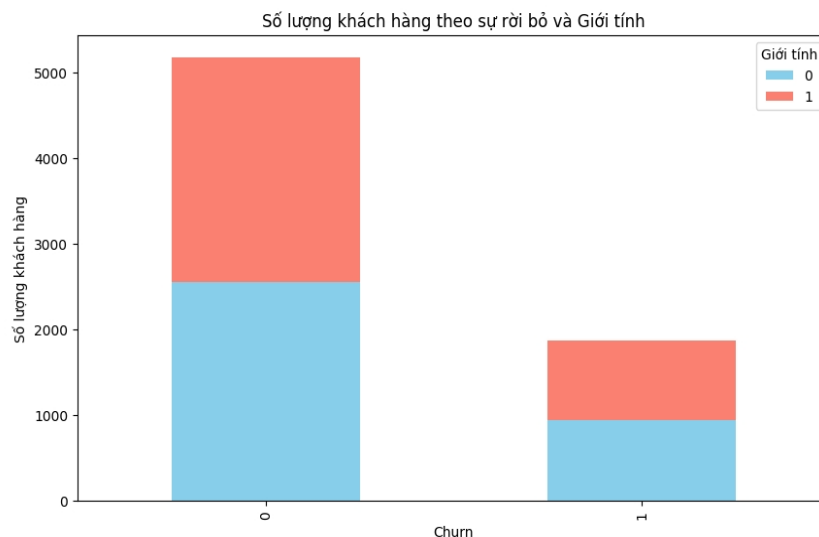
```
total_customers = len(Telco_data)
churned_customers = Telco_data['Churn'].sum()

churn_rate = (churned_customers / total_customers) * 100

print(f"Tỷ lệ rời bỏ: {churn_rate:.2f}%")
```

Tỷ lệ rời bỏ: 26.54%

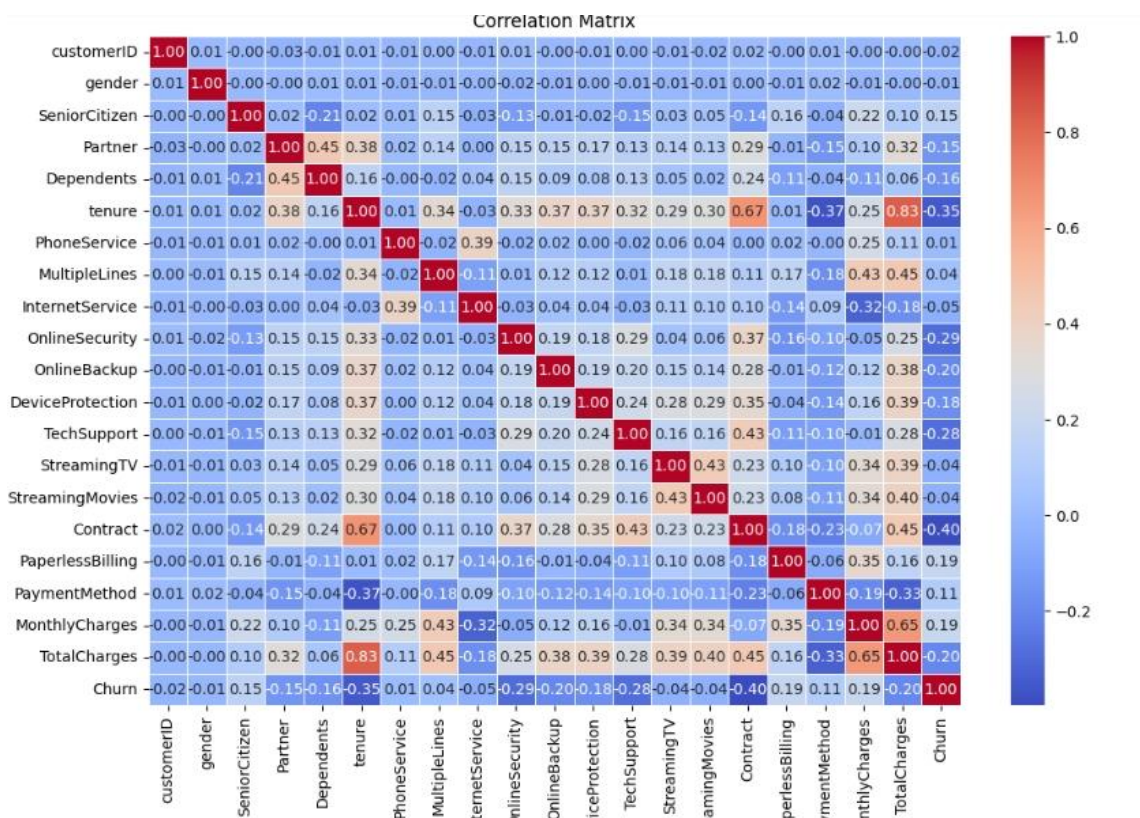
Hình 14: Tỷ lệ rời bỏ của khách hàng



Hình 15: Số lượng khách hàng rời bỏ và giới tính

Tỷ lệ rời bỏ của khách hàng có khoảng hơn 2000 khách hàng, đây là một con số đáng kể trong toàn bộ tập khách hàng. Tỷ lệ nam và nữ cân bằng, cho thấy không có sự tác động của giới tính trong việc sử dụng hay rời bỏ dịch vụ.

Mối quan hệ giữa các biến:



Hình 16: Bảng tương quan các thuộc tính của bộ dữ liệu

Từ ma trận tương quan, có thể thấy rằng hành vi rời bỏ dịch vụ (Churn) của khách hàng chịu ảnh hưởng mạnh mẽ từ các yếu tố liên quan đến thời gian và chi phí. Cụ thể, khách hàng có xu hướng trung thành hơn khi họ đã gắn bó lâu dài với dịch vụ (thể hiện qua tương quan âm với tenure) và có hợp đồng dài hạn (thể hiện qua tương quan âm mạnh với Contract). Tuy nhiên, chi phí hàng tháng chưa hợp lý (MonthlyCharges) có thể làm tăng khả năng rời bỏ. Điều thú vị là các dịch vụ bổ sung như StreamingTV, StreamingMovies, OnlineSecurity và TechSupport có mối tương quan dương với nhau, cho thấy khách hàng thường có xu hướng đăng ký nhiều dịch vụ cùng lúc, và việc này tất yếu dẫn đến tương quan dương giữa số lượng dịch vụ với chi phí hàng tháng. Trong khi đó, yếu tố giới tính (gender) gần như không có ảnh hưởng đến hành vi sử dụng dịch vụ. Những điều này gợi ý rằng chiến lược giữ chân khách hàng nên tập trung vào việc xây

dựng mối quan hệ lâu dài thông qua các hợp đồng dài hạn và gói dịch vụ tích hợp với mức giá hợp lý, thay vì phân khúc theo đặc điểm giới tính.

- Phân bố giá trị của các thuộc tính trong bộ dữ liệu:



Hình 17: Phân phối dữ liệu các thuộc tính

Phân tích tổng hợp từ các biểu đồ phân phối cho thấy một bức tranh toàn diện về cơ sở khách hàng và hành vi sử dụng dịch vụ.

Về đặc điểm giới tính, có sự cân bằng về giới tính, có sự cân bằng về đối khách dịch vụ, đa phần đều độc lập về tài chính và không phải người cao tuổi. Thời gian sử dụng dịch vụ (tenure) phân bố tương đối đều từ 0-72 tháng, với các đỉnh ở hai đầu phổ cho thấy sự đa dạng giữa khách hàng mới và khách hàng trung thành.

Về xu hướng sử dụng dịch vụ, PhoneService là dịch vụ phổ biến nhất, trong khi InternetService thể hiện ba phân khúc rõ rệt (không sử dụng, DSL và Fiber optic). Đặc

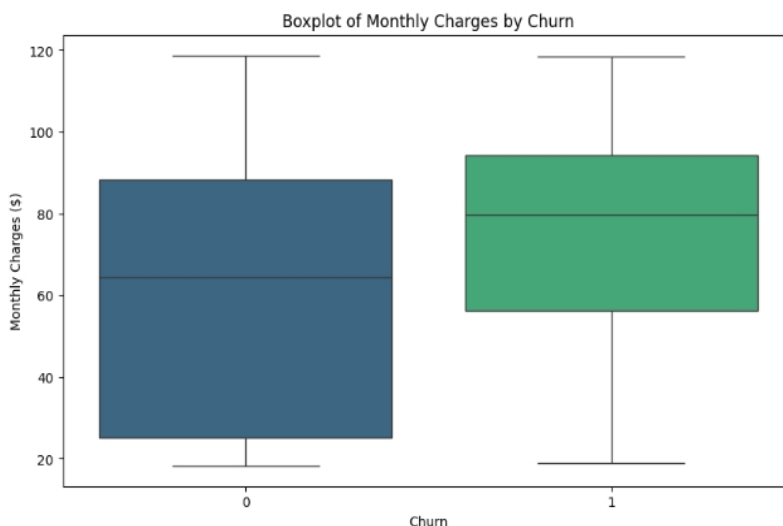


biệt, với các dịch vụ bổ sung như OnlineBackup, DeviceProtection, TechSupport, StreamingTV và StreamingMovies, khách hàng thường có xu hướng hoặc sử dụng toàn bộ hoặc không sử dụng, rất ít trường hợp sử dụng một phần.

MonthlyCharges có phân phối tương đối đều trong khoảng 20-120, trong khi TotalCharges có phân phối lệch phải. Về phương thức thanh toán và hóa đơn, xu hướng sử dụng hóa đơn điện tử (PaperlessBilling) chiếm ưu thế, trong khi các phương thức thanh toán có sự phân bố khá đồng đều. Tỷ lệ Churn thấp hiện tại là một tín hiệu tích cực, tuy nhiên vẫn có cơ hội cải thiện doanh thu thông qua việc thúc đẩy khách hàng sử dụng nhiều dịch vụ bổ sung hơn và phát triển các gói dịch vụ phù hợp với từng phân khúc, đặc biệt là nhóm khách hàng độc thân vốn chiếm đa số trong cơ sở dữ liệu.

#### b. Phân tích chẩn đoán:

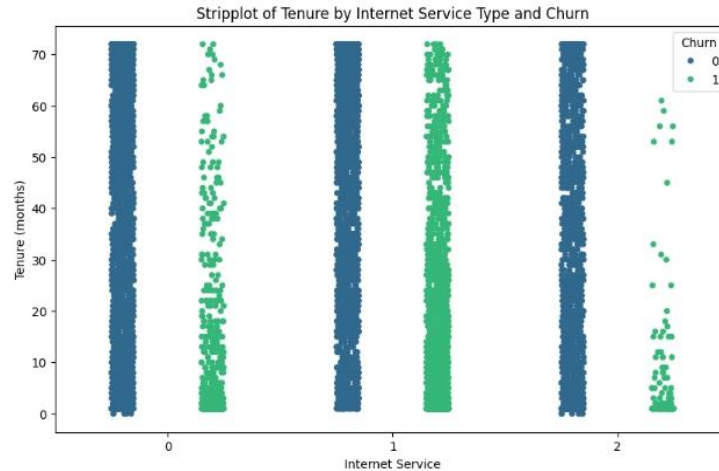
- Tỷ lệ rời bỏ và số tiền khách hàng thanh toán hàng tháng tỉ lệ thuận hay không?



Hình 18: Mối quan hệ giữa sự rời bỏ và số tiền thanh toán hàng tháng

Từ hình ảnh, có thể thấy phí hàng tháng trung bình cao hơn ở nhóm rời đi cho thấy khả năng cao khách hàng có phí cao dễ rời đi hơn. Cả hai nhóm đều có mức phí dao động từ khoảng 20 đến 120 USD, cho thấy không có sự khác biệt lớn về phạm vi phí hàng tháng. Phí hàng tháng ở cả hai nhóm có sự biến động tương đối giống nhau, không có quá nhiều điểm bất thường. Như vậy, khách hàng có phí hàng tháng cao có thể dễ rời đi, doanh nghiệp nên cân nhắc các chương trình giữ chân hoặc ưu đãi cho nhóm khách hàng này để giảm tỷ lệ rời đi.

- Khách hàng rời đi có hài lòng với dịch vụ so với mức giá đã trả ra hay không?



Hình 19: Mối quan hệ giữa sự rời bỏ và dịch vụ

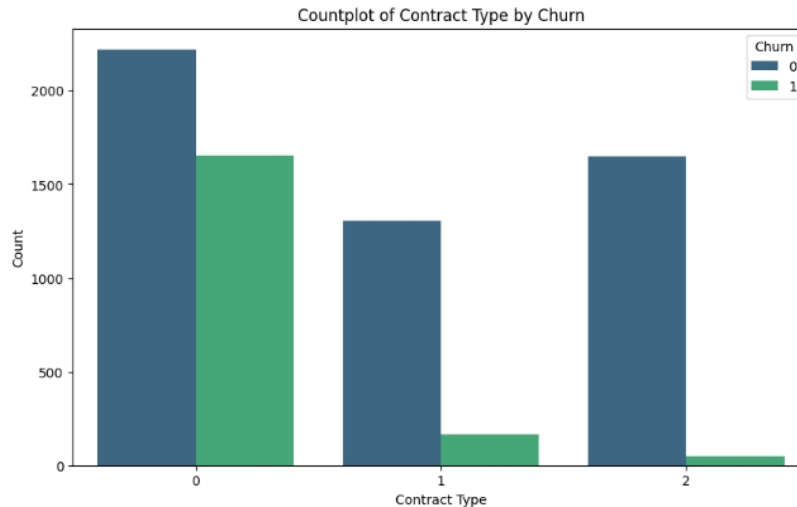
Từ hình ảnh, có thể thấy

- Dịch vụ DSL (0): Khách hàng dùng dịch vụ DSL có tỷ lệ rời đi thấp và có xu hướng gắn bó dài hạn (tenure cao). Điều này cho thấy họ có thể hài lòng với giá trị dịch vụ so với mức giá.
- Dịch vụ Fiber optic (1): Khách hàng sử dụng dịch vụ Fiber optic có tỷ lệ rời đi cao hơn so với DSL, đặc biệt ở nhóm có thời gian sử dụng ngắn (tenure thấp). Điều này có thể chỉ ra rằng họ không cảm thấy giá trị dịch vụ tương xứng với chi phí, dẫn đến việc rời đi sớm.
- Không sử dụng dịch vụ internet (2): Nhóm không sử dụng dịch vụ internet có thời gian sử dụng ngắn và tỷ lệ rời đi rất cao. Đây có thể là nhóm không nhận được giá trị rõ ràng từ dịch vụ hoặc không cần đến dịch vụ internet, nên không tiếp tục duy trì.

Như vậy khách hàng rời đi có khả năng không hài lòng với dịch vụ Fiber optic vì chi phí cao mà giá trị cảm nhận chưa tương xứng, dẫn đến tỷ lệ rời đi cao.

- Tỷ lệ rời bỏ khách hàng phụ thuộc vào thời gian gắn bó và chi phí hàng tháng như thế nào?

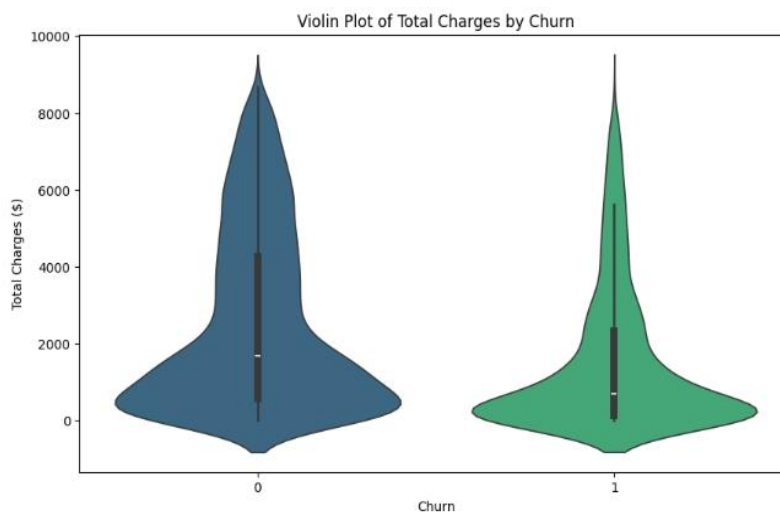




Hình 20: Sự phụ thuộc của thời gian gắn bó và chi phí

Dựa vào hình ảnh, hợp đồng Month-to-month (0) Có tỷ lệ rời bỏ cao nhất Khoảng 1600 khách hàng trong tổng số 2100 khách hàng đã rời bỏ, điều này dễ hiểu khi khách hàng không bị ràng buộc dài hạn để chuyển đổi sang nhà cung cấp khác. Đồng thời, những khách hàng có thời gian sử dụng dịch vụ càng lâu dài thì tỉ lệ rời bỏ càng thấp. Có thể thấy rằng việc thiếu ràng buộc hợp đồng dài hạn khiến khách hàng dễ chuyển đổi, nguyên nhân từ chi phí chuyển đổi thấp ở hợp đồng ngắn hạn hoặc do chính sách giá không hấp dẫn cho hợp đồng ngắn hạn. Như vậy, cần thiết kế ưu đãi đặc biệt cho khách hàng chuyển từ month-to-month lên hợp đồng dài hạn, cần cải thiện trải nghiệm khách hàng trong 3 tháng đầu để tăng tỷ lệ chuyển đổi lên hợp đồng dài hạn, đồng thời xem xét điều chỉnh giá để khuyến khích khách hàng chọn gói dài hạn.

- Tại sao một số khách hàng có tổng số tiền chi trả thấp dù thời gian gắn bó lâu dài?



Hình 21: Mối quan hệ giữa sự rời bỏ và tổng số tiền sử dụng dịch vụ

Từ hình ảnh, dễ thấy có sự phân tách rõ rệt về mức chi phí giữa nhóm rời bỏ (1) và không rời bỏ (0) một số khách hàng trung thành (0) có mức chi tiêu thấp bất thường, và phân phối chi phí không đồng đều ở cả hai nhóm. Điều này xảy ra có thể do một số nguyên nhân như khách hàng đang sử dụng gói dịch vụ cơ bản với giá thấp, hoặc họ được hưởng chương trình khuyến mãi/ưu đãi đặc biệt dài hạn, do công ty thiếu chiến lược up-sell/cross-sell hiệu quả hoặc chính sách giá không linh hoạt theo thời gian sử dụng.

Có thể thấy các yếu tố chính dẫn đến việc rời bỏ khách hàng trong ngành viễn thông bao gồm chi phí hàng tháng cao, thời gian gắn bó ngắn và sự không hài lòng với loại dịch vụ. Khách hàng có hợp đồng Month-to-month có tỷ lệ rời bỏ cao do không bị ràng buộc lâu dài, trong khi những người sử dụng dịch vụ Fiber optic có xu hướng rời bỏ nhiều hơn so với dịch vụ DSL, đặc biệt là khi họ có thời gian sử dụng ngắn. việc thiếu các chương trình giữ chân hiệu quả và các chiến lược upsell/cross-sell cũng góp phần làm tăng tỷ lệ rời bỏ.

#### 4.2.3. Phân tích dự đoán và đánh giá mô hình:

Việc sử dụng các mô hình dự đoán và trực quan hóa kết quả dữ liệu là một phương pháp quan trọng trong việc phân tích và ra quyết định. Các mô hình dự đoán giúp doanh nghiệp và tổ chức dự đoán kết quả tương lai bằng cách phân tích dữ liệu lịch sử và xu hướng hiện tại. Với các kỹ thuật như hồi quy, cây quyết định và học máy, nhóm có thể tạo ra các mô hình dự đoán có độ chính xác cao, phản ánh sát thực tế.

Nhóm đã tiến hành chọn ra 3 mô hình phù hợp nhất cho việc dự đoán và đánh giá các kết quả khả dĩ của bộ dữ liệu, bao gồm: Logistic Regression, Random Forest và K-Nearest Neighbors, sau đó tiến hành trực quan hóa dữ liệu qua các mô hình này.

*So sánh, đánh giá mô hình có độ hiệu quả cao nhất:*

- Đầu tiên, nhóm tiến hành chia tập dữ liệu thành các tập riêng biệt, bao gồm tập huấn luyện, xác thực, và thử nghiệm.

```
[97] x_train, x_temp, y_train, y_temp = train_test_split(x, y, test_size=0.3, random_state=42)
     x_val, x_test, y_val, y_test = train_test_split(x_temp, y_temp, test_size=0.5, random_state=42)
```

- Tỷ lệ của các tập dữ liệu cũng được chia như sau:

training dataset : 70%

validation dataset: 15%

testing dataset: 15%

- Tiếp tục quy trình chuẩn hóa dữ liệu:

```
[73] scaler = StandardScaler()
      x_train = scaler.fit_transform(x_train)
      x_val = scaler.transform(x_val)
      x_test = scaler.transform(x_test)
```

- + *StandardScaler*: Nhằm chuẩn hóa các biến đầu vào để đưa chúng về cùng thang đo (phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1), giúp mô hình dự đoán hiệu quả hơn.
  - + *fit\_transform*: Được áp dụng cho *x\_train* để tính toán và áp dụng chuẩn hóa dựa trên các giá trị trong tập huấn luyện.
  - + *transform*: Được áp dụng cho *x\_val* và *x\_test* để áp dụng cùng một chuẩn hóa, đảm bảo tính nhất quán.
- Khởi tạo một dictionary chứa các mô hình dự đoán với tên gọi tương ứng và các tham số khởi tạo ban đầu:

```
[120] models = {
      "Logistic Regression": LogisticRegression(max_iter=1000),
      "K-Nearest Neighbors": KNeighborsClassifier(),
      "Random Forest": RandomForestClassifier(n_estimators=100)
    }
```

- Thực hiện huấn luyện và đánh giá những mô hình dự đoán kể trên và lưu lại kết quả độ chính xác của mô hình:

```
[184] for name, model in models.items():
      model.fit(x_train, y_train)

      y_pred = model.predict(x_val)

      accuracy = accuracy_score(y_val, y_pred)

      report = classification_report(y_val, y_pred, output_dict=True)

      confusion = confusion_matrix(y_val, y_pred)

      print(f"\n{name} \n Validation Accuracy: {accuracy}")
      print(classification_report(y_val, y_pred))
      print(confusion)
      print("\n")
      print("_"*20)

      results.append({
          'Model': name,
          'Accuracy': accuracy,
      })
      plt.figure(figsize=(3,3))
      sns.heatmap(confusion, annot=True, fmt='d', cmap='Blues', cbar=False,
                  annot_kws={'fontsize': 12, 'fontweight': 'bold'})
      plt.title('Confusion Matrix')
      plt.xlabel('Predicted Labels')
      plt.ylabel('True Labels')
      plt.show()
```

Kết quả đánh giá được thể hiện như sau:

Logistic Regression				
Validation Accuracy: 0.8125				
	precision	recall	f1-score	support
0	0.86	0.90	0.88	778
1	0.67	0.58	0.62	278
accuracy			0.81	1056
macro avg	0.76	0.74	0.75	1056
weighted avg	0.81	0.81	0.81	1056
[[697 81]				
[117 161]]				

Hình 22: Mô hình dự đoán Logistic Regression

K-Nearest Neighbors					Random Forest				
Validation Accuracy: 0.7859848484848485					Validation Accuracy: 0.8115530303030303				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.84	0.87	0.86	778	0	0.84	0.92	0.88	778
1	0.60	0.55	0.57	278	1	0.70	0.50	0.58	278
accuracy			0.79	1056	accuracy			0.81	1056
macro avg	0.72	0.71	0.72	1056	macro avg	0.77	0.71	0.73	1056
weighted avg	0.78	0.79	0.78	1056	weighted avg	0.80	0.81	0.80	1056
[[678 100]					[[719 59]				
[126 152]]					[140 138]]				

Hình 23: Mô hình dự đoán K-Nearest Neighbors và Random Forest

Các mô hình dự đoán đã cho ra các kết quả khá tin cậy, độ chính xác của hai mô hình Logistic Regression và Random Forest là tương đương nhau (khoảng 81%), trong khi K-Nearest Neighbors có độ tin cậy thấp hơn một chút, nhưng bấy nhiêu đó là đủ trong việc đánh giá mức độ trung thành của các tệp khách hàng, sự chênh lệch giữa các yếu tố dù nhỏ nhưng vẫn rất quan trọng. Dễ dàng nhận thấy, tuy độ chính xác chênh lệch không nhiều đối với mô hình Logistic Regression, nhưng Random Forest có thể là một lựa chọn tốt hơn do số lượng lỗi nhầm lẫn (false positives và false negatives) thấp hơn và có khả năng phân biệt khách hàng rời bỏ tốt hơn một chút so với Logistic Regression. Đây là một điểm cộng lớn để doanh nghiệp có thể đưa ra quyết định chuẩn xác, cũng như giảm bớt chi phí cho các trường hợp cảnh báo sai.

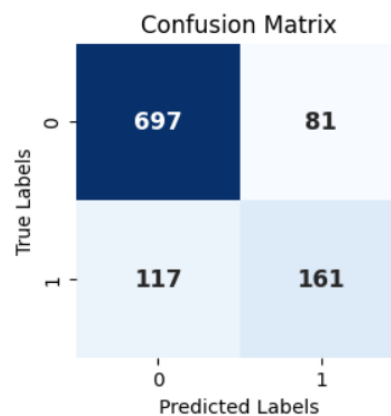
Nhận xét một cách tổng quan rằng, trong ba mô hình dự đoán trên, Logistic Regression là mô hình dự đoán có độ chính xác cao nhất, với độ chính xác 81.25%, một con số khá tốt trong việc dự đoán tổng thể. Tuy nhiên, chỉ số này không phản ánh đầy đủ hiệu quả của mô hình khi quan tâm đến vấn đề rời bỏ doanh nghiệp, bởi để đánh giá đầy đủ, cần xét nhiều yếu tố và điểm số khác nhau.

Cụ thể, trong mô hình này, với Precision của class 1 có tỉ lệ số điểm là 0.67, chỉ có khoảng 67% là chính xác. Tức là đôi khi mô hình sẽ dự đoán nhầm một khách hàng ở lại

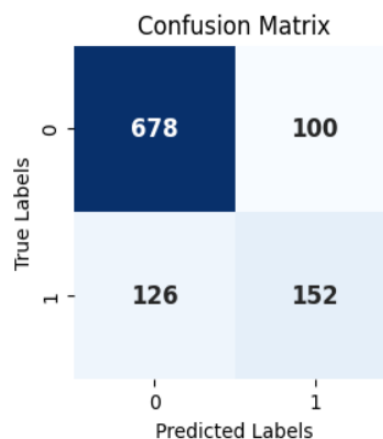
với doanh nghiệp thành rồi bỏ. Điểm Recall của class 0 và class 1 đạt lần lượt là 0.9 và 0.58, điều này có ý nghĩa rằng mô hình có thể dự đoán chính xác đến 90% số khách hàng ở lại, nhưng lại chỉ phát hiện được 58% tổng số khách hàng thực sự rời bỏ doanh nghiệp.

Đây là một hạn chế lớn vì mục tiêu của mô hình này là dự đoán được càng nhiều khách hàng có nguy cơ rời bỏ càng tốt. Đối với F1-score của mô hình này, điểm số của class 0 là 0.88, một con số chấp nhận được đối với tỉ lệ tệp khách hàng trung thành, ngược lại ở class 1 chỉ nằm ở mức 0.62, cho thấy sự cân bằng giữa Precision và Recall chưa cao đối với nhóm khách hàng rời bỏ, cần cải thiện để đạt được độ chính xác và độ bao phủ tốt hơn. Dựa trên kết quả từ các mô hình, Random Forest là lựa chọn hợp lý nhất để hỗ trợ các quyết định giữ chân khách hàng, tối ưu hóa tài nguyên và đảm bảo tính hiệu quả trong các chiến dịch kinh doanh. Dù vậy, Logistic Regression vẫn là mô hình dự đoán có độ chính xác tổng quan tốt nhất trong cả ba mô hình kể trên, thế nên mức hiệu quả của nó đem lại doanh nghiệp vẫn ở thang điểm tương đối hài lòng.

Các ma trận nhầm lẫn, tương ứng với các mô hình dự đoán:



Hình 24: Ma trận nhầm lẫn của mô hình dự đoán Logistic Regression



Hình 25: Ma trận nhầm lẫn của mô hình dự đoán K-Nearest Neighbors

Confusion Matrix

True Labels	0	719	59
	1	140	138
		0	1
		Predicted Labels	

Hình 26: Ma trận nhầm lẫn của mô hình dự đoán Random Forest

Các số liệu được thể hiện trong ma trận nhầm lẫn đã cho nhóm một kết quả thiết thực hơn về tính chính xác của chúng. Cụ thể ở mô hình dự đoán Logistic Regression, có tới 117 khách hàng rời bỏ nhưng bị nhầm lẫn thành ở lại, điều này có thể ảnh hưởng tiêu cực đến việc giữ chân khách hàng, vì những khách hàng này sẽ không được tiếp cận với các chiến dịch khuyến mãi hoặc chăm sóc đặc biệt để ngăn họ rời bỏ. Trong khi ở chiều hướng ngược lại, có 81 trường hợp nhầm lẫn khách hàng ở lại thành rời bỏ, dẫn đến việc cảnh báo sai và có thể gây lãng phí tài nguyên khi chăm sóc những khách hàng này.

#### 4.3. Thảo luận và đánh giá kết quả để hỗ trợ ra quyết định kinh doanh

Việc cả ba mô hình dự đoán đều gặp khó khăn trong việc phát hiện tệp khách hàng rời bỏ với độ chính xác không cao (khoảng 50-58%), nghĩa là các mô hình rất có thể bỏ sót một số lượng lớn khách hàng có nguy cơ rời bỏ. Vì vậy, doanh nghiệp trước tiên cần tập trung vào tệp khách hàng bị đánh giá là có nguy cơ rời đi cao, cần xây dựng các chiến lược chăm sóc và tiếp thị dành riêng cho tệp khách hàng này, chẳng hạn như gửi ưu đãi cá nhân, cải thiện chất lượng dịch vụ hoặc có các dịch vụ chăm sóc đặc biệt, nhằm làm tăng mức độ hài lòng của họ. Bên cạnh đó cũng cần cân nhắc sử dụng mô hình phù hợp nhất như đã phân tích và đánh giá ở trên cho việc này. Lựa chọn đúng công cụ sẽ rất hữu ích trong việc xác định khách hàng mục tiêu cho các chiến dịch giữ chân.

Về vấn đề dự đoán sai tệp khách hàng, doanh nghiệp cũng cần phải giảm thiểu lãng phí tài nguyên trong các chiến dịch tiếp thị, cũng như giữ chân khách hàng. Điều này giúp tối ưu hóa chi phí và tập trung các tài nguyên vào các khách hàng có khả năng rời bỏ cao hơn. Doanh nghiệp cũng có thể dùng nguồn lực đã tiết kiệm để đầu tư vào các chương trình thúc đẩy lòng trung thành của khách hàng hiện tại (như các chương trình ưu đãi hoặc phần thưởng cho khách hàng trung thành).

Cuối cùng, cần tối ưu hóa độ chính xác với chiến lược cải tiến mô hình, cần lựa chọn sáng suốt mô hình dự đoán để đưa ra quyết định, đồng thời tinh chỉnh thêm dữ liệu

hoặc xử lý mất cân bằng dữ liệu để cải thiện độ nhạy cho tệp khách hàng rời bỏ. Sau khi triển khai mô hình vào các quyết định kinh doanh, cần theo dõi hiệu suất của mô hình và đánh giá lại định kỳ. Việc này giúp đảm bảo rằng mô hình luôn cập nhật và phù hợp với các thay đổi của thị trường hoặc hành vi khách hàng. Khi có thêm dữ liệu mới hoặc mô hình cho thấy dấu hiệu sụt giảm hiệu suất, doanh nghiệp nên tái huấn luyện mô hình hoặc thử nghiệm các mô hình khác để đảm bảo độ chính xác và độ nhạy luôn đạt mức tối ưu.

## CHƯƠNG 5: KẾT LUẬN

---

### 5.1. Kết quả và hạn chế:

Dự án đã thành công trong việc phân tích và dự đoán tỷ lệ rời bỏ của khách hàng trong ngành viễn thông thông qua các mô hình học máy. Kết quả phân tích dữ liệu cho thấy rằng các yếu tố như chi phí hàng tháng cao, hợp đồng ngắn hạn và loại dịch vụ (đặc biệt là dịch vụ Fiber optic) là những nguyên nhân chính dẫn đến việc rời bỏ của khách hàng. Mô hình **Logistic Regression** đạt độ chính xác cao nhất với **81.25%**, cung cấp một góc nhìn toàn diện về xu hướng và nguyên nhân của việc rời bỏ khách hàng, đồng thời hỗ trợ doanh nghiệp xác định các chiến lược giữ chân phù hợp.

Mặc dù mô hình Logistic Regression đạt độ chính xác tốt nhất, nhưng độ nhạy đối với khách hàng rời bỏ còn thấp (58%), chưa tối ưu cho mục tiêu phát hiện sớm các khách hàng có nguy cơ rời bỏ. Điều này có thể ảnh hưởng đến hiệu quả của chiến lược giữ chân khách hàng khi có khả năng bỏ sót một số khách hàng tiềm năng rời bỏ. Ngoài ra, do thời gian và tài nguyên hạn chế, nhóm chỉ sử dụng ba mô hình cơ bản và chưa triển khai các phương pháp nâng cao hơn hay các kỹ thuật tối ưu hóa khác, để cải thiện kết quả.

### 5.2. Hướng phát triển chủ đề

Trong tương lai, nhóm có thể mở rộng nghiên cứu theo các hướng sau:

- Xây dựng các dashboard mang tính dự đoán
- Tinh chỉnh mô hình học máy dựa trên các mô hình hiện tại để cải thiện độ chính xác và tin cậy
- Áp dụng các mô hình nâng cao hơn: Thử nghiệm các mô hình phức tạp hơn, như Deep Learning, hoặc các thuật toán tối ưu hóa, để cải thiện độ chính xác và độ nhạy trong việc dự đoán khách hàng rời bỏ.
- Tăng cường thu thập dữ liệu về hành vi khách hàng: Việc bổ sung các yếu tố khác như lịch sử phản hồi từ dịch vụ chăm sóc khách hàng, mức độ hài lòng, hay phản hồi từ khảo sát sẽ giúp mô hình dự đoán chính xác hơn.
- Thực hiện phân tích theo thời gian thực (Real-Time Analytics): Triển khai mô hình dự đoán trong thời gian thực để nắm bắt tức thì những thay đổi và đưa ra phản hồi kịp thời cho các khách hàng có khả năng rời bỏ.



## THAM KHẢO

---

- [1] Wouter Buckinx & Dirk Van den Poel (2003). Customer Base Analysis: Partial Defection of Behaviorally-Loyal Clients in a Non-Contractual FMCG Retail Setting, *European Journal of Operational Research* (pp 164(1):252-268 DOI:10.1016/j.ejor.2003.12.010). Tại [https://www.researchgate.net/publication/24125684\\_Customer\\_Base\\_Analysis\\_Partial\\_Defection\\_of\\_Behaviorally-Loyal\\_Clients\\_in\\_a\\_Non-Contractual\\_FMCG\\_Retail\\_Setting](https://www.researchgate.net/publication/24125684_Customer_Base_Analysis_Partial_Defection_of_Behaviorally-Loyal_Clients_in_a_Non-Contractual_FMCG_Retail_Setting), truy cập ngày 19/10/2024)
- [2] Norizan M. Kassim & Nor Asiah Abdullah (2008). Customer Loyalty in e-Commerce Settings: An Empirical Study, *Electronic Markets* (pp 275 - 292) volume 18, Issue 33. Tại [https://www.researchgate.net/publication/220505715\\_Customer\\_Loyalty\\_in\\_e-Commerce\\_Settings\\_An\\_Empirical\\_Study](https://www.researchgate.net/publication/220505715_Customer_Loyalty_in_e-Commerce_Settings_An_Empirical_Study), truy cập ngày 19/10/2024)
- [3] Alan S. Dick & Kunal Basu (1994). Customer Loyalty: Toward an Integrated Conceptual Framework, *Journal of the Academy of Marketing Science* ( pp 99 - 113, Volume 22, Issue 2) Tại <https://sci-hub.se/10.1177/0092070394222001>, truy cập ngày 19/10/2024)
- [4] Aysha Shereen (2021). Customer Retention Measurement – Metrics to Calculate and Tips to Improve Retention, *Freshworks.com*. Tại <https://www.freshworks.com/live-chat-software/customer-engagement/customer-retention-cheaper-than-new-acquisitions-blog/?q=ecommerce-chat/customer-retention-is-5x-cheaper-than-new-acquisitions-blog/>, truy cập ngày 19/10/2024)
- [5] FPT IS (2024). Customer Churn là gì? Cách quản lý và giảm thiểu Customer Churn, *fpt-is.com* Tại <https://fpt-is.com/goc-nhin-so/customer-churn/>, truy cập ngày 22/10/2024)
- [6] Alex Danchenko (2022). Customer Churn Rate: Definition, Calculation & Tips to Reduce, *Reteno.com*. Tại <https://reteno.com/blog/customer-churn-rate-definition-calculation-tips-to-reduce>, truy cập ngày 22/10/2024)
- [7] GeeksforGeeks, (updated 2024). K-Nearest Neighbor(KNN) Algorithm, *geeksforgeeks.org*. Tại <https://www.geeksforgeeks.org/k-nearest-neighbours/>, truy cập ngày 22/10/2024)
- [8], [9] Viswa (updated 2024). Logistic Regression, *Medium.com*. Tại <https://medium.com/@vk.viswa/logistic-regression-d001d0bce6c7>, truy cập ngày 22/10/2024)

[10] GeeksforGeeks, (updated 2024). Random Forest Algorithm in Machine Learning. geeksforgeeks.org. Tại <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>, truy cập ngày 22/10/2024)