

**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

# Thành viên

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

1. **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%** |

1. **Nội dung chi tiết**
2. **Mô hình dự đoán Covid**
3. 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.

1. 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](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series).
* 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.

1. 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:



Trong đó tại mỗi thời điểm  với 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.
* : Tỷ lệ tiếp xúc mỗi người trong nhóm S(t) với người trong nhóm I(t) (hệ số lây nhiễm).
* : Tỷ lệ phục hồi sau bệnh (hệ số phục hồi).
* : Tỷ lệ tử vong khi nhiễm bệnh (hệ số tử vong).
* N(t): Tổng dân số trong khu vực đó. Ta có

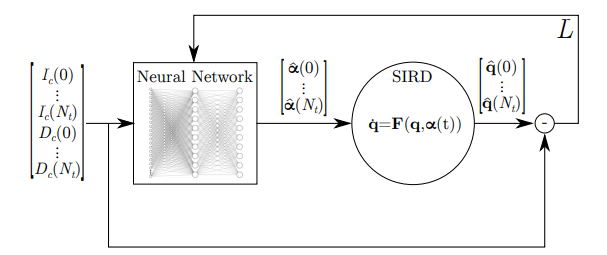


* Từ dữ liệu về COVID-19 đã cho, ta đi tìm các hệ số  của mô hình, để từ đó thực hiện dự đoán. Để tìm các hệ số này trước hết ta nói đến phương pháp xấp xỉ Euler trong việc giải hệ SIRD.
* Phương pháp Euler là một phương pháp thường được sử dụng trong việc giải các hệ phương trình vi phân thông thường.
* Giả sử ta có phương trình vi phân bậc nhất . Khi đó, ý tưởng của phương pháp Euler là xấp xỉ nghiệm y bằng dãy {yn} sao cho  với  là bước xấp xỉ đủ nhỏ. Tổng quát, một hệ phương trình vi phân bậc một được viết dưới dạng



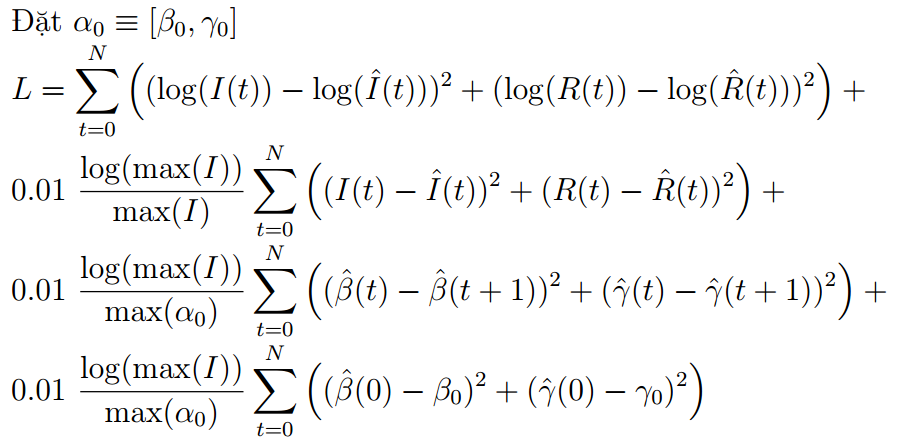
Trong đó  là các hàm số thực phụ thuộc vào biến t () và  là các hàm số thực phụ thuộc vào biến t ().

* Ở đây với dữ liệu COVID-19 đã cho ta chọn  là 1 (dữ liệu được cập nhật qua từng ngày), và từ hệ trên ta tính ra được các hệ số  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  (hệ số lây nhiễm),  (hệ số phục hồi),  (hệ số tử vong nếu là SIRD) để tính I, R, D tại thời điểm . 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  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.

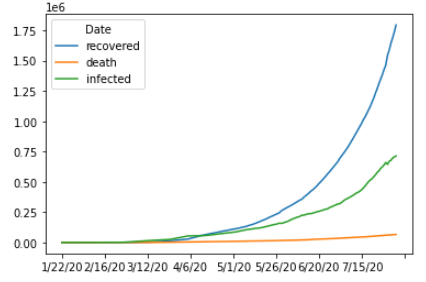
* Bài báo gốc [First-principles machine learning modelling of COVID-19](https://arxiv.org/pdf/2004.09478.pdf).
* Ở đây để cài đặt minh họa nhóm sẽ dùng mô hình SIR (cho kết quả tốt), mô hình SIRD đã thử nghiệm nhưng không tối ưu được (do hàm loss khá phức tạp) dẫn đến underfitting. Do đó tham số phải tìm là .
* 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 , hàm activation là ReLU.
* Hàm loss được lấy ý tưởng từ bài báo trên như sau:



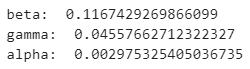
* Huấn luyện mô hình chúng ta sẽ tối ưu hàm loss này.

1. Kết quả

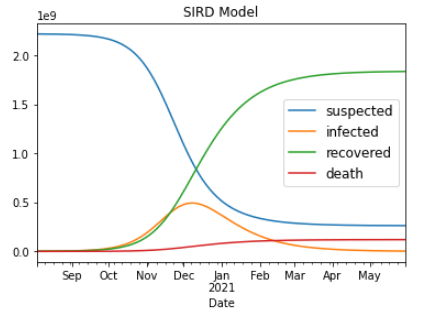
* Kết quả dùng phương pháp xấp xỉ Euler để tìm bộ tham số  của hệ SIRD và dự đoán số người nhiễm, số người không nhiễm, số người hồi phục và số người chết.
* Dữ liệu COVID-19 tính đến ngày 04/08/2020 sau khi tiền xử lý của một số các quốc gia châu Á [Việt Nam, Thái Lan, Malaysia, Nhật Bản, South Korea,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



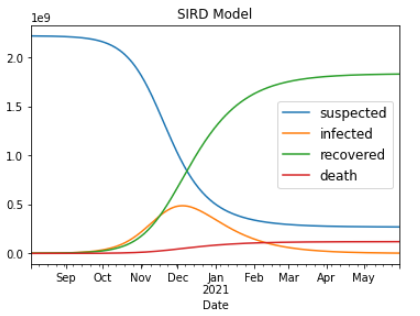
* Từ bộ dữ liệu COVID-19 được cập nhật đến ngày 04/08/2020 thì theo phương pháp nêu ở trên ta tìm được trung bình của từng tham số như sau:



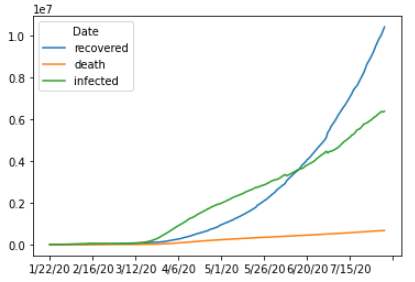
* Khi có được các tham số này ta có thể áp dụng thuật toán xấp xỉ Euler để dự đoán số ca nhiễm (infected), không nhiễm và có nguy cơ nhiễm (suspected), hồi phục (recovered), số ca tử vong (death) trong tương lai theo mô hình SIRD.
* Kết quả minh họa với số ngày là 300 bắt đầu từ ngày “04-08-2020” như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



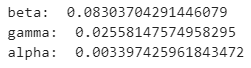
* Ta có thể sử dụng hàm odeint trong thư viện scipy để giải hệ phương trình vi phân, kết quả cho ra hoàn toàn tương tự.



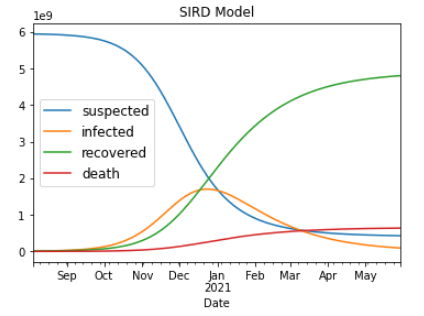
* Thử nghiệm trên 50 quốc gia có số ca nhiễm lớn nhất và Việt Nam (tổng cộng 51 quốc gia) cũng cho kết quả tương tự.
* Dữ liệu thể hiện sau khi xử lý của 51 quốc gia.



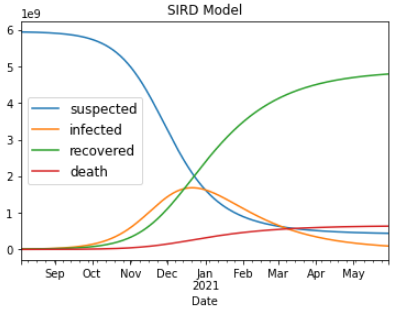
* Kết quả tìm được các tham số



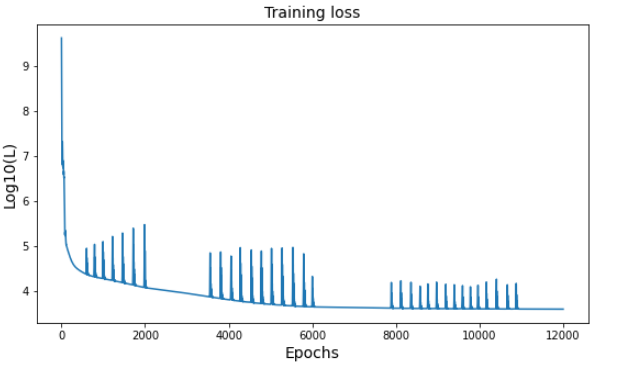
* Kết quả tìm được dựa theo phương pháp xấp xỉ Euler



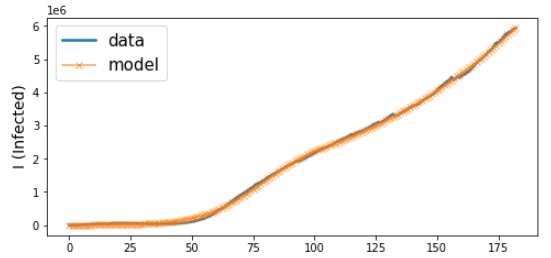
* Kết quả tìm được sử dụng hàm odeint của thư viện scipy.

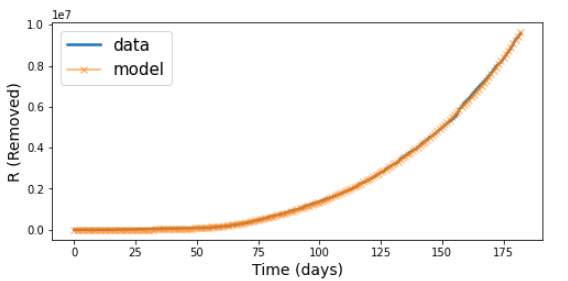


* Kết quả thử nghiệm với cách tiếp cận Machine Learning (thử nghiệm trên mô hình SIR như đã nói ở trên).
* Biễu diễn hàm loss trong quá trình huấn luyện mô hình.

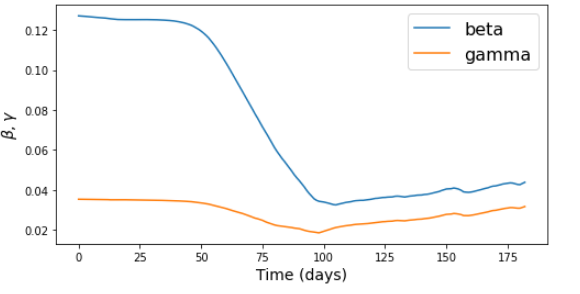


* Kết quả huấn luyện mô hình.





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



* Giá trị trung bình của từng tham số.



* 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.

1. Nhận xét

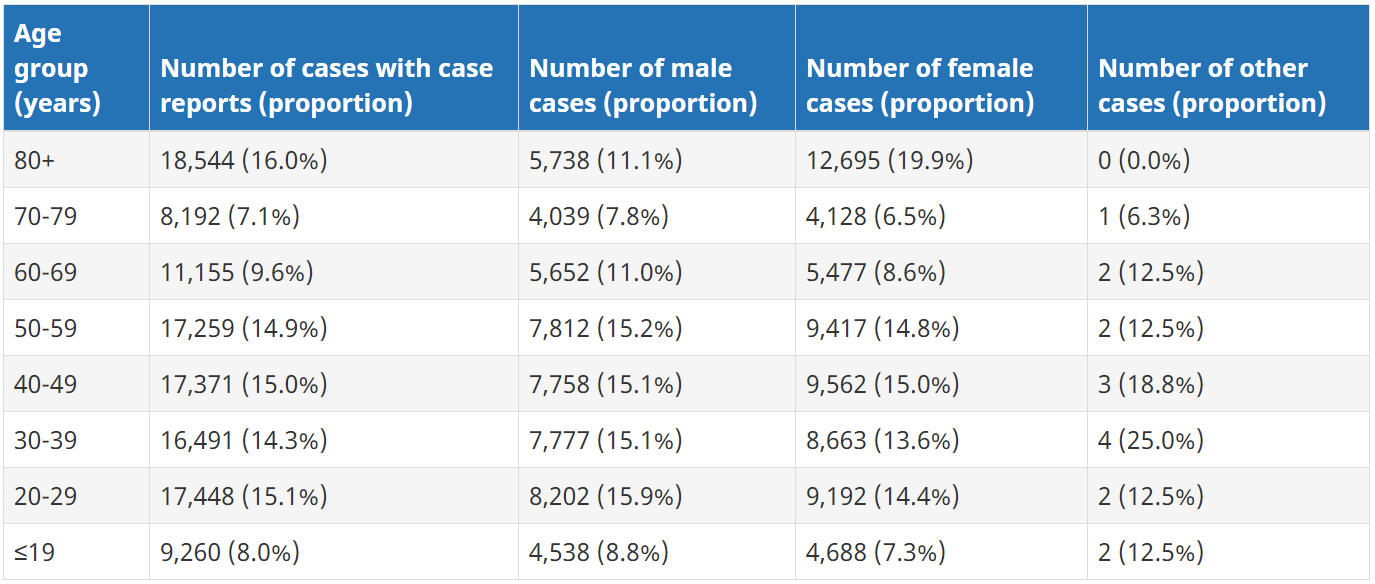
* Phương pháp xấp xỉ Euler cho kết quả tốt, thể hiện đúng sự dịch chuyển của dịch bệnh theo mô hình SIRD.
* Phương pháp dùng Machine Learning là một hướng tiếp cận khác, bản chất là tìm các tham số của một hệ phương trình vi phân khi biết trước nghiệm (các cặp (input, ouput) của các biến trong hệ phương 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.

1. **Xác định đối tượng dễ nhiễm nhất**
2. Mô tả bài toán

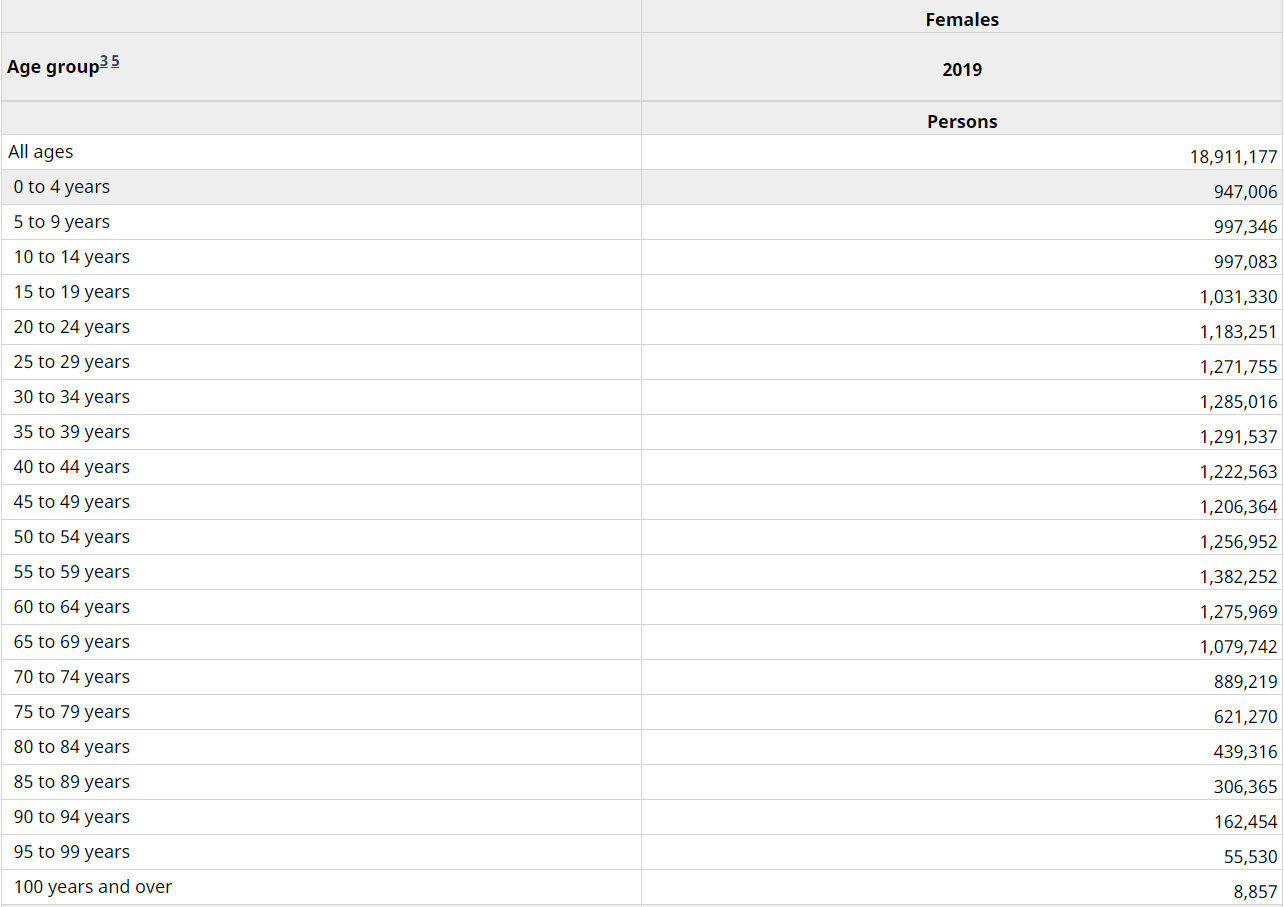
Khi dịch bệnh bùng nổ, một trong những điều mà mọi người thường quan tâm nhất chính là đối tượng nào dễ bị nhiễm bệnh nhất. Nhờ việc xác định này, chúng ta có thể lên kế hoạch một cách cụ thể, tỉ mỉ để phòng/chống dịch bệnh (chẳng hạn, ta có thể phân tích các đặc điểm của đối tượng này để tìm ra cách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

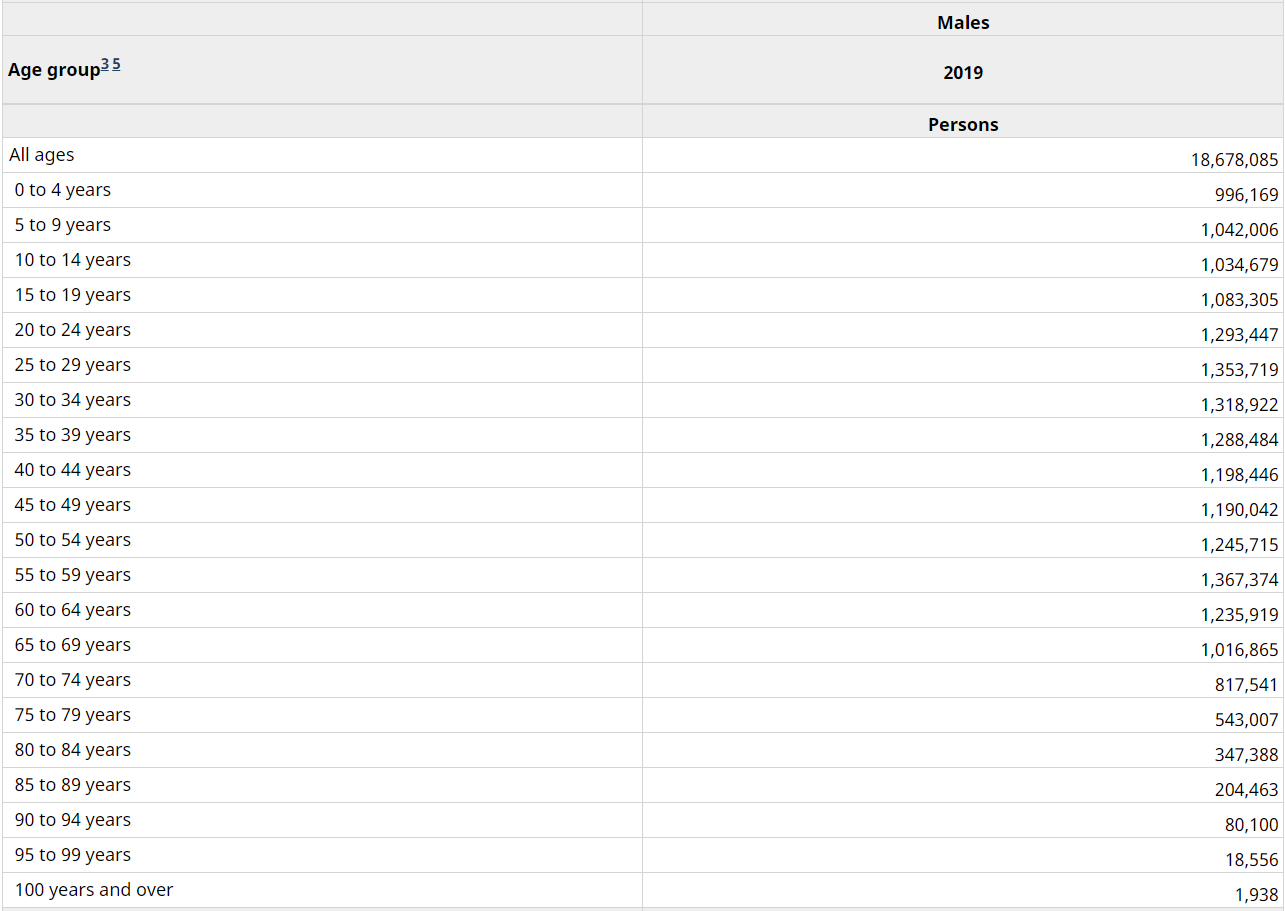
1. Bộ dữ liệu

* 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).



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



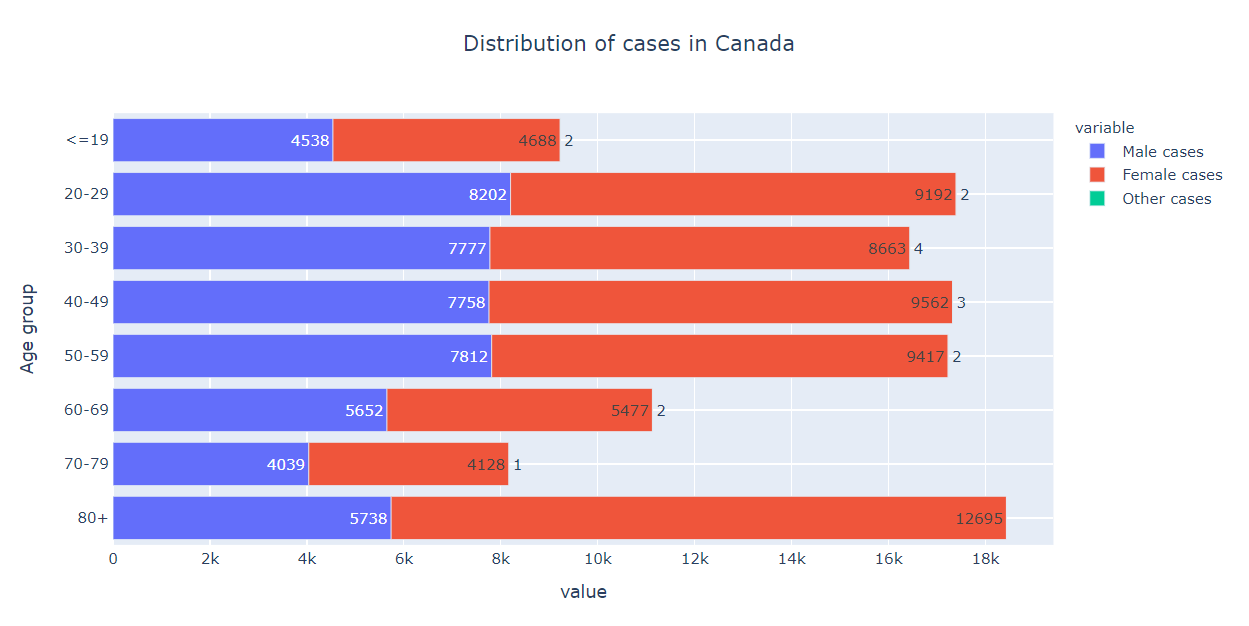


1. Chi tiết Mô hình/Thuật toán sử dụng

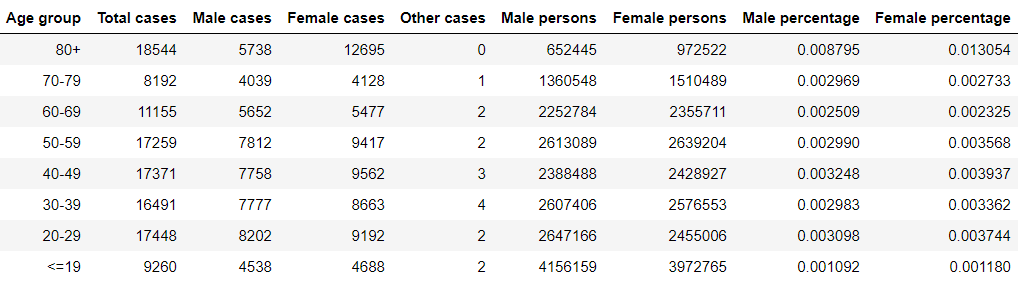
* Tính xác suất lây nhiễm của từng nhóm tuổi và giới tính.
* So sánh các xác suất này và chọn ra nhóm đối tượng nào có xác suất lớn nhất.

1. Kết quả

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



* Tính xác suất lây nhiễm của từng nhóm.



* 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).



1. Nhận xét

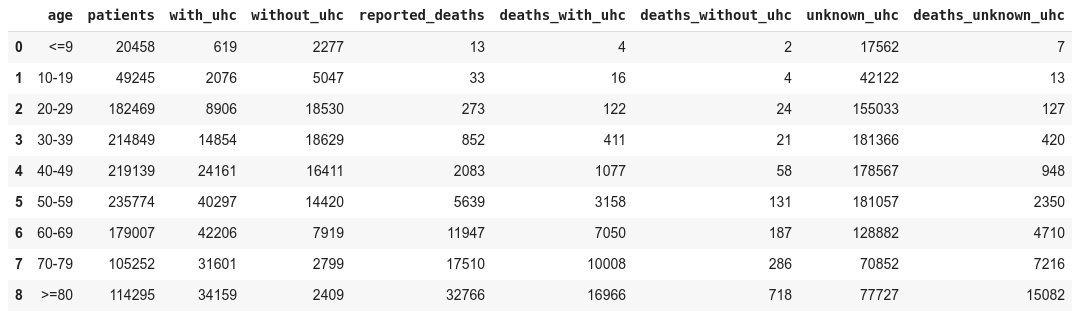
* Từ dữ liệu đã tính toán, ta dễ thấy nhóm tuổi từ **20-59** cho xác suất lây nhiễm khá cao và gần như nhau. Vì nhóm tuổi này thuộc độ tuổi lao động, nên việc tiếp xúc nhau nhiều đã tạo cơ hội cho dịch bệnh lây lan nhanh.
* Những người thuộc độ tuổi **80+** cho kết quả cao nhất. Vì ở độ tuổi này, các 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.

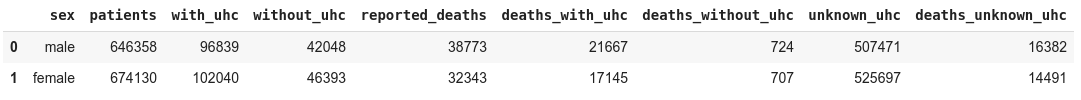
1. **Xác định đối tượng nhiễm mà dễ chết nhất**
2. Mô tả bài toán

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.

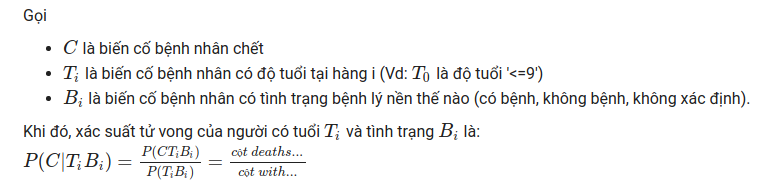
1. Chi tiết Mô hình/Thuật toán sử dụng

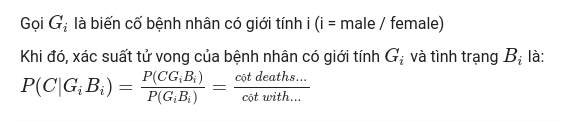
* Dữ liệu thống kê các ca tử vong theo độ tuổi và theo giới tính tại Mỹ trong thời gian từ ngày 22 tháng 1 đến ngày 30 tháng 05 (2020).

****



* 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:

****



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

1. 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 nhât 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.

1. **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://arxiv.org/pdf/2004.09478.pdf)

<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>