ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CNTT

TOÁN ỨNG DỤNG & THỐNG KÊ

COVID-19 MODELING

04 THÁNG TÁM 2020

HCMUS

I. Thành viên

MSSV	Họ và tên
1712135	Nguyễn Xuân Anh Quân
1712475	Cao Nhơn Hưng
1712791	Lâm Bá Thịnh

II. Bảng đánh giá

STT	Yêu cầu	Ghi chú
1	Xây dựng mô hình dự đoán COVID.	**100%**
2	Xác định đối tượng dễ nhiễm nhất.	**100%**
3	Xác định đối tượng nhiễm mà dễ chết nhất.	**100%**

III. Nội dung chi tiết

1. Mô hình dự đoán Covid

a. Mô tả bài toán

- Yêu cầu bài toán: Từ dữ liệu về COVID-19 cho trước, xây dựng mô hình dự đoán số người nhiễm, không nhiễm, hồi phục, tử vong tại một thời điểm t bất kỳ.
- Ý tưởng: Mô hình phổ biến từ trước đến nay dùng để mô tả dịch bệnh là SIR hoặc SIRD, trong yêu cầu bài toán có yêu cầu tính số người tử vong nên sẽ chọn mô hình SIRD.

b. Bộ dữ liệu

- Dữ liệu được đọc trực tiếp từ link github COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

- Dữ liệu được biểu diễn ở dạng time series, được cập nhật hàng ngày với mỗi quốc gia trên toàn cầu.
- Ta sẽ đọc 3 file sau :
  - `time_series_covid19_confirmed_global.csv`: Số ca xác nhận.
  - `time_series_covid19_deaths_global.csv`: Số ca tử vong.
  - `time_series_covid19_recovered_global.csv`: Số ca hồi phục.

c. Chi tiết Mô hình/Thuật toán sử dụng

- Mô hình SIRD (Susceptible - Infectious – Recovered - Death) là một trong những mô hình toán học để mô tả dịch bệnh COVID-19 hiện nay.
- Mô hình thể hiện 4 trạng thái (có nguy cơ mắc bệnh – mắc bệnh – hồi phục – tử vong) cho nhóm người trong một khu vực nào đó với giả định rằng tổng dân số khu vực đó không thay đổi.
- Mô hình SIRD là một hệ gồm các phương trình vi phân sau:

$$\frac{dS}{dt} = -\frac{\beta}{N} IS$$
$$\frac{dI}{dt} = \frac{\beta}{N} IS - \gamma I - \alpha I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{\mathrm{dD}}{dt} = \alpha I$$

Trong đó tại mỗi thời điểm $t \geq t_0 \geq 0$ với t_0 là thời điểm đầu ghi nhận,

- S(t): Số người có nguy cơ mắc bệnh.
- I(t): Số người nhiễm bệnh.
- R(t): Số người hồi phục sau khi nhiễm bệnh.
- D(t): Số người chết khi nhiễm bệnh.
- $\beta(t)$: Tỷ lệ tiếp xúc mỗi người trong nhóm S(t) với người trong nhóm I(t) (hệ số lây nhiễm).
- $\gamma(t)$: Tỷ lệ phục hồi sau bệnh (hệ số phục hồi).
- $\alpha(t)$: Tỷ lệ tử vong khi nhiễm bệnh (hệ số tử vong).
- N(t): Tổng dân số trong khu vực đó. Ta có

$$N(t) = S(t) + I(t) + R(t) + D(t)$$

- Từ dữ liệu về COVID-19 đã cho, ta đi tìm các hệ số α, β, γ của mô hình, để từ đó thực hiện dự đoán. Để tìm các hệ số này trước hết ta nói đến phương pháp xấp xỉ Euler trong việc giải hệ SIRD.

- Phương pháp Euler là một phương pháp thường được sử dụng trong việc giải các hệ phương trình vi phân thông thường.
- Giả sử ta có phương trình vi phân bậc nhất $y' = f(t, y(t))$. Khi đó, ý tưởng của phương pháp Euler là xấp xỉ nghiệm y bằng dãy $\{y_n\}$ sao cho $y_{n+1} := y_n + f(t_n, y_n)\Delta t$ với Δt là bước xấp xỉ đủ nhỏ. Tổng quát, một hệ phương trình vi phân bậc một được viết dưới dạng

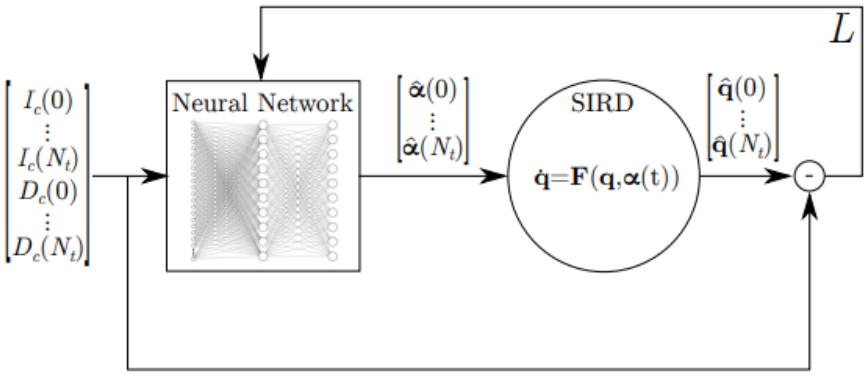
$$y_1' = f_1(t, y_1, ..., y_N)$$

$$...$$

$$y_N' = f_N(t, y_1, ..., y_N)$$

  Trong đó y_i là các hàm số thực phụ thuộc vào biến t ($t \geq 0$) và f_i là các hàm số thực phụ thuộc vào biến t ($t \geq 0$).
- Ở đây với dữ liệu COVID-19 đã cho ta chọn Δt là 1 (dữ liệu được cập nhật qua từng ngày), và từ hệ trên ta tính ra được các hệ số α, β, γ của mỗi ngày, từ đó tính trung bình để tìm được giá trị chung của từng hệ số. Khi tính được bộ hệ số rồi có thể đưa vào mô hình SIRD để dự đoán với ngày bắt đầu nhập vào.

- [Một cách tiếp cận khác dùng Học Máy (Machine Learning)]
  - Các mô hình học máy mới đề xuất gần đây cũng được đưa vào dự đoán tình hình dịch bệnh COVID-19.
  - Ý tưởng: Khởi tạo SIR hoặc SIRD với các tham số ban đầu β (hệ số lây nhiễm), γ (hệ số phục hồi), α (hệ số tử vong nếu là SIRD) để tính I, R, D tại thời điểm $t \geq t_0$. Sau đó dùng một Neural Network và dữ liệu về số ca nhiễm, ca tử vong và hồi phục đã công bố trước đó để ước tính lại các hệ số β, γ, α tại mỗi thời điểm công bố dịch bệnh COVID-19. Từ đó có thể tính trung bình các giá trị này để lấy giá trị chung của mô hình SIRD hoặc SIR.



Hình: Mô hình minh họa hướng tiếp cận dùng Học máy trong mô tả dịch bệnh COVID-19, được trích từ bài báo gốc.

- Bài báo gốc First-principles machine learning modelling of COVID-19.
- Ở đây để cài đặt minh họa nhóm sẽ dùng mô hình SIR (cho kết quả tốt), mô hình SIRD đã thử nghiệm nhưng không tối ưu được (do hàm loss khá phức tạp) dẫn đến underfitting. Do đó tham số phải tìm là β, γ.
- Kiến trúc của mạng Neural Network
  - Thêm một layer Batch Normalization để chuẩn hóa dữ liệu (về trạng thái zero-mean với độ lệch chuẩn 1) giúp cho quá trình tối ưu ổn định và nhanh hơn.
  - Gồm chỉ có 1 fully connected layer gồm 8 node (theo như trong bài báo đề nghị), hàm activation là ReLU.
  - Layer đầu ra gồm có 2 node là β, γ, hàm activation là ReLU.
- Hàm loss được lấy ý tưởng từ bài báo trên như sau:

  Đặt $\alpha_0 \equiv [\beta_0, \gamma_0]$

$$L = \sum_{t=0}^{N} \left((\log(I(t)) - \log(\hat{I}(t)))^2 + (\log(R(t)) - \log(\hat{R}(t)))^2 \right) +$$

$$0.01 \frac{\log(\max(I))}{\max(I)} \sum_{t=0}^{N} \left((I(t) - \hat{I}(t))^2 + (R(t) - \hat{R}(t))^2 \right) +$$

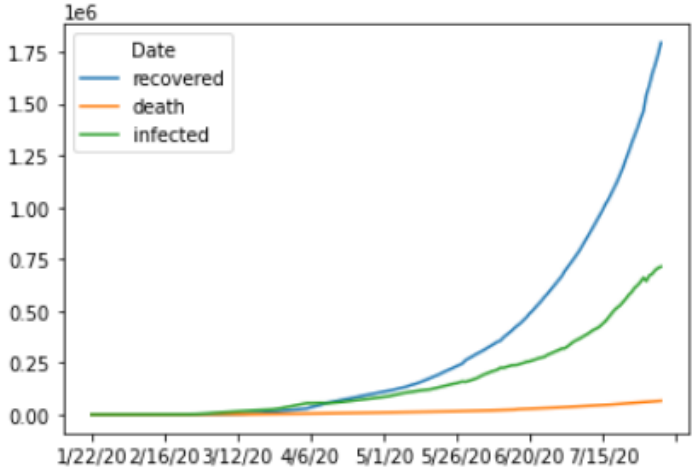
$$0.01 \frac{\log(\max(I))}{\max(\alpha_0)} \sum_{t=0}^{N} \left((\hat{\beta}(t) - \hat{\beta}(t+1))^2 + (\hat{\gamma}(t) - \hat{\gamma}(t+1))^2 \right) +$$

$$0.01 \frac{\log(\max(I))}{\max(\alpha_0)} \sum_{t=0}^{N} \left((\hat{\beta}(0) - \beta_0)^2 + (\hat{\gamma}(0) - \gamma_0)^2 \right)$$

- Huấn luyện mô hình chúng ta sẽ tối ưu hàm loss này.

d. Kết quả

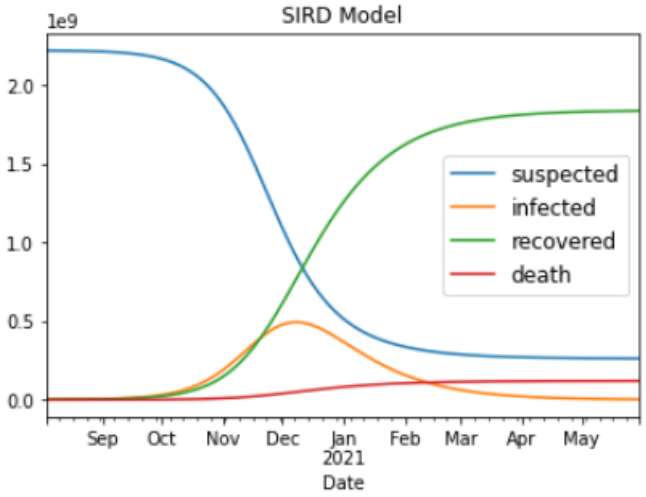
- Kết quả dùng phương pháp xấp xỉ Euler để tìm bộ tham số β, γ, α của hệ SIRD và dự đoán số người nhiễm, số người không nhiễm, số người hồi phục và số người chết.
    - Dữ liệu COVID-19 tính đến ngày 04/08/2020 sau khi tiền xử lý của một số các quốc gia châu Á [Việt Nam, Thái Lan, Malaysia, Nhật Bản, South Korea, Indonesia, Singapore, Philippines, Ấn Độ, Iran] được biểu diễn như sau:



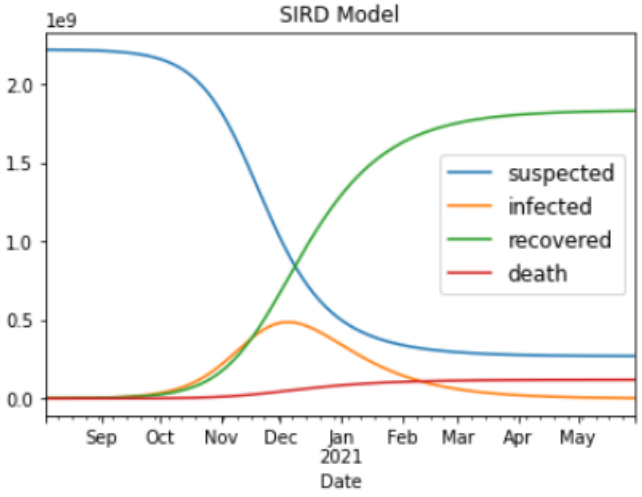
    - Từ bộ dữ liệu COVID-19 được cập nhật đến ngày 04/08/2020 thì theo phương pháp nêu ở trên ta tìm được trung bình của từng tham số như sau:

```
beta:  0.1167429269866099
gamma:  0.04557662712322327
alpha:  0.002975325405036735
```

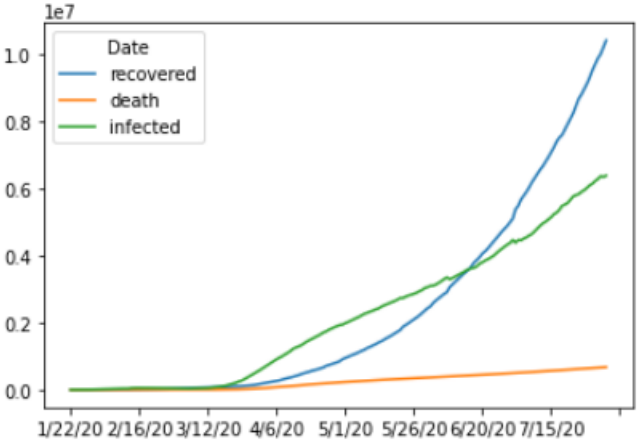
    - Khi có được các tham số này ta có thể áp dụng thuật toán xấp xỉ Euler để dự đoán số ca nhiễm (infected), không nhiễm và có nguy cơ nhiễm (suspected), hồi phục (recovered), số ca tử vong (death) trong tương lai theo mô hình SIRD.
    - Kết quả minh họa với số ngày là 300 bắt đầu từ ngày "04-08-2020" như sau:



- Ta có thể sử dụng hàm odeint trong thư viện scipy để giải hệ phương trình vi phân, kết quả cho ra hoàn toàn tương tự.



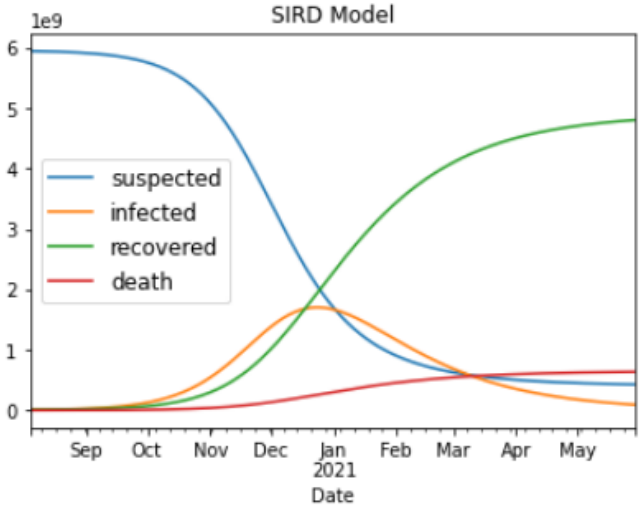
- Thử nghiệm trên 50 quốc gia có số ca nhiễm lớn nhất và Việt Nam (tổng cộng 51 quốc gia) cũng cho kết quả tương tự.
- Dữ liệu thể hiện sau khi xử lý của 51 quốc gia.



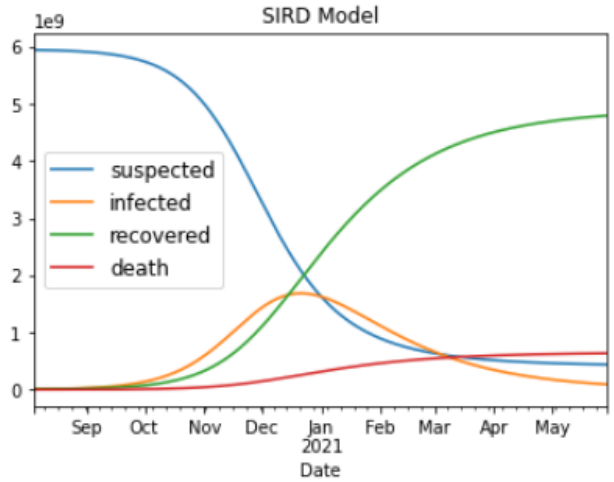
- Kết quả tìm được các tham số

```
beta:  0.08303704291446079
gamma:  0.02558147574958295
alpha:  0.003397425961843472
```

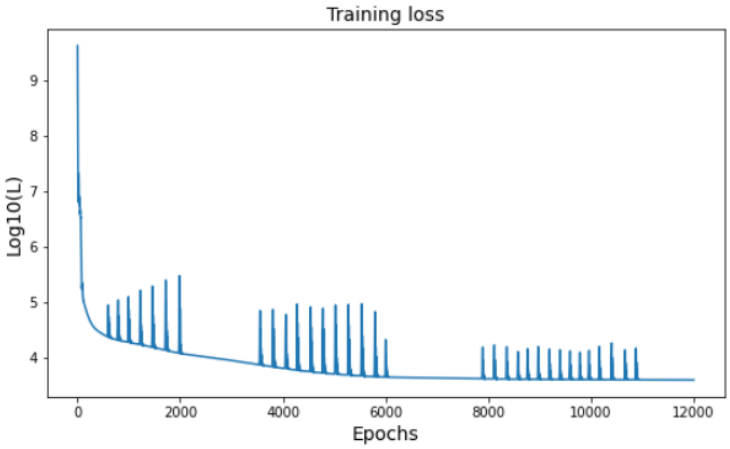
- Kết quả tìm được dựa theo phương pháp xấp xỉ Euler



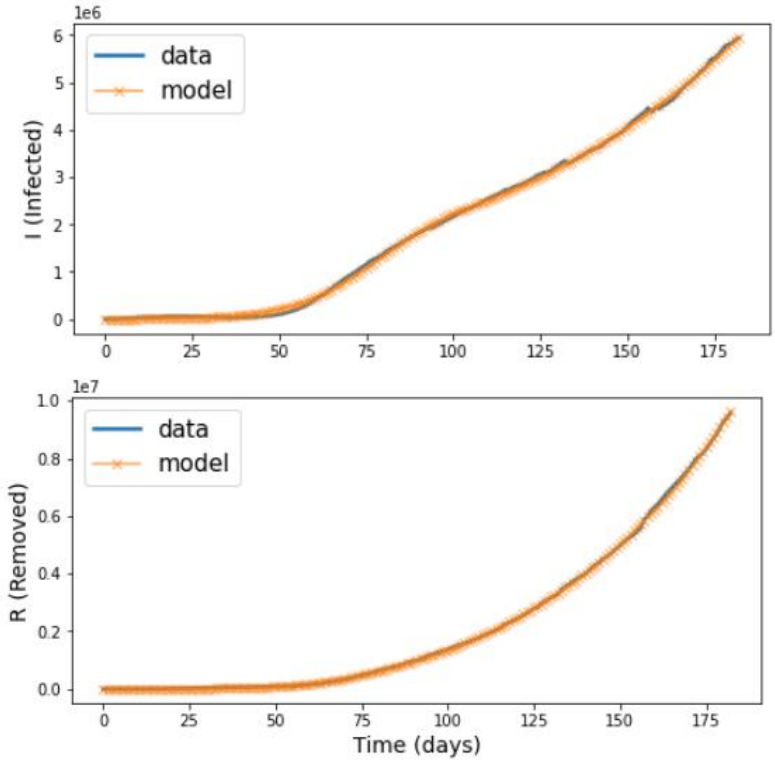
- Kết quả tìm được sử dụng hàm odeint của thư viện scipy.



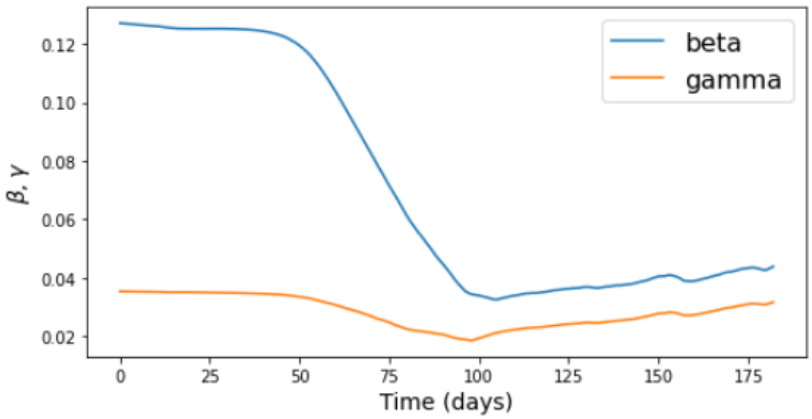
- Kết quả thử nghiệm với cách tiếp cận Machine Learning (thử nghiệm trên mô hình SIR như đã nói ở trên).
    - Biểu diễn hàm loss trong quá trình huấn luyện mô hình.



- Kết quả huấn luyện mô hình.



- Kết quả cho model fit với data khá tốt.
- Đồ thị biểu diễn hai tham số β và γ qua từng ngày để mô hình fit với data.



- Giá trị trung bình của từng tham số.

```
beta:  0.07176673
gamma 0.028299239
```

- Từ đó dùng 2 tham số này vào mô hình SIR để dự đoán số người nhiễm (infected), số người không nhiễm và có nguy cơ nhiễm bệnh (suspected), số người chết và số người hồi phục (recovered) tại một thời điểm nào đó.

e. Nhận xét

- Phương pháp xấp xỉ Euler cho kết quả tốt, thể hiện đúng sự dịch chuyển của dịch bệnh theo mô hình SIRD.
- Phương pháp dùng Machine Learning là một hướng tiếp cận khác, bản chất là tìm các tham số của một hệ phương trình vi phân khi biết trước nghiệm (các cặp (input, ouput) của các biến trong hệ phương trình vi phân đó) sử dụng một mạng Neural Network. Kết quả cho ra các tham số mô hình fit với dữ liệu khá tốt. Tuy nhiên theo thử nghiệm thì mô hình này không ổn định, việc tối ưu khá khó khăn do hàm loss phức tạp, đôi khi không tìm được điểm tối ưu toàn cục.

2. Xác định đối tượng dễ nhiễm nhất

a. Mô tả bài toán

Khi dịch bệnh bùng nổ, một trong những điều mà mọi người thường quan tâm nhất chính là đối tượng nào dễ bị nhiễm bệnh nhất. Nhờ việc xác định này, chúng ta có thể lên kế hoạch một cách cụ thể, tỉ mỉ để phòng/chống dịch bệnh (chẳng hạn, ta có thể phân tích các đặc điểm của đối tượng này để tìm ra cách thức lây lan của dịch bệnh, từ đó đưa ra biện pháp để hạn chế con đường lây lan).

b. Bộ dữ liệu

- Dữ liệu thống kê số lượng ca nhiễm theo từng nhóm tuổi và giới tính tại Canada theo từng ngày (hình dưới là dữ liệu được cập nhật vào ngày 04/08/2020).

Age group (years)	Number of cases with case reports (proportion)	Number of male cases (proportion)	Number of female cases (proportion)	Number of other cases (proportion)
80+	18,544 (16.0%)	5,738 (11.1%)	12,695 (19.9%)	0 (0.0%)
70-79	8,192 (7.1%)	4,039 (7.8%)	4,128 (6.5%)	1 (6.3%)
60-69	11,155 (9.6%)	5,652 (11.0%)	5,477 (8.6%)	2 (12.5%)
50-59	17,259 (14.9%)	7,812 (15.2%)	9,417 (14.8%)	2 (12.5%)
40-49	17,371 (15.0%)	7,758 (15.1%)	9,562 (15.0%)	3 (18.8%)
30-39	16,491 (14.3%)	7,777 (15.1%)	8,663 (13.6%)	4 (25.0%)
20-29	17,448 (15.1%)	8,202 (15.9%)	9,192 (14.4%)	2 (12.5%)
≤19	9,260 (8.0%)	4,538 (8.8%)	4,688 (7.3%)	2 (12.5%)

- Dữ liệu thống kê dân số theo từng nhóm tuổi và giới tính tại Canada trong năm 2019.

	Females
Age group[3 5]	2019
	Persons
All ages	18,911,177
0 to 4 years	947,006
5 to 9 years	997,346
10 to 14 years	997,083
15 to 19 years	1,031,330
20 to 24 years	1,183,251
25 to 29 years	1,271,755
30 to 34 years	1,285,016
35 to 39 years	1,291,537
40 to 44 years	1,222,563
45 to 49 years	1,206,364
50 to 54 years	1,256,952
55 to 59 years	1,382,252
60 to 64 years	1,275,969
65 to 69 years	1,079,742
70 to 74 years	889,219
75 to 79 years	621,270
80 to 84 years	439,316
85 to 89 years	306,365
90 to 94 years	162,454
95 to 99 years	55,530
100 years and over	8,857

Age group[3][5]	Males
	2019
	Persons
All ages	18,678,085
0 to 4 years	996,169
5 to 9 years	1,042,006
10 to 14 years	1,034,679
15 to 19 years	1,083,305
20 to 24 years	1,293,447
25 to 29 years	1,353,719
30 to 34 years	1,318,922
35 to 39 years	1,288,484
40 to 44 years	1,198,446
45 to 49 years	1,190,042
50 to 54 years	1,245,715
55 to 59 years	1,367,374
60 to 64 years	1,235,919
65 to 69 years	1,016,865
70 to 74 years	817,541
75 to 79 years	543,007
80 to 84 years	347,388
85 to 89 years	204,463
90 to 94 years	80,100
95 to 99 years	18,556
100 years and over	1,938

c. Chi tiết Mô hình/Thuật toán sử dụng

- Tính xác suất lây nhiễm của từng nhóm tuổi và giới tính.

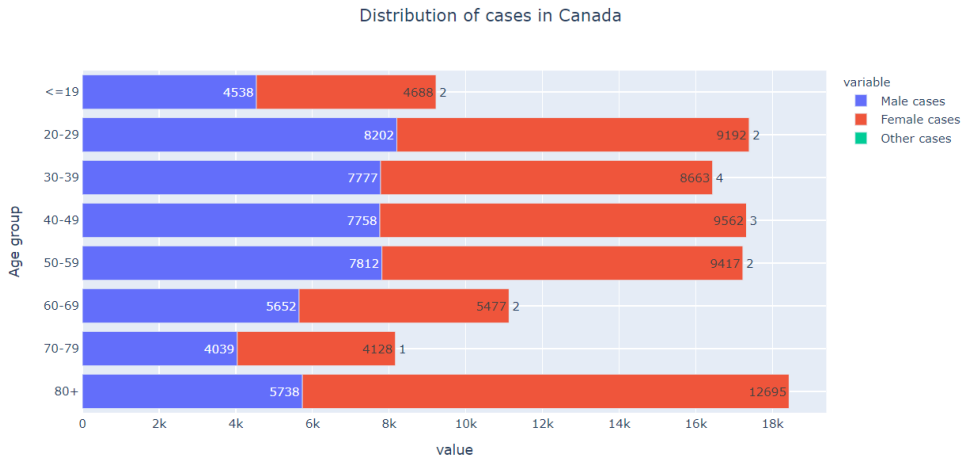
$$P(age, gender) = \frac{Infected(age, gender)}{Population(age, gender)}.$$

- So sánh các xác suất này và chọn ra nhóm đối tượng nào có xác suất lớn nhất.

$$(age, gender)_{optimal} = argmax_{age,gender} P(age, gender).$$

d. Kết quả

- Trực quan hóa bộ dữ liệu qua biểu đồ bar.



- Tính xác suất lây nhiễm của từng nhóm.

Age group	Total cases	Male cases	Female cases	Other cases	Male persons	Female persons	Male percentage	Female percentage
80+	18544	5738	12695	0	652445	972522	0.008795	0.013054
70-79	8192	4039	4128	1	1360548	1510489	0.002969	0.002733
60-69	11155	5652	5477	2	2252784	2355711	0.002509	0.002325
50-59	17259	7812	9417	2	2613089	2639204	0.002990	0.003568
40-49	17371	7758	9562	3	2388488	2428927	0.003248	0.003937
30-39	16491	7777	8663	4	2607406	2576553	0.002983	0.003362
20-29	17448	8202	9192	2	2647166	2455006	0.003098	0.003744
<=19	9260	4538	4688	2	4156159	3972765	0.001092	0.001180

- Xác định nhóm đối tượng dễ bị nhiễm nhất (theo kết quả đã tính, ta thấy xác suất người có giới tính **Nữ,** thuộc nhóm tuổi **80+** cho kết quả cao nhất so với những đối tượng còn lại).

```
Doi tuong de nhiem nhat la Female , do tuoi 80+.
```

e. Nhận xét

- Từ dữ liệu đã tính toán, ta dễ thấy nhóm tuổi từ **20-59** cho xác suất lây nhiễm khá cao và gần như nhau. Vì nhóm tuổi này thuộc độ tuổi lao động, nên việc tiếp xúc nhau nhiều đã tạo cơ hội cho dịch bệnh lây lan nhanh.
- Những người thuộc độ tuổi **80+** cho kết quả cao nhất. Vì ở độ tuổi này, các đối tượng thường có sức khỏe yếu, thường xuyên đến bệnh viện, mà đây là nơi dễ tiếp xúc với những người mắc bệnh nhất, do đó việc bị lây nhiễm bệnh là điều khó tránh khỏi.

3. Xác định đối tượng nhiễm mà dễ chết nhất

a. Mô tả bài toán

Khi dịch bệnh bùng nổ, số ca nhiễm tăng cao nhưng nguồn tài nguyên và cơ sở vật chất để chữa trị cho các bệnh nhân lại hạn chế. Do đó ta cần xác định được các đối tượng có nguy cơ tử vong cao nhất để tập trung chữa trị, đồng thời phân bố hợp lý được các tài nguyên cho các nhóm bệnh nhân với các yếu tố, tình trạng khác nhau.

b. Chi tiết Mô hình/Thuật toán sử dụng

- Dữ liệu thống kê các ca tử vong theo độ tuổi và theo giới tính tại Mỹ trong thời gian từ ngày 22 tháng 1 đến ngày 30 tháng 05 (2020).

	age	patients	with_uhc	without_uhc	reported_deaths	deaths_with_uhc	deaths_without_uhc	unknown_uhc	deaths_unknown_uhc
0	<=9	20458	619	2277	13	4	2	17562	7
1	10-19	49245	2076	5047	33	16	4	42122	13
2	20-29	182469	8906	18530	273	122	24	155033	127
3	30-39	214849	14854	18629	852	411	21	181366	420
4	40-49	219139	24161	16411	2083	1077	58	178567	948
5	50-59	235774	40297	14420	5639	3158	131	181057	2350
6	60-69	179007	42206	7919	11947	7050	187	128882	4710
7	70-79	105252	31601	2799	17510	10008	286	70852	7216
8	>=80	114295	34159	2409	32766	16966	718	77727	15082

	sex	patients	with_uhc	without_uhc	reported_deaths	deaths_with_uhc	deaths_without_uhc	unknown_uhc	deaths_unknown_uhc
0	male	646358	96839	42048	38773	21667	724	507471	16382
1	female	674130	102040	46393	32343	17145	707	525697	14491

- Ta thực hiện tính xác suất tử vong của bệnh nhân theo nhóm tuổi + trình trạng bệnh nền và theo giới tính + tình trạng bệnh nền:

Gọi

- C là biến cố bệnh nhân chết
- T_i là biến cố bệnh nhân có độ tuổi tại hàng i (Vd: T_0 là độ tuổi '<=9')
- B_i là biến cố bệnh nhân có tình trạng bệnh lý nền thế nào (có bệnh, không bệnh, không xác định).

Khi đó, xác suất tử vong của người có tuổi T_i và tình trạng B_i là:

$$P(C|T_iB_i) = \frac{P(CT_iB_i)}{P(T_iB_i)} = \frac{cột\ deaths...}{cột\ with...}$$

Gọi G_i là biến cố bệnh nhân có giới tính i (i = male / female)

Khi đó, xác suất tử vong của bệnh nhân có giới tính G_i và tình trạng B_i là:

$$P(C|G_iB_i) = \frac{P(CG_iB_i)}{P(G_iB_i)} = \frac{cột\ deaths...}{cột\ with...}$$

- Thông tin chi tiết về bảng dữ liệu và quá trình xử lý được trình bày trong notebook **Ex03.ipynb**

c. Kết quả và nhận xét

- Từ dữ liệu đã tính toán, ta dễ thấy nhóm tuổi từ **0-39** có tỉ lệ tử vong thấp bởi nhóm tuổi này thuộc độ tuổi phát triển - trưởng thành, có hệ miễn dịch tương đối tốt.
- Những người thuộc độ tuổi **80+ và có bệnh lý nền** cho tỉ lệ tử vong cao nhất vì ở độ tuổi này, các đối tượng thường có sức đề kháng yếu, đồng thời có thể mang trước nhiều bệnh lý nền do lão hóa, do đó khi nhiễm bệnh khó có khả năng hồi phục.
- Đồng thời các bệnh nhân **có bệnh lý nền** khiến tỉ lệ tử vong **cao gần gấp đôi** khi không có bệnh lý, chứng tỏ yếu tố này ảnh hưởng khá lớn đến khả năng hồi phục của bệnh nhân.

4. Nguồn tham khảo

Dữ liệu thống kê số ca nhiễm, tử vong và hồi phục của JHU:

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Dữ liệu thống kê các nhiễm theo độ tuổi và giới tính tại Canada:
https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html?stat=num&measure=total#a2

Dữ liệu thống kê dân số theo từng độ tuổi và giới tính tại Canada:

https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501&pickMembers%5B0%5D=1.1&pickMembers%5B1%5D=2.2&cubeTimeFrame.startYear=2019&cubeTimeFrame.endYear=2019&referencePeriods=20190101%2C20190101

Paper: First-principles machine learning modelling of COVID-19.

https://diendantoanhoc.net/topic/169781-m%C3%B4-h%C3%ACnh-lan-truy%E1%BB%81n-c%E1%BB%A7a-d%E1%BB%8Bch-b%E1%BB%87nh/

https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model?fbclid=IwAR33zXJbRaIM6shc4b3u-dASiIJuXw2EcbgkygjPinsFMjqY8wKvaW0sV64

Dữ liệu thống kê các ca tử vong theo độ tuổi và giới tính tại Mỹ trong thời gian từ ngày 22 tháng 1 đến ngày 30 tháng 05 (2020) - table 3:

https://www.cdc.gov/mmwr/volumes/69/wr/mm6924e2.htm?s_cid=mm6924e2_w#T3_down