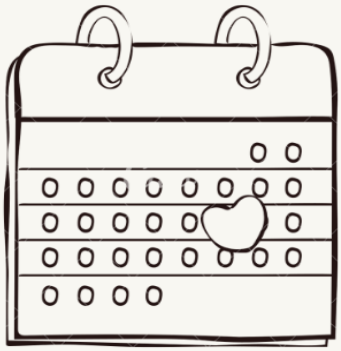
BÁO CÁO



PHÂN CÔNG CÔNG VIỆC VÀ TỔNG KẾT

ĐỒ ÁN THỰC HÀNH
NHẬP MÔN KHOA HỌC DỮ LIỆU

NHÓM 10

THỰC HIỆN

20120037 - Trần Thị Minh Anh
20120128 - Nguyễn Thị Cẩm Lai
20120166 - Nguyễn Dương Tuấn Phương
20120547 - Võ Thành Phong

MỤC LỤC

I. THÔNG TIN CHUNG

1. Thông tin thành viên

Họ và Tên	MSSV	Vai trò	Github
Nguyễn Thị Cẩm Lai	20120128	Nhóm trưởng	ntclai
Trần Thị Minh Anh	20120037	Thành viên	anhtran-chowmuaii
Võ Thành Phong	20120547	Thành viên	VTaPo
Nguyễn Dương Tuấn Phương	20120166	Thành viên	phuongnguyen1110

2. Các công cụ tổ chức quản lý

Link Github:

https://github.com/VTaPo/NMKHDL_Project_Gr10

Link Notion:

https://www.notion.so/bb88f97e2fc54c608375996db1463f2c?v=40ae32d5b2b840aca02eaa3f0490fd65

II. CHI TIẾT PHÂN CÔNG CÔNG VIỆC

Phần 1: Quản lý

Nhiệm vụ	Thành viên thực hiện	Thời gian
Tổ chức và quản lý Github	Võ Thành Phong	Toàn bộ quá trình thực hiện đồ án
Tổ chức và quản lý Notion	Nguyễn Thị Cẩm Lai	Toàn độ quá trình thực hiện đồ án

Phần 2: Quy trình Khoa học Dữ liệu

A. Thu thập dữ liệu (Data collection)

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

C. Đặt các câu hỏi có ý nghĩa cần trả lời

	Nhiệm vụ	Thành viên thực hiện	Thời gian
A	Lựa chọn chủ đề, nguồn cung cấp dữ liệu. (Các thành viên tự tìm hiểu và lựa chọn, sau đó thảo luận với nhau để đưa ra quyết định chung)	○ Trần Thị Minh Anh ○ Nguyễn Thị Cẩm Lai ○ Võ Thành Phong ○ Nguyễn Dương Tuấn Phương	Bắt đầu: 7/11 Kết thúc: 12/11
	Thu thập dữ liệu đã thống nhất theo cách thủ công bằng các công cụ hỗ (selenium, request hoặc API,...)	○ Trần Thị Minh Anh ○ Nguyễn Dương Tuấn Phương	Bắt đầu: 12/11 Kết thúc: 21/11
B	Thực hiện khám phá dữ liệu, đan xen với tiền xử lý dữ liệu (đáp ứng các nội dung được mô tả trong đồ án)	○ Võ Thành Phong ○ Nguyễn Thị Cẩm Lai	Bắt đầu: 22/11 Kết thúc 26/11

C	Đưa ra ít nhất 5 câu hỏi có thể được trả lời bằng dữ liệu này (đáp ứng nội dung được mô tả trong đồ án)	○ Trần Thị Minh Anh ○ Nguyễn Dương Tuấn Phương	Bắt đầu: 27/11 Kết thúc: 5/12

Phần 3: Mô hình hóa

Nhiệm vụ	Thành viên thực hiện	Thời gian
Mô hình hóa dữ liệu và đánh giá mô hình	○ Võ Thành Phong ○ Nguyễn Thị Cẩm Lai	Bắt đầu: 27/11 Kết thúc: 5/12

Phần 4: Báo cáo và tổng hợp

Nhiệm vụ	Thành viên thực hiên	Thời gian
Báo cáo cá nhân, trả lời các câu hỏi sau khi hoàn thành đồ án: ○ Bạn đã gặp những khó khăn gì? ○ Bạn đã học được gì? ○ Nhóm của bạn: Bạn sẽ làm gì nếu có nhiều thời gian hơn?	○ Trần Thị Minh Anh ○ Nguyễn Thị Cẩm Lai ○ Võ Thành Phong ○ Nguyễn Dương Tuấn Phương	Bắt đầu: 6/12 Kết thúc: 7/12
Tổng hợp và thiết kế slide báo cáo dạng .pdf	○ Trần Thị Minh Anh ○ Nguyễn Dương Tuấn Phương	Bắt đầu: 10/12 Kết thúc: 12/12
Thiết kế lại các file jupyter	○ Nguyễn Thị Cẩm Lai ○ Võ Thành Phong	Bắt đầu: 6/12 Kết thúc: 10/12
Làm file .pdf cho phân công công việc của nhóm/ tổng hợp báo cáo để đánh giá lại công việc	○ Nguyễn Thị Cẩm Lai	Bắt đầu: 6/12 Kết thúc: 8/12

III. ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH NHIỆM VỤ

Họ và tên	Mức độ hoàn thành	Phần chưa làm được
Trần Thị Minh Anh	100%	Không có
Nguyễn Thị Cẩm Lai	100%	Không có
Nguyễn Dương Tuấn Phương	100%	Không có
Võ Thành Phong	100%	Không có

IV. BÁO CÁO CÁ NHÂN

20120037 – Trần Thị Minh Anh	
Khó khăn gặp phải	- Ở nhiệm vụ Thu thập dữ liệu: o Gặp khó khăn trong việc cân nhắc lựa chọn giữa các công cụ (Selenium, BeautifulSoup, HTML Request) o Trong khi để bot chạy tự động, trong việc handle các đường link không mở được (chưa rõ nguyên nhân lúc được lúc không), nên đã bỏ qua một vài dòng dữ liệu có thể thu thập được. o Chưa tối ưu được thời gian khi crawl dữ liệu. Nói cách khác là thời gian chạy tương đối lâu nhưng số lượng dữ liệu thu được còn chưa cao. - Ở nhiệm vụ Trả lời câu hỏi: o Tự nhận thấy chưa tìm được những câu hỏi thực sự hay và có ý nghĩa. o Trong trực quan hóa dữ liệu còn lúng túng khi sử dụng các biểu đồ kết hợp (như grouped bar chart, box plot).

Kỹ năng học được	- Sử dụng các công cụ để crawl dữ liệu tự động từ trang web. - Những kinh nghiệm từ các bạn trong nhóm về việc thực hiện một đồ án KHDL: o Cách quản lý github repository, notion và phân công nhiệm vụ. o Trình bày một file ipynb trực quan, đẹp mắt: + Cách quản lý các đề mục + Sử dụng thêm hình vẽ trực quan, màu sắc thích hợp, tăng độ tập trung vào những mục cần chú ý. - Học cách sử dụng sklearn trong mô hình hóa dữ liệu và dự đoán giá chung cư.

20120128 - Nguyễn Thị Cẩm Lai	
Khó khăn gặp phải	- Mất nhiều thời gian cho việc tìm kiếm bộ dữ liệu phù hợp. Đa số trong các nguồn dữ liệu tìm được, nguồn phù hợp để phân tích khám phá các insight thì không phù hợp để xây dựng mô hình học máy và ngược lại. - Còn gặp nhiều khó khăn khi dùng các thư viện và công cụ nên khi thực hiện đồ án mất nhiều thời gian để nghiên cứu, đặt biết là ở phần mô hình hóa. - Có quá nhiều đồ án của tất cả các môn trong học kỳ, việc phân chia thời gian để cân bằng mọi đồ án là một thách thức lớn.

Kỹ năng học được	- Sử dụng Notion để tổ chức quản lý công việc chung cho nhóm (phân chia nhiệm vụ, cập nhật tiến độ). - Sử dụng Github để quản lý nội dung đồ án. - Học hỏi được nhiều kiến thức từ các thành viên khác trong nhóm. - Học được kỹ năng quản lý nhóm và phối hợp làm việc với các thành viên trong nhóm để mang lại hiệu quả cao, đáp ứng tiến độ đồ án qua các giai đoạn. - Khám phá được nhiều công cụ và thư viện khác nhau trong quá trình làm tiền xử lý dữ liệu, trực quan hóa và mô hình học máy. - Hiểu sâu hơn về các khái niệm trong mô hình học máy.

20120166 – Nguyễn Dương Tuấn Phương	
Khó khăn gặp phải	- Thu thập dữ liệu: o Tìm kiếm được trang web để lấy dữ liệu mà đáp ứng được đầy đủ các yêu cầu của đồ án. o Sử dụng những thư viện mới để giúp cho việc thu thập dữ liệu được nhanh hơn. - Đặt và trả lời câu hỏi: o Khó khăn trong việc đặt ra các câu hỏi có ý nghĩa từ dữ liệu minh đã thu thập được để đáp ứng nhu cầu của đồ án. o Dữ liệu được thu thập về đáp ứng được yêu cầu của đề án tuy nhiên một vài trường dữ liệu lại bị thiếu dữ liệu khá là nhiều và nó gây ảnh hưởng đến việc trực quan hoá dữ liệu khi trả lời các câu hỏi được đặt ra. - Học cách vẽ các biểu đồ khác ngoài bar chart. - Câu hỏi chưa thực sự sâu sắc và còn khá khuôn mẫu. - Ngoài ra còn khó khăn trong việc sử dụng Canva để làm slide báo cáo cho buổi thuyết trình.

Kỹ năng học được	- Chịu trách nhiệm với phần công việc mà mình đã được giao khi làm đồ án. - Quá trình làm việc nhóm hiệu quả nhờ phân công rõ ràng qua Notion. - Sử dụng các công cụ, thư viện mới để thu thập dữ liệu từ trang web. - Vẽ các biểu đồ thể hiện được đầy đủ thông tin cần thiết và cách đọc biểu đồ để từ đó rút ra được thông tin cần tìm. - Những kinh nghiệm từ các bạn trong nhóm: - Dùng GitHub để quản lý/chia sẻ các file ipynb cho đồ án. - Cách trình bày nội dung bên trong các file sao cho dễ hiểu, sạch đẹp. - Sử dụng sklearn trong mô hình hóa dữ liệu - Kỹ năng phân chia công việc nhóm của bạn nhóm trưởng sao cho khoa học và hợp lý.

20120547 – Võ Thành Phong	
Khó khăn gặp phải	- Ở nhiệm vụ khám phá dữ liệu: o 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. o Việc tìm hiểu những ý tưởng chính để khám phá còn chậm và dư thừa, có những phần làm bài bị dư thừa hoặc chưa hợp lí về mặt thứ tự và phải nhờ các thành viên khác chỉnh sửa. - Ở nhiệm vụ học máy:

	○ Do em chưa được học môn học môn nhập máy nên những kiến thức trình bày trong đồ án lần này đều là tự học. ○ Mất nhiều thời gian tìm hiểu và tìm hiểu kiến thức bằng tiếng Anh nên chưa thể hiểu sâu sắc, do tự học và nhóm cũng chưa có ai học qua các môm học về học máy nên cũng không chắc chắn 100% ở quá trình đánh giá tính đúng đắn. ○ Hàm chi phí còn cao và chưa thể tối ưu hơn nữa cho mô hình học máy.
Kỹ năng học được	- Sử dụng notion tổ chức không gian làm việc nhóm. - Những kinh nghiệm từ các bạn trong nhóm về việc thực hiện một đồ án KHDL: ○ Cách thu thập dữ liệu bằng thư viện BeautifulSoup và selenium từ hai bạn Minh Anh và Tuấn Phương. ○ Cách lựa chọn những thuộc tính cần thiết để thu thập nên một tập dữ liệu thô ban đầu. ○ 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. Nhiều cách vẽ dạng đồ thị khác nhau. ○ Cách tạo table content trên jupyter notebook và edit file jupyter notebook đẹp hơn từ bạn Cẩm Lai. ○ Tiền xử lý dữ liệu thiếu bằng thư viện sklearn cho phần học máy từ bạn Cẩm Lai.

V. NẾU CÓ THÊM THỜI GIAN NHÓM SẼ LÀM GÌ?

1. Thu thập dữ liệu

- Nhận thấy một khó khăn lớn trong bước đầu của nhóm là tìm kiếm một nguồn dữ liệu phù hợp để đáp ứng tất cả các nội dung đồ án đưa ra. Tuy nhiên trong quá trình tìm kiếm thì nhận thấy hầu hết các nguồn dữ liệu chỉ đáp ứng tốt một phần nào đó, chẳng hạn như: nguồn dữ liệu phù hợp cho phân tích khám phá dữ liệu để tìm ra các insight có ý nghĩa thì không phù hợp để xây dựng mô hình học máy và ngược lại.
- Do vậy, nếu có nhiều thời gian hơn thì nhóm sẽ tìm kiếm và cân nhắc kỹ lưỡng hơn để lựa chọn ra nguồn dữ liệu và chủ đề phù hợp nhất cho đồ án.

2. Đặt và trả lời các câu hỏi có ý nghĩa

- Cố gắng khai thác thêm nhiều khía cạnh hơn nữa từ bộ dữ liệu để tìm ra các câu chuyện bên trong, các mối tương quan của các cặp thuộc tính với nhau.
- Cố gắng áp dụng thêm nhiều kỹ thuật tiền xử lý dữ liệu bằng các công cụ khác như thư viện sklearn.
- Thực hiện trực quan hóa dữ liệu bằng nhiều loại đồ thị, thư viện khác nhau.

3. Mô hình hóa

- Nghiên cứu kĩ hơn về phần hồi quy đa biến khi đường tuyến tính không còn phù hợp với bộ dữ liệu.
- Học qua kiến thức về model selection để lựa chọn các thuộc tính cần thiết và bậc cho hồi quy đa thức.
- Thay vì sử dụng hàm PCA có sẵn từ sklearn để giảm chiều dữ liệu, nhóm sẽ thử tự cài đặt lại thuật toán trên.
- Tìm hiểu kĩ hơn về lớp Linear Regression của sklearn để có thể so sánh, thấy được những cách tối ưu kết quả hơn nữa thông qua việc điều chỉnh các hyper parameters mà thư viện cung cấp để có thể áp dụng vào các hàm tự cài đặt của nhóm.