Bản phân công, tổng kết công việc
Nhóm 15

I. Tổng quan về đồ án, các công cụ sử dụng

1. Mục tiêu đồ án:

Trong đồ án này, sinh viên được cung cấp những kiến thức cơ bản về khoa học dữ liệu. Từ đó đưa ra và giải quyết những vấn đề thực tế có thể trả lời được bằng dữ liệu, thông qua các quy trình: thu thập, tiền xử lý, phân tích thống kê, trực quan và mô hình hóa dữ liệu.

2. Chủ đề

Phân tích thị trường lao động Việt Nam

Nguồn dữ liệu: https://www.vietnamworks.com

3. Các công cụ sử dụng

Crawl dữ liệu: HTTP Requests

Thư viện học máy: Scikit Learn

Quản lý source code: GitHub

https://github.com/HoangPhuc0304/NMKHDL-2022.git

Quản lý lịch biểu: Notion

https://www.notion.so/886f73683c94463ba8a7ce46dce43048?v=e0d29a8932484efca4b07f3b3c924516

4. Khung thời gian

5 tuần (7/11/2022 – 11/12/2022)

II. Thông tin thành viên

STT	Họ tên	MSSV	Email
1	Phạm Hoàng Nam Anh	18120278	18120278@student.hcmus.edu.vn
2	Nguyễn Thị Châu Ngọc	20120146	20120146@student.hcmus.edu.vn
3	Lê Hoàng Phúc	20120549	20120549@student.hcmus.edu.vn
4	Nguyễn Hoàng Thịnh	20120587	20120587@student.hcmus.edu.vn

III. Bảng phân công công việc, đánh giá mức độ hoàn thành

1. Nội dung công việc

STT	Mã công việc	Tên công việc	Khung thời gian
1	T00	Chọn chủ đề	7/11 – 11/11
2	T01	Thu thập dữ liệu	15/11 – 18/11
3	T02	Khám phá, tiền xử lý dữ liệu	19/11 – 23/11
4	T03	Đặt câu hỏi	27/11 – 30/11
5	T04	Mô hình hóa dữ liệu	3/12 – 6/12
6	T05	Đánh giá mô hình	7/12 – 10/12
7	T06	Tổng hợp kết quả	11/12 – 14/12

2. Phân công công việc

STT	Họ tên	Mã công việc	Mô tả công việc	Mức độ hoàn thành

1	Phạm Hoàng Nam Anh	T00	Tìm hiểu, đề xuất các đề tài thú vị đem lại nhiều ý nghĩa	100%
		T01	Thu thập dữ liệu, cân nhắc sử dụng các công cụ phù hợp, lưu vào file .csv	100%
		T05	Đánh giá, so sánh 2 mô hình đã xây dựng ở T04	100%
2	Nguyễn Thị Châu Ngọc	T00	Tìm hiểu, đề xuất các đề tài thú vị đem lại nhiều ý nghĩa	100%
		T02	Khám phá đan xen tiền xử lý cho dữ liệu đã thu thập ở T01	100%
		T03	Đặt 2 câu hỏi ý nghĩa có thể trả lời được bằng dữ liệu, kết hợp trực quan hóa	100%
		T05	Đánh giá, so sánh 2 mô hình đã xây dựng ở T04	100%
3	Lê Hoàng Phúc	T00	Tìm hiểu, đề xuất các đề tài thú vị đem lại nhiều ý nghĩa	100%
		T03	Đặt 2 câu hỏi ý nghĩa có thể trả lời được bằng dữ liệu, kết hợp trực quan hóa	100%
		T04	Xây dựng mô hình hồi quy cho bài toán (dự đoán số lượng ứng viên đăng kí một công việc dựa trên các đặc trưng của công việc), sử dụng Random Forest Regression	100%
		T06	Tổng hợp kết quả, lập bảng phân công	100%
4	Nguyễn Hoàng Thịnh	T00	Tìm hiểu, đề xuất các đề tài thú vị đem lại nhiều ý nghĩa	100%
		T03	Đặt 2 câu hỏi ý nghĩa có thể trả lời được bằng dữ liệu, kết hợp trực quan hóa	100%

		T04	Xây dựng mô hình hồi quy cho bài toán (dự đoán số lượng ứng viên đăng kí một công việc dựa trên các đặc trưng của công việc), sử dụng Multiple Layer Perception Regression	100%
		T06	Tổng hợp slide báo cáo	100%

IV. Những bài học và khó khăn gặp phải

STT	Họ tên	Mã công việc	Học được	Những khó khăn
1	Phạm Hoàng Nam Anh	T00	Tìm hiểu được thêm các lĩnh vực thú vị. Cách để lựa chọn được một nguồn dữ liệu phù hợp.	Khó khăn khi tìm một nguồn phù hợp. Phải đánh đổi trong việc lựa chọn các nguồn. Có những trang web có rất nhiều thông tin thú vị nhưng không thể khai thác vì vấn đề bảo mật hay không phù hợp cho xây dựng mô hình học máy sau này và ngược lại.
		T01	Kĩ năng sử dụng Dev Tool của trình duyệt, sử dụng các công cụ khác nhau để crawl data như scrapy, selenium...	Khó khăn trong việc truy cập những dữ liệu không được chia sẻ chính thống.
		T05	Hiểu về những các độ đo trong đánh giá mô hình. Vận dụng những độ đo phù hợp để đánh giá các mô hình đang sử dụng, từ đó rút ra những mô hình phù	Đánh giá những mô hình nào thực sự là phù hợp với dữ liệu. Hiểu về những nguyên nhân khiến mô hình kém phù hợp với dữ liệu đang xét.

			hợp đối với dữ liệu thực hành.	
2	Nguyễn Thị Châu Ngọc	T00	Tìm hiểu được thêm các lĩnh vực thú vị. Cách để lựa chọn được một nguồn dữ liệu phù hợp.	Khó khăn khi tìm một nguồn phù hợp. Phải đánh đổi trong việc lựa chọn các nguồn. Có những trang web có rất nhiều thông tin thú vị nhưng không thể khai thác vì vấn đề bảo mật hay không phù hợp cho xây dựng mô hình học máy sau này và ngược lại.
		T02	Các kĩ năng trong việc khám phá và tiền sử lí dữ liệu. Cách để hiểu hơn về dữ liệu đang làm việc và biết khai thác những khía cạnh thú vị của dữ liệu để phục vụ cho mục đích sau này.	Khó khăn trong việc tìm ra những khía cạnh thú vị của dữ liệu để tiến hành xử lý và khám phá.
		T03	Kĩ năng đặt câu hỏi và xử lí để trả lời các câu hỏi bằng dữ liệu. Biết về đa dạng các biểu đồ và nâng cao kĩ năng trực quan hóa dữ liệu một cách tốt nhất .	Lựa chọn các câu hỏi thật sự có ý nghĩa cho các đối tượng khác nhau. Chọn ra những biểu đồ thật sự phù hợp kết hợp với việc xử lí dữ liệu để trực quan hóa trả lời câu hỏi một cách dễ hiểu.
		T05	Hiểu về những các độ đo trong đánh giá mô hình. Vận dụng những độ đo phù hợp để đánh giá các mô hình đang sử dụng, từ đó rút ra	Đánh giá những mô hình nào thực sự là phù hợp với dữ liệu. Hiểu về những nguyên nhân khiến mô hình kém phù hợp với dữ liệu đang xét.

			những mô hình phù hợp đối với dữ liệu thực hành.	
3	Lê Hoàng Phúc	T00	Tìm hiểu được thêm các lĩnh vực thú vị. Cách để lựa chọn được một nguồn dữ liệu phù hợp	Gặp khó khăn trong việc tìm một nguồn phù hợp. Phải đánh đổi trong việc lựa chọn các nguồn. Có những web có rất nhiều thông tin thú vị nhưng không thể khai thác vì vấn đề bảo mật hay không phù hợp cho xây dựng mô hình học máy sau này và ngược lại.
		T03	Các kĩ năng để đưa ra và trả lời các câu hỏi bằng dữ liệu Sử dụng các biểu đồ hình vẽ để trực quan cho phần trình bày	Lựa chọn các câu hỏi thực sự có ý nghĩa cho các đối tượng khác nhau Lựa chọn các loại biểu đồ phù hợp nhất để biểu diễn dữ liệu.
		T04	Có cái nhìn tổng quan về các loại mô hình Nhận biết cách chọn mô hình, và thuật toán nào trong mô hình đó cho phù hợp với bài toán đưa ra	Khó khăn trong việc đưa ra bài toán phù hợp nhất với dữ liệu hiện có Có quá nhiều mô hình và thuật toán khác nhau. Mất nhiều thời gian để tìm hiểu và lựa chọn.
		T06	Cách tổ chức lịch biểu, phân công công việc	Không có
4	Nguyễn Hoàng Thịnh	T00	Tìm hiểu được thêm các lĩnh vực thú vị. Cách để lựa chọn được một nguồn dữ liệu phù hợp	Gặp khó khăn trong việc tìm một nguồn phù hợp. Phải đánh đổi trong việc lựa chọn các nguồn. Có những web có rất nhiều thông tin thú vị nhưng không thể

				khai thác vì vấn đề bảo mật hay không phù hợp cho xây dựng mô hình học máy sau này và ngược lại.
		T03	Kỹ năng tìm câu hỏi để trả lời và xử lý câu hỏi đó bằng dữ liệu Cách để trực quan hóa câu hỏi đó bằng biểu đồ	Mất nhiều thời gian để tìm câu hỏi sao cho phù hợp với dữ liệu hiện tại Lựa chọn biểu đồ trình bày sao cho dễ hiểu và rõ rang nhất có thể
		T04	Nhận biết việc chọn câu hỏi sao cho phù hợp và cách chọn mô hình và thuật toán nào trong mô hình đó cho bài toán đó	Khó khăn trong việc đưa ra bài toán phù hợp với dữ liệu hiện có. Có quá nhiều mô hình và thuật toán khác nhau. Mất nhiều thời gian để tìm hiểu và lựa chọn
		T06	Cách trình bày báo cáo	Không có

V.Những điều nhóm sẽ thực hiện nếu có thêm thời gian

STT	Nội dung	Sẽ phát triển thêm nếu có nhiều thời gian hơn
1	Thu thập dữ liệu	Tìm thêm những nguồn khác nhau để làm phong phú hơn cho tập dữ liệu. Tập dữ liệu thu thập cần phải đảm bảo, đáp ứng đủ trường để thuận tiện cho việc phát triển mô hình sau này.
2	Đặt câu hỏi	Sử dụng đa dạng, nâng cao hơn các loại biểu đồ để trực quan cho câu trả lời. Đưa ra thêm các câu hỏi mới lạ có ý nghĩa thực tiễn cao.
3	Mô hình hóa dữ liệu	Với bài toán dự đoán số lượng ứng viên sẽ đăng kí một công việc nào đó. Nhóm cần học và tìm hiểu thêm các loại mô hình hồi quy khác

		nhau hơn như (KNN, Lasso Regression, Gaussian Regression, Polynomial Regression,...). Từ đó lựa chọn thuật toán tối ưu nhất cho bài toán đang xét.