ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

====oOo====

BÁO CÁO MÔN HỆ HỖ TRỢ QUYẾT ĐỊNH

ĐỀ TÀI:

HỆ THỐNG PHÂN TÍCH NGUY CƠ TAI NẠN GIAO THÔNG CÁC KHU VỰC TRONG THÀNH PHỐ

Giảng viên hướng dẫn:	**TS. LÊ CHÍ NGỌC**
Sinh viên thực hiện:	**HOÀNG VĂN THÀNH**
MSSV:	**20173586**
Lớp:	**HTTTQL - K62**

Hà Nội, 2020

Mục lục

	Trang
Lời mở đầu	*3*
Chương 1: Khảo sát, đánh giá	*4*
Chương 2: Phân tích thiết kế hệ thống	*7*
Chương 3: Thuật toán, đánh giá mô hình	*10*
Chương 4: Giao diện hệ thống	*14*
Kết luận	*16*
Tài liệu tham khảo	*17*

Lời mở đầu

Tai nạn giao thông đường bộ là một bài toán nan giải làm đau đầu các nhà quản lý ở mọi quốc gia, khu vực. Trong thời đại 4.0, ta có thể áp dụng các thành quả của khoa học kĩ thuật đương thời trong việc hỗ trợ giảm thiểu nguy cơ xảy ra tai nạn. Bài báo cáo này tôi xin trình bày về một giải pháp có thể giúp góp phần giải quyết vấn đề trên thông qua một số mô hình máy học đơn giản để phân tích các dữ liệu về các vụ tai nạn giao thông.

Bài báo cáo này gồm 4 phần:

- **Chương 1: Khảo sát, đánh giá**
- **Chương 2: Phân tích thiết kế hệ thống**
- **Chương 3: Thuật toán, đánh giá mô hình**
- **Chương 4: Giao diện hệ thống**

Mặc dù đã rất cố gắng nhưng khó tránh khỏi sai sót. Rất mong được sự góp ý của thày cô, bạn bè. Em xin cảm ơn thày **Lê Chí Ngọc** đã giúp đỡ em rất nhiều để có thể hoàn thành báo cáo môn học này.

Chương 1: Khảo sát đánh giá

1. Khảo sát

Tai nạn giao thông đường bộ là nguyên nhân chính gây tử vong trên toàn cầu, dẫn đến khoảng 1,25 triệu ca tử vong hàng năm. Các cơ quan vận tải trên toàn thế giới đã cố gắng thực hiện các chiến lược để giảm thiểu tai nạn bằng cách đưa ra các bộ luật an toàn giao thông khác nhau. Tuy nhiên, giảm thiểu tai nạn giao thông là một mục tiêu khó thực hiện vì do nhiều yếu tố mà không thể kiểm soát hoàn toàn được bằng luật.

Một phần xuất phát từ những khó khăn trong việc dự đoán tai nạn giao thông có thể xảy ra khi nào và ở đâu. Mặc dù khó khăn, nhưng dự đoán không phải là không thể. Tai nạn giao thông không xảy ra theo cách hoàn toàn ngẫu nhiên. Một số tài liệu khẳng định rằng xác suất của tai nạn giao thông bị ảnh hưởng bởi vô số yếu tố có thể đo lường được. Những yếu tố dễ dàng thấy được như tốc độ lái xe, tình trạng giao thông, cấu trúc đường và thời tiết.

Ta có thể xây dựng một hệ thống có thể dựa trên các vụ tai nạn trước đó để phân tích các yếu tố trên, đưa ra dự báo về nguy cơ xảy ra tai tạn cho từng điểm mỗi ngày. Từ đó đưa ra một số giải pháp tăng cường quản lý hệ thống giao thông ở các địa điểm nguy cơ cao giúp giảm thiểu tai nạn xảy ra.

2. Khảo sát dữ liệu

Bộ dữ liệu chính sử dụng cho bài toán này là dữ liệu được lấy từ Kaggle. Bộ dữ liệu này chứa thông tin chi tiết về khoảng 1,6 triệu vụ tai nạn giao thông xảy ra ở Anh trong khoảng thời gian từ năm 2005 đến 2014. Những chi tiết này bao gồm địa điểm, thời gian, mức độ nghiêm trọng của các vụ tai nạn cũng như thời tiết và một số yếu tố về cơ sở hạ thầng giao thông. Trong bài toán này, tôi chỉ sử dụng phần dữ liệu ghi lại các vụ tai nạn từ năm 2010 tới 2014 để tránh bị ảnh hưởng bởi nhiễu của cuộc khủng hoảng kinh tế năm 2008.

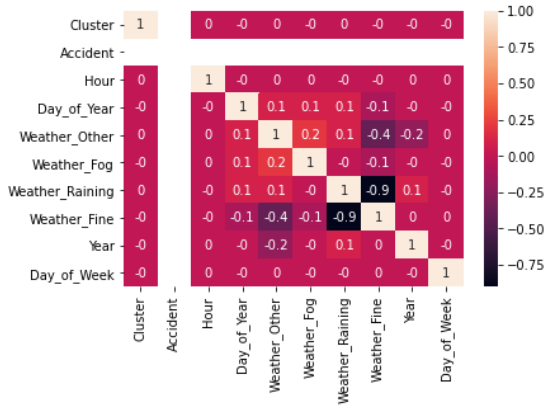
Sau khi tiền xử lí dữ liệu, ta được bộ dữ liệu như sau:

Hour	Day_of_Year	Weather_Other	Weather_Fog	Weather_Raining	Weather_Fine	Year	Day_of_Week
7.0	305	1.0	0.0	0.0	0.0	2010	2
18.0	305	0.0	0.0	1.0	0.0	2010	2
10.0	335	0.0	0.0	0.0	1.0	2010	3
21.0	32	0.0	0.0	0.0	1.0	2010	7
20.0	91	0.0	0.0	0.0	1.0	2010	2
14.0	18	0.0	0.0	0.0	1.0	2010	2
8.0	60	0.0	0.0	0.0	1.0	2010	1
23.0	91	0.0	0.0	0.0	1.0	2010	2
6.0	91	1.0	0.0	0.0	0.0	2010	2
18.0	91	0.0	0.0	0.0	1.0	2010	2
18.0	121	0.0	0.0	0.0	1.0	2010	3
21.0	182	0.0	0.0	0.0	1.0	2010	5
16.0	15	0.0	0.0	0.0	1.0	2010	6
13.0	13	0.0	0.0	0.0	1.0	2010	4
20.0	14	0.0	0.0	0.0	1.0	2010	5
17.0	13	0.0	0.0	0.0	1.0	2010	4
18.0	15	0.0	0.0	0.0	1.0	2010	6
23.0	17	0.0	0.0	0.0	1.0	2010	1
7.0	18	0.0	0.0	0.0	1.0	2010	2
19.0	17	0.0	0.0	0.0	1.0	2010	1

Bộ dữ liệu bao gồm 756 934 bản ghi với 8 cột:

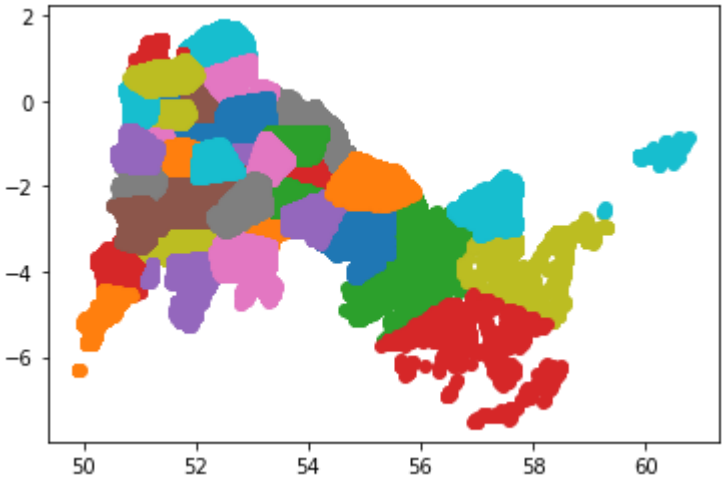
- **Hour**: giờ xảy ra tai nạn
- **Day_of_Year**: ngày trong năm xảy ra tai nạn
- **Day_of_Week**: thứ trong tuần xảy ra tai nạn
- **Year**: năm sảy ra tai nạn
- **Weather_Fine**: thời điểm xảy ra tai nạn có thời tiết tốt
- **Weather_Raining**: thời điểm xảy ra tai nạn có mưa
- **Weather_Fog**: thời điểm xảy ra tai nạn có sương mù
- **Weather_Other**: điều kiện thời tiết khác

Ta có biểu đồ mức độ tương quan giữa các cột:



Thật không may, bộ dữ liệu nói trên chỉ chứa các mẫu "dương tính" - nghĩa là chỉ có các bản ghi về các vụ tai nạn và các đặc điểm khác nhau của các vụ tai nạn này. Bộ dữ liệu không cung cấp bất kỳ mẫu "âm tính" nào. Vì tôi muốn dự đoán khả năng xảy ra tai nạn, nên tôi phải lấy mẫu âm tính. Bằng phương pháp được mô tả trong **Chương 3** tôi đã tạo ra tập dữ liệu âm tính gộp lại với tập dương tính trên để xác định khả năng xảy ra tai nạn tại các vị trí mà không làm thay đổi bản chất bài toán(sẽ được chứng minh trong **Chương 3**)

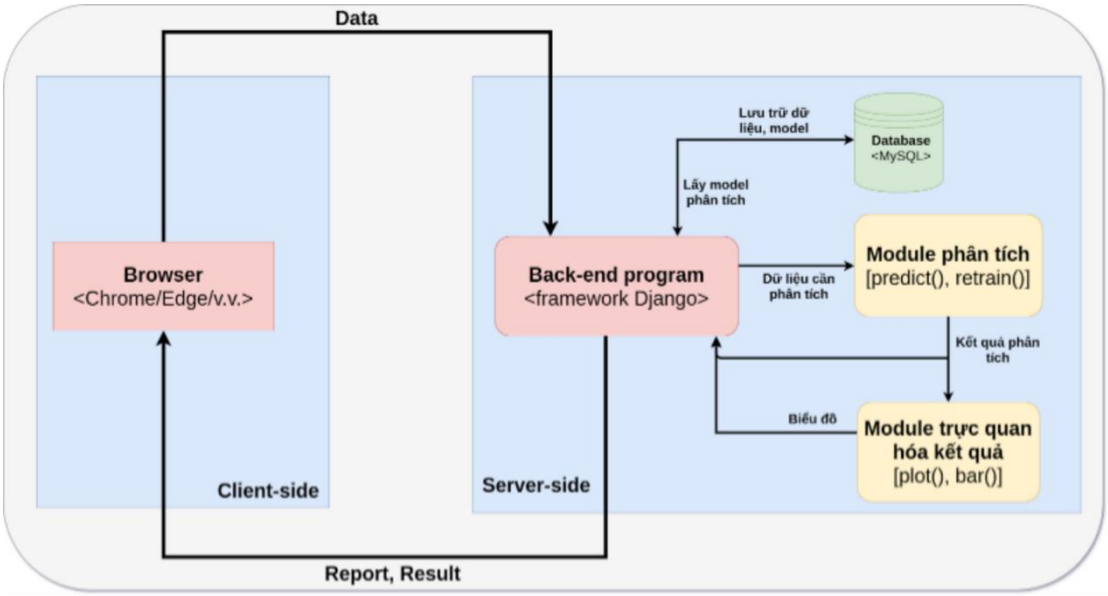
Khi khám phá bộ dữ liệu, tôi đã phát hiện ra rằng có những địa điểm thường xuyên xảy ra tai nạn. Sau đó, tôi đã sử dụng thuật toán Kmeans Clustering để nhóm các vị trí này thành các cụm, gọi là "điểm nóng tai nạn". Kết quả bước này ta được feature mới là **Cluster** cho biết mỗi vụ tai nạn thuộc vào cụm nào.



Sau đó tôi phân tích thêm về bộ dữ liệu này và phát hiện ra rằng xác suất xảy ra tai nạn có tương quan với một số đặc tính, bao gồm các ngày trong tuần, thời gian trong ngày và địa điểm. Sau khi xử lí xong tập dữ liệu, tôi tiến hành thử một số thuật toán và thấy Random Forest cho kết quả tốt nhất.

Chương 2: Phân tích thiết kế hệ thống

1. Sơ đồ tổng quan hệ thống:



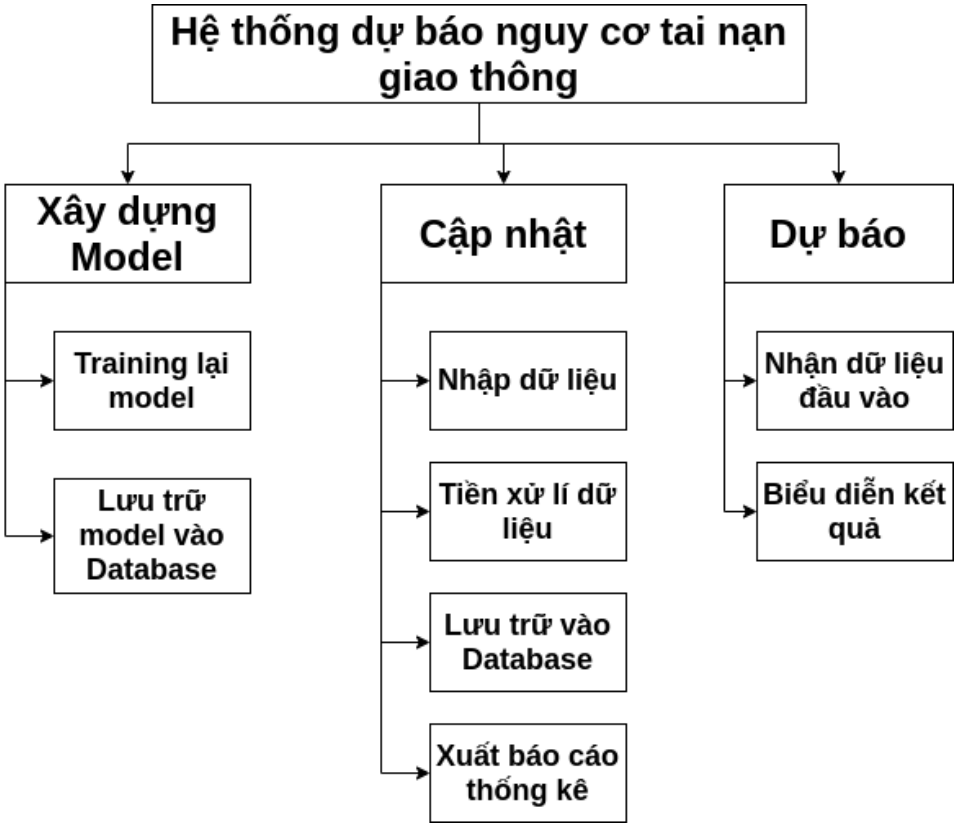
Phần Client-side:

- HTML: Là ngôn ngữ đánh dấu siêu văn bản xây dựng các trang giao diện web có thể mở được bằng bất cứ nền tảng hay thiết bị nào có trình duyệt.
- CSS: Thành phần bổ trợ cho HTML để định dạng lại các phần tử trong HTML đẹp hơn.
- JavaScript: Là ngôn ngữ lập trình dạng script giúp cho các phần tử trong giao diện web có chức tăng khi người dùng tương tác với nó.

Phần Server-side:

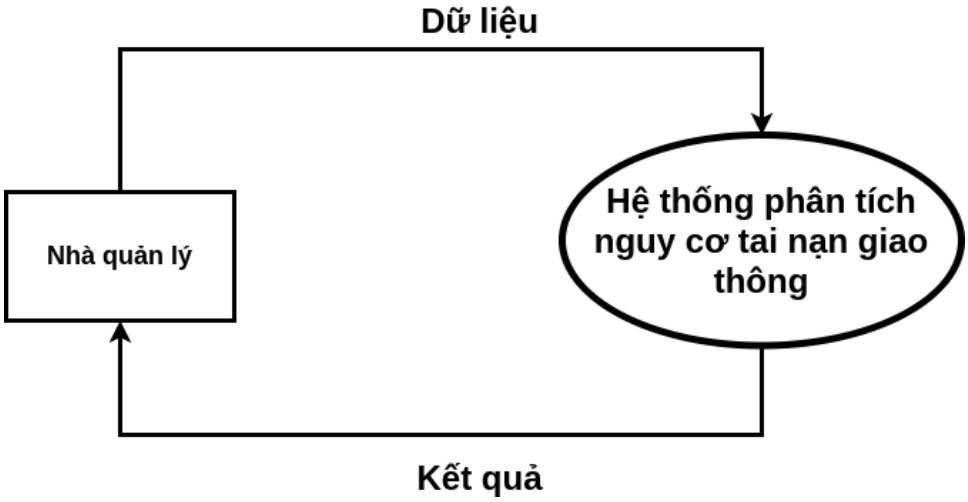
- Python: Là một ngôn ngữ lập trình dạng script phổ biến với sự ngắn gọn, tiện dụng, linh hoạt và cộng đồng người dùng đông đảo. Có các Framework lớn về khoa học dữ liệu và hệ thống web.
- Django: Là một Framework bằng ngôn ngữ python giúp xây dựng hệ thống server lớn. Thường dùng trong các hệ thống thương mại.

2. Sơ đồ phân cấp chức năng

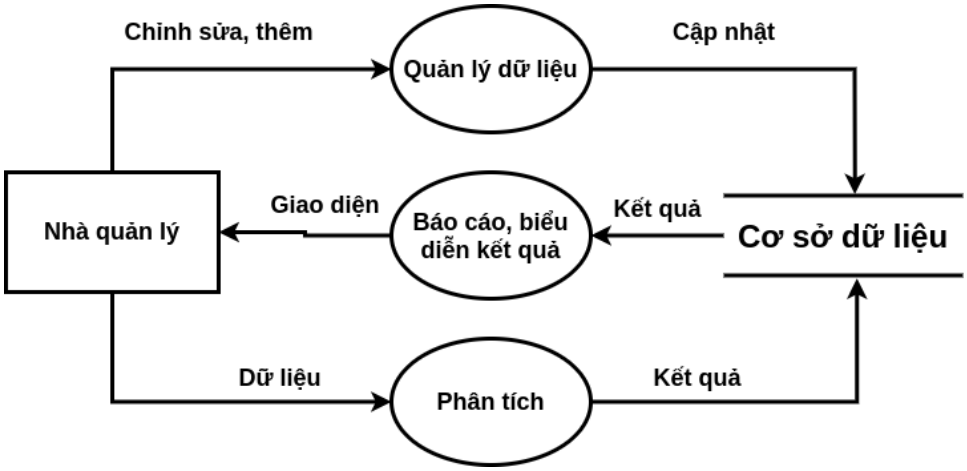


3. Sơ đồ luồng dữ liệu

3.1 Sơ đồ luồng dữ liệu mức ngữ cảnh



3.2 Sơ đồ luồng dữ liệu mức đỉnh


Chỉnh sửa, thêm
Quản lý dữ liệu
Cập nhật
Nhà quản lý
Giao diện
Báo cáo, biểu diễn kết quả
Kết quả
Cơ sở dữ liệu
Dữ liệu
Phân tích
Kết quả

Chương 3: Thuật toán, đánh giá mô hình

3.1 Tạo tập dữ liệu âm tính

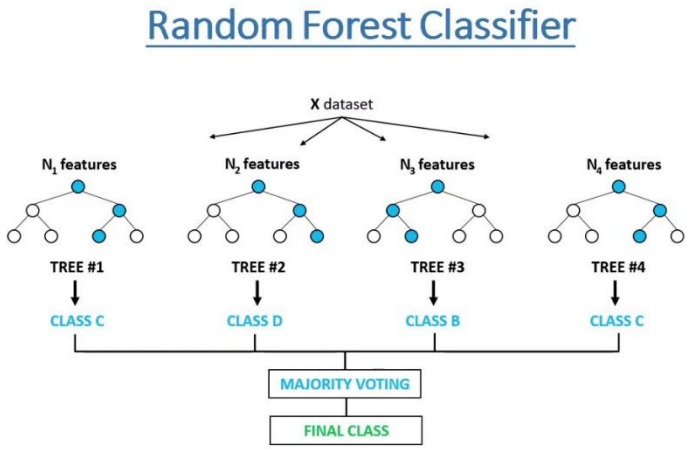
Để xây dựng được mô hình phân lớp nhị phân dự báo xảy ra tai nạn hay không tai nạn, ta cần có các tập mẫu về hai lớp đó. Các mẫu về các vụ tai nạn ta đã có ở trên. Còn tập mẫu về lớp không xảy ra tai nạn ta có thể sinh ra với một số điều kiện nhất định nhưng không trùng với các dữ liệu của các vụ tai nạn. Tại mọi con đường khi không xảy ra tai nạn đều có thể là một mẫu âm tính. Như vậy ta có thể xây dựng tập mẫu âm tính rất lớn. Nhưng rất dễ gây mất cân bằng nếu ta lấy quá nhiều.

Ý tưởng là tại mỗi mẫu về xảy ra tai nạn, ta thay đổi các thông tin về ngày, giờ, vị trí. Nếu thông tin mới đó không trùng với vụ tai nạn nào khác thì ta có một mẫu mới về không xảy ra tai nạn. Cụ thể, với từng vụ tai nạn xảy ra:

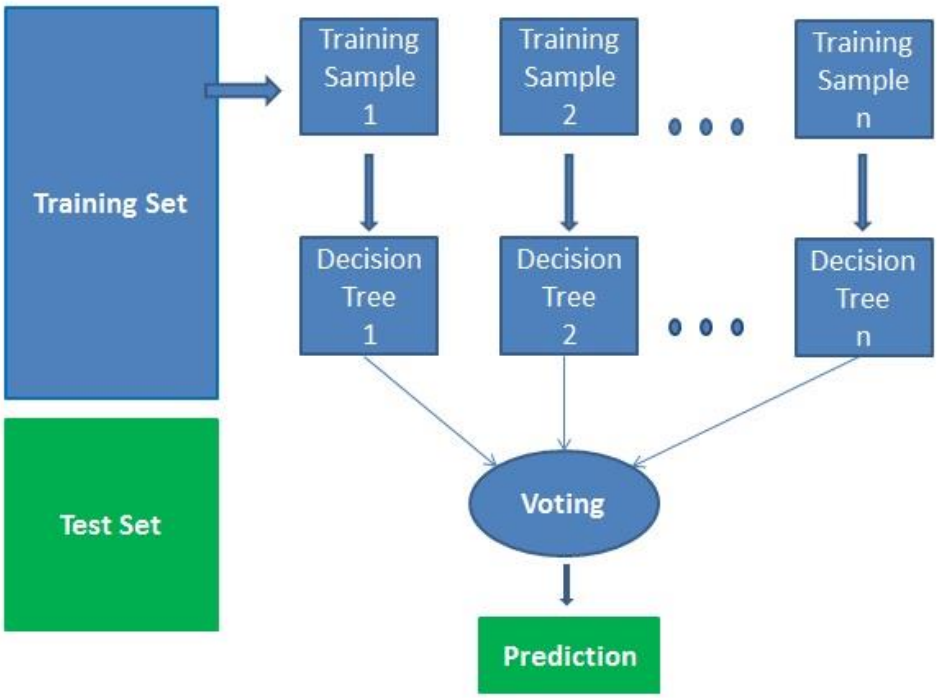
- **Hour**: chọn một giờ khác ngẫu nhiên trong khoảng [0;23] mà kết quả thu được không trùng với các vụ tai nạn khác
- **Cluster**: chọn một cụm khác ngẫu nhiên mà kết quả thu được không trùng với vụ tai nạn nào khác.
- **Day_of_Year:** chọn một ngày khác ngẫu nhiên trong năm mà kết quả thu được không trùng với các vụ tai nạn khác.

Sau bước này, từ hơn 700000 vụ tai nạn, ta sinh ra được hơn 1200000 mẫu âm tính mới. Ta cần quan tâm tới một vấn đề khác là liệu việc thêm tập dữ liệu âm tính tự tạo này vào có làm thay đổi bản chất của bài toán dự báo tai nạn giao thông mà ta đang cần làm hay không? Điều này sẽ được chứng minh trong phần đánh giá kết quả sau khi chạy xong mô hình ở mục **3.3**

3.2 Xây dựng mô hình Random Forest Classifier



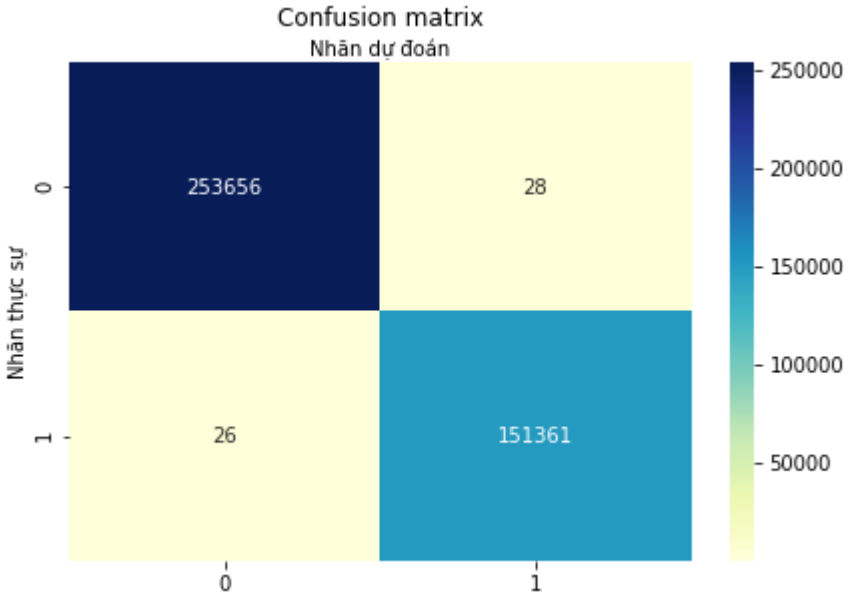
Rừng ngẫu nhiên(Random Forest) là một thuật toán học tập có giám sát. Nó có thể được sử dụng cả để phân loại và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một rừng ngẫu nhiên bao gồm nhiều cây quyết định(Decision Tree). Người ta nói rằng càng có nhiều cây, rừng càng mạnh. Rừng ngẫu nhiên tạo cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, nhận dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng phương thức bỏ phiếu. Nó cũng cung cấp một chỉ số khá tốt về tầm quan trọng của tính năng.



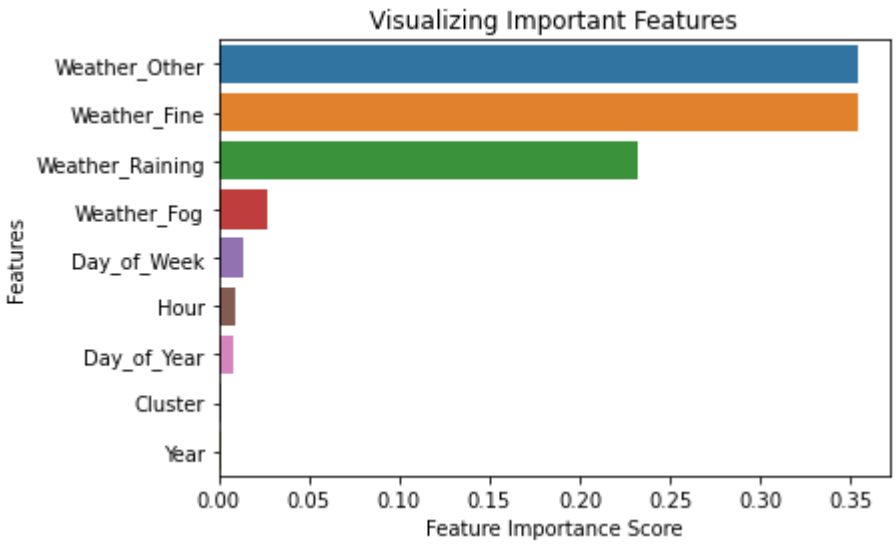
Sau khi trộn mẫu dương tính với âm tính, ta cho mô hình học từ tập dữ liệu đó và đạt được một số kết quả như sau:

- **Accuracy**: 0.999867
- **Precision**: 0.999815
- **Recall**: 0.999828

Kết quả phân loại trên tập test:



Mức độ quan trọng của các thuộc tính:



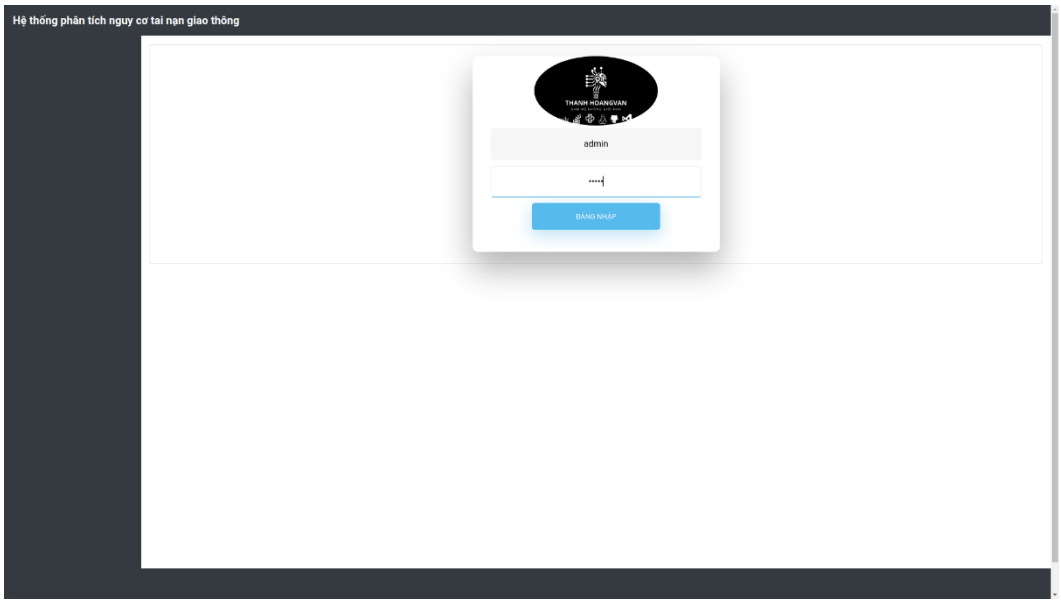
3.3 Đánh giá kết quả

Từ kết quả thu được trên, ta thấy rằng các đặc trưng như Hour, Day_of_Year, Cluster mà ta dùng để tạo bộ dữ liệu âm tính có mức độ quan trọng không cao, như vậy mô hình đã thực sự học được dữ liệu về bài toán mà ta cần tìm chứ không phải học được điều kiện mà ta sinh ra tập âm tính.

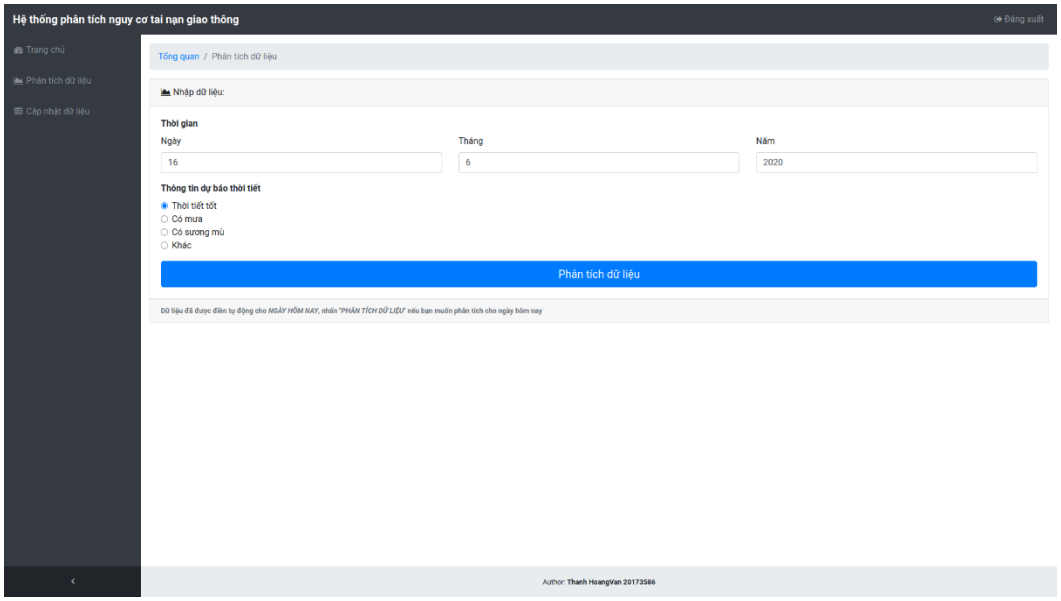
Kết quả mô hình về các chỉ số **Accuracy, Precision, Recall** cũng rất cao, như vậy có thể kết luận mô hình này có thể tin tưởng được.

Chương 4: Giao diện hệ thống

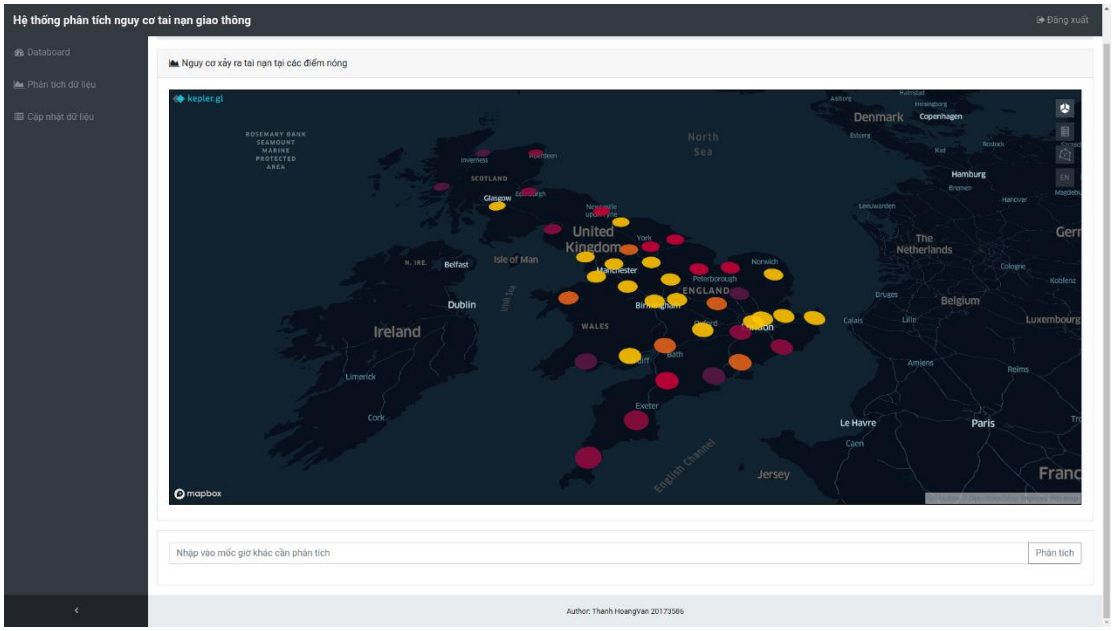
1. Giao diện đăng nhập



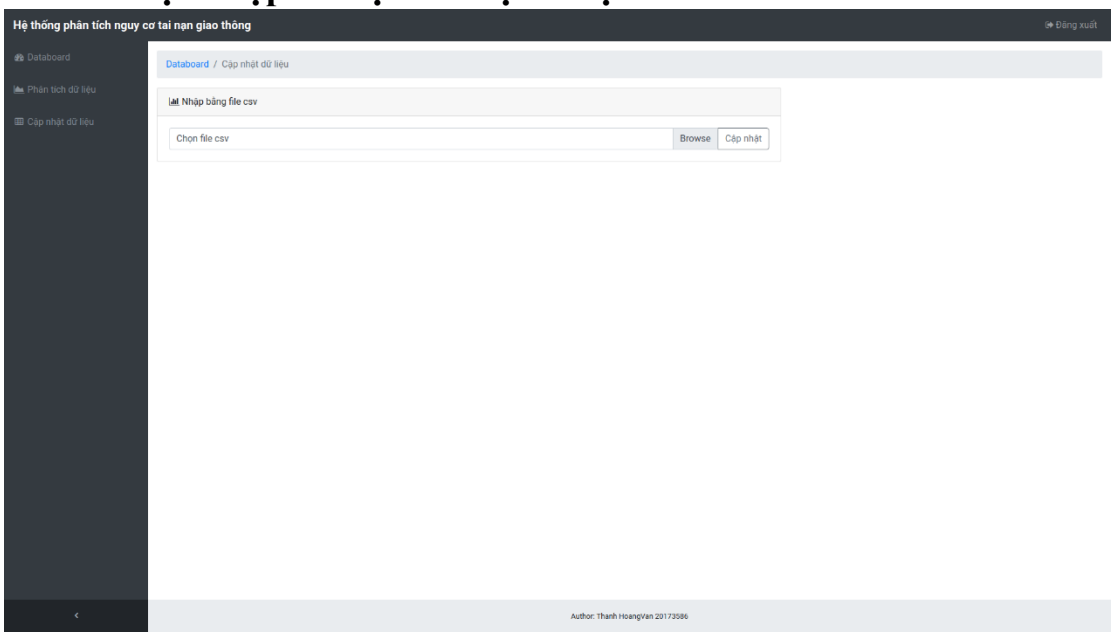
2. Giao diện nhập dữ liệu cho mô hình



3. Giao diện kết quả phân tích



4. Giao diện nhập dữ liệu các vụ tai nạn mới



Kết luận

Trên đây là giải pháp của tôi cho bài toán xác định nguy cơ tai nạn giao thông. Bằng một mô hình đơn giản như Random Forest Classifier kết hợp với Kmeans Clustering cũng đem lại kết quả với độ tin cậy cao.

Tuy nhiên, do số đặc trưng của dữ liệu còn nhỏ nên mô hình chưa thể khái quát được tổng quan về bài toán nên hạn chế về sự khả thi khi triển khai thực tế. Nên cần kết hợp thêm các dữ liệu của nhiều bên liên quan để cải thiện độ hiệu quả của hệ thống.

Mặt khác, mô hình chỉ mang tính chất hỗ trợ, gợi ý cho nhà quản lý ra quyết định. Hệ thống không thể thay thế hoàn toàn nhà quản lý trong việc ra quyết định. Các quyết định còn phụ thuộc vào rất nhiều yếu tố không biết trước như thiên tai, dịch bệnh,v.v. nhưng lại có ảnh hưởng rất lớn tới kết quả của bài toán.

Tài liệu tham khảo

- Machine Learning Cơ Bản - Vũ Hữu Tiệp
- “Why Should I Trust You?” Explaining the Predictions of Any Classifier - Marco Tulio Ribeiro - Sameer Singh - Carlos Guestrin
- Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study - Zhuoning Yuan - Xun Zhou - Tianbao Yang