

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## TIỂU LUẬN

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

### Phương pháp nghiên cứu khoa học

***Đề tài:*** Xây dựng mô hình dự báo về diễn biến của Covid-19 trên địa bàn tỉnh Thừa Thiên Huế

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 20 tháng 01 năm 2022



ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## TIỂU LUẬN

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

### Phương pháp nghiên cứu khoa học

***Đề tài:***Xây dựng mô hình dự báo về diễn biến của Covid-19 trên địa bàn tỉnh Thừa Thiên Huế

###### Giảng viên hướng dẫn: Nguyễn Quang Lịch

###### Sinh viên thực hiện:Phan Thế Minh Châu

###### Lớp:Khoa học dữ liệu và Trí tuệ nhân tạo*.*

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 20 tháng 01 năm 2022

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ



PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN

**Học kỳ I, năm học 2021 - 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT1:**  Bằng số: .........................................  Bằng chữ: ....................................... | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT2:**  Bằng số: .........................................  Bằng chữ: ....................................... |

Điểm kết luận: ...........................................................................................................

Bằng số:.....................................................................................................................

Bằng chữ: .................................................................................................................

*Thừa Thiên Huế, ngày tháng năm 2021*

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1**  *(Ký và ghi rõ họ và tên)* | **Cán bộ chấm thi 2**  *(Ký và ghi rõ họ và tên)* |

# I/ Đặt vấn đề

### 1/ Giới thiệu về Covid-19:

COVID-19 bây giờ đã trở thành một loại virus chết người lan rộng và ảnh hưởng đến hàng chục nghìn người trên khắp thế giới. Nguồn gốc của nó là thành phố Vũ Hán, Trung Quốc vào tháng 12 năm 2019. Khi mọi người không biết về virus, COVID-19 bắt đầu lây lan từ người này sang người khác; nó đã từ từ lan đến hầu hết các quốc gia và đã trở thành một đại dịch toàn cầu.

COVID-19 là dạng viết tắt của bệnh coronavirus 2019, một bệnh do một loại coronavirus mới (nCoV) gây ra hiện nay được gọi là coronavirus 2 (SARS-CoV-2) gây hội chứng hô hấp cấp tính nghiêm trọng; trước đây được gọi là 2019-nCoV. COVID-19 không phải là tên chính thức của loại virus này; nó được gọi là SARS-CoV-2 bởi Ủy ban quốc tế về phân loại vi rút vì các triệu chứng của nó liên quan đến vi rút gây ra dịch SARS vào năm 2003. Tuy nhiên, trước đây vi rút này chưa xuất hiện ở người, và lần này, chúng đã lây nhiễm lên người nên để tránh nhầm lẫn với các virut khác, Tổ chức Y tế Thế giới (WHO) đã đặt tên là COVID-19.

Coronavirus lây truyền từ người sang người khi họ tiếp xúc trực tiếp với nhau hoặc khi người bị nhiễm bệnh hắt hơi hoặc ho. Là bệnh đường hô hấp nên ảnh hưởng trực tiếp đến hệ hô hấp. Theo CDC, nCoV được báo cáo là rất dễ lây lan, có nghĩa là nó dễ dàng lây lan giữa người với người. Nó cũng có thể lây lan khi một người chạm vào bề mặt hoặc các vật dụng ăn được đã tiếp xúc với người bị bệnh.

Các triệu chứng phổ biến nhất của COVID-19 là ho và hắt hơi, sốt và các vấn đề về hô hấp. Ngoài các triệu chứng này, còn có thể bị tiêu chảy, các vấn đề về thính giác, mất khứu giác, đau ngực và nghẹt mũi.

WHO đã đưa ra các biện pháp phòng ngừa để tránh lây nhiễm vi rút COVID-19. Bao gồm che mặt bằng khẩu trang hoặc khăn, giản cách xã hội và thực thi nguyên tắc 5K.

### 2/ Tình hình hiện tại:

Tính đến ngày 15/01/2022, tỉnh Thừa Thiên Huế ghi nhận tổng cộng 17.480 ca nhiễm Covid-19; tổng số bệnh nhân điều trị khỏi: 14.133; Tổng số bệnh nhân tử vong:85 ca.

Tình hình diễn biến đang ngày càng phức tạp, dự báo sẽ còn tiếp tục diễn biến trong thời gian tiếp theo do có sự xuất hiện của nhiều biến chủng mới. Qúa trình phòng, chống dịch Covid trên địa bàn gặp nhiều khó khăn, thách thức.

Một trong những khó khăn đó là chưa thể nhìn nhận dự báo, đánh giá tình hình chính xác để kịp thời đáp ứng các nhu cầu cấp bách, mức độ bao phủ các dịch vụ thiết yếu về phát hiện, dự phòng, quản lý điều trị còn chưa cao.

Vì vậy, cần phải có một mô hình để dự báo chính xác diễn biến của dịch bệnh để sẵn sàng đáp ứng các yêu cầu cấp bách; thực hiện tốt công tác khoanh vùng, truy vết, xét nghiệm, cách ly; triển khai nghiêm túc các quy định về cách ly y tế tập trung đối với các đối tượng có nguy cơ theo chỉ đạo của Bộ Y tế và Ban Chỉ đạo COVID-19 tỉnh; nâng cao năng lực xét nghiệm tại địa phương đáp ứng đòi hỏi mới trong công tác tầm soát cộng đồng để phòng, chống dịch. Chủ động phương án phòng, chống dịch lây lan diện rộng trên địa bàn tỉnh; chuẩn bị đầy đủ cơ sở vật chất, trang thiết bị, vật tư y tế... huy động kịp thời mọi nguồn lực phục vụ công tác phòng, chống dịch.

# II/ Tổng quan nghiên cứu

Sự hiểu biết toàn diện về đặc điểm dịch tễ học của virus Covid-19 là rất quan trọng để hạn chế sự lây lan của nó. Do đó, nghiên cứu này áp dụng các mô hình dựa trên trí tuệ nhân tạo để dự đoán mức độ phổ biến của đợt bùng phát COVID-19 ở tỉnh Thừa Thiên Huế.

Mục tiêu của nghiên cứu này là tìm ra mô hình học máy (Machine learning) hoạt động tốt nhất để dự đoán và dự báo COVID-19. Sau đó, người đọc sẽ có được cái nhìn sơ lược về một số nguyên tắc cơ bản của học máy và cách học máy có thể được sử dụng để dự đoán và dự báo COVID-19, điều này có thể giúp ích trong các nhiệm vụ tự động hóa dự đoán lây nhiễm COVID-19 và giúp dự báo khả năng lây nhiễm COVID-19 trong tương lai.

# III/ Phương pháp và nội dung nghiên cứu

ML được sử dụng trong nhiều lĩnh vực khác nhau, bao gồm y học để dự đoán bệnh tật và dự báo kết quả của nó. Trong y học, chẩn đoán đúng và đúng thời điểm là chìa khóa để điều trị thành công. Nếu phương pháp điều trị có tỷ lệ sai sót cao, nó có thể gây ra một số trường hợp tử vong. Do đó, các nhà nghiên cứu đã bắt đầu sử dụng các ứng dụng trí tuệ nhân tạo để chữa bệnh. Nhiệm vụ rất phức tạp vì các nhà nghiên cứu phải chọn đúng công cụ: đó là vấn đề sống hay chết.

Với nhiệm vụ này, ML đã đạt được một cột mốc quan trọng trong lĩnh vực chăm sóc sức khỏe. Các kỹ thuật ML được sử dụng để giải thích và phân tích các tập dữ liệu lớn và dự đoán đầu ra của chúng. Các công cụ ML này được sử dụng để xác định các triệu chứng của bệnh và phân loại mẫu thành các nhóm điều trị. ML giúp các bệnh viện duy trì các quy trình hành chính và điều trị bệnh truyền nhiễm

Kỹ thuật ML trước đây được sử dụng để điều trị ung thư, viêm phổi, tiểu đường, bệnh Parkinson, viêm khớp, rối loạn thần kinh cơ, và nhiều bệnh khác; chúng cho kết quả dự đoán và dự đoán chính xác hơn 90%.

Để dự báo thông qua ML, có thể sử dụng phân tích chuỗi thời gian, đây là một phần quan trọng của ML. Đây là kiểu hồi quy đơn biến trong đó đặc trưng mục tiêu (đặc trưng phụ thuộc) được dự báo chỉ sử dụng một đặc điểm đầu vào (đặc trưng độc lập), đó là thời gian

Nó được sử dụng để dự báo các giá trị sự kiện trong tương lai, và nó có vai trò quan trọng trong việc dự báo sự tồn tại của các bệnh đường hô hấp như COVID-19. Các trường hợp khả quan đang tăng lên hàng ngày, vì vậy cần phải dự báo liệu tỷ lệ mà số lượng đang tăng có tiếp tục dựa trên các quan sát trước đó hay không. Nó rất hữu ích cho chính phủ, bởi vì dựa trên dự báo, chính phủ có thể lập kế hoạch về các nguồn lực để kiểm soát sự lây lan của dịch bệnh và hành động cho tương lai để tốc độ phát triển của bệnh nhiễm trùng giảm mà không ảnh hưởng đến nhiều người hơn

Dự báo phụ thuộc hoàn toàn vào các xu hướng trong quá khứ, do đó không thể đảm bảo được độ chính xác tuyệt đối. Tuy nhiên, dự báo gần đúng về các sự kiện có thể giúp các cơ quan tổ chức có đánh giá việc lập kế hoạch nhiệm vụ sắp tới để kịp thời ứng phó với bất kỳ tình huống đại dịch nào như COVID-19.

Nghiên cứu này sẽ sử dụng phương pháp dự báo được sử dụng rộng rãi nhất, được gọi là mô hình ARIMA để dự báo chuỗi thời gian. ARIMA được sử dụng cho dữ liệu chuỗi thời gian để dự đoán xu hướng trong tương lai.

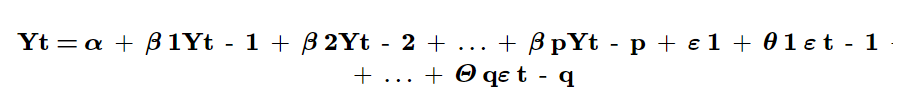
ARIMA là một dạng phân tích hồi quy đơn biến dự đoán các giá trị trong tương lai dựa trên sự khác biệt giữa các giá trị hơn là các giá trị thực tế. Nó kết hợp ba thuật ngữ:

* tự động hồi phục (AR), hiển thị việc thay đổi một biến hoàn nguyên các giá trị trên cơ sở các giá trị trước đó của nó;
* tích hợp (I), hoặc thay thế các giá trị dữ liệu trên cơ sở sự khác biệt giữa các giá trị dữ liệu và các giá trị trước đó
* đường trung bình động (MA), mức trung bình liên tiếp được lấy trên các khung thời gian liên tiếp có kích thước không đổi của một chuỗi thời gian có sẵn trước đó.

ARIMA sử dụng phương trình dự báo pdq trong đó các tham số này được xác định là:

* p là số quan sát bị trễ;
* d xác định thời gian (tức là, các quan sát thô khác nhau bao nhiêu lần); và
* q xác định kích thước của cửa sổ MA.

Trong mô hình ARIMA, giá trị của Yt phụ thuộc vào giá trị trễ của chính nó. Trong mô hình MA, giá trị của Yt phụ thuộc vào sai số dự báo trễ. Do đó, phương trình tổng quát của ARIMA là:



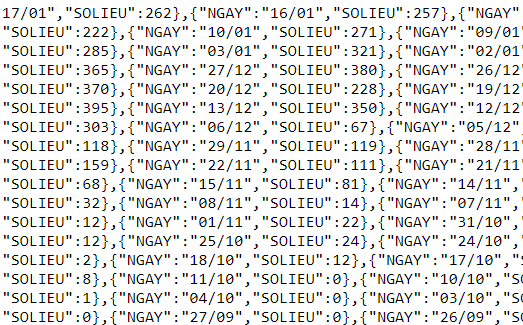
trong đó Yt là mục tiêu được dự đoán, α là giá trị không đổi, β1Yt - 1 là độ trễ kết hợp tuyến tính của Y được tính đến p độ trễ và ϴ1εt - 1 là kết hợp tuyến tính của lỗi dự báo trễ được tính đến q trễ.

## Thu thập dữ liệu:

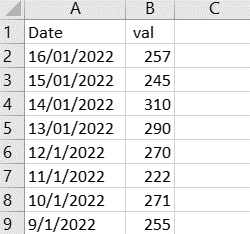
Dữ liệu được thu thập từ trang web Cập nhật tình hình dịch bệnh Covid-19 trên địa bàn tỉnh Thừa Thiên Huế <http://covid.thuathienhue.gov.vn/>

Dữ liệu gồm 1 tập dataset ghi lại số ca mắc mới trên địa bàn toàn tỉnh Thừa Thiên Huế trong ngày từ ngày 2/8/2021 đến ngày 16/1/2022.

Dữ liệu thô lấy về được dưới dạng file .json



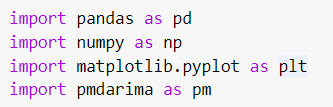
Từ đó sẽ chuyễn sang file .csv đễ dễ dàng phân tích



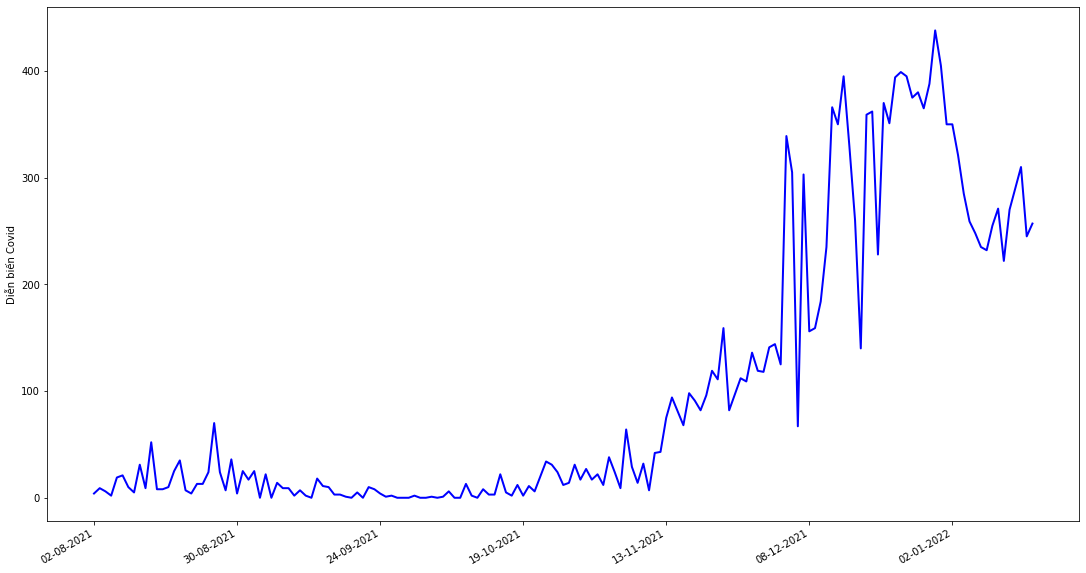
## Phân tích dữ liệu:

Sử dụng ngôn ngữ lập trình Python để tiến hành xử lý dữ liệu và xây dựng mô hình dự báo với phương pháp ARIMA.

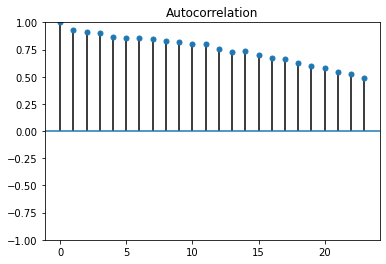
Import các thư viện cần thiết



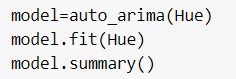
Biểu đồ diễn biến covid:

****

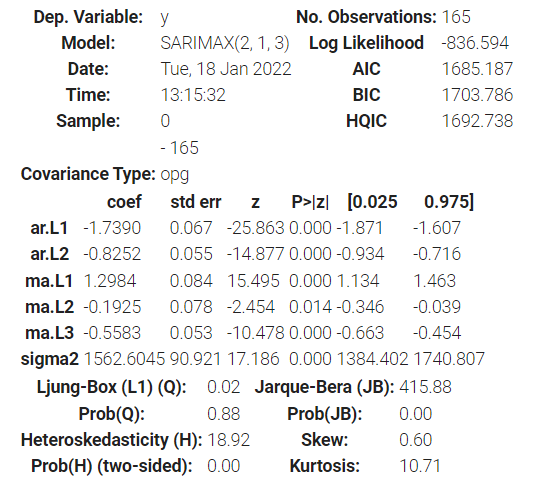
Biểu đồ tương quan của ACF:



## Thiết lập mô hình và đào tạo nó trên dữ liệu đầu vào



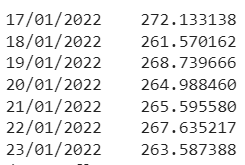
Kết quả của SARIMAX:

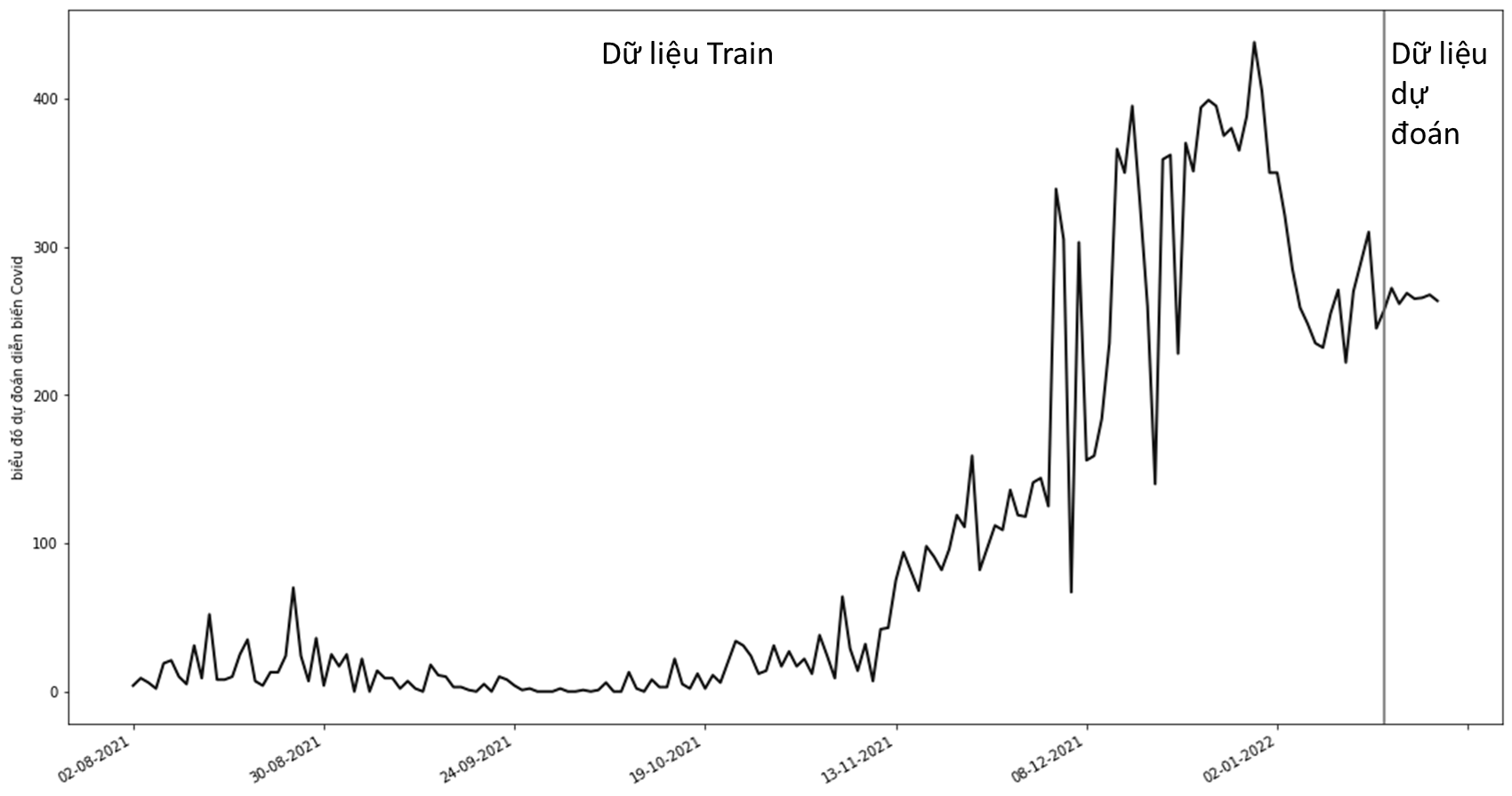


## Dựa trên mô hình đã trade được, tiến hành dự báo COVID-19 ở tỉnh Thừa Thiên Huế:

\*\*Dự báo 7 ngày tiếp theo:

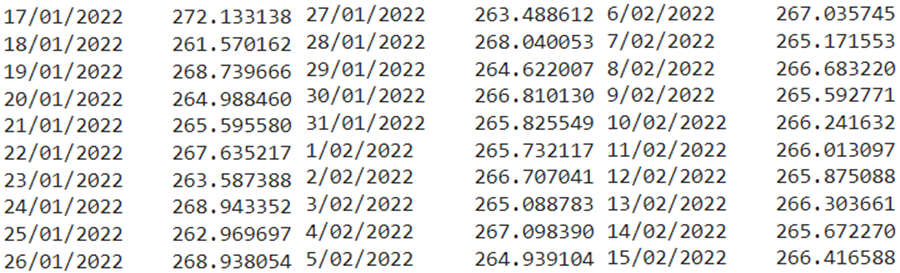
Kết quả:

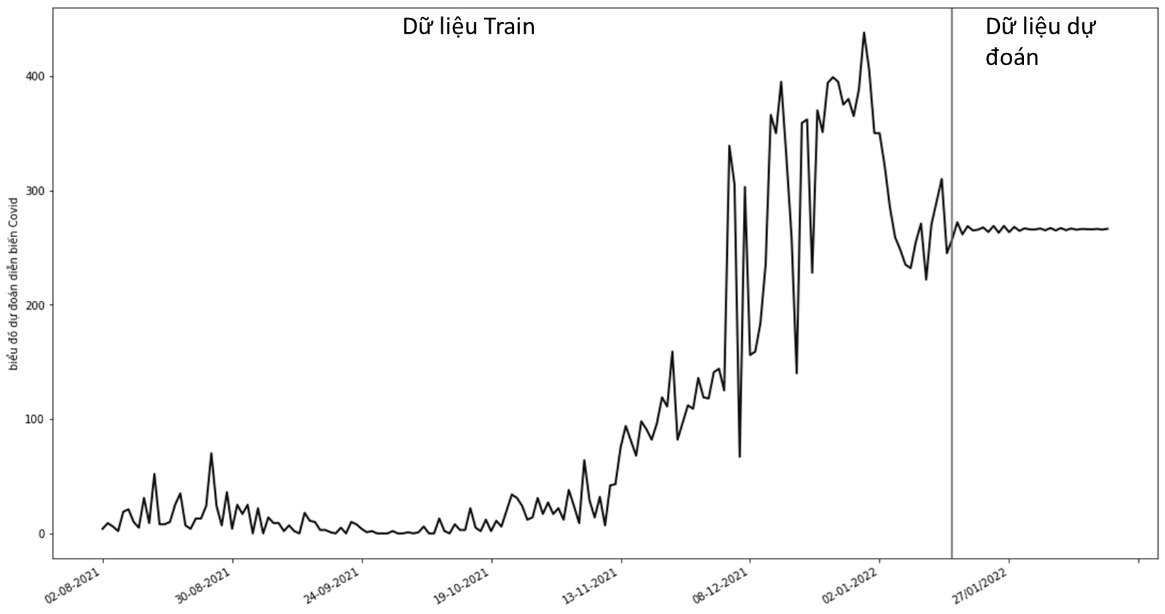




\*\*Dự báo 30 ngày tiếp theo:

Kết quả:

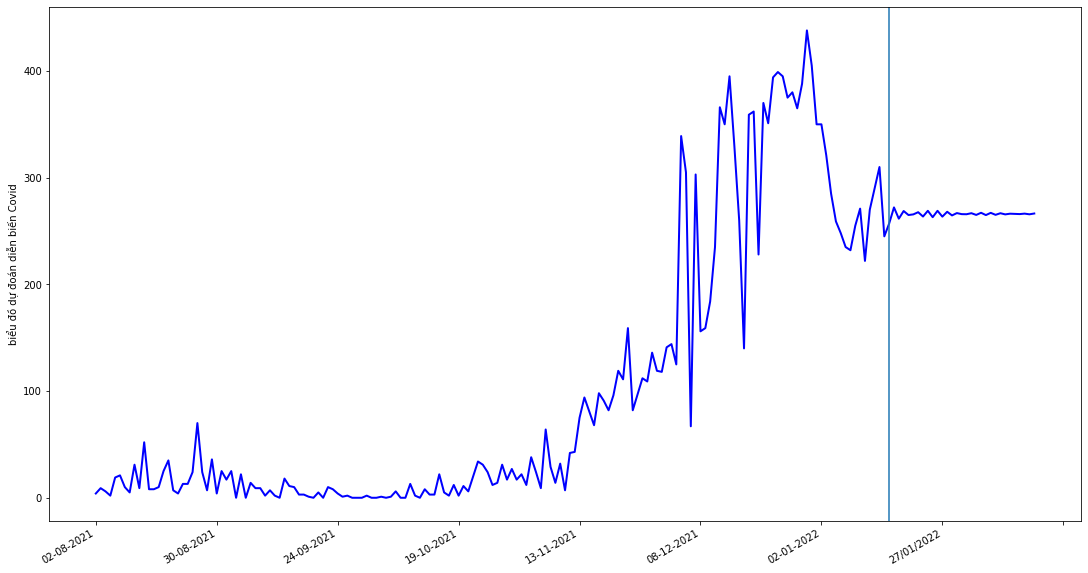




# IV/ Kết quả nghiên cứu và thảo luận

Kết quả, mô hình dự báo COVID-19 trên địa bàn tỉnh Thừa Thiên Huế đã đưa ra được kết quả dự báo dựa trên dữ liệu đầu vào với phương pháp học máy ARIMA.

Nhìn vào kết quả dự báo của 2 trường hợp trên ta có thể thấy kết quả dự báo 7 ngày có biến động nhiều hơn so với kết quả dự báo 30 ngày. Ở kết quả dự báo 30 ngày, đường thẳng khá duỗi thẳng càng về sau, ta có thể nhận định được ở trường hợp này là độ chính xác không cao.



Với phương pháp ARIMA, mô hình dự báo có hiệu quả khi kích thước dự đoán nhỏ hơn hoặc bằng 5% so với kích thước dữ liệu dùng để Train mô hình. Ví dụ, ở mô hình trên, dữ liệu dùng để Train mô hình là từ ngày 01/08/2021 đến ngày 16/01/2022 nên kích thước của data Train là 167. Vì vậy mô hình trên chỉ có hiệu quả khi dự đoán trong vòng 8 ngày. Do đó ta chỉ có thể chấp nhận được kết quả dự báo 7 ngày và loại bỏ kết quả 30 ngày do kích thước dự báo lớn hơn 5% so với dữ liệu Train.

Để tăng mức độ hiệu quả của mô hình dự báo bằng phương pháp ARIMA, ta cần có một tập dữ liệu tương đối lớn. Khi đó kết quả dự báo sẽ có ý nghĩa đối với những dự báo có kích thước dài hơn.

# V/ Kết luận và kiến nghị

Đại dịch COVID-19 đã gây ảnh hưởng lớn đến nhiều mặt trong xã hội từ nền kinh tế đến sức khỏe của người dân cả nước nói chung và tỉnh Thừa Thiên Huế nói riêng. Nó đã lây nhiễm cho 17.480 người và gây tư vong cho 85 người trên địa bàn tỉnh Thừa Thiên Huế tính đến ngày 15/1/2022. Các nhà khoa học đã nỗ lực hết sức để tìm ra giải pháp cho nó. Việc sử dụng máy tính của các nhà khoa học để dự đoán sớm đã phổ biến. Rất nhiều nghiên cứu đang được tiến hành sử dụng học máy để chống lại COVID-19. Các nhà nghiên cứu khác nhau có thể sử dụng chương này để tìm hiểu cách học máy có thể được sử dụng để dự báo không chỉ tình huống này mà còn các trường hợp khác. Nghiên cứu này đặc biệt sử dụng phương pháp ARIMA về thời gian để dự báo tình hình và diễn biến của COVID-19 trên địa bàn tỉnh Thừa Thiên Huế. Nhiều quốc gia đã chứng kiến ​​tổng số ca tử vong cao do COVID-19. Người ta tin rằng hiệu suất của mô hình có thể được cải thiện hoặc mô hình có thể cung cấp dữ liệu chính xác hơn nếu có nhiều bộ dữ liệu hơn. Mô hình cho kết quả trên cơ sở dữ liệu được xây dựng bởi thông tin do các cơ quan y tế đưa ra. Do đó, dự báo có thể không chính xác 100%, nhưng nó chắc chắn có thể được sử dụng như một biện pháp điều chỉnh. Đối với công việc trong tương lai, việc nâng cao hơn nữa có thể được thực hiện bằng cách kết hợp các yếu tố và thuật toán mới với ARIMA để có được kết quả chính xác hơn.

# VI/ Tài liệu tham khảo

* <http://covid.thuathienhue.gov.vn/>

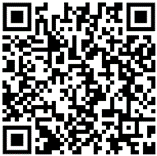


* https://en.wikipedia.org/wiki/Autoregressive\_integrated\_moving\_average



* <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>



**\*\*\* Source code mô hình:** [https://colab.research.google.com/drive/1uki7nsdCuVSJeStjYfazOaLDXyvLouJo](https://colab.research.google.com/drive/1uki7nsdCuVSJeStjYfazOaLDXyvLouJo?usp=sharing)