|  |  |  |
| --- | --- | --- |
|  | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN

KHOA HỌC DỮ LIỆU

**TÊN ĐỀ TÀI**

DỰ ĐOÁN THỜI TIẾT TẠI ĐÀ NẴNG

Giảng viên chấm:

|  |  |  |
| --- | --- | --- |
| HỌ VÀ TÊN SINH VIÊN | LỚP HỌC PHẦN | ĐIỂM BẢO VỆ  (do GV chấm ghi) |
| Nguyễn Phước Quốc | 18N14 |  |
| Huỳnh Trần Khánh Toàn | 18N14 |  |
| Mai Thê Viễn | 18N14 |  |

ĐÀ NẴNG, 05/2021

# TÓM TẮT

Vấn đề cần giải quyết :

Khi nói về thời tiết mai sẽ ra sau ông bà ta sẽ nhìn lên trời và dự đoán ngày mai sẽ như thế này như thế kia sẽ có những lúc chính xác những lúc sai lầm, nên việc dự đoán thời tiết vào những ngày tiếp theo sẽ rất quan trọng nhất là cho các ngành nông lâm ngư nghiệp.

Vậy mục tiêu của nhóm là dự đoán thời tiết những ngày tiếp theo ở TP. Đà Nẵng để có thể giúp cho các bác nông dân, ngư dân có một mùa thu hoạch thành công.

Phương pháp giải quyết :

Nhóm cần dự đoán được nhiệt độ những ngày tiếp theo dựa trên chuỗi nhiệt độ thời gian trước đó để giúp cho ngành nông lâm ngư nghiệp có một kế hoạch tốt cho mùa vụ. Ở đây chúng em sử dụng dụng 2 mô hình để dự đoán nhiệt độ ngày tiếp theo :

* RNN (Recurrent neural network)
* ARIMA (Autoregressive Intergrated Moving Average)

Kết quả đạt được :

MAPE: Sai số giữa giá trị dự báo so với giá trị thực tế sấp sỉ 3,5%(ARIMA) và 2% (RNN) giá trị thực tế.

# BẢNG PHÂN CÔNG NHIỆM VỤ

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Mai Thế Viễn | * Thu thập dữ liệu từ trang web https://freemeteo.vn/thoi-tiet/. * Làm sạch dữ liệu sau khi craw. * Thống kê mô tả trực quan dữ liệu. * Đưa dữ liệu ra file excel. | Đã hoàn thành |
| Huỳnh Trần Khánh Toàn | * Lựa chọn đặc trưng. * Làm sạch dữ liệu. * Kiểm tra và tiền xử lý dữ liệu. * Trực quan dữ liệu. | Đã hoàn thành. |
| Nguyễn Phước Quốc | * Lựa chọn mô hình. * Chia dữ liệu thành train/ test phù hợp. * Trình bày các đồ thị thể hiện hiệu quả của các mô hình trong quá trình huấn luyện/Kiểm thử. * So sánh hiệu quả của các mô hình bằng đồ thị và bảng dữ liệu. | Đã hoàn thành |

**MỤC LỤC**

[TÓM TẮT 2](#_Toc73010503)

[BẢNG PHÂN CÔNG NHIỆM VỤ 3](#_Toc73010504)

[1. Giới thiệu 6](#_Toc73010505)

[2. Thu thập và mô tả dữ liệu 7](#_Toc73010506)

[2.1. Thu thập dữ liệu 7](#_Toc73010507)

[Phương pháp craw dữ liệu : 7](#_Toc73010508)

[2.2. Mô tả dữ liệu 8](#_Toc73010509)

[3. Trích suất dữ liệu 9](#_Toc73010510)

**Danh sách hình ảnh**

[Hình 1: Dữ liệu ở dạng JSON 7](#_Toc73011436)

[Hình 2: Dữ liệu dạng file Excel 8](#_Toc73011437)

[Hình 3: Bảng dữ liệu sau khi craw 9](#_Toc73011438)

[Hình 4: Dữ liệu sau khi làm sạch 9](#_Toc73011439)

[Hình 5: Dữ liệu sau khi đã xử lý 10](#_Toc73011440)

[Hình 6: Kiểm tra tính phân bố của dữ liệu 10](#_Toc73011441)

[Hình 7: Đồ thị dữ liệu 11](#_Toc73011442)

[Hình 8: Mô hình RNN 11](#_Toc73011443)

[Hình 9: Đồ thị dự báo bằng mô hình RNN 14](#_Toc73011444)

[Hình 10: Phóng to phần dự báo 14](#_Toc73011445)

[Hình 11: Đồ thị của dữ liệu 15](#_Toc73011446)

[Hình 12: Chuyển dữ liệu về dạng tháng 16](#_Toc73011447)

[Hình 13: Phân rã dữ liệu theo phân rã cộng 17](#_Toc73011448)

[Hình 14: Đồ thị dự báo của mô hình ARIMA 18](#_Toc73011449)

[Hình 15: Bảng so sánh kết quả đạt được cảu 2 mô hình 18](#_Toc73011450)

# Giới thiệu

Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Sở dĩ chúng ta gọi lớp mô hình này là chuỗi thời gian (*time series*) là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Một mô hình chuỗi thời gian thường dự báo dựa trên giả định rằng các qui luật trong quá khứ sẽ lặp lại ở tương lai. Do đó xây dựng mô hình chuỗi thời gian là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Do là dữ liệu chịu ảnh hưởng bởi tính chất thời gian nên chuỗi thời gian thường xuất hiện những qui luật đặc trưng như: yếu tố chu kỳ, mùa vụ và yếu tố xu hướng. Đây là những đặc trưng thường thấy và xuất hiện ở hầu hết các chuỗi thời gian :

* Yếu tố chu kỳ, mùa vụ là những đặc tính lặp lại theo chu kỳ. Ví dụ như nhiệt độ trung bình các tháng trong năm sẽ chịu ảnh hưởng bởi các mùa xuân, hạ, thu, đông. Hay xuất nhập khẩu của một quốc gia thường có chu kỳ theo các quí.
* Yếu tố xu hướng (*trend*) thể hiện đà tăng hoặc giảm của chuỗi trong tương lai. Chẳng hạn như lạm phát là xu hướng chung của các nền kinh tế, do đó giá cả trung bình của giỏ hàng hóa cơ sở hay còn gọi là chỉ số CPI luôn có xu hướng tăng và xu hướng tăng này đại diện cho sự mất giá của đồng tiền.

Và với bài toán dự báo như thế này, chúng em sẽ quy nó về Time Series Analysis để có thể dễ hiểu và dễ áp dụng.

Input: thông số nhiệt độ của những ngày hôm trước.

Output: thông số nhiệt độ của những ngày hôm sau.

# 2. Thu thập và mô tả dữ liệu

## 2.1. Thu thập dữ liệu

Hiện nay ứng dụng web được xây dựng theo 2 loại chính đó là:

* SPA (single page application): khi view-source ra thì toàn thấy các đọan mã script. Các trang SPA thì việc crawl sẽ nhận dữ liệu từ API.
* MPA( multiple page application) : khi view-source thì chúng ta sẽ thấy tất cả dữ liệu ở trong source. Để crawl các trang web được xây dựng theo MPA thì cần phải bóc tách dữ liệu từ HTML khá là phức tạp.

### Phương pháp craw dữ liệu :

* Công nghệ sử dung: NodeJS, cheerio, request-promise, fs
* Cheerio: hỗ trợ parse DOM giống như là jQuery, để truy xuất đối tượng
* Request-promise: hợ trợ việc lấy mã nguồn của trang cần cào,
* Fs: thư viện hỗ trợ làm việc với file trong Node.js
* Các bước xử lí
* Bước 1: Lấy dữ liệu trừ trang web thông qua request-promise
* Bước 2: Xử lý dữ liệu, lấy nhưng dữ liệu cần..
* Bước 3: Xuất file json
* Bước 4: Chuyển đổi các file json thành file excel
  + Dữ liệu được craw là kết quả dự báo thời tiết qua các ngày từ tháng 4 năm 2021 đến tháng 11 năm 2012 (1984 kết quả)

Text

Description automatically generated

Hình 1: Dữ liệu ở dạng JSON

* + Chuyển đổi các file json thành excel

A picture containing text, window

Description automatically generated

Hình 2: Dữ liệu dạng file Excel

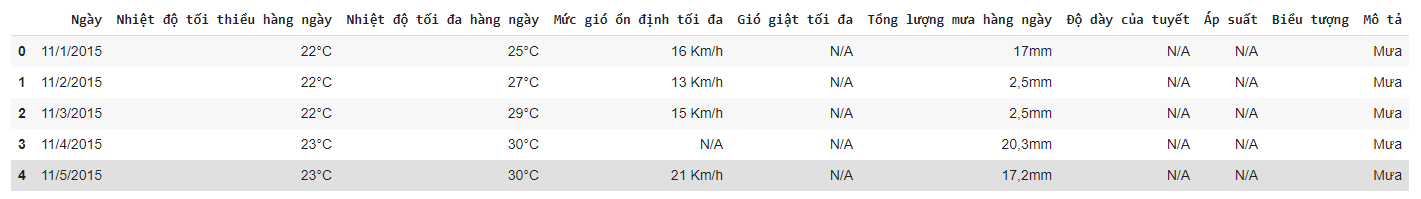
## 2.2. Mô tả dữ liệu

Có tất cả 1984 mẫu đã được craw.

Dữ liệu nhóm em craw bao gồm các trường sau :

1. Date (ngày)
2. Temperature (Nhiệt độ)
3. Relative temperature (Nhiệt độ tối đa)
4. Wind (độ giật tối đa)
5. Relative humidity (Độ ẩm)
6. Dewpoint (Điểm sương)
7. Pressure (áp suất)
8. Detailed description (Mô tả chi tiết)

# Trích suất dữ liệu

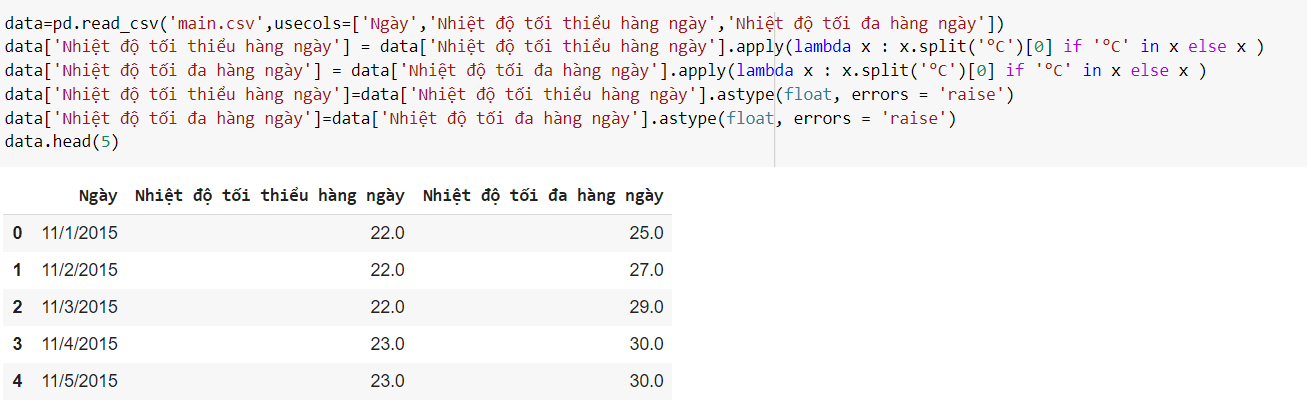


Hình 3: Bảng dữ liệu sau khi craw

Vì nhóm sử dụng dữ liệu chuỗi thời gian để dự báo nhiệt độ những ngày tiếp theo nên đăc trưng chính nhóm chọn sẽ là nhiệt độ tối thiểu hằng ngày và nhiệt độ tối đa hằng ngày.

* 1. **Làm sạch dữ liệu:**

Trước khi xử lý và làm sạch dữ liệu chúng em đã kiểm tra giá trị trong 2 “cột nhiệt độ tối thiểu hàng ngày” và “nhiệt độ tối đa hàng ngày” và kết quả cho thấy không có vấn đề về “missing value”. Dữ liệu thu thập được khá đầy đủ nhưng ta cần đưa giá trị này về số thực để có thể sử dụng mô hình xử lý số liệu và bỏ những cột không cần thiết.



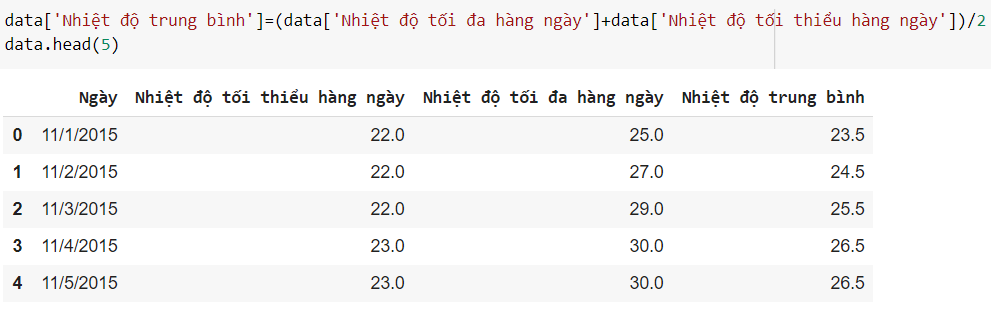
Hình 4: Dữ liệu sau khi làm sạch

* 1. **Tiền xử lý và kiểm tra phân bố dữ liệu:**

Đưa dữ liệu cột “Nhiệt độ tối thiểu hàng ngày” và “Nhiệt độ tối đa hàng ngày” về kiểu số thực để thuận tiện cho việc tính toán cũng như áp dụng các mô hình của chuỗi thời gian.

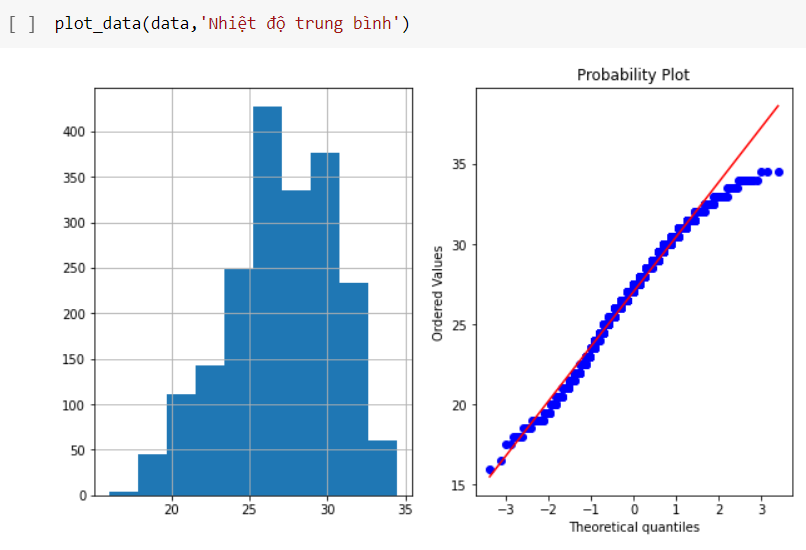
Chuyển dữ liệu cột “Ngày” sang kiểu dữ liệu datetime.

Ta sẽ tính trung bình của cột “Nhiệt độ tối thiểu hàng ngày” và “Nhiệt độ tối đa hàng ngày” để tạo thành cột “Nhiệt độ trung bình”.



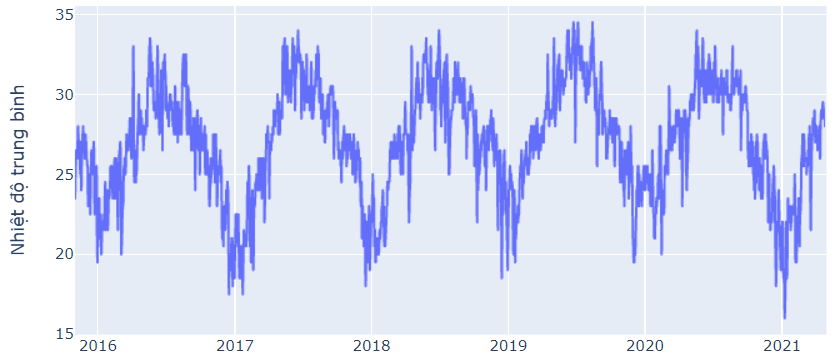
Hình 5: Dữ liệu sau khi đã xử lý

Kiểm tra dữ liệu ở cột nhiệt độ trung bình ta thấy dữ liệu đã có dạng phân bố chuẩn thích hợp để áp dụng các mô hình của chuỗi thời gian.



Hình 6: Kiểm tra tính phân bố của dữ liệu

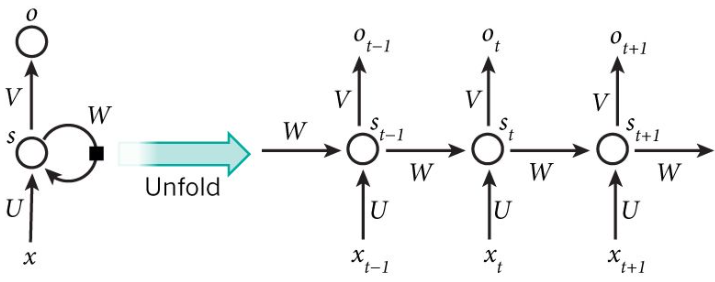
* 1. **Trực quang dữ liệu sau khi đã xử lý:**



Hình 7: Đồ thị dữ liệu

1. **Mô hình hóa dữ liệu**
   1. **Cơ sở lý thuyết mô hình học máy**
      1. **Mô hình RNN:**

Ý tưởng chính của RNN (Recurrent Neural Network) là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau. Tức là chúng không liên kết thành chuỗi với nhau. Nhưng các mô hình này không phù hợp trong rất nhiều bài toán. Ví dụ, nếu muốn đoán từ tiếp theo có thể xuất hiện trong một câu thì ta cũng cần biết các từ trước đó xuất hiện lần lượt thế nào chứ nhỉ? RNN được gọi là hồi quy (Recurrent) bởi lẽ chúng thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó. Trên lý thuyết, RNN có thể sử dụng được thông tin của một văn bản rất dài, tuy nhiên thực tế thì nó chỉ có thể nhớ được một vài bước trước đó mà thôi. Về cơ bản một mạng RNN có dạng như sau:



Hình 8: Mô hình RNN

Mô hình trên mô tả phép triển khai nội dung của một RNN. Triển khai ở đây có thể hiểu đơn giản là ta vẽ ra một mạng nơ-ron chuỗi tuần tự. Ví dụ ta có một câu gồm 5 chữ “Đẹp trai lắm gái theo”, thì mạng nơ-ron được triển khai sẽ gồm 5 tầng nơ-ron tương ứng với mỗi chữ một tầng. Lúc đó việc tính toán bên trong RNN được thực hiện như sau:

* *xt*​ là đầu vào tại bước *t*. Ví dụ, *x*1​ là một vec-tơ one-hot tương ứng với từ thứ 2 của câu (trai).
* *st*​ là trạng thái ẩn tại bước *t*. Nó chính là **bộ nhớ** của mạng. *st*​ được tính toán dựa trên cả các trạng thái ẩn phía trước và đầu vào tại bước đó: *st*​=*f*(*Uxt*​+*Wst*−1​). Hàm *f* thường là một hàm phi tuyến tính như [tang hyperbolic (tanh)](https://vi.wikipedia.org/wiki/H%C3%A0m_hypebolic) hay [ReLu](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)). Để làm phép toán cho phần tử ẩn đầu tiên ta cần khởi tạo thêm *s*−1​, thường giá trị khởi tạo được gắn bằng 0.
* *ot*​ là đầu ra tại bước *t*. Ví dụ, ta muốn dự đoán từ tiếp theo có thể xuất hiện trong câu thì *ot*​ chính là một vec-tơ xác xuất các từ trong danh sách từ vựng của ta:  *ot*=softmax(*Vst*)
  + 1. **Mô hình ARIMA:**

ARIMA model là viết tắt của cụm từ Autoregressive Intergrated Moving Average. Mô hình sẽ biểu diễn phương trình hồi qui tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 2 thành phần chính:

* **Auto regression**: Kí hiệu là AR. Đây là thành phần tự hồi qui bao gồm tợp hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình AR(p) của chuỗi xt được biểu diễn như bên dưới:

AR(p)=ϕ0+ϕ1xt−1+ϕ2xt−2+⋯+ϕpxt−p

* **Moving average**: Qúa trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuổi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Qúa trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên ϵt (stochastic term). Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất:

Vế (1) có nghĩa rằng kì vọng của chuỗi bằng 0 để đảm bảo chuỗi dừng không có sự thay đổi về trung bình theo thời gian. Vế (2) là phương sai của chuỗi không đổi. Do kì vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution) và được kí hiệu là ϵt∼WN(0,σ2). Nhiễu trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của model và không có tính qui luật. Qúa trình trung bình trượt được biểu diễn theo nhiễu trắng như sau:

MA(q)=μ+

Quá trình này có thể được biểu diễn theo [dịch chuyển trễ - backshift operator](https://en.wikipedia.org/wiki/Lag_operator) B như sau:

MA(q)=μ+(1+θ1B+⋯+θqBq)ϵt

Như vậy bạn đọc đã hình dung ra moving average là gì rồi chứ? Về mặt ý tưởng thì đó chính là quá trình hồi qui tuyến tính của giá trị hiện tại theo các giá trị hiện tại và quá khứ của sai số nhiễu trắng (white noise error term) đại diện cho các yếu tố shock ngẫu nhiên, những sự thay đổi không lường trước và giải thích bởi mô hình.

* **Intergrated**: Là quá trình đồng tích hợp hoặc lấy sai phân. Yêu cầu chung của các thuật toán trong time series là chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Một số chuỗi tài chính còn qui đổi sang logarit hoặc lợi suất. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration). Quá trình sai phân bậc d của chuỗi được thực hiện như sau:
* Sai phân bậc 1: I(1)=Δ(xt)=xt−xt−1
* Sai phân bậc d: I(d)=Δd(xt)=Δ(Δ(…Δ(xt))) d times

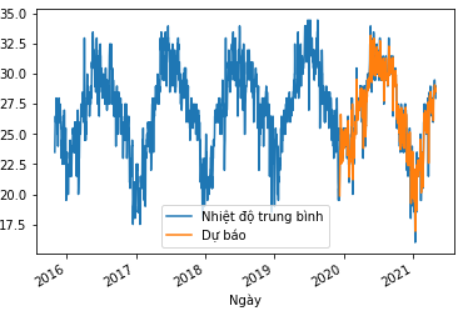
Thông thường chuỗi sẽ dừng sau quá trình đồng tích hợp I(0) hoặc I(1). Rất ít chuỗi chúng ta phải lấy tới sai phân bậc 2. Một số trường hợp chúng ta sẽ cần biến đổi logarit hoặc căn bậc 2 để tạo thành chuỗi dừng. Phương trình hồi qui ARIMA(p, d, q) có thể được biểu diễn dưới dạng:

Δxt=ϕ1Δxt−1+ϕ2Δxt−2+...+ϕpΔxt−p+θ1ϵt−1+θ2ϵt−2+...+θqϵt−q

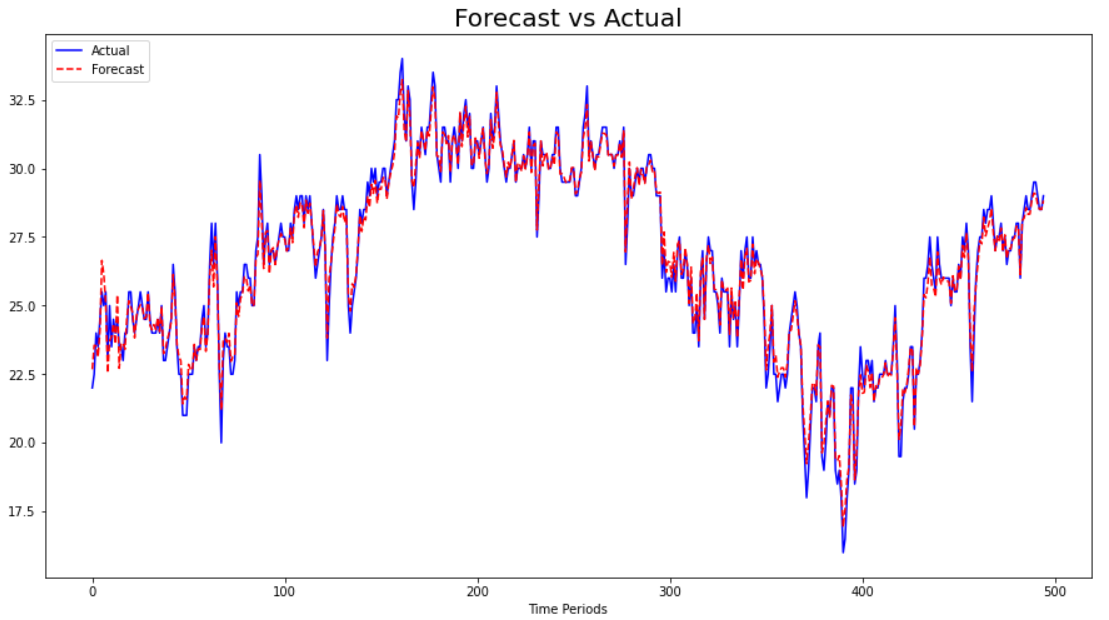
Trong đó Δxt là giá trị sai phân bậc d và ϵt là các chuỗi nhiễu trắng.

* 1. **Chia mô hình thành train và test**
     1. **Mô hình RNN:**
* Chia dữ liệu thành tập train và test để input vào thuật toán:
* Dữ liệu tập test: từ ngày 10/12/2019 đến 31/4/2021
* Dữ liệu tập train: từ ngày 1/11/2015 đến 10/12/2019
* Các thông số của mô hình:
* Một mạng RNN bao gồm 1 lớp đầu vào, 1 lớp đầu ra và 100 lớp ẩn
* Hàm activation: ReLU
* learning\_rate = 0.001
* Hàm loss: Hàm MAE
* Hàm optimizer: Adam

Kết quả đạt được:



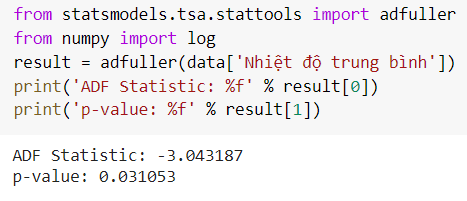
Hình 9: Đồ thị dự báo bằng mô hình RNN



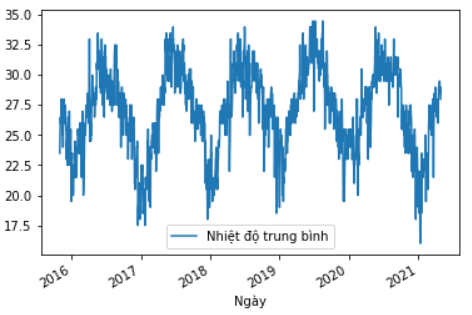
Hình 10: Phóng to phần dự báo

* + 1. **Mô hình ARIMA**

Kiểm tra tính ổn định của dữ liệu:

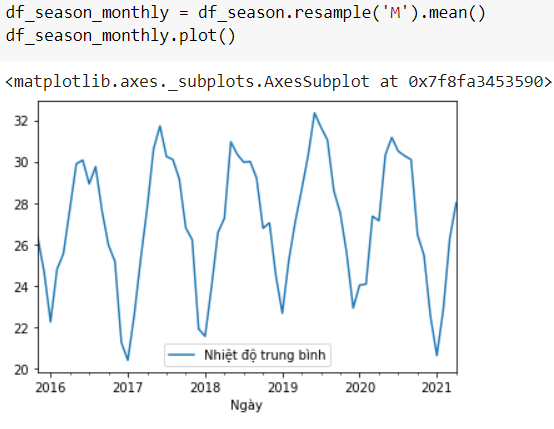


Thông qua Augmented Dickey Fuller test ta thấy p-value < 0.05 ta sẽ loại bỏ giả thuyết null, chấp nhận giả thuyết thay thế. Khi đó ta có thể khẳng định rằng chuỗi không có nghiệm đơn vị và có tính ổn định.



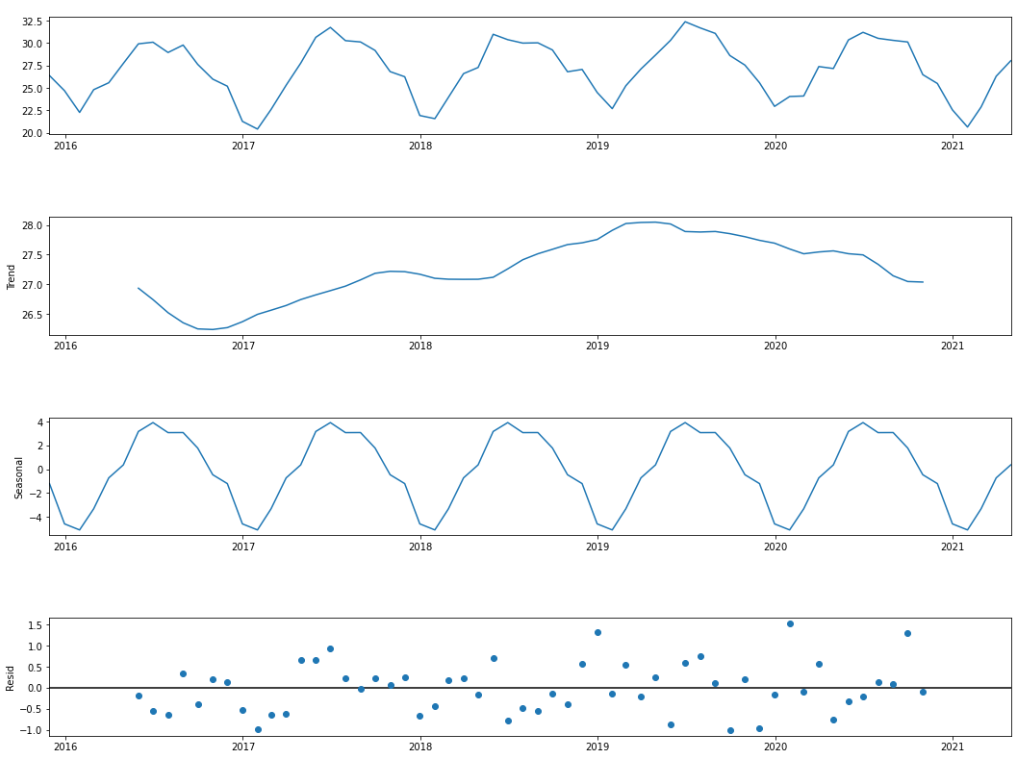
Hình 11: Đồ thị của dữ liệu

Nhìn sơ qua dữ liệu ta thấy dữ liệu có tính thời vụ theo năm (m=12) ta sẽ làm trơn dữ liệu bằng cách chuyển tất cả dữ liệu của ngày về tháng.



Hình 12: Chuyển dữ liệu về dạng tháng

Ta nhìn thấy chuỗi có chu kì là một năm. Nhiệt độ hằng năm có tính chu kì ổn định. Ta sẽ dùng phép phân rã thời vụ ( Phân rã cộng để trích lọc các thành phần cấu tạo nên chuỗi bao gồm: xu hướng (trend), mùa vụ (seasonal) và phần dư (Residual) như bên dưới:



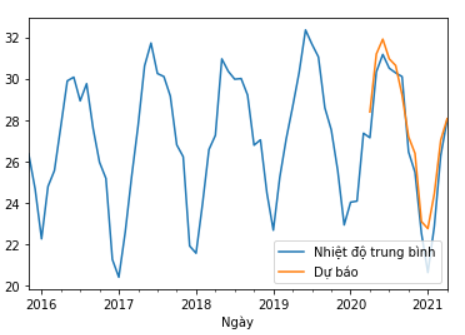
Hình 13: Phân rã dữ liệu theo phân rã cộng

Và sử dụng phương pháp auto ARIMA để tìm ra 3 thông số (p,d,q) và thiết lập seasonal = True và m=12 (vì ta đã tìm ra chu kì thông qua đồ thị phân rã trên).

Chia dữ liệu thành 2 phần:

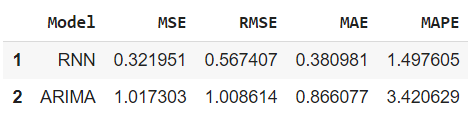
* Tâp train : từ ngày 30/11/2015 – 31/3/2020
* Tập test : từ ngày 31/3/2020 – 31/4/2021

Kết quả dự báo



Hình 14: Đồ thị dự báo của mô hình ARIMA

Kết quả triển khai:



Hình 15: Bảng so sánh kết quả đạt được của 2 mô hình

1. **Kết luận**

Thông qua các kiến thức đã học trên lớp, nhóm em đã áp dụng đầy đủ những kiến thức cần thiết như cách để craw dữ liệu, làm sạch và chuẩn hóa dữ liệu , mô hình hóa dữ liệu. Áp dụng vào bài này bằng 2 mô hình học máy dự đoán chuỗi thời gian là RNN và ARIMA để dự đoán nhiệt độ những ngày tiếp theo một cách trực quan.

1. **Tài liệu tham khảo**
2. Mô hình Arima trong time series:

<https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html>

1. Mô hình RNN trong time series:

<https://viblo.asia/p/machine-learning-thu-lam-nha-thien-van-du-bao-thoi-tiet-djeZ1xYmKWz?fbclid=IwAR0kElTHj2hZVIMrrGNDiW4behiefAIXPLL_c_CN3NEuzCw4lfEl1SuGn8M>

1. Dữ liệu thời tiết ở Đà Nẵng:

<https://freemeteo.vn/thoi-tiet/da-nang/history/daily-history/?gid=6301169&station=11397&date=2021-04-01&language=vietnamese&country=vietnam&fbclid=IwAR1z-tmydXTZYq44sBKSRxhIsXk-aQbVnVlw5OUpiPxKxXMAJcsgwcEBJps>