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
LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

NHÓM 1

Nhóm trưởng	20120547 – Võ Thành Phong
Thành viên	20120128 – Nguyễn Thị Cẩm Lai
Thành viên	20120489 – Võ Phi Hùng
Thành viên	20120125 – Bùi Anh Kiệt

Giáo viên hướng dẫn: Thầy Lê Nhựt Nam.

Thành phố Hồ Chí Minh – 12/2022

MỤC LỤC

I. Phân công công việc

1. Phân công

Nhiệm vụ	Chi tiết	Thời gian
Thu thập dữ liệu	Tìm kiếm tập dữ liệu phù hợp.	13/11/2022 8:00 AM – 15/11/2022 11:59 PM
	Mô tả về chủ đề của bộ dữ liệu, nguồn gốc của bộ dữ liệu, cách bộ dữ liệu được thu thập.	
	Tổng quan về bộ dữ liệu: số lượng thuộc tính, giới thiệu từng thuộc tính.	
	Trình bày trên file jupyter notebooks ở dạng markdown.	
Khám phá dữ liệu 1	Một số code tổng quan cho bộ dữ liệu như .shape(), .head(), .describe(), unique(),. isnull().sum()….	16/11/2022 8:00 AM – 17/11/2022 11:59 PM
	Có vấn đề các dòng có ý nghĩa khác nhau không? Mỗi cột hiện đang có kiểu dữ liệu gì?	
	Trình bày trên file jupyter notebook.	
Khám phá dữ liệu 2	Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp hay không? Giải thích?	18/11/2022 8:00 AM – 19/11/2022 11:59 PM
	Chuyển về kiểu dữ liệu phù hợp nếu cần thiết. Giải thích lý do chọn kiểu dữ liệu.	
	Tiền xử lý dữ liệu: Loại bỏ giá trị nan, thay đổi tên viết tắt, ….	
	Trình bày trên file jupyter notebook.	
Khám phá dữ liệu 3	Với các cột kiểu dữ liệu dạng numeric, các giá trị được phân bố như thế nào (Phần trăm giá trị thiếu, giá trị lớn nhất, giá trị nhỏ nhất).	20/11/2022 8:00 AM – 22/11/2022 11:59 PM
	Với các cột kiểu dữ liệu dạng categorical, các giá trị được phân bố như thế nào (Phần trăm giá trị thiếu, số lượng giá trị khác nhau không tính nan, danh sách các giá trị khác nhau không tính nan).	
	Trình bày trên file jupyter notebook.	
Đặt câu hỏi + Trực quan 1	Đặt ít nhất 3 câu hỏi. Trình bày rõ nếu trả lời được câu hỏi sẽ mang lại lợi ích gì.	24/11/2022 11:00 PM – 02/12/2022 11:59 PM
	Trực quan hóa để minh họa cho câu trả lời. Trình bày trên file jupyter.	
Đặt câu hỏi + Trực quan 2	Đặt ít nhất 3 câu hỏi. Trình bày rõ nếu trả lời được câu hỏi sẽ mang lại lợi ích gì.	24/11/2022 11:00 PM – 02/12/2022 11:59 PM
	Trực quan hóa để minh họa cho câu trả lời. Trình bày trên file jupyter.	
Báo cáo cá nhân	Viết báo cáo cá nhân trình bày: gặp những khó khăn gì, đã học được gì.	30/11/2022 11:59 PM – 02/12/2022 11:59 PM
	Nhóm trưởng viết phần nếu có thêm thời gian thì nhóm sẽ làm gì.	

Làm slide báo cáo	Làm toàn thể nội dung trong jupyter thành file .ppt để cho ngày vấn đáp. Mọi ý kiến phát sinh thêm thảo luận trên nhóm để cùng giải quyết.	03/12/2022 11:59 PM – 09/12/2022 11:59 PM
Tổng hợp thành file notebook hoàn chỉnh	Tổng hợp, thiết kế lại các file notebook của các giai đoạn thành một file hoàn chỉnh và commit lên nhánh main trên github của nhóm.	03/12/2022 11:59 PM – 09/12/2022 11:59 PM
Viết báo cáo	Viết báo cáo cuối cùng cho phân công công việc, cảm nhận của các thành viên và kế hoạch của nhóm nếu có thêm nhiều thời gian.	03/12/2022 11:59 PM – 09/12/2022 11:59 PM

2. Đánh giá

Nhiệm vụ	Thực hiện	Mức độ hoàn thành	Chưa làm được
Thu thập dữ liệu	Cẩm Lai	100%	Không
Khám phá dữ liệu 1	Phi Hùng	100%	Không
Khám phá dữ liệu 2	Cẩm Lai Anh Kiệt	100%	Không
Khám phá dữ liệu 3	Thành Phong	100%	Không
Đặt câu hỏi+Trực quan 1	Phi Hùng Anh Kiệt	100%	Không
Đặt câu hỏi+Trực quan 2	Cẩm Lai Thành Phong	100%	Không
Báo cáo cá nhân	Tất cả thành viên	100%	Không
Làm slide báo cáo	Phi Hùng Anh Kiệt	100%	Không
Tổng hợp thành file jupyter notebook hoàn chỉnh	Cẩm Lai	100%	Không
Viết báo cáo	Thành Phong	100%	Không

II. Báo cáo cá nhân

1. 20120547 – Võ Thành Phong

- Khó khăn gặp phải:

+ Ở nhiệm vụ khám phá dữ liệu: Do chưa thành thạo nhiều hàm trong các thư viện hỗ trợ như Pandas, Numpy, Matplotlib, Seaborn nên quá trình làm việc của bản thân mất khá nhiều thời gian để tìm hiểu những hàm muốn sử dụng để giải quyết các vấn đề được đặt ra.

+ Ở nhiệm vụ trực quan hóa dữ liệu:

* Nhận thấy việc tìm ý tưởng ban đầu cho phân tích dữ liệu và trực quan của bản thân chưa có tính sáng tạo.

* Có câu hỏi chưa thể trực quan bằng cách vẽ đồ thị mà chỉ có thể trực quan bằng cách tạo dataframe, chưa đặc sắc cho phần trực quan.

- Những kinh nghiệm học được:

+ Sử dụng notion tổ chức không gian làm việc nhóm.

+ Những trang web tìm kiếm dữ liệu uy tín, chẳng hạn như bạn Cẩm Lai giới thiệu cho em biết về kho lưu trữ học máy UCI (UCI Machine Learning Repository).

+ Cách lựa chọn một tập dữ liệu đã được công bố để có thể làm tốt các quy trình của đồ án khoa học dữ liệu: dữ liệu với số lượng mẫu lớn, số thuộc tính vừa phải không quá ít cũng không quá nhiều, bộ dữ liệu nên là dữ liệu thô chưa được làm sạch quá kĩ càng đểcó thể thấy được một cách chính xác nhất tình hình thực tế của vấn đề đăng sau bộ dữ liệu.

+ Những cú pháp, hàm của các thư viện từ các bạn trong nhóm.

+ Cách sử dụng sức mạnh của hàm apply: dùng apply trong tiền xử lý dữ liệu để tăng tốc độ khi làm việc với các dataframes.

+ Học hỏi cách dùng hệ màu khác nhau trực quan cho đồ thị.

+ Biết cách sử dụng đồ thị choropleth từ thư viện plotly do bạn Cẩm Lai thực hiện.

2. 20120128 – Nguyễn Thị Cẩm Lai

- Khó khăn gặp phải:

+ Còn bỡ ngỡ với các thư viên và công cụ nên khi thực hiện đồ án có gặp nhiều lúng túng gây mất thời gian để nghiên cứu.

+ Có quá nhiều đồ án của tất cả các môn trong kỳ học, việc phân chia thời gian để cân bằng mọi đồ án là một thách thức lớn.

- Những kinh nghiệm học được:

+ Cách sử dụng các thư viện và công cụ cần thiết một cách thuần thục hơn.

+ Biết cách sắp xếp thời gian để hoàn thành tốt các quy trình được giao trong đồ án.

+ Biết cách phối hợp làm việc với các thành viên trong nhóm để mang lại hiệu quả trong công việc.

3. 20120489 – Võ Phi Hùng

- Khó khăn gặp phải:

+ Vì không học môn Nhập môn Khoa học dữ liệu, nên bản thân cảm thấy cực kỳ thiếu kiến thức. Em đã phải tự tra cứu các thông tin cũng như hỏi các thành viên trong nhóm rất nhiều. Cũng may nhờ kiến thức qua 3 bài lap thực hành mà khó khăn này nhẹ nhàng đi rất nhiều.

+ Mặc dù deadline làm việc mà nhóm trường giao không hề gắt gao, nhưng vì phải thực hiện nhiều đồ án ở các môn khác cùng lúc nên em liên tục trễ deadline. Rất may, các

bạn trong nhóm đã nhắc nhở kịp thời nên không ảnh hưởng nghiêm trọng đến kế hoạch của nhóm.

- Những kinh nghiệm học được:

+ Về mặt kiến thức, em đã học được rất nhiều về quy trình khoa học dữ liệu. Từ việc thu thập, khám phá đến tiền xử lý, trực quan hoá dữ liệu. Đây là một trải nghiệm thú vị với một người chưa từng tiếp xúc với khoa học dữ liệu như em.

+ Về mặt kỹ năng, em được nâng cao khả năng làm việc nhóm và sử dụng GitHub và Jupyter Notebook.

+ Về mặt thái độ, em được rèn luyện một thái độ nghiêm túc hơn trong việc nộp bài đúng hạn cũng như thái độ tự học, tìm hiểu những kiến thức, kĩ năng không thuộc chuyên môn.

4. 20120125 – Bùi Anh Kiệt

- Khó khăn gặp phải:

+ Bản thân em, khi thực hiện đồ án, không cảm thấy khó khăn về mặt kiến thức do đã được trang bị đầy đủ qua những gì đã học trên lớp, các Lab 0, 1, 2 và 3 cũng như từ môn Nhập môn khoa học dữ liệu.

+ Tuy nhiên, đối với quá trình làm việc nhóm, em đã gặp phải khó khăn khi làm quen với Github vì đây là lần đầu tiên có cơ hội tiếp cận và sử dụng.

+ Bên cạnh đó, mặc dù kế hoạch cho từng nhiệm vụ đã được lập cẩn thận và phù hợp cho từng thành viên, nhưng em vẫn phải đợi nhóm trưởng và ứng dụng Notion nhắc nhở nhiều bởi không có sự sắp xếp hợp lý trong kế hoạch hoạt động cá nhân, may mà không làm chậm tiến độ của nhóm.

- Những kinh nghiệm học được:

+ Về mặt kiến thức, em có cơ hội và đã học hỏi được từ các bạn về cách sử dụng Seaborn để Trực quan hóa dữ liệu. Nhiều dạng biểu đồ mà mình chưa từng biết đến, cũng đã được tìm hiểu và áp dụng trong đồ án này.

+ Về mặt kỹ năng làm việc nhóm, em đã biết sử dụng Github, hiểu rõ cách để thực hiện tốt quy trình khoa học dữ liệu và hơn hết là có tinh thần luôn luôn bình tĩnh, khách quan, thành thật và trách nhiệm hơn như một nhà khoa học dữ liệu thực thụ.

III. Nếu có thêm thời gian nhóm sẽ làm gì

- Tập trung hơn nữa vào phần C và D: Đặt câu hỏi và phân tích, trực quan để trả lời câu hỏi:

+ Khai thác nhiều khía cạnh hơn nữa từ bộ dữ liệu như là cố gắng tìm hiểu về toàn bộ những câu chuyện có thể có về mối tương quan từng đôi một của các cặp thuộc tính trong bộ dữ liệu.

+ Cố gắng áp dụng những kĩ thuật tiền xử lý dữ liệu một cách cụ thể bằng cách áp dụng thêm một số kiến thức nâng cao của thư viện sklearn.

+ Trực quan bằng nhiều loại đồ thị khác nhau hơn nữa để phù hợp hơn về mặt biểu diễn đối với từng loại dữ liệu: liên tục, rời rạc, tỉ lệ phần trăm, …

+ Cố gắng tìm kiếm những câu hỏi để có thể trực quan đa biến nhiều hơn.