ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN VỀ COVID-19 SỬ DỤNG PROPHET

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Trần Quốc Khánh	18520908
2	Lê Trần Hoài Ân	18520426

TP. HỒ CHÍ MINH – 12/2020

1. GIỚI THIỆU

Trong phạm vi đồ án này, chúng tôi dựa vào những kỹ thuật khai thác, phân tích, trực quan dữ liệu đã được học để tìm hiểu về dữ liệu Covid-19 và tiến hành thử nghiệm xây dựng mô hình dự đoán tình hình Covid-19 (số ca nhiễm, số ca tử vong, …) cho từng quốc gia, khu vực riêng biệt.

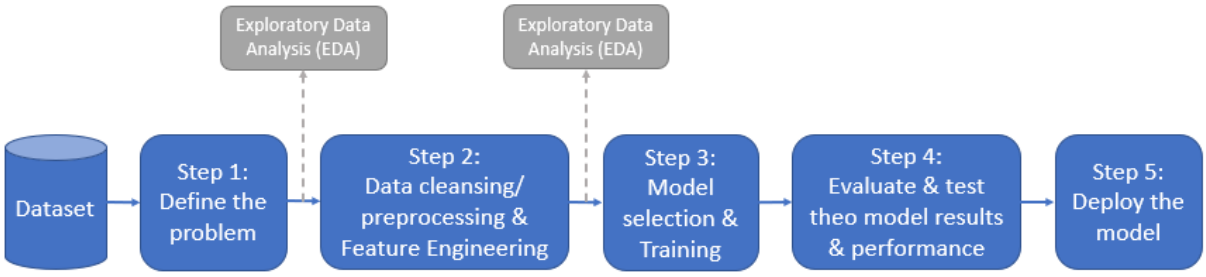
Chúng tôi nghiên cứu và sử dụng packet Prophet của Facebook để xây dựng mô hình liên quan đến dãy số thời gian (time series), đây là packet được đội ngũ nhân viên của Facebook phát triển để dành cho những mô hình dự đoán giá trị theo thời gian. Và chính bởi vì dữ liệu Covid-19 ở đây cũng đều được ghi nhận theo ngày, nên chúng tôi quyết định tiến hành thực nghiệm và so sánh với các mô hình hồi quy khác để kiểm chứng xem Facebook Prophet có thật sự hiệu quả để xây dựng mô hình dự đoán tình hình dịch bệnh dựa trên dữ liệu này hay không.

Hiện nay, đã có nhiều bài toán liên quan về Covid-19 trong Khoa học dữ liệu xuất hiện kể từ khi đại dịch Covid-19 bắt đầu bùng nổ trên thế giới, nhưng nhìn chung đây vẫn là một đề tài rất mới, còn nhiều khía cạnh, phương thức để tiếp cận và khai thác.

Sau quá trình nghiên cứu bộ dữ liệu và xây dựng thành công mô hình, nhóm chúng tôi thu được các kết quả tương đối khả quan cho thấy được phần nào mức độ phù hợp của Facebook Prophet khi dùng để xây dựng mô hình dự đoán bài toán về dữ liệu Covid-19 cũng như các bài toán về dữ liệu liên quan đến dòng thời gian (time series).

2. NỘI DUNG

Trong phần này, chúng tôi tiến hành tiếp cận dựa theo quy trình thực hiện phân tích đã được học như ở Hình 1 bên dưới:



Hình 1. *Quy trình Phân tích dữ liệu.*

2.1. Giới thiệu bộ dữ liệu

- **Tên bộ dữ liệu:** Novel Corona Virus 2019 Dataset
- **Nguồn dữ liệu:** Bộ dữ liệu được xây dựng tổng hợp từ bộ dữ liệu Novel Corona Virus 2019 của Kaggle và dữ liệu nhóm thu thập bổ sung từ trang web https://www.worldometers.info/.
- **Ý nghĩa bộ dữ liệu:** Bộ dữ liệu chứa thông tin về tình hình dịch Covid-19 tại các quốc gia trên thế giới (số ca nhiễm, số ca tử vong, số ca phục, …) theo dòng thời gian.
- **Số thuộc tính:** 8

- **Số điểm dữ liệu:** 156.292
- **Dữ liệu bị khuyết:** Province/State – 44313 điểm dữ liệu (28,35%)
- **Codebook mô tả thông tin các thuộc tính của bộ dữ liệu:**

Bảng 1. Bảng thống kê thông tin các thuộc tính bộ dữ liệu Novel Corona Virus 2019

STT	Tên thuộc tính	Ý nghĩa	Kiểu dữ liệu	Miền giá trị
1	SNo	Số thứ tự	Numeric	1-156292
2	ObservationDate	Ngày ghi nhận	Date	22/01/2020-15/11/2020
3	Province/State	Tỉnh thành/ Tiểu bang	String	Anhui, Beijing, Fujian, Gansu,…
4	Country/Region	Quốc gia/ Vùng lãnh thổ	String	Mainland China, US, Japan, South Korea,...
5	Last Update	Ngày cập nhật	Date	22/09/2020-16/11/2020
6	Confirmed	Số ca nhiễm	Numeric	0–1867721
7	Deaths	Số ca tử vong	Numeric	-178(*)–45974
8	Recovered	Số ca khỏi bệnh	Numeric	-854405(*)–4174884 (**)

(*) (**): Trong bộ dữ liệu có tồn tại 3 điểm dữ liệu với số ca tử vong và/hoặc số ca hồi phục mang giá trị âm. Có thể do quá trình ghi nhận và tạo dữ liệu đã xảy ra sai sót hoặc nhầm lẫn nên một số điểm dữ liệu chứa những giá trị không phù hợp.

Hầu hết các điểm dữ liệu sẽ mô tả số ca Covid-19 theo thành phố, khu vực. Nhưng với một số quốc gia (Hoa kỳ, Colombia) sẽ là số liệu tổng hợp số ca khỏi bệnh trên toàn quốc gia, thay vì theo từng bang/tỉnh. Việc này dẫn đến sự không đồng thuận ở một số điểm dữ liệu thuộc trường hợp nói trên so với các điểm dữ liệu khác.

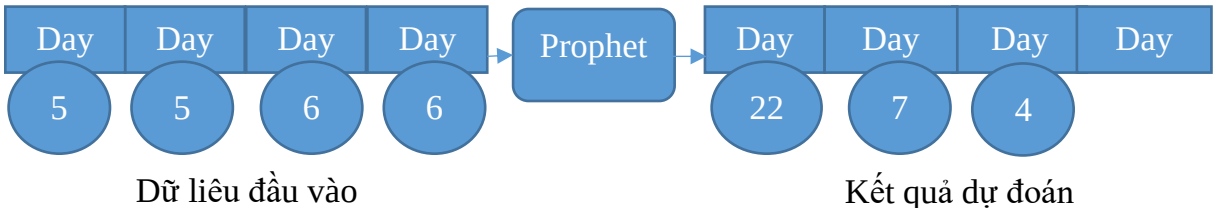
Do đây là dữ liệu được ghi nhận và tổng hợp từ nhiều quốc gia, khu vực trên toàn thế giới với lượng dữ liệu tương đối lớn nên xuất hiện những hạn chế, nhưng đổi lại bộ dữ liệu có nguồn gốc chính thống, khách quan.

2.2. Xác định vấn đề bài toán (Time series và Covid-19)

Với sự bùng nổ của đại dịch Covid-19 trên toàn cầu, đã có không ít bài toán dự đoán các vấn đề liên quan xuất hiện. Nhưng nếu không có nhiều dữ liệu và bỏ qua các lĩnh vực liên quan đến tình hình bệnh dịch (chính sách quốc gia, xã hội, y tế, …) thì dữ liệu Covid-19 đơn thuần còn lại là thời gian và địa điểm ghi nhận tình hình. Đó cũng là những gì thuộc về bộ dữ liệu chúng tôi đang sử dụng. Với bài toán có một biến giá trị được nghiên cứu theo thời gian chúng tôi đã quyết định chọn thử nghiệm mô hình dãy số thời gian (Time series) với Facebook Prophet (trình bày ở phần sau).

Dãy số thời gian [3] thường được dùng để nghiên cứu biến động của kinh tế xã hội theo thời gian. Dãy số thời gian là dãy các trị số của chỉ tiêu thống kê được sắp xếp theo thứ tự thời gian. Dãy số thời gian cho phép thống kê học nghiên cứu đặc điểm biến động của hiện tượng theo thời gian vạch rõ xu hướng và tính quy luật của sự biến động, đồng thời dự đoán các mức độ của hiện tượng trong tương lai. Dãy số thời gian gồm hai thành phần: thời gian và chỉ tiêu của hiện tượng được nghiên cứu (cụ thể trong bài toán này là số ca nhiễm/ tử vong/ hồi phục).

Dựa vào những thuộc tính sẵn có của bộ dữ liệu, kết hợp với các thuộc tính mới có liên quan đến nhiệm vụ dự đoán được chúng tôi tạo ra và những đặc tính của dãy số thời gian, nhóm đưa ra quyết định tiến hành xây dựng mô hình dự đoán với đầu vào là chuỗi thời gian, số ca nhiễm/ tử vong/ hồi phục và đầu ra dự đoán sẽ là số ca nhiễm/ tử vong/ hồi phục trong một khoảng thời gian dự đoán cho toàn thế giới và cho từng quốc gia, khu vực riêng biệt.



Hình 1. Mô tả bài toán

2.3. Tiền xử lý và phân tích thăm dò (EDA)

2.3.1. Tiền xử lý dữ liệu

Với tính tiện lợi và đơn giản trong việc xây dựng mô hình dự đoán từ các mô hình Hồi quy đơn giản và đặc biệt là packet Facebook Prophet thì việc chuẩn bị dữ liệu nhóm chúng tôi sẽ tập trung vào việc làm sạch và xử lý dữ liệu theo khu vực để tăng thông tin thu được từ dữ liệu nhằm phát triển hiệu suất mô hình.

Đối với một số ít các điểm dữ liệu lỗi (thiếu, sai về giá trị, ý nghĩa) nhóm chúng tôi quyết định sẽ loại bỏ khỏi bộ dữ liệu, vì số lượng điểm dữ liệu lỗi là rất ít, khó có khả năng gây ảnh hưởng nhiều đến hiệu suất huấn luyện và một phần đây là dữ liệu số ca được ghi nhận và con số này có thể biến động rất khó lường nên nhóm đã không chọn cách thay thế bằng một giá trị khác.

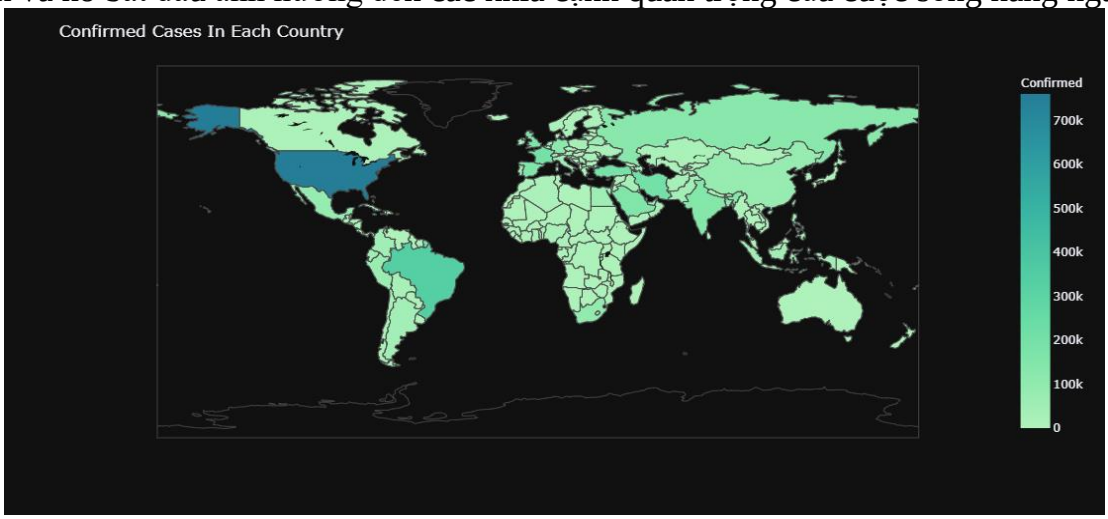
Đối với thuộc tính **Province/State** có 28,35% điểm dữ liệu bị khuyết bằng cách thay thế chúng với giá trị **“Unknown”.**

Bộ dữ liệu hiện có đã có đủ các yếu tố để mang vào huấn luyện. Tuy nhiên, để tăng thông tin thu được, chúng tôi đã tiến hành tạo thêm các thuộc tính mới dựa trên các thuộc tính có sẵn của dữ liệu:

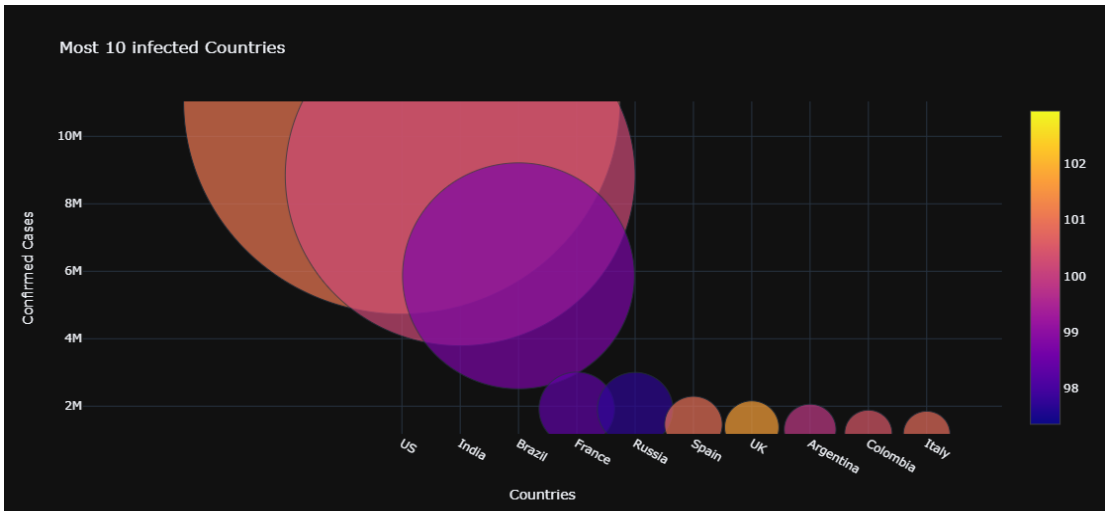
- **New_Daily_case:** Số ca nhiễm mới mỗi ngày
- **New_Deaths:** Số tử vong mới mỗi ngày
- **New_Recovered:** Số ca hồi phục mới mỗi ngày
- **Active_case:** Số ca bệnh hiện có

2.3.2. Phân tích thăm dò

Sự bùng phát COVID-19 đang phát triển thành một cuộc khủng hoảng quốc tế lớn và nó bắt đầu ảnh hưởng đến các khía cạnh quan trọng của cuộc sống hàng ngày.



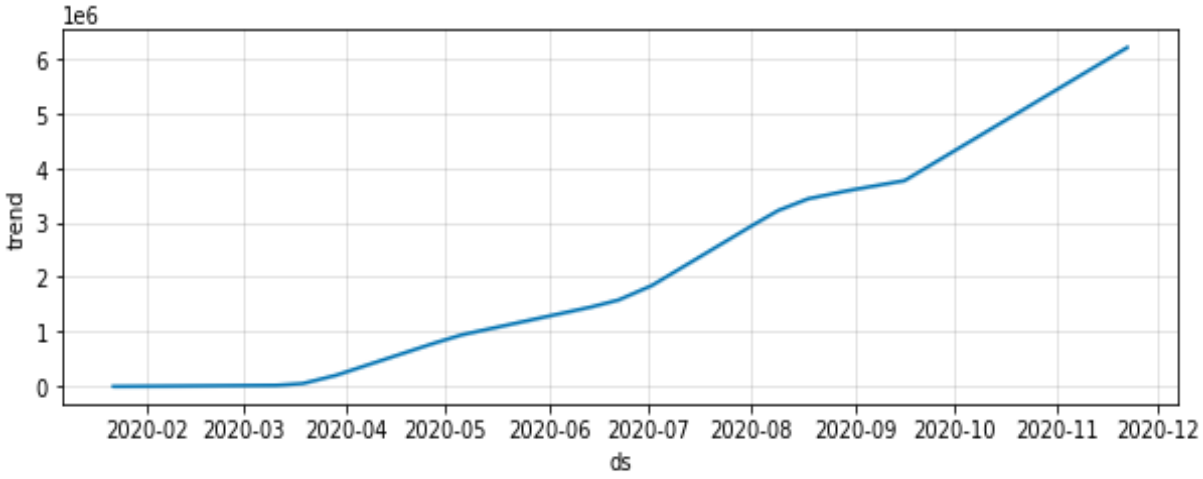
Hình 2. Biểu đồ thể hiện số ca nhiễm theo từng quốc gia trên thế giới



Hình 3. Top 10 quốc gia bị ảnh hưởng nhiều nhất bởi Covid-19

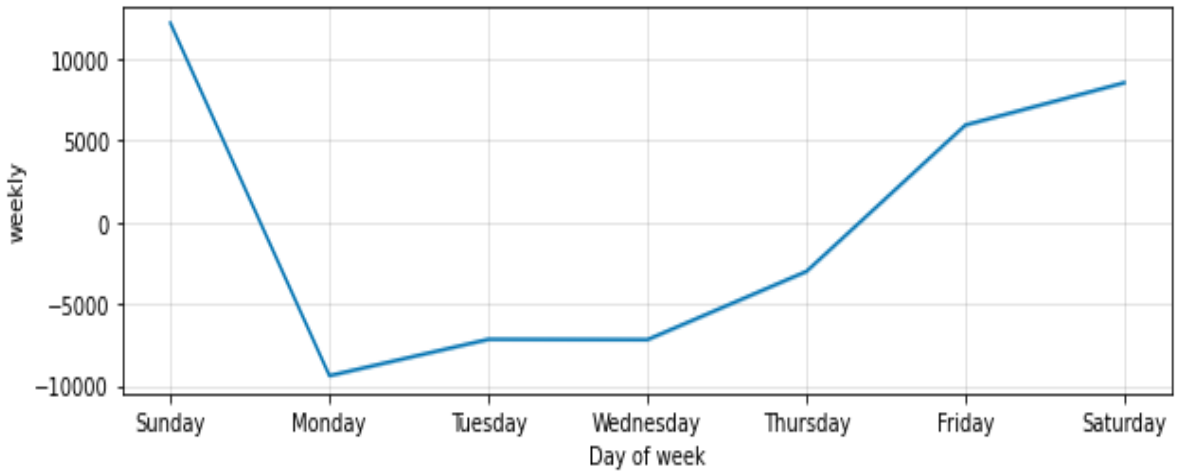
Quan sát các Hình 2 và 3 chúng ta có thể thấy các quốc gia như Hoa kỳ, Ấn Độ, Brazil,… đang bị ảnh hưởng nhiều nhất bởi tình hình Covid-19. Nguyên nhân cho điều

này là các quốc gia kể trên đều là các quốc gia tập trung đông dân cư, có lịch sử xuất nhập cảnh phức tạp. Đồng thời, chính sách đối phó chủ quan với Covid-19 cũng là 1 trong những nguyên nhân dẫn đến tình trạng dịch bệnh không ngừng gia tăng ở các quốc gia này.



Hình 4. Biểu đồ thể hiện xu hướng phát triển của Covid-19 trên thế giới

Ở Hình 4, chúng ta dễ dàng nhận thấy rằng xu hướng phát triển của Covid-19 vẫn không ngừng tăng nhanh và khó kiểm soát. Ngày càng có nhiều chủng virus liên quan đến Covid-19 được phát hiện, chúng đều có tốc độ lây lan nhanh, mức độ nguy hiểm đáng quan ngại. Đặc biệt, các tháng mùa Đông là các tháng có nhiệt độ thấp, tạo điều kiện thuận lợi cho virus tồn tại, lây lan dẫn đến tình trạng số ca mắc có xu hướng tăng vọt vào các tháng này.



Hình 5. Biểu độ thể hiện xu hướng phát triển của Covid-19 theo ngày

Ở Hình 5, số lượng ca nhiễm Covid-19 được ghi nhận nhiều nhất vào các ngày cuối tuần (từ thứ 6 đến chủ nhật). Điều này có thể giải thích là do vào các ngày này, người dân thường có xu hướng tụ tập vui chơi, mua sắm, du lịch bên ngoài bất chấp các lệnh giới nghiêm đã làm cho tình hình dịch bệnh diễn biến thêm phần phức tạp.

Chính vì vậy, chúng ta cần có các mô hình dự đoán đủ mạnh, đủ chính xác và đáng tin cậy để có thể đưa ra các chính sách, các chủ trương thích hợp nhằm góp phần làm giảm nguy cơ dịch bệnh lan rộng, sớm đưa trật tự đời sống trở về lại với bình thường.

2.3.3. Lựa chọn các đặc trưng

Nhóm chúng tôi loại bỏ hai thuộc tính **SNo** và **Last Update** vì đây là hai thuộc tính sẽ không tham gia hay sử dụng trong quá trình phân tích, huấn luyện. Đối với hai thuộc tính **Province/State** và **Country/Region** sẽ không dùng trong việc đưa vào huấn luyện trực tiếp nhưng chúng tôi sẽ dùng để chia dữ liệu thành dữ liệu ở các khu vực để có những điều chỉnh tham số mô hình phù hợp với khu vực, cũng như việc phân tích, nghiên cứu cụ thể hơn về tình hình ở quốc gia, khu vực đó.

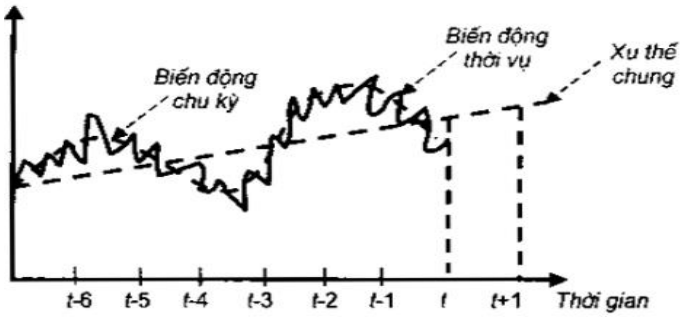
Sau khi xử lý, dữ liệu cuối cùng được đưa vào trực tiếp để huấn luyện mô hình sẽ là thuộc tính **ObservationDate** (ngày ghi nhận) và biến được lựa chọn để dự đoán trong các thuộc tính: **Confirmed, Deaths, Recovered, New_Daily_case, New_Deaths, New_Deaths, Active_case.**

2.4. Phương pháp tiếp cận

Nhóm chúng tôi sử dụng **Facebook Prophet [1]** – một packet được phát triển tuân theo mô hình API của sklearn. Prophet giúp hầu hết mọi người có thể dự đoán các giá trị của chuỗi thời gian ngay cả khi có rất ít hoặc không có kinh nghiệm trong lĩnh vực thực hiện. Với việc đầu vào luôn là một khung dữ liệu bao gồm chỉ hai thành phần là ngày và một biến số, Facebook Prophet sẽ là phương tiện thích hợp để tiến hành xây dựng mô hình dự đoán trên bộ dữ liệu mà chúng tôi có.

Facebook Prophet dựa vào các thuật toán phân tích những biến động trong dãy số thời gian và sự ảnh hưởng lên dữ liệu của nó. Cụ thể một số thành tố tác động tới dãy số thời gian:

- **Xu thế chung:** Biểu thị hướng di chuyển tương đối trơn tru của dãy số thời gian trong thời gian dài
- **Biến động chu kì:** bao gồm những loại biến động trung hạn, lặp đi lặp lại và nhìn chung gắn với chu kì kinh doanh.
- **Biến động thời vụ:** bao gồm những loại biến động trung hạn, lặp đi lặp lại và gắn liền với các mùa khác trong năm. Những biến động thời vụ này chồng lên xu thế chung và biến động chu kì.
- **Biến động bất thường:** là những thay đổi thất thường trong dãy số thời gian do những biến cố ngẫu nhiên không dự báo trước được gây ra. Những biến động thất thường này chồng lên xu thế chung, biến động chu kì và biến động thời vụ.



Hình 6. Minh họa về Time series

Với việc yêu cầu đầu vào đơn giản chỉ có hai thành phần thì đi kèm Facebook Prophet sẽ có khá nhiều các tham số tinh chỉnh do người dùng cung cấp nhằm giúp mô hình phù hợp hơn khi dùng cho các bộ dữ liệu, bài toán khác nhau. Cụ thể một vài tham số quan trọng như:

- **growth:** Tham số có hai giá trị là **"linear"** và **"logistic"** chỉ xu thế chung của dữ liệu.
- **changepoints:** Chỉ những điểm trong dữ liệu có sự thay đổi đột ngột (biến động bất thường), ngoài ra còn có các tham số **n_changepoints, changepoint_range** và **changepoint_prior_scale.**
- **seasonality_mode:** Cho biết các thành phần thời vụ nên được tích hợp như thế nào với những dự đoán. Tham số có hai giá trị **"additive"** và **"multiplicative".**

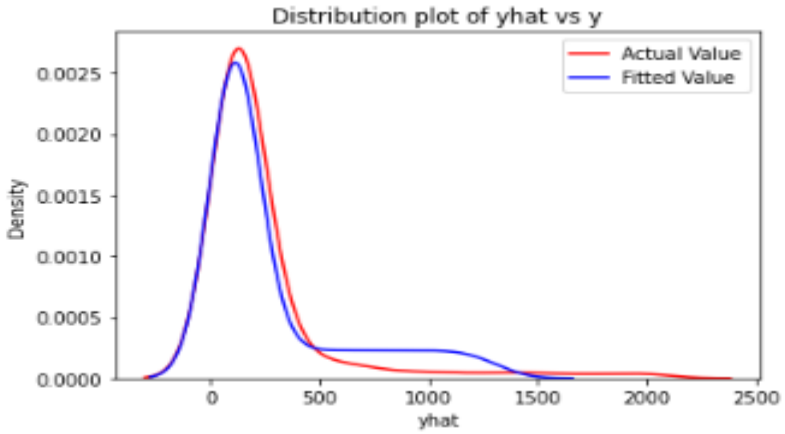
Các tham số về dãy số thời gian trong Facebook Prophet nhìn chung sẽ quyết định các đặc điểm, tính chất mà thời gian có khả năng tác động đến dữ liệu được sử dụng. Nhưng một số yếu tố thời gian như ngày lễ, các mùa trong năm ở tùy khu vực, quốc gia sẽ có những biến động và sức ảnh hưởng khác nhau, điều này tạo ra khó khăn trong việc xây dựng mô hình dự đoán chung. Thay vào đó để tối ưu hiệu suất sẽ có những lựa chọn thông số phù hợp riêng với các quốc gia, khu vực.

2.5. Phát triển mô hình, so sánh hiệu suất và đánh giá

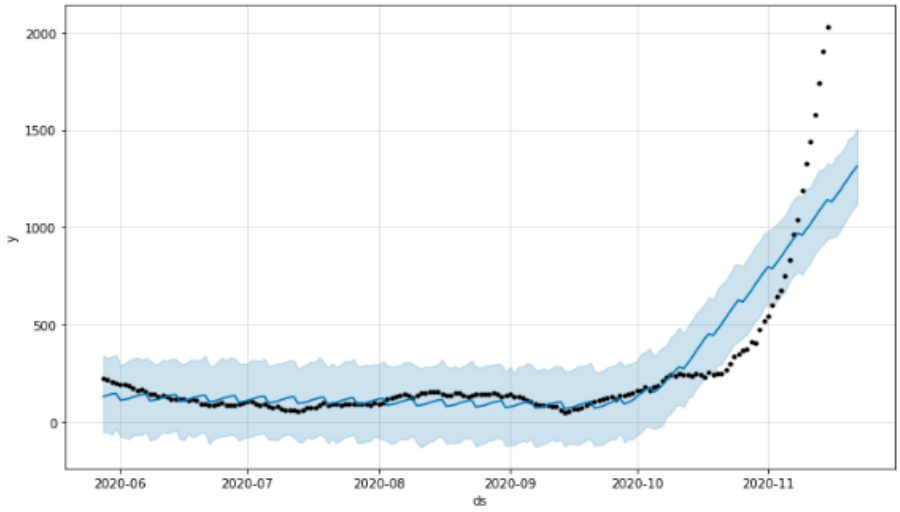
Để đánh giá, phân tích hiệu suất của các mô hình nhóm chúng tôi chọn độ đo đánh giá R – square đánh giá tỷ lệ giải thích của mô hình ước lượng, hệ số này nằm giữa 0 và 1, càng gần 1 tỷ lệ giải thích được của mô hình càng tốt. Bên cạnh đó, tiến hành trực quan kết quả thu được bằng các biểu đồ để tạo điều kiện dễ dàng quan sát và nhận định tình hình.

Như đã đề cập ở phần trên, dữ liệu sẽ được trích ra để huấn luyện tùy thuộc vào khu vực, quốc gia để có những tham số mô hình phù hợp với từng trường hợp cụ thể. Ở phần này nhóm chúng tôi sẽ trình bày một số kết quả thu được từ việc huấn luyện dữ liệu ở một số quốc gia, cụ thể sẽ là toàn lãnh thổ **Hoa Kỳ** và tỉnh **Tokyo** ở **Nhật Bản**, với biến số dự đoán là **Active_case.**

2.5.1. *Kết quả dự đoán thu được từ dữ liệu thuộc Tokyo*

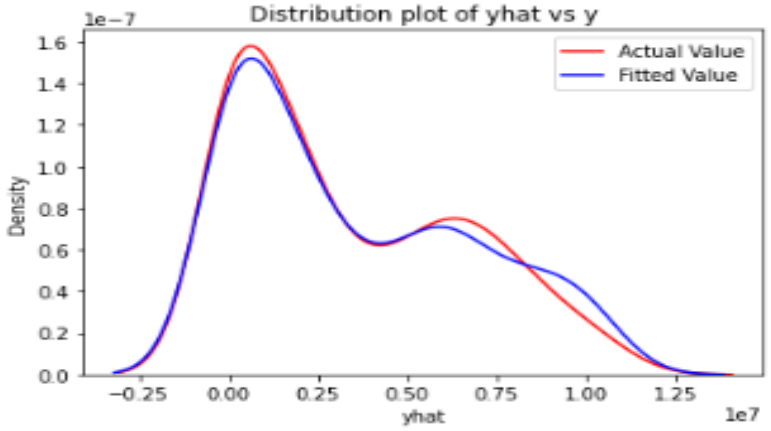


Hình 7. Distribution plot kết quả dự đoán tình hình Covid-19 của Tokyo – Nhật Bản

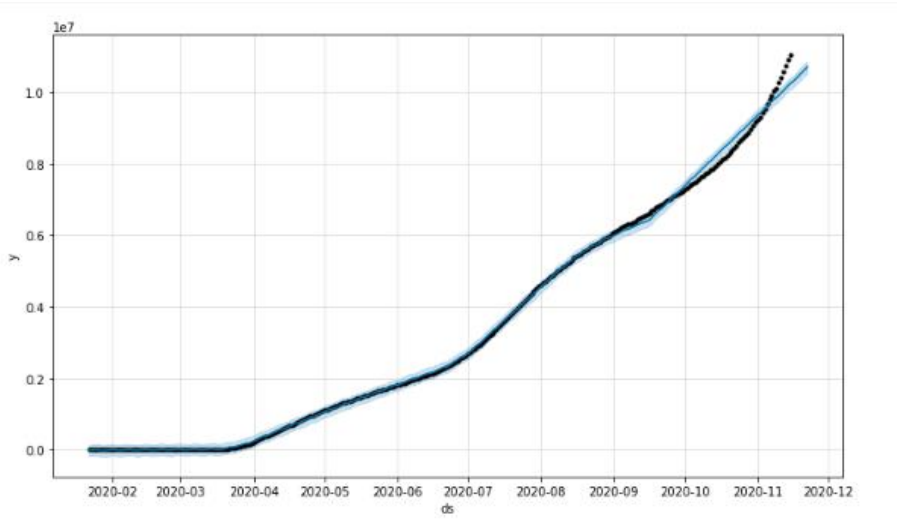


Hình 8. Kết quả dự đoán tình hình Covid-19 của Tokyo – Nhật Bản

2.5.2. *Kết quả dự đoán thu được từ dữ liệu thuộc Hoa kỳ*



Hình 9. Distribution plot kết quả dự đoán tình hình Covid-19 của Hoa kỳ

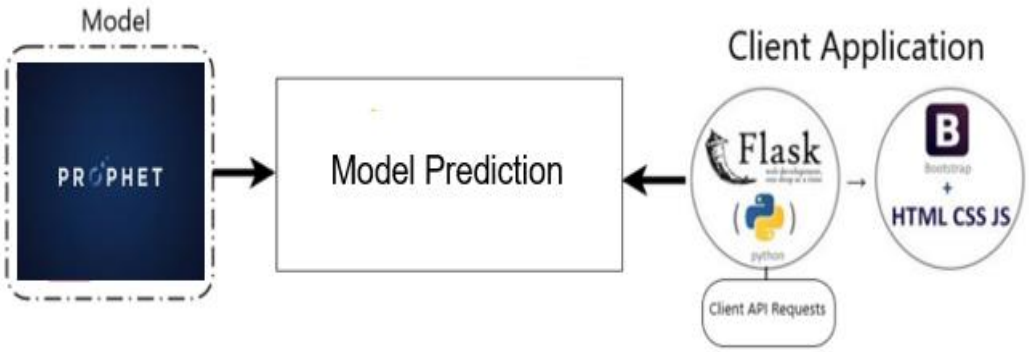


Hình 10. Kết quả dự đoán tình hình Covid-19 của Hoa Kỳ

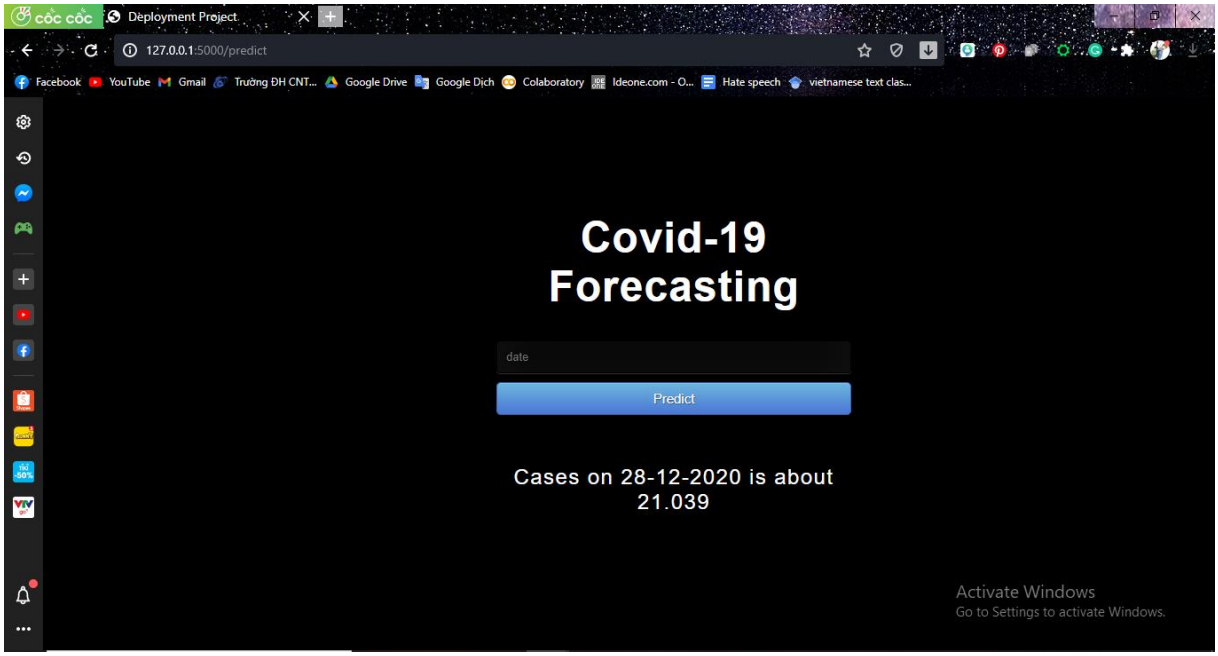
Kết quả đạt được là tương đối khả quan, mặc dù tình hình dịch bệnh còn bị ảnh hưởng bởi rất nhiều yếu tố khác và thời gian không phải là một yếu tố trọng yếu quyết định. Nhưng ngoài kết quả từ hai khu vực được trình bày ở trên thì nhóm chúng tôi cũng thu được những kết quả khả quan tương tự đối với các khu vực khác. Dựa vào các kết quả R2 score đó và hình dạng của mô hình với dữ liệu, nhóm đưa ra kết luận Facebook Prophet có những khả năng, tính chất phù hợp để sử dụng trong việc dự đoán về tình hình dịch bệnh Covid-19.

2.6. Triển khai mô hình (Deploy model)

Sau khi đã xây dựng thành công các mô hình dự đoán tình hình dịch bệnh Covid-19 cho các quốc gia, khu vực khác nhau trên thế giới, chúng tôi tiến hành xây dựng các công cụ để có thể dễ dàng đưa ra các dự đoán, phát triển mô hình trong tương lai dựa trên framework Flask [4] của ngôn ngữ Python.



Hình 11. Nguyên lý hoạt động của framework Flask



Hình 12. *Kết quả triển khai mô hình bằng framework Flask*

3. KẾT LUẬN

Sau khi tiến hành phân tích và nghiên cứu bộ dữ liệu, nhận thấy sự tương đồng với các bài toán về dãy số thời gian và bài toán dự đoán về tình hình Covid-19 thì nhóm chúng tôi đã quyết định sử dụng Facebook Prophet để tiến hành thực nghiệm xây dựng mô hình dự đoán. Việc huấn luyện trên toàn bộ dữ liệu khu vực, quốc gia sẽ mang lại kết quả không tốt vì mỗi quốc gia, khu vực sẽ có rất nhiều yếu tố khác nhau tác động đến tình hình của dịch bệnh, do đó nhóm chúng tôi tiến hành huấn luyện mô hình riêng cho mỗi khu vực, quốc gia. Mô hình sau khi được huấn luyện đã đạt được những kết quả tương đối khả quan (R2_score > 0.9).

Đề tài này có thể xem là một trong các giải pháp mới và hiệu quả cho bài toán về Covid-19. Với bộ dữ liệu hiện có thì việc nghiên cứu các phương pháp khác để xây dựng mô hình dự đoán cũng như quan sát đặc tính của của bộ dữ liệu là có rất nhiều hướng. Song, sau quá trình nghiên cứu trong phạm vi đồ án nhóm chúng tôi đưa ra nhận định có thể sử dụng Facebook Prophet trong bài toán xây dựng mô hình dự đoán về Covid-19.

Trong tương lai, chúng tôi sẽ tiếp tôi tìm hiểu, nghiên cứu sâu hơn về các mô hình có thể hỗ trợ cho việc dự đoán diễn biến tình hình dịch bệnh Covid-19 với hiệu quả tốt hơn, độ tin cậy cao hơn nhằm chỉ ra các giải pháp đúng đắn, kịp thời, góp phần vào việc kiểm soát sự lây lan của dịch bệnh này.

TÀI LIỆU THAM KHẢO

[1] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." *The American Statistician* 72.1 (2018): 37-45.

[2] Aggarwal, S. (2019). Flask Framework Cookbook: Over 80 proven recipes and techniques for Python web development with Flask. Packt Publishing Ltd.

[3] Nathan, "Cơ Bản Về Lớp Giải Thuật Time Series Forecasting," 2020. [Trực tuyến]. Available: https://insights.magestore.com/posts/giai-thuat-time-series-forecasting. [Đã truy cập 2020].

[4] A. Sagar, "How to Easily Deploy Machine Learning Models Using Flask," 2019. [Trực tuyến]. Available: https://towardsdatascience.com/how-to-easily-deploy-machine-learning-models-using-flask-b95af8fe34d4. [Đã truy cập 2020].

Nhóm sinh viên đã có áp dụng các kiến thức về phân tích, trực quan dữ liệu đã được học và có bài báo được đăng tại Hội nghị Khoa học trẻ và Nghiên cứu sinh UIT 2020: "Xây Dựng Giải Pháp Chấm Điểm Tín Dụng Tại Thị Trường Việt Nam".

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Quốc Khánh	- Thực hiện phân tích, trực quan dữ liệu. - Thực hiện deploy model. - Viết báo cáo.
2	Lê Trần Hoài Ân	- Tìm hiểu, xây dựng các mô hình dự đoán. - Viết báo cáo và soạn slide báo cáo.