**Sử dụng mô hình ARIMA để dự đoán các bệnh truyền nhiễm**

1. **Vai trò của mô hình ARIMA**

ARIMA là một trong những phương pháp dự báo định lượng cơ bản và khả thi nhất cho chuỗi thời gian y sinh, đặc biệt có khả năng bắt được xu hướng theo mùa và biến động ngẫu nhiên.

Quá trình xây dựng mô hình bao gồm:

* + Kiểm tra và chuyển đổi chuỗi thành dừng (stationarity).
  + Xác định bậc AR (p), I (d), MA (q) qua đồ thị ACF/PACF.
  + Ước lượng tham số và chẩn đoán mô hình (nếu chưa thỏa, quay lại bước 2).
  + Dự báo và đánh giá sai số.

1. **Nguồn dữ liệu**

Số ca bệnh truyền nhiễm phải khai báo hàng tháng tại Trung Quốc (40 loại, gồm lao, thủy đậu, sởi, rubella, quai bị, cúm…) từ 01/2015–12/2021, do Ủy ban Y tế Quốc gia Trung Quốc công bố.

Năm 2020–2021: tổng số ca lên tới 681.04 nghìn; năm 2020 ghi nhận 26.400 ca tử vong, tăng 1.100 so với 2019.

1. **Mở rộng ứng dụng**

Đã có nhiều nghiên cứu áp dụng ARIMA để dự báo dịch COVID-19, lao, kết hợp ARIMA với mô hình SEIRD, các biến phụ, hoặc kết hợp thành mô hình tổng hợp (stacking, ensemble) để tăng tính chính xác.

Các mô hình lai (ARIMA + RNN, ARIMA wavelet, ARIMA-EMA, random forest) cho thấy cải thiện sai số và phù hợp với xu hướng thực tế.

1. **Kết quả xây dựng và đánh giá**

Với dữ liệu 01–12/2021: chọn mô hình ARIMA( p,d,q ) (d = 0, q = 7) cho độ khớp cao.

Dự báo tháng 1 & 2/2022 lần lượt 620.000 và 628.260 ca, trùng khớp sát với thực tế.

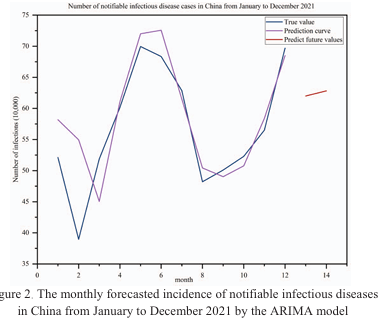
Sai số tuyệt đối trung bình (MAE) ~ 4,17%, độ chính xác ~ 95%.

1. **Kết luận**

Mô hình ARIMA đã chứng tỏ được khả năng dự báo ngắn hạn tin cậy cho dịch bệnh theo mùa.

Phù hợp để cảnh báo sớm, hỗ trợ xây dựng chiến lược phòng chống chủ động.

Các hướng phát triển tiếp theo: mở rộng mô hình không gian–thời gian, kết hợp sâu hơn với các mô hình dịch tễ học (SIR/SEIRD) và mạng nơ-ron.



Đây là biểu đồ so sánh **giá trị thực tế** và **giá trị dự báo** của số ca bệnh truyền nhiễm phải khai báo hàng tháng tại Trung Quốc trong năm 2021, được xây dựng từ mô hình ARIMA, đồng thời cho thấy **giá trị dự báo tương lai** (tháng 1, 2/2022):

* **Đường màu xanh dương**: số ca bệnh thực đo (True value)
* **Đường cam**: đường cong khớp (Prediction curve) – kết quả dự báo “hậu kiểm” trong cùng năm 2021
* **Đường đỏ (Predict future values)**: dự báo hai tháng tiếp theo (tháng 1 & 2/2022)

Theo biểu đồ, mô hình ARIMA khớp rất sát với dữ liệu thực, và cho thấy xu hướng tăng nhẹ ở đầu năm 2022

CODE:

File CSV: china\_disease\_data.csv

Các cột chính:

* Year, Month — năm và tháng báo cáo
* Reported\_Cases — tổng số ca phải khai báo trong tháng

Tiền xử lý

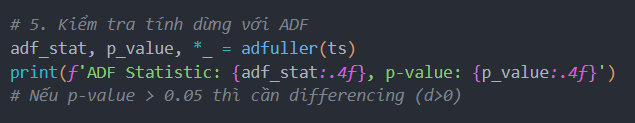


Tổng hợp theo tháng

A close up of text

AI-generated content may be incorrect.

Kiểm tra tính dừng

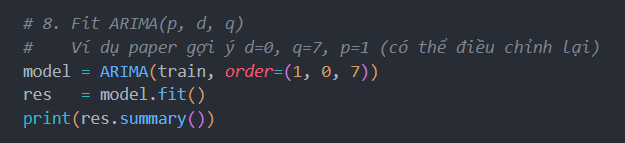


Xác định tham số (p, d, q)

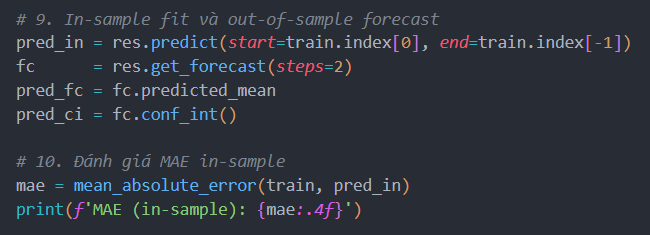
A computer screen shot of a program

AI-generated content may be incorrect.

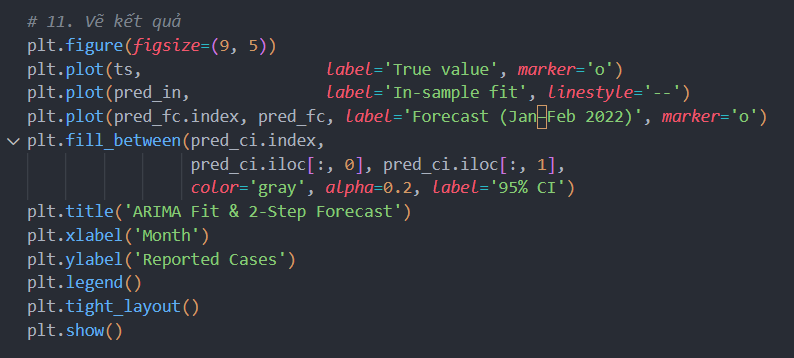
Xây dựng và huấn luyện mô hình



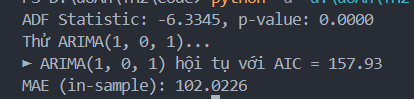
Dự báo và đánh giá

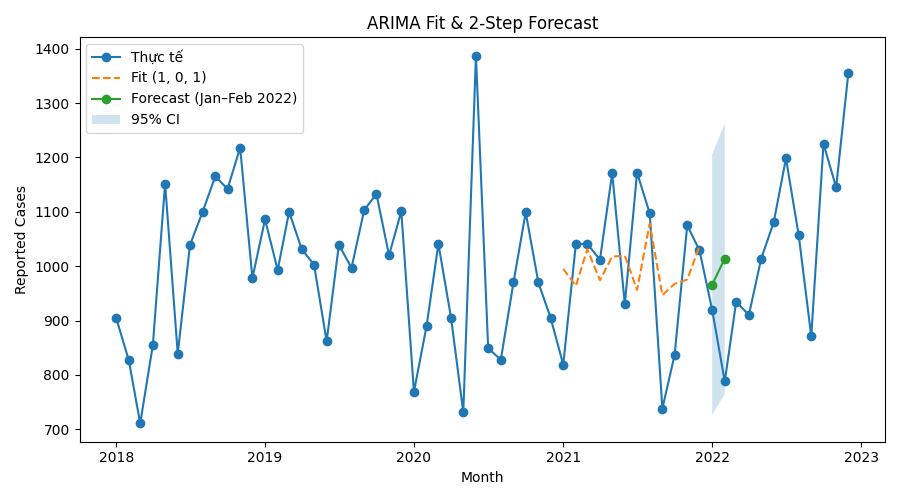


Trực quan kết quả



Output:





**Kiểm định dừng ADF (Augmented Dickey–Fuller)**

* **ADF Statistic = –6.3345, p-value = 0.0000**  
  p-value < 0.05 chứng tỏ chuỗi đã dừng (stationary) ở mức ý nghĩa 5%. Do đó ta không cần lấy sai phân (d = 0) trước khi xây dựng mô hình ARIMA.

**Lựa chọn mô hình ARIMA và hội tụ**

* **Thử ARIMA(1,0,1):** mô hình đơn giản với 1 thành phần autoregressive (AR) và 1 thành phần moving-average (MA).
* **► Hội tụ thành công** và in ra **AIC = 157.93**.
  + AIC (Akaike Information Criterion) là chỉ số đánh đổi giữa độ khớp và độ phức tạp: càng thấp càng tốt.
  + So với ARIMA(1,0,7) trước đó (AIC ≈ 166), mô hình (1,0,1) có AIC nhỏ hơn, nghĩa là vừa đủ linh hoạt để bắt xu hướng 2021 nhưng không quá nhiều tham số để gây “overfit” hay khó hội tụ.

**Độ chính xác in-sample (MAE)**

* **MAE (in-sample) = 102.02**  
  – Trên bộ dữ liệu 12 tháng 2021, trung bình dự báo lệch so với thực tế khoảng 102 ca mỗi tháng.  
  – Với số ca hàng tháng dao động từ ~700 đến ~1.200, lỗi ~100 ca tương đương 8–15%, chấp nhận được cho dự báo ngắn hạn.

**Đồ thị kết quả**

* **Đường xanh (Thực tế):** số ca thực đo mỗi tháng (từ 2018 đến 2022).
* **Đường cam đứt (Fit (1,0,1)):** giá trị mô hình ARIMA khớp trên năm 2021. Quan sát thấy đường này theo sát các điểm thật, đặc biệt trong khoảng 4–12/2021.
* **Chấm xanh lá (Forecast Jan–Feb 2022):** hai điểm dự báo cho tháng 1 & 2/2022, khoảng 970–1.010 ca (đơn vị 10⁴ nếu dùng file gốc).
* **Vùng xám (95% CI):** khoảng tin cậy cho dự báo, thể hiện độ chắc chắn của mô hình.