ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH BỘ DỮ LIỆU VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN VỤ ÁN TẠI MỸ TRONG NĂM 2023

Nhóm 17			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Phạm Mạnh Hùng	21520901	HTTT
2	Phùng Thiên Phúc	21521297	HTTT
3	Nguyễn Anh Dĩ	21521952	CNTT

TP. HỒ CHÍ MINH – 12/2024

1.GIỚI THIỆU

Trong đề tài này, chúng tôi tập trung vào việc phân tích bộ dữ liệu phạm tội tại nước Mỹ trong năm 2023 và lựa chọn mô hình thích hợp nhất để dự đoán các thông tin cần được biết như kiểu vụ án dựa trên các đặc trưng. Vấn đề phạm tội luôn là một vấn đề nhức nhói bởi dù dưới sự can thiệp của chính quyền hay ràng buộc của luật pháp, các vụ án vẫn luôn xảy ra hàng ngày với số lượng thậm chí không hề nhỏ. Vì vậy, việc phân tích và dự đoán các yếu tố liên quan đến tội phạm, chẳng hạn như thời gian, địa điểm, loại hình tội phạm và các đặc trưng liên quan, có thể giúp các cơ quan chức năng và cộng đồng chuẩn bị tốt hơn trong việc phòng ngừa và ứng phó. Điều này không chỉ góp phần giảm thiểu rủi ro mà còn mang lại lợi ích về mặt an toàn xã hội.

Để hiện thực được toàn bộ đề tài, chúng tôi sử dụng ngôn ngữ lập trình chính và đóng vai trò nền tảng cho toàn bộ quy trình là Python. Python - một ngôn ngữ gần gũi đối với các lập trình viên một phần cú pháp dễ dàng nắm bắt nhưng điểm quan trọng nhất để chúng tôi quyết định chọn sử dụng chính là có nhiều thư viện hỗ trợ lập trình thuật toán, trực quan hóa dữ liệu, các mô hình máy học, các phương pháp tiền xử lý dữ liệu,… Một số thư viện được chúng tôi sử dụng trong đề tài như: Numpy, Pandas, Matplotlib, Seaborn, Sklearn, Random, Scipy,… . Phương pháp chúng tôi lựa chọn để xây dựng mô hình dự đoán giá chính là áp dụng các thuật toán máy học như: Random Forest Classification, Support Vector Classifier và Gradient Boosting Classifier để dựa trên các đặc trưng truyền vào và đưa ra kết quả dự đoán.

Nội dung bài báo cáo của chúng tôi bao gồm quá trình thu thập bộ dữ liệu có sẵn từ trang web **catalog.data.gov**, tiến hành tiền xử lý và phân tích thăm dò để tìm ra các biến phụ thuộc có ảnh hưởng đến biến mục tiêu, sau đó trực quan hóa các cột thuộc tính liên quan trong bộ dữ liệu. Kế đến, chúng tôi tiến hành thử nghiệm trên nhiều mô hình máy học và chọn ra mô hình có độ chính xác cao nhất trong việc dự đoán vụ án.

Để đảm bảo tính khách quan và giá trị của bộ dữ liệu khi phân tích, chúng tôi đã thảo luận thống nhất chọn bộ dữ liệu trên trang web của chính phủ Mỹ.

2. MÔ TẢ BỘ DỮ LIỆU

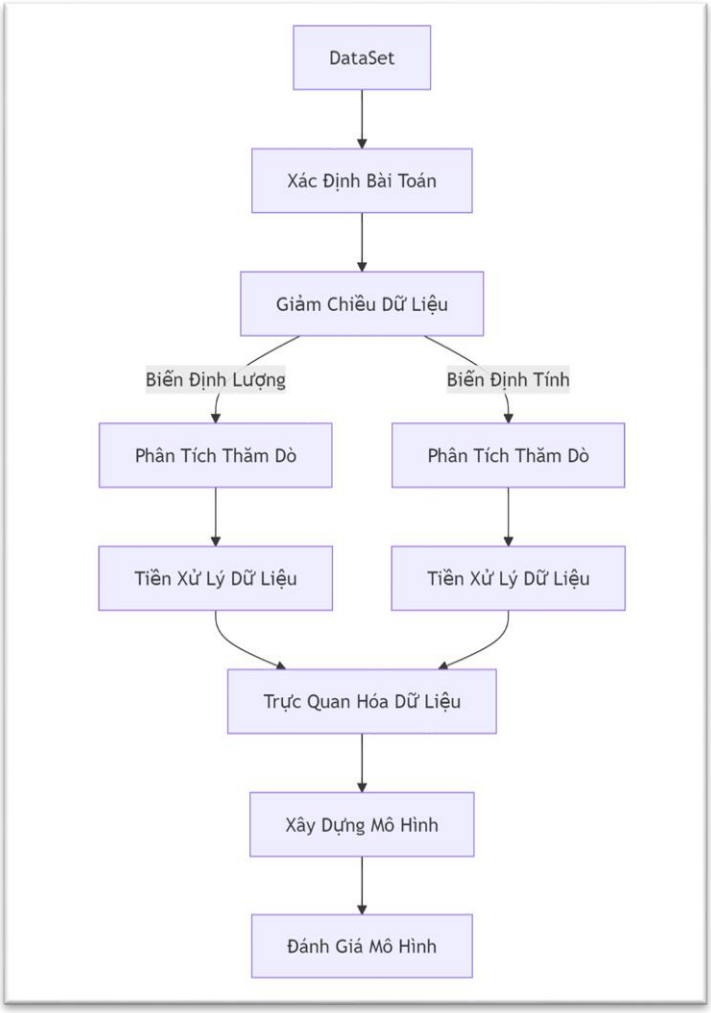
Bộ dữ liệu thu thập cung cấp thông tin chi tiết các vụ phạm tội khắp nước Mĩ từ năm 2020 cho đến 2024 tính từ thời điểm thu lại. Bộ dữ liệu bao quát được nhiều dạng phạm tội khác nhau, từ việc phạm đến thế xác như sát hại đến việc phạm đến tài sản như trộm hay cướp bóc. Dữ liệu được sắp xếp một cách tỉ mỉ, cung cấp những thông tin chi tiết về xu hướng tội phạm, sự khác biệt địa lý và xu hướng theo thời gian. Bộ dữ liệu được thu thập từ website **[1]** thể hiện chi tiết như sau:

Tên cột dữ liệu	Mô tả	Kiểu dữ liệu

DR_NO	Mã định danh duy nhất cho mỗi vụ án	int64
Date Rptd	Ngày vụ án được báo cáo	object
DATE OCC	Ngày vụ án xảy ra	object
TIME OCC	Thời gian vụ án xảy ra	int64
AREA	Khu vực hoặc quận nơi vụ án xảy ra	int64
AREA NAME	Tên mô tả của khu vực	object
Rpt Dist No	Số quận báo cáo	int64
Part 1-2	Chỉ định vụ án là Loại 1 (nghiêm trọng) hay Loại 2 (ít nghiêm trọng hơn).	int64
Crm Cd	Mã hoặc số phân loại vụ án.	int64
Crm Cd Desc	Mô tả mã vụ án.	object
Mocodes	Động cơ hoặc hoàn cảnh liên quan đến vụ án.	object
Vict Age	Tuổi của nạn nhân.	int64
Vict Sex	Giới tính của nạn nhân.	object
Vict Descent	Dân tộc hoặc nguồn gốc chủng tộc của nạn nhân.	object
Premis Cd	Mã cơ sở (ví dụ: dân cư, thương mại)	float64
Premis Desc	Mô tả cơ sở	object
Weapon Used Cd	Mã của vũ khí được sử dụng (nếu có)	float64
Weapon Desc	Mô tả vũ khí	object
Status	Tình trạng hiện tại của vụ án (ví dụ: mở, đóng)	object
Status Desc	Mô tả tình trạng của vụ án	object
Crm Cd 1	Mã tội phạm bổ sung nếu có	float64
Crm Cd 2	Mã tội phạm bổ sung nếu có	float64
Crm Cd 3	Mã tội phạm bổ sung nếu có	float64
Crm Cd 4	Mã tội phạm bổ sung nếu có	float64
LOCATION	Vị trí chung của vụ án	object
Cross Street	Ngã tư hoặc đường gần đó	object
LAT	Tọa độ vĩ độ của vị trí tội phạm	float64
LON	Tọa độ kinh độ của vị trí tội phạm	float64

3. PHƯƠNG PHÁP PHÂN TÍCH

Sau khi thu thập dữ liệu từ catalog.data.gov, quá trình phân tích dữ liệu được thể hiện chi tiết qua sơ đồ sau:



Hình 1. Quy trình PTDL.

3.1. Đánh giá bộ dữ liệu thô

Bộ dữ liệu thô ban đầu có 28 cột với 955339 dòng, trong đó có 15 biến số và 13 biến phân loại. Bộ dữ liệu thô cũng chứa một số lượng các giá trị khuyết tương đối lớn, cụ thể:

Tên cột dữ liệu	Số lượng các giá trị khuyết
Mocodes	136675
Vict Sex	130045
Vict Descent	130055
Premis Cd	12
Premis Desc	569
Weapon Used Cd	630320
Weapon Desc	630320

Status	1
Crm Cd 1	11
Crm Cd 2	886873
Crm Cd 3	953045
Crm Cd 4	955275
Cross Street	806439

Với các giá trị còn lại đều không chứa null

3.2. Thu giảm dữ liệu

Bộ dữ liệu nguyên mẫu có 986873 dòng và 28 cột (220Mb) ➔ Thu giảm dữ liệu chỉ lấy các cột cần thiết và chỉ lấy trong năm 2023: "DATE ","Date Rptd", "AREA", "TIME OCC", "Vict Sex", "Vict Descent", "Weapon Used Cd", "Vict Age", "Crm Cd","Premis Cd","Time Period". Sau thu giảm, bộ dữ liệu chỉ còn 230719 dòng và 11 cột với tên là **"crime data in USA 2023.csv".**

Ngoài ra, đối với các bước sau này chúng tôi đã dùng hàm **random** để lọc ngẫu nhiên chỉ còn lại **14000** dòng nhằm dễ dàng trong việc phân tích sau này.

3.3. Phân tích thăm dò

Bước đầu, chúng tôi thực hiện khảo sát các cột với thuộc tính describe để rút ra các thông tin từng cột: count, mean, std, min, max và tứ phân vị. Đối với các thuộc tính định tính, thu được kết thông tin về unique, top và freq. Thêm vào đó là khảo sát kiểu dữ liệu của các cột để có thể đưa về datatype hợp lý

Thực hiện thăm dò các giá trị null của cột ➔ phát hiện 3 cột chứa null ("Vict Sex", "Vict Descent", "Weapon Used Cd") ➔ thực hiện tiền xử lý

Để khi xây dựng mô hình có kết quả tốt nhất, chúng tôi cũng xử lý các giá trị ngoại lai bằng phương pháp Z_score

Dùng tương quan pearson để đánh giá độ tương quan giữa các biến số

Sau quá trình kiểm tra sử dụng tương quan, có thể nhận thấy p-value và correlation không tốt → các biến không phải làm giá trị liên tục

3.4. Tiền xử lý dữ liệu

3.4.1. Xử lý các giá trị bị khuyết

Các cột chứa giá trị khuyết như: 'Vict Sex', 'Vict Descent', 'Weapon Used Cd', chúng tôi lược bỏ các giá trị khuyết bằng cách xóa đi các hàng trong bộ dữ liệu.

Tuy vậy, đối với cột 'Vict Age' và 'Weapon Used Cd' có chứa dữ liệu là số 0 cần được lược bỏ để tránh gây nhiễu dữ liệu trong việc phân tích sau này

3.4.2. Định dạng dữ liệu

Để dữ liệu dễ dàng được phân tích, chúng tôi đã đổi kiểu các biến 'Part 1-2' và 'Weapon Type Cd' thành kiểu Object.

Đối với thuộc tính Date OCC, chúng tôi thực hiện chuyện về kiểu dữ liệu datetime để có dạng YEAR – MONTH – DAY để có định dạng nhất quán cho thuộc tính thời gian.

3.4.3. Thêm cột dữ liệu

Vì một số có giá trị là từ viết tắt hay giá trị dữ liệu phân hóa quá đa dạng nên chúng tôi thêm một số cột để dữ liệu thể hiện ra rành mạch và dễ thao tác hơn

- Thêm cột 'Weapon Type' được ánh xạ từ cột 'Weapon Used Cd' còn 5 giá trị
- Thêm cột 'Age Group' được ánh xạ từ cột 'Vict Age' còn 6 giá trị
- Thêm cột 'TIME SPAN' được ánh xạ từ cột 'Time OCC' còn 23 giá trị
- Thêm cột 'Descent Desc' được ánh xạ từ cột 'Vict Descent' còn 19 giá trị
- Thêm cột 'Sex' được ánh xạ từ cột 'Vict Sex' còn 3 giá trị

3.5. Lựa chọn và chuẩn hóa thuộc tính

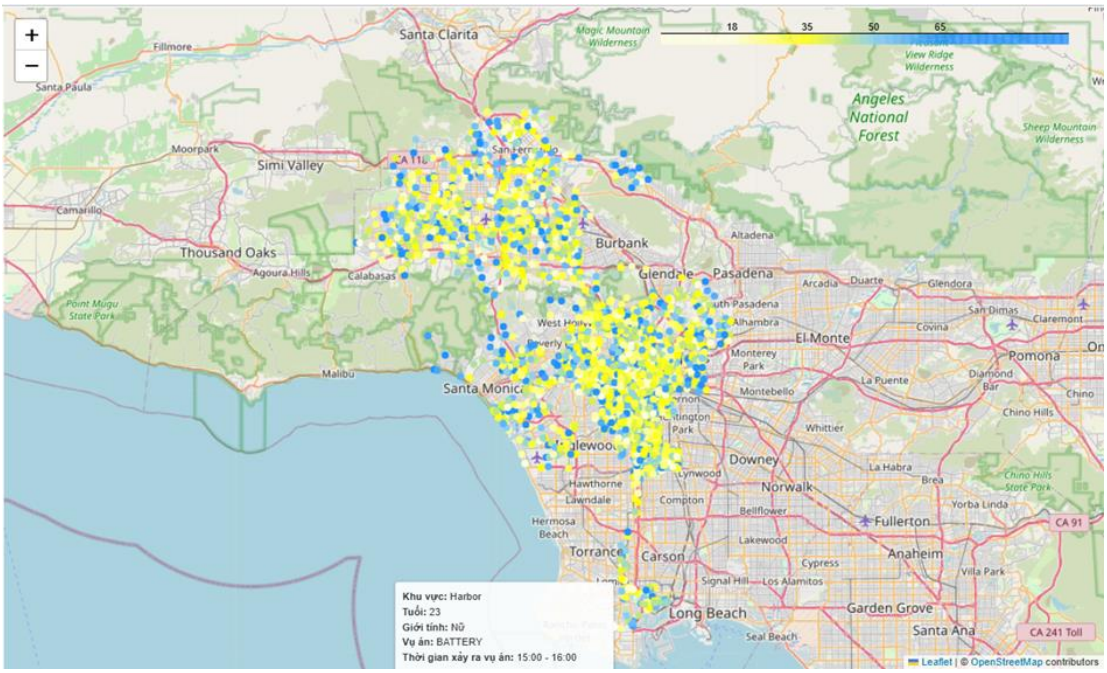
Dùng phân tích Chi-Squared (đối với các biến kiểu phân loại), chúng tôi thấy được mức độ ảnh hưởng hưởng của biến 'Crm Cd Desc' vơi các biến phân loại còn lại đều đáng kể. Tuy nhiên, đối với biến 'Part 1-2', 'Weapon Type', 'Vict Sex' cho ra p-value 'ảo' nên cần được xem kĩ hơn trong việc phân tích.

Chúng tôi lựa chọn các thuộc tính dựa trên kinh nghiệm và hiểu biết bao gồm: 'TIME SPAN', 'Age Group','AREA NAME', 'Descent Desc', 'Weapon Type', 'Vict Sex'.

Đối với các biến kiểu số, để đảm bảo tính đúng đắn của giá trị dự đoán, chúng tôi đã sử dụng phương pháp Standard Scaler để đưa chúng về cùng một giá trị trước khi đưa vào mô hình. Còn đối với các biến phân loại, chúng tôi tiến hành sử dụng phương pháp Label Encoding toàn bộ các giá trị chuỗi thành số để mô hình có thể hiểu được.

4. TRỰC QUAN HÓA

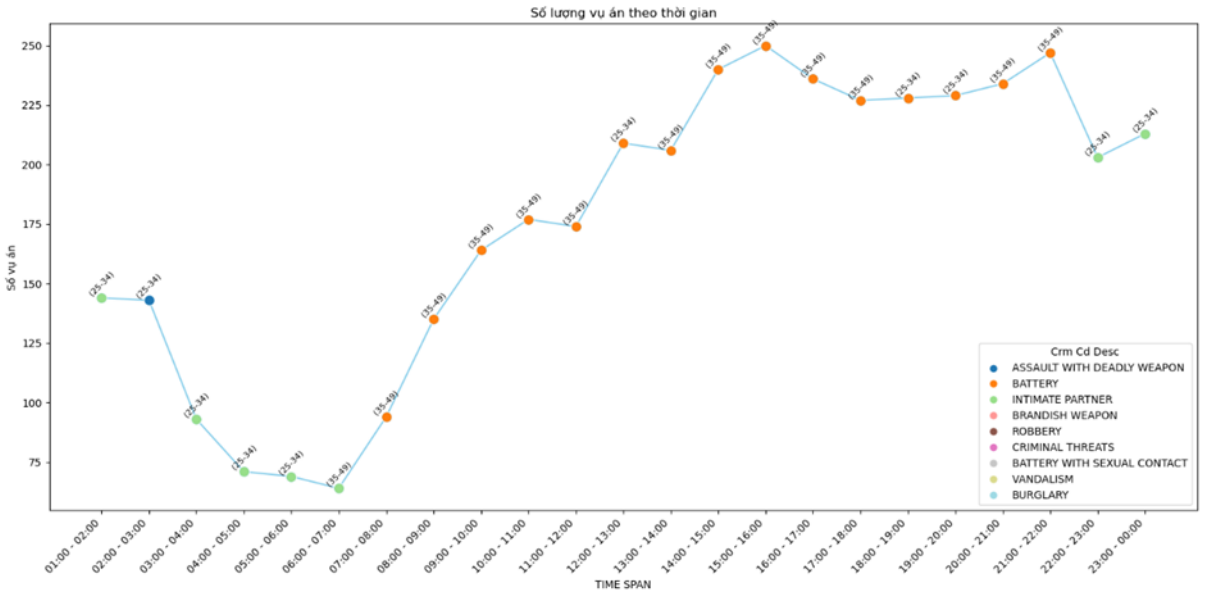
Sau việc tiền xử lý và thăm dò dữ liệu hoàn tất, chúng tôi thực hiện trực quan hóa để có thể phân tích chuyên sâu hơn cho bộ dữ liệu vụ án tại Mỹ này



Biểu đồ heatmap dang bản đồ thể hiện được mật độ phân bố của các vụ án theo khu vực trên bản đồ (cụ thể ở đây là bang California tại Mỹ), mật độ dày tại các khu vực là thành phố nổi bật như Hollywood, Van Nuys, 77th Street và dày nhất là khu vực trung tâm Los Angeles. (các chấm phân bố theo LAT và LON có sẵn trên bộ dữ liệu)

Thấy được chấm có màu là màu vàng / vàng nhạt có mật độ phân bố nổi bật nhất ám chỉ độ tuổi của nạn nhân thường gặp phải trong các vụ án có độ tuổi từ 25 đến 49.

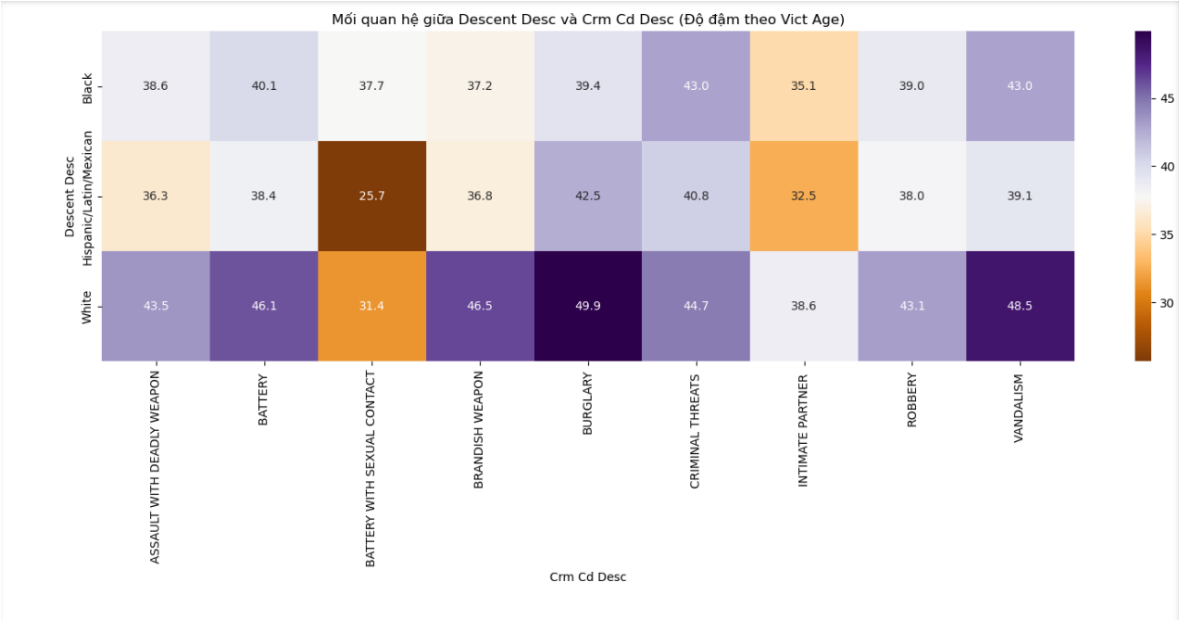
Ngoài ra, khi rê chuột vào từng chấm sẽ hiện các thông tin liên quan đến vụ án như: Khu vực xảy ra vụ án, tuổi nạn nhân, giới tính nạn nhân, loại vụ án và thời gian xảy ra vụ án.



Biểu đồ đường thể hiện số lượng vụ án diễn ra trong khung giờ 1 tiếng. Khung giờ 06:00 – 07:00 có số lượng vụ án ít nhất và 15:00 – 16:00 cùng với 21:00 – 22:00 có số lượng vụ án xảy ra nhiều nhất

Về xu hướng vụ án, từ 21:00 số lượng vụ án có xu hướng tăng lên cho đến 06:00 với kiểu vụ án thường xảy ra là INTIMATE PARTNER (vụ án liên quan đến người thân), từ 07:00 số lượng vụ án có xu hướng tăng lên cho đến 20:00 với kiểu vụ án thường xảy ra là BATTERY (dùng vũ lực) (với kiểu vụ án được thể hiện bằng chấm màu trên từng gấp khúc)

Đối với độ tuổi, từ 21:00 – 06:00 nạn nhân có xu hướng ở độ tuổi từ 25 – 34 tuổi, từ 07:00 – 20:00 nạn nhân có xu hướng ở độ tuổi từ 35 – 49 tuổi (với độ tuổi được thể hiện bằng cách ghi chú trên từng chấm màu)



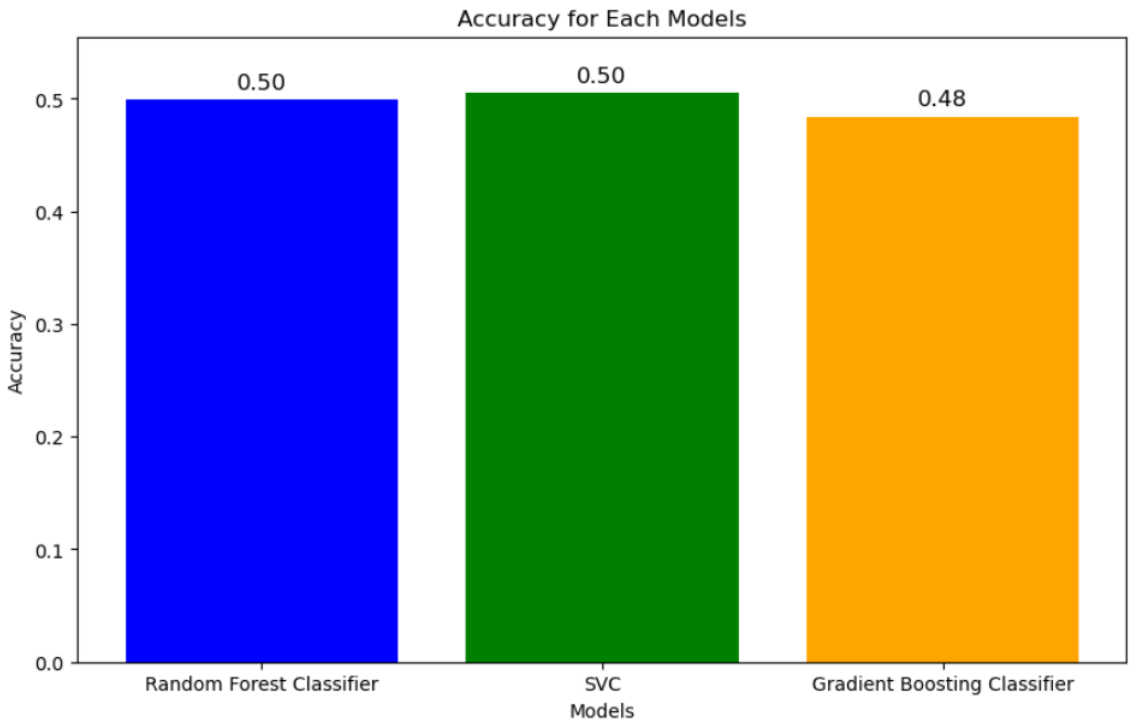
Biểu đồ heatmap cho thấy quan hệ chủng tộc nạn nhân và các loại tội phạm thể hiện bằng độ tuổi trung bình. Nhìn chung, nạn nhân là người da trắng có độ tuổi trung bình cao hơn (31 – 50), nạn nhân là người da màu hay người Latin có độ tuổi trung bình thấp hơn (26 – 43)

BATTERY WITH SEXUAL CONTACT (vũ lực có tác động tình dục), INTIMATE PARTNER (vụ án liên quan đến người thân) có xu hướng xảy ra ở độ tuổi trẻ hơn (cao nhất là 39 tuổi), BATTERY (vũ lực), BURGLARY (trộm cắp), VANDALISM (phá hoại) có xu hướng xảy ra ở độ tuổi lớn hơn (cao nhất là 50 tuổi)

5. KẾT QUẢ PHÂN TÍCH

Sau khi tiến hành chuẩn hóa dữ liệu và thực nghiệm nhiều lần trên các mô hình máy học như: Random Forest Classification, Support Vector Classifier và Gradient Boosting Classifier, chúng tôi đã thu thập được kết quả dự đoán vụ án theo biểu đồ bên dưới. Độ chính xác của mô hình được tính toán dựa trên các thang đo Accuracy.

Dưới đây là biểu đồ thể hiện Accuracy khi dự đoán từng mô hình



Thấy được, 3 mô hình đều cho kết quả Accuracy gần như nhau (0.48 - 0.5), tức các mô hình đạt chuẩn baseline ít nhất là 0.5 cả cho thấy các mô hình này chưa hoàn toàn dự đoán đúng nhưng vẫn có tiềm năng để tiếp tục đưa vào thêm các đặc trưng để nâng cao hiệu suất dự đoán sau này.

6. KẾT LUẬN

Thông qua toàn bộ nội dung được trình bày trong đồ án, chúng tôi đã xử lý, phân tích và xây dựng mô hình dự đoán cũng như đánh giá các yếu tố ảnh hưởng đến các vụ án phạm tội. Người đọc có thể xem bài báo cáo này như một nguồn tham khảo giúp hiểu rõ hơn về đặc điểm chung, thời gian, địa điểm và các yếu tố liên quan đến các vụ án. Thông qua việc trực quan hóa dữ liệu, chúng tôi đã phát hiện ra các đặc điểm phổ biến nhất của các vụ án, từ đó sử dụng các thuật toán và phương pháp phân tích để xác định những yếu tố có ảnh hưởng nhất đến tần suất và tính chất của các vụ phạm tội.

Trong số các mô hình được sử dụng, các mô hình chúng tôi sử dụng đều ở ngưỡng baseline là 50%. Hiệu suất còn hạn chế này có thể do nhiều nguyên nhân, như: dữ liệu thu thập còn chưa đủ trường, sai sót trong quá trình xử lý các thuộc tính, hoặc các mô hình máy học chưa đủ phức tạp hay chưa phù hợp. Trong tương lai, chúng tôi sẽ tiếp tục nghiên cứu và cải tiến mô hình để đạt hiệu suất cao hơn, hướng đến việc cung cấp một nguồn tham khảo đáng tin cậy cho cộng đồng phân tích trong việc phân tích, dự đoán và phòng ngừa các vụ án phạm tội. Bài cáo cáo này nên chỉ được xem với mục đích tìm hiểu chuyên sâu theo bộ dự liệu có sẵn.

TÀI LIỆU THAM KHẢO

[1] Website dataset

Link: Crime Data from 2020 to Present (12/11/2024).

[2] Website Seaborn

Link: seaborn: statistical data visualization — seaborn 0.13.2 documentation

[3] Website Matplotlib

Link: Matplotlib — Visualization with Python

[4] Website Scikit-learn

Link: scikit-learn: machine learning in Python — scikit-learn 1.6.0 documentation

[5] Website folium

Link: https://python-visualization.github.io/folium/latest/reference.html

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Phạm Mạnh Hùng	Thu giảm dữ liệu – EDA – Tiền xử lý dữ liệu, Word, Canva.
2	Nguyễn Anh Dĩ	EDA – Tiền xử lý – Trực quan hóa dữ liệu, Word, Canva.
3	Phùng Thiên Phúc	Xây dựng mô hình – Đánh giá mô hình, Canva