ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN THỰC HÀNH: NHẬP MÔN KHOA HỌC DỮ LIỆU

Môn Học: Nhập môn khoa học dữ liệu

GVHD:

Thầy Trần Đại Chí - Thầy Nguyễn Bảo Long

Thầy Lê Nhựt Nam - Thầy Nguyễn Thái Vũ

Nhóm: 16

Danh sách sinh viên thực hiện:

1. Phan Phong Lưu 20120326
2. Bùi Thanh Tùng 20120398
3. Đỗ Tấn Tài 20120408
4. Trần Khắc Bình 20120437

Thành phố Hồ Chí Minh - 2022

Phần 1: Bảng phân công công việc

STT	Công việc	Người thực hiện	Ngày hoàn thành	Nội dung
1	Tìm nguồn dữ liệu, đề tài.	Tùng, Tài, Bình, Lưu	6-11-2022	Chọn đề tài Causes of death
2	Thu thập dữ liệu.	Tùng, Bình	20-11-2022	Dùng selenium
3	Tiền xử lý, khám phá dữ liệu	Tài, Lưu	1-12-2022	Thực hiện như các bước trong file hướng dẫn đồ án
4	Đặt câu hỏi 1	Tài	5-12-2022	Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?
5	Đặt câu hỏi 2, 3	Tùng	5-12-2022	Số ca tử vong do các bệnh truyền nhiễm ở các khu vực? Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?
6	Đặt câu hỏi 4	Bình	5-12-2022	Diễn biến của các nguyên nhân gây tử vong theo thời gian
7	Đặt câu hỏi 5, 6	Lưu	5-12-2022	Mối quan hệ giữa một số nguyên nhân tử vong So sánh phân phối tuổi tử vong
8	Mô hình hóa	Tài, Bình	13-12-2022	Hồi quy tuyến tính
9	Mô hình hóa	Tùng, Lưu	13-12-2022	Hồi quy đa thức
10	Làm slide	Tùng, Tài, Bình, Lưu	14-12-2022	
11	Viết báo cáo	Tùng, Tài, Bình, Lưu	14-12-2022	

Kế hoạch và lịch trình công việc: https://painted-almond-8b8.notion.site/Nh-p-M-n-Khoa-H-c-D-Li-u-a26856d28ea34bdebeb902b84d82f754

Phần 2: Đánh giá công việc

STT	Họ và tên	Khó khăn mắc phải	Bài học rút ra
1	Bùi Thanh Tùng	Khó khăn trong việc xây dựng mô hình vì tập dữ liệu chúng em thu thập chỉ có 1 trường là numeric.	Nên thu thập đa dạng dữ liệu trước khi giải quyết một bài toán.
2	Đỗ Tấn Tài	Vì tập dữ liệu chỉ nói về một trường thông tin là số ca tử vong nên việc đặt câu hỏi bị khó khăn, dễ bị trùng lặp.	Lựa chọn dữ liệu có tính đa dạng hơn.
3	Trần Khắc Bình	Dữ liệu quá nhiều dòng nên việc thu thập tốn rất nhiều thời gian. Mặc khác, dữ liệu có ít dòng nên không có nhiều thứ để khác thác cho phần đặt câu hỏi và mô hình hóa	Lần sau sẽ thu thập dữ liệu có số dòng và cột phù hợp hơn.
4	Phan Phong Lưu	Khó khăn trong việc tìm câu hỏi vì một số quốc gia không có số liệu.	Tìm cách điền đầy đủ dữ liệu, thu thập dữ liệu ít bị thiếu hơn.

Nếu có nhiều thời gian hơn, nhóm em sẽ:

- Cùng nhau bàn bạc, thảo luận nhiều hơn để tìm hiểu tường minh thông tin dữ liệu mà nhóm đã thu thập. Từ đó tối đa hóa việc xử lý các vấn đề phát sinh trong việc tính toán và tiền xử lý dữ liệu. Loại bỏ các dữ liệu gây nhiễu hoặc các dữ liệu không có giá trị về mặt ý nghĩa.
- Đưa ra các câu hỏi mang tính sâu sắc hơn nhằm phân tích hiệu quả giá trị mà dữ liệu đã cung cấp.
- Hoàn thiện file notebook, tuân thủ clear-coding, viết code ngắn gọn và đơn giản nhất có thể, chú thích tường minh cho từng bước xử lý.
- Tìm hiểu và làm kỹ hơn về phần mô hình hóa.

Tài liệu tham khảo

[1]. Tamas Ujhelyi, (2021), *Polynomial Regression in Python using scikit-learn (with a practical example)*, truy cập ngày 10/12/2022 tại https://data36.com/polynomial-regression-python-scikit-learn/

[2]. AlindGupta, (2022), *Python | Linear Regression using sklearn*, truy cập ngày 10/12/2022 tại https://www.geeksforgeeks.org/python-linear-regression-using-sklearn