Data Science - Final Project

18120296 cuong091200 18120339 viplazylmht Open in Colab Github

Quy trình làm việc

28/12/2020

Mục tiêu

- Xác định chủ đề sẽ thực hiện.
- Thu tập dữ liệu về chủ đề đó.

Công việc

- Cả nhóm cùng thảo luận và đưa ra chủ đề muốn thực hiện.
- Duy: tìm API về chủ đề đó.
- Cường: tìm hiểu cách lấy dữ liệu từ api đó, đưa file notebook lên github.

Kết quả

- Chủ đề thực hiện liên quan đến dữ liệu `covid-19` của nước Mỹ.
- Lấy được đầy đủ dữ liệu từ API.

29/12/2020

Mục tiêu

- Đặt được câu hỏi.
- Khám phá dữ liệu.

Công việc

- Cường: Khám phá dữ liệu vừa thu thập được làm tiền đề cho việc đặt câu hỏi.
- Duy: dựa vào những kết quả của bước khám phá dữ liệu đặt câu hỏi liên quan đến dữ liệu vừa thu thập được, đồng thời nêu ra các ý nghĩa nếu như trả lời được câu hỏi đó.

Kết quả

- Đưa ra câu hỏi cần trả lời: Số ca mắc mới trong một ngày bất kì (hợp lệ) là bao nhiêu? Nó phụ thuộc như thế nào vào số ca mắc của những ngày trước đó?

30/12/2020

Mục tiêu

- Bộ dữ liệu thô phía trên chưa có sự kết nối giữa những ngày liền kề nhau nên nhóm muốn tạo ra sự liên kết giữa các ngày liền kề nhau.

Công việc

- Duy: Viết code tạo thêm thuộc tính mới cho tập dữ liệu (số ca mắc của `i` ngày trước của mỗi dòng dữ liệu sẽ được thêm vào cột `new_case_his_i`.
- Cường: Tìm hiểu Lab3 của thầy để hiểu rõ hơn các bước tiền xử lý và mô hình hóa dữ liệu.

Kết quả

- Tạo ra một bộ dữ liệu mới trong đó mỗi dòng có thêm các thuộc tính là số lượng ca nhiễm mới của ngày thứ `i` trước đó (thu thập 28 ngày).

31/12/2020

Mục tiêu

- Chia tập train, validation, test.
- Xây dựng pipeline để tiền xử lý dữ liệu.

Công việc

- Duy: Viết 1 pipeline để xóa các cột không cần thiết, điền missing value (với mỗi thuộc tính có cách điền khác nhau), và tiêu chuẩn hóa dữ liệu.
- Cường: Đề xuất giảm chiều dữ liệu bằng cách tạo ra 1 thuộc tính mới là tổng của tất các các thuộc tính được thêm vào ở ngày 30/12 và xóa các thuộc tính đó đi.

Kết quả

- Hoàn thành bước tiền xử lý dữ liệu.

01/01/2021

- Nghỉ tết dương lịch.

02/01/2021

Mục tiêu

- Tìm kiếm mô hình phù hợp với bài toán để bắt đầu đào tạo mô hình.

Công việc

- Cả nhóm tìm hiểu các mô hình và các siêu tham số của mô hình. Phải hiểu rõ lý thuyết của mô hình và thực hiện việc đào tạo mô hình.

Kết quả

- Nhóm chọn được 2 mô hình đó là MLP regressor và Linear Regression.
- Nhưng việc đào tạo không diễn ra thuận lợi (không tìm được mô hình tốt).

03/01/2021

Mục tiêu

- Tìm ra nguyên nhân tại sao mô hình không fit được tập train.
- Tìm cách khắc phục vấn đề trên.

Công việc

- Cường: tìm ra được nguyên nhân do tập train có chứa ngoại lai và đã xóa dòng đó đi.
- Duy: xây dựng bộ dữ liệu test đối với cả nước Mỹ (thêm một tập nữa để đánh giá mô hình).

Kết quả

- Khắc phục được vấn đề ở ngày 02/01.

05/01/2021

Mục tiêu

- Đánh giá lại đồ án.

Công việc

- Duy: Huấn luyện nhiều mô hình đồng thời chọn ra siêu tham số tốt nhất, vẽ biểu đồ, viết slide báo cáo.
- Cường: Nhận xét biểu đồ và đưa ra mô hình tốt nhất cho bài toán, cập nhật .

Kết quả

- Hoàn thành các nội dung huấn luyện và nhận xét.
- Phát hiện yếu điểm ở bước tiền xử lí dữ liệu: Việc xóa các dòng dữ liệu có giá trị âm đi là chưa hợp lý.

07/01/2021

Mục tiêu

Tìm giải pháp giải quyết vấn đề đặt ra ở ngày 05/01.

Công việc

- Cường: Đề xuất phương pháp giải quyết các dòng dữ liệu có giá trị âm.
- Duy: Cài đặt hàm để giải quyết bài toán trên.

Kết quả

- Giải quyết xong phần tiền xử lý dữ liệu.

08/01/2021

Mục tiêu

Thêm phần trực quan kết quả dự đoán.

Công việc

- Duy: Trực quan bộ dữ liệu test và bộ dữ liệu Mỹ
- Cường: Cài đặt và trực quan phần dự đoán khả năng diễn tiến của dịch bệnh trong tương lai.

Kết quả

- Vẽ được các biểu đồ cần thiết phục vụ cho việc đánh giá mô hình.
- Làm rõ ý nghĩa của câu hỏi (dự đoán tương lai).

09/01/2021

Mục tiêu

Kiểm tra và tối ưu lại mô hình.

Công việc

- Cường: Đề xuất phương pháp đánh giá độ lỗi mới (MAE) dùng trong huấn luyện và đánh giá.
- Duy: Kiểm tra lại các tập dữ liệu, các siêu tham số và nhận xét.

Kết quả

- Loại bỏ ý tưởng MSE vì dữ liệu đã phù hợp (đã được xử lý ở ngày 03/01).
- Chia lại tập dữ liệu theo tỉ lệ thích hợp hơn (70% 15% 15%).

11/01/2021

Mục tiêu

Hoàn thành đồ án.

Công việc

- Cường: Kiểm tra lại mô hình, các siêu tham số và nhận xét.
- Duy: Cập nhật nội dung file quy trình làm việc và file báo cáo.

Kết quả

- Hoàn thành đồ án.
- Sẵn sàng mở issues để cho phép các nhóm khác nhận xét.

15/01/2021

Mục tiêu

- Hoàn thành đồ án dựa trên đóng góp của nhóm khác.
- Thêm hệ số tương quan để giải thích cho việc lựa chọn thuộc tính cho mô hình.
- Biểu diễn cụ thể hơn về dữ liệu và làm mượt dữ liệu.

Công việc

- Cường: Kiểm tra lại mô hình, nhận xét thêm (về cả hệ số tương quan).
- Duy: Thay biểu đồ `boxplot` bằng `scatter` để biểu diễn dữ liệu huấn luyện trực quan hơn, cập nhật nội dung file quy trình làm việc và file báo cáo.

Kết quả

Hoàn thành đồ án.