Khóa học Big Data

Capstone Project   
Báo cáo

ⓒ2023 SAMSUNG.

Văn phòng trách nhiệm xã hội doanh nghiệp của Samsung Electronics nắm giữ bản quyền của tài liệu này.

Tài liệu này là tài sản văn học được bảo vệ theo luật bản quyền nên nghiêm cấm việc in lại và sao chép mà không được phép.

Để sử dụng tài liệu này ngoài chương trình giảng dạy của Samsung Innovation Campus, bạn phải nhận được sự đồng ý bằng văn bản từ chủ sở hữu bản quyền.

| COVID-19 Analytics and Prediction |
| --- |

14/08/2024

BD01 - Nhóm 4

Võ Thị Ngọc Châu

Lê Quang An

Nguyễn Minh Gia Bảo

Phạm Thành Đạt

Đặng Vạn Như Ý

Hà Vân Yến

**Nội dung**

1. Giới thiệu

1.1. Thông tin chung

1.2. Động lực, mục tiêu và tiềm năng

1.3. Vai trò của các thành viên

1.4. Kế hoạch thực hiện

2. Thực hiện dự án

2.1. Mô tả bộ dữ liệu

2.2. Tìm kiếm dữ liệu

2.3. Quy trình xử lý dữ liệu

2.4. Ứng dụng

3. Kết quả đạt được

3.1. Tập lệnh và mã tiền xử lý dữ liệu

3.2. Mô tả bộ dữ liệu đã chuyển đổi

3.3. Trực quan hóa dữ liệu

3.4. Huấn luyện mô hình

4. Projected Impact

4.1. Thành tựu và lợi ích của dự án

4.2. Cải tiến trong tương lai

5. Đánh giá và nhận xét các thành viên của nhóm trưởng

6. Đánh giá và nhận xét của giảng viên

# 1. Giới thiệu

## 1.1. Thông tin chung

**Dịch COVID-19 là gì?**

Bệnh viêm đường hô hấp (COVID-19) là bệnh truyền nhiễm do một loại coronavirus mới được phát hiện gây ra.

Hầu hết những người bị nhiễm virus COVID-19 sẽ gặp phải bệnh về đường hô hấp nhẹ đến trung bình và hồi phục mà không cần điều trị đặc biệt. Người già và những người có vấn đề về sức khỏe như bệnh tim mạch, tiểu đường, bệnh hô hấp mãn tính và ung thư có nhiều khả năng phát triển bệnh nặng hơn.

Cách tốt nhất để phòng ngừa và làm chậm sự lây truyền là hiểu biết rõ về bệnh COVID-19, hậu quả mà nó gây ra và cách lây truyền. Bảo vệ bản thân và người khác khỏi nhiễm bệnh bằng cách rửa tay thường xuyên hoặc sử dụng nước rửa tay sát khuẩn và không chạm vào mặt.

Virus COVID-19 lây truyền chủ yếu qua các nước bọt hoặc dịch mũi khi người bị nhiễm bệnh ho hoặc hắt hơi, vì vậy điều quan trọng là bạn cũng phải thực hiện vệ sinh đường hô hấp (ví dụ: ho vào khuỷu tay).

Lưu ý:

Dữ liệu thu được là từ chính phủ Mexico và do đó, phân tích có giá trị đối với người dân Mexico hoặc có thể là Bắc Mỹ. Thống kê và hành vi đại dịch rất khác nhau đối với các nước châu Á so với các nước Bắc Mỹ hoặc châu Âu do tỷ lệ tử vong thấp hơn nhiều so với các nước châu Á.

## 1.2. Động lực, mục tiêu và tiềm năng

### Động lực:

* **Phát hiện sớm:** Phát hiện sớm các trường hợp nghi nhiễm COVID-19 giúp ngăn chặn sự lây lan của dịch bệnh trong cộng đồng, bảo vệ sức khỏe cho mọi người.
* **Phân bổ nguồn lực:** Dự đoán xu hướng dịch bệnh giúp các cơ quan y tế phân bổ nguồn lực hợp lý, như giường bệnh, thiết bị y tế, nhân lực y tế.
* **Nghiên cứu:** Tìm hiểu các yếu tố nguy cơ và các yếu tố liên quan đến việc mắc bệnh COVID-19 để đưa ra các biện pháp phòng ngừa và điều trị hiệu quả hơn.
* **Hỗ trợ quyết định:** Cung cấp thông tin hữu ích cho các nhà hoạch định chính sách để đưa ra các quyết định kịp thời và hiệu quả.

### Mục tiêu:

* **Xây dựng mô hình dự đoán:** Sử dụng các thuật toán học máy (Logistic Regression, Random Forest, Gradient Boosting, XGBoost,...) để xây dựng một mô hình có khả năng dự đoán chính xác khả năng một người mắc bệnh COVID-19 dựa trên các dữ liệu đầu vào như:
  + Thông tin bệnh nhân: Tuổi, giới tính, loại bệnh nhân
  + Bệnh nền: tiểu đường, hen suyễn, tim mạch, viêm phổi, cao huyết áp, ức chế miễn dịch,...
  + Tiếp xúc gần với các bệnh nhân lây nhiễm khác
* **Đánh giá hiệu suất mô hình:** Đánh giá độ chính xác, độ nhạy và độ đặc hiệu của mô hình để đảm bảo mô hình có độ tin cậy cao.
* **Triển khai mô hình:** Triển khai mô hình vào một hệ thống thực tế, ví dụ như một ứng dụng di động hoặc một nền tảng trực tuyến, để người dùng có thể tự đánh giá nguy cơ mắc bệnh.
* **Cập nhật mô hình:** Thường xuyên cập nhật mô hình để đảm bảo tính chính xác và phù hợp với tình hình dịch bệnh thay đổi.

### Các lợi ích tiềm năng:

* **Nâng cao hiệu quả công tác phòng chống dịch:** Giúp các cơ quan y tế phát hiện sớm các ca bệnh, giảm tải cho hệ thống y tế.
* **Bảo vệ sức khỏe cộng đồng:** Giúp người dân chủ động phòng ngừa và bảo vệ bản thân và gia đình.
* **Hỗ trợ quá trình ra quyết định:** Cung cấp thông tin khoa học để hỗ trợ các nhà hoạch định chính sách đưa ra quyết định.
* **Thúc đẩy nghiên cứu:** Góp phần vào việc nghiên cứu về COVID-19 và các bệnh truyền nhiễm khác.

## 1.3. Thành viên nhóm và vai trò của từng thành viên

| Thành viên | Vai trò |
| --- | --- |
| Võ Thị Ngọc Châu | Nhóm trưởng, lên kế hoạch, phân chia công việc, hỗ trợ các bạn thực hiện deadline (nếu có), đánh giá kết quả và góp ý chỉnh sửa, thực hiện các nhiệm vụ được giao. |
| Lê Quang An | Tìm kiếm dữ liệu, góp ý xây dựng dự án và thực hiện các nhiệm vụ được giao. |
| Nguyễn Minh Gia Bảo | Tìm kiếm dữ liệu, góp ý xây dựng dự án và thực hiện các nhiệm vụ được giao. |
| Phạm Thành Đạt | Tìm kiếm dữ liệu, góp ý xây dựng dự án và thực hiện các nhiệm vụ được giao. |
| Đặng Vạn Như Ý | Tìm kiếm dữ liệu, góp ý xây dựng dự án và thực hiện các nhiệm vụ được giao. |
| Hà Vân Yến | Tìm kiếm dữ liệu, góp ý xây dựng dự án và thực hiện các nhiệm vụ được giao. |

## **1.4.** Kế hoạch thực hiện:

| Công việc | Thời gian thực hiện |
| --- | --- |
| 1. Xác định đề tài và thu thập/tìm kiếm dữ liệu | 20/06/2024 - 04/07/2024 |
| 1. Tiền xử lý dữ liệu | 05/07/2024 - 12/07/2024 |
| 1. Trực quan hóa dữ liệu | 13/07/2024 - 20/07/2024 |
| 1. Huấn luyện mô hình bằng nhiều thuật toán | 21/07/2024 - 30/07/2024 |
| 1. Đánh giá và lựa chọn mô hình | 31/07/2024 - 07/08/2024 |
| 1. Viết báo cáo + Thiết kế slide + Thuyết trình thử | 08/08/2024 - 14/08/2024 |

# 2. Thực hiện dự án

## 2.1. Mô tả bộ dữ liệu:

* Dữ liệu ban đầu có 566602 hàng và 23 cột:
* id: Mã số nhận dạng của bệnh nhân.
* sex: Xác định giới tính của bệnh nhân, 1 là nữ và 2 là nam.
* patient\_typeại: Loại bệnh nhân, 1 là không nhập viện và 2 là nhập viện
* entry\_date: Ngày bệnh nhân đến bệnh viện.
* date\_symptoms: Ngày bệnh nhân bắt đầu xuất hiện các triệu chứng
* date\_died: Ngày bệnh nhân qua đời
* intubed: chỉ bệnh nhân có đặt nội khí quản không. 1 biểu thị bệnh nhân đã sử dụng máy thở và 2 biểu thị bệnh nhân không sử dụng.
* pneumonia - viêm phổi: Cho biết bệnh nhân đã bị viêm túi khí hay chưa. 1 là có, 2 là không
* age: Tuổi của bệnh nhân
* pregnancy - mang thai: Cho biết bệnh nhân có thai hay không, “1” là có, “2” là không
* diabetes - tiểu đường: Cho biết bệnh nhân có bị tiểu đường hay không, “1” là có, “2” là không
* copd: Cho biết bệnh nhân có mắc bệnh phổi tắc nghẽn mãn tính (COPD) hay không, “1” là có, “2” là không,
* asthma - hen suyễn: Cho biết bệnh nhân có bị hen suyễn hay không, “1” là có, “2” là không
* inmsupr: Cho biết bệnh nhân có bị ức chế miễn dịch hay không, “1” là có, “2” là không
* hypertension - tăng huyết áp: Cho biết bệnh nhân có bị tăng huyết áp hay không, “1” là có, “2” là không
* other\_disease: Cho biết bệnh nhân có mắc bệnh khác hay không, “1” là có, “2” là không.
* cardiovascular - tim mạch: Cho biết liệu bệnh nhân có mắc bệnh tim hoặc mạch máu hay không, “1” cho có, “2” cho không
* obesity - béo phì: Cho biết bệnh nhân có béo phì hay không, “1” là có, “2” là không
* renal\_chronic - thận mãn tính: Cho biết bệnh nhân có bệnh thận mãn tính hay không, “1” là có, “2” là không
* tobacco - thuốc lá: Cho biết liệu bệnh nhân có phải là người sử dụng thuốc lá hay không, “1” là có, “2” là không.
* contact\_other\_covid: Cho biết liệu bệnh nhân có tiếp xúc với bệnh nhân nhiễm Covid19 khác hay không.
* icu: Cho biết liệu bệnh nhân có được đưa vào Đơn vị Chăm sóc Đặc biệt (ICU) hay không, “1” cho có, “2” cho không
* covid\_res: 1 cho biết người đó dương tính với covid, 2 cho biết người đó là âm tính với covid, 3 cho biết kết quả đang chờ xử lý

## 2.2. Tìm kiếm dữ liệu

Bộ dữ liệu được nhóm tìm kiếm từ trang web Kaggle (https://www.kaggle.com/code/ritikmeena/covid-19-prediction/input)

## **2.**3**.** Quy trình xử lý dữ liệu

* Xử lý dữ liệu trùng
* Loại bỏ các đặc trưng không liên quan, không ảnh hưởng đến kết quả dự đoán
* Xử lý dữ liệu trống
* Do bộ dữ liệu có rất nhiều dữ liệu, nên ta sẽ xét số lượng dữ liệu trống của từng thuộc tính và có cách xử lý tương ứng
* Dữ liệu trống ở cột intubed là 78.5% , icu là 78.5% thiếu -> xóa cột
* Tỷ lệ rỗng < 5%: pneumonia, diabetes, copd, asthma, inmsupr, hypertension, other\_disease, cardiovascular, obesity, renal\_chronic, tobacco.*.* -> xóa các hàng chứa giá trị rỗng.
* Dữ liệu trống ở cột pregnancy là 50% , contact\_other\_covid là 25% -> dự đoán giá trị thiếu bằng thuật toán random forest.
* Kiểm tra và loại bỏ outliers
* Chuẩn hóa thuộc tính phân loại thành kiểu dữ liệu Binary

## 2.4. Ứng dụng

### *Trong lĩnh vực y tế:*

* Phát hiện sớm các ca nhiễm: Mô hình có thể giúp phát hiện sớm các trường hợp nghi ngờ nhiễm COVID-19 dựa trên các triệu chứng và dữ liệu y tế, từ đó giúp ngăn chặn sự lây lan của dịch bệnh.
* Phân loại mức độ nghiêm trọng: Mô hình có thể dự đoán mức độ nghiêm trọng của bệnh, giúp các bác sĩ đưa ra phác đồ điều trị phù hợp và phân bổ nguồn lực y tế hiệu quả.
* Dự báo diễn biến bệnh: Mô hình có thể dự đoán khả năng bệnh nhân tiến triển nặng hoặc tử vong, giúp các bác sĩ theo dõi sát sao và có biện pháp can thiệp kịp thời.
* Hỗ trợ chẩn đoán: Mô hình có thể được sử dụng như một công cụ hỗ trợ chẩn đoán, giúp các bác sĩ đưa ra quyết định chính xác hơn.

### *Trong lĩnh vực quản lý y tế:*

* Lập kế hoạch: Mô hình giúp các cơ quan y tế lập kế hoạch về nguồn lực y tế, giường bệnh, thiết bị y tế và nhân lực để ứng phó với các đợt dịch bệnh.
* Quản lý dịch bệnh: Mô hình có thể được sử dụng để theo dõi và đánh giá hiệu quả của các biện pháp phòng chống dịch bệnh, từ đó điều chỉnh các chiến lược phù hợp.
* Xây dựng các chính sách: Mô hình cung cấp thông tin khoa học để xây dựng các chính sách phòng chống dịch bệnh hiệu quả.

### *Trong lĩnh vực xã hội:*

### Nâng cao nhận thức: Mô hình giúp người dân hiểu rõ hơn về bệnh COVID-19 và các yếu tố nguy cơ, từ đó chủ động phòng ngừa.

### Hỗ trợ cộng đồng: Mô hình có thể được sử dụng để xây dựng các ứng dụng di động hoặc các nền tảng trực tuyến giúp người dân tự đánh giá nguy cơ mắc bệnh và tìm kiếm sự hỗ trợ y tế khi cần thiết.

### *Các ứng dụng tiềm năng khác:*

### Nghiên cứu khoa học: Mô hình có thể được sử dụng để nghiên cứu các yếu tố nguy cơ và các yếu tố liên quan đến việc mắc bệnh COVID-19, từ đó tìm ra các mục tiêu mới cho việc phát triển thuốc và vắc xin.

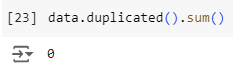
### Phát triển các ứng dụng y tế khác: Mô hình có thể được mở rộng để dự đoán các bệnh truyền nhiễm khác.

# 3. Kết quả đạt được

## 3.1. Tập lệnh và mã tiền xử lý dữ liệu

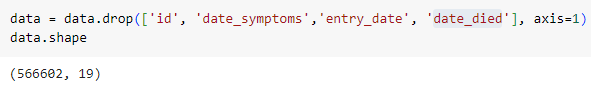
**3.1.1. Dữ liệu trùng:**

Kiểm tra dữ liệu trùng:

  
 => Nhận xét: Bộ dữ liệu không có dữ liệu trùng

### 3.1.2. Loại bỏ các cột không liên quan:

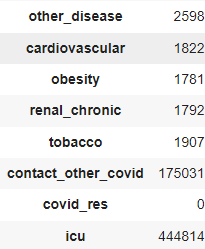
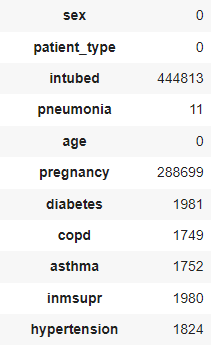
Nhóm hướng đến giải quyết bài toán dự đoán một bệnh nhân có mắc covid-19 hay không. Nhóm nhận thấy rằng các cột id (mã định danh bệnh), entry\_date (ngày nhập viện), date\_symptoms (ngày bệnh nhân xuất hiện các triệu chứng), date\_died (ngày bệnh nhân mất) không liên quan đến khả năng mắc bệnh nên nhóm quyết định loại bỏ các cột này.



### 3.1.3. Dữ liệu trống:

3.1.3.1. Kiểm tra các giá trị trống:



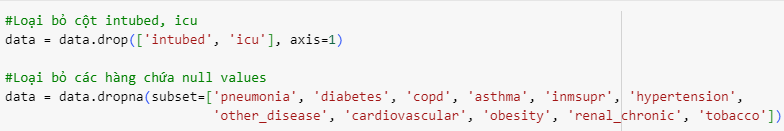


**=> Nhận xét:**

* Tỉ lệ null values ở cột intubed, icu chiếm 78,5% nên nhóm quyết định sẽ bỏ 2 cột này
* Các cột có null values chiếm dưới 5% nhóm sẽ thực hiện loại bỏ các records đó: pneumonia(0.002%), diabetes(0.35%), copd(0.31%), asthma(0.31%), inmsupr(0.35%), hypertension(0.27%), other\_disease(0.46%), cardiovascular(0.32%), obesity(0.32%), renal\_chronic(0.32%), tobacco(0,34%)
* Các cột có null values còn lại đều là biến phân loại và thiếu rất nhiều (pregnancy 50% và contact\_other\_covid 25%) nên nhóm quyết định thực hiện hồi quy để dự đoán các giá trị còn thiếu.

3.1.3.2: Xử lý dữ liệu trống:

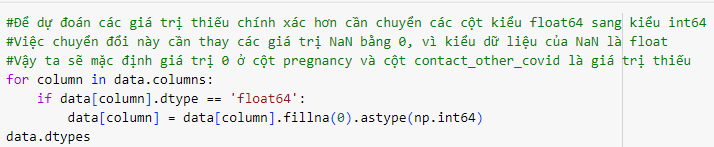
* Xóa cột intubed, icu và xóa các hàng chứa dữ liệu trống ở các cột pneumonia, diabetes, copd, asthma, inmsupr, hypertension, other\_disease, cardiovascular, obesity, renal\_chronic, tobacco:

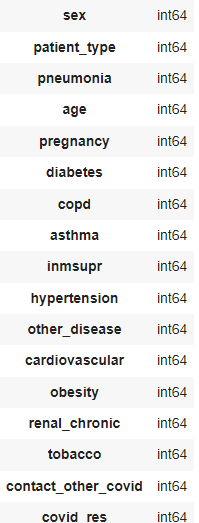




=> Dữ liệu thiếu chỉ còn ở 2 cột pregnancy, contact\_other\_covid. Ta sẽ tiến hành xử lý dữ liệu thiếu 2 cột này tiếp theo.

* Chuyển kiểu dữ liệu từ float64 sang int64:



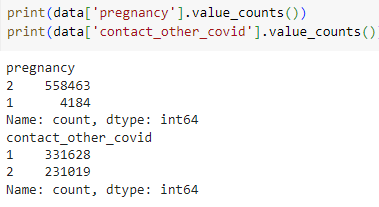


* Sử dụng random forest dự đoán các giá trị trống cho cột pregnancy, contact\_other\_covid:



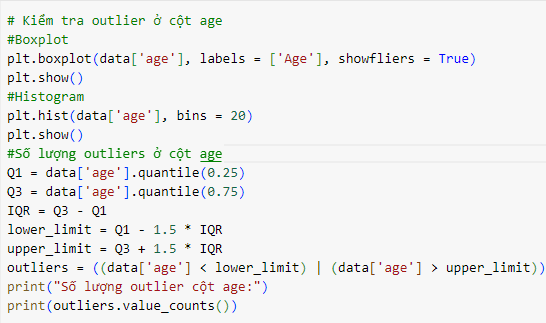


* Cột pregnancy, contact\_other\_covid sau khi xử lý:

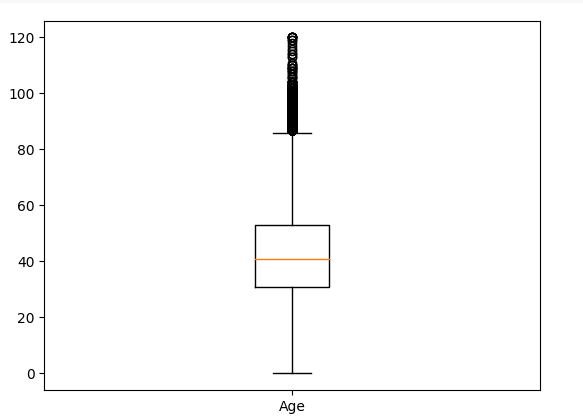


### 3.1.4. Outliers:

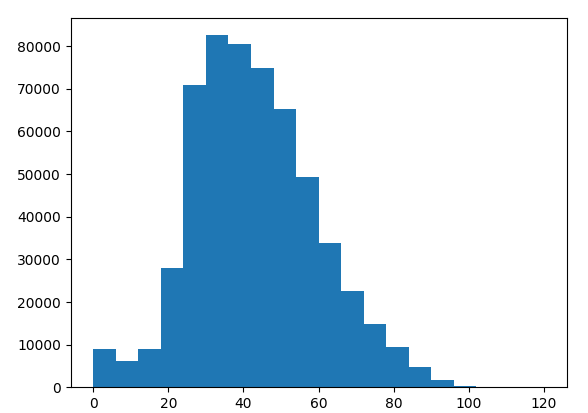
Kiểm tra outliers;

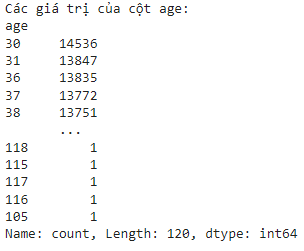
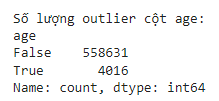


**\* Boxplot:**



**\* Histogram:**





**\*\* Nhận xét:**

* Từ histogram ta thấy dữ liệu cột age bị lệch phải (có điểm ngoại lệ lệch nhiều về bên phải, hoặc “đuôi” của histogram nằm ở bên phải). Từ boxplot ta thấy có khá nhiều điểm được coi là ngoại lệ (4016). Việc có nhiều giá trị ngoại lệ là do các giá trị từ 30 - 50 chiếm phần lớn, và các giá trị lớn hơn 100 rất ít, gần như chỉ có 1 giá trị, cũng như là các giá trị dưới 20, từ đó ảnh hưởng đến các giá trị tứ vị phân.

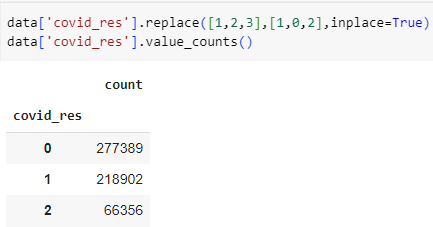
**\*\*Kết luận:** Mặc dù các giá trị outlier là rất nhiều tuy nhiên ta có thể giải thích được chúng và với số lượng outlier nhiều như thế, việc loại bỏ hết các outlier sẽ dẫn đến khi train model có thể không rõ ràng, chính xác vì chúng chứa nhiều tri thức và vẫn có thể là những giá trị hợp lệ

### 3.1.5. Chuẩn hóa dữ liệu cột covid\_res:

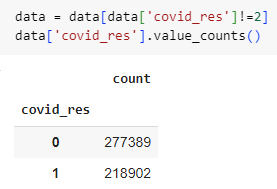
* Giải thích các giá trị ở cột covid\_res

1. positive: mắc bệnh
2. negative: chưa mắc bệnh
3. awaitinga: chờ kết quả

* Để thuận tiện hơn ta sẽ thay 1, 2, 3 thành 1, 0, 2:

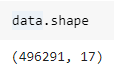


* Vì nhóm đang tập trung vào xem xét các yếu tố bị ảnh hưởng bởi corona, nên trọng tâm chính của chúng ta sẽ là mắc bệnh hoặc không mắc bệnh, do đó chúng ta sẽ bỏ qua giá trị chờ kết quả.
* Bây giờ tôi sẽ xóa tất cả các hàng có giá trị là 2:

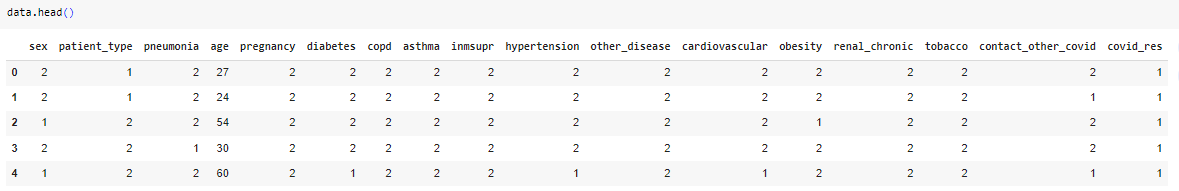


## 3.2. Mô tả bộ dữ liệu đã chuyển đổi

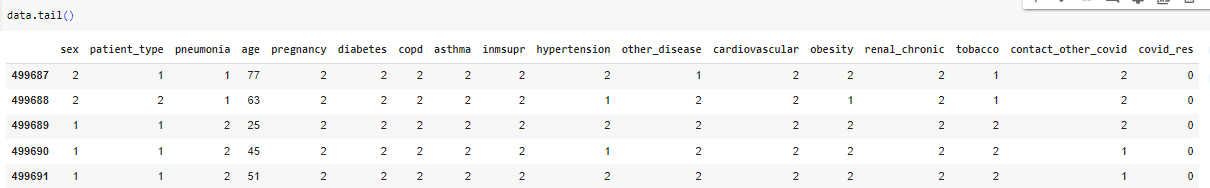
* Số hàng và cột dữ liệu sau xử lý:



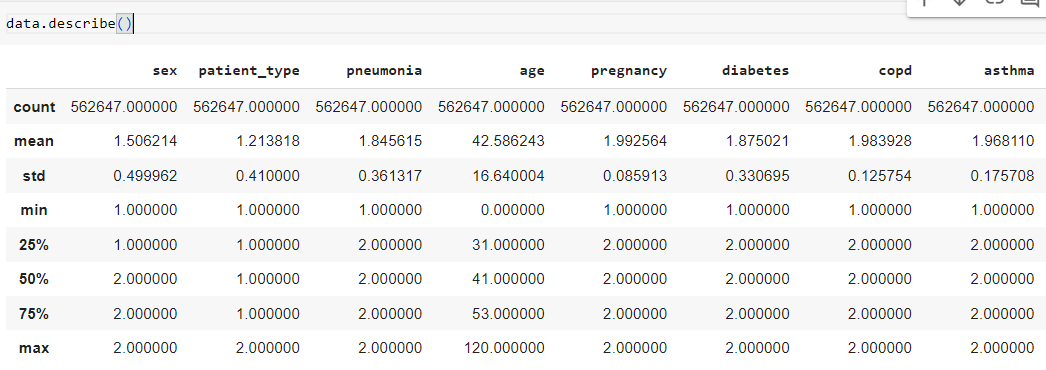
* 5 hàng dữ liệu đầu:

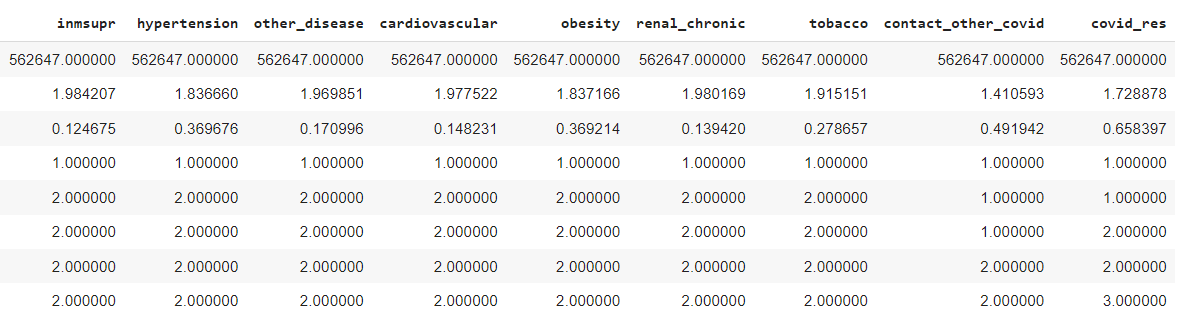


* 5 hàng dữ liệu cuối:



* Các thông số thống kê:

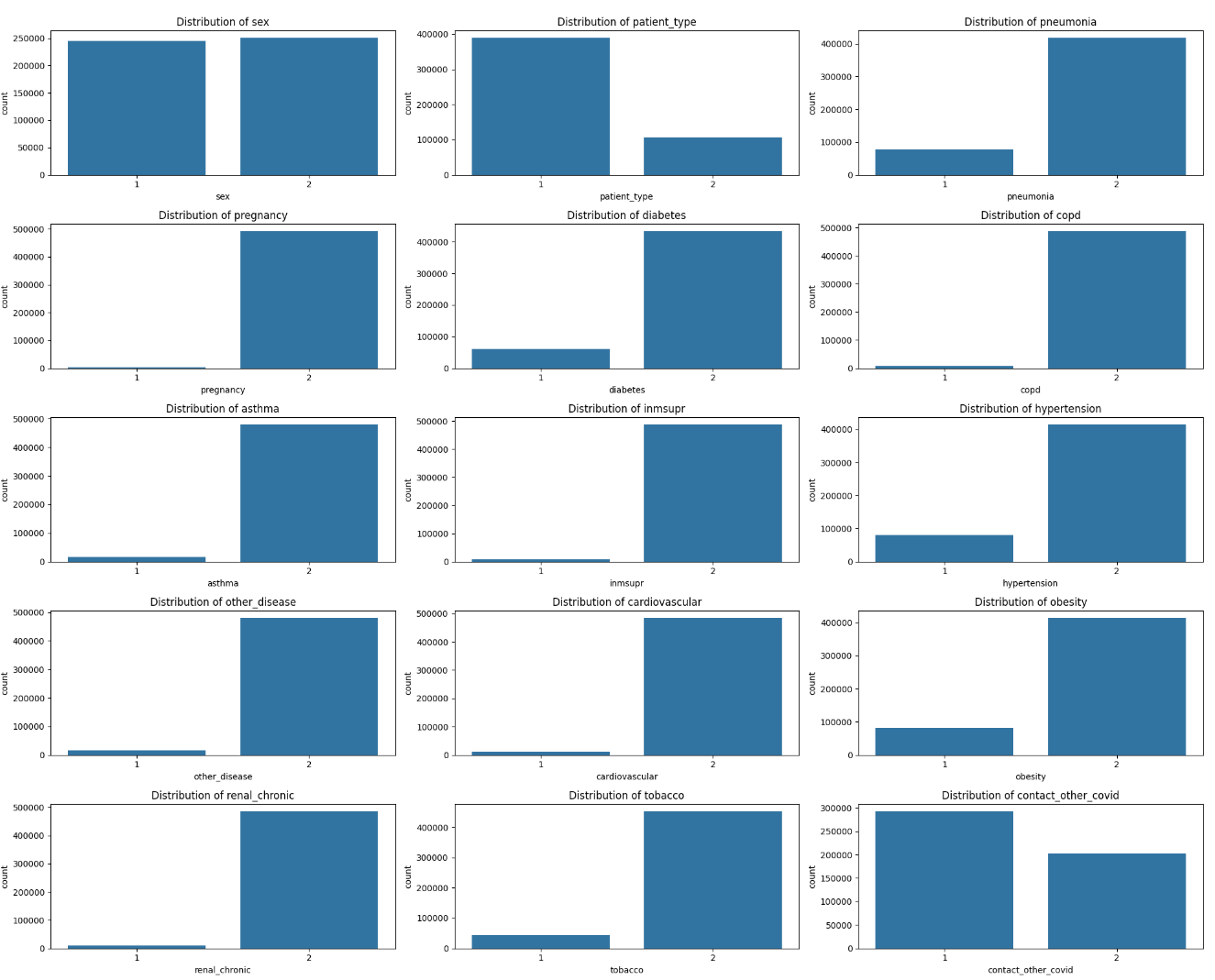




\* Kết luận: Bộ dữ liệu sau xử lý đã khá sạch, có thể phục vụ tốt cho việc trực quan và huấn luyện mô hình

## **3.3.** Trực quan hóa dữ liệu

3.3.1. Ma trận biểu đồ cột (Column chart matrix) thể hiện số lượng bệnh nhân của từng thuộc tính ảnh hưởng đến bệnh

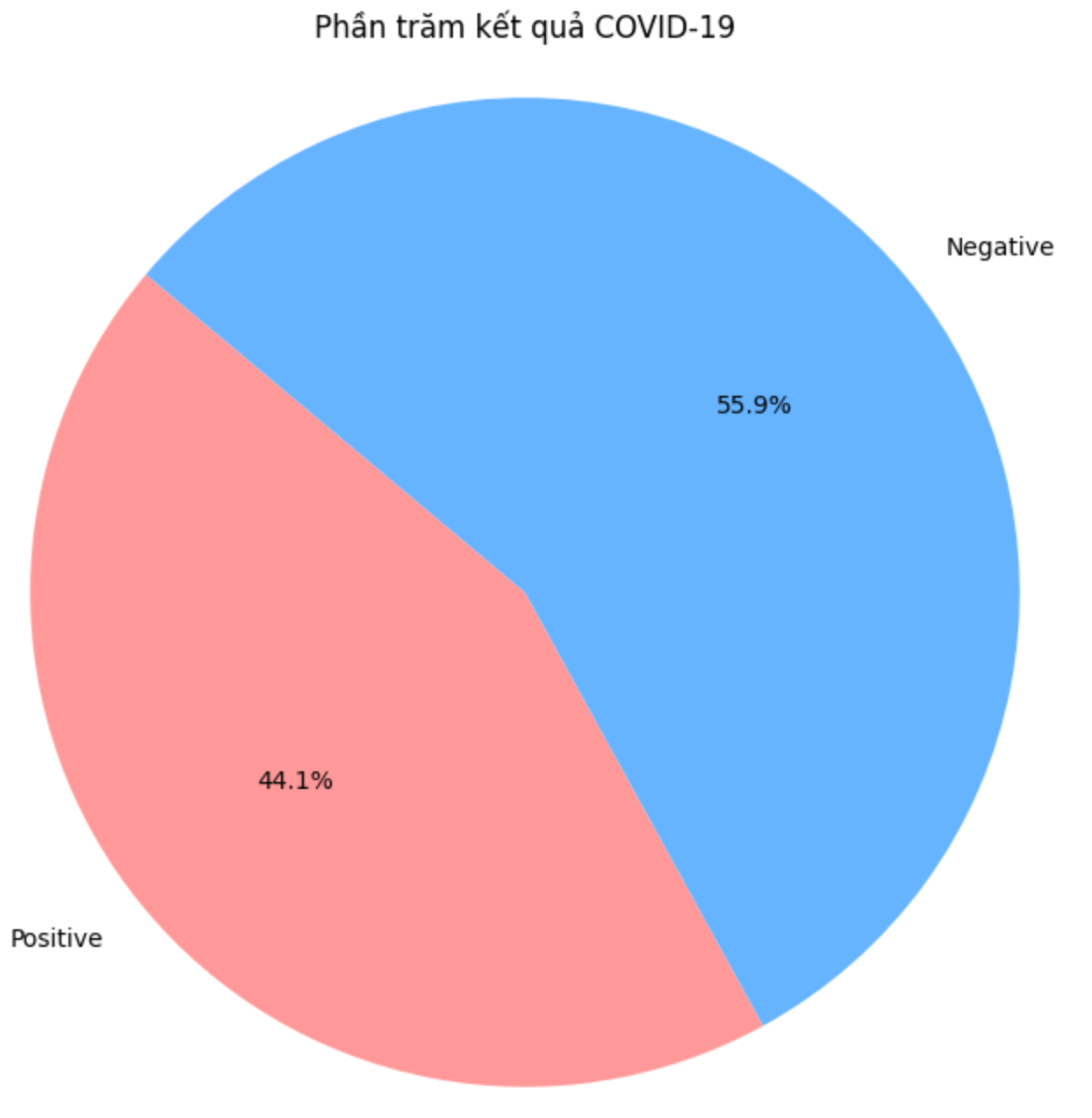


***=> Nhận xét:***

1. **sex**:
   * Số lượng bệnh nhân nam (1) và nữ (2) gần như bằng nhau, với một chút chênh lệch nhỏ.
2. **patient\_type**:
   * Loại bệnh nhân nội trú (1) chiếm phần lớn, trong khi bệnh nhân ngoại trú (2) ít hơn đáng kể.
3. **pneumonia**:
   * Số lượng bệnh nhân không bị viêm phổi (2) chiếm ưu thế, trong khi số lượng bệnh nhân bị viêm phổi (1) ít hơn nhiều.
4. **pregnancy**:
   * Số lượng bệnh nhân không mang thai (2) chiếm phần lớn, với rất ít bệnh nhân mang thai (1).
5. **diabetes**:
   * Số lượng bệnh nhân không bị tiểu đường (2) rất lớn, trong khi bệnh nhân bị tiểu đường (1) chiếm tỷ lệ nhỏ.
6. **copd**:
   * Số lượng bệnh nhân không bị COPD (2) cũng rất lớn, với bệnh nhân bị COPD (1) chiếm tỷ lệ nhỏ.
7. **asthma**:
   * Số lượng bệnh nhân không bị hen suyễn (2) chiếm ưu thế, với bệnh nhân bị hen suyễn (1) ít hơn đáng kể.
8. **inmsupr**:
   * Số lượng bệnh nhân không bị ức chế miễn dịch (2) rất lớn, trong khi bệnh nhân bị ức chế miễn dịch (1) chiếm tỷ lệ nhỏ.
9. **hypertension**:
   * Số lượng bệnh nhân không bị tăng huyết áp (2) chiếm phần lớn, với bệnh nhân bị tăng huyết áp (1) ít hơn đáng kể.
10. **other\_disease**:
    * Số lượng bệnh nhân không có bệnh lý khác (2) chiếm phần lớn, với bệnh nhân có bệnh lý khác (1) ít hơn nhiều.
11. **cardiovascular**:
    * Số lượng bệnh nhân không bị bệnh tim mạch (2) rất lớn, với bệnh nhân bị bệnh tim mạch (1) chiếm tỷ lệ nhỏ.
12. **obesity**:
    * Số lượng bệnh nhân không bị béo phì (2) chiếm ưu thế, với bệnh nhân bị béo phì (1) ít hơn đáng kể.
13. **renal\_chronic**:
    * Số lượng bệnh nhân không bị bệnh thận mãn tính (2) rất lớn, trong khi bệnh nhân bị bệnh thận mãn tính (1) chiếm tỷ lệ nhỏ.
14. **tobacco**:
    * Số lượng bệnh nhân không hút thuốc (2) chiếm phần lớn, với bệnh nhân hút thuốc (1) ít hơn nhiều.
15. **contact\_other\_covid**:
    * Số lượng bệnh nhân không tiếp xúc với người bị COVID-19 (2) chiếm ưu thế, với bệnh nhân có tiếp xúc (1) ít hơn đáng kể.

=> Các biểu đồ trên cho thấy rằng trong hầu hết các thuộc tính, số lượng bệnh nhân không có tình trạng bệnh lý chiếm tỷ lệ lớn hơn nhiều so với số lượng bệnh nhân có tình trạng bệnh lý.

3.3.2. Biểu đồ tròn (Pie chart) thể hiện phần trăm số lượng người mắc covid và người không mắc covid

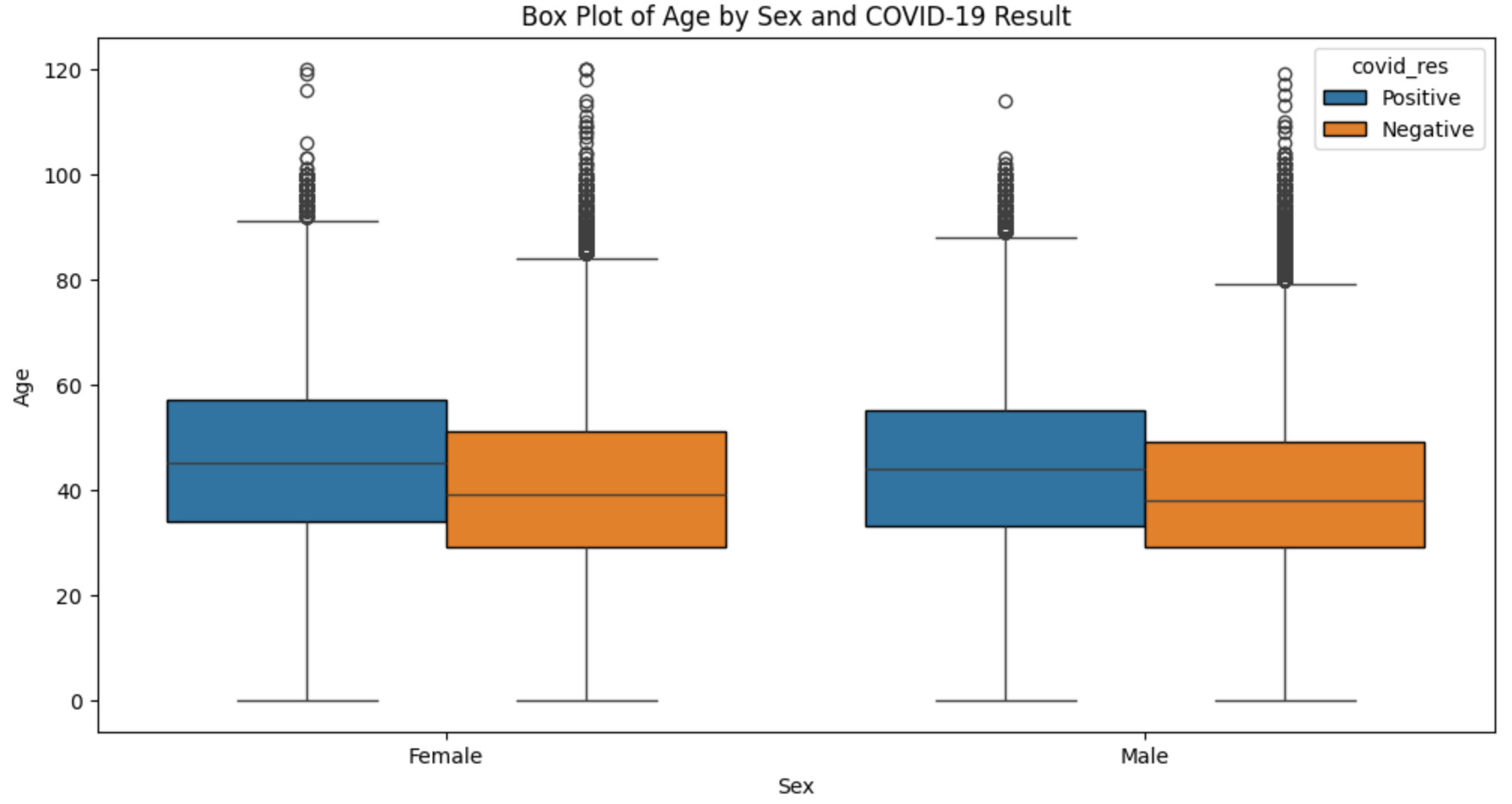
****

***=> Nhận xét:***

* 55.8% số người có kết quả âm tính (Negative) với COVID-19.
* 44.2% số người có kết quả dương tính (Positive) với COVID-19.

=> Tỷ lệ số lượng người mắc COVID-19 chiếm gần một nửa (44.2%) so với tổng số người được xét nghiệm, trong khi đó số lượng người không mắc COVID-19 chiếm hơn một nửa (55.8%). Điều này cho thấy tình hình lây nhiễm COVID-19 vẫn còn đáng kể trong mẫu xét nghiệm này.

3.3.3. Biểu đồ hộp (Box plot) mối quan hệ giữa giới tính của bệnh nhân theo kết quả xét nghệm covid



***=> Nhận xét:***

1. Độ tuổi trung bình:

* Ở cả nam và nữ, độ tuổi trung bình của những người có kết quả dương tính và âm tính đều tương đối gần nhau.
* Cụ thể, độ tuổi trung bình của nam giới có kết quả dương tính là khoảng 45 tuổi, còn kết quả âm tính là gần 50 tuổi.
* Đối với nữ giới, độ tuổi trung bình của cả hai nhóm kết quả dương tính và âm tính cũng gần tương đương, dao động quanh mức 45-50 tuổi.

1. Phân phối độ tuổi:

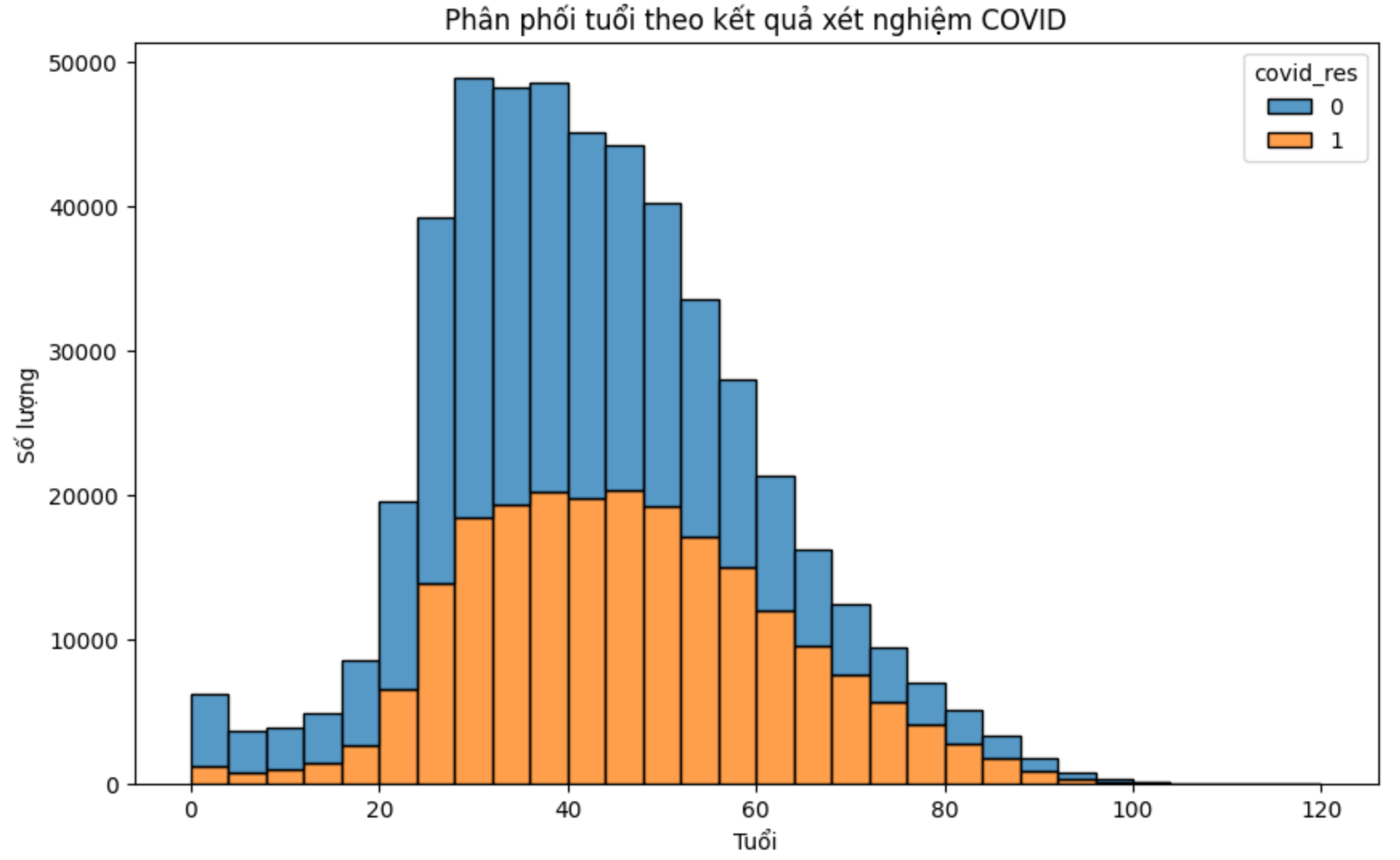
* Phân phối độ tuổi ở cả nam và nữ đều trải dài từ rất trẻ (dưới 20 tuổi) cho đến người cao tuổi (trên 100 tuổi).
* Ở cả hai giới, có những trường hợp ngoại lệ (outliers) với độ tuổi trên 100, cho thấy một số ít bệnh nhân rất lớn tuổi.

1. So sánh giữa giới tính: Phân phối độ tuổi giữa nam và nữ có sự tương đồng về mức độ trải rộng. Tuy nhiên, có vẻ như ở cả hai giới, những người có kết quả âm tính có độ tuổi trung bình hơi cao hơn so với những người có kết quả dương tính.

**Tóm lại:**

* Độ tuổi trung bình của những người có kết quả xét nghiệm dương tính và âm tính với COVID-19 là khá tương đương ở cả hai giới.
* Sự khác biệt không quá lớn giữa hai giới, cho thấy kết quả xét nghiệm COVID-19 không quá phụ thuộc vào độ tuổi hay giới tính, mà có thể còn phụ thuộc vào nhiều yếu tố khác như tình trạng sức khỏe và khả năng tiếp xúc với virus.

3.3.4. Biểu đồ mật độ (Histogram) mối quan hệ giữa tuổi và kết quả xét nghiệm covid



***=>Nhận xét:***

1. Phân phối tuổi tổng quát:

* Độ tuổi của người được xét nghiệm trải rộng từ trẻ em dưới 20 tuổi đến người lớn tuổi trên 100 tuổi.
* Phần lớn người xét nghiệm tập trung ở độ tuổi từ 20 đến 60, với đỉnh phân phối nằm ở khoảng từ 30 đến 40 tuổi.

1. Kết quả xét nghiệm theo độ tuổi:

* Ở nhóm tuổi dưới 20 và trên 60, số lượng người dương tính với COVID-19 có xu hướng giảm so với nhóm tuổi từ 20 đến 60.
* Nhóm tuổi từ 20 đến 60 có tỷ lệ dương tính cao hơn, đặc biệt là trong độ tuổi từ 30 đến 50, điều này có thể do họ nằm trong nhóm tuổi lao động, có nhiều khả năng tiếp xúc với virus hơn.

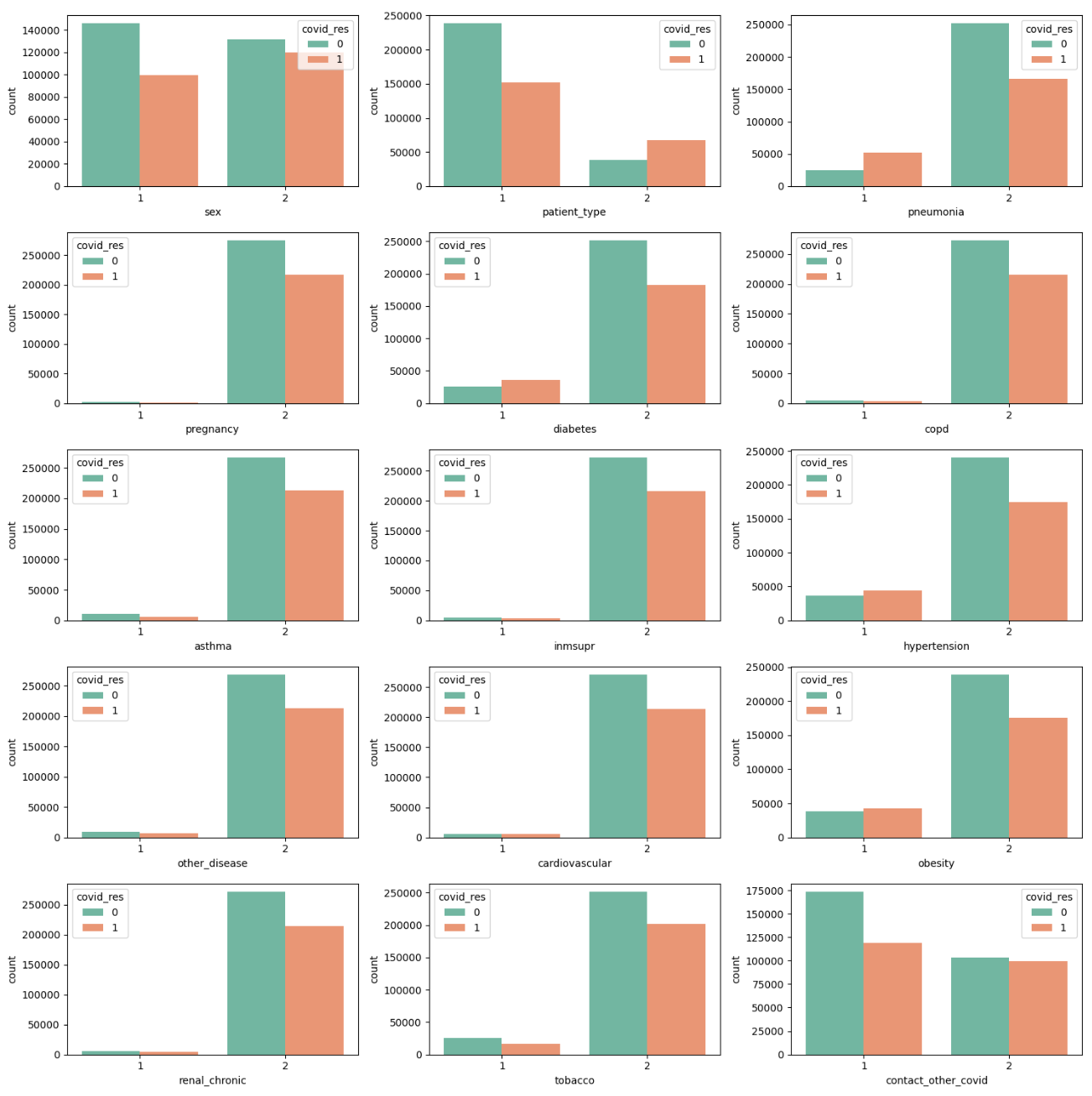
1. So sánh giữa dương tính và âm tính:

* Số lượng người có kết quả âm tính nhiều hơn đáng kể so với số lượng người có kết quả dương tính trong mọi độ tuổi.
* Mặc dù có một tỷ lệ người dương tính với COVID-19 ở tất cả các độ tuổi, nhưng tỷ lệ này giảm dần khi độ tuổi tăng lên, đặc biệt là sau 60 tuổi.

**Tóm lại:**

* Độ tuổi từ 30 đến 50 có tỷ lệ mắc COVID-19 cao hơn so với các độ tuổi khác, điều này có thể phản ánh mức độ tiếp xúc xã hội và hoạt động kinh tế của nhóm tuổi này.
* Tuy nhiên, ở các nhóm tuổi khác, tỷ lệ dương tính vẫn tồn tại nhưng ở mức thấp hơn.

3.3.5. Ma trận biểu đồ cột thể hiện số lượng người mắc covid với người không mắc covid theo từng thuộc tính

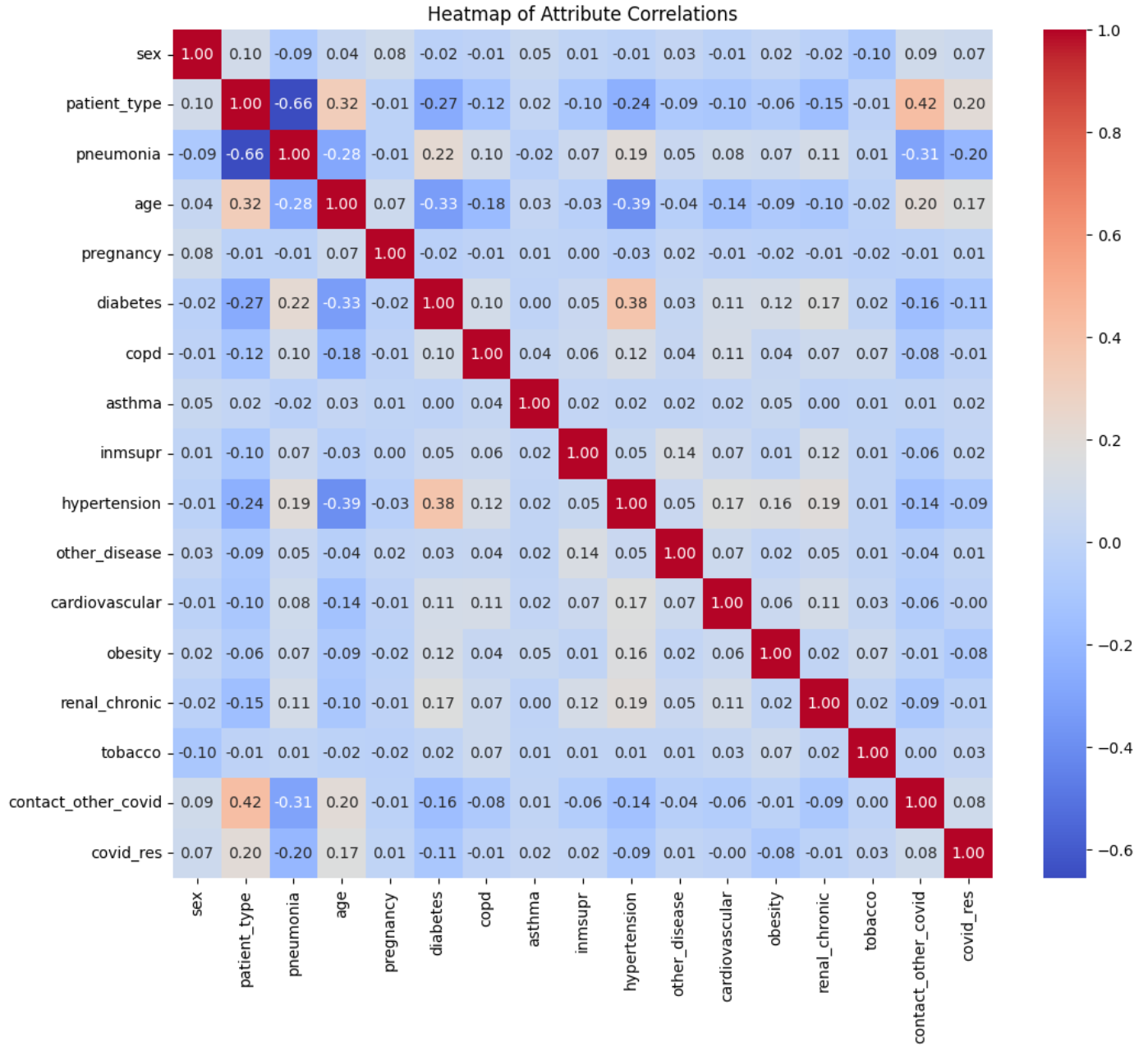


***=> Nhận xét:***

1. **Giới tính (sex)**: Số lượng nam giới (1) và nữ giới (2) mắc COVID-19 khá đồng đều, tuy nhiên, tỷ lệ mắc COVID-19 ở nữ giới có vẻ cao hơn một chút so với nam giới.
2. **Loại bệnh nhân (patient\_type)**: Bệnh nhân nội trú (2) có tỷ lệ mắc COVID-19 cao hơn nhiều so với bệnh nhân ngoại trú (1). Điều này cho thấy bệnh nhân nội trú có nguy cơ mắc COVID-19 cao hơn.
3. **Viêm phổi (pneumonia)**: Những người bị viêm phổi (2) có tỷ lệ mắc COVID-19 cao hơn đáng kể so với những người không bị viêm phổi (1). Điều này cho thấy viêm phổi là một yếu tố nguy cơ cao liên quan đến COVID-19.
4. **Mang thai (pregnancy)**: Những người không mang thai (2) có tỷ lệ mắc COVID-19 cao hơn so với những người mang thai (1). Tuy nhiên, số lượng người mang thai trong mẫu khá ít, vì vậy cần thận trọng khi đưa ra kết luận.
5. **Bệnh tiểu đường (diabetes)**: Những người bị tiểu đường (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị tiểu đường (1). Điều này cũng là một dấu hiệu cho thấy bệnh tiểu đường có thể là yếu tố nguy cơ đối với COVID-19.
6. **Bệnh phổi tắc nghẽn mãn tính (COPD)**: Những người không bị COPD (2) có số lượng mắc COVID-19 cao hơn, tuy nhiên tỷ lệ người bị COPD (1) mắc COVID-19 cũng đáng chú ý.
7. **Hen suyễn (asthma)**: Số lượng người không bị hen suyễn (2) có tỷ lệ mắc COVID-19 cao hơn so với những người bị hen suyễn (1).
8. **Suy giảm miễn dịch (immsupr)**: Những người bị suy giảm miễn dịch (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị suy giảm miễn dịch (1).
9. **Tăng huyết áp (hypertension)**: Những người bị tăng huyết áp (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị tăng huyết áp (1).
10. **Bệnh khác (other\_disease)**: Những người không có bệnh khác (2) có tỷ lệ mắc COVID-19 cao hơn so với những người có bệnh khác (1).
11. **Bệnh tim mạch (cardiovascular)**: Những người bị bệnh tim mạch (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị bệnh tim mạch (1).
12. **Béo phì (obesity)**: Những người bị béo phì (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị béo phì (1).
13. **Bệnh mãn tính về thận (renal\_chronic)**: Những người bị bệnh thận mãn tính (2) có tỷ lệ mắc COVID-19 cao hơn so với những người không bị bệnh thận mãn tính (1).
14. **Hút thuốc lá (tobacco)**: Những người không hút thuốc lá (2) có tỷ lệ mắc COVID-19 cao hơn so với những người hút thuốc lá (1).
15. **Tiếp xúc với người mắc COVID-19 (contact\_other\_covid)**: Những người đã tiếp xúc với người mắc COVID-19 (1) có tỷ lệ mắc COVID-19 cao hơn so với những người không tiếp xúc với người mắc COVID-19 (2).

Nhìn chung, các thuộc tính liên quan đến các bệnh mãn tính, suy giảm miễn dịch, và các yếu tố nguy cơ sức khỏe khác như béo phì, hút thuốc lá, và tiếp xúc với người mắc COVID-19 đều có xu hướng làm tăng tỷ lệ mắc COVID-19.

3.3.6. Biểu đồ nhiệt (Head map) thể hiện độ tương quan giữa các thuộc tính



***=> Nhận xét:***

### 1. Độ tương quan giữa các thuộc tính:

* **Patient\_type (Loại bệnh nhân)** có mối tương quan cao nhất với thuộc tính **contact\_other\_covid** (tiếp xúc với người mắc COVID-19) với giá trị 0.42. Điều này cho thấy rằng những người tiếp xúc với người mắc COVID-19 có khả năng cao sẽ trở thành bệnh nhân nội trú.
* **Hypertension (Tăng huyết áp)** có mối tương quan trung bình với **age (tuổi)** với giá trị 0.38. Điều này hợp lý vì người lớn tuổi thường có nguy cơ cao bị tăng huyết áp.
* **Age (Tuổi)** cũng có mối tương quan âm với **pneumonia (viêm phổi)** với giá trị -0.28, cho thấy người lớn tuổi thường có nguy cơ mắc viêm phổi cao hơn.
* **Pneumonia (Viêm phổi)** và **patient\_type** có mối tương quan âm khá mạnh (-0.66). Điều này có thể do những người bị viêm phổi nặng thường được chăm sóc tại nhà hoặc không phải nhập viện nội trú.

### 2. Độ tương quan giữa các thuộc tính covid\_res (kết quả xét nghiệm covid):

* Thuộc tính **pneumonia (viêm phổi)** có mối liên hệ nghịch biến mạnh nhất với **covid\_res** (-0.20), cho thấy rằng những người bị viêm phổi có khả năng mắc COVID-19 thấp hơn so với người không bị viêm phổi. Tuy nhiên, cần xem xét kỹ kết quả này vì nó có thể liên quan đến các yếu tố khác như mức độ nghiêm trọng của bệnh hoặc sự chẩn đoán khác nhau.
* **Patient\_type (Loại bệnh nhân)** và **age (tuổi)** có mối tương quan dương với **covid\_res**, lần lượt là 0.20 và 0.17, cho thấy bệnh nhân nội trú và người lớn tuổi có khả năng cao mắc COVID-19.
* Các thuộc tính khác như **tăng huyết áp (hypertension)**, **tiểu đường (diabetes)**, **COPD**, **immsupr (suy giảm miễn dịch)** có mối tương quan khá thấp hoặc không có mối tương quan rõ ràng với **covid\_res**, điều này cho thấy chúng không có tác động lớn đến khả năng mắc COVID-19 trong mẫu dữ liệu này.

### Kết luận:

* Biểu đồ nhiệt cho thấy một số thuộc tính có mối tương quan nhất định với nhau và với khả năng mắc COVID-19, nhưng hầu hết các mối tương quan này đều không quá mạnh.
* Các yếu tố như loại bệnh nhân, tuổi tác, và viêm phổi có tác động đáng kể đến khả năng mắc COVID-19, trong khi các yếu tố khác dường như không có tác động lớn.

## 3.4. Huấn luyện các mô hình

### 3.4.1. Phân chia tập dữ liệu

Chia bộ dữ liệu thành hai phần training data và test data nhằm đảm bảo tính công bằng trên mọi model:



### **3.4.2. Lựa chọn thuật toán:**

3.4.2.1. Linear Regression:

* "Hồi quy tuyến tính" là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại.
* Đây cũng là 1 mô hình phù hợp trong việc phân tích sự tác động của từng biến độc lập lên biến phụ thuộc (covid\_res)

3.4.2.2. Logistic Regression:

* Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhóm
* Bài toán của chúng ta là **binary classification** và **Logistic Regression** thường được coi là thuật toán cơ bản nhất cho các bài toán dạng này.
* Thuật toán **Logistic Regression** có thời gian thực thi nhanh và cách cài đặt đơn giản

3.4.2.3. Random Forest:

* Random Forest xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.
* Random Forest cũng phù hợp với bài toán binary classification, nâng cao độ chính xác của mô hình và ngăn chặn vấn đề overfitting. 3.4.2.4. Gradient Boosting:
* Gradient Boost là mô hình tuần tự kết hợp của nhiều Cây quyết định khác nhau, trong đó kết quả của một cây quyết định được sử dụng để đào tạo cây tiếp theo, được đánh giá là một thuật toán học máy mạnh mẽ được tạo thành từ Gradient descent và Boosting

3.4.2.5. XGBoost:

* XGBoost là một trong những thuật toán học máy phổ biến và được sử dụng rộng rãi nhất hiện nay vì đơn giản là nó rất mạnh mẽ. Nó tương tự như Gradient Boost nhưng có một vài tính năng bổ sung làm cho nó mạnh hơn

3.4.2.6. LightGBM:

* Sử dụng một kỹ thuật độc đáo được gọi là Lấy mẫu một phía dựa trên Gradient (GOSS – Gradient-based One-Side Sampling) để lọc ra các cá thể dữ liệu nhằm tìm ra giá trị phân tách

### **3.4.3. Lựa chọn kỹ thuật đánh giá:**

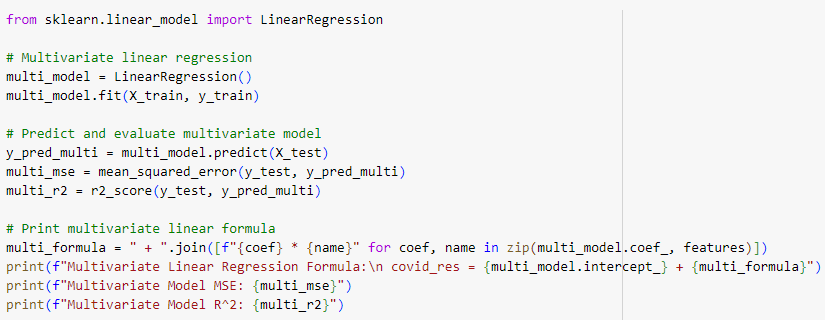
* Phân chia tập dữ liệu: Sử dụng kỹ thuật Train-test split: Chia dữ liệu thành hai tập: tập huấn luyện (train set) để huấn luyện mô hình và tập kiểm tra (test set) để đánh giá mô hình.
* Đối với mô hình Linear Regression:
* **Mean Squared Error (MSE):** Bình phương trung bình của sai số dự đoán.
* **R-squared:** Đo lường phần phương sai của biến phụ thuộc được giải thích bởi mô hình.
* Đối với các mô hình Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM:

Dùng các chỉ số:

* **Accuracy:** Tỷ lệ dự đoán đúng trên toàn bộ dữ liệu.
* **Precision:** Tỷ lệ các dự đoán dương tính đúng trên tổng số dự đoán dương tính.
* **Recall:** Tỷ lệ các dự đoán dương tính đúng trên tổng số mẫu thực tế dương tính.
* **F1-score:** Trung bình điều hòa của precision và recall.

### **3.4.4. Huấn luyện mô hình:**

3.4.4.1. Linear Regression:



\* Kết quả:

* Multivariate Linear Regression Formula:

***covid\_res*** = - 0.4853684519771498 + 0.05426961919884633 \* sex + 0.1310541973287605 \* patient\_type - 0.13901549156718981 \* pneumonia + 0.0031056561320815523 \* age - 0.02691589119834419 \* pregnancy - 0.04077180905051631 \* diabetes + 0.1057455368016624 \* copd + 0.04504165618608364 \* asthma + 0.1125158028001424 \* inmsupr - 0.006377996236076359 \* hypertension + 0.06632017602932726 \* other\_disease + 0.08875247012256358 \* cardiovascular + -0.07727118683664853 \* obesity + 0.08350280794731875 \* renal\_chronic + 0.06481496239181288 \* tobacco + -0.016589694052934284 \* contact\_other\_covid

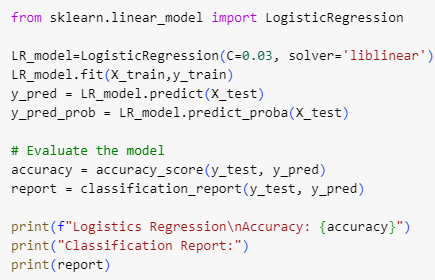
* 0 = - 0.4853684519771498
* 1 = 0.05426961919884633 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 5,4% nếu là nữ, tăng 10,8% nếu là nam (với các yếu tố còn lại không đổi)
* 2 = 0.1310541973287605 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là 13,1% nếu không nhập viện, 26,2% nếu nhập viện (với các yếu tố còn lại không đổi)
* *3 = -0.13901549156718981 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 13,9% nếu không bị viêm phổi, giảm 27,8% nếu bị viêm phổi (với các yếu tố còn lại không đổi)*
* 4 = 0.0031056561320815523 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 3% khi tuổi càng cao (với các yếu tố còn lại không đổi)
* 5 = - 0.02691589119834419 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 2,7% nếu có thai, giảm 5,4% nếu không mang thai (với các yếu tố còn lại không đổi)
* 6 = - 0.04077180905051631 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 4,1% nếu không mắc bệnh tiểu đường, giảm 8,2% nếu mắc bệnh tiểu đường (với các yếu tố còn lại không đổi)
* 7 = 0.1057455368016624 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 10,5% nếu không mắc bệnh nghẽn phổi mãn tính, tăng 21% nếu mắc bệnh nghẽn phổi mãn tính (với các yếu tố còn lại không đổi)
* 8 = 0.04504165618608364 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 4,5% nếu không mắc bệnh hen xuyễn, 9% nếu mắc bệnh hen xuyễn (với các yếu tố còn lại không đổi)
* 9 = 0.1125158028001424 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 11,3% nếu không mắc bệnh ức chế miẽn dịch, 22,6% nếu mắc bệnh ức chế miễn dịch (với các yếu tố còn lại không đổi)
* 10 = - 0.006377996236076359 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 0,6% nếu không có tăng huyết áp, giảm 1,2% nếu tăng huyết áp (với các yếu tố còn lại không đổi)
* 11 = 0.06632017602932726 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 6,6% nếu không có mắc bệnh khác, tăng 13,2% nếu tăng huyết áp (với các yếu tố còn lại không đổi)
* 12 = 0.08875247012256358 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 8.8% nếu không mắc bệnh khác, 17,6% nếu mắc bệnh khác (với các yếu tố còn lại không đổi)
* 13 = -0.07727118683664853 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 7,7% nếu không mắc bệnh tim mạch, giảm 15,4% nếu mắc bệnh tim mạch (với các yếu tố còn lại không đổi)
* 14 = 0.08350280794731875 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 8,3% nếu không bị béo phì, tăng 16,6% nếu bị béo phì (với các yếu tố còn lại không đổi)
* 15 = 0.06481496239181288 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là tăng 6,4% nếu không hút thuốc, tăng 12,8% nếu hút thuốc (với các yếu tố còn lại không đổi)
* 16 = -0.016589694052934284 có nghĩa là khả năng mắc bệnh covid19 được kì vọng là giảm 1,6% nếu không tiếp xúc bệnh nhân covid khác, giảm 3,2% nếu tiếp xúc với bệnh nhân mắc covid khác(với các yếu tố còn lại không đổi)

=> Biến có tác động lớn nhất đối biến covid\_res là biến pneumonia, nghĩa là việc bệnh nhân có bị viêm phổi hay không có ảnh hưởng nhiều nhất đến khả năng mắc covid19

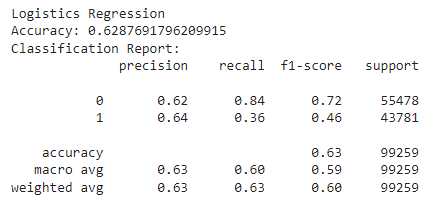
* Multivariate Model MSE: 0.22935301030677166
* Multivariate Model R^2: 0.06966842173680632

\* Thời gian thực thi: 3 giây

3.4.4.2. Logistics Regression:

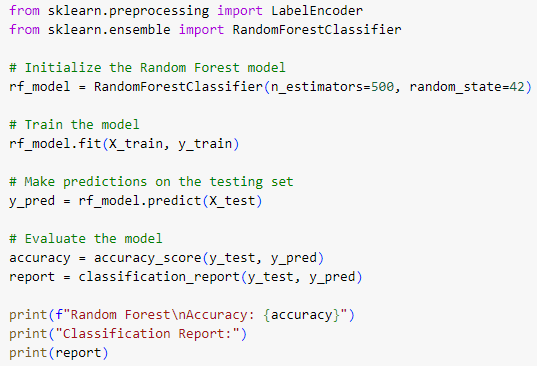


\* Kết quả:

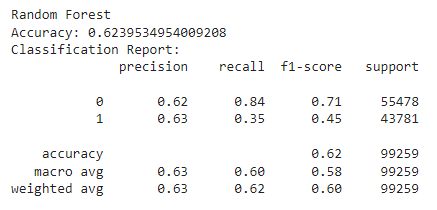


\* Thời gian thực thi: 5 giây

3.4.4.3. Random Forest:

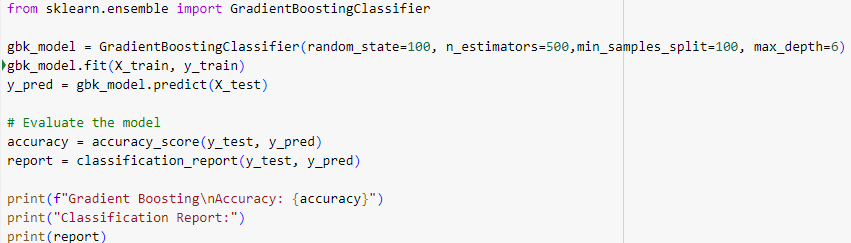


\* Kết quả:

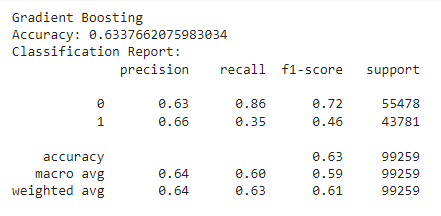


\* Thời gian thực thi: 4 phút

3.4.4.4. Gradient Boosting:

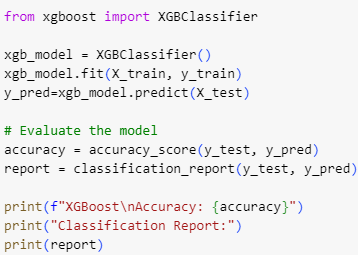


\* Kết quả:

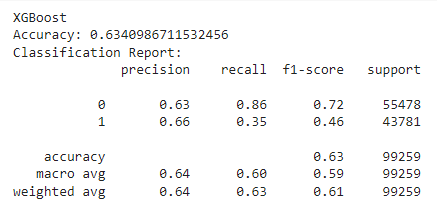


\* Thời gian thực thi: 7 phút

3.4.4.5. XGBoost:

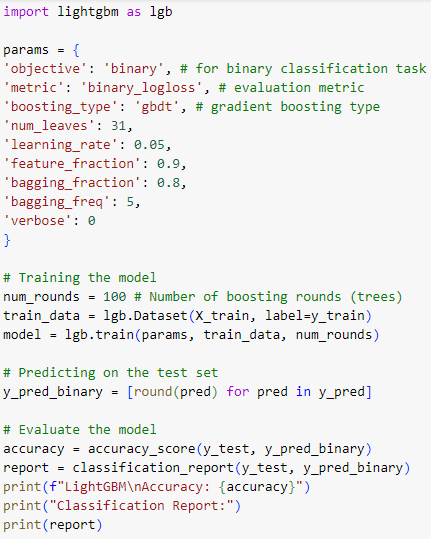


\* Kết quả:

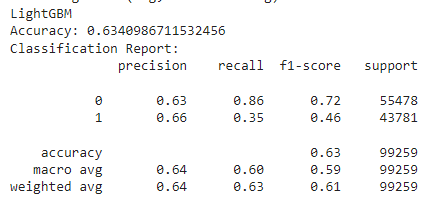


\* Thời gian thực thi: 3 giây

3.4.4.6. LightGBM:



\* Kết quả:



\* Thời gian thực thi: 4-5 giây

### 3.4.5. Lựa chọn mô hình:

Dựa vào kết quả huấn luyện mô hình:

| **Mô hình** | **Độ chính xác (Accuracy)** | **Precision (Class 0)** | **Precision (Class 1)** | **Recall (Class 0)** | **Recall (Class 1)** | **F1-Score (Class 0)** | **F1-Score (Class 1)** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Hồi quy Logistic** | 0.6288 | 0.62 | 0.64 | 0.84 | 0.36 | 0.72 | 0.46 |
| **Random Forest** | 0.6240 | 0.62 | 0.63 | 0.84 | 0.35 | 0.71 | 0.45 |
| **Gradient Boosting** | 0.6338 | 0.63 | 0.66 | 0.86 | 0.35 | 0.72 | 0.46 |
| **XGBoost** | 0.6341 | 0.63 | 0.66 | 0.86 | 0.35 | 0.72 | 0.46 |
| **LightGBM** | 0.6240 | 0.62 | 0.63 | 0.84 | 0.35 | 0.71 | 0.45 |

1. **Mô hình Linear Regression:** Dùng mô hình hồi quy đa biến

* MSE: Một giá trị MSE càng nhỏ thì mô hình càng phù hợp với dữ liệu. Với giá trị MSE = 0.229, có nghĩa là trung bình bình phương sai số giữa giá trị dự đoán và giá trị thực tế của biến phụ thuộc (covid\_res) là 0.229. Tuy nhiên, để đánh giá mức độ tốt xấu của giá trị này, chúng ta cần so sánh với MSE của các mô hình khác hoặc với một ngưỡng MSE được xác định trước. Nếu MSE này nhỏ so với các mô hình khác hoặc nhỏ hơn ngưỡng cho phép, thì có thể coi mô hình này có độ chính xác khá tốt.
* R-square: Với giá trị R-squared = 0.069, điều này có nghĩa là chỉ có khoảng 6.9% sự biến động của biến covid\_res được giải thích bởi các biến độc lập trong mô hình. Đây là một giá trị R-squared rất thấp. Điều này cho thấy mô hình hồi quy tuyến tính đa biến này không phù hợp để dự đoán biến covid\_res. Các biến độc lập trong mô hình không có khả năng giải thích tốt sự biến động của biến phụ thuộc.

**=> Kết luận:**

* Mô hình hồi quy tuyến tính đa biến này không phù hợp để dự đoán biến covid\_res.
* Các biến độc lập trong mô hình không có mối quan hệ tuyến tính mạnh với biến phụ thuộc.
* Có thể có các yếu tố khác, không được bao gồm trong mô hình, ảnh hưởng lớn đến biến covid\_res.

1. **Mô hình Logistic Regression:**

* **Hiệu suất chung:** Mô hình đạt được độ chính xác tổng thể là 0.63, nghĩa là mô hình dự đoán đúng khoảng 63% các mẫu với thời gian thực hiện là 5-7 giây. Đây là một kết quả khá ổn, tuy nhiên vẫn còn nhiều không gian để cải thiện.
* **Phân tích theo lớp:**
* **Class 0 (Negative):** Mô hình có độ chính xác (precision) và độ phủ (recall) tương đối cao, đặc biệt là recall đạt 0.84, cho thấy mô hình khá tốt trong việc phát hiện các mẫu thuộc lớp 0.
* **Class 1 (Positive):** Mô hình có độ chính xác và độ phủ thấp hơn so với lớp 0, đặc biệt là recall chỉ đạt 0.36. Điều này cho thấy mô hình gặp khó khăn trong việc phát hiện các mẫu thuộc lớp 1.
* **F1-score:** F1-score của class 1 thấp hơn class 0 (0.46 < 0.72, phản ánh sự mất cân bằng trong việc dự đoán hai lớp.

### => Tổng quan: Dựa trên các kết quả trên, có thể đưa ra một số nhận xét sau:

* **Mô hình đã đạt được một số kết quả khả quan:** Mô hình đã có thể phân loại khá tốt các mẫu thuộc class 0.
* **Vẫn còn nhiều không gian để cải thiện:** Mô hình gặp khó khăn trong việc phân loại các mẫu thuộc class 1. Điều này có thể do một số nguyên nhân như:
* **Dữ liệu không cân bằng:** Số lượng mẫu trong hai lớp có thể khác nhau đáng kể, dẫn đến mô hình thiên về lớp có nhiều mẫu hơn.
* **Các đặc trưng không đủ hoặc không phù hợp:** Các đặc trưng sử dụng để huấn luyện mô hình có thể chưa đủ để phân biệt rõ ràng giữa hai lớp.
* **Tham số của mô hình chưa được tối ưu:** Các tham số của mô hình như tỉ lệ học, số lượng cây quyết định (nếu sử dụng Random Forest),... có thể chưa được điều chỉnh tối ưu.

1. **Mô hình Random Forest:**

* **Hiệu suất chung:** Tương tự với mô hình Logistic Regression, mô hình Random Forest đạt được độ chính xác tổng thể là 0.62. Điều này có nghĩa là mô hình đã dự đoán đúng khoảng 62% các mẫu trong tập dữ liệu. Tuy nhiên thời gian thực hiện huấn luyện mô hình đến 4 phút.

**Phân tích chi tiết theo từng class:**

* **Class 0:**
* **Precision (Độ chính xác):** 0.62: Tức là trong số các mẫu mà mô hình dự đoán là thuộc lớp 0, có khoảng 62% là dự đoán đúng.
* **Recall (Độ phủ):** 0.84: Điều này có nghĩa là mô hình đã tìm thấy 84% các mẫu thực sự thuộc lớp 0 trong tổng số các mẫu thuộc lớp 0.
* **F1-score:** 0.71: Đây là giá trị trung bình điều hòa giữa precision và recall, cho thấy sự cân bằng giữa độ chính xác và độ phủ của lớp 0.
* **Class 1:**
* **Precision (Độ chính xác):** 0.63: Tương tự lớp 0, trong số các mẫu mà mô hình dự đoán là thuộc lớp 1, có khoảng 63% là dự đoán đúng.
* **Recall (Độ phủ):** 0.35: Tuy nhiên, mô hình chỉ tìm thấy 35% các mẫu thực sự thuộc lớp 1 trong tổng số các mẫu thuộc lớp 1. Điều này cho thấy mô hình gặp khó khăn hơn trong việc phân loại chính xác các mẫu thuộc lớp 1.
* **F1-score:** 0.45: Giá trị F1-score thấp hơn so với lớp 0, cho thấy sự mất cân bằng trong việc dự đoán hai lớp.

**=> Tổng quan:** Mô hình Random Forest đang hoạt động tốt hơn trên class 0 (negative) so với class 1 (positive). Điều này thể hiện rõ qua chỉ số recall và F1-score của class 1 thấp hơn đáng kể so với class 0.

1. Mô hình Gradient Boosting:

* **Hiệu suất chung:** Tương tự với 2 mô hình trên, mô hình Gradient Boosting đạt được độ chính xác tổng thể là 0.63. Điều này có nghĩa là mô hình đã dự đoán đúng khoảng 63% các mẫu trong tập dữ liệu. Cao hơn 0.1% so với mô hình Random Forest. Tuy nhiên, mô hình tốn khá nhiều thời gian để huấn luyện, thời gian thực thi lên đến 7 phút.
* **Phân tích theo từng class**: Mô hình Gradient Boosting có các chỉ số tương đương với mô hình Random Forest, mô hình hoạt động tốt hơn trên class 0 (negative) so với class 1 (positive). Điều này thể hiện rõ qua chỉ số recall và F1-score của class 1 thấp hơn đáng kể so với class 0. Vì vậy, vẫn còn nhiều khía cạnh để cải thiện chất lượng của mô hình.

1. Mô hình XGBoost:

* **Hiệu suất chung:** Mô hình XGBoost đạt được độ chính xác tổng thể là 0.63. Điều này có nghĩa là mô hình đã dự đoán đúng khoảng 63% các mẫu trong tập dữ liệu tương tự với mô hình Gradient Boosting nhưng thời gian thực thi huấn luyện mô hình nhanh hơn khá nhiều so với mô hình Gradient Boosting, thời gian thực thi chỉ có 3 giây.
* **Phân tích theo từng class**: Mô hình XGBoost cũng có các chỉ số tương đương với mô hình Gradient Boosting, mô hình hoạt động tốt hơn trên class 0 (negative) so với class 1 (positive). Điều này thể hiện rõ qua chỉ số recall và F1-score của class 1 thấp hơn đáng kể so với class 0. Vì vậy, vẫn còn nhiều khía cạnh để cải thiện chất lượng của mô hình.

1. Mô hình LightGBM

* **Hiệu suất chung:** Mô hình LightGBM đạt được độ chính xác tổng thể là 0.62. Điều này có nghĩa là mô hình đã dự đoán đúng khoảng 62% các mẫu trong tập dữ liệu xấp xỉ với các mô hình đã được huấn luyện trên, thời gian thực thi huấn luyện mô hình nhanh xấp xỉ so với mô hình XGBoost, có thể chậm hơn vài giây.
* **Phân tích theo từng class**: Mô hình LightGBM cũng có các chỉ số tương đương với mô hình XGBoost, mô hình hoạt động tốt hơn trên class 0 (negative) so với class 1 (positive). Điều này thể hiện rõ qua chỉ số recall và F1-score của class 1 thấp hơn đáng kể so với class 0.

**=> Lựa chọn mô hình:** Sau khi đánh giá cả 6 mô hình, thì ta có 3 mô hình có hiệu suất khá ổn nhất, đó là: Logistic Regresstion, XGBoost, LightGBM. Không chọn mô hình Linear Regression là vì có mô hình hồi quy đa biến có vẻ không phù hợp với bộ dữ liệu, còn mô hình Random Forest và Gradient Boosting thì tốn quá nhiều thời gian để huấn luyện mô hình.

Ta sẽ tiếp tục so sánh 3 mô hình trên để tìm ra được mô hình phù hợp nhất:

1. **Accuracy:**
   1. XGBoost: 0.6341
   2. LightGBM: 0.6239
   3. Logistic Regression: 0.6288

=> XGBoost có độ chính xác cao nhất.

1. **Precision, Recall, F1-score:**
   1. Đối với class 0 (Negative class):

* Precision: Logistic Regression = LightGBM (0.62) < XGBoost (0.63)
* Recall: Logistic Regression = LightGBM (0.84) < XGBoost (0.86)
* F1-score: Logistic Regression (0.72) = XGBoost (0.72) > LightGBM (0.71)
  1. Đối với class 1 (Positive class):
* Precision: Logistic Regression (0.64) < XGBoost (0.66) > LightGBM (0.63)
* Recall: Logistic Regression = LightGBM (0.35) < XGBoost (0.36)
  + 1. F1-score: Logistic Regression = XGBoost (0.46) > LightGBM (0.45)

1. **Train Time:**
   1. XGBoost: 3 giây
   2. LightGBM: 4 giây
   3. Logistic Regression: 5-7 giây

=> XGBoost có thời gian huấn luyện nhanh hơn Logistic Regression và LightGBM

**Tổng kết:**

* **XGBoost** có độ chính xác, Precision, Recall, và F1-score cao nhất trên cả hai lớp. Ngoài ra, thời gian huấn luyện của nó cũng ngắn hơn Logistic Regression.
* **Logistic Regression** cũng là một mô hình tốt, với các chỉ số gần như tương đương XGBoost, nhưng thua kém một chút về Precision, Recall, và thời gian huấn luyện.
* **LightGBM** có kết quả tương đối thấp hơn so với XGBoost và Logistic Regression trong hầu hết các chỉ số.

**Lựa chọn mô hình:** **XGBoost** là mô hình tốt nhất trong số ba mô hình này dựa trên các chỉ số phân loại và thời gian huấn luyện.



# 4. Projected Impact

## 4.1. Thành tựu và lợi ích của dự án

## 4.2. Cải tiến trong tương lai

5. Đánh giá và nhận xét các thành viên của nhóm trưởng 

| Họ và tên | Đánh giá, nhận xét |
| --- | --- |
| Võ Thị Ngọc Châu | 1.Ưu điểm:  1.1. Về kiến thức: có kiến thức cơ bản về phân tích dữ liệu  1.2. Về kinh nghiệm: sử dụng các công cụ hỗ trợ hiệu quả  1.3. Hoàn thành các công việc được giao đúng hạn  2. Nhược điểm:  2.1. Các điểm còn sai, lỗi chưa khắc phục được: phân chia thời gian thực hiện các công việc chưa thực sự hợp lý  2.2. Kiến thức và kỹ năng: kỹ năng phân chia công việc chưa tốt  3. Tổng kết: Hoàn thành khá tốt, cần trao dồi thêm khả năng quản lý và phân chia công việc. |
| Lê Quang An | 1.Ưu điểm:  1.1. Về kiến thức: tiếp thu tốt kiến thức về phân tích dữ liệu  1.2. Hoàn thành các công việc được giao đúng hạn  2. Nhược điểm:  Kỹ năng: mới tiếp xúc các công cụ hỗ trợ phân tích nên chưa thực sự thành thạo  3. Tổng kết: Hoàn thành tốt trong phạm vi dự án, nên thực hành nhiều để trao dồi thêm |
| Nguyễn Minh Gia Bảo | 1.Ưu điểm:  1.1. Về kiến thức: tiếp thu tốt kiến thức về phân tích dữ liệu  1.2. Hoàn thành các công việc được giao đúng hạn  1.3. Có sự chủ động trong việc đóng góp ý kiến cũng như học hỏi  2. Nhược điểm:  Kỹ năng: mới tiếp xúc các công cụ hỗ trợ phân tích nên chưa thực sự thành thạo  3. Tổng kết: Hoàn thành tốt trong phạm vi dự án, nên thực hành nhiều để trao dồi thêm |
| Phạm Thành Đạt | 1.Ưu điểm:  1.1. Về kiến thức: tiếp thu tốt kiến thức về phân tích dữ liệu  1.2. Hoàn thành các công việc được giao đúng hạn  1.3. Có sự chủ động trong việc đóng góp ý kiến cũng như học hỏi  2. Nhược điểm:  Kỹ năng: mới tiếp xúc các công cụ hỗ trợ phân tích nên chưa thực sự thành thạo  3. Tổng kết: Hoàn thành tốt trong phạm vi dự án, nên thực hành nhiều để trao dồi thêm |
| Đặng Vạn Như Ý | 1.Ưu điểm:  1.1. Về kiến thức: có kiến thức cơ bản về phân tích dữ liệu  1.2. Về kinh nghiệm: sử dụng các công cụ hỗ trợ hiệu quả  1.3. Hoàn thành các công việc được giao đúng hạn  1.4. Có sự chủ động trong việc xây dựng dự án  2. Nhược điểm:  2.1. Kiến thức và kỹ năng: nên trao dồi thêm để phát triển bản thân hơn  3. Tổng kết: Hoàn thành tốt |
| Hà Vân Yến | 1.Ưu điểm:  1.1. Về kiến thức: có kiến thức cơ bản về phân tích dữ liệu  1.2. Về kinh nghiệm: sử dụng các công cụ hỗ trợ hiệu quả  1.3. Hoàn thành các công việc được giao đúng hạn  1.4. Có sự chủ động trong việc xây dựng dự án  2. Nhược điểm:  2.1. Kiến thức và kỹ năng: nên trao dồi thêm để phát triển bản thân hơn  3. Tổng kết: Hoàn thành tốt |

6. Đánh giá và nhận xét của giảng viên

| Danh mục | Điểm | Đánh giá và nhận xét |
| --- | --- | --- |
| Ý tưởng | \_\_/10 |  |
| Ứng dụng | \_\_/30 |  |
| Kết quả | \_\_/30 |  |
| Quản lý dự án | \_\_/10 |  |
| Thuyết trình và báo cáo | \_\_/20 |  |
| Tổng | \_\_/100 |  |