**PYTHON FOR DATA SCIENCE**

**GV LT: Hà Văn Thảo**



**BÁO CÁO ĐỒ ÁN**

Tên đồ án: Traffic Accidents in Canada

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 THÔNG TIN CHUNG 3](#_Toc122703275)

[1.1. Thông tin thành viên 3](#_Toc122703276)

[1.2. Các công tụ tổ chức quản lý 3](#_Toc122703277)

[2 NỘI DUNG BÀI BÁO CÁO 4](#_Toc122703278)

[3 BÁO CÁO QUÁ TRÌNH THỰC HIỆN 5](#_Toc122703279)

[4 NẾU CÓ THỜI GIAN NHÓM SẼ LÀM GÌ? 6](#_Toc122703280)

[1. Thu thập dữ liệu 6](#_Toc122703281)

[2. Đặt và trả lời các câu hỏi có ý nghĩa 6](#_Toc122703282)

[5 CÁC TÀI LIỆU THAM KHẢO 7](#_Toc122703283)

# C:\Users\tdqua_000\Desktop\self_reflection.jpg THÔNG TIN CHUNG

## Thông tin thành viên

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** | **Github** |
| 01 | 20280016 |  |  |
| 02 | 20280083 | Lại Toàn Thắng |  |
| 03 |  |  |  |
| 04 |  |  |  |
|  |  | **Tổng:** | 100% |

## Các công tụ tổ chức quản lý

Link Github:

https://github.com/laitoanthang/KHDL-PythonForDS

Link Notion:

https://www.notion.so/thangdumbest/N-CU-I-K-2dd34951d9fe4c269adf020471cd16c7

Link câu hỏi: https://docs.google.com/spreadsheets/d/1UG6YbiDmWzxuseNBtZs652whpZgNG6vaiT6ynEhZRBM/edit#gid=0

# NỘI DUNG BÀI BÁO CÁO

## 2.1. Mô Tả Vấn Đề (Problem Overview) Tai nạn giao thông đường bộ gia tăng mỗi năm. Chúng ta có thể ngăn chặn nếu tất cả chúng ta thận trọng hơn khi tham gia giao thông, việc đó có thể làm được nếu chúng ta tìm hiểu kĩ càng nguyên nhân gây ra tai nạn. Có những yếu tố góp phần vào tai nạn nhiều hơn những người khác, nếu vậy thì làm thế nào để chúng ta tìm ra những điều mà hầu hết các vụ tai nạn đường bộ đều có điểm chung và sau đó sử dụng những điều đó để đưa ra quyết định tốt hơn, điều này có thể dẫn đến việc giảm sự cố. Nhóm nghĩ tại sao không sử dụng học máy trên bộ dữ liệu trong thế giới thực như vậy và tìm ra mọi thứ nào xác định yếu tố nào có mối tương quan lớn nhất với tai nạn thực tế và cũng sử dụng phương pháp tương tự này để dự đoán tai nạn thời gian thực.

## 2.2. Dataset sử dụng

<https://open.canada.ca/data/en/dataset/8dd0ab9b-d45d-4526-9256-c598fbc4ff3a>

Nhóm đã sử dụng dữ liệu có sẵn từ một page trên facebook được crawl từ [trang web trên](https://open.canada.ca/data/en/dataset/8dd0ab9b-d45d-4526-9256-c598fbc4ff3a).

## 2.3. EDA (Exploratory data analysis)

Nhóm đã làm một chút EDA để có được một insight về tập dữ liệu.

Dữ liệu có tất cả là 10 cột, tức nhóm phải deal với tổng cộng 10 features.

…

## 2.4. Model building và feature engineering

# 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. Vì vậy nhóm đã không thể làm phần modeling.
* Cuối cùng, 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 hơn 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

## Mô hình hoá

* Học qua kiến thức về model selection để lựa chọn các thuộc tính cần thiết
* Label encoding tất cả các categorical features
* Sampling với tất cả các label encoded features
* Logistic regression với response encoding
* Tìm hiểu về LightGBM with response coding
* Tìm hiểu thêm về stacking classifiers.

Một số việc cần làm:

* Thống kê tai nạn theo từng thời gian trong ngày -> thêm mức độ severity để dễ nhận biết
* Tương tự với severity với weather
* Tạo new features để dự đoán model
* Lable encoding tất cả các categorical features
* Sampling with all label encoded features
* Logistic regression với response encoding
* Model 2: LightGBM with response coding
* Tìm hiểu về Logistic Regression, Random Forest, Decision Trees, LightGBM và stacking classifiers. Xe cái nào đưa ra kết quả tốt nhất trong chúng

# CÁC TÀI LIỆU THAM KHẢO