TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA ĐIỆN – ĐIỆN TỬ**



**PHẠM HOÀNG TẤN DUY**

**UPDATE DAILY COVID-19 DATA,**

**LƯU TRỮ VÀ XÂY DỰNG**

**MÔ HÌNH DỰ ĐOÁN**

**ĐỒ ÁN TỔNG HỢP**

**KỸ THUẬT ĐIỆN TỬ - VIỄN THÔNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA ĐIỆN – ĐIỆN TỬ**



**PHẠM HOÀNG TẤN DUY**

**UPDATE DAILY COVID-19 DATA,**

**LƯU TRỮ VÀ XÂY DỰNG**

**MÔ HÌNH DỰ ĐOÁN**

**ĐỒ ÁN TỔNG HỢP**

**KỸ THUẬT ĐIỆN TỬ - VIỄN THÔNG**

Người hướng dẫn

**TS. Nguyễn Thành Nam**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

**LỜI CẢM ƠN**

Tôi xin chân thành cảm ơn Thầy Nguyễn Thành Nam đã giúp đỡ em trong quá trình thực hiện và hoàn thành Đồ án tổng hợp.

*TP. Hồ Chí Minh, ngày 21 tháng 12 năm 2021*

*Tác giả*

Công trình được hoàn thành tại Trường Đại học Tôn Đức Thắng

Cán bộ hướng dẫn khoa học: TS. Nguyễn Thành Nam

*(Ghi rõ học hàm, học vị, họ tên và chữ ký)*

Đồ án tổng hợp được bảo vệ tại **Hội đồng đánh giá Đồ án tốt nghiệp/tổng hợp của Trường Đại học Tôn Đức Thắng** vào ngày… /…/……

Xác nhận của Chủ tịch Hội đồng đánh giá Đồ án tốt nghiệp/tổng hợp và Trưởng khoa quản lý chuyên ngành sau khi nhận Đồ án tốt nghiệp/tổng hợp đã được sửa chữa (nếu có).

**CHỦ TỊCH HỘI ĐỒNG TRƯỞNG KHOA**

**…………………………. ………………………………**

**CÔNG TRÌNH ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS Nguyễn Thành Nam. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Đồ án tốt nghiệp/ tổng hợp còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Đồ án tốt nghiệp/ tổng hợp của mình.** Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 21 tháng 12 năm 2021*

*Tác giả*

*Phạm Hoàng Tấn Duy*

(Trang này dùng để đính kèm Nhiệm vụ Đồ án tốt nghiệp có chữ ký của Giảng viên hướng dẫn)

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG  **KHOA ĐIỆN –ĐIỆN TỬ**  ------------------- | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  Độc lập – Tự do – Hạnh phúc  ---------------------- |

**LỊCH TRÌNH LÀM ĐỒ ÁN TỐT NGHIỆP**

Họ tên sinh viên: PHẠM HOÀNG TẤN DUY ............................................................

Lớp: 15040201 MSSV: 41502101.................................

Tên đề tài: Update daily COVID-19 data, lưu trữ và xây dựng mô hình dự đoán. ..............................................................

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuần/Ngày** | **Khối lượng** | | **GVHD ký** |
| **Đã thực hiện** | **Tiếp tục thực hiện** |
| **1** |  |  |  |
| **2** |  |  |  |
| **3** |  |  |  |
| **4** |  |  |  |
| **5** |  |  |  |
| **6** |  |  |  |
| **7** |  |  |  |
| **Kiểm tra giữa kỳ** | Đánh giá khối lượng hoàn thành……..%  được tiếp tục/không tiếp tục thực hiện ĐATN | | |
| **9** |  |  |  |
| **10** |  |  |  |
| **11** |  |  |  |
| **12** |  |  |  |
| **13** |  |  |  |
| **14** |  |  |  |
| **15** |  |  |  |
| Nộp Đồ án tốt nghiệp | Đã hoàn thành……..% Đồ án tốt nghiệp  được bảo vệ/không được bảo vệ ĐATN | | |

**UPDATE DAILY COVID-19 DATA, LƯU TRỮ**

**VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN**

**TÓM TẮT**

Trong bối cảnh tình hình dịch COVID-19 đang diễn biến phức tạp trên toàn cầu và ở Việt Nam, gây ra nhiều sự mất mát, khó khăn và ảnh hưởng tiêu cực đến nền kinh tế toàn cầu và ở Việt Nam nên em quyết định sử dụng sự kiện này để làm đề tài nghiên cứu cho Đồ án Tổng hợp của em cho HK1 năm học 2021-2022. Tuy nhiên, đề tài lúc đăng ký có tên là “Streaming COVID-19 data, lưu trữ và xây dựng mô hình dự đoán”, do đề tài của em chỉ cập nhật data mỗi ngày vào khung giờ nhất định chứ không cập nhật real-time nên Giáo viên phản biện có yêu cầu em đổi tên đề tài cho phù hợp là: Update COVID-19 data, lưu trữ và xây dựng mô hình dự đoán.

**[THESIS TITLE]**

**ABSTRACT**

**(BOLD, SIZE 16)**

(Times New Roman – 13) ……………………………………………………………

………………………………………………………………………………………...

………………………………………………………………………………………...

………………………………………………………………………………………...

………………………………………………………………………………………...

MỤC LỤC

DANH MỤC HÌNH VẼ xiii

DANH MỤC BẢNG BIỂU xvi

DANH MỤC CÁC CHỮ VIẾT TẮT xvii

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1

1.1 Giới thiệu chung 1

1.1.1 Giới thiệu về COVID-19 1

1.1.2 Giới thiệu về Vắc-xin COVID-19 2

1.2 Mục đích thực hiện đề tài 3

1.3 Yêu cầu của đề tài 3

1.4 Phương pháp thực hiện 3

CHƯƠNG 2. LÝ THUYẾT ĐỀ TÀI 4

2.1 Giới thiệu về Big Data và Khoa học Dữ liệu 4

2.1.1 Big Data 4

2.1.2 Khoa Học Dữ Liệu (Data Science) 7

2.2 Giới thiệu về Data Mining và Machine Learning 8

2.2.1 Data Mining 8

2.2.2 Machine Learning 10

2.3 Giới thiệu Python và các thư viện Python sử dụng trong đề tài 12

2.3.1 Python 12

2.3.2 Các thư viện của Python 13

2.4 Giới thiệu SQL và Cơ sở dữ liệu MySQL 14

2.4.1 SQL 14

2.4.2 Cơ sở dữ liệu MySQL 16

2.5 Giới thiệu Jupyter Notebook, API và Kaggle 17

2.5.1 Jupyter Notebook 17

2.5.2 API (Application Programming Interface) 17

2.5.3 Kaggle 18

2.6 Giới thiệu về Docker và Apache Airflow 19

2.6.1 Docker 19

2.6.2 Apache Airflow 20

CHƯƠNG 3. THU THẬP, PHÂN TÍCH VÀ BIỂU ĐỒ 21

3.1 Thiết kế hệ thống 21

3.1.1 Kiến trúc hệ thống 21

3.1.2 Sơ đồ khối 22

3.2 Tổng quát về dịch COVID-19 23

3.2.1 Thu thập dữ liệu từ RapidAPI và lưu trữ vào Cơ sở dữ liệu MySQL 23

3.2.2 Thiết kế Dashboard bằng phần mềm Power BI 25

3.3 Chi tiết về dịch COVID-19 32

3.3.1 Giới thiệu 32

3.3.2 Biểu đồ và phân tích với Python 32

3.3.2.1 So sánh Việt Nam và Châu Á 36

3.3.2.2 So sánh Việt Nam và Đông Nam Á (South East Asia) 38

3.4 Tình hình tiêm ngừa Vắc-xin COVID-19 40

3.4.1 Tình hình tiêm ngừa Vắc-xin COVID-19 của Thế giới 40

3.4.2 So sánh Việt Nam và Châu Á 43

3.4.3 So sánh Việt Nam và Đông Nam Á (South East Asia) 45

CHƯƠNG 4. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN 47

4.1 Thuật toán SVM, Linear Regression và Bayesian Ridge Regression 47

4.1.1 Thuật toán SVM (Support Vector Machines) 47

4.1.2 Thuật toán Linear Regression 51

4.1.2.1 Giới thiệu Linear Regression 51

4.1.2.2 GIới thiệu OLS (Ordinary Least Square): 53

4.1.3 Thuật toán Bayesian Ridge Regression 54

4.2 Áp dụng vào mô hình dự đoán 56

4.2.1 Áp dụng thuật toán SVM 56

4.2.2 Áp dụng thuật toán Linear Regression 57

4.2.3 Áp dụng thuật toán Bayesian Ridge Regression 59

4.3 Kết quả dự đoán và so sánh 60

4.4 Áp dụng Polynomial Regression dự đoán số ca nhiễm ở Việt Nam 61

CHƯƠNG 5. KẾT LUẬN 63

5.1 Kết luận 63

5.2 Hướng phát triển 64

TÀI LIỆU THAM KHẢO 65

PHỤ LỤC A MÃ NGUỒN PYTHON 1

# DANH MỤC HÌNH VẼ

[Hình 1‑1. Hình ảnh của COVID-19 1](#_Toc92290756)

[Hình 1‑2. Các loại Vắc-xin COVID-19 2](#_Toc92290757)

[Hình 2‑1. Big Data 4](#_Toc92290758)

[Hình 2‑2. 5Vs of Big Data 6](#_Toc92290759)

[Hình 2‑3. Data Science và các lĩnh vực liên quan 8](#_Toc92290760)

[Hình 2‑4. Các bước thực hiện Data Mining 9](#_Toc92290761)

[Hình 2‑5. Các loại thuật toán được sử dụng trong Machine Learning 11](#_Toc92290762)

[Hình 2‑6. Lựa chọn thuật toán phù hợp 11](#_Toc92290763)

[Hình 2‑7. Hình ảnh định nghĩa SQL 15](#_Toc92290764)

[Hình 2‑8. Định nghĩa về API 18](#_Toc92290765)

[Hình 2‑9. Quy trình thực thi của hệ thống sử dụng Docker. 20](#_Toc92290766)

[Hình 2‑10. Kiến trúc của Apache Airflow 20](#_Toc92290767)

[Hình 3‑1. Kiến trúc hệ thống 21](#_Toc92290768)

[Hình 3‑2. Sơ đồ khối 22](#_Toc92290769)

[Hình 3‑3. Định dạng JSON của Response 23](#_Toc92290770)

[Hình 3‑4. Dữ liệu trong MySQL 24](#_Toc92290771)

[Hình 3‑5. Tính năng của Power BI 25](#_Toc92290772)

[Hình 3‑6. Dashboard về số ca nhiễm, ca tử vong... 26](#_Toc92290773)

[Hình 3‑7. Dashboard về tỷ lệ tử vong, tỷ lệ mắc bệnh... 27](#_Toc92290774)

[Hình 3‑8. Dashboard về tổng ca nhiễm của các châu lục 29](#_Toc92290775)

[Hình 3‑9. Dashboard về tổng ca tử vong của các châu lục 30](#_Toc92290776)

[Hình 3‑10. Dashboard về tổng ca phục hồi của các châu lục 31](#_Toc92290777)

[Hình 3‑11. Tương quan giữa các khu vực trên Thế giới 32](#_Toc92290778)

[Hình 3‑12. Số ca nhiễm, tử vong và đường Moving Average của Thế giới 32](#_Toc92290779)

[Hình 3‑13. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Thế giới 33](#_Toc92290780)

[Hình 3‑14. Tỉ lệ tử vong theo ngày và trung bình của Thế giới 33](#_Toc92290781)

[Hình 3‑15. Top 10 các Tỉnh có số ca cao nhất trên Thế giới 34](#_Toc92290782)

[Hình 3‑16. Top 10 Bang của Mỹ có số ca cao nhất 34](#_Toc92290783)

[Hình 3‑17. Top 10 nước và phần còn lại về số ca nhiễm 35](#_Toc92290784)

[Hình 3‑18. Top 10 tỉnh và phần còn lại về số ca nhiễm 35](#_Toc92290785)

[Hình 3‑19. Số ca nhiễm, tử vong và đường Moving Average của Châu Á 36](#_Toc92290786)

[Hình 3‑20. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Châu Á 36](#_Toc92290787)

[Hình 3‑21. Số ca nhiễm và tử vong giữa Việt Nam và các nước Châu Á 37](#_Toc92290788)

[Hình 3‑22. Biểu đồ cột giữa Việt Nam và các nước châu Á 37](#_Toc92290789)

[Hình 3‑23. Số ca nhiễm, tử vong và đường Moving Average của Đông Nam Á 38](#_Toc92290790)

[Hình 3‑24. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Đông Nam Á 38](#_Toc92290791)

[Hình 3‑25. Số ca nhiễm và tử vong giữa Việt Nam và các nước Đông Nam Á 39](#_Toc92290792)

[Hình 3‑26. Biểu đồ cột giữa Việt Nam và các nước Đông Nam Á 39](#_Toc92290793)

[Hình 3‑27. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin 40](#_Toc92290794)

[Hình 3‑28. Biểu đồ số lượng Quốc gia sử dụng Vắc-xin đơn 40](#_Toc92290795)

[Hình 3‑29. Biểu đồ top 10 Quốc gia có tổng số Vắc-xin đã tiêm 41](#_Toc92290796)

[Hình 3‑30. Biểu đồ top 10 Quốc gia có số Vắc-xin tiêm mỗi ngày 41](#_Toc92290797)

[Hình 3‑31. Biểu đồ top 10 Quốc gia có số người đã tiêm đầy đủ 41](#_Toc92290798)

[Hình 3‑32. Biểu đồ số lượng Quốc gia sử dụng Vắc-xin kết hợp 42](#_Toc92290799)

[Hình 3‑33. Biểu đồ số lượng Vắc-xin mỗi ngày của Thế giới 42](#_Toc92290800)

[Hình 3‑34. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin ở Châu Á 43](#_Toc92290801)

[Hình 3‑35. Biểu đồ số lượng Vắc-xin mỗi ngày của các nước Châu Á 43](#_Toc92290802)

[Hình 3‑36. Biểu đồ số lượng Vắc-xin cộng dồn của các nước Châu Á 44](#_Toc92290803)

[Hình 3‑37. Biểu đồ phân phối tiêm Vắc-xin mỗi ngày của Việt Nam 44](#_Toc92290804)

[Hình 3‑38. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin ở Đông Nam Á 45](#_Toc92290805)

[Hình 3‑39. Biểu đồ số lượng Vắc-xin mỗi ngày của Đông Nam Á 45](#_Toc92290806)

[Hình 3‑40. Biểu đồ số lượng Vắc-xin cộng dồn của Đông Nam Á 46](#_Toc92290807)

[Hình 3‑41. Biểu đồ Scatter về tổng số lượng Vắc-xin của Đông nam Á 46](#_Toc92290808)

[Hình 4‑1. Siêu phẳng (hyper-plane) trong SVM 47](#_Toc92290809)

[Hình 4‑2. Margin của siêu phẳng trong SVM 48](#_Toc92290810)

[Hình 4‑3. Siêu phẳng có chứa tín hiệu nhiễu 49](#_Toc92290811)

[Hình 4‑4. Non-linear trong SVM 50](#_Toc92290812)

[Hình 4‑5. Mô hình Hồi quy tuyến tính Linear Regression 52](#_Toc92290813)

[Hình 4‑6. R-Square (R2) của Mô hình hồi quy tuyến tính 52](#_Toc92290814)

[Hình 4‑7. Phương pháp Ordinary Least Square 53](#_Toc92290815)

[Hình 4‑8. So sánh các trọng số của mô hình 55](#_Toc92290816)

[Hình 4‑9. Curve Fitting với Hồi quy Bayesian Ridge 55](#_Toc92290817)

[Hình 4‑10. Kết quả dự đoán của thuật toán SVR 56](#_Toc92290818)

[Hình 4‑11. Kết quả dự đoán của thuật toán Polynomial Regression 57](#_Toc92290819)

[Hình 4‑12. Các mô hình Underfitted, well-fitted and overfitted 58](#_Toc92290820)

[Hình 4‑13. Kết quả dự đoán của thuật toán Bayesian Ridge Regression 59](#_Toc92290821)

[Hình 4‑14. Dự đoán số ca nhiễm ở Việt Nam 61](#_Toc92290822)

# DANH MỤC BẢNG BIỂU

Bảng 4‑1. Bảng so sánh kết quả dự đoán 53

Bảng 4‑2. Bảng kết quả dự đoán số ca nhiễm của Việt Nam 55

# DANH MỤC CÁC CHỮ VIẾT TẮT

SQL Broadband Digital Terminal

SMTP Fast Fourier Transform

API Multi-Input Multi-Output

# GIỚI THIỆU ĐỀ TÀI

## Giới thiệu chung

### Giới thiệu về COVID-19

COVID-19 (viết tắt của **co**rona**vi**rus **d**isease 20**19)** là một bệnh đường hô hấp cấp tính truyền nhiễm gây ra bởi chủng virus corona SARS-CoV-2 và các biến thể của nó. Virus là 1 loại mới, được phát hiện điều tra ổ dịch bắt nguồn từ khu chợ hải sản và động vật ở Vũ Hán (Wuhan), tỉnh Hồ Bắc (Hubei), Trung Quốc vào tháng 12/2019. Virus gây viêm đường hô hấp cấp ở người và lây lan từ người sang người.



Hình 1‑1. Hình ảnh của COVID-19

Phương thức lây truyền chính là từ người sang người bằng 3 cách: truyền trực tiếp (qua các giọt dịch hô hấp khi tiếp xúc gần), truyền tiếp xúc (mầm bệnh bám trên 1 vật thể rồi người khác chạm vào đưa vào hệ hô hấp như miệng, mũi), truyền khí dung (mầm bệnh bay lửng lơ trong không khí, người khác hít phải hoặc tiếp xúc qua niêm mạc). Do cách thức lây nhiễm qua đường hô hấp khiến COVID-19 là căn bệnh rất dễ lây lan trong những không gian hẹp, kín, những chỗ đông người và tiếp xúc gần.

Để phòng chống và hạn chế việc lây lan của dịch COVID-19, WHO (World Health Organization: tổ chức sức khỏe Thế giới) và Bộ Y tế gửi đến thông điệp 5K (Khẩu trang, Khử khuẩn, Khoảng cách, Không tụ tập và Khai báo Y tế) trong trạng thái “bình thường mới”.

### Giới thiệu về Vắc-xin COVID-19

Vắc-xin COVID-19 là Vắc-xin truyền tin RNA (mRNA) dạy các tế bào cách tạo protein giúp kích hoạt phản ứng miễn dịch bên trong cơ thể của chúng ta, mang lại lại lợi ích cho người được tiêm chủng bằng các cung cấp khả năng bảo vệ các loại virus mà không phải chịu những hậu quả nghiêm trọng có thể xảy ra khi mắc bệnh.

Để phòng chống, tiếp cận an toàn và hiệu quả COVID-19 là yếu tố quan trọng để chấm dứt đại dịch COVID-19, hiện tại đã có nhiều Vắc-xin đã được chứng minh và đang phát triển, WHO cùng với các đối tác đang làm việc để phát triển, sản xuất và triển khai Vắc-xin an toàn và hiệu quả.

Hiện nay, đã có nhiều loại Vắc-xin được WHO phê duyệt và chứng nhận an toàn để cấp phép sử dụng như AstraZeneca, Moderna, Pfizer/BioNTech, Sputnik V, Sinopharm, Janssen, SinoVac… Vào đầu tháng 1/2020, ở nhiều quốc gia trên Thế giới đã bắt đầu triển khai việc tiêm Vắc-xin để phòng chống COVID-19, tiêu biểu như Anh, Pháp, Mỹ… Và cho đến nay, đã có nhiều quốc gia tham gia tiêm ngừa Vắc-xin, trong đó có Việt Nam để sống chung với dịch và đưa Thế giới trở lại trạng thái “bình thường mới”.



Hình 1‑2. Các loại Vắc-xin COVID-19

## Mục đích thực hiện đề tài

Trong bối cảnh tình hình dịch COVID-19 đang diễn biến rất phức tạp cùng với nhiều biến thể và đặc biệt là biến thể Omicron đã gây ra không ít sự khó khăn, mất mát và ảnh hưởng tiêu cực đến kinh tế cũng như gây nên sự rối loạn trong cuộc sống của Thế giới nói chung và Việt Nam nói riêng.

Bên cạnh đó cùng với sự phát triển mạnh về Khoa Học Công Nghệ, nhiều quốc gia trên Thế giới như Anh, Mỹ … đã nghiên cứu và sản xuất nhiều loại Vắc-xin nhằm ngăn chặn sự lây lan nhanh của các biến thể COVID và đưa Thế giới trở lại “Bình thường mới”.

Mục đích thực hiện đề tài là đem đến một cái nhìn tổng quát và chi tiết về tình hình dịch COVID-19 cùng với tình hình tiêm chủng Vắc-xin ngừa COVID-19 ở trên Thế giới, bên cạnh đó là Việt Nam so với khu vực Châu Á và Đông Nam Á.

## Yêu cầu của đề tài

Yêu cầu đề tài là tìm hiểu về Dữ liệu lớn (Big Data), tìm hiểu về Data Mining, sử dụng phần mềm Power BI để vẽ biểu đồ để thể hiện được cái nhìn tổng quát về dịch COVID-19, sử dụng ngôn ngữ lập trình Python cùng các thư viện đi kèm để thực hiện phân tích và vẽ biểu đồ để có cái nhìn chi tiết hơn về dịch COVID-19 và tình hình tiêm ngừa Vắc-xin của các quốc gia trên Thế giới nói chung và Việt Nam nói riêng, lưu trữ dữ liệu bằng Cơ sở dữ liệu MySQL và sử dụng thuật toán Machine Learning để dự đoán số ca nhiễm của Thế giới và Việt Nam trong thời gian gần.

## Phương pháp thực hiện

Phương pháp thực hiện của em là:

* Sử dụng ngôn ngữ lập trình Python để trích xuất dữ liệu từ API và nguồn dữ liệu từ các trang web uy tín, tin cậy như Kaggle.
* Sử dụng MySQL là Cơ sở dữ liệu để lưu trữ dữ liệu.
* Sử dụng các thư viện trong Python và phần mềm Power BI để vẽ các biểu đồ về tình hình dịch COVID-19 và Vắc-xin.

# LÝ THUYẾT ĐỀ TÀI

## Giới thiệu về Big Data và Khoa học Dữ liệu

### Big Data

Big Data hay còn được gọi là Dữ liệu lớn dùng để mô tả khối lượng dữ liệu lớn, bao gồm có cấu và không có cấu trúc. Dữ liệu này cung cấp thông tin cho doanh nghiệp, tổ chức theo cơ sở hằng ngày, hằng giờ, thậm chí hằng phút.

Big Data có thể dùng để phân tích để tìm ra những insight trong kinh doanh dẫn đến các quyết định tốt hơn và các chiến lược kinh doanh, sử dụng trong ngành Y tế để dự đoán, theo dõi tình hình sức khỏe của bệnh nhân.

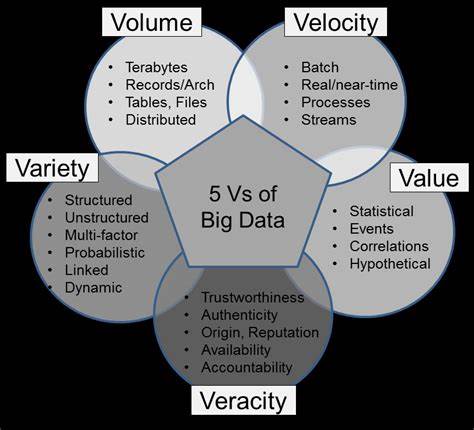
Tuy nhiên, do những bộ dữ liệu này quá lớn đến nỗi các phần mềm xử lý dữ liệu truyền thống không quản lý được, nhưng những khối dữ liệu khổng lồ này có thể giải quyết được rất nhiều vấn đề về kinh doanh hay sức khỏe mà trước đây không thể giải quyết được.



Hình 2‑1. Big Data

Big Data có 5 đặc trưng cơ bản (5V) gồm:

* Volume: là sự tăng trưởng về mặt khối lượng, Big Data không ngừng phát triển về mặt kích thước (khối lượng) khi mà nhu cầu thu thập dữ liệu ngày càng lớn dẫn đến khối lượng sẽ phát triển ngày càng khổng lồ theo thời gian. Dữ liệu có thể thu thập dưới định dạng hình ảnh, âm thành, video … trên các kênh truyền thông, mạng xã hội, … Khối lượng dữ liệu của 1 hệ thống thông tin có thể lên đến hàng Terabyte và Petabyte.
* Velocity: là sự tăng trưởng về mặt tốc độ. Bên cạnh sự tăng trưởng về mặt khối lượng thì tốc độ tăng trưởng của dữ liệu là rất nhanh. Ở các nền tảng mạng xã hội người ta thu thập thông tin qua các bài post, qua các thông báo chỉ trong vài giây. Sự chuyển động của dữ liệu hiện tại hầu như là thực tế (real time) và tốc độ cập nhật thông tin chỉ còn đơn vị hàng Mili giây.
* Variety: là sự tăng trưởng về tính đa dạng của dữ liệu. Ngày nay dữ liệu không còn đơn giản là dạng số hay chữ như truyền thông nữa, mà đã phát triển thành gồm cả dữ liệu cấu trúc và phi cấu trúc như video, hình ảnh, âm thanh hay là từ các cảm biến, từ các file log của các thiết bị thu. Dữ liệu của 1 doanh nghiệp hay hệ thống thông tin không còn đơn giản chỉ có 1 hay một vài dữ liệu mà tính đa dạng cũng ngày càng tăng lên theo thời gian làm cho tính phức tạp của dữ liệu ngày càng phức tạp hơn.
* Veracity: là tính xác thực của dữ liệu, cùng với sự tăng trưởng nhanh về mặt khối lượng, tăng trưởng và đa dạng của dữ liệu, thì đảm bảo tính xác thực của dữ liệu là rất quan trọng. Cũng với sự phát triển của xu hướng thời đại 4.0 như hiện nay, xu hướng mạng xã hội và người dùng di động làm cho việc xác định độ tin cậy và xác thực của dữ liệu ngày một khó khăn hơn. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu là rất quan trọng của Big Data.
* Value: cuối cùng nhưng là quan trọng nhất, giá trị thông tin mà Big Data mang lại là tính chất quan trọng nhất của xu hướng công nghệ Big Data. Ví dụ về doanh nghiệp phải hoạch định được những giá trị thông tin của Big Data cho vấn đề, bài toán hoặc mô hình hoạt động kinh doanh của mình.



Hình 2‑2. 5Vs of Big Data

Ứng dụng của Big Data: Big Data mang lại nhiều giá trị cho rất nhiều ngành khác nhau, tiêu biểu như Ngân Hàng, Y Tế, Thương Mại Điện Tử, Digital Marketing, ngăn chặn nội dung xấu/đen…

* Ngành Ngân Hàng: sử dụng kỹ thuật phân cụm để tìm ra khu vực có nhu cầu giao dịch để thành lập chi nhánh mới, Big Data là nền tảng của ngân hàng kỹ thuật số, áp dụng Machine Learning và AI để phát hiện các hoạt động gian lận, hỗ trợ xử lý, báo cáo và phân tích lượng dữ liệu khổng lồ từ các hoạt động giao dịch hằng ngày nhằm đảm bảo an ninh cho ngân hàng …
* Ngành Y Tế: Theo dõi tình trạng bệnh nhân, đánh giá các triệu chứng và xác định nhiều bệnh thời gian đầu, lưu trữ hồ sơ bệnh lý, thông tin nhạy cảm được bảo mật, dự đoán trước về dấu hiệu phát dịch, bùng dịch …
* Ngành Thương Mại Điện Tử: Thu thập dữ liệu về nhu cầu mua hàng của Khách hàng, tạo mô hình tiếp thị hiệu suất cao, quản lý các sản phẩm được xem và mua nhiều nhất để đưa ra các chương trình khuyến mãi, lập báo cáo theo nhiều tiêu chí, phân tích hành vi mua hàng, xu hướng mua hàng để cung cấp dịch vụ tốt nhất …
* Ngành Retail (Bán lẻ): Giúp nhà quản lý xây dựng mô hình chi tiêu của từng Khách hàng, nhờ sự trợ giúp của Big Data ngành công nghiệp có thể so sánh tỷ lệ cung – cầu đề sản xuất tránh việc tung ra sản phẩm không được sử dụng, bố trí sản phẩm trưng bày dựa vào thói quen mua hàng và nhu cầu của Khách hàng…
* Ngành Digital Marketing: Phân tích thị trường, đối thủ cạnh tranh và đánh giá mục tiêu kinh doanh, tạo báo cáo sau chiến dịch quảng cáo để tăng hiệu suất, tập trung vào chủ đề tìm kiếm nâng cao và tư vấn cho chủ doanh nghiệp để thực hiện chiến lược xếp hạng trên Google (SEO)
* Ngăn chặn nội dung xấu/đen: Các Extension của các nền tảng trình duyệt như Chrome, Edge, Firefox … có nhiều Addin phục vụ cho việc lọc nội dung (content filtering) có phù hợp hay không, ngăn chặn các tổ chức xấu lợi dụng để trục lợi.

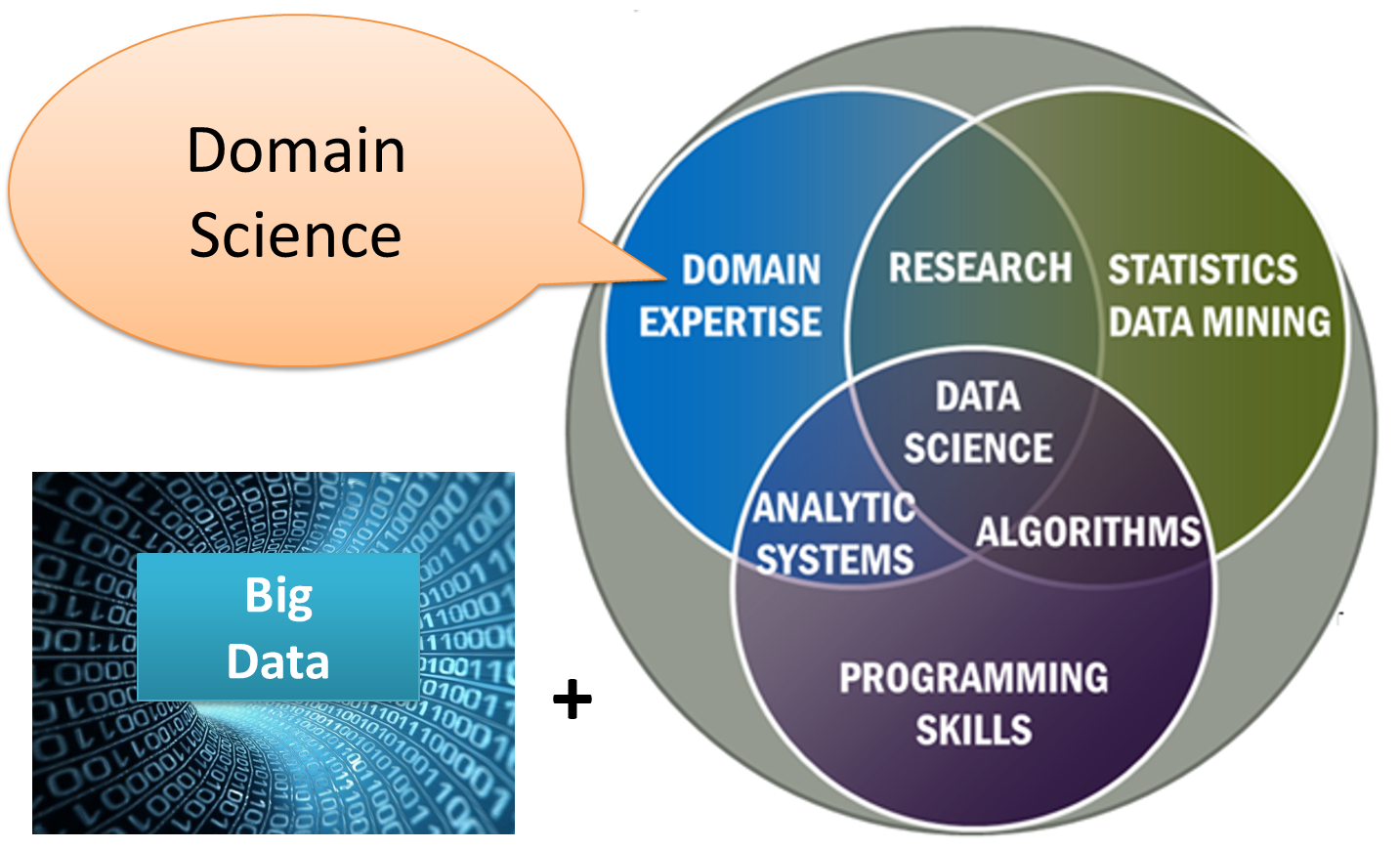
### Khoa Học Dữ Liệu (Data Science)

Khoa Học Dữ Liệu là khoa học về việc quản trị và phân tích dữ liệu, trích xuất các giá trị từ dữ liệu để tìm ra các insight, các tri thức hành động, các quyết định.

Khoa học dữ liệu gồm 3 phần chính: Tạo ra và quản trị, phân tích dữ liệu và chuyển kết quả phân tích thành giá trị hành động. Việc phân tích lại dựa vào 3 nguồn tri thức: Toán học (Thống kê), Công nghệ Thông tin (Máy học) và tri thức của lĩnh vực ứng dụng.

Các lĩnh vực của Khoa học dữ liệu: Khai thác dữ liệu (Data mining), Thống kê (Statistic), Học máy (Machine Learning), Phân tích (Analyzing) và Lập trình (Programming).

Khoa học dữ liệu yêu cầu thực hiện các quan sát, đặt câu hỏi, hình thành giả thuyết, phân tích kết quả và đưa ra một khuyến nghị. Mục đích chính là biến đổi một lượng lớn dữ liệu chưa qua xử lý, làm thế nào để định vị được thành mô hình kinh doanh, ừ đó giúp giảm chi phí, gia tăng hiệu quả làm việc, nhìn nhận cơ hội, rủi ro trên thị trường và làm gia tăng lợi thế cạnh tranh của doanh nghiệp.



Hình 2‑3. Data Science và các lĩnh vực liên quan

## Giới thiệu về Data Mining và Machine Learning

### Data Mining

Data Mining – khai phá dữ liệu là quá trình phân loại, trích xuất, sắp xếp tập dữ liệu lớn để xác định mẫu, thiết lập các mối liên hệ để đưa ra hướng giải quyết vấn đề nhờ phân tích dữ liệu, sử dụng những dữ liệu có giá trị tiềm ẩn từ bên trong lượng lớn dữ liệu được lưu trữ trong Cơ sở dữ liệu.

Quá trình khai phá dữ liệu là quá trình phức tạp bao gồm kho dữ liệu lớn, chuyên sâu cùng với các công nghệ tính toán, Data Mining không chỉ giới hạn trong việc trích xuất mà còn được dùng để làm sạch, phân tích mẫu, làm sạch và chuyển đổi. Trong Data Mining có nhiều tham số quan trọng khác nhau như phân cụm, phân loại, dự đoán và quy tắc kết hợp.

Để thực hiện Data Mining bao gồm 7 bước:

* Bước 1: Làm sạch dữ liệu, làm sạch để dữ liệu không có dư thừa và bất thường.
* Bước 2: Tích hợp dữ liệu, kết hợp nhiều nguồn dữ liệu lại làm một để dễ quản lý và sử dụng.
* Bước 3: Lựa chọn dữ liệu, trích xuất dữ liệu từ Cơ sở dữ liệu đa chiều.
* Bước 4: Chuyển đổi dữ liệu, chuyển đổi dữ liệu để phân tích tóm tắt và cho các hoạt động tổng hợp.
* Bước 5: Khai phá dữ liệu, trích xuất dữ liệu từ dữ liệu hiện có.
* Bước 6: Đánh giá mẫu, phân tích mẫu trong dữ liệu.
* Bước 7: Trình bày thông tin, dữ liệu sẽ được thể hiện. trình bày dưới dạng cây, biểu đồ, bảng và ma trận …



Hình 2‑4. Các bước thực hiện Data Mining

Data Mining có rất nhiều ứng dụng để phục vụ, đáp ứng nhu cầu cho nhiều lĩnh vực:

* Phân tích và quản lý rủi ro trong doanh nghiệp.
* Phát hiện gian lận.
* Phát hiện bệnh cho bệnh nhân ở nhiều giai đoạn.
* Phân tích thị trường và chứng khoán.

### Machine Learning

Machine Learning là lĩnh vực con của Trí tuệ nhân tạo (Artificial Intelligence) hay còn được viết tắt là AI, sử dụng thuật toán để máy tính có thể học từ dữ liệu để thực hiện công việc thay vì được lập trình để thực hiện công việc một cách rõ ràng.

Trên thực tế Machine Learning có rất nhiều ứng dụng đã mang lại rất nhiều lợi ích cho các doanh nghiệp, tổ chức, cho ngành Y Tế … Dưới đây là một số lĩnh vực phổ biến mà Machine Learning góp mặt:

* Xử lý ảnh (Image Processing): Gắn thẻ hình ảnh (Image Tagging), Nhận dạng ký tự (Optical Character Recognition), Xe tự hành (Self-driving vehicle) …
* Phân tích văn bản (Text Analysis): Lọc spam (Spam filtering), Phân tích ngữ nghĩa (Sentiment Analysis), Khai thác thông tin(Information Extraction) …
* Khai phá dữ liệu (Data Mining): Phát hiện bất thường (Anomaly Detection), Phát hiện các quy luật (Association rules), Gom nhóm(Grouping), Dự đoán(Predictions) …
* Trò chơi điện tử và Robot

Trong Machine Learning có 2 loại thuật toán cơ bản: Học có giám sát (Supervised learning) và Học không có giám sát (Unsupervised learning)

* Học có giám sát: Trong thuật toán Học có giám sát được chia thành là Phân loại (Classification) và Hồi quy (Regression). Các thuật toán thường được sử dụng là Support Vector Machines (SVMs), Linear Regression, Logistic Regression, k-Nearest Neighbors (kNN), Neural Network…
* Học không có giám sát: Trong thuật toán Học không có giám sát được được chia thành là Phân cụm (Clustering), Biểu diễn và giảm số chiều (Representation and dimensional reduction). Các thuật toán thường được sử dụng là K-mean, Spectral Clustering, Hierarchical clustering- Apriori (Association Rule Mining) …



Hình 2‑5. Các loại thuật toán được sử dụng trong Machine Learning



Hình 2‑6. Lựa chọn thuật toán phù hợp

## Giới thiệu Python và các thư viện Python sử dụng trong đề tài

### Python

Python là ngôn ngữ lập trình bậc cao được tạo ra bởi Guido van Rossum và ra mắt lần đầu vào năm 1991, là ngôn ngữ thông dịch (Interpreter), mã nguồn mở (Open-source), dễ học dễ đọc, chạy đa nền tảng. Python có cấu trúc gọn gàng, hỗ trợ nhiều mâu đa lập trình như: lập trình hàm (functional programming), lập trình hướng đối tượng (object-oriented programming) …

* Ưu điểm của Python:
* Là ngôn ngữ dễ học cho người mới, và là được nhiều lập trình viên sử dụng.
* Cú pháp đơn giản dễ làm quen.
* Có khả năng mở rộng, thích hợp cho các chương trình nhỏ và dự án lớn.
* Sử dụng được trên nhiều Hệ điều hành: Windows, UNIX, Mac OS, Linux.
* Nhược điểm của Python:
* Python không có một số thuộc tính như: public, private, protected …
* Mặc dù tốc độ xử lý nhanh, tuy nhiên vẫn chậm hơn Java, C++ …

Cùng với Ưu và Nhược điểm, Python có nhiều tính năng nổi bật được nhiều lập trình viên sử dụng:

* Phát triển phần mềm (Software Development)
* Phát triển Web phía máy chủ (Backend Development)
* Python được sử dụng để xử lý dữ liệu lớn, cùng với lượng thư viện hỗ trợ như Pandas, PySpark, Numpy …
* Python cũng được sử dụng cho lĩnh vực Trí tuệ nhân tạo, Machine Learning, Deep Learning cùng với nhiều thư viện hỗ trợ như Tensorflow, Scikit-learn, Keras, PyTorch …

### Các thư viện của Python

Python được biết đến là ngôn ngữ mã nguồn mở, có nhiều thư viện hỗ trợ nhằm giúp lập trình viên dễ dàng hơn trong việc lập trình. Trong đề tài của em, em có sử dụng một số thư viện để hỗ trợ:

* **Pandas**: công cụ dùng để phân tích và thao tác dữ liệu mã nguồn mở nhanh, mạnh, dễ sử dụng và linh hoạt được xây dựng trên ngôn ngữ lập trình Python, cung cấp rất nhiều tính năng hữu ích cho người dùng, đặc biệt là trong ngành Khoa học Dữ liệu.
* **Numpy (Numeric Python)**: là thư viện toán học mạnh mẽ và phổ biến của Python. Numpy cho phép người sử dụng làm việc, thao tác hiệu quả với ma trận 1 chiều, 2 chiều hay đa chiều (multidimensional) và mảng với tốc độ xử lý nhanh gấp nhiều lần so với sử dụng “core Python” đơn thuần. Numpy hoạt động nhanh trên mảng, các phép toán học, logic, sắp xếp, các phép biến đổi Fourier, đại số tuyến tính, mô phỏng ngẫu nhiên ….
* **Matplotlib**: trực quan hóa là một bước quan trọng trong lĩnh vực Khoa học dữ liệu và Matplotlib là giải pháp hiệu quả cho người dùng Python. Matplotlib là thư viện vẽ đồ thị, biểu đồ, hình ảnh trực quan tĩnh và động rất mạnh mẽ cho người làm việc với Python và Numpy.
* **Seaborn**: bên cạnh Matplotlib thì Python còn có Seaborn cũng là thư viện dùng để trực quan hóa dữ liệu, vẽ đồ thị, biểu đồ. Seaborn được xây dựng dựa trên Matplotlib và tích hợp chặt chẽ với cấu trúc dữ liệu của Pandas. Seaborn cung cấp cho người dùng Python nhiều function để vẽ nhiều dạng biểu đồ, thống kê đẹp hơn so với Matplotlib. Các chức năng vẽ biểu đồ của Seaborn hoạt động trên dataframe và mảng chứa tập dữ liệu.
* **Requests**: là thư viện HTTP đơn giản của Python, dùng để gửi lệnh HTTP qua các server, dịch vụ Web API. Thư viện Requests cung cấp cho người dùng nhiều gửi các loại lệnh như GET (để lấy dữ liệu), POST (để thêm dữ liệu), PUT (để cập nhật dữ liệu), DELETE (để xóa dữ liệu) …
* **Scikit-learn**: là thư viện mạnh mẽ dành cho các thuật toán Machine Learning được biết bằng ngôn ngữ Python. Scikit-learn cung cấp các công cụ để xử lý các bài toán Machine Learning và mô hình thống kê (Statistical Modeling) như: classification, clustering, regression, và dimensionality reduction. Scikit-learn là một công cụ đơn giản và hiệu quả cho phân tích dự đoán dữ liệu, xây dựng trên tư viện Numpy, SciPy và Matplotlib.

## Giới thiệu SQL và Cơ sở dữ liệu MySQL

### SQL

SQL là viết tắt của Structured Query Language, nghĩa là ngôn ngữ truy vấn dữ liệu mang tính cấu trúc theo tiêu chuẩn ANSI/ISO. SQL là ngôn ngữ để giao tiếp với bất cứ hệ thống cơ sở dữ liệu quan hệ RDBMS (Relational Database Management System) như Oracle, MySQL, PostgreSQL, Microsoft SQL Server …SQL được sử dụng rất nhiều ở các công ty vừa và lớn để xây dựng hệ thống lưu trữ cơ sở dữ liệu.

Cơ sở dữ liệu này được quy thành nhiều bảng, mỗi bảng gồm nhiều cột và dòng, các bảng này có mối liên kết chặt chẽ với nhau do người dùng định nghĩa. SQL giúp truy vấn thông tin nhanh và quản lý hiệu quả hơn, giúp bào trì, bảo mật thông tin dễ dàng hơn.

Không giống như các dạng lưu trữ dữ liệu truyền thống như tài liệu giấy hay Excel sẽ có nhiều rủi ro như bị mất, chỉnh sửa như xóa, sửa … và dữ liệu không được bảo mật cao, dễ bị rò rỉ thông tin bảo mật ra bên ngoài. Còn với SQL, sau khi lưu thông tin vào cơ sở dữ liệu, người dùng chỉ cần sử dụng câu query (truy vấn) SQL ngắn để tương tác với các bảng là đã trích xuất được thông tin cần thiết, việc thêm, sửa, xóa cũng dễ dàng hơn rất nhiều.



Hình 2‑7. Hình ảnh định nghĩa SQL

SQL có nhiều Ưu điểm giúp cho người dùng dễ dàng thao tác với cơ sở dữ liệu:

* Xác định dữ liệu trong cơ sở dữ liệu và thao tác dữ liệu đó.
* Truy cập dữ liệu trong các hệ thống quản lý cơ sở dữ liệu quan hệ.
* Tạo chế độ view, thủ tục lưu trữ, chức năng trong cơ sở dữ liệu.
* Nhúng trong các ngôn ngữ khác sử dụng mô-đun SQL, thư viện và trình biên dịch trước.
* Thiết lập quyền trên các bảng, thủ tục và view.

Các lệnh SQL cơ bản được chia thành 3 loại: DDL, DML và DCL

* DDL (Data Definition Language): Ngôn ngữ định nghĩa dữ liệu.
* DML (Data Manipulation Language): Ngôn ngữ thao tác dữ liệu.
* DCL (Data Control Language): Ngôn ngữ điều khiển dữ liệu.

### Cơ sở dữ liệu MySQL

MySQL là Hệ quản trị cơ sở dữ liệu mà nguồn miễn phí (free-source) phở biến thế giới và được nhiều nhà phát triển ưa chuộng sử dụng. MySQL là Hệ quản trị cơ sở dữ liệu sử dụng SQL, có tốc độ cao, ổn định, dễ dàng sử dụng, hoạt động được trên nhiều Hệ điều hành như Window, Ubuntu, Linux, MacOS, có tính khả chuyển. Hệ quản trị cơ sở dữ liệu này hoạt động dựa trên mô hình tiêu chuẩn Client (Máy khách) – Server (Máy chủ).

Nhờ tốc độ và tính bảo mật cao, MySQL thích hợp cho các ứng dụng có sử dụng truy cập Cơ sở dữ liệu trên Internet. Hiện nay, nhiều Website lớn như Facebook, Google hay Twitter … đều đang sử dụng MySQL để hỗ trợ lưu trữ dữ liệu.

Ưu và nhược điểm của MySQL:

* Ưu điểm:
* Dễ dàng sử dụng: Tốc độ xử lý cao và ổn định, hoạt động trên nhiều Hệ điều hành cung cấp nhiều hàm tiện ích.
* Tính bảo mật cao: MySQL phù hợp với các ứng dụng có truy cập Internet nhờ sở hữu nhiều tính năng bảo mật cấp cao.
* Cho phép khôi phục: MySQL cho phép các giao dịch (transaction) khôi phục, cam kết (commit) và phục hồi sự cố
* Khả năng mở rộng mạnh mẽ: MySQL có khả năng xử lý với lượng dữ liệu lớn và có thể mở rộng.
* Nhược điểm:
* MySQL bị hạn chế dung lượng, tốc độ truy vấn sẽ giảm dần nếu lượng dữ liệu truy vấn lớn, để cải thiện cần áp dụng nhiều biện pháp như chia sẻ database ra nhiều server hoặc tạo cache MySQL.

## Giới thiệu Jupyter Notebook, API và Kaggle

### Jupyter Notebook

Jupyter Notebook là một ứng dụng Web mã nguồn mở cho phép người dùng tạo và chia sẻ code, phương trình, hình ảnh hóa (visualization). Ứng dụng của Jupyter Notebook: làm sạch, chuyển đổi dữ liệu, lập mô hình thống kê, trực quan hóa dữ liệu mô phỏng … Jupyter Notebook cho phép người dùng thực thi code theo từng ô lệnh (cell) mà không cần phải chạy hoàn toàn 1 file như những ứng dụng viết code khác.

### API (Application Programming Interface)

API được gọi là Giao diện Lập trình Ứng dụng, là trung gian phần mềm cho phép 2 ứng dụng giao tiếp với nhau. API có thể sử dụng cho hệ điều hành (Operating System), hệ thống cơ sở dữ liệu (Database System), phần cứng máy tính, hệ thống trên nền tảng web (Web-based System) … API là một tập hợp các quy trình, giao thức và công cụ để xây dựng ứng dụng phần mềm. API có tính bảo mật, an toàn cao.

Hiện nay có rất nhiều API được ra đời nhằm giúp cho các lập trình viên như Google Maps API, Youtube API, Twitter API, Amazon Product Advertising API …

Ưu điểm của API:

* Giao tiếp 2 chiều phải được xác nhận.
* Kết nối mọi lúc nhờ Internet.
* Hỗ trợ chức năng RESTful.
* Hỗ trợ đầy đủ thành phần MVC: controller, filter, IoC container, routing…
* Mã nguồn mở, giao tiếp API là 2 chiều nên thông tin đáng tin cậy.

Nhược điểm của API:

* Đòi hỏi kiến thức chuyên sâu.
* Tốn nhiều chi phí vận hành, nâng cấp, chỉnh sửa và phát triển.
* Gặp vấn đề về bảo mật nếu bị tấn công hệ thống.



Hình 2‑8. Định nghĩa về API

### Kaggle

Kaggle là một công ty con dưới sự điều hành của Google LLC – là một tập đoàn công nghệ đa quốc gia được thành lập tại Mỹ, là doanh nghiệp lớn chuyên về cung cấp các dịch vụ và sản phẩm liên quan đến mạng Internet như dịch vụ công nghệ thông tin, quảng bá trực tuyến, phần mềm, phần cứng cũng như nổi bất nhật là công cụ tìm kiếm thông tin.

Kaggle là một cộng đồng trực tuyến, là nơi những nhà Khoa học dữ liệu cùng với mọi đối tượng có thể sử dụng để thực hành về Machine Learning, Deep Learning và tương tác với nhau như một cộng đồng. Kaggle cung cấp rất nhiều tập dữ liệu (được gọi là Dataset) về mọi lĩnh vực như thể thao, y tế, tài chính … và một nền tảng (platform) trực tuyến để mọi người có thể chia sẻ kiến thức, cũng như code của bản thân. Kaggle còn cung cấp nơi để mọi người có thể luyện thêm kỹ năng bằng tham gia thi đấu các cuộc thi do Kaggle tổ chức.

## Giới thiệu về Docker và Apache Airflow

### Docker

Docker là nền tảng mã nguồn mở, là phần mềm cho phép lập trình viên triển khai, dựng và kiểm thử ứng dụng một cách nhanh chóng. Docker đóng gói phần mềm vào các đơn vị tiêu chuẩn hóa được gọi là container có mọi thứ mà phần mềm cần để chạy, trong đó bao gồm có thư viện, mã, công cụ hệ thống và thời gian chạy. Docker giúp lập trình viên nhanh chóng triển khai và thay đổi quy mô ứng dụng vào bất kỳ môi trường nào và đảm bảo rằng mã của bạn sẽ chạy được bằng cách sử dụng Docker.

* Cách thức hoạt động của Docker: Docker hoạt động bằng cách cung cấp phương thức tiêu chuẩn để chạy mã, là hệ điều hành của container (trên nền tảng ảo hóa). Docker được cài đặt trên từng máy chủ và cung cấp các lệnh đơn giản dễ dàng sử dụng để xây dựng, khởi động hoặc dừng container.
* Lý do nên sử dụng Docker: cho phép vận chuyển mã nhanh hơn, tiêu chuẩn hóa các hoạt động ứng dụng, tiết kiệm tiền bằng cách cải thiện khả năng vận dụng tài nguyên.
* Những thành phần chính của Docker: Docker Engine (thành phần chính của Docker, là công cụ để đóng gói ứng dụng), Docker Hub (chứa hàng ngàn public images có thể sử dụng dễ dàng từ cộng đồng), Images (là khuôn mẫu để tạo container), Container (là 1 instance của 1 image, có thể thao tác dựa trên Docker API hoặc Docker CLI), Docker Client (giúp người dùng giao tiếp với Docker host), Docker Daemon (lắng nghe yêu cầu từ Docker Client để quản lý Container, Image, Network và Volumes thông qua REST API), Dockerfiles (tập tin để chỉ dẫn để build một image)và Volumes (phần dữ liệu được tạo ra khi container được khởi tạo).
* Khi nào nên sử dụng Docker: Khi build ứng dụng cần scale một cách linh hoạt, thiết lập máy local và server trên cùng một môi trường nhanh chóng, triển khai kiến trúc Microservices …



Hình 2‑9. Quy trình thực thi của hệ thống sử dụng Docker.

### Apache Airflow

Apache Airflow là phần mềm mã nguồn mở, được sử dụng để sắp xếp các luồng làm việc dưới dạng chu kỳ theo hướng DAG (Directed Acyclic Graph) của những tác vụ theo hướng lập trình. Một lịch trình Airflow được sử dụng để lên lịch những quy trình làm việc và đường ống dữ liệu (data pipeline). Giao diện Airflow cho phép người dùng quản lý, theo dõi những đường ống chạy, giám sát tiến trình và xử lý sự cố khi cần thiết.

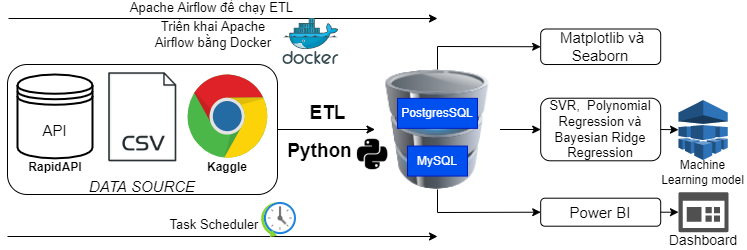


Hình 2‑10. Kiến trúc của Apache Airflow

# THU THẬP, PHÂN TÍCH VÀ BIỂU ĐỒ

## Thiết kế hệ thống

### Kiến trúc hệ thống

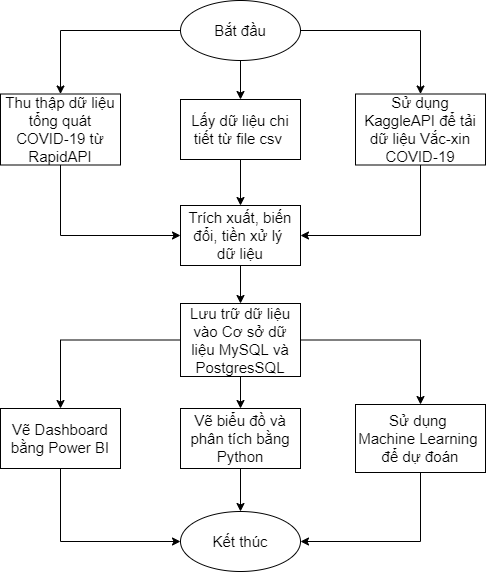


Hình 3‑1. Kiến trúc hệ thống

Hình 3-1 là kiến trúc hệ thống của Đồ án để giới thiệu những công nghệ, thuật toán Machine Learning, phần mềm được sử dụng trong Đồ án và quy trình công việc:

* Sử dụng Task Scheduler để lập lịch tự động và triển khai Apache Aiflow bằng Docker để chạy tác vụ ETL (Extract-Transfrom-Load)
* Thu thập dữ liệu từ các trang web được cung cấp bởi bên thứ 3 (third-party) uy tín, chính xác và được xác nhận.
* Thu thập dữ liệu thông qua API bằng phương thức HTTP.
* Thu thập dữ liệu từ nền tảng đa người dùng Kaggle.
* Sử dụng ngôn ngữ lập trình Python để trích xuất, tiền xử lý (preprocessing)
* Lưu trữ dữ liệu vào Cơ sở dữ liệu MySQL và PostgresSQL bằng Python.
* Sử dụng thư viện Matplotlib và Seaborn để phân tích và vẽ biểu đồ.
* Sử dụng thuật toán SVM, Polynomial Regression và Bayesian Ridge Regression để dự đoán.
* Sử dụng phần mềm Power BI để vẽ Dashboard báo cáo.

### Sơ đồ khối



Hình 3‑2. Sơ đồ khối

Hình 3-2 là sơ đồ khối mô tả chi tiết quy trình của Đồ án: Thu thập dữ liệu tổng quát về dịch COVID-19 từ trang web RapidAPI, thu thập dữ liệu từ nền tảng trực tuyến Kaggle về dữ liệu Vắc-xin và thu thập dữ liệu chi tiết về dịch COVID-19 từ Trường Đại học Johns Hopkins. Trích xuất, tiền xử lý, biến đổi dữ liệu bằng ngôn ngữ lập trình Python và lưu trữ dữ liệu bằng Cơ sở dữ liệu MySQL. Sử dụng phần mềm Power BI để vẽ các Dashboard, sử dụng thư viện Matplotlib và Seaborn để phân tích và vẽ biểu đồ và sử dụng Machine Learning để dự đoán số ca nhiễm trong tương lai gần của Thế giới và Việt Nam.

## Tổng quát về dịch COVID-19

Để có được cái nhìn tổng quát về dịch COVID-19, trong chương này em sử dụng Python để thu thập dữ liệu, lưu trữ dữ liệu vào Cơ sở dữ liệu MySQL và vẽ Dashboard bằng phần mềm Power BI.

### Thu thập dữ liệu từ RapidAPI và lưu trữ vào Cơ sở dữ liệu MySQL

Để thu thập được dữ liệu về dịch COVID-19 như tổng số ca nhiễm, tổng số ca tử vong, tổng số người khỏi bệnh, tỷ lệ tử vong, tỷ lệ mắc bệnh … trong đồ án này em sử dụng nền tảng cung cấp API là RapidAPI.

RapidAPI là được phát triển từ cuộc thi Hackathon năm 2015, được tạo ra bởi các nhà phát triển dành cho các nhà phát triển để họ có một nơi để truy cập vào các API và Microservices cũng như xây dựng các ứng dụng, phần mềm hiệu quả và dễ dàng hơn. RapidAPI là nơi các nhà phát triển có tể tìm, thử nghiệm và kết nối với các API – tất cả chỉ cần 1 tài khoản, 1 API key và 1 SDK(Software Development Kit). Các nhà phát triển có thể chia sẻ và cộng tác làm việc, RapidAPI cho phép các nhà phát triển chia sẻ và đăng ký API công khai. RapidAPI cung cấp rất nhiều API như dữ liệu về Thể thao, dữ liệu về Y tế, dữ liệu về Khoa học …

Trong Đồ án này, em sử dụng “VACCOVID - coronavirus, Vắc-xin and treatment tracker” của RapidAPI, là 1 API miễn phí và đã được xác thực thông tin bởi RapidAPI để phân tích và vẽ Dashboard.

Để trích xuất được dữ liệu từ API, em sử dụng ngôn ngữ lập trình Python cùng với thư viện Requests. Để sử dụng thư viện Requests, em cần cung cấp 2 thông số (parameter) là URL, phương thức (method) “GET” và đưa phản hồi (response) trả về thành định dạng lưu trữ JSON(JavaScript Object Notation).



Hình 3‑3. Định dạng JSON của Response

Sau khi có được đầy đủ dữ liệu về COVID-19 của các quốc gia trên thế giới, em sử dụng Python để giao tiếp với Cơ sở dữ liệu MySQL để lưu trữ. Em sử dụng 2 từ khóa “TRUNCATE” (xóa tất cả dữ liệu của 1 bảng) và “INSERT” (thêm dữ liệu vào bảng) để cập nhật dữ liệu mới cho bảng “SUMMARY\_COVID\_DATA” của Cơ sở dữ liệu.



Hình 3‑4. Dữ liệu trong MySQL

Hình 3-4 là dữ liệu thống kê về dịch COVID-19 được cập nhật vào MySQL, đây là dữ liệu được tính đến ngày 24/12/2021 cùng với những cột như:

* NAME: tên quốc gia.
* TOTAL\_CASES: tổng số ca nhiễm.
* TOTAL\_DEATHS: tổng số ca tử vong.
* TOTAL\_RECOVERED: tổng số ca phục hồi.
* FATALITY\_RATE: tỷ lệ tử vong.
* INFECTION\_RISK: nguy cơ lây nhiễm.
* SERIOUS\_CRITICAL: số ca nguy kịch.
* TOTAL\_TESTS: tổng số xét nghiệm.
* TEST\_PERCENTAGE: tỷ lệ xét nghiệm
* RECOVERY\_PROPORATION: tỷ lệ hồi phục.
* POPULATION: dân số.

### Thiết kế Dashboard bằng phần mềm Power BI

Power BI là phần mềm thuộc nhóm kinh doanh thông minh (Business Intelligence) của Microsoft, Power BI được phát triển với mục đích sử dụng dữ liệu để đưa ra các báo cáo và hiển thị thông tin chi tiết của doanh nghiệp một cách trực quan. Power BI có ở cả 3 nền tảng Website, Desktop và Thiết bị di động.



Hình 3‑5. Tính năng của Power BI

Power BI có khả năng kết nối được với nhiều loại dữ liệu khác nhau như Excel, Cơ sở dữ liệu (MySQL, Oracle, PostgreSQL, SQL Server …), nền tảng Azure, dịch vụ Online (Google Analytics, Salesforce) … Những tính năng của Power BI: Trí tuệ nhân tạo, Modelling, Tự động làm sạch dữ liệu, kết nối với API, hỗ trợ triển khai Hybrid …

* Ưu điểm của Power BI:
* Tích hợp Python và R để nâng cao chất lượng trực quan hóa.
* Tính năng Power Query giúp dễ dàng sử dụng và làm sạch dữ liệu.
* Được xây dựng trên nền tảng Trí tuệ nhân tạo (AI) và Machine Learning.
* Cài đặt lịch trình tự động cập nhật.

Đầu tiên, để có cái nhìn tổng quát về tình hình dịch COVID-19 trên thế giới, em có thiết kế 1 Dashboard (Hình 3-6) về tổng số ca nhiễm, tổng số ca tử vong, … ở các châu lục và quốc gia.



Hình 3‑6. Dashboard về số ca nhiễm, ca tử vong...

Tính đến ngày 26/12/2021 trên toàn thế giới, tổng số ca nhiễm là 279 triệu ca, tổng số ca tử vong là 5 triệu ca, tổng số ca hồi phục là 250 triệu ca.

Em sử dụng biểu đồ hình bánh (pie chart) (Hình 3-6[1], Hình 3-6[2], Hình 3-6[3]) để thể hiện tỷ lệ về tổng số ca nhiễm, ca tử vong và ca hồi phục của từng khu vực châu lục: châu Á, châu Âu, châu Mỹ (Nam Mỹ, Bắc Mỹ), Phi và châu Đại Dương:

* Về tỷ lệ số ca nhiễm, qua biểu đồ em thấy được châu Á và châu Âu là 2 châu lục có số ca nhiễm lớn nhất lần lượt là 84.1 và 83.3 triệu ca chiếm 30.05% và 29.75% so với toàn thế giới.
* Về tỷ lệ số ca tử vong, qua biểu đồ em thấy được châu Á và châu Âu là 2 châu lục có số ca tử vong lớn nhất lần lượt là 1.5 và 1.25 triệu ca chiếm 27.85% và 23.06% so với toàn thế giới.
* Về tỷ lệ số ca hồi phục, qua biểu đồ em thấy được châu Á và châu Âu là 2 châu lục có số ca hồi phục lớn nhất lần lượt là 81.4 và 72.6 triệu ca chiếm 32.51% và 29.01% so với toàn thế giới.

Em sử dụng biểu đồ cột (bar chart) (Hình 3-6[4], Hình 3-6[5], Hình 3-6[6]) cùng với đường trung bình (average line) để so sánh tổng số ca nhiễm, tử vong và phục hồi của 15 quốc gia có số ca nhiều nhất:

* Ở biểu đồ tổng số ca nhiễm: Mỹ, Ấn Độ và Brazil là 3 quốc gia có số ca nhiễm cao nhất lần lượt là 53, 34.8 và 22.2 triệu ca cùng với đường trung bình là 12.88.
* Ở biểu đồ tổng số ca tử vong: Mỹ, Brazil và Ấn Độ là 3 quốc gia có số ca tử vong cao nhất lần lượt là 0.83, 0.62 và 0.48 triệu ca cùng với đường trung bình là 0.258.
* Ở biểu đồ tổng số ca hồi phục: Mỹ, Ấn Độ và Brazil là 3 quốc gia có số ca hồi phục cao nhất lần lượt là 41, 34.2 và 21.4 triệu ca cùng với đường trung bình là 11.44.

Dashboard dưới đây (Hình 3-7) thể hiện về số ca dương tính, số ca nguy kịch của ngày 24/12/2021 và tỷ lệ tử vong, tỷ lệ mắc bệnh tính đến ngày 24/12/2021.



Hình 3‑7. Dashboard về tỷ lệ tử vong, tỷ lệ mắc bệnh...

Hình 3-7[1] thể hiện mối tương quan giữa tỷ lệ tử vong, tỷ lệ lây nhiễm và dân số của 10 quốc gia có tỷ lệ lây nhiễm cao nhất. Bar chart và line chart để thể hiện tỷ lệ lây nhiễm (INFECTION\_RISK), tỷ lệ tử vong (FATALITY\_RATE) và dân số (POPULATION)

* Andorra, Montenegro và Seychelles là 3 quốc gia có tỷ lệ lây nhiễm cao nhất lần lượt là 28.06%, 25.92% và 24.38%, có tỷ lệ tử vong lần lượt là 0.64%, 1.47% và 0.54%. Tỉ lệ mắc bệnh cao do dân số của 3 quốc gia này rất nhỏ, với tỷ lệ giữa số ca mắc và dân số là: Andorra (21571/77447), Montenegro (161944/ 628183) và Seychelles (24047/ 99245).

Hình 3-7[2] thể hiện tổng số ca nhiễm, tử vong và hồi phục của 10 quốc gia có tổng số ca nhiễm cao nhất:

* Biểu đồ cột xếp chồng (stacked column chart) để thể hiện sự tương quan giữa tổng số ca nhiễm, tử vong và hồi phục của từng quốc gia.
* Mỹ, Ấn Độ và Brazil là 3 quốc gia có tổng số ca nhiễm + tổng số ca tử vong + tổng số ca phục hồi cao nhất trên thế giới, với thống kê lần lượt là 95, 69 và 44 triệu ca.

Hình 3-7[3] thể hiện mối tương quan giữa dân số, số ca nhiễm và tỷ lệ tử vong của 10 quốc gia có số ca nhiễm cao nhất:

* Nga là quốc gia có tỷ lệ tử vong cao nhất với tỷ lệ tử vong 2.92%, tiếp đến là Brazil và Iran có tỷ lệ tử vong là 2.78% và 2.12%.

Hình 3-7[4] thể hiện mối tương quan giữa dân số, tổng số xét nghiệm và tỷ lệ xét nghiệm của 10 quốc gia có tổng số xét nghiệm nhiều nhất:

* “Stacked column và line chart” để thể hiện dân số (POPULATION), tổng số xét nghiệm (TOTAL\_TESTS) và tỷ lệ xét nghiệm (TEST\_PERCENTAGE).
* 3 quốc gia có dân số nhiều nhất thế giới là Trung Quốc (1.44 tỷ), Ấn Độ (1.4 tỷ) và Mỹ (333.8 triệu) có tỷ lệ xét nghiệm lần lượt là 11.12%, 48% và 239%.

Dashboard dưới đây (Hình 3-8) thể hiện tỷ trọng về tổng số ca nhiễm của các quốc gia tại châu Âu, châu Á, châu Mỹ (Nam và Bắc Mỹ), châu Phi và châu Đại Dương.



Hình 3‑8. Dashboard về tổng ca nhiễm của các châu lục

* Ở châu Âu (Hình 3-8[1]): Vương quốc Anh, Nga và Pháp có số ca nhiễm cao nhất là 11.9, 10.4 và 9.1 triệu ca chiếm tỷ trọng 16.85%, 14.7% và 12.88%.
* Ở châu Á (Hình 3-8[2]): Ấn Độ, Thổ Nhĩ Kỳ và Iran có số ca nhiễm cao nhất là 34.8, 9.3 và 6.2 triệu ca chiếm tỷ trọng 46.98%, 12.54% và 8.35%.
* Ở Bắc Mỹ (Hình 3-8[3]): Mỹ, Mexico và Canada có số ca nhiễm cao nhất là 53, 3.95 và 1.95 triệu ca chiếm tỷ trọng 84.42%, 6.29% và 3.12%.
* Ở Nam Mỹ (Hình 3-8[4]): Brazil, Argentina và Colombia có số ca nhiễm cao nhất là 22.2, 5.45 và 5.12 triệu ca chiếm tỷ trọng 56.35%, 13.82% và 12.98%.
* Ở châu Phi (Hình 3-8[5]): Nam Phi, Morocco và Tunisia có số ca nhiễm cao nhất là 3.4, 0.95 và 0.72 triệu ca chiếm tỷ trọng 42.2%, 11.83% và 8.95%.
* Ở châu Đại Dương (Hình 3-8[6]): Úc, Fiji và Polynesia thuộc Pháp có số ca nhiễm cao nhất là 301.2, 52.8 và 46.4 nghìn ca chiếm tỷ trọng 65%, 11.4% và 10.01%.

Dashboard dưới đây (Hình 3-9) thể hiện tỷ trọng về tổng số ca tử vong của các quốc gia tại châu Âu, châu Á, châu Mỹ (Nam và Bắc Mỹ), châu Phi và châu Đại Dương.



Hình 3‑9. Dashboard về tổng ca tử vong của các châu lục

* Ở châu Âu (Hình 3-9[1]): Nga, Vương quốc Anh và Ý có số ca tử vong cao nhất là 33.2, 147.8 và 136.5 nghìn ca chiếm tỷ trọng 23.53%, 11.47% và 10.59%.
* Ở châu Á (Hình 3-9[2]): Ấn Độ, Indonesia và Iran có số ca tử vong cao nhất là 479.7, 144.05 và 131.3 nghìn ca chiếm tỷ trọng 43.42%, 13.04% và 11.89%.
* Ở Bắc Mỹ (Hình 3-9[3]): Mỹ, Mexico và Canada có số ca tử vong cao nhất là 837.8, 298.7 và 30.1 nghìn ca chiếm tỷ trọng 68.02%, 24.26% và 2.45%.
* Ở Nam Mỹ (Hình 3-9[4]): Brazil, Peru và Colombia có số ca tử vong cao nhất là 618.4, 202.5 và 129.7 nghìn ca chiếm tỷ trọng 51.95%, 17.01% và 10.9%.
* Ở châu Phi (Hình 3-9[5]): Nam Phi, Morocco và Ai Cập có số ca tử vong cao nhất lần lượt là 90.7, 25.5 và 21.5 nghìn ca chiếm tỷ trọng 45.92%, 12.91% và 10.91%.
* Ở châu Đại Dương (Hình 3-9[6]): Úc, Fiji và Polynesia thuộc Pháp có số ca tử vong cao nhất là 2.2, 0.7 và 0.4 nghìn ca chiếm tỷ trọng 49.21%, 15.66% và 14.29%.

Dashboard dưới đây (Hình 3-10) thể hiện tỷ trọng về tổng số ca phục hồi của các quốc gia tại châu Âu, châu Á, châu Mỹ (Nam và Bắc Mỹ), châu Phi và châu Đại Dương.



Hình 3‑10. Dashboard về tổng ca phục hồi của các châu lục

* Ở châu Âu (Hình 3-10[1]): Vương quốc Anh, Nga và Pháp có số ca phục hồi cao nhất là 9.96, 9.22 và 7.79 triệu ca chiếm tỷ trọng 16.13%, 14.94% và 12.62%.
* Ở châu Á (Hình 3-10[2]): Ấn Độ, Thổ Nhĩ Kỳ và Iran có số ca phục hồi cao nhất là 34.2, 8.92 và 6.02 triệu ca chiếm tỷ trọng 47.55%, 12.4% và 8.37%.
* Ở Bắc Mỹ (Hình 3-10[3]): Mỹ, Mexico và Canada có số ca phục hồi cao nhất là 41, 3.3 và 1.81 triệu ca chiếm tỷ trọng 82.8%, 6.66% và 1.23%.
* Ở Nam Mỹ (Hình 3-10[4]): Brazil, Argentina và Colombia có số ca phục hồi cao nhất là 21.4, 5.26 và 4.96 triệu ca chiếm tỷ trọng 56.72%, 13.93% và 13.13%.
* Ở châu Phi (Hình 3-10[5]): Nam Phi, Ma-rốc và Tunisia có số ca phục hồi cao nhất là 3.1, 0.93 và 0.69 triệu ca chiếm tỷ trọng 42.14%, 12.73% và 9.44%.
* Ở châu Đại Dương (Hình 3-10[6]): Úc, Fiji và Polynesia thuộc Pháp có số ca phục hồi cao nhất lần lượt là 229.3, 51.2 và 45.7 nghìn ca chiếm tỷ trọng 59.31%, 13.25% và 11.83%.

## Chi tiết về dịch COVID-19

### Giới thiệu

Ở phần 3.2, là cái nhìn tổng quát về tình hình dịch COVID-19 của Thế giới khi mà virus SARS-CoV-2 đã hoành hành khắp cả 5 châu lục và khiến hàng trăm triệu người lây nhiễm, hàng triệu người tử vong ảnh hưởng tiêu cực đến Y tế, Kinh tế …

Nguồn dữ liệu sử dụng đến từ Kho dữ liệu của Trung tâm Hệ thống Khoa học và Kỹ thuật (CSSE) tại trường Đại học Johns Hopkins.

### Biểu đồ và phân tích với Python



Hình 3‑11. Tương quan giữa các khu vực trên Thế giới

Hình 3-11 sử dụng bảng dạng gradient để thể hiện mối tương quan giữa các khu vực về tổng số ca nhiễm, ca tử vong, ca phục hồi, ca dương tính … Màu càng đậm cho thấy số lượng càng lớn và ngược lại. Tính đến ngày 26/12/2021: Châu Âu và Châu Á có tổng số ca nhiễm, tử vong và hồi phục cao nhất. Nam Mỹ là nơi có ti lệ tự vong cao nhất với 3.02%, tiếp đến là Châu Phi (2.39%) và Bắc Mỹ (1.96%).



Hình 3‑12. Số ca nhiễm, tử vong và đường Moving Average của Thế giới

Hình 3-12 sử dụng biểu đồ đường (line chart) để thể hiện số lượng ca nhiễm và ca tử vong cộng dồn (cumulative) theo thời gian cùng với đường Moving Average 7 ngày (MA7). Đường MA7 đi sát với đường số ca nhiễm và tử vong cho thấy tăng đều, nhanh và ít biến động.



Hình 3‑13. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Thế giới

Hình 3-13 sử dụng biểu đồ thanh (bar chart) để thể hiện số ca nhiễm và tử vong theo từng ngày của Thế giới cùng với đường MA7. Số ca nhiễm cao nhất trong 1 ngày hơn 1.4 triệu ca và số ca tử vong cao nhất trong 1 ngày là hơn 20 nghìn ca.



Hình 3‑14. Tỉ lệ tử vong theo ngày và trung bình của Thế giới

Hình 3-14 sử dụng biểu đồ đường để thể hiện tỷ lệ tử vong theo ngày và trung bình của Thế giới. Tỉ lệ tử vong trung bình là hơn 0.03, thấp hơn nhiều so với những dịch bệnh khác. Tỉ lệ tử vong tăng nhanh từ tháng 2/2020 đến 6/2020 và giảm mạnh đến gần cuối năm 2020, cùng với việc các quốc gia trên Thế giới triển khai việc tiêm Vắc-xin giúp cho tỷ lệ tử vong duy trì mức 0.02 và giảm dần theo thời gian.

Bên cạnh thông tin về các Quốc gia thì thông tin về các Tỉnh/Bang của các Quốc gia này cũng khai thác để có được cái nhìn chi tiết hơn.



Hình 3‑15. Top 10 các Tỉnh có số ca cao nhất trên Thế giới

Hình 3-15 sử dụng bảng gradient để thể hiện sự tương quan về tỷ lệ lây nhiễm, tỷ lệ tử vong, số ca nhiễm và tử vong giữa các Tỉnh/Bang của các Quốc gia. Anh, Maharashtra và California là 3 tỉnh có số cao nhiễm cao nhất (10.1, 6.65 và 5.29 triệu ca).



Hình 3‑16. Top 10 Bang của Mỹ có số ca cao nhất

Hình 3-16 sử dụng bảng gradient để thể hiện sự tương quan về tỷ lệ lây nhiễm, tỷ lệ tử vong, số ca nhiễm và tử vong giữa các Bang của Mỹ. California, Texas và Florida là 3 bang có số ca nhiễm và tử vong cao nhất tại Mỹ (5.3, 4,48 và 3.83 triệu ca). Texas có tỷ lệ lây nhiễm cao nhất hơn 4.1 triệu người cùng tỷ lệ tử vong là 0.016%, New Jersey có tỷ lệ lây nhiễm thấp nhất 328.2 nghìn người cùng tỷ lệ tử vong cao là 0.02%.



Hình 3‑17. Top 10 nước và phần còn lại về số ca nhiễm

Hình 3-17 sử dụng biểu đồ cột (bar chart) để thể hiện số ca nhiễm của top 10 nước có số ca nhiễm cao nhất và các nước còn lại sẽ là “Others”.



Hình 3‑18. Top 10 tỉnh và phần còn lại về số ca nhiễm

Hình 3-18 sử dụng biểu đồ cột (bar chart) để thể hiện số ca nhiễm của top 10 tỉnh có số ca nhiễm cao nhất và các nước còn lại sẽ là “Others”.

#### So sánh Việt Nam và Châu Á



Hình 3‑19. Số ca nhiễm, tử vong và đường Moving Average của Châu Á

Hình 3-19 sử dụng biểu đồ đường (line chart) để thể hiện số lượng ca nhiễm và ca tử vong cộng dồn (cumulative) theo thời gian cùng với đường Moving Average 7 ngày (MA7). Số ca nhiễm và tử vong tăng nhanh vào tháng 5/2021 và vẫn chưa có dấu hiệu dừng lại. Đường MA7 đi sát với đường số ca nhiễm và tử vong cho thấy tăng đều, nhanh và ít biến động.



Hình 3‑20. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Châu Á

Hình 3-20 sử dụng biểu đồ đường và cột (bar and line chart) để thể hiện số ca nhiễm và tử vong theo từng ngày của Châu Á cùng với đường MA7. Đầu tháng 3/2021 là thời điểm dịch COVID-19 bùng phát mạnh ở Ấn Độ dẫn đến số ca nhiễm và tử vong tăng mạnh. Số ca nhiễm cao nhất trong 1 ngày gần 900 nghìn ca và số ca tử vong cao nhất trong 1 ngày là hơn 8 nghìn ca.



Hình 3‑21. Số ca nhiễm và tử vong giữa Việt Nam và các nước Châu Á

Hình 3-21 sử dụng biểu đồ đường thể hiện tốc độ gia tăng số ca nhiễm và tử vong cộng dồn theo ngày giữa Việt Nam và các quốc gia khác ở châu Á. Ấn Độ là quốc gia có dịch bùng phát mạnh vào tháng 4/2021 dẫn đến số ca nhiễm và tử vong tăng nhanh. So sánh giữa Việt Nam và các nước châu Á, Việt Nam có tốc độ gia tăng ít hơn về cả số ca nhiễm và tử vong, đây là 1 biểu đồ tốt dành cho Việt Nam.



Hình 3‑22. Biểu đồ cột giữa Việt Nam và các nước châu Á

Hình 3-22 sử dụng biểu đồ cột để so sánh về số ca nhiễm, tử vong và hồi phục giữa Việt Nam và các nước Châu Á. Ấn Độ là quốc gia có số ca nhiễm, tử vong và phục hồi cao nhất ở Châu Á.

#### So sánh Việt Nam và Đông Nam Á (South East Asia)



Hình 3‑23. Số ca nhiễm, tử vong và đường Moving Average của Đông Nam Á

Hình 3-23 sử dụng biểu đồ đường (line chart) để thể hiện số lượng ca nhiễm và ca tử vong cộng dồn (cumulative) theo thời gian cùng với đường Moving Average 7 ngày (MA7). Số ca nhiễm và tử vong tăng nhanh vào tháng 6/2021 và vẫn chưa có dấu hiệu dừng lại. Đường MA7 đi sát với đường số ca nhiễm và tử vong cho thấy tăng đều, nhanh và ít biến động.



Hình 3‑24. Số ca nhiễm, tử vong mỗi ngày và đường MA7 của Đông Nam Á

Hình 3-24 sử dụng biểu đồ đường và cột (bar and line chart) để thể hiện số ca nhiễm và tử vong theo từng ngày của Đông Nam Á cùng với đường MA7. Số ca nhiễm cao nhất trong 1 ngày gần 900 nghìn ca và số ca tử vong cao nhất trong 1 ngày là hơn 8 nghìn ca. Số ca nhiễm và số ca tử vong luôn ở mức rất cao từ tháng 5/2021 đến tháng 9/2021.



Hình 3‑25. Số ca nhiễm và tử vong giữa Việt Nam và các nước Đông Nam Á

Hình 3-25 sử dụng biểu đồ đường thể hiện tốc độ gia tăng số ca nhiễm và tử vong cộng dồn theo ngày giữa Việt Nam và các quốc gia Đông Nam Á. Đầu tháng 6/2021, dịch bùng phát mạnh ở khu vực Đông Nam Á, nhiều quốc gia bị ảnh hưởng như Indonesia, Philippines, Malaysia, Thailand và Việt Nam. So sánh giữa Việt Nam và các nước Đông Nam Á, Việt Nam có tốc độ gia tăng về cả số ca nhiễm và tử vong chỉ sau Indonesia, Philippines, Malaysia và Thailand.



Hình 3‑26. Biểu đồ cột giữa Việt Nam và các nước Đông Nam Á

Hình 3-26 sử dụng biểu đồ cột để so sánh về số ca nhiễm, tử vong và hồi phục giữa Việt Nam và các nước Đông Nam Á. Indonesia là quốc gia có số ca nhiễm, tử vong và phục hồi cao nhất ở Đông Nam Á. Việt Nam đứng thứ 5 trong khu vực.

## Tình hình tiêm ngừa Vắc-xin COVID-19

### Tình hình tiêm ngừa Vắc-xin COVID-19 của Thế giới

Các bảng và biểu đồ được thu thập và cập nhật đến ngày 27/12/2021.



Hình 3‑27. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin

Hình 3-27 là bảng sơ lược về tiến độ tiêm chủng Vắc-xin COVID-19 trên toàn Thế giới, đã có 223 quốc gia tham gia tiêm ngừa, 18 loại Vắc-xin được sử dụng, hơn 9105 triệu liều Vắc-xin đã được tiêm, hơn 3274 triệu người đã được tiêm và hơn 2591 triệu người đã tiêm đủ 2 mũi (đối với loại Vắc-xin yêu cầu 2 mũi).



Hình 3‑28. Biểu đồ số lượng Quốc gia sử dụng Vắc-xin đơn

Hình 3-28 sử dụng biểu đồ cột ngang (horizontal bar chart) để thể hiện số ượng quốc gia sử dụng Vắc-xin đơn. Oxford/AstraZeneca (của Anh) được 23 quốc gia sử dụng, Pfizer/BioTech (của Mỹ và Đức) được 7 quốc gia sử dụng và Sinopharm/Beijing (của Trung Quốc) được 3 quốc gia sử dụng.



Hình 3‑29. Biểu đồ top 10 Quốc gia có tổng số Vắc-xin đã tiêm



Hình 3‑30. Biểu đồ top 10 Quốc gia có số Vắc-xin tiêm mỗi ngày



Hình 3‑31. Biểu đồ top 10 Quốc gia có số người đã tiêm đầy đủ

Hình 3-29 sử dụng biểu đồ cột để thể hiện tổng số lượng Vắc-xin đã được tiêm của top 10 quốc gia. Trung Quốc là quốc gia dẫn đầu với hơn 2.7 tỷ liều, tiếp đến là Ấn Độ và Mỹ với 1.5 tỷ và 500 triệu liều.

Hình 3-30 sử dụng biểu đồ cột để thể hiện số lượng Vắc-xin tiêm mỗi ngày cao nhất của top 10 quốc gia. Trung Quốc là quốc gia dẫn đầu với hơn 22 triệu liều, tiếp đến là Ấn Độ và Bangladesh với 10 và 4 triệu liều.

Hình 3-31 sử dụng biểu đồ cột để thể hiện tổng số người đã được tiêm đầy đủ của top 10 quốc gia. Trung Quốc là quốc gia dẫn đầu với hơn 1.2 tỷ người, tiếp đến là Ấn Độ và Mỹ với 577 và 200 triệu người.



Hình 3‑32. Biểu đồ số lượng Quốc gia sử dụng Vắc-xin kết hợp

Hình 3-32 sử dụng biểu đồ cột ngang (horizontal bar chart) để thể hiện số ượng quốc gia sử dụng Vắc-xin đơn. Kết hợp giữa Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioTech được 24 quốc gia sử dụng, kết hợp giữa Moderna, Oxford/AstraZeneca, Pfizer/BioTech được 15 quốc gia sử dụng và kết hợp giữa Oxford/AstraZeneca, Pfizer/BioTech được 11 quốc gia sử dụng.



Hình 3‑33. Biểu đồ số lượng Vắc-xin mỗi ngày của Thế giới

Hình 3-33 sử dụng biểu đồ đường để thể hiện số lượng Vắc-xin được tiêm mỗi ngày của Thế giới. Số lượng Vắc-xin tăng mạnh vào đầu tháng 3/2021 và đạt định vào đầu tháng 7/2021 với số lượng 43.9 triệu liều 1 ngày.

### So sánh Việt Nam và Châu Á



Hình 3‑34. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin ở Châu Á

Hình 3-34 là bảng sơ lược về tiến độ tiêm chủng Vắc-xin COVID-19 ở Châu Á, đã có 50 quốc gia tham gia tiêm ngừa, 18 loại Vắc-xin được sử dụng, hơn 6104 triệu liều Vắc-xin đã được tiêm, hơn 1803 triệu người đã được tiêm và hơn 1326 triệu người đã tiêm đủ 2 mũi (đối với loại Vắc-xin yêu cầu 2 mũi).



Hình 3‑35. Biểu đồ số lượng Vắc-xin mỗi ngày của các nước Châu Á

Hình 3-35 là biểu đồ đường của các quốc gia tại châu Á có số lượng Vắc-xin tiêm mỗi ngày cùng với Việt Nam. Trung Quốc là quốc gia có số lượng mũi tiêm một ngày cao nhất khi đạt đinh 22.4 triệu liều/ngày. Tiếp đến là Ấn Độ với gần 10 triệu liều/ngày. Cuộc chạy đua Vắc-xin ở Châu Á bắt đầu từ tháng 5 cũng là lúc dịch bùng phát mạnh ở khu vực châu Á và Việt Nam.



Hình 3‑36. Biểu đồ số lượng Vắc-xin cộng dồn của các nước Châu Á

Hình 3-36 là biểu đồ đường thể hiện số lượng Vắc-xin cộng dồn của Việt Nam và các nước châu Á. Trung Quốc là quốc gia có xu hướng tăng mạnh nhất so với phần còn lại, tăng mạnh vào đầu tháng 7/2021 đến giữa tháng 9/2021, tính cho đến ngày 26/12/2021 đã tiêm được hơn 2.5 tỷ liều. Tiếp đến là Ấn Độ, có xu hướng tăng từ tháng 8/21 cho đến nay và đạt được hơn 1 tỷ mũi tiêm.



Hình 3‑37. Biểu đồ phân phối tiêm Vắc-xin mỗi ngày của Việt Nam

Hình 3-37 là biểu đồ cột thể hiện số lượng Vắc-xin phân phối mỗi ngày của Việt Nam. Có hơn 140 ngày mà Việt Nam tiêm được 0.2 triệu liều/ngày và khoảng 5 ngày Việt Nam tiêm được từ 1.6 đến 1.8 triệu liều/ngày. Từ 0.6 đến 1.4 triệu liều/ngày là số lượng tiêm đạt được trong 20 ngày.

### So sánh Việt Nam và Đông Nam Á (South East Asia)



Hình 3‑38. Bảng sơ lược về tiến độ tiêm chủng Vắc-xin ở Đông Nam Á

Hình 3-38 là bảng sơ lược về tiến độ tiêm chủng Vắc-xin COVID-19 trên toàn Thế giới, đã có 10 quốc gia tham gia tiêm ngừa, 10 loại Vắc-xin được sử dụng, hơn 754 triệu liều Vắc-xin đã được tiêm, hơn 352 triệu người đã được tiêm và hơn 263 triệu người đã tiêm đủ 2 mũi (đối với loại Vắc-xin yêu cầu 2 mũi).



Hình 3‑39. Biểu đồ số lượng Vắc-xin mỗi ngày của Đông Nam Á

Hình 3-39 là biểu đồ đường thể hiện số lượng Vắc-xin tiêm mỗi ngày của các Quốc gia Đông Nam Á. Indonesia là quốc gia có số lượng Vắc-xin tiêm mỗi ngày cao nhất (đạt đỉnh 1.9 triệu 1 ngày), đứng thứ 2 là Việt Nam khi đạt đỉnh là 1.6 triệu liều. Số liều Vắc-xin tăng mạnh vào đầu tháng 7/2021 cũng chính là thời điểm dịch bùng phát mạnh ở khu vực Đông Nam Á, dẫn đến các quốc gia triển khai tiêm ngừa nhanh chóng để chống lại dịch COVID-19.



Hình 3‑40. Biểu đồ số lượng Vắc-xin cộng dồn của Đông Nam Á

Hình 3-40 là biểu đồ đường thể hiện số lượng Vắc-xin cộng dồn của Việt Nam và các nước Đông Nam Á. Indonesia là quốc gia có xu hướng tăng mạnh nhất so với phần còn lại, tăng mạnh vào đầu tháng 7/2021 đến nay, tính cho đến ngày 26/12/2021 đã tiêm được hơn 250 triệu liều. Tiếp đến là Việt Nam, có xu hướng tăng từ tháng 9/2021 đến nay và đạt được gần 150 triệu mũi tiêm.



Hình 3‑41. Biểu đồ Scatter về tổng số lượng Vắc-xin của Đông nam Á

Hình 3-41 là biểu đồ Scatter thể hiện tổng số lượng Vắc-xin của các nước Đông Nam Á. Biểu đồ này có tê tương tác trực tiếp để xem số lượng Vắc-xin đã tiêm được của từng quốc gia. Ví dụ ngày 4/7/2021, Indonesia có 46.33 triệu liều, Philippines có 11.7 triệu liều, Thái Lan có 11.06 triệu liều và Việt Nam có 3.903 triệu liều.

# XÂY DỰNG MÔ HÌNH DỰ ĐOÁN

## Thuật toán SVM, Linear Regression và Bayesian Ridge Regression

### Thuật toán SVM (Support Vector Machines)

SVM là thuật toán máy học (Machine Learning) có giám sát, được sử dụng phổ biến trong các bài toán phân lớp (classification) và hồi quy (regression). SVMs được chia thành 2 loại:

* SVM: dùng cho các bài toán phân lớp.
* SVR (Support Vector Regression): dùng cho các bài toán hồi quy.

Trong thực tế, SVM có nhiều loại bài toán con như: linear, non-linear, soft-margin, hard-margin, binary-class và multi-class.

Ý tưởng của SVM là tìm một siêu phẳng (hyper-plane) đê phân tách các điểm dữ liệu, siêu phẳng này sẽ chia không gian dữ liệu thành các vùng khác nhau và mỗi vùng sẽ chứa một loại dữ liệu.



Hình 4‑1. Siêu phẳng (hyper-plane) trong SVM

Trong một không gian có rất nhiều siêu phẳng (Hình 4-1), để tối ưu hóa và mang lại kết quả chính xác cần phải tìm ra siêu phẳng tối ưu nhất. Siêu phẳng được biểu diễn bằng hàm số  (với W và X là vector <W.X> là tích vô hướng) hay  (WT là ma trận chuyển vị). Cách chọn siêu phẳng tối ưu:



Hình 4‑2. Margin của siêu phẳng trong SVM

Trong hình 4-2, siêu phẳng phân tách tập dữ liệu thành 2 lớp (chấm đỏ và chấm xanh) H0 thỏa mãn <W.X> + b =0, siêu phẳng này tạo ra thêm 2 nửa không gian (half-space) dữ liệu: Không gian chứa dữ liệu lớp chấm đỏ Xi thỏa mãn và không gian chứa dữ liệu chấm xanh Xj thỏa mãn.

Chọn hai siêu phẳng lề H1 đi qua điểm thuộc lớp chấm đỏ và H2 đi qua điểm thuộc lớp chấm xanh đều song song với H0.

* H1 : <W.X> + b =-1
* H2 : <W.X> + b =1

Khoảng cách từ H1 đến H0 là d-

Khoảng cách từ H2 đến H0 là d+

m = d- + d+ được gọi là mức lề

* **Siêu phẳng tối ưu** là siêu phẳng phân tách có lề (margin) lớn nhất, trong lý thuyết học máy cho rằng một siêu phẳng tối ưu sẽ cực tiểu hóa giới hạn lỗi mắc phải.

Để tìm được siêu phẳng tối ưu, trong thuật toán SVM cần phải tìm được margin:

Khoảng cách từ 1 điểm Xk đến siêu phẳng H0: trong đó ||W|| là độ dài vector W: ||W||=<W.W>= 

Khoảng cách từ một điểm Xi nằm trên H1 và H0: 

Khoảng cách từ một điểm Xj nằm trên H2 và H0: 

🡪 Các điểm nằm trên 2 mặt phẳng H1 và H2 được gọi là Support Vector.

Cực tiểu hóa (minimization): với điều kiện:

* , nếu yi=-1
* , nếu yi=1
* Bài toán **Hard-margin** của SVM.

Ở hình 4-2, việc xác định siêu phẳng trong điều kiện lý tưởng khi mà tập dữ liệu có thể phân tách tuyến tính, tìm được 2 siêu phẳng lề H1 và H2 mà không có điểm nhiễu nào nằm giữa chúng. Vậy trong trường hợp có nhiễu, bài toán cần sử dụng biến **slack.**



Hình 4‑3. Siêu phẳng có chứa tín hiệu nhiễu

Với trường hợp có tín hiệu nhiễu, cần sử dụng biến slack:

* , nếu yi=-1
* , nếu yi=-1

Cực tiểu hóa: 

Với điều kiện:

* 
* 

Trong đó, C là tham số xác định mức độ phạt (penalty degree) đối với lỗi, là một tham số quan trọng mà cần phải tối ưu trong các bài toán SVM.

* Bài toán **Soft-margin** của SVM

Một bài toán có không gian dữ liệu phi tuyến tính (non-linear) khi không tìm được siêu phẳng H0­.



Hình 4‑4. Non-linear trong SVM

Chúng ta cần phải biểu diễn (ánh xạ) dữ liệu từ dữ liệu từ không gian ban đầu X sang không gian F bằng một hàm ánh xạ phi tuyến để giải quyết bài toán:

* 
* 
* Bài toán **Non-linear** của SVM

### Thuật toán Linear Regression

#### Giới thiệu Linear Regression

Thuật toán Linear Regression (Hồi quy tuyến tính) là phương pháp tiếp cận tuyến tính để dự đoán mối quan hệ giữa biến phụ thuộc Y (biến kết cục) trên trục tung Y dự trên các biến độc lập X (biến giải thích) trên trục hoành X trong mô hình. Hồi quy tuyến tính được phát triển thành Mô hình hồi quy tuyến tính (Linear Regression Model) là công cụ quan trọng trong lĩnh vực Kinh tế lượng và là phương pháp thống kê hồi quy và dự báo dữ liệu.

Phương pháp Hồi quy Tuyến tính tổng quát: Mô hình Hồi quy Tuyến tính bội (Multiple Regression model) với dạng Mô hình Hồi quy tổng thể (Population Regression model) với n-1 biến giải thích có dạng:

(1)

* Y là biến phụ thuộc (dependent variable).
* X là các biến giải thích hay biến độc lập (explanatory/independent variables).
* βn là hệ số của các biến độc lập, trong đó B1 là hệ số tự do.
* u là hạng nhiễu hay sai số ngẫu nhiên (random hay stochastic error term).
* i là ký hiệu cho quan sát thứ i trong tổng thể.

Phương trình (1) rút gọn thành mô hình tổng thể (population model):

 Với 

* βx là thành phần tất định (deterministic component), là 1 trung bình có điều kiện (conditional mean) của Yi, nghĩa là E(Yi|X): giá trị trung bình của Y khi X là giá trị được cho trước.
* Ui là thành phần phi hệ thống (nonsystematic component).

Phương trình: Y = β0 + β1\*X, Với β1 là độ dốc (slope); β0 là chặn (intercept)

* β1 = Tương quan \* ( Độ lệch chuẩn của Y / Độ lệch chuẩn của X)
* β0 = Trung bình (Y) – B1 \* Trung bình (X)



Hình 4‑5. Mô hình Hồi quy tuyến tính Linear Regression

Với: Độ lệch chuẩn của X.

Với : Độ lệch chuẩn của Y.

Độ mạnh của Mô hình hồi quy tuyến tính: được đo lường thông qua **hệ số xác định r2** (coefficient of determination) hay **R-Square (R2)** với dao động từ 0 đến 1.

Công thức tính R-Square (R2): **R2 = (TSS – RSS)/TSS.**



Hình 4‑6. R-Square (R2) của Mô hình hồi quy tuyến tính

* Tổng các diện tích (TSS): phép đo tổng hợp biến thiên trong tỷ lệ đáp ứng / biến phụ thuộc và có thể được coi là số lượng biến thiên vốn có trong đáp ứng trước khi hồi quy được thực hiện.
* Sum of Squares (RSS): RSS đo lường lượng biến đổi còn lại không giải thích được sau khi thực hiện hồi quy.
* Hiệu số TSS – RSS đo lường mức độ thay đổi trong đáp ứng được giải thích.

#### GIới thiệu OLS (Ordinary Least Square):

Là mô hình hồi quy bằng phương pháp bình phương nhỏ nhất hay cực tiểu hay còn được gọi là mô hình Pooled OLS (Pool) là một trường hợp đặc biệt của phương pháp bình phương cực tiểu tổng quát (generalized least squares method – GLS) để tìm đường hồi quy gần nhất với giá trị liên tục của biến phụ thuộc (tổng bình phương của các nhiễu hay sai số (∑ui2 ) nhỏ nhất).

Công thức tổng quát: Yi = b1 + b2X2i+ b3X3i +…+ bkXki+ui

Phương sai và sai số chuẩn: , trong đó:

* Phương sai: tỷ lệ của tổng bình phương phần dư (RSS) chia cho (n – k) với bậc tự do, n là cỡ mẫu và k là số tham số hồi quy ước lượng.
* 𝜎̂: sai số chuẩn của hồi quy (standard error of the regression, SER), là độ lệch chuẩn của các giá trị Y xoay quanh đường hồi quy.



Hình 4‑7. Phương pháp Ordinary Least Square

### Thuật toán Bayesian Ridge Regression

Kỹ thuật hồi quy Bayes (Bayesian Regression) được sử dụng để bao gồm các tham số chính quy hóa trong quá trình ước lượng: tham số chính quy hóa không được định cứng mà sẽ được điều chỉnh theo dữ liệu có sẵn. Điều này thực hiện bằng cách giới thiệu cò mồi không thông tin trên các siêu tham số của mô hình. Các *l2* chính quy hóa được sử dụng trong hồi quy và phân loại Ridge bằng với việc tìm 1 ước lượng hậu kỳ tối đa Gaussian trên các hệ số**với độ chính xác ****. Hồi quy Ridge giúp giải quyết một số vấn đề của OLS (Ordinary Least Squares) bằng cách áp đặt hình phạt (penalty) với kích cỡ của các hệ số, hệ số Ridge tối thiểu hóa tổng bình phương (RSS) còn lại:



Thay vì đặt  thủ công, cho nó như 1 biến ngẫu nhiên ước lượng từ dữ liệu. Với mô hình xác suất đầy đủ, đầu ra ***y***là Gaussian được phân phối xung quanh:  với là biến ngẫu nhiên được ước tính từ dữ liệu.

* Ưu điểm của Hồi quy Bayes:
* Thích ứng với dữ liệu trong tầm tay.
* Sử dụng để bao gồm các tham số chính quy hóa trong quy trình ước tính.
* Nhược điểm của Hồi quy Bayes:
* Việc suy diễn mô hình có thể tốn nhiều thời gian.

Hồi Quy Bayesian Ridge ước tính một mô hình xác suất của bài toán Hồi quy, giá trị trước cho hệ số**đưa ra bởi Gaussian hình cầu:. Các mồivàđược chọn là phân bố Gamma, liên hợp cho độ chính xác Gaussian, mô hình kết quả gọi là Hồi quy Bayesian Ridge. Các tham số, và được ước tính phù hợp với mô hình, các tham số chính quy hóa và  ước tính bằng tối đa hóa khả năng cận biên log (log marginal likelihood). Giá trị ban đầu của tối đa hóa có được đặt bằng siêu tham số alpha\_init và lambda\_init. Có 4 siêu tham số , ,  và  của các phân phối Gamma và , mặc định .

So sánh với ước lượng OLS, trọng số có xu hướng dịch chuyển về phía số không, điểu giúp ổn định thuật toán. Việc ước lượng hóa mô hình được lặp đi lặp lại tối đa hóa khả năng log biên của các quan sát.



Hình 4‑8. So sánh các trọng số của mô hình

Hình 4-8 là biểu đồ so sánh các dự đoán và độ không chắc chắn của Hồi quy Bayesian Ridge cho hồi quy một chiều bằng phương pháp Polynomial Feature và ước lượng OLS.

Việc lựa chọn các tham số khởi tạo tham số chính quy hóa (alpha, lambda) khá quan trọng. Do các tham số chính quy hóa được xác định bởi quy trình lặp đi lặp lại phụ thuộc vào biến khởi tạo khi điều chỉnh đường cong Hồi quy Bayesian Ridge.



Hình 4‑9. Curve Fitting với Hồi quy Bayesian Ridge

Hình 4-9 là ví dụ hình sin được tính bằng đa thức được sử dụng các cặp giá trị ban đầu. Với cặp mặc định (=1.9, =1) độ lệch đường cong quá lớn và phương sai nhỏ. Thay vào đó là =1, =10-3 để giảm độ lệch.

## Áp dụng vào mô hình dự đoán

Trong phần này, em sử dụng 3 thuật toán: SVR, Linear Regression và Linear Regression để dự đoán số ca nhiễm của Thế giới trong 15 ngày tiếp theo.

### Áp dụng thuật toán SVR



Hình 4‑10. Kết quả dự đoán của thuật toán SVR

Hình 4-10 là biểu đồ đường thể hiện kết quả dự đoán của thuật toán SVR. Đường màu xanh là số ca nhiễm thực tế và đường màu tím gạch đứt là số ca nhiễm dự đoán.

Tham số được sử dụng trong mô hình SVR:

* kernel: “poly”, gamma (hệ số Kernel) = 0.01, epsilon=1, degree (hàm của đa thức Kernel)=2, C (tham số chính quy)=0.1

Thông số kỹ thuật (metric) của mô hình thuật toán Polynomial Regression:

* Lỗi trung bình tuyệt đối (Mean Absolute Error): MAE=11830855.88
* Lỗi bình phương trung bình (Mean Squared Error): MSE= 150007501802146
* R-Square= 0.98
* Hệ số chặn: intercept=5111133.31

### Áp dụng thuật toán Linear Regression



Hình 4‑11. Kết quả dự đoán của thuật toán Polynomial Regression

Hình 4-11 là biểu đồ đường thể hiện kết quả dự đoán của thuật toán Polynomial Regression. Đường màu xanh là số ca nhiễm thực tế và đường màu cam gạch đứt là số ca nhiễm dự đoán.

Do số lượng feature đầu vào để training nhỏ (hơn 700) nên cần sử dụng Polynomial Features để cải thiện các feature của thuật toán.

* Giới thiệu về **Polynomial Feature** và công dụng:

Trong các bài toán với số lượng dữ liệu đầu vào nhỏ, tạo Polynomial Feature để cải thiện thuật toán. Để cải thiện các feature của thuật toán bằng nhiều cách khác nhau, có thể kết hợp nhiều feature lại thành 1. Cách thường được sử dụng nhất là thêm vào các bậc cao hơn cho cùng 1 feature để nâng bậc cho hàm: *y=f(x)=b0+b1x1* thành hàm bậc 2: *y=f(x)=b0+b1x1+b2x12*hay bậc 3: *y=f(x)=b0+b1x1+b2x12+ b3x13.*

Một điều quan trọng trong triển khai hồi quy đa thức (Polynomial Regression) là việc lựa chọn mức độ tối ưu của hàm hồi quy đa thức. Việc lựa chọn này không có quy tắc cụ thể, tùy thuộc vào từng bài toán, có 2 vấn đề cần chú ý khi lựa chọn bậc cho hàm (degree): underfitting and overfitting.



Hình 4‑12. Các mô hình Underfitted, well-fitted and overfitted

Hình 4-12 là các ví dụ về các trường hợp Underftting, Overfitting và Well-fitting khi ử dụng bậc 1, 2, 3 và 5. Underfitting là khi một mô hình có kết quả dự đoán không chính xác, không thể nắm bắt chính xác các yếu tố phụ thuộc của dữ liệu. Overfitting là khi một mô hình học quá tốt dữ liệu hiện có, tuy nhiên không khái quát tốt và có R2 thấp khi có dữ liệu mới.

Trong bài toán dự đoán này, dữ liệu training đầu vào là số ngày nên em dùng Polynomial Features để nâng lên bậc 3 (degree=3) nên thuật toán này có tên Polynomial Regression, phương trình có dạng: *y=f(x)=b0+b1x1+b2x12+ b3x13.*

Thông số kỹ thuật (metric) của mô hình thuật toán Polynomial Regression:

* Lỗi trung bình tuyệt đối (Mean Absolute Error): MAE=2903897.15
* Lỗi bình phương trung bình (Mean Squared Error): MSE=16114561353282
* R-Square=0.998
* Hệ số chặn: intercept=0
* Hệ số góc: coefficients=(6.901e+6, -2.335e+5, 1.738e+3, -1.24e+0)
* y=f(x)= 6.901e+6 -2.335e+5*x1*+1.738e+3*x12*-1.24e+0*x13*

### Áp dụng thuật toán Bayesian Ridge Regression



Hình 4‑13. Kết quả dự đoán của thuật toán Bayesian Ridge Regression

Hình 4-13 là biểu đồ đường thể hiện kết quả dự đoán của thuật toán Bayesian Ridge Regression. Đường màu xanh là số ca nhiễm thực tế và đường màu xanh lá gạch đứt là số ca nhiễm dự đoán.

Sử dụng tham số:

* n\_iter (số lần lặp tối đa)=40
* =[1e-7, 1e-6, 1e-5, 0.0001,0.001]
* tol = [1e-6, 1e-5, 1e-4, 1e-3, 1e-2]
* normalize = [True, False]

Thông số kỹ thuật (metric) của mô hình thuật toán Bayesian Ridge Regression:

* Lỗi trung bình tuyệt đối (Mean Absolute Error): MAE= 2873488.36
* Lỗi bình phương trung bình (Mean Squared Error): MSE=15752742048978
* R-Square=0.998
* Hệ số chặn: intercept=0
* Hệ số góc: coefficients=(6.802e+6, -2.326e+5, 1.737e+3, -1.24e+0)

Trong cả 3 mô hình dự đoán: SVR, Polynomial Regression và Bayesian Ridge Regression đều sử dụng để dự đoán số ca nhiễm chỉ trong 15 ngày tiếp theo vì số ca nhiễm tăng hay giảm còn phụ thuộc vào nhiều yếu tố khác như khả năng tiêm ngừa, mức an sinh, điều kiện kinh tế và thói quen sinh hoạt của từng khu vực trên Thế giới.

## Kết quả dự đoán và so sánh

|  |  |  |  |
| --- | --- | --- | --- |
| Ngày | SVR | Polynomial | Bayesian Ridge |
| 30/12/2021 | 299493324 | 272985695 | 273070005 |
| 31/12/2021 | 300325500 | 273348117 | 273433573 |
| 01/01/2022 | 301158850 | 273708740 | 273795348 |
| 02/01/2022 | 301993374 | 274067556 | 274155325 |
| 03/01/2022 | 302829073 | 274424558 | 274513495 |
| 04/01/2022 | 303665947 | 274779737 | 274869851 |
| 05/01/2022 | 304503995 | 275133087 | 275224386 |
| 06/01/2022 | 305343218 | 275484600 | 275577092 |
| 07/01/2022 | 306183615 | 275834268 | 275927962 |
| 08/01/2022 | 307025187 | 276182085 | 276276988 |
| 09/01/2022 | 307867933 | 276528043 | 276624163 |
| 10/01/2022 | 308711854 | 276872133 | 276969480 |
| 11/01/2022 | 309556950 | 277214350 | 277312931 |
| 12/01/2022 | 310403220 | 277554685 | 277654508 |
| 13/01/2022 | 311250665 | 277893131 | 277994205 |

Bảng 4‑1. Bảng so sánh kết quả dự đoán

Bảng 4-1 là bảng kết quả dự đoán của 3 thuật toán SVR, Polynomial Regression và Bayesian Ridge Regression. Thời gian dự đoán là 15 ngày từ 30/12/2021 đến 13/1/2022. Với thuật toán SVR, số ca nhiễm dự đoán tăng từ 299.5 triệu đến 311.2 triệu với độ dao động ∆=11.757.341, trung bình 783.822 ca nhiễm/ngày. Với thuật toán Polynomial Regression, số ca nhiễm dự đoán tăng từ 273 triệu đến 277.9 triệu với độ dao động ∆=4.907.436, trung bình 327.162 ca nhiễm/ngày. Với thuật toán Bayesian Ridge Regression, số ca nhiễm dự đoán tăng từ 273 triệu đến 278 triệu với độ dao động ∆=4.924.200, trung bình 328.280 ca nhiễm/ngày. Thuật toán SVR là thuật toán cho ra dự đoán có số ca nhiều nhất gấp 2 lần so với Polynomial Regression và Bayesian Ridge Regression.

## Áp dụng Polynomial Regression dự đoán số ca nhiễm ở Việt Nam

Trong phần này, em sử dụng thuật toán Polynomial Regression để dự đoán về số ca nhiễm của Việt Nam trong 30 ngày tiếp theo.



Hình 4‑14. Dự đoán số ca nhiễm ở Việt Nam

Hình 4-14 là biểu đồ đường thể hiện kết quả dự đoán số ca nhiễm trong 30 ngày tiếp theo của Việt Nam sử dụng thuật toán Polynomial Regression. Đường màu xanh là số ca nhiễm thực tế và đường màu đỏ gạch đứt là số ca nhiễm dự đoán.

Thông số kỹ thuật (metric) của mô hình thuật toán Polynomial Regression:

* Lỗi trung bình tuyệt đối (Mean Absolute Error): MAE=88066.58
* Lỗi bình phương trung bình (Mean Squared Error): MSE=12591878628
* R-Square=0.905
* Hệ số chặn: intercept=0
* Hệ số góc: coefficients=(-1.167e+5, 2.97e+3, -1.508e+1, 1.99e-2)
* y=f(x)= -1.167e+5+2.97e+3*x1*-1.508e+1*x12*+1.99e-2*x13*

|  |  |  |  |
| --- | --- | --- | --- |
| Ngày | Kết quả | Ngày | Kết quả |
| 30/12/2021 | 1489298 | 14/01/2022 | 1668546 |
| 31/12/2021 | 1500863 | 15/01/2022 | 1680941 |
| 01/01/2022 | 1512482 | 16/01/2022 | 1693392 |
| 02/01/2022 | 1524156 | 17/01/2022 | 1705899 |
| 03/01/2022 | 1535885 | 18/01/2022 | 1718463 |
| 04/01/2022 | 1547668 | 19/01/2022 | 1731083 |
| 05/01/2022 | 1559507 | 20/01/2022 | 1743760 |
| 06/01/2022 | 1571401 | 21/01/2022 | 1756494 |
| 07/01/2022 | 1583349 | 22/01/2022 | 1769285 |
| 08/01/2022 | 1595353 | 23/01/2022 | 1782133 |
| 09/01/2022 | 1607413 | 24/01/2022 | 1795039 |
| 10/01/2022 | 1619528 | 25/01/2022 | 1808001 |
| 11/01/2022 | 1631699 | 26/01/2022 | 1821021 |
| 12/01/2022 | 1643925 | 27/01/2022 | 1834099 |
| 13/01/2022 | 1656208 | 28/01/2022 | 1847234 |

Bảng 4‑2. Bảng kết quả dự đoán số ca nhiễm của Việt Nam

Bảng 4-2 là bảng kết quả dự đoán sử dụng thuật toán Polynomial Regression. Thời gian dự đoán là 30 ngày từ 30/12/2021 đến 28/01/2022. Số ca nhiễm dự đoán tăng từ 1.489 triệu đến 1.847 triệu với độ dao động ∆=357.936, trung bình 11.931 ca nhiễm/ngày.

# KẾT LUẬN

## Kết luận

Dịch COVID-19 đang diễn ra 1 cách phức tạp và khó lường ở khắp mọi nơi trên Thế giới, Châu Âu và Châu Á là 2 châu lục bị ảnh hưởng nhiều nhất cùng với số ca nhiễm và số ca tử vong rất cao. Tiếp đến là khu vực Bắc Mỹ cũng có só ca nhiễm và tử vong cao, tuy nhiên Mỹ là quốc gia chiếm tỷ trọng cao nhất về số ca nhiễm (chiếm hơn 80%) và số ca tử vong (chiếm gần 70%). Bên cạnh số ca nhiễm và tử vong cao thì số ca hồi phục cũng là 1 tín hiệu tốt khi đã có gần 90% số người mắc bệnh đã hồi phục. Tuy số ca nhiễm cao nhưng tỷ lệ tử vong của dịch COVID-19 so với các đại dịch khác là nhỏ hơn nhiều, khi mà tỷ lệ tử vong chỉ 1.78% (so với dịch SARS-2003 có tỷ lệ tử vong 10%, dịch MERS-2012 có tỷ lệ tử vong 34.4%, dịch Ebola có tỷ lệ tử vong 43.9%). Sự nguy hiểm của dịch COVID-19 đó là tốc độ lây lan rất nhanh và chuyển biến xấu khiến cho cả Thế giới lo sợ và hoang mang, tuy nhiên cùng với sự phát triển của Vắc-xin COVID-19 đã đem đến sự an tâm và giúp cho Thế giới trở lại được trạng thái “bình thường mới”.

Việt Nam cũng là quốc gia bị ảnh hưởng không ít bởi đại dịch COVID-19 khi số ca nhiễm và cả tử vong ngày một tăng và đã trải qua nhiều đợt giãn cách xã hội nhằm truy vết và ngăn chặn sự lây lan của dịch. Ở khu vực châu Á, Việt Nam thuộc top 10 quốc gia bị ảnh hưởng chỉ đứng sau Ấn Độ, Iran, Nhật Bản … và ở khu vực Đông Nam Á, Việt Nam thuộc top 5 quốc gia có số ca nhiễm và tử vong cao trong khu vực. Bên cạnh đó, Việt Nam đang thực hiện rất tốt về quy định phòng chống dịch, cũng như khả năng ứng phó và điều trị nhanh của các lực lượng tuyến đầu.

Để có được trạng thái “bình thường mới” như hiện nay, Vắc-xin đóng vai trò rất quan trọng. Vắc-xin đã được triển khai và sử dụng bởi rất nhiều quốc gia trên Thế giới, trong đó có Việt Nam. Tỷ lệ tiêm ngừa (1 mũi và 2 mũi) của các quốc gia tình tới thời điểm hiện tại là rất cao (từ 70-150% dân số). Việt Nam đang triển khai việc tiêm Vắc-xin nhanh khi đã có hơn 130% dân số đã được tiêm ngừa so với khu vực châu Á nói chung và Đông Nam Á nói riêng.

* Ưu điểm: Đồ án này giúp em có cái nhìn rõ hơn về đại dịch COVID-19 đang diễn ra về mặt tổng quát và chi tiết, so sánh được Việt Nam với các quốc gia ở Châu Á và khu vực Đông Nam Á, theo dõi được tình trạng tiêm chủng Vắc-xin COVID-19 của Thế giới nói chung, giữa Việt Nam và Châu Á, Đông Nam Á nói riêng. Nhờ vào các thuật toán Machine Learning, em có thể xây dựng được mô hình dự đoán về số ca nhiễm trong thời gian gần của Thế giới và Việt Nam.
* Nhược điểm: Do số lượng dữ liệu đầu vào không đủ lớn, dẫn đến sai số khiến cho kết quả dự đoán không được chính xác như mong đợi.

## Hướng phát triển

Trong đề tài này, hướng phát triển của em là:

* Sử dụng ngôn ngữ Python để tạo ra các báo cáo bằng file pdf.
* Gửi email cho những người đã được cài đặt sẵn thông qua giao thức SMTP.
* Sử dụng Power BI để tạo ra các Dashboard tự động cập nhật.
* Triển khai các báo cáo bằng các nền tảng web/app cho Desktop và Mobile.

# TÀI LIỆU THAM KHẢO

**Tiếng Việt**

Vũ Hữu Tiệp (2018), *Machine Learning cơ bản.*

**Tiếng Anh**

D. J. C. MacKay(1992),*Bayesian Interpolation, Computation and Neural Systems, Vol. 4, No. 3*

M. E. Tipping (2001), *Sparse Bayesian Learning and the Relevance Vector Machine,* Journal of Machine Learning Research, Vol. 1

Platt, John (1999), *Probabilistic outputs for support vector machines and comparison to regularizedlikelihood methods.*

Chih-Chung Chang and Chih-Jen Lin (2001), *A Library for Support Vector Machines.*

Andreas C. Müller and Sarah Guido (2016), *Introduction to Machine Learning with Python: A Guide for Data Scientists,* O'Reilly Media*.*

O Theobald (2017), *Machine Learning For Absolute Beginners: A Plain English Introduction (Second Edition) (Machine Learning From Scratch Book 1)*

O Theobald (2019), *Machine Learning with Python: A Practical Beginners’ Guide (Machine Learning From Scratch Book 2).*

O Theobald (2017), *Data Analytics for Absolute Beginners: Make Decisions Using Every Variable: (Introduction to Data, Data Visualization, Business Intelligence & Machine Learning).*

Brady Ellison (2021), *Python for Data Analysis: A Beginners Guide to Master the Fundamentals of Data Science and Data Analysis by Using Pandas, Numpy and Ipython.*

Wes McKinney (2012), *Python for Data Analysis: A Beginners Guide to Master the Fundamentals of Data Science and Data Analysis by Using Pandas, Numpy and Ipython.*

Cole Nussbaumer Knaflic (2015), *Storytelling with Data: A Data Visualization Guide for Business Professionals*, Wiley; 1st edition.

1. MÃ NGUỒN PYTHON

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn.