**PYTHON FOR DATA SCIENCE**

**GV LT: Hà Văn Thảo**



**BÁO CÁO ĐỒ ÁN**

Tên đồ án: Traffic Accidents in Singapore

MSSV 1:

MSSV 2:

MSSV 3: 20280083 Họ tên: Lại Toàn Thắng

MSSV 4: Họ tên:

Họ tên giáo viên: Hà Văn Thảo

C:\Users\tdqua_000\Dropbox\SS-Slides\DeCuong-CDIO\Template CDIO v4.2\Templates\Hinh anh\LogoTruong.png

Ngành Khoa Học Dữ Liệu

Đại học Khoa học tự nhiên TP HCM

Mục Lục

[1 TỰ ĐÁNH GIÁ ĐỒ ÁN 3](#_Toc122602910)

[1.1 Mô tả dự án 4](#_Toc122602911)

[1.2 Đóng góp của các thành viên cho dự án 5](#_Toc122602912)

[1.3 Thông tin cần thiết 6](#_Toc122602913)

[2 BÁO CÁO QUÁ TRÌNH THỰC HIỆN 7](#_Toc122602914)

[3 NẾU CÓ THỜI GIAN NHÓM SẼ LÀM GÌ? 8](#_Toc122602915)

[1. Thu thập dữ liệu 8](#_Toc122602916)

[2. Đặt và trả lời các câu hỏi có ý nghĩa 8](#_Toc122602917)

[3. Mô hình hoá 8](#_Toc122602918)

[4 CÁC THAM KHẢO 8](#_Toc122602919)

# C:\Users\tdqua_000\Desktop\self_reflection.jpg TỰ ĐÁNH GIÁ ĐỒ ÁN

Điểm tự đánh giá cho toàn đồ án: 9.5 / **10**

**Các Tiêu Chí Đánh Giá**

## Mô tả dự án

## Đóng góp của các thành viên cho dự án

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** | **Tỉ lệ** |
| 01 | 20280016 | Visualization |  |
| 02 | 20280083 | Visualization | 25% |
| 03 |  | Visualization |  |
| 04 |  | Visualization |  |
|  |  | **Tổng:** | 100% |

## Thông tin cần thiết

Phần 1: Quản lý Nhiệm vụ

|  |  |  |
| --- | --- | --- |
| **Nhiệm Vụ** | **Thành viên thực hiện** | **Thời gian** |
| Tổ chức và quản lý Github | Lại Toàn Thắng | Toàn bộ quá trình thực hiện |

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

Phần 3: Mô hình hóa

|  |  |  |
| --- | --- | --- |
| **Nhiệm Vụ** | **Thành viên thực hiện** | **Thời gian** |
|  |  |  |

# C:\Users\tdqua_000\Desktop\checklist (1).jpg BÁO CÁO QUÁ TRÌNH THỰC HIỆN

|  |  |
| --- | --- |
| Ở nhiệm vụ Thu thập dữ liệu: | *Nhóm đã có dataset có sẵn. Đã trải qua quá trình preprocess nên ta thấy dữ liệu tốt và đẹp.*  *Tuy nhiên, những khó khăn có thể gặp trong quá trình này là việc lựa chọn trang web để thực hiện crawl dữ liệu, Việc cân nhắc lựa chọn giữa các công cụ (Selenium, BeautifulSoup, HTML Request), bỏ qua một vai dòng dữ liệu do không mở được link, chưa tối ưu được thời gian khi crawl dữ liệu.* |
| Ở nhiệm vụ đặt câu hỏi | *Có chút khó khăn trong việc đặt các câu hỏi có ý nghĩa từ dữ liệu mình có* |
| *Ở nhiệm vụ trả lời câu hỏi* | *Có một số đồ thị cần các kiến thức liên quan để quản lý cơ sở dữ liệu như aggregate, group by,… thì chưa làm được tốt* |
| *Kỹ năng học được* | *Kinh nghiệm từ các các bạn trong nhóm về việc thực hiện một đồ án KHDL:*   * *Cách quản lý github repository và phân công nhiệm vụ.* * *Trình bày một file ipynb trực quan đẹp mắt.* |
|  |  |

# C:\Users\tdqua_000\Desktop\exercise-clipart.jpg NẾU CÓ THỜI GIAN NHÓM SẼ LÀM GÌ?

## 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.
* Mục tiêu của hồi quy tuyến tính và các mô hình dự đoán khác là dự đoán giá trị của một biến phụ thuộc vào các biến độc lập (chủ yếu là dạng numeric). Thực tế dự liệu của mình phần lớn là object.
* Có một sự thật là dữ liệu nào dễ trực quan thì khó để mô hình hoá 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.

## 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

# CÁC TÀI LIỆU THAM KHẢO