MỤC LỤC

Thông tin nhóm

MSSV	Họ Tên
20120234	Ngô Nguyễn Quang Tú
20120236	Phạm Tấn Anh Vũ
20120237	Hà Nguyễn Thảo Vy
20120194	Nguyễn Hữu Thiện

Mức độ hoàn thành

Mức độ hoàn thành tổng thể của mỗi yêu cầu

Yêu cầu	Tỉ lệ hoàn thành	Sinh viên thực hiện
Viết báo cáo	100%	20120234, 20120194
Tiền xử lý dữ liệu	98%	20120234, 20120236, 20120194
Phân phối dữ liệu	96%	20120234, 20120237
Trực quan hoá dữ liệu	100%	20120234, 20120236, 20120237
Thuật toán Machine Learning	100%	20120234

Mức độ hoàn thành của từng thành viên

MSSV	Mức độ hoàn thành
20120234	98%
20120236	97%
20120237	97%
20120194	95%

Thu thập dữ liệu

1. Ngữ cảnh, câu chuyện gì khiến nhóm sinh viên thực hiện việc tìm kiếm dữ liệu?

- Nhóm quan tâm đến tình hình dịch bệnh COVID-19 tại châu Á và muốn tìm kiếm dữ liệu liên quan đến số ca nhiễm và số người chết do COVID-19 trong khu vực này để phục vụ cho mục đích nghiên cứu và phân tích, cụ thể là tập dữ liệu COVID-19 tại Ấn Độ nơi mà đang là quốc gia có số ca tử vong lớn nhất khu vực châu Á.

2. Dữ liệu mà nhóm sinh viên là về chủ đề gì và được lấy từ nguồn nào?

- Dữ liệu được nhóm sinh viên chọn là "COVID-19 in India" từ trang Kaggle datasets: https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india.
- Bộ dữ liệu này cung cấp thông tin về số ca nhiễm COVID-19 ở Ấn Độ theo các bang và quận từ ngày 30/01/2020 đến 11/08/2021.

3. Người ta có cho phép sử dụng dữ liệu như thế này hay không? Ví dụ: cần kiểm tra thử License của dữ liệu là gì?

- Bộ dữ liệu "COVID-19 in India" có giấy phép CC0 1.0 Universal, nghĩa là dữ liệu có sẵn cho mọi người sử dụng, chỉnh sửa và chia sẻ mà không cần yêu cầu sự cho phép hay trả tiền nên hoàn toàn có quyền sử dụng bộ dữ liệu này.

4. Người ta đã thu thập dữ liệu này như thế nào? Phương pháp thực hiện là gì?

- Bộ dữ liệu "COVID-19 in India" từ Kaggle là tập hợp các báo cáo chính thức của Chính phủ Ấn Độ và các cơ quan y tế địa phương. Các số liệu được thu thập thông qua hệ thống Sức khỏe Đối tác Ấn Độ (IDSP) và được cập nhật hàng ngày. Các trường dữ liệu, bao gồm số ca nhiễm xác nhận, số người hồi phục, và số ca tử vong, được cập nhật theo ngày và theo bang/quận.

Khám phá dữ liệu (thường đan xen với pha tiền xử lý dữ liệu)

1. Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?

- Mỗi dòng trong tập dữ liệu "COVID-19 in India" có ý nghĩa là một bản ghi về số ca nhiễm COVID-19 trong một bang hoặc một quận của Ấn Độ vào một ngày cụ thể. Mỗi bản ghi bao gồm các thông tin sau:
  - `State/UnionTerritory`: tên bang hoặc quận.
  - `Confirmed`: số ca nhiễm xác nhận COVID-19.
  - `Deaths`: số ca tử vong do COVID-19.
  - `Recovered`: số ca đã hồi phục sau khi mắc COVID-19.
  - `Date`: ngày bản ghi được cập nhật.
- Không có vấn đề về các dòng có ý nghĩa khác nhau trong tập dữ liệu này. Tất cả các dòng đều có ý nghĩa giống nhau và cùng cấu trúc.

2. Mỗi cột có ý nghĩa gì?

- Mỗi cột trong tập dữ liệu "COVID-19 in India" có ý nghĩa như sau:
  - `State/UnionTerritory`: tên bang hoặc quận.
  - `Confirmed`: số ca nhiễm xác nhận COVID-19.
  - `Deaths`: số ca tử vong do COVID-19.
  - `Recovered`: số ca đã hồi phục sau khi mắc COVID-19.
  - `Date`: ngày bản ghi được cập nhật.
- Các cột này cung cấp thông tin về tình hình dịch bệnh COVID-19 tại các bang hoặc quận khác nhau của Ấn Độ trong suốt thời gian từ ngày đầu tiên ghi nhận ca nhiễm đến thời điểm cuối cùng của tập dữ liệu.

3. Mỗi cột hiện đang có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp hay không?

- Mỗi cột trong tập dữ liệu "COVID-19 in India" hiện đang có kiểu dữ liệu như sau:
  - `State/UnionTerritory`: chuỗi ký tự (string).
  - `Confirmed`: số nguyên (integer).
  - `Deaths`: số nguyên (integer).
  - `Recovered`: số nguyên (integer).
  - `Date`: chuỗi ký tự (string).
- Không có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp. Tuy nhiên, cột `Date` hiện đang có kiểu dữ liệu chuỗi ký tự (string) nên sẽ cần chuyển đổi thành kiểu dữ liệu ngày-tháng (date) để có thể thực hiện được một số thao tác phân tích thời gian.

4. Với mỗi cột, các giá trị (dạng số, dạng phân loại) được phân bố như thế nào?

- Với mỗi cột trong tập dữ liệu "COVID-19 Cases", các giá trị được phân bố như sau:
  - Cột `State/UnionTerritory`: là cột phân loại (categorical) vì nó chứa tên các bang hoặc quận ở Ấn Độ. Tên của mỗi bang hoặc quận là duy nhất và không bị trùng lặp.
  - Cột `Confirmed`, `Deaths`, và `Recovered`: là các cột dạng số (numeric) vì chúng chứa số lượng ca nhiễm, số lượng ca tử vong và số lượng ca hồi phục tương ứng tại mỗi bang hoặc quận của Ấn Độ. Giá trị trong các cột này là các số nguyên dương.
  - Cột `Date`: là cột phân loại (categorical) vì chúng chứa các giá trị ngày tháng dưới dạng chuỗi ký tự (string). Mỗi giá trị của cột Date là một ngày khác nhau trong thời gian từ ngày đầu tiên ghi nhận ca nhiễm đến thời điểm cuối cùng của tập dữ liệu.

5. Có cần phải tiền xử lý dữ liệu hay không và nếu có thì nhóm sinh viên cần phải xử lý như thế nào?

- Có thể cần phải tiền xử lý dữ liệu để chuẩn bị dữ liệu cho việc phân tích và khám phá thêm. Các xử lý cần được thực hiện như sau:

  - Chuyển đổi kiểu dữ liệu cột `Date` từ chuỗi ký tự (string) sang kiểu dữ liệu ngày-tháng (date).
  - Kiểm tra dữ liệu cột `State` và xử lý các dữ liệu bị dính ký tự rác và lọc thông tin các thông tin cùng một bang nhưng sai chính tả
  - Kiểm tra và xử lý dữ liệu khuyết (nếu có).
  - Xoá các cột thừa như `Time`, `Sno`,`ConfirmedIndianNational`,`ConfirmedForeignNational`
- Sau khi tiền xử lý xong, các bước khám phá và phân tích dữ liệu có thể được thực hiện để tìm ra những thông tin quan trọng từ dữ liệu này.

Khám phá mối quan hệ trong dữ liệu

- Xem chi tiết trong phần 5 **Data Visulization** của file notebook Covid-19 in India

Chi tiết Thuật toán

- Xem chi tiết trong phần 6 **Machine Learning Algorithmns Applications** của file notebook Covid-19 in India

Tài liệu tham khảo

- Nguyen Bao Long, Le Nhut Nam, Nguyen Ngoc Duc; Lab01 - Linear Regression (Julia); Môn học: Intro to ML (2023)