****

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

KHOA CÔNG NGHỆ THÔNG TIN

**TOÁN ỨNG DỤNG VÀ THỐNG KÊ**

BÁO CÁO ĐỒ ÁN CUỐI KỲ

**LỚP CỬ NHÂN TÀI NĂNG**

**Nhóm thực hiện:**

1712858 - Nguyễn Ngọc Tú

1712902 - Phạm Cao Vỉ

1712906 - Nguyễn Hoàng Việt

[Phần I Các mô hình máy học dự báo thời điểm hết dịch COVID 19](#_rzgem7h0oeks)

[I. Hồi quy phi tuyến](#_haryzpmjlpdi)

[1. Giới thiệu mô hình](#_iw5dakfzauz5)

[Stochastic Gradient Descent](#_ywsnwpe6237n)

[2. Kết quả dự đoán](#_mnenm5gneij2)

[II. Mô hình SIRD](#_uxt4d1l3rlw0)

[Giới thiệu mô hình](#_8657ljp54pvu)

[Xây dựng mô hình](#_5gipdwpiksqe)

[Áp dụng Linear Regression - xây dựng hàm dự đoán](#_ljzhk0i715t6)

[II. Mô hình LSTM](#_u1spy5jfrn1k)

[1. Giới thiệu mô hình](#_hi3z4rl1dgvz)

[2. Bên trong 1 đoạn của LSTM](#_ombgnscmgy6h)

[3. Mô hình sử dụng](#_dbxz9v34gbpl)

[4. Kết quả chạy trên dữ liệu train va test](#_2ctjywdslfrk)

[5. Kết quả dự đoán ngày hết dịch](#_1aviu344myw)

[Phần II Thống kê tìm đối tượng dễ nhiễm COVID 19, nếu nhiễm thì dễ chết nhất](#_3iy5judc113g)

[I. Công thức sử dụng](#_3og20cqm3q80)

[II. Kết quả thống kê xác suất nhiễm bệnh](#_rejco7n145yj)

[III. Kết quả thống kê xác suất chết vì bệnh](#_qg0fphj2kp70)

[IV. Kết luận](#_f34yfcm5csu0)

[Tài liệu tham khảo](#_2ps74hq1r8y1)

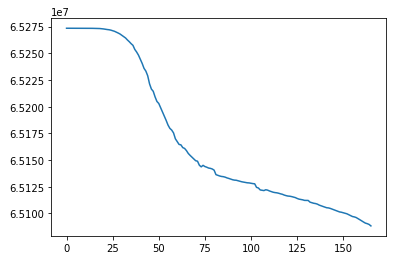
# 

# **Phần I Các mô hình máy học dự báo thời điểm hết dịch COVID 19**

## I. Hồi quy phi tuyến

### **1. Giới thiệu mô hình**

Sau khi visualize số lượng người có khả năng nhiễm bệnh theo thời gian



Ta thấy có thể mô phỏng đường cong này bằng 1 hàm số lan truyền virus trong sinh học:

y(t)=

Ta sẽ dùng phương pháp gradient descent xây dựng hàm mất mát và tìm hệ số a,b,c cho hàm số.

Hàm mất mát : L(t)=

Tại mỗi (i, yi)

=

=

=

### **Stochastic Gradient Descent**

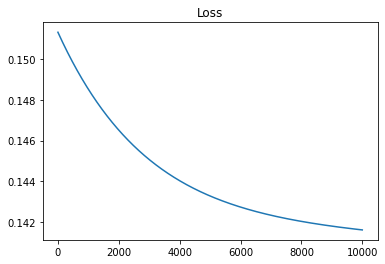
Lặp các dữ liệu (xi,yi)

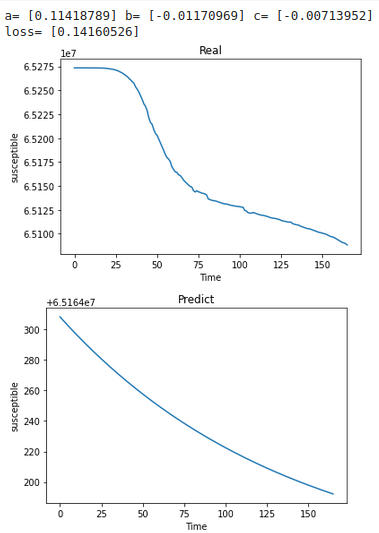
a= a - lr\*

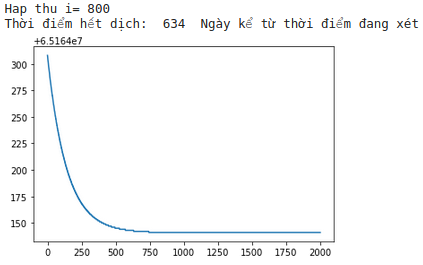
b= b - lr\*

c= c - lr\*

### **2. Kết quả dự đoán**







## **II. Mô hình SIR**D

### **Giới thiệu mô hình**

* Mô hình SIR là một mô hình toán học cơ bản về dịch bệnh, được giới thiệu trong bài báo kinh điển của Kermack và McKendrick. Trong mô hình này, dân số được chia thành 3 nhóm, dựa theo trạng thái đối với bệnh: 1) những người có khả năng mắc bệnh (Susceptible), 2) những người đang nhiễm bệnh và có thể lây cho người khác (Infected), và 3) những người không còn khả năng mắc bệnh (Recovered). Nhưng trong thực tế, dân số còn được chia ra thành 4 loại. Ngoài 3 loại kể trên, dân số được chia thêm thành 4) Những người đã tử vong (Death).
* Với mô hình trên, một người ở trạng thái S chỉ có thể chuyển sang I. Từ I có thể chuyển sang R hoặc D và không có chiều ngược lại,

### **Xây dựng mô hình**

* Gọi số lượng người thuộc mỗi nhóm tại thời điểm t lần lượt là S(t), I(t), R(t), D(t). Với tổng dân số P không đổi. Ta có

**S(t) +I(t) + R(t) +D(t) = P**

**=> I(t)= P- (S(t) + R(t) +D(t))**

**⇔ I(t)- I(t-1)=-S(t) +S(t-1) -R(t) +R(t-1) -D(t) +D(t-1)**

* Nhiệm vụ của ta là tìm các hàm sao cho
* S(t) = h(S(t-1),I(t-1))
* R(t) = g(I(t-1))
* D(t) =f(I(t-1))
* Khi đó I(t) có thể viết lại được như sau:

**I(t) =S(t-1) +R(t-1) + D(t-1) + I(t-1) - h(S(t-1),I(t-1)) -g(I(t-1))- f(I(t-1))**

### **Áp dụng Linear Regression - xây dựng hàm dự đoán**

* Xây dựng hàm mất mát

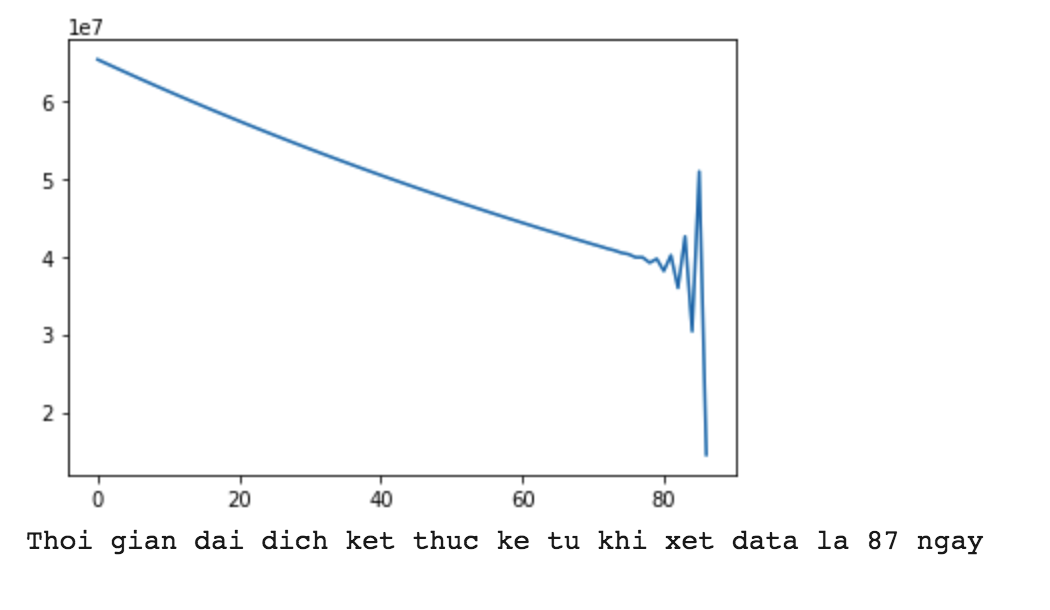
**L(t)=**

* Với dự liệu N đủ lớn

**L(t)=**

* Sử dụng gradient descent để tối ưu nhất hàm L(t). Từ đó ta tìm được các trong số W và bias với mỗi hàm f(), g(), h() tương ứng

Kết quả dự đoán :



Vì dữ liệu lớn và có nhiều điểm khá nhiễu. Kết quả dự đoán kết thúc là khoảng 87 ngày từ khi xét data

## **II. Mô hình LSTM**

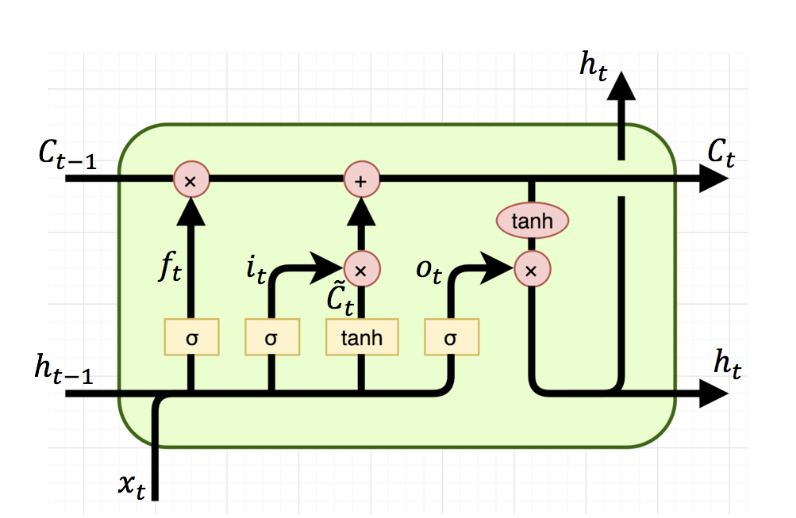
### 

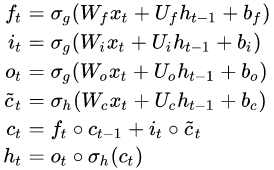
### **1. Giới thiệu mô hình**

Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một dạng đặc biệt của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi [Hochreiter & Schmidhuber (1997)](http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa (long-term dependency). Việc nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện nó để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

### 2. **Bên trong 1 đoạn của LSTM**





∈ : input vector to the LSTM unit

∈ : forget gate's activation vector

∈ : input/update gate's activation vector

∈ : output gate's activation vector

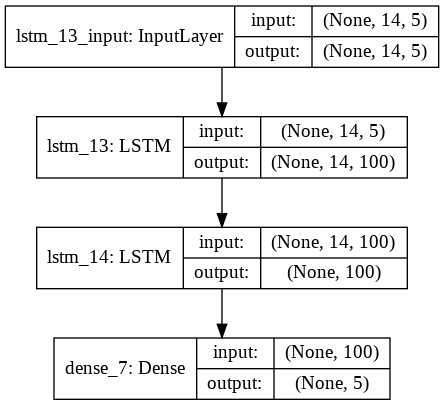
∈ : hidden state vector also known as output vector of the LSTM unit

∈ : cell input activation vector

∈ : cell state vector

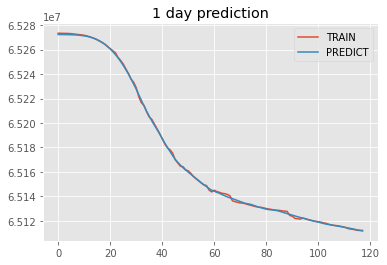
W ∈ , U ∈ and b ∈ : weight matrices and bias vector parameters which need to be learned during training

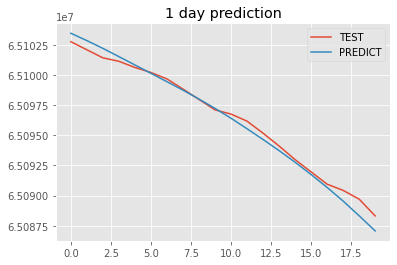
### 3. Mô hình sử dụng



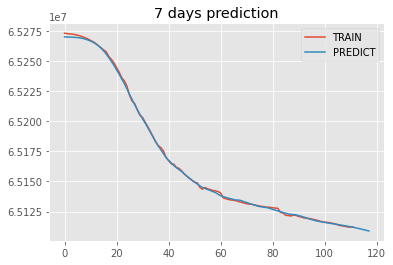
### 4. Kết quả chạy trên dữ liệu train va test

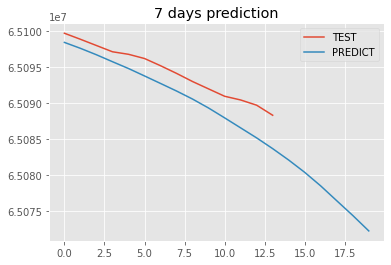
1. Dự đoán ngắn hạn 1 ngày



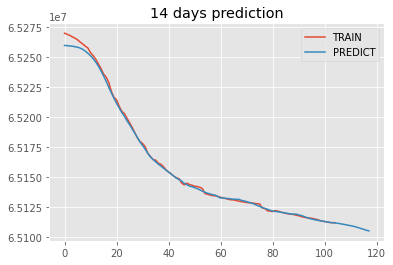


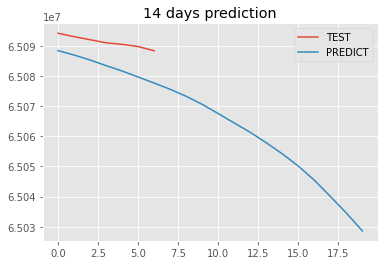
1. Dự đoán dài hạn 7 ngày



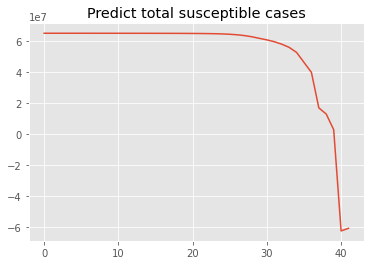


1. Dự đoán dài hạn 14 ngày





### 5. Kết quả dự đoán ngày hết dịch



Số lượng người có thể bị nhiễm sau **41 nữa kể từ thời điểm đang xét** sẽ về 0

# **Phần II Thống kê tìm đối tượng dễ nhiễm COVID 19, nếu nhiễm thì dễ chết nhất**

## I. Công thức sử dụng

Ta sẽ thống kê theo khu vực và theo độ tuổi của New York city.

*Xét ví dụ sau :*

Chọn ngẫu nhiên 1 người thuộc New York city

Xét trường hợp sau:

A: Người này thuộc độ tuổi : 0 - 17

B: Người này sống ở quận Brooklyn

I: Người này bị bệnh

D: Người này bị chết vì COVID 19

Ta có :

P(I|AB) = = =

Với :

n(AB|I): Số người ở quận Brooklyn ở độ tuổi 0-17 mắc bệnh

n(AB): Số người ở quận Brooklyn

n(I) : Tổng số người nhiễm bệnh

n(

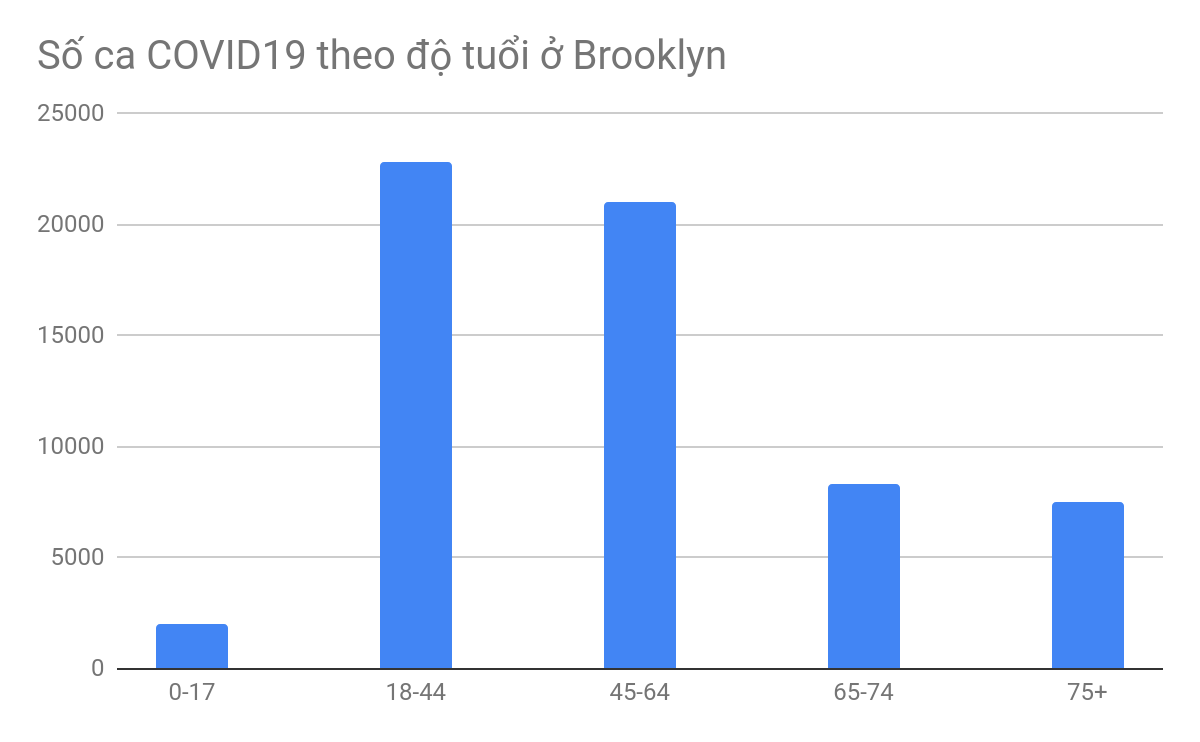
Ta có n(AB|I) = 2058, n(AB) = 584100

=> Xác suất một người nhiễm bệnh nếu người này ở quận Brooklyn, độ tuổi 0-17 :

P(I|AB)= = 0.0035

Để tìm đối tượng dễ bệnh/chết vì COD19, ta sẽ tìm đối tượng có xác suất bệnh/chết vì COD19 cao nhất, bằng cách thống kê và tính toán như trên.

Kết quả thống kê ở quận Brooklyn



|  |  |
| --- | --- |
| **group** | **B: BK\_CASE\_COUNT** |
|  |  |
| A: 0-17 | 2058 |
| n(AB) | 584100 |
| P(I|AB) | 0,003523369286 |
|  |  |
| A:18-44 | 22801 |
| n(AB) | 1076500 |
| P(I|AB) | 0,02118067812 |
|  |  |
| A:45-64 | 21071 |
| n(AB) | 599700 |
| P(I|AB) | 0,03513590128 |
|  |  |
| A:65-74 | 8305 |
| n(AB) | 192900 |
| P(I|AB) | 0,04305339554 |
|  |  |
| A: 75+ | 7498 |
| n(AB) | 103700 |
| P(I|AB) | 0,07230472517 |

Ta thấy ở quận Brooklyn thì có độ tuổi 75+ dễ bị nhiễm nhất.

## II. Kết quả thống kê xác suất nhiễm bệnh

Bảng tổng hợp các quận của NYC

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **group** | **BK\_CASE\_COUNT** | **BX\_CASE\_COUNT** | **MN\_CASE\_COUNT** | **QN\_CASE\_COUNT** | **SI\_CASE\_COUNT** |
| 0-17 | 2058 | 1860 | 553 | 1742 | 465 |
| n(AB) | 584100 | 351750 | 236250 | 451300 | 100750 |
| P(I|AB) | 0,003523369286 | 0,005287846482 | 0,002340740741 | 0,003859960115 | 0,004615384615 |
|  |  |  |  |  |  |
| A=18-44 | 22801 | 18694 | 10749 | 25389 | 5874 |
| n(AB) | 1076500 | 570950 | 750250 | 894100 | 168150 |
| P(I|AB) | 0,02118067812 | 0,03274192136 | 0,01432722426 | 0,02839615256 | 0,03493309545 |
|  |  |  |  |  |  |
| A=45-64 | 21071 | 18323 | 9251 | 25418 | 5508 |
| n(AB) | 599700 | 340600 | 388600 | 612300 | 132100 |
| P(I|AB) | 0,03513590128 | 0,05379624193 | 0,02380597015 | 0,04151233056 | 0,04169568509 |
|  |  |  |  |  |  |
| 65-74 | 8305 | 5636 | 3706 | 7696 | 1441 |
| n(AB) | 192900 | 97600 | 141600 | 189800 | 43200 |
| P(I|AB) | 0,04305339554 | 0,05774590164 | 0,02617231638 | 0,04054794521 | 0,03335648148 |
|  |  |  |  |  |  |
| 75+ | 7498 | 5065 | 4359 | 7127 | 1312 |
| n(AB) | 103700 | 54400 | 76800 | 103700 | 21300 |
| P(I|AB) | 0,07230472517 | 0,09310661765 | 0,0567578125 | 0,0687270974 | 0,06159624413 |

Từ đây ta thấy đối tượng dễ nhiễm bệnh nhất là 75 tuổi trở lên ở quận Bronx

## III. Kết quả thống kê xác suất chết vì bệnh

Bảng tổng hợp các quận của NYC trong trường hợp khả năng chết vì bệnh

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **group** | **BK\_DEATH\_COUNT** | **BX\_DEATH\_COUNT** | **MN\_DEATH\_COUNT** | **QN\_DEATH\_COUNT** | **SI\_DEATH\_COUNT** |
| 0-17 | 5 | 5 | 1 | 2 | 0 |
| n(AB) | 584100 | 351750 | 236250 | 451300 | 100750 |
| P(D|AB) | 0,000008560178052 | 0,00001421464108 | 0,000004232804233 | 0,000004431641923 | 0 |
|  |  |  |  |  |  |
| A=18-44 | 203 | 171 | 62 | 250 | 31 |
| n(AB) | 1076500 | 570950 | 750250 | 894100 | 168150 |
| P(D|AB) | 0,0001885740827 | 0,0002995008319 | 0,00008263912029 | 0,0002796107818 | 0,0001843592031 |
|  |  |  |  |  |  |
| A=45-64 | 1274 | 949 | 384 | 1449 | 184 |
| n(AB) | 599700 | 340600 | 388600 | 612300 | 132100 |
| P(D|AB) | 0,002124395531 | 0,002786259542 | 0,0009881626351 | 0,002366487016 | 0,001392884179 |
|  |  |  |  |  |  |
| 65-74 | 1432 | 1013 | 572 | 1466 | 203 |
| n(AB) | 192900 | 97600 | 141600 | 189800 | 43200 |
| P(D|AB) | 0,007423535511 | 0,01037909836 | 0,004039548023 | 0,007723919916 | 0,004699074074 |
|  |  |  |  |  |  |
| 75+ | 2713 | 1789 | 1488 | 2788 | 478 |
| n(AB) | 103700 | 54400 | 76800 | 103700 | 21300 |
| P(D|AB) | 0,02616200579 | 0,03288602941 | 0,019375 | 0,0268852459 | 0,02244131455 |

Từ đây ta thấy đối tượng dễ chết vì bệnh nhất là 75 tuổi trở lên ở quận Bronx

## IV. Kết luận

Đối tượng dễ bệnh nhất : 75 tuổi trở lên sống ở quận Bronx

Đối tượng dễ chết vì bệnh nhất nhất : 75 tuổi trở lên sống ở quận Bronx

# Tài liệu tham khảo

[https://machinelearningmastery.com](https://machinelearningmastery.com/how-to-develop-lstm-models-for-time-series-forecasting/)