SAMSUNG

COVID-19
Analytics and Prediction

OU_BD_01 Nhóm 4

Big Data Course

Covid19 Analysis and Prediction of

BỐI CẢNH DỊCH BỆNH COVID-19

- **COVID-19:** Bệnh viêm đường hô hấp do coronavirus mới gây ra.
- **Triệu chứng:** Hầu hết mắc bệnh từ nhẹ đến trung bình, hồi phục mà không cần điều trị đặc biệt. Người già và những người có bệnh nền có nguy cơ cao mắc bệnh nặng.
- **Phòng ngừa:**
  - Hiểu biết về bệnh, hậu quả và cách lây truyền.
  - Rửa tay thường xuyên, sử dụng nước rửa tay sát khuẩn.
  - Tránh chạm vào mặt, thực hiện vệ sinh đường hô hấp (ho vào khuỷu tay).
- **Lây truyền:** Chủ yếu qua giọt bắn từ nước bọt hoặc dịch mũi khi người bệnh ho hoặc hắt hơi.
- ***Lưu ý: Dữ liệu phân tích từ chính phủ Mexico, áp dụng chủ yếu cho Bắc Mỹ.***

MỤC TIÊU

- Xây dựng mô hình dự đoán khả năng mắc COVID-19 của bệnh nhân dựa vào các yếu tố đầu vào cụ thể.
- Trực quan hóa giữa các yếu tố đầu vào và khả năng mắc bệnh để tìm ra sự tương quan giữa chúng, từ đó lựa chọn các thuật toán để xây dựng mô hình. Tiến hành đánh giá và lựa chọn mô hình tốt nhất.

PHƯƠNG PHÁP VÀ CÔNG CỤ

- **Công cụ hỗ trợ:**
  - Google colab
  - Jupyter Notebook
- **Phương pháp thực hiện:**
  - ***Pandas:*** Tải dữ liệu vào DataFrame, xử lý giá trị thiếu, chuẩn hóa, mã hóa biến phân loại, và loại bỏ dữ liệu không cần thiết.
  - ***Seaborn & Matplotlib:*** Trực quan hóa dữ liệu với biểu đồ cột, tròn, heatmap, xu hướng, boxplot và histogram để phát hiện ngoại lệ và phân tích phân phối.
  - ***Xây dựng và huấn luyện mô hình:*** XGBoost, Logistic Regression, LightGBM, Linear Regression, Random Forest, Gradient Boosting, …

MÔ TẢ BỘ DỮ LIỆU

- Bộ dữ liệu được nhóm tìm kiếm từ Kaggle
- Dữ liệu ban đầu có 566602 hàng và 23 cột:

Thông số	Ý nghĩa	Thông số	Ý nghĩa
id	Mã bệnh nhân	**age**	Tuổi của bệnh nhân
sex	Giới tính (1: Nữ, 2: Nam)	**pregnancy**	Mang thai (1: Có, 2: Không)
patient_type	Loại bệnh nhân (1: Không nhập viện, 2: Nhập viện)	**diabetes**	Tiểu đường (1: Có, 2: Không)
entry_date	Ngày đến bệnh viện	**copd**	Bệnh phổi tắc nghẽn mãn tính (1: Có, 2: Không)
date_symptoms	Ngày xuất hiện triệu chứng	**asthma**	Hen suyễn (1: Có, 2: Không)

Thông số	Ý nghĩa	Thông số	Ý nghĩa
date_died	Ngày qua đời (nếu có)	**inmsupr**	Ức chế miễn dịch (1: Có, 2: Không)
intubed	Sử dụng máy thở (1: Có, 2: Không)	**hypertension**	Tăng huyết áp (1: Có, 2: Không)
pneumonia	Viêm phổi (1: Có, 2: Không)	**other_disease**	Bệnh khác (1: Có, 2: Không)
cardiovascular	Bệnh tim mạch (1: Có, 2: Không)	**obesity**	Béo phì (1: Có, 2: Không)
renal_chronic	Bệnh thận mãn tính (1: Có, 2: Không)	**tobacco**	Sử dụng thuốc lá (1: Có, 2: Không)
contact_other_covid	Tiếp xúc với bệnh nhân COVID-19 khác (1: Có, 2: Không)	**icu**	Đưa vào ICU (1: Có, 2: Không)
covid_res	Kết quả COVID-19 (1: Dương tính, 2: Âm tính, 3: Chờ xử lý)		

XỬ LÝ GIÁ TRỊ THIẾU

❌ **Loại bỏ cột:**

- **intubed** (78.5% thiếu)
- **icu** (78.5% thiếu)

✅ **Loại bỏ records:**

- Tỷ lệ rỗng < 5%: **pneumonia**, **diabetes**, **copd**, **asthma**, **inmsupr**, **hypertension**, **other_disease**, **cardiovascular**, **obesity**, **renal_chronic**, **tobacco**..

intubed	444813
pneumonia	11
age	0
pregnancy	288699
diabetes	1981
copd	1749
asthma	1752

inmsupr	1980
hypertension	1824
other_disease	2598
cardiovascular	1822
obesity	1781
renal_chronic	1792
tobacco	1907
contact_other_covid	175031
covid_res	0
icu	444814

XỬ LÝ GIÁ TRỊ THIẾU

🔄 **Dự đoán giá trị thiếu:**

- **pregnancy** (50% thiếu)
- **contact_other_covid** (25% thiếu)

-> Sử dụng hồi quy để dự đoán các giá trị còn thiếu.

🔄 **Chuyển đổi kiểu dữ liệu:**

- Chuyển từ **float64** sang **int64** sau khi thay **NaN** bằng **0**.
- **pregnancy** và **contact_other_covid**: 0 được coi là giá trị thiếu.

pregnancy	286249
contact_other_covid	173792

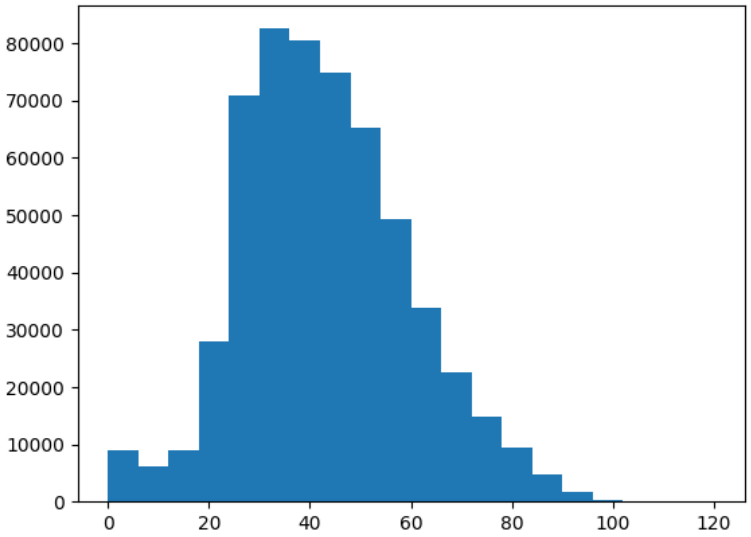
```
pregnancy
2    558463
1      4184
Name: count, dtype: int64
contact_other_covid
1    331628
2    231019
Name: count, dtype: int64
```

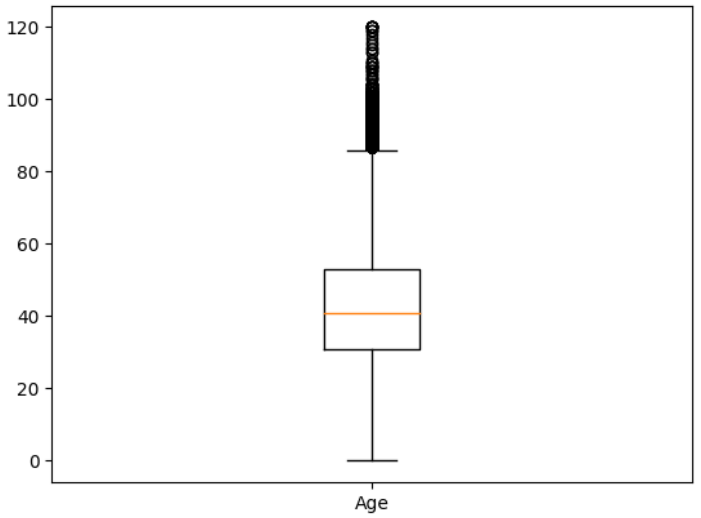
OUTLIERS

🔍 **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

```
Số lượng outlier cột age:
age
False    558631
True       4016
Name: count, dtype: int64
```

```
Các giá trị của cột age:
age
30     14536
31     13847
36     13835
37     13772
38     13751
       ...
118        1
115        1
117        1
116        1
105        1
Name: count, Length: 120, dtype: int64
```





CHUẨN HÓA & MÃ HÓA DỮ LIỆU

✅ **Xử Lý cột covid_res:**

- **Giải thích giá trị**:
  - **1**: 🟢 Positive (Mắc bệnh)
  - **0**: 🟡 Awaiting (Chờ kết quả)
  - **2**: 🔵 Negative (Chưa mắc bệnh)
- ❌ **Loại bỏ giá trị 2**: Tập trung vào phân tích mắc bệnh và không mắc bệnh, loại bỏ hàng có giá trị 2.

covid_res	
0	277389
1	218902
2	66356

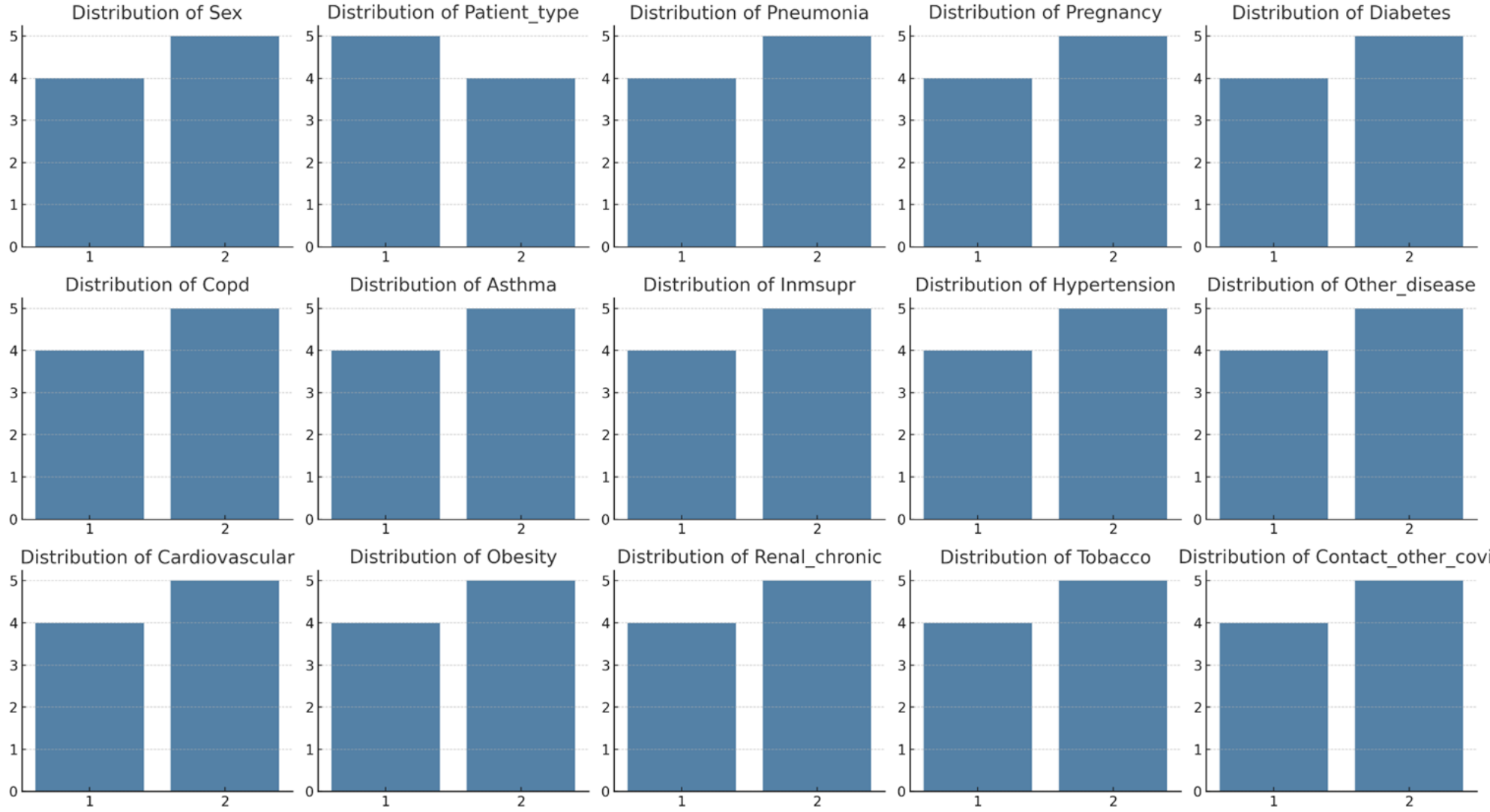
dtype: int64

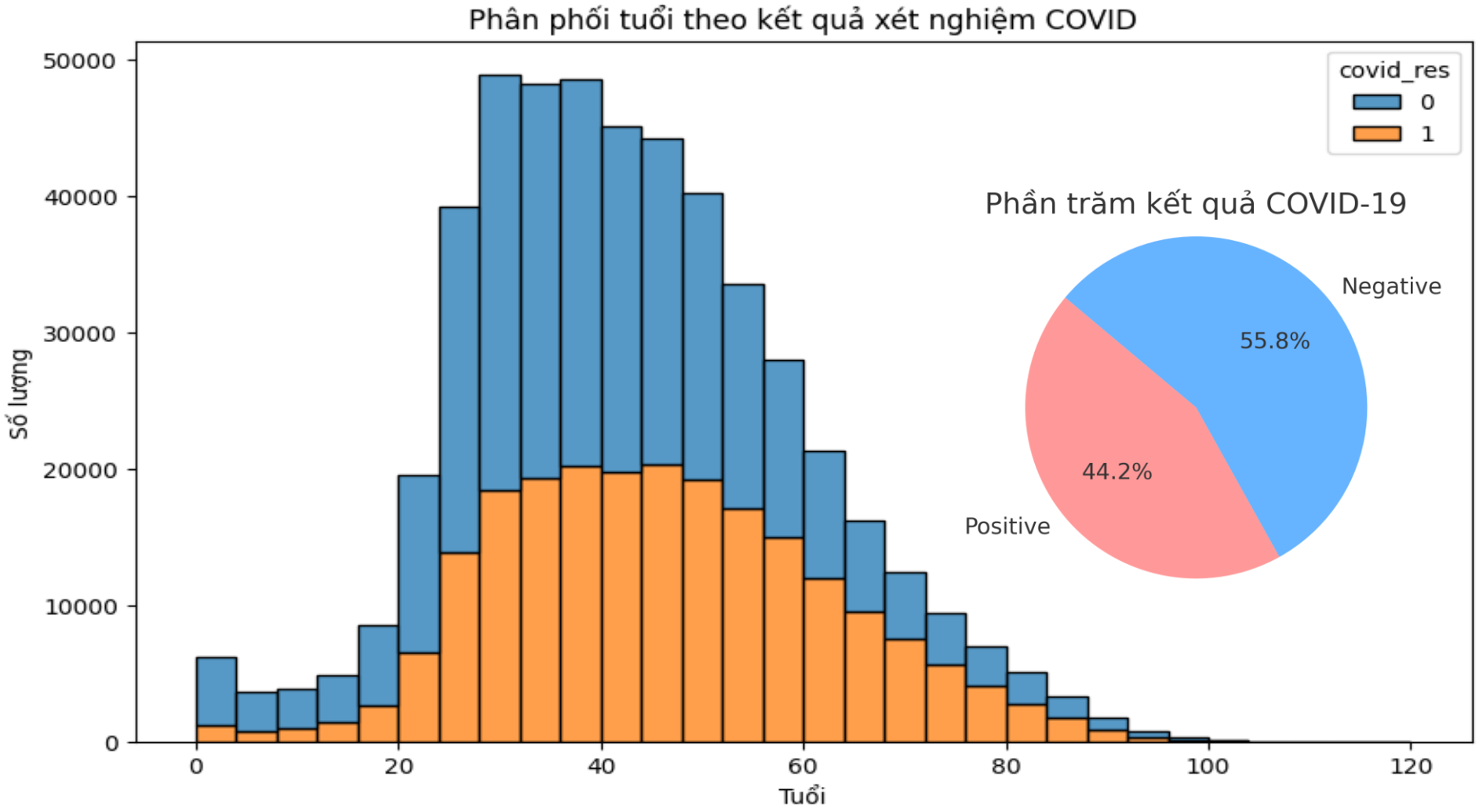
→

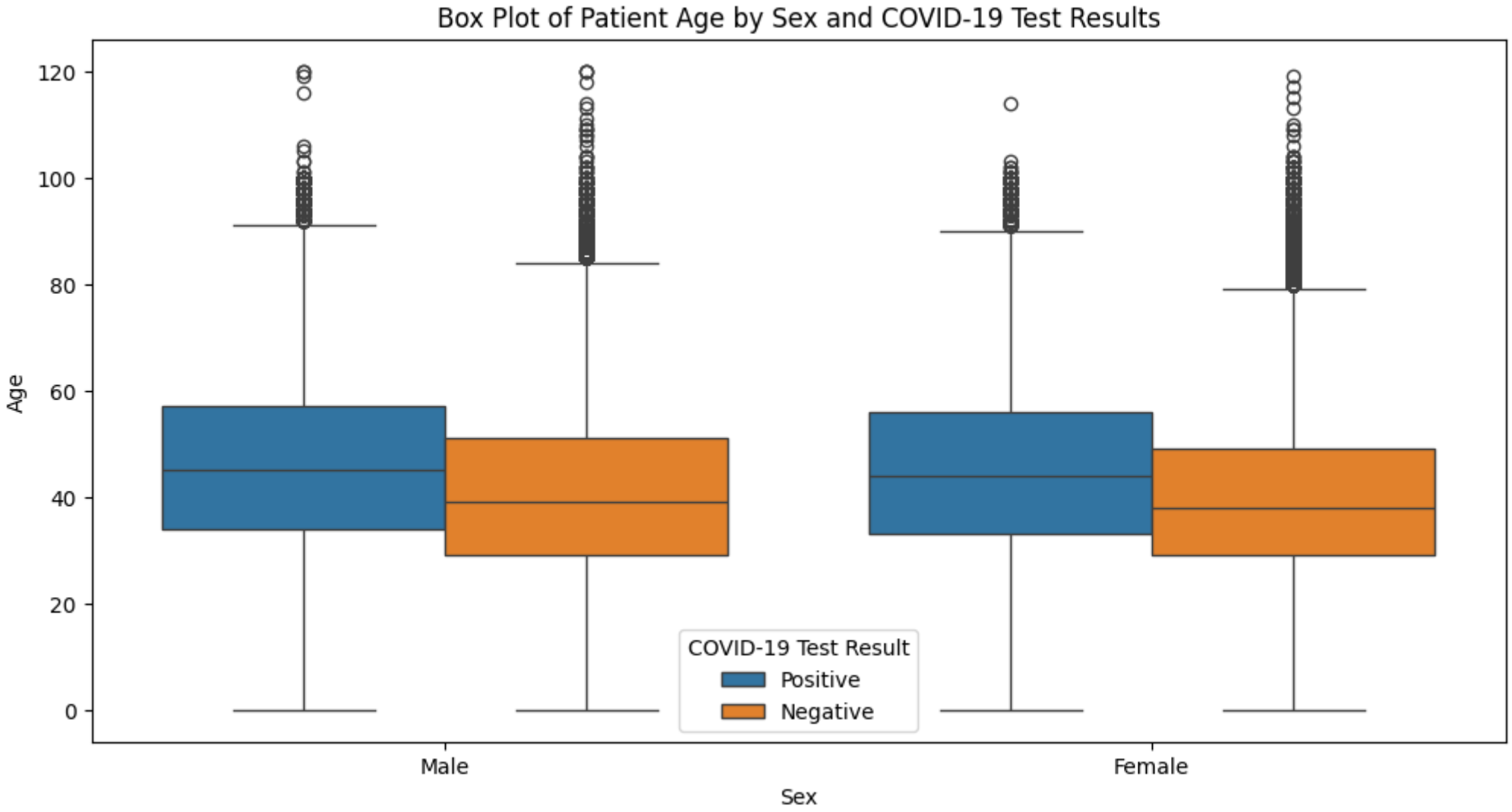
covid_res	
0	277389
1	218902

dtype: int64

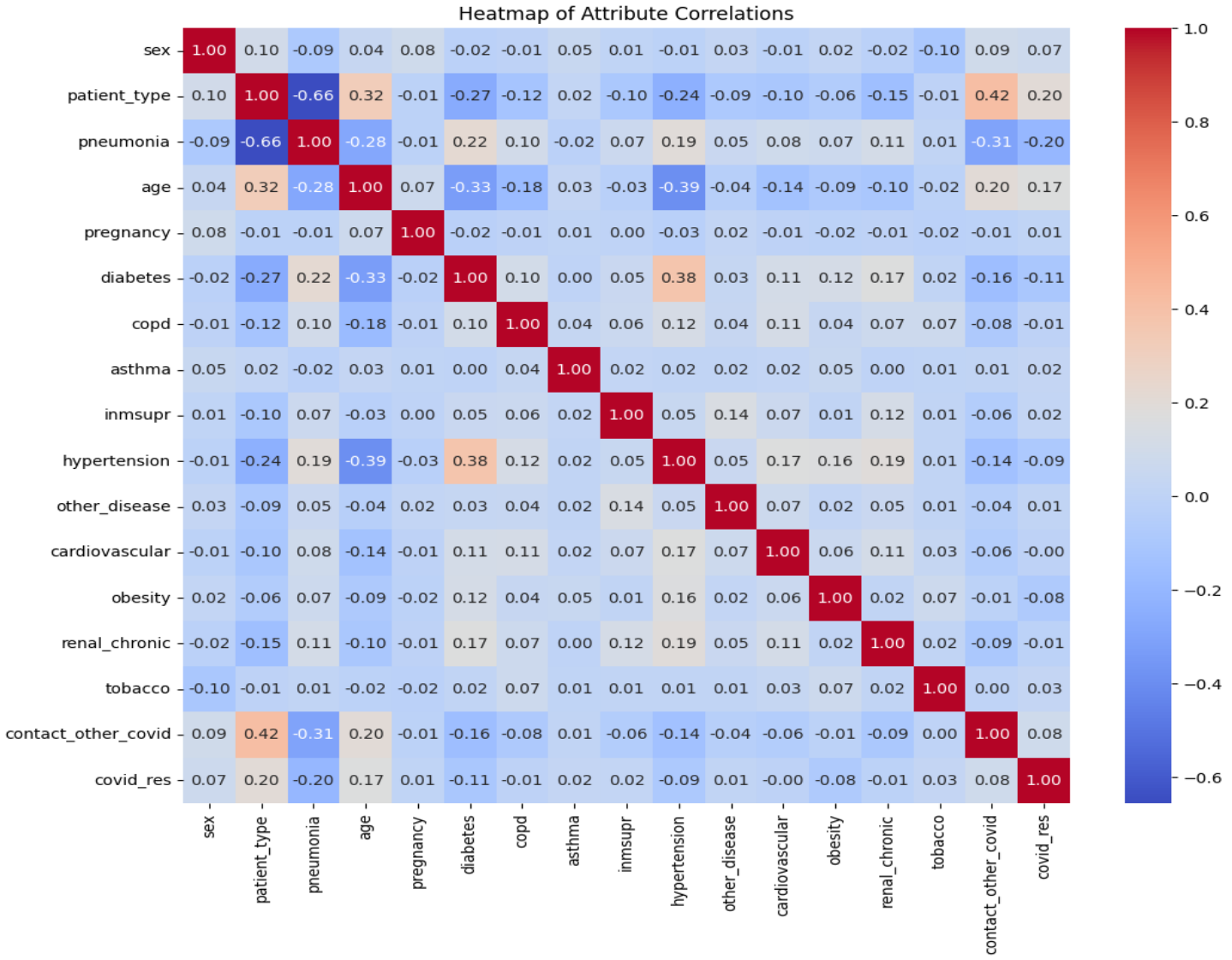
3.2 Trực quan hóa dữ liệu



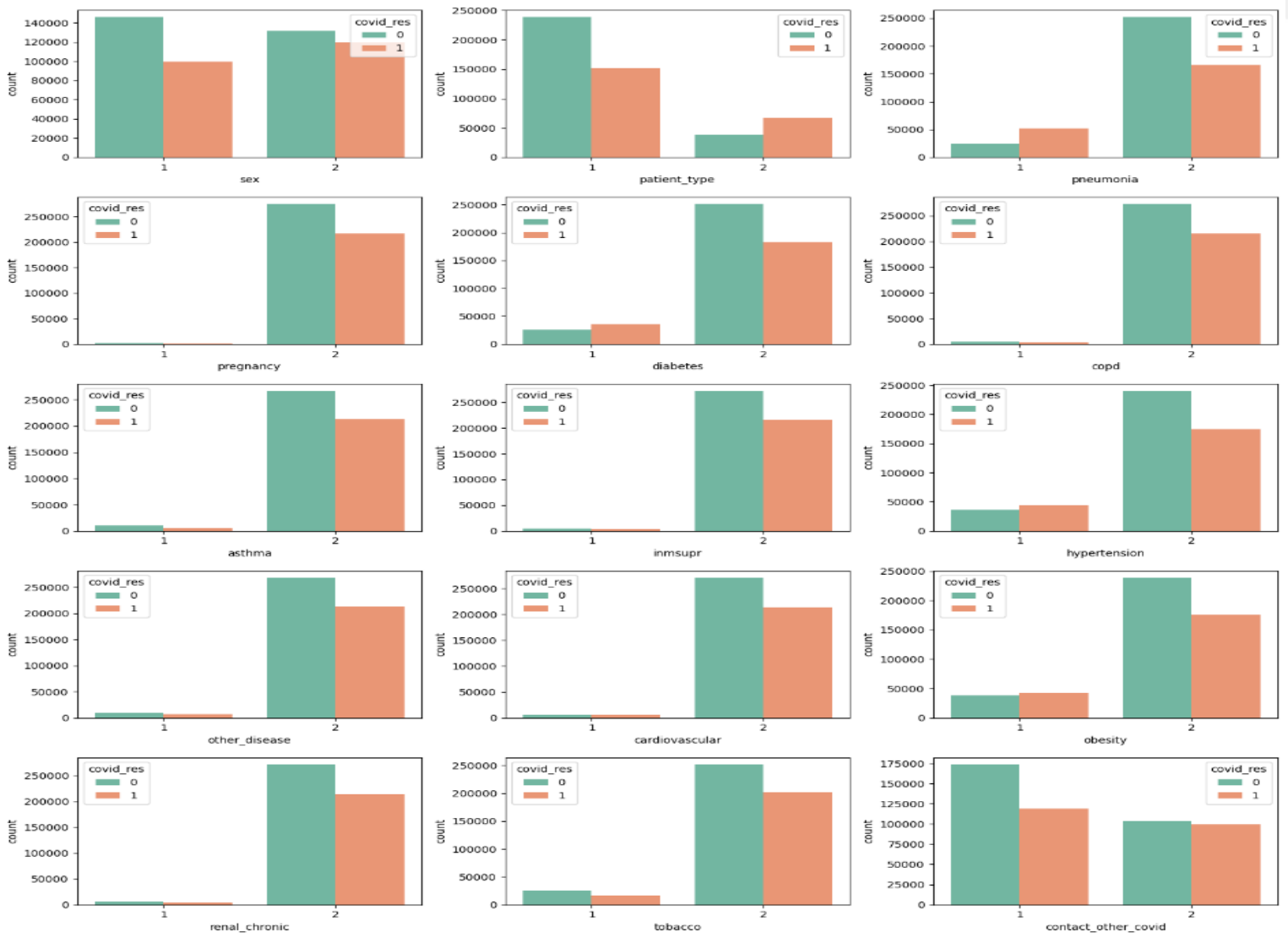

Phân phối tuổi theo kết quả xét nghiệm COVID
covid_res
0
1
Số lượng
50000
40000
30000
20000
10000
0
0
20
40
60
80
100
120
Tuổi
Phần trăm kết quả COVID-19
Negative
55.8%
44.2%
Positive


Box Plot of Patient Age by Sex and COVID-19 Test Results
Age
120
100
80
60
40
20
0
Male
Female
Sex
COVID-19 Test Result
Positive
Negative

3.2 Trực quan hóa dữ liệu



3.2 Trực quan hóa dữ liệu



TẢI VÀ TIỀN XỬ LÝ DỮ LIỆU

1. Xử lý dữ liệu trùng

2. Loại bỏ các đặc trưng không liên quan, không ảnh hưởng đến kếtquả dự đoán

3. 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.

4. Kiểm tra và loại bỏ outliers

5. Chuẩn hóa thuộc tính phân loại thành kiểu dữ liệu Binary

LINEAR REGRESSION

Công thức mô hình:

covid_res = -0.49 + 0.05*sex + 0.13*patient_type - 0.14*pneumonia + 0.003*age - 0.03*pregnancy - 0.04*diabetes + 0.106*copd + 0.045*asthma + 0.113*inmsupr - 0.006*hypertension + 0.066*other_disease + 0.09*cardiovascular + -0.08*obesity + 0.08*renal_chronic + 0.065*tobacco + -0.017*contact_other_covid

Các chỉ số chính:

- **MSE:** 0.229 *(Chỉ ra sai số bình phương trung bình giữa giá trị dự đoán và thực tế)*
- **R-squared:** 0.069 *(Chỉ 6.9% sự biến động của phản ứng COVID được giải thích bởi mô hình)*

MODELING

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

TÓM TẮT KẾT QUẢ

🏆 XGBoost đạt độ chính xác cao nhất, 0.6341. Mô hình này duy trì được sự cân bằng hợp lý giữa khả năng phân loại dương tính và âm tính.

⚠️ **Hạn chế của mô hình:**

- Tất cả các mô hình đều gặp khó khăn trong việc xác định các ca dương tính với COVID-19.
- **Recall cho Class 1** vẫn ở mức thấp, có nhiều ca dương tính bị bỏ sót. Điều này rất quan trọng vì nhận diện chính xác ca dương tính là mục tiêu chính trong bối cảnh dịch bệnh.
- Mặc dù mô hình trên phân loại các ca âm tính (Class 0) khá tốt, nhưng hiệu suất với các ca dương tính (Class 1) vẫn chưa đạt yêu cầu, đặc biệt về khả năng phát hiện (Recall).

💡 **Kết luận:**

- Dù XGBoost là mô hình tốt nhất trong loạt thử nghiệm này dựa trên các chỉ số phân loại và thời gian huấn luyện. Nhưng vẫn cần có những cải tiến để tăng cường khả năng phát hiện các ca dương tính với COVID-19.

ĐỀ XUẤT HƯỚNG PHÁT TRIỂN

- **Tinh chỉnh mô hình**: Điều chỉnh tham số của XGBoost để cải thiện khả năng phát hiện ca dương tính (Class 1).
- **Kết hợp mô hình**: Áp dụng kỹ thuật ensemble để tăng cường hiệu suất tổng thể.
- **Xử lý mất cân bằng**: Sử dụng oversampling hoặc undersampling để cải thiện recall cho Class 1.
- **Tăng cường đặc trưng**: Tạo thêm đặc trưng mới hoặc giảm chiều dữ liệu để giúp mô hình phân biệt tốt hơn giữa các ca âm tính và dương tính.

Kết luận:

XGBoost cho kết quả tốt nhất nhưng vẫn gặp khó khăn trong việc phát hiện các ca dương tính (Class 1). Việc cải thiện recall cho các ca dương tính là ưu tiên, đặc biệt quan trọng trong bối cảnh dịch bệnh. Tập trung vào tinh chỉnh mô hình và xử lý mất cân bằng dữ liệu sẽ giúp nâng cao hiệu quả trong phát hiện và kiểm soát COVID-19.