ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Thông tin sinh viên

1. **20120089** **LÊ XUÂN HOÀNG**
2. **20120422** **NGUYỄN THỊ ÁNH TUYẾT**
3. **20120460** **LÊ NGUYỄN HẢI DƯƠNG**
4. **20120494** **LÊ XUÂN HUY**

Đồ án:
Nhập môn khoa học dữ liệu

Thành phố Hồ Chí Minh – 12/2022

THÔNG TIN THÀNH VIÊN

MSSV	Họ tên	Email
20120089	Lê Xuân Hoàng	20120089@student.hcmus.edu.vn
20120422	Nguyễn Thị Ánh Tuyết	20120422@student.hcmus.edu.vn
20120460	Lê Nguyễn Hải Dương	20120460@student.hcmus.edu.vn
20120494	Lê Xuân Huy	20120494@student.hcmus.edu.vn

PHÂN CÔNG VÀ MỨC ĐỘ HOÀN THÀNH

I. Thu thập dữ liệu

Thành viên	Công việc
Lê Xuân Hoàng	Tìm link của từng con vật.
Nguyễn Thị Ánh Tuyết	Từ link từng con vật, cào dữ liệu của từng con vật.
Lê Nguyễn Hải Dương	Tìm lỗi, bổ sung, chỉnh sửa phần thu thập dữ liệu.
Lê Xuân Huy	Tìm lỗi, bổ sung, chỉnh sửa phần thu thập dữ liệu.

II. Khám phá dữ liệu và tiền xử lý

Thành viên	Công việc
Lê Xuân Hoàng	Tìm lỗi, bổ sung, chỉnh sửa phần khám phá và tiền xử lý.
Nguyễn Thị Ánh Tuyết	Tìm lỗi, bổ sung, chỉnh sửa phần khám phá và tiền xử lý.
Lê Nguyễn Hải Dương	Khám phá và tiền xử lý dữ liệu.
Lê Xuân Huy	Khám phá và tiền xử lý dữ liệu.

III. Đặt câu hỏi và trả lời

Thành viên	Công việc
Lê Xuân Hoàng	Tự đặt câu hỏi và trả lời câu hỏi của bản thân, sau đó minh họa cho câu hỏi của mình
Nguyễn Thị Ánh Tuyết	Tự đặt câu hỏi và trả lời câu hỏi của bản thân, sau đó minh họa cho câu hỏi của mình
Lê Nguyễn Hải Dương	Tự đặt câu hỏi và trả lời câu hỏi của bản thân, sau đó minh họa cho câu hỏi của mình
Lê Xuân Huy	Tự đặt câu hỏi và trả lời câu hỏi của bản thân, sau đó minh họa cho câu hỏi của mình

Mỗi người có mỗi file đặt và trả lời câu hỏi trên Github. Folder **dat_va_tra_loi_cau_hoi.**

IV. Mô hình và đánh giá

Thành viên	Công việc
Lê Xuân Hoàng	Mô hình dự đoán cân nặng của bộ chim sẻ dựa trên chiều dài của nó – Linear regression model.
Nguyễn Thị Ánh Tuyết	Bổ sung, đánh giá và xác định bậc tối ưu cho Polynomial regression model.
Lê Nguyễn Hải Dương	Mô hình dự đoán cân nặng của bộ chim sẻ dựa trên chiều dài của nó – Polynomial regression model.
Lê Xuân Huy	Tổng hợp tất cả thông tin, làm slide

V. Công việc khác

- Làm slide: Lê Xuân Huy
- Phân công: Lê Xuân Hoàng

VI. Tổng hợp kết quả

1. Bạn đã gặp khó khăn gì

DƯƠNG:

- Các cột thú vị ban đầu nhóm định đặt câu hỏi không sử dụng được. Nguyên nhân:
  - Nhiều cột missing ratio lên đến 90 thậm chí là 95%.
  - Khó xử lý.
- Dữ liệu là ngôn ngữ tự nhiên, không có format chung nên rất khó xử lý:

- Cùng một đơn vị nhưng cách ghi khác nhau, nhiều loại đơn vị khác nhau.
- Dòng các dữ liệu cách nhau bằng dấu ‘,’ thì lại xuất hiện thêm dấu cách sau dấu phẩy hay ‘ ‘ (2 dấu cách)…
- Chữ lúc được ghi hoa lúc được ghi thường, có chữ có thêm ký tự đặc biệt như “dấu nháy kép” hoặc “dấu phẩy”.

HOÀNG:

- Phần cào từ dữ liệu selenium khá lag, chạy rất lâu, khi thực hiện thao quá nhiều thao tác thì thậm chí sẽ không tiếp tục chạy đc
- Trước khi nghĩ đến regex, đã thử chuyển dữ liệu bằng tay, rất khó khăn
- Tốn nhiều thời gian để suy nghĩ ra bài toán để áp dụng mô hình học máy

HUY:

- Phần suy nghĩ để đặt câu hỏi khá là khó.
- Xử lý dữ liệu của cột kiểu phân loại khá phiền phức, còn có cả cột có dữ liệu kiểu dict lồng 1 cái dict khác.
-

TUYẾT:

- Ban đầu dùng selenium để cào dữ liệu của từng con vật, tuy nhiên phải chờ khá lâu để chạy hết được 1 con vật nên em đổi sang dùng scrapy thì thời gian chạy rất nhanh.
- Dữ liệu ở các cột Height, Weight, Length,.. không đồng nhất về mặt hình thức, có chỗ thì có 1 số, chỗ thì 2 số, chỗ thì dấu ‘.’ dùng để đánh dấu là thập phân nhưng cũng có chỗ dấu ‘,’ để đánh dấu là thập phân. Dấu ‘ ’ xuất hiện không có quy tắc và có rất nhiều đơn vị khác nhau nên cần xử lí phức tạp.
- Dữ liệu thiếu nhiều nên khó để chọn lọc được dữ liệu để phân tích, đặt câu hỏi.

2. Bạn đã học được điều gì

DƯƠNG:

- Tìm được dữ liệu thì nên kiểm tra kỹ càng hơn.
- Các kỹ năng mới trong tiền xử lý (split multiple delimiter, merge,…)
- Chọn những cột kiểu phân loại thì nên chọn những cột có ít giá trị khác nhau, hoặc có cách để giảm số lượng do nhiều quá thì khó theo dõi và khó chọn biểu đồ.
- Cách để làm ra 1 mô hình máy học và đánh giá bằng sklearn.

HOÀNG:

- Cần xác định rõ các vấn đề muốn giải quyết trước khi tìm dữ liệu, để chọn dữ liệu phù hợp.

- Một số thuật toán học máy từ sklearn
- Nên sử dụng regex nếu nó có định dạng (sau khi xem code của đồng đội thì thấy làm cách này thật sự rất khỏe)
- Khi dùng selenium mà lag quá thì có thể chia ra nhiều lần để chạy, không nhất thiết phải chạy 1 lần, miễn sao ra kết quả để dùng là được.
- Cách dùng Trello để làm nhóm

TUYẾT

- Cần chọn lọc những nguồn dữ liệu đầy đủ hơn.
- Xem xét tất cả các trường hợp có thể có của dữ liệu để tiền xử lí, tránh bỏ sót một số trường hợp đặc biệt.
- Biết cách tạo ra một số mô hình máy học bằng sklearn.
- Sử dụng một số phương pháp để đánh giá mô hình như K-Fold, Leave-one-out (1 nhánh của K-Fold), chọn bậc tối ưu nhất cho Polynomial regression.
- Được thực hành quy trình làm việc với dữ liệu (thu thập -> khám phá và tiền xử lí -> phân tích, đặt câu hỏi, khai thác câu trả lời từ dữ liệu -> xây dựng mô hình)

HUY

- Thao tác với github, các thư viên trong python đã thuần thục hơn.
- Hiểu hơn về cách sử dụng các thư viện máy học.
- Hiểu rõ hơn về các quy trình dữ liệu.

3. Bạn sẽ làm gì nếu có nhiều thời gian hơn

- Tìm kiếm thêm các mô hình hay hơn, xử lí được các ngôn ngữ tự nhiên chứ không chỉ tạo mô hình với các dữ liệu số.
- Xử lý các cột Continents, Countries để thống kê đa dạng sinh học cho từng khu vực.
- Làm mô hình dự đoán con vật dựa trên thông tin về tốc độ, cân nặng, chiều cao và điều kiện khu vực sống. (có nhiều vụ phá hoại của động vật chỉ thấy được bóng dáng chứ không biết là con gì để thực hiện phòng tránh, mô hình trên sẽ giúp chúng ta dự đoán chúng).
- Làm mô dự đoán lớp , bộ hoặc họ của động vật dựa trên thông tin các cột.

Công việc chưa làm được: Không có.