BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÁO CÁO

ĐỒ ÁN MÁY HỌC ỨNG DỤNG

Đề tài

PHÂN TÍCH DỮ LIỆU TỘI PHẠM Ở LOS ANGELES DỰA TRÊN GIẢI THUẬT GÔM CỤM

LOS ANGELES CRIME DATA ANALYSIS BASED ON CLUSTER ALGORITHM

Giảng viên hướng dẫn:

TS. Lưu Tiến Đạo

Nhóm sinh viên thực hiện:

Nguyễn Phú Lâm B2105548

Phan Trần Thảo Duy B2111789

Trần Tấn Đạt B2105571

Huỳnh Thanh Phong B2207555

Cần Thơ, 11/2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÁO CÁO

DỰ ÁN MÁY HỌC ỨNG DỤNG

Đề tài

PHÂN TÍCH DỮ LIỆU TỘI PHẠM Ở LOS ANGELES DỰA TRÊN GIẢI THUẬT GÔM CỤM

LOS ANGELES CRIME DATA ANALYSIS BASED ON CLUSTER ALGORITHM

Giảng viên hướng dẫn:

TS. Lưu Tiến Đạo

Nhóm sinh viên thực hiện:

Nguyễn Phú Lâm B2105548

Phan Trần Thảo Duy B2111789

Trần Tấn Đạt B2105571

Huỳnh Thanh Phong B2207555

Cần Thơ, 11/2024

LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến quý thầy cô Trường Đại học Cần Thơ, Trường Công nghệ thông tin và Truyền thông đặc biệt là thầy TS. Lưu Tiến Đạo vì đã tận tình hỗ trợ nhóm hoàn thành đồ án.

Do lần đầu tiếp xúc với chủ đề này nên nhóm còn gặp nhiều khó khăn, do đó việc sai sót là không thể tránh khỏi. Mong thầy cô có thể xem xét và góp ý để bài báo cáo được tốt hơn.

Cuối lời nhóm xin gửi lời chúc sức khỏe đến quý thầy cô, chúc quý thầy cô gặp nhiều niềm vui và nhiều thành đạt.

Xin chân thành cảm ơn!

Cần Thơ, ngày 11 tháng 11 năm 2024

Thay lời cảm ơn

(Nhóm trưởng)

Nguyễn Phú Lâm

NHẬN XÉT CỦA GIẢNG VIÊN

……

……

……

……

……

……

Cần Thơ, ngày tháng năm 2024

Giảng viên hướng dẫn

(Ký và ghi rõ họ tên)

Lưu Tiến Đạo

PHÂN CÔNG CÔNG VIỆC

Họ tên - MSSV	Tên công việc	Đánh giá
Nguyễn Phú Lâm B2105548	Nhóm trưởng Phân công công việc, hỗ trợ và điều hành dự án nhóm. Thực hiện Chương 1, Chương 2. Tổng hợp dự án và đưa ra kết luận.	100%
Phan Trần Thảo Duy B2111789	Thành viên Hỗ trợ hoàn thành dự án: Đưa ra ý kiến đóng góp. Thực hiện Chương 6 và tham gia thiết kế slide.	100%
Trần Tấn Đạt B2105571	Thành viên Hỗ trợ hoàn thành dự án: Đưa ra ý kiến đóng góp. Thực hiện Chương 4 và thiết kế slide.	100%
Huỳnh Thanh Phong B2207555	Thành viên Hỗ trợ hoàn thành dự án: Đưa ra giải pháp trong mã nguồn. Thực hiện Chương 3, 5 và tham gia thiết kế slide.	100%

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Thuật ngữ đầy đủ
LADP	Los Angeles Police Department
BJA	Bureau of Justice Assistance
IACA	International Association of Crime Analysts

TÓM TẮT

Tội phạm ở Los Angeles ngày càng đa dạng và phức tạp, bao gồm tội phạm bạo lực và tội phạm tài sản. Chúng thay đổi theo thời gian và đạt đỉnh điểm vào những năm 1970 và 1990. Kể từ đầu những năm 2020, tội phạm đã gia tăng đáng kể ở Los Angeles.

Dự án này phân tích các báo cáo tội phạm từ năm 2020 từ tất cả các khu vực lân cận của Los Angeles bằng giải thuật gôm cụm để đưa ra giải pháp dự đoán loại tội phạm đã xảy ra, thời gian và địa điểm nhất định.

ABSTRACT

Crime in Los Angeles is increasingly diverse and complex, including violent crime and property crime. They changed over time and peaked in the 1970s and 1990s. Since the early 2020s, crime has increased dramatically in Los Angeles.

This project analyzes crime reports from 2020 from all Los Angeles neighborhoods using clustering algorithms to come up with a solution to predict the type of crime that has occurred, when and where it has occurred.

PHẦN NỘI DUNG

CHƯƠNG 1

MÔ TẢ BÀI TOÁN

Trong bối cảnh hiện nay, tội phạm đang trở nên tinh vi về công nghệ trong việc phạm tội. Thách thức mà các cơ quan tình báo và thực thi pháp luật phải đối mặt là khó khăn trong việc phân tích khối lượng lớn dữ liệu liên quan đến tội phạm và các hoạt động khủng bố, do đó các cơ quan cần biết kỹ thuật để bắt tội phạm và vẫn dẫn đầu trong cuộc đua không hồi kết giữa tội phạm và cơ quan thực thi pháp luật.

1.1. Tội phạm ở Los Angeles

Tội phạm là một hiện tượng xã hội lâu đời như chính các xã hội, và mặc dù sẽ không bao giờ có một xã hội thoát khỏi tội phạm - chỉ vì nó cần mọi người trong xã hội đó phải suy nghĩ và hành động giống nhau - các xã hội luôn tìm cách giảm thiểu nó và ngăn chặn nó.

Trong lịch sử Hoa Kỳ hiện đại, tỷ lệ tội phạm gia tăng sau Thế chiến thứ hai, đạt đỉnh điểm từ những năm 1970 đến đầu những năm 1990. Tội phạm bạo lực tăng gần gấp bốn lần từ năm 1960 đến đỉnh điểm vào năm 1991. Tội phạm tài sản tăng hơn gấp đôi trong cùng thời kỳ. Tuy nhiên, kể từ những năm 1990, tội phạm ở Hoa Kỳ đã giảm đều đặn. Cho đến gần đây, công tác phòng chống tội phạm vẫn được nghiên cứu dựa trên các phương pháp xã hội và hành vi nghiêm ngặt, nhưng những phát triển gần đây trong Phân tích dữ liệu đã cho phép một cách tiếp cận mang tính định lượng hơn trong chủ đề này.

Năm 2012, Sở cảnh sát Los Angeles báo cáo rằng tội phạm đã giảm trong thành phố trong năm thứ 10 liên tiếp[1]. Năm 2013, Los Angeles báo cáo **296** vụ giết người trong thành phố, tương ứng với tỷ lệ 6,3 trên 100.000 dân—giảm đáng kể so với năm 1980, khi tỷ lệ giết người mọi thời đại là 34,2 trên 100.000 dân được báo cáo trong năm.

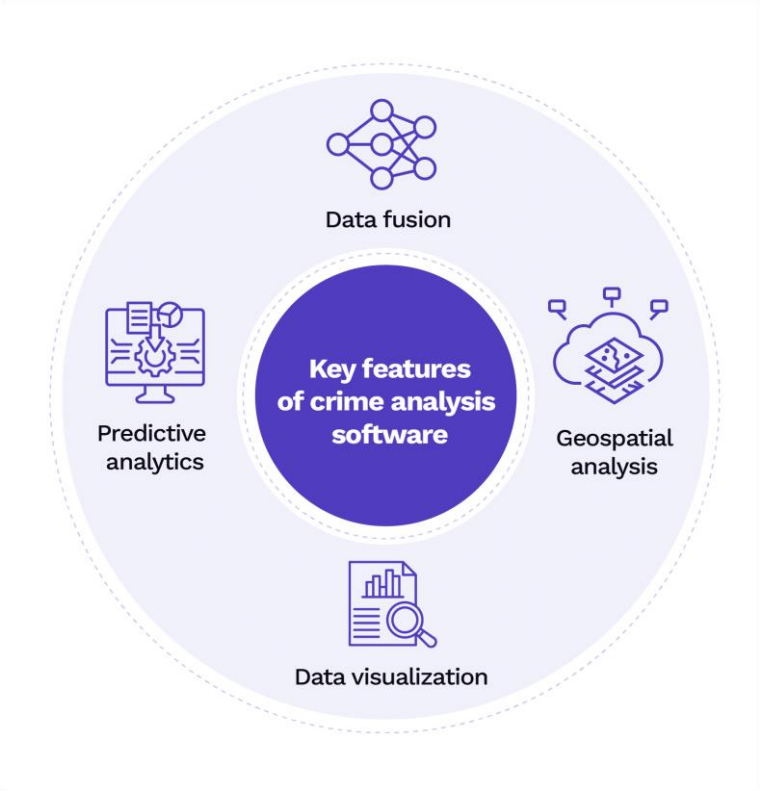
Dự án này sẽ khai phá bộ dữ liệu gồm các báo cáo tội phạm từ năm 2020 từ tất cả các khu vực lân cận của Los Angeles và chúng tôi sẽ tạo ra một mô hình dự đoán loại tội phạm đã xảy ra, theo thời gian và địa điểm.

[1] https://www.nydailynews.com/2013/01/08/crime-rate-in-los-angeles-falls-for-10th-straight-year-making-it-the-safest-big-city-in-america-but-cell-phone-thefts-are-way-up/

1.2. Phân tích tội phạm

Phân tích tội phạm được định nghĩa là các quá trình phân tích cung cấp thông tin liên quan đến mô hình tội phạm và mối tương quan xu hướng để hỗ trợ cơ quan chức năng lập kế hoạch triển khai các nguồn lực để ngăn chặn và trấn áp các hoạt động tội phạm. Điều quan trọng là phải phân tích tội phạm vì những lý do sau:

- Phân tích tội phạm để thông báo kịp thời cho cơ quan thực thi pháp luật về xu hướng tội phạm chung và cụ thể.
- Phân tích tội phạm để tận dụng nhiều thông tin hiện có trong hệ thống tư pháp và phạm vi công cộng.



Hình 1. Đặc điểm then chốt của một phần mềm phân tích tội phạm

Tỷ lệ tội phạm đang thay đổi nhanh chóng và phân tích được cải thiện tìm thấy các mô hình tội phạm ẩn, nếu có, mà không có bất kỳ kiến thức rõ ràng nào trước về các mô hình này. Các mục tiêu chính của phân tích tội phạm bao gồm:

1. Trích xuất các mô hình tội phạm bằng cách phân tích dữ liệu tội phạm và tội phạm có sẵn

2. Dự đoán tội phạm dựa trên sự phân bố không gian của dữ liệu hiện có và dự đoán tỷ lệ tội phạm bằng các kỹ thuật khai thác dữ liệu khác nhau

3. Phát hiện tội phạm

Theo Hiệp hội các nhà phân tích tội phạm quốc tế (IACA)[2], phân tích tội phạm được định nghĩa như sau: Một nghề nghiệp và quy trình trong đó một tập hợp các kỹ thuật định lượng và định tính được sử dụng để phân tích dữ liệu có giá trị cho các cơ quan cảnh sát và cộng đồng của họ. Nó bao gồm phân tích tội phạm và tội phạm, nạn nhân tội phạm, rối loạn, vấn đề chất lượng cuộc sống, vấn đề giao thông và hoạt động nội bộ của cảnh sát, và kết quả của nó hỗ trợ điều tra và truy tố tội phạm, hoạt động tuần tra, chiến lược phòng ngừa và giảm tội phạm, giải quyết vấn đề và đánh giá các nỗ lực của cảnh sát.

Mục đích chính của phân tích tội phạm là hỗ trợ (tức là hỗ trợ) các hoạt động của sở cảnh sát. Các chức năng này bao gồm điều tra hình sự, bắt giữ và truy tố; hoạt động tuần tra; chiến lược phòng, chống tội phạm; giải quyết vấn đề; và đánh giá và trách nhiệm giải trình của các nỗ lực của cảnh sát. Nếu không có cơ quan cảnh sát, phân tích tội phạm sẽ không tồn tại. Mặc dù chung chung, định nghĩa này bao gồm một loạt các hoạt động trong đó các nhà phân tích tội phạm hỗ trợ các cơ quan cảnh sát. Một ấn phẩm của Cục Hỗ trợ Tư pháp (BJA) cung cấp một cái nhìn tổng quan về cách một chức năng phân tích mang lại lợi ích cho các cơ quan thực thi pháp luật theo chín cách (Cục Hỗ trợ Tư pháp, 2005):

1. Giúp giải quyết các cuộc điều tra hình sự.

2. Tăng khả năng khởi tố tội phạm:

3. Hỗ trợ giám đốc điều hành và nhiệm vụ của cơ quan

4. Chủ động thông báo cho cán bộ thực thi pháp luật về xu hướng tội phạm và phát triển các đánh giá về mối đe dọa, tính dễ bị tổn thương và rủi ro

5. Đào tạo nhân viên thực thi pháp luật và nhân viên tình báo khác

6. Hỗ trợ phát triển cơ sở dữ liệu trên máy vi tính để tổ chức thông tin và tình báo .

7. Thúc đẩy mối quan hệ có ý nghĩa với các nhân viên thực thi pháp luật khác.

[2] Hiệp hội các nhà phân tích tội phạm quốc tế được thành lập vào năm 1990 để giúp các nhà phân tích tội phạm trên toàn thế giới cải thiện kỹ năng của họ và tạo ra các liên hệ có giá trị, giúp các cơ quan thực thi pháp luật sử dụng tốt nhất phân tích tội phạm và ủng hộ các tiêu chuẩn về hiệu suất và kỹ thuật trong chính nghề nghiệp. Chúng tôi hoàn thành những mục tiêu này thông qua đào tạo, kết nối mạng và xuất bản.

8. Đảm bảo tuân thủ các luật và quy định của địa phương, tiểu bang, bộ lạc và liên bang.

9. Cung cấp hỗ trợ cho các trung tâm hợp nhất.

1.3. Hướng giải quyết của bài toán

Như vậy, bài báo cáo này ứng dụng các kỹ thuật phân tích, bài toán sử dụng tập dữ liệu tội phạm **Crime data from 2020 to present**, Los Angeles. Sử dụng các giải thuật gôm cụm và vẽ biểu đồ trực quan để có cái nhìn chi tiết hơn, giúp du khách và cơ quan chức năng có thể nắm rõ và dự đoán tội phạm có thể xảy ra, góp phần đảm bảo an toàn và chất lượng cuộc sống.

Để giải quyết vấn đề này, nhóm chúng tôi áp dụng chu trình Khoa học dữ liệu đầy đủ bao gồm các bước sau:

- Kiểm tra chất lượng của dữ liệu và thực hiện tất cả các hành động cần thiết để làm sạch tập dữ liệu.
- Khám phá dữ liệu để hiểu các biến và tạo trực giác về dữ liệu.
- Chuẩn hóa dữ liệu và Chuyển đổi dữ liệu để chuẩn bị tập dữ liệu cho các thuật toán học tập (nếu cần).
- Đào tạo / Kiểm tra việc tạo dữ liệu để đánh giá hiệu suất của các mô hình và tinh chỉnh các siêu tham số của chúng.
- Lựa chọn và đánh giá mô hình. Đây sẽ là mục tiêu cuối cùng; Tạo ra một mô hình dự đoán xác suất của từng loại tội phạm dựa trên vị trí và ngày.

Dự án gồm 6 chương được mô tả như sau:

CHƯƠNG 1: MÔ TẢ BÀI TOÁN

CHƯƠNG 2: MÔ TẢ DỮ LIỆU VÀ Ý NGHĨA CỦA DỮ LIỆU

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

CHƯƠNG 4: XỬ LÝ DỮ LIỆU

CHƯƠNG 5: PHÂN TÍCH DỮ LIỆU

CHƯƠNG 6: CÀI ĐẶT VÀ CẤU HÌNH

CHƯƠNG 7: ĐÁNH GIÁ MÔ HÌNH

CHƯƠNG 2
MÔ TẢ DỮ LIỆU, Ý NGHĨA CỦA DỮ LIỆU

Tập dữ liệu tội phạm từ năm 2020 đến nay cho chúng ta biết xu hướng và mô hình trong hoạt động tội phạm. Tập dữ liệu rất có giá trị cho các cơ quan thực thi pháp luật, các nhà hoạch định chính sách, nhà nghiên cứu và công chúng để đưa ra quyết định sáng suốt về các chiến lược can thiệp và phòng ngừa tội phạm. Khi làm việc với bộ dữ liệu tội phạm kéo dài khung thời gian này, phải chú ý xem xét một số yếu tố để xác định xem nó có phù hợp để phân tích tội phạm hay không.

2.1. Giới thiệu tập dữ liệu

Tập dữ liệu cung cấp[3] là một đoạn trích của tập dữ liệu liên quan đến các tội phạm được báo cáo ở Los Angeles từ năm 2020. Nó chứa thông tin về các sự cố tội phạm khác nhau, bao gồm các chi tiết như số báo cáo, ngày và giờ xảy ra, địa điểm, loại tội phạm, đặc điểm của các bên liên quan và các thông tin liên quan khác.

Tập dữ liệu thuộc sở hữu của Sở cảnh sát Los Angeles, được tạo vào ngày 11 tháng 02 năm 2020 được cập nhật liên tục mỗi hai tháng một lần.

2.2. Mô tả dữ liệu

Tập dữ liệu đang khai thác gồm 990293 dòng, 28 thuộc tính và mỗi dòng là một vụ án. Được cập nhật gần nhất vào ngày 30 tháng 10 năm 2024 (sẽ được cập nhật thêm nữa).

Bảng dưới đây mô tả chi tiết thông tin các cột của tập dữ liệu:

Bảng 1.1. Bảng mô tả các đặc trưng tập dữ liệu tội phạm

STT	Tên trường	Mô tả	Loại dữ liệu
1	DR_NO	Số bộ phận hồ sơ: Số hồ sơ chính thức được tạo thành từ năm có 2 chữ số, ID khu vực và 5 chữ số	Text
2	Date Rptd	Ngày báo cáo sự việc. Có định dạng: MM/DD/YYYY	Floating Timestamp

[3] https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

STT	Tên trường	Mô tả	Loại dữ liệu
3	DATE OCC	Ngày sự việc xảy ra. Có định dạng: MM/DD/YYYY	Floating Timestamp
4	TIME OCC	Thời gian sự việc xảy ra (định dạng 24 giờ quân sự).	Text
5	AREA	LAPD có 21 Đồn Cảnh sát Cộng đồng được gọi là Khu vực Địa lý trong bộ. Các khu vực địa lý này được đánh số tuần tự từ 1-21.	Text
6	AREA NAME	Tên của khu vực hoặc trạm tuần tra, thường liên quan đến một địa danh hoặc cộng đồng lân cận. **21 Khu vực Địa lý** hoặc Bộ phận Tuần tra cũng được chỉ định tên tham chiếu đến một địa danh hoặc cộng đồng xung quanh mà nó chịu trách nhiệm. Ví dụ, 77th Street Division nằm ở giao lộ của South Broadway và 77th Street, phục vụ các khu phố ở Phía Nam Los Angeles.	Text
7	Rpt Dist No	Mã gồm bốn chữ số đại diện cho một tiểu khu vực trong Khu vực địa lý. Tất cả các hồ sơ tội phạm tham chiếu đến "RD" mà nó xảy ra để so sánh thống kê[4].	Text
8	Part 1-2	Phân loại tội phạm theo mức độ nghiêm trọng.	Number
9	Crm Cd	Mã số đại diện cho loại tội phạm. Cho biết tội ác đã thực hiện. (Tương tự như Bộ luật Hình sự 1)	Text
10	Crm Cd Desc	Diễn giải mã tội phạm. Loại tội phạm được xác định Bộ luật hình sự được cung cấp.	Text

[4] Tìm các quận báo cáo LAPD trên LA City GeoHub tại http://geohub.lacity.org/datasets/c4f83909b81d4786aa8ba8a74a4b4db1_4

STT	Tên trường	Mô tả	Loại dữ liệu
11	Mocodes	Modus Operandi: Các hoạt động liên quan đến nghi phạm thực hiện tội phạm. [5]	Text
12	Vict Age	Tuổi của nạn nhân	Text
13	Vict Sex	Giới tính nạn nhân: Nam, nữ, không xác định	Text
14	Vict Descent	Mã gốc: A - Khác Châu Á B - Đen C - Trung Quốc D - Campuchia F - Philippines G - Guamania H - Tây Ban Nha / Latinh / Mexico I - Người Mỹ da đỏ / Alaska bản địa J - Nhật Bản K - Hàn Quốc L - Lào O - P khác - Thái Bình Dương Đảo S - Samoa U - Hawaii V - Việt Nam W - Trắng X - Không xác định Z - Người Ấn Độ gốc Á	Text
15	Premis Cd	Loại cấu trúc, phương tiện hoặc vị trí xảy ra vụ án.	Number
16	Premis Desc	Xác định Mã tiền đề được cung cấp.	Text
17	Weapon Used Cd	Loại hung khí được sử dụng trong vụ án.	Text
18	Weapon Desc	Xác định Mã vũ khí được sử dụng được cung cấp.	Text
19	Status	Tình trạng vụ án. (IC là mặc định)	Text
20	Status Desc	Xác định Mã trạng thái được cung cấp.	Text
21	Crm Cd 1	Cho biết tội ác đã thực hiện. Bộ luật hình sự 1 là bộ luật chính và nghiêm trọng nhất. Bộ luật hình sự 2, 3 và 4 lần lượt là các tội phạm ít nghiêm trọng. Số lượng lớp tội phạm thấp hơn nghiêm trọng hơn.	Text

[5] https://data.lacity.org/api/views/y8tr-7khq/files/3a967fbd-f210-4857-bc52-60230efe256c?download=true&filename=MO%20CODES%20(numerical%20order).pdf

STT	Tên trường	Mô tả	Loại dữ liệu
22	Crm Cd 2	Có thể chứa mã cho một tội phạm bổ sung, ít nghiêm trọng hơn Bộ luật Hình sự 1.	Text
23	Crm Cd 3	Có thể chứa mã cho một tội phạm bổ sung, ít nghiêm trọng hơn Bộ luật Hình sự 1.	Text
24	Crm Cd 4	Có thể chứa mã cho một tội phạm bổ sung, ít nghiêm trọng hơn Bộ luật Hình sự 1.	Text
25	LOCATION	Địa chỉ đường phố của vụ án phạm tội được làm tròn đến hàng trăm khối gần nhất để duy trì danh tính.	Text
26	Cross Street	Đường giao cắt với địa chỉ chính.	Text
27	LAT	Vĩ độ	Number
28	LON	Kinh độ	Number

2.3. Ý nghĩa của tập dữ liệu

Tập dữ liệu tội phạm từ **Los Angeles Police Department (LAPD)** cung cấp thông tin chi tiết về các vụ án đã được báo cáo trong khu vực thành phố Los Angeles. Tập dữ liệu này không chỉ hữu ích trong việc nghiên cứu và phân tích tội phạm mà còn mang lại nhiều ý nghĩa và giá trị trong các lĩnh vực như quản lý xã hội, quy hoạch đô thị, và chính sách an ninh công cộng.

Những ý nghĩa và tầm quan trọng của tập dữ liệu tội phạm – Crime Data:

2.3.1. Xác định xu hướng và mô hình tội phạm

Tập dữ liệu giúp các cơ quan chức năng và nhà nghiên cứu xác định các xu hướng tội phạm theo thời gian và không gian. Ví dụ:

- Thời gian xảy ra tội phạm: Xác định những thời điểm cao điểm xảy ra tội phạm như ban đêm hoặc cuối tuần.
- Phân bố địa lý: Phân tích khu vực nào có tỷ lệ tội phạm cao để ưu tiên phân bổ nguồn lực.
- Loại tội phạm phổ biến: Giúp phát hiện các loại tội phạm thường xuyên xảy ra như trộm cắp, tấn công hay sử dụng vũ khí.

2.3.2. Hỗ trợ công tác phòng ngừa và giảm thiểu tội phạm

Phân tích dữ liệu giúp xây dựng chiến lược phòng ngừa tội phạm hiệu quả hơn. Các cơ quan an ninh có thể sử dụng dữ liệu này để:

- Phân bổ cảnh sát tuần tra vào các khu vực có tỉ lệ tội phạm cao.
- Đưa ra các khuyến cáo an toàn cho người dân dựa trên loại tội phạm và khu vực xảy ra.
- Xác định điểm nóng tội phạm (crime hotspots) và triển khai các biện pháp can thiệp phù hợp.

2.3.3. Nâng cao hiệu quả của chính sách công và quản lý đô thị

Tập dữ liệu có thể được sử dụng bởi các nhà hoạch định chính sách để:

- Cải thiện an ninh cộng đồng bằng cách điều chỉnh chính sách hoặc chương trình dựa trên phân tích tội phạm.
- Hỗ trợ quy hoạch đô thị nhằm giảm bớt các điều kiện tạo thuận lợi cho tội phạm (ví dụ: cải thiện ánh sáng công cộng, tăng cường giám sát an ninh ở các khu vực công cộng).
- Đánh giá hiệu quả của các chiến dịch an ninh và các biện pháp can thiệp xã hội.

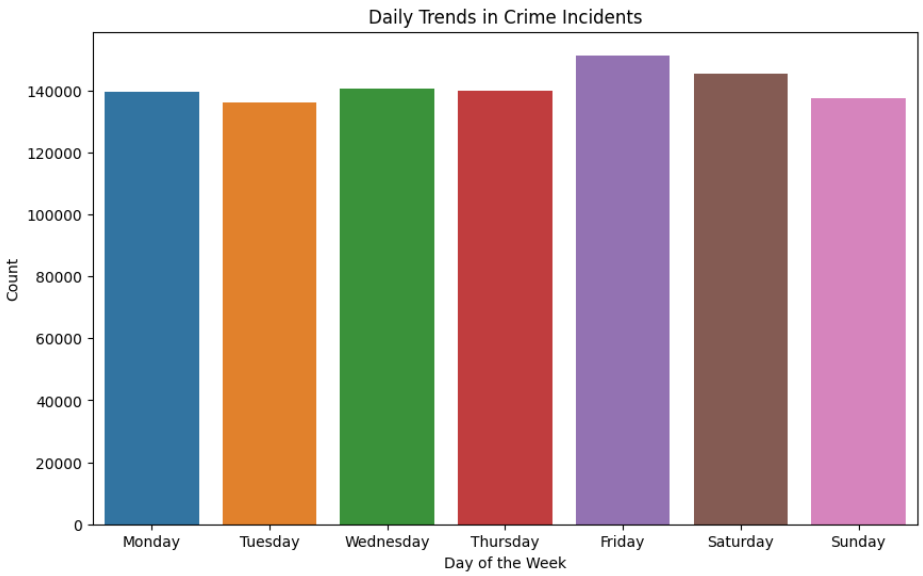
2.3.4. Nghiên cứu khoa học và giáo dục

Các nhà nghiên cứu, sinh viên và chuyên gia trong các lĩnh vực như tội phạm học, khoa học xã hội, khoa học dữ liệu có thể sử dụng tập dữ liệu này để:

- Phân tích các yếu tố liên quan đến tội phạm như độ tuổi, giới tính, dân tộc của nạn nhân.
- Sử dụng các kỹ thuật học máy (machine learning) và trí tuệ nhân tạo (AI) để dự đoán và ngăn ngừa tội phạm.
- Tạo ra các mô hình dự báo tội phạm để nâng cao an toàn cộng đồng.

2.3.5. Tăng cường tính minh bạch và trách nhiệm giải trình

Đồ thị dưới đây mô tả một cách nhìn về tập dữ liệu:



Hình 2.1. Những ngày thường xảy ra vụ án ở Los Angeles

Tóm lại, tập dữ liệu tội phạm của thành phố Los Angeles không chỉ có giá trị trong việc **phân tích tội phạm** mà còn có thể được sử dụng để hỗ trợ **công tác phòng ngừa, hoạch định chính sách, và nâng cao chất lượng sống của người dân**. Việc tận dụng dữ liệu này một cách hiệu quả có thể mang lại những thay đổi tích cực trong việc bảo vệ an ninh và an toàn cộng đồng.

Chính vì lẽ đó, dự án của chúng tôi tập trung phân tích tập dữ liệu này với các giải thuật gôm cụm nhằm đưa ra giải pháp giảm thiểu tội phạm trong khu vực Los Angeles. Các chương sau sẽ trình bày chi tiết điều này.

CHƯƠNG 3

CƠ SỞ LÝ THUYẾT - KMEANS

3.1. Giới thiệu thuật toán Kmeans

Thuật toán KMeans là một phương pháp phân cụm không giám sát được sử dụng phổ biến trong khai thác dữ liệu và học máy. Mục tiêu của thuật toán là chia một tập hợp dữ liệu thành K cụm khác nhau, sao cho các điểm dữ liệu trong cùng một cụm có sự tương đồng cao hơn với nhau so với các điểm dữ liệu trong các cụm khác.

Cơ sở toán học của thuật toán KMeans liên quan đến việc tối ưu hóa khoảng cách giữa các điểm dữ liệu và centroid của các cụm.

3.2. Một số khái niệm trong giải thuật gôm cụm - Kmeans

3.2.1 Định nghĩa về cụm (Cluster)

Mỗi cụm C_k chứa một tập hợp các điểm dữ liệu $X=\{x_1,x_2,...,x_n\}$. Mục tiêu là tối ưu hóa việc phân chia các điểm dữ liệu vào K cụm.

Một số phương pháp tìm số cụm tối ưu:

Bảng 3.1. Các phương pháp xác định cụm của Kmeans

STT	Thuật toán	Đặc điểm
1	Phương pháp Elbow	Định số cụm tối ưu bằng cách tính toán hàm mục tiêu J(K) cho nhiều giá trị K và tìm điểm khuỷu tay trên đồ thị.
2	Phương pháp Silhouette	Phương pháp này đo lường độ tương đồng của mỗi điểm dữ liệu với các cụm mà nó thuộc về so với các cụm khác.
3	Phương pháp Gap Statistic	Phương pháp này so sánh tổng khoảng cách bình phương của các cụm thực tế với các cụm ngẫu nhiên.
4	Phương pháp Davies-Bouldin Index	Dùng chỉ số Davies-Bouldin (DBI) đo lường chất lượng của các cụm bằng cách tính tỷ lệ giữa khoảng cách giữa các centroid và khoảng cách trung bình trong cùng một cụm

3.2.2. Centroid của cụm

Centroid là điểm đại diện cho một cụm, được tính bằng cách lấy trung bình tọa độ của tất cả các điểm dữ liệu thuộc về cụm đó.

Centroid được sử dụng để xác định xem một điểm dữ liệu mới có thuộc về cụm nào hay không.

Vị trí của các centroid ảnh hưởng lớn đến kết quả phân cụm cuối cùng. Nếu centroid được tính toán chính xác, các cụm sẽ phân tách tốt và có ý nghĩa hơn.

Đối với cụm C_k centroid μ_k được tính như sau:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \epsilon C_k} x_i$$

Trong đó $|C_k|$ là số lượng điểm trong cụm C_k.

3.2.3. Khoảng cách

Các cách khoảng cách trong máy học:

Bảng 3.2. Khoảng cách trong máy học

STT	Tên	Công thức
1	Khoảng cách Euclidean	$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$
2	Khoảng cách Manhattan	$d(p,q) = \sum_{i=1}^{n} \lvert p_i - q_i \rvert$
3	Khoảng cách Cosine	$d(p,q) = 1 - \frac{p.q}{\lvert\lvert p \rvert\rvert . \lvert\lvert q \rvert\rvert}$
4	Khoảng cách Minkowski	$d(p,q) = (\sum_{i=1}^{n} \lvert p_i - q_i \rvert^p)^{\frac{1}{p}}$

3.2.4. Hàm mục tiêu

Mục tiêu của thuật toán KMeans là giảm thiểu tổng khoảng cách bình phương giữa các điểm dữ liệu và centroid của các cụm. Hàm mục tiêu được định nghĩa như sau:

$$J = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} d(x_i, \mu_k)^2$$

Hàm này tính tổng khoảng cách bình phương từ tất cả các điểm dữ liệu đến centroid tương ứng của chúng. KMeans cố gắng tìm ra các giá trị của C_k và μ_k để tối thiểu hóa J.

3.2.5. Quy trình lặp lại

Thuật toán KMeans lặp lại hai bước cho đến khi không có thay đổi đáng kể trong centroid hoặc không có điểm nào bị thay đổi cụm:

+ Gán cụm: Gán mỗi điểm dữ liệu vào cụm có centroid gần nhất.
+ Cập nhật centroid: Cập nhật các centroid dựa trên điểm dữ liệu mới đã được gán.

3.2.6. Kết thúc thuật toán

Quá trình lặp lại cho đến khi một trong các điều kiện sau xảy ra:

+ Centroid không thay đổi (hoặc thay đổi rất ít).
+ Tất cả các điểm dữ liệu không bị thay đổi cụm.
+ Đạt đến số lần lặp tối đa đã chỉ định.

CHƯƠNG 4
XỬ LÝ DỮ LIỆU

3.1. Giới thiệu

Trong quá trình phân tích dữ liệu tội phạm, việc xử lý dữ liệu đóng vai trò quan trọng trong việc đảm bảo tính chính xác và độ tin cậy của các kết quả phân tích. Chương này sẽ trình bày các bước xử lý dữ liệu đã thực hiện trên tập dữ liệu tội phạm của Sở Cảnh sát Los Angeles (LAPD).

3.2. Khám phá dữ liệu

Trước khi tiến hành các bước xử lý, chúng tôi thực hiện các bước khám phá dữ liệu cơ bản để hiểu rõ cấu trúc và đặc điểm của tập dữ liệu. Các phương thức như *df.info()* và *df.describe()* được sử dụng để đánh giá số lượng dòng và cột, cũng như các thuộc tính thống kê của dữ liệu. Kết quả cho thấy rằng tập dữ liệu có hơn 990 nghìn dòng và 28 cột với các cột chứa thông tin đa dạng về các vụ án, điều này cung cấp cái nhìn tổng quát về chất lượng và tính khả thi của dữ liệu cho phân tích. Số lượng này vẫn còn tiếp tục cập nhật theo Sở cảnh sát Los Angeles.

3.3. Làm sạch dữ liệu

Một trong những vấn đề phổ biến trong tập dữ liệu là sự tồn tại của các giá trị thiếu (NaN) và dữ liệu không hợp lệ. Để đảm bảo tính chính xác và toàn vẹn của tập dữ liệu, chúng tôi áp dụng hai phương thức hiệu quả: loại bỏ các dòng chứa giá trị thiếu và thay thế giá trị thiếu bằng số không.

Loại bỏ các dòng chứa giá trị thiếu: Chúng tôi sử dụng hàm *df.dropna()* để loại bỏ tất cả các dòng mà có ít nhất một giá trị thiếu. Phương pháp này giúp đảm bảo rằng chỉ những dòng dữ liệu đầy đủ mới được giữ lại, từ đó nâng cao độ chính xác của các phân tích sau này.

Thay thế giá trị thiếu bằng số không: Để giữ lại tất cả các dòng trong tập dữ liệu, chúng tôi áp dụng hàm *df.fillna(0)* nhằm thay thế các giá trị thiếu bằng 0. Phương pháp này đặc biệt hữu ích trong những trường hợp mà giá trị không xác định có thể coi là không xảy ra, từ đó giữ nguyên cấu trúc của dữ liệu.

3.4. Xử lý dữ liệu trùng lặp

Một vấn đề khác có thể gặp phải trong dữ liệu là sự trùng lặp. Để đảm bảo tính độc nhất của mỗi dòng trong tập dữ liệu, chúng tôi tiến hành kiểm tra và loại bỏ các cột bị trùng lặp bằng cách sử dụng lệnh df = df.loc[:, ~df.columns.duplicated()].

Phương pháp này giúp giữ lại các cột duy nhất trong tập dữ liệu, đảm bảo rằng không có thông tin nào bị lặp lại không cần thiết.

Ngoài ra, chúng tôi cũng sử dụng hàm *df.duplicated()* để xác định các dòng trùng lặp trong tập dữ liệu. Nếu phát hiện bất kỳ dòng nào bị trùng lặp, chúng sẽ được xử lý và loại bỏ, nhằm cải thiện độ chính xác của các mô hình phân tích.

3.5. Xử lý ngày tháng năm

Trong tập dữ liệu tội phạm, các cột chứa thông tin về thời gian cần được chuyển đổi từ dạng chuỗi (string) sang kiểu dữ liệu ngày tháng (datetime) để có thể thực hiện các phép toán và phân tích thống kê chính xác. Để thực hiện việc này, chúng tôi sử dụng phương thức pd.to_datetime() của thư viện Pandas. Phương thức này tự động nhận diện và chuyển đổi các chuỗi ngày tháng với các định dạng khác nhau thành đối tượng datetime64[ns], giúp chuẩn hóa dữ liệu và đảm bảo tính nhất quán cho các bước phân tích sau.

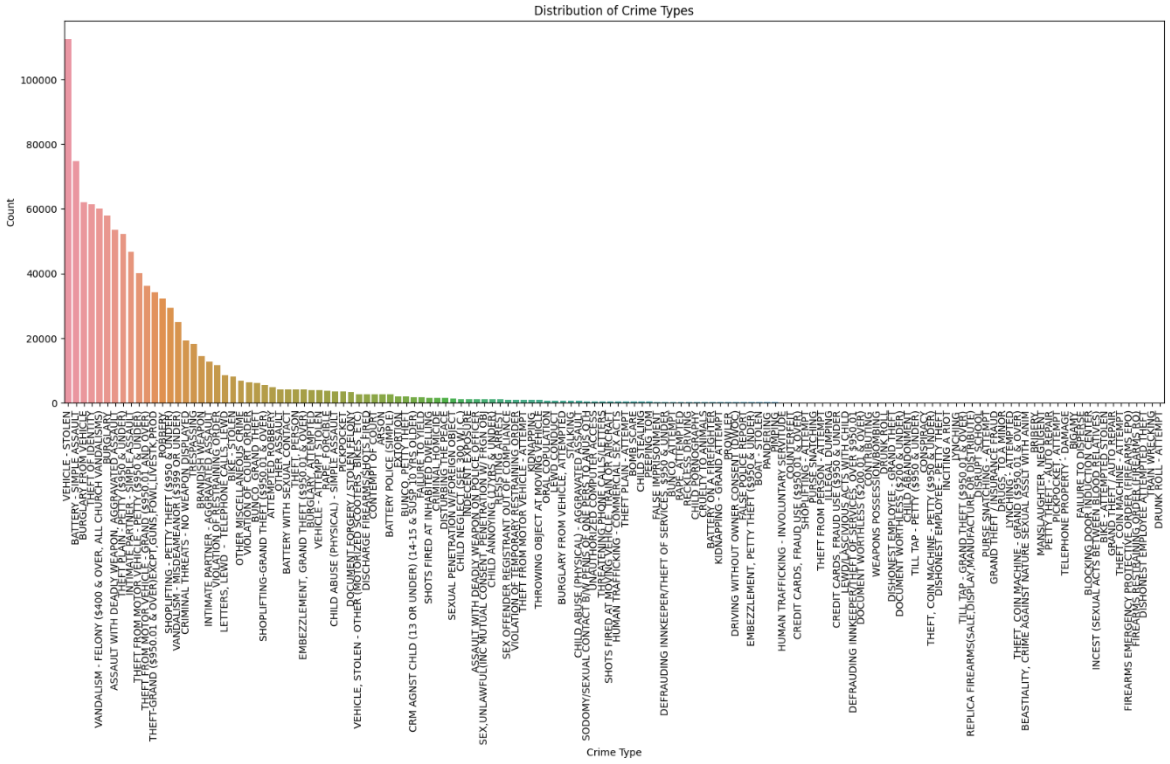
3.6. Kết luận

Thông qua quá trình xử lý dữ liệu, chúng tôi đã khắc phục được các vấn đề về giá trị thiếu và dữ liệu trùng lặp, đồng thời đã làm sạch dữ liệu một cách hiệu quả. Những bước xử lý này đã tạo nền tảng vững chắc cho việc áp dụng các phương pháp tiếp theo, từ đó giúp nâng cao tính chính xác và độ tin cậy của các kết quả phân tích.

CHƯƠNG 5
PHÂN TÍCH DỮ LIỆU

Tập dữ liệu Tội phạm ở Los Angeles từ năm 2020 đến nay có thể ứng dụng nghiên cứu ở nhiều lĩnh vực, đặc biệt là phục vụ cho công tác phòng chống Tội phạm ở Los Angeles. Áp dụng các thư viện có sẵn của kaggle, ta có thể trực quan hoá dữ liệu và rút ra được một số dự đoán cũng như kết luận sau đây:

5.1. Biểu đồ thanh thể hiện sự phân bổ các loại tội phạm



Hình 5.1. Biểu đồ thanh thể hiện sự phân bổ các loại tội phạm

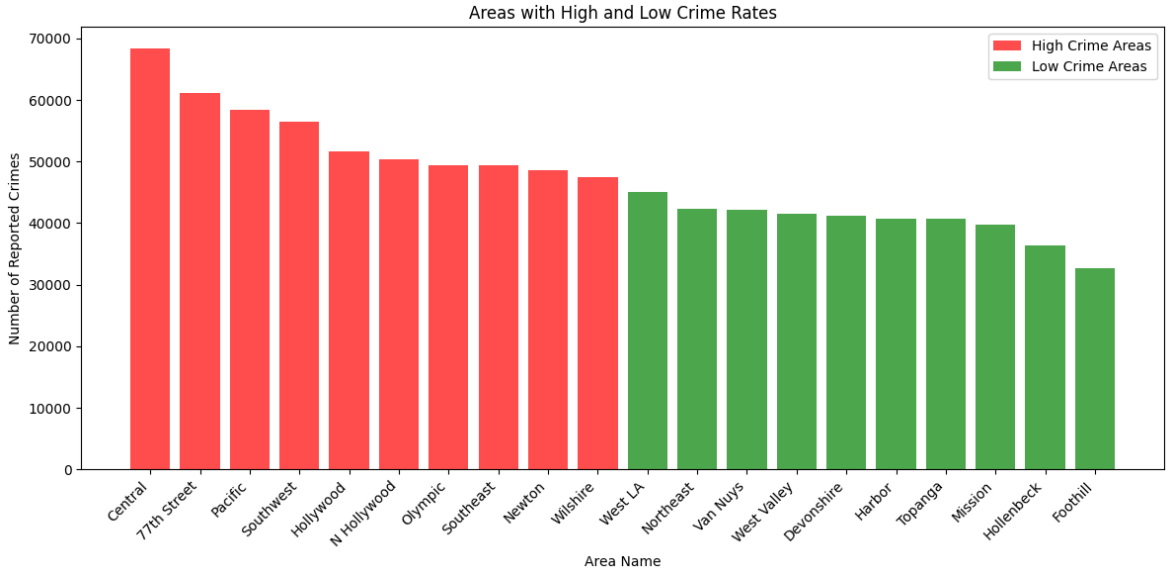
- Trục X (Crime Type) : Hiển thị nhiều loại tội phạm khác nhau, mỗi nhãn đại diện cho một loại tội phạm khác nhau.
- Trục Y (Count) : Hiển thị số lượng, tần suất của từng loại tội phạm, cho biết tần suất xảy ra của từng loại trong tập dữ liệu.

Qua biểu đồ trên ta có thể thấy được ở Los Angeles, loại tội phạm bị tố cáo nhiều nhất chính là: Trộm phương tiện giao thông (Vehical - stolen)

Theo một số bản tin ở Los Angles, người dân có thói quen đậu xe ô tô một thời gian dài trên phố, cộng với sự lan truyền về thông tin có lỗi an toàn ảo mật xảy ra ở

các loại xe như *Kia* và *Hyundai*[6]. Thậm chí còn có 1 số video clip hướng dẫn về việc mở khoá để trộm các hãng xe trên. Mặc dù các hãng xe đã có cập nhật về tính bảo mật và an toàn, có sự giảm đáng kể nhưng tình trạng trộm cắp xe vẫn là một vấn đề nhức nhối ở Los Angeles.

5.2. Biểu đồ cột thể hiện sự tương quan giữa khu vực và số lượng tội phạm mỗi khu vực



Hình 5.2. Sự tương quan giữa khu vực và số lượng tội phạm

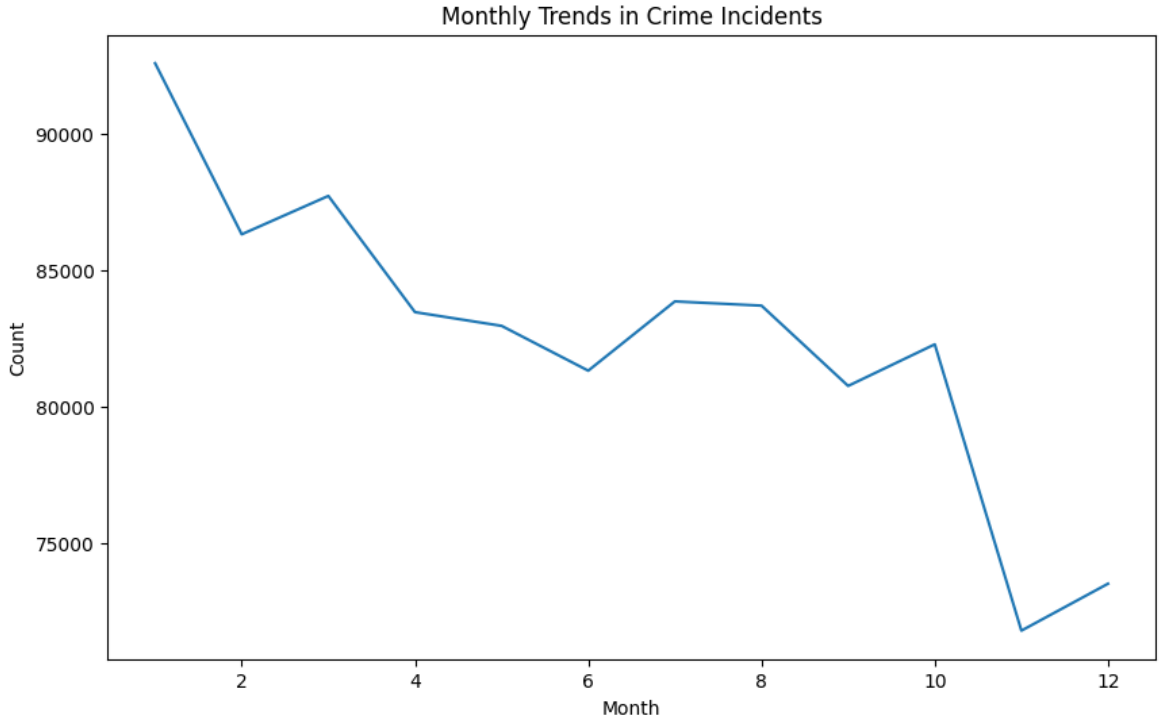
Trục X: Hiển thị tên các khu vực ở Los Angeles có dữ liệu về tội phạm

Trục Y: Hiển thị số lượt báo cáo tội phạm ở các khu vực.

Biểu đồ cột trên ta có thể nhận thấy được các khu vực có số lượng báo cáo tội phạm cao, cần được chú ý nhiều hơn là các khu vực có màu đỏ (trên 50000), các khu vực dưới 50000 là các cột màu xanh. Dù là một thành phố nổi tiếng khắp thế giới về một thành phố hiện đại, "Kinh đô điện ảnh", nhưng chính Central (Trung tâm thành phố) lại chính là nơi có số lượng tội phạm cao nhất bậc nhất. Theo một số nguồn tin Los Angeles có một trong những cộng đồng người vô gia cư lớn nhất nước Mỹ, và nhiều người trong số này tập trung ở khu vực trung tâm, đặc biệt là khu Skid Row. Tình trạng vô gia cư, cùng với những vấn đề về sức khỏe tâm thần và nghiện ngập, có thể góp phần vào sự gia tăng của một số loại tội phạm.

[6] https://www.yahoo.com/news/los-angeles-car-thefts-remain-204302765.html

5.3. Biểu đồ đường thể hiện xu hướng tội phạm ở các tháng



Hình 5.3. Biểu đồ đường thể hiện xu hướng tội phạm

Biểu đồ đường này cho thấy xu hướng số vụ phạm tội tại Los Angeles theo từng tháng trong năm. Dưới đây là một số điểm đáng chú ý:

- Tổng số vụ phạm tội có xu hướng giảm dần từ đầu năm đến cuối năm. Tháng đầu năm có số vụ phạm tội cao nhất, khoảng hơn 90,000 vụ, sau đó giảm dần.

- Mức thấp nhất: Số vụ phạm tội giảm xuống mức thấp nhất vào tháng 11, với số lượng khoảng dưới 75,000 vụ.

- Biến động nhẹ: Mặc dù có một xu hướng giảm chung, vẫn có một số biến động nhẹ trong các tháng giữa năm, với một số tháng có số vụ tăng nhẹ trước khi tiếp tục giảm.

Biểu đồ này phản ánh những yếu tố như sự gia tăng kiểm soát tội phạm vào cuối năm, thay đổi trong điều kiện xã hội hoặc thời tiết, hoặc các chương trình an ninh tăng cường của thành phố Los Angeles.

4. Biểu đồ nhiệt thể hiện mối tương quan giữa loại tội phạm và địa điểm gây án

Hình 5.4. Biểu đồ nhiệt thể hiện mối tương quan giữa loại tội phạm và địa điểm

Dựa trên biểu đồ nhiệt thể hiện sự phân bố của các loại tội phạm ở các khu vực ta có thể thấy được một số thông tin sau

- 3 loại tội phạm có sự phân bố cao chính là các vụ trộm cắp, trộm phương tiện giao thông, tội hành hung

Một số khu vực nhất định nổi bật với các ô sáng ở nhiều dòng, cho thấy các khu vực này có mức độ tội phạm tổng thể cao hơn. Ví dụ, khu Central và Southwest thường có màu sáng trong các loại tội phạm phổ biến, cho thấy đây là các khu vực có tần suất tội phạm cao. Loại tội phạm phân bố ở nhiều khu vực với số lượng cao cũng chính là Trộm phương tiện giao thông.

Như vậy, việc phân lớp và gom nhóm các tội phạm giúp cho việc điều tra và quản lý an ninh các khu vực ở Los Angeles được dễ dàng hơn. Từ đó, có thể đưa ra được các chính sách cũng như biện pháp cho từng loại tội phạm trong các khu vực, cụ thể là nhóm đối tượng trộm phương tiện giao thông – một vấn đề cũng khá nhức nhối tại Los Angeles. Ngoài ra, còn giúp cho các hãng xe kịp thời phát hiện các lỗi về phần mềm để cải thiện sản phẩm ngày càng tốt hơn.

CHƯƠNG 6
CÀI ĐẶT VÀ CẤU HÌNH

6.1. Cài đặt Kmeans cho tập dữ liệu tội phạm

Gôm cụm dữ liệu dựa trên Kinh độ (LON) và Vĩ độ (LAT) để phân tích mối quan hệ giữa vị trí địa lý và số lượng tội phạm cũng như là các đặc điểm về tội phạm ở mỗi khu vực.

Yêu cầu: Dữ liệu ở cột LAT và LON đã được phân tích và xử lý sạch ở các bước trước.

Sử dụng ngôn ngữ Python để cái đặt máy học.

6.1.2. Cài đặt và Import các thư viện cần thiết.

Có thể sử dụng Pip Installs Packages(Pip) để cài đặt các thư viện cho Python.

Bảng 6.1. Các thư viện cần thiết

STT	Thư viện Python	Chức năng
1	Pandas	Pandas là thư viện giúp dễ dàng thao tác và phân tích dữ liệu dạng bảng, hỗ trợ các cấu trúc dữ liệu như DataFrame (bảng dữ liệu) và Series (cột hoặc hàng trong bảng).
2	matplotlib	Matplotlib là thư viện vẽ đồ thị mạnh mẽ, trong đó pyplot là một mô-đun cung cấp các hàm để vẽ các biểu đồ (như biểu đồ đường, biểu đồ thanh, biểu đồ phân tán). Nó là công cụ hữu ích trong việc trực quan hóa dữ liệu, giúp bạn hiểu rõ hơn về các mẫu hoặc xu hướng trong tập dữ liệu.
3	seaborn	seaborn là thư viện trực quan hóa dữ liệu xây dựng dựa trên matplotlib, giúp tạo ra các biểu đồ đẹp mắt và dễ đọc hơn. Nó cung cấp các biểu đồ thống kê nâng cao (như biểu đồ phân tán với phần đánh dấu mật độ dữ liệu, biểu đồ phân bố, biểu đồ tương quan), hữu ích cho phân tích dữ liệu.
4	sklearn	Thư viện phát triển mạnh về máy học của Python KMeans là thuật toán phân cụm được sử dụng từ thư viện scikit-learn

5	numpy	Numpy là thư viện toán học hỗ trợ làm việc với mảng đa chiều (ndarrays) và cung cấp các công cụ tính toán số học hiệu quả. Nó rất phổ biến trong khoa học dữ liệu, cung cấp các hàm tính toán vector hóa giúp tăng tốc độ xử lý so với cách tính toán thông thường trong Python.

6.1.3. Chuẩn bị dữ liệu

Sử dụng hàm *pandas.read_csv(filepath_or_buffer)* có sẵn trong thư viện Pandas để đọc dữ liệu từ file CSV (Comma-Separated Values) và chuyển đổi dữ liệu đó thành một DataFrame và chọn những cột cần thiết cho giải thuật K-Means.

6.1.4. Tìm số cụm tối ưu bằng phương pháp Elbow

Sử dụng *khoảng cách Euclidean* cho tập giá trị phân cụm.

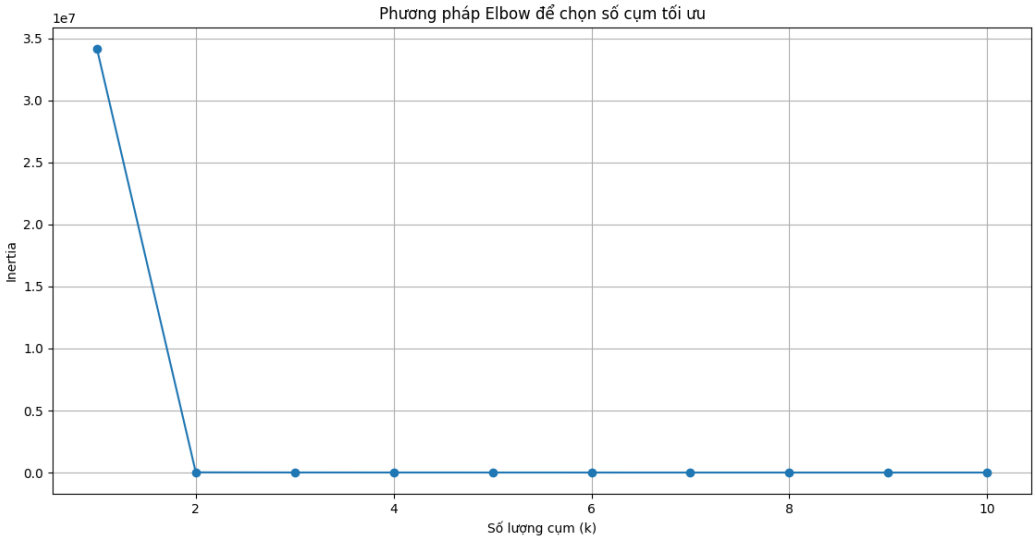
Chọn dải giá trị cho số cụm k: Thử nghiệm với một dải giá trị cho k(1,2,…,11), vì ta thấy k càng tăng giá trị Inertia không đổi nữa.

Chạy K-means: Với mỗi giá trị k, thực hiện phân cụm bằng K-Means và tính toán Inertia cho kết quả.

6.1.5. Vẽ biểu đồ Elbow:

+ Trục x: Số cụm k.
+ Trục y: Giá trị Inertia cho từng k.

Xác định "elbow": Vị trí elbow trên biểu đồ là giá trị k mà từ đó, Inertia giảm chậm lại rõ rệt.

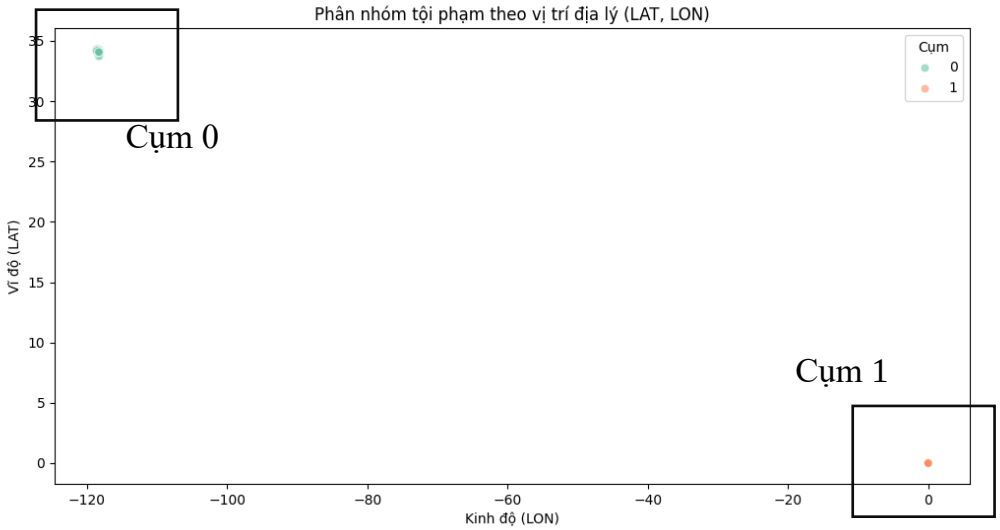


Hình 6.1. Phương pháp Elbow để chọn số cụm tối ưu

Từ biểu đồ Elbow ta chọn được K = 2. Vì khi K>2 giá trị Inertia không thay đổi.

6.2. Chạy thuật toán K-Means

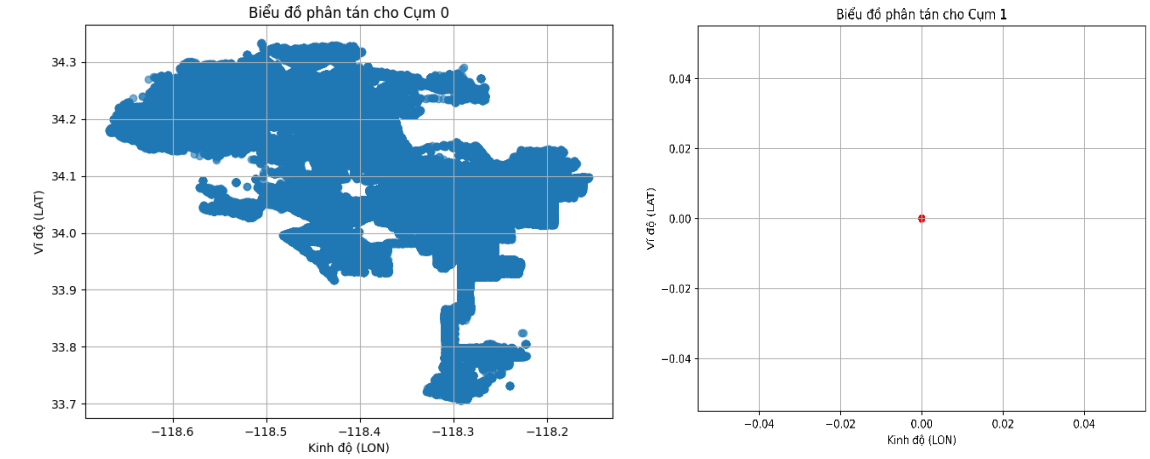
Huấn luyện mô hình K-Means với số cụm bằng 2 ta đạt được kết quả và hiển thị qua biểu đồ phân tán như sau:



Hình 6.2. Biểu đồ phân tán theo cụm

Trong đó:

Các biểu đồ phân tán của từng cụm



Hình 6.3. Cụm 0 theo Kinh độ và Cụm 1 theo Kinh độ

Bảng dưới đây trình bày đặc trưng theo cụm.

Bảng 6.2. Đặc trưng theo cụm

Đặc trưng	**Cụm 0**	**Cụm 1**
Tội phạm phổ biến nhất (Crm Cd Desc)	VEHICLE - STOLEN	BATTERY - SIMPLE ASSAULT
Địa điểm xảy ra phổ biến nhất (Premis Desc)	STREET	SINGLE FAMILY DWELLING
Tổng số vụ	750128	2260

6.3. Cấu hình máy tính

Bảng 6.3. Cấu hình máy tính

Yêu cầu phần cứng	Bộ xử lý (CPU)	Tối thiểu: 4 lõi (quad-core) với tốc độ xung nhịp từ 2.0 GHz trở lên. Khuyến nghị: 6-8 lõi (hexacore hoặc octacore) với tốc độ xung nhịp từ 3.0 GHz trở lên, để xử lý nhanh hơn, đặc biệt là với các tập dữ liệu lớn.
	Bộ nhớ (RAM)	Tối thiểu: 8 GB. Khuyến nghị: 16 GB hoặc hơn, đặc biệt nếu bạn làm việc với các tập dữ liệu lớn hoặc sử dụng các thuật toán yêu cầu nhiều bộ nhớ.
	Bộ lưu trữ	Tối thiểu: 256 GB HDD.

		Khuyến nghị: 512 GB SSD hoặc hơn. SDD giúp tăng tốc độ đọc/ghi dữ liệu, điều này rất hữu ích khi làm việc với các tập dữ liệu lớn.
Yêu cầu phần mềm	Hệ điều hành	Bạn nên chọn một hệ điều hành mà bạn quen thuộc và có hỗ trợ tốt cho các thư viện mà bạn sẽ sử dụng.
	Ngôn ngữ	Python: Một trong những ngôn ngữ phổ biến nhất cho phân tích dữ liệu và machine learning.
	Môi trường	IDE như PyCharm hoặc Visual Studio Code: Nếu bạn thích phát triển trong môi trường lập trình tích hợp.

PHẦN KẾT LUẬN

1. Kết quả đạt được

Sinh viên vận dụng được kiến thức học trên lớp về máy học để giải quyết vấn đề trong thực tiễn. Kết quả đạt được:

- Sử dụng thành thạo ngôn ngữ python để huấn luyện các mô hình.
- Hiểu và phối hợp các phương pháp khác nhau để xử lý dữ liệu
- Phân tích dữ liệu: sử dụng các biểu bản đồ được vẽ bằng thư viện của máy học.
- Tư duy trong việc vận dụng kiến thức xã hội và kiến thức máy.
- Kmeans và FastMarkerCluster là một trong những hướng tốt để gôm cụm và đưa ra hướng giải quyết mới.

2. Hướng phát triển

- Tốc độ xử lý dữ liệu lớn cần được cải tiến.
- Phân tích nhiều hơn và kỹ hơn các khía cạnh tâm lý tội phạm
- Tạo ra các mô hình nhanh hiệu quả
- Phát triển phần mềm đưa ra giải pháp xử lý tội phạm

TÀI LIỆU THAM KHẢO

[1] FBI, "Crime Data from 2020 to Present," LADP, Los Angeles, 2020.

[2] P. N. Khang, Giáo trình khai khoáng dữ liệu, Trường Đại học Cần Thơ , 2022.

[3] Do Thanh Nghi, Pham Nguyen Khang, Giáo trình Nguyên lý máy học, Trường Đại học Cần Thơ.

[4] "w3school," W3.CSS, 2024. [Online]. Available: https://www.w3schools.com/python/. [Accessed 1999].

[5] "Machine Learning cơ bản," 2017. [Online]. Available: https://machinelearningcoban.com/2017/01/01/kmeans/.

PHỤ LỤC

Các biểu đồ được phân tích và vẽ tại:

https://www.kaggle.com/code/lacoss/ct294-crimeanalys-losangeles