**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**DỰ ĐOÁN MỨC ĐỘ THIỆT HẠI TỪ CÁC VỤ TRỘM CẮP Ở SAN FRANCISCO, HOA KỲ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | | |
| STT | Họ tên | MSSV |
| 1 | Phạm Thảo Nhi | 19520815 |
| 2 | Bùi Hồng Thiên | 19520970 |

**TP. HỒ CHÍ MINH – 12/2020**

# GIỚI THIỆU

Tội phạm là một vấn đề nghiêm trọng ảnh hưởng đến tất cả các quốc gia trên thế giới. Nhằm giảm thiểu tỷ lệ tội phạm, ngày nay có rất nhiều công nghệ được sử dụng để dự đoán các chi tiết liên quan đến một vụ án cụ thể. Công ty hàng đầu về công nghệ này là PredPol. PredPol xác định ngày và địa điểm có thể xảy ra vụ án cụ thể và cử một đội tuần tra đến những khu vực này để ngăn chặn vụ án xảy ra. Tương tự như vậy, trong báo cáo này, nhóm chúng em sẽ tìm hiểu bộ dữ liệu The San Francisco Police Department’s(SFPD) Incident Report Dataset để xây dựng một mô hình dự đoán mức thiệt hại của nạn nhân của một vụ trộm cướp tại San Francisco, CA. Nhóm chúng em sẽ làm các bước như phân tích thăm dò, tiền xử lý dữ liệu, tối ưu các tham số và đánh giá trên các mô hình như KNN, RandomForest, LogisticRegression. Thông qua thực nghiệm, nhóm chúng em thấy rằng mô hình RandomForest cho kết quả khả quan nhất và vẫn còn có thể cải thiện. Chúng em xin chân thành cảm ơn.

# NỘI DUNG

Đặt vấn đề: Dự đoán mức độ thiệt hại của nạn nhân của một vụ trộm cướp tại San Francisco, CA.

Biến mục tiêu: ‘Amount’, được trích xuất từ cột ‘Incident Description’

Bảng 2.1 Tạo biến mục tiêu Amount từ Incident Description

|  |  |
| --- | --- |
| **Incident Description** | **Amount** |
| Theft, Other Property, $50-$200 | $50-$200 |

A picture containing text, gallery, screenshot

Description automatically generated

Hình 2.1: Quy trình triển khai phân tích bộ dữ liệu và phát triển mô hình

## Về bộ dữ liệu

### Thông tin bộ dữ liệu

* Tên: The San Francisco Police Department’s(SFPD) Incident Report Dataset
* Nguồn dữ liệu: https://data.sfgov()org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783
* Gồm 509811 sample và 34 feature
* Ngày đầu tiên được ghi nhận là 01/01/2018. Được cập nhật lần cuối vào 10/04/2021.
* Là một bộ dữ liệu được tổng hợp từ kho dữ liệu tội phạm để cung cấp thông tin về báo cáo các vụ án.
* Missing: có tồn tại

Thông tin thêm về các thuộc tính:

Bảng 2.2 Bảng giá trị và ý nghĩa các thuộc tính trong bộ dữ liệu

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tên thuộc tính** | **Giá trị** | **Giải thích** |  | **Tên thuộc tính** | **Giá trị** | **Giải thích** |
| Incident Datetime | [1/1/2018 0:00;9/9/2021 9:50] | Ngày và giờ xảy ra vụ án |  | Incident Code | Gồm 811 giá trị số hoặc nhiều số. | Mã hệ thống để mô tả vụ án. Một vụ án có thể có nhiều Incident Code |
| Incident Date | [1/1/2018; 12/31/2020] | Ngày xảy ra vụ án |  | Incident Category\* | Gồm 50 loại. Sẽ ghi rõ hơn bên dưới. | Được ánh xạ tới Incident Code để phân loại vụ án |
| Incident Time | [0:00, 23:59] | Thời gian xảy ra vụ án |  | Incident Subcategory\* | Gồm 72 Loại. Sẽ ghi rõ hơn bên dưới. | Được ánh xạ đến Incident Category để phân loại vụ án chi tiết |
| Incident Year | {2018,2019,  2020} | Năm xảy ra vụ án |  | Incident Description | Text | Mô tả chi tiết vụ án |
| Incident Day of Week | {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday} | Ngày trong tuần xảy ra vụ án |  | Resolution | {Cite or Arrest Adult, Exceptional Adult, Open or Active, Unfounded} | Phương án giải quyết tại chỗ |
| Report Datetime | [1/1/2018 0:00;  12/31/2020 23:59] | Thời điểm vụ án được ghi nhận |  | Intersection | Tên đường/Tên đường /… | Nút giao thông gần nhất nơi xảy ra vụ án |
| Row ID | [61868707041.0;  107720000000.0] | Giá trị duy nhất ngẫu nhiên đại diện cho sample |  | CNN | Số định danh được quy định | Mã định danh của từng Intersection |
| Incident ID | [618687; 1077195] | Mã định danh từng vụ án do hệ thống sinh ra ngẫu nhiên |  | Police District | Sẽ bổ sung thêm | Cảnh sát của quận nơi sự cố xảy ra |
| Incident Number | [0; 981424262] | Mã số vụ án được cấp ghi trên báo cáo giấy |  | Analysis Neighborhood | Tên định danh.  Sẽ bổ sung thêm | Khu vực lân cận nơi sự cố xảy ra |
| CAD Number | [1.0; 999999999.0] | Mã điều động nhân lực. Từng vụ án riêng sẽ có mã riêng. |  | Supervisor District | [1;11], nan | Số đại diện cho từng Police District |
| Report Type Code | {II, VI, VS, IS} | Mã hệ thống cho các vụ án được báo cáo |  | Latitude | [37.70798826; 37.82999075] | Vĩ độ của Intersection |
| Report Type Description | {Initial, Initial Suplement,Vehicle Initial,Vehicle Suplement, Coplogic Initial, Coplogic Supplement,} | Mô tả của các mã hệ thống cho các vụ án được báo cáo |  | Longitude | [-122.5112949;-122.3637428] | Kinh độ của Intersection |
| Filed Online | {True, nan} | Tickbox có phải được khai báo online không |  | Point | (Kinh độ; Vĩ độ) | Tọa độ của Intersection |

Incident Category\* gồm các giá trị: {'Arson', 'Assault', 'Burglary', 'Case Closure', 'Civil Sidewalks', 'Courtesy Report', 'Disorderly Conduct', 'Drug Offense', 'Drug Violation', 'Embezzlement', 'Fire Report', 'Forgery And Counterfeiting', 'Fraud', 'Gambling', 'Homicide', 'Human Trafficking (A), Commercial Sex Acts', 'Human Trafficking (B), Involuntary Servitude', 'Human Trafficking, Commercial Sex Acts', 'Larceny Theft', 'Liquor Laws', 'Lost Property', 'Malicious Mischief', 'Miscellaneous Investigation', 'Missing Person', 'Motor Vehicle Theft', 'Motor Vehicle Theft?', 'Non-Criminal', 'Offences Against The Family And Children', 'Other', 'Other Miscellaneous', 'Other Offenses', 'Prostitution', 'Rape', 'Recovered Vehicle', 'Robbery', 'Sex Offense', 'Stolen Property', 'Suicide', 'Suspicious', 'Suspicious Occ', 'Traffic Collision', 'Traffic Violation Arrest', 'Vandalism', 'Vehicle Impounded', 'Vehicle Misplaced', 'Warrant', 'Weapons Carrying Etc', 'Weapons Offence', 'Weapons Offense', nan}

Incident Subcategory\* gồm các giá trị: {'Aggravated Assault', 'Arrest', 'Arson', 'Bad Checks', 'Bribery', 'Burglary - Commercial', 'Burglary - Hot Prowl', 'Burglary - Other', 'Burglary - Residential', 'Case Closure', 'Courtesy Report', 'Disorderly Conduct', 'Drug Violation', 'Drunkenness', 'Embezzlement', 'Extortion-Blackmail', 'Fire Report', 'Forgery And Counterfeiting', 'Fraud', 'Gambling', 'Homicide', 'Human Trafficking, Commercial Sex Acts', 'Human Trafficking, Involuntary Servitude', 'Intimidation', 'Kidnapping', 'Larceny - Auto Parts', 'Larceny - From Vehicle', 'Larceny Theft - Bicycle', 'Larceny Theft - From Building', 'Larceny Theft - Other', 'Larceny Theft - Pickpocket', 'Larceny Theft - Purse Snatch', 'Larceny Theft - Shoplifting', 'Liquor Law Violation', 'Loitering', 'Lost Property', 'Manslaughter', 'Miscellaneous Investigation', 'Missing Adult', 'Missing Person', 'Motor Vehicle Theft', 'Motor Vehicle Theft (Attempted)', 'Non-Criminal', 'Other', 'Other Offenses', 'Prostitution', 'Rape', 'Rape - Attempted', 'Recovered Vehicle', 'Robbery - Carjacking', 'Robbery - Commercial', 'Robbery - Other', 'Robbery - Residential', 'Robbery - Street', 'Sex Offense', 'Simple Assault', 'Stalking', 'Stolen Property', 'Suicide', 'Suspicious Occ', 'Suspicious Package', 'Theft From Vehicle', 'Traffic Collision', 'Traffic Collision - Hit & Run', 'Traffic Violation Arrest', 'Trespass', 'Vandalism', 'Vehicle Impounded', 'Vehicle isplaced', 'Warrant', 'Weapons Offense', nan}

Analysis Neighborhood\* gồm các giá trị: {'Bayview Hunters Point', 'Bernal Heights', 'Castro/Upper Market', 'Chinatown', 'Excelsior', 'Financial District/South Beach', 'Glen Park', 'Golden Gate Park', 'Haight Ashbury', 'Hayes Valley', 'Inner Richmond', 'Inner Sunset', 'Japantown', 'Lakeshore', 'Lincoln Park', 'Lone Mountain/USF', 'Marina', 'McLaren Park', 'Mission', 'Mission Bay', 'Nob Hill', 'Noe Valley', 'North Beach', 'Oceanview/Merced/Ingleside', 'Outer Mission', 'Outer Richmond', 'Pacific Heights', 'Portola', 'Potrero Hill', 'Presidio', 'Presidio Heights', 'Russian Hill', 'Seacliff', 'South of Market', 'Sunset/Parkside', 'Tenderloin', 'Treasure Island', 'Twin Peaks', 'Visitacion Valley', 'West of Twin Peaks', 'Western Addition', nan}

### Tiền xử lý dữ liệu:

* Có 3 cách để xử lý missing được sử dụng: thay thế giá trị missing bằng giá trị khác, loại bỏ hoàn toàn điểm missing, giữ nguyên không xử lý. Và tùy từng feature sẽ có từng cách xử lý khác nhau. Sau khi xử lí đã trả về dữ liệu khá sạch.
* Do mục đích là dự đoán mức độ thiệt hại do các vụ chiếm đoạt tài sản(được thể hiện qua trường Incident Category với giá trị là ‘Larceny Theft’) nên sẽ lọc dữ liệu và lấy các sample có Incident Category’ là ‘Larceny Theft’. Sau đó dựa vào feature ‘Incident Category Description’ ta sẽ có một feature mới là ‘Amount’ - Đây chính là biến mục tiêu mà chúng ta sẽ nhắm tới.

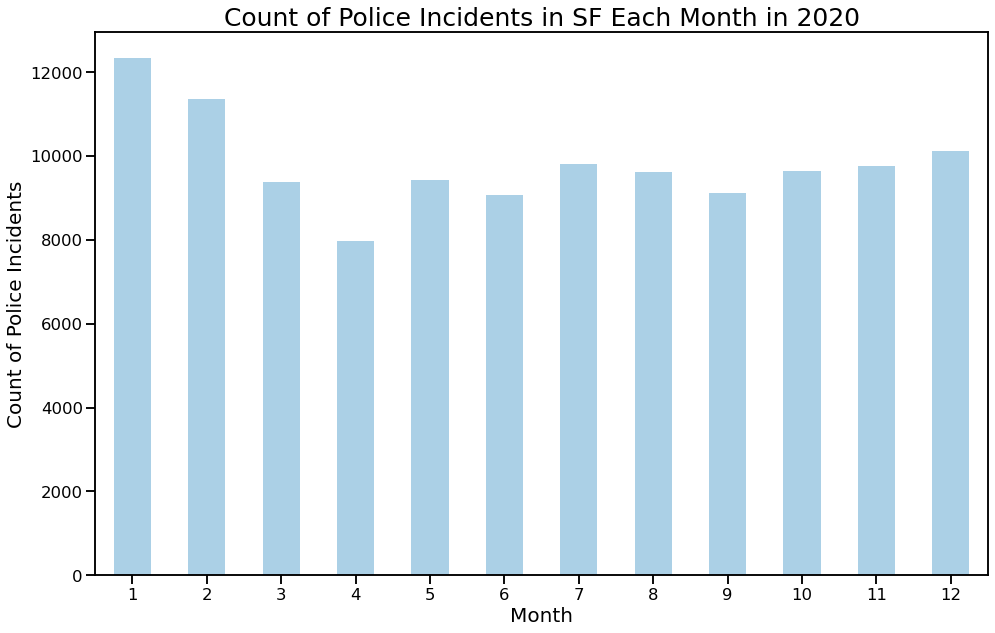
Bảng 2.3 Giá trị và ý nghĩa của biến mục tiêu ‘Amount’

|  |  |  |
| --- | --- | --- |
|  | **Giá trị** | **Giải thích** |
| Amount | ' $200-$950', ' $50-$200', ' <$50', ' >$950' | Mức độ thiệt hại trong các vụ chiếm đoạt tài sản. |

## 2.2. Phân tích thăm dò

### 2.2.1. Những insight cơ bản của bộ dữ liệu

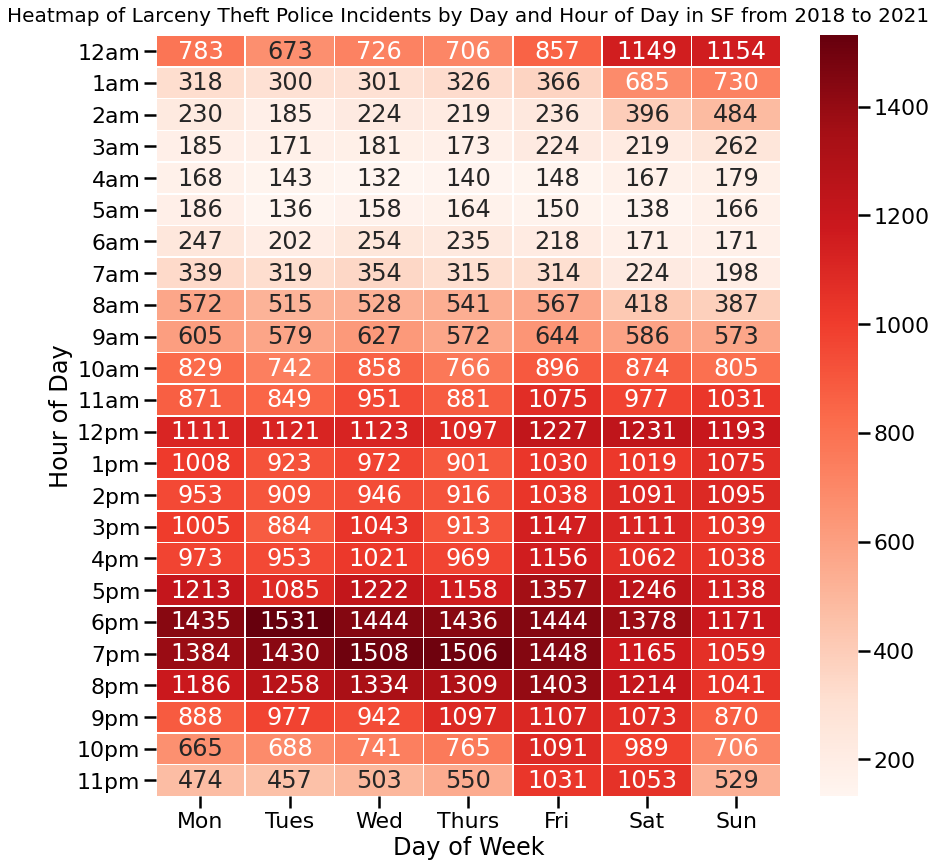
Trong năm 2020, số ca phạm tội được ghi nhận tại San Francisco giảm khoảng 21% so với năm 2019, chỉ còn 117602 ca. Hai tháng đầu năm, số ca phạm tội vẫn còn ở mức trên dưới 12000 ca. Nhưng những tháng sau đó, con số này giảm đến khoảng 20%, dao động ở mức 10000. Trong đó, tháng 4 là tháng có số ca phạm tội thấp nhất với chỉ khoảng 8000 ca. Nguyên nhân của sự sụt giảm này là do vào đầu năm 2020, đại dịch Covid-19 bắt đầu xuất hiện và bùng nổ tại San Francisco, dẫn đến việc chính thức thực hiện giãn cách toàn thành phố vào ngày 17/3/2020.



Hình 2.2 Số ca phạm tội nói chung theo từng tháng trong năm 2020

Loại hình phạm tội phổ biến trong ba năm từ 2018 đến 2020 tại San Francisco là “Larceny Theft” (trộm cắp) với hơn 128000 ca. Trong đó, loại trộm cắp xảy ra nhiều nhất là “Larceny - From Vehicle” (trộm cắp từ các phương tiện di chuyển như xe hơi, xe tải, …). (Hình 1 – Phụ lục)

Các vụ trộm cắp thường xảy ra vào khung giờ 10am - 9pm tất cả các ngày trong tuần. Ngoài ra, số vụ trộm cắp vào thứ 6, 7 trong khung giờ 10am - 12am sáng hôm sau nhiều hơn các ngày còn lại.



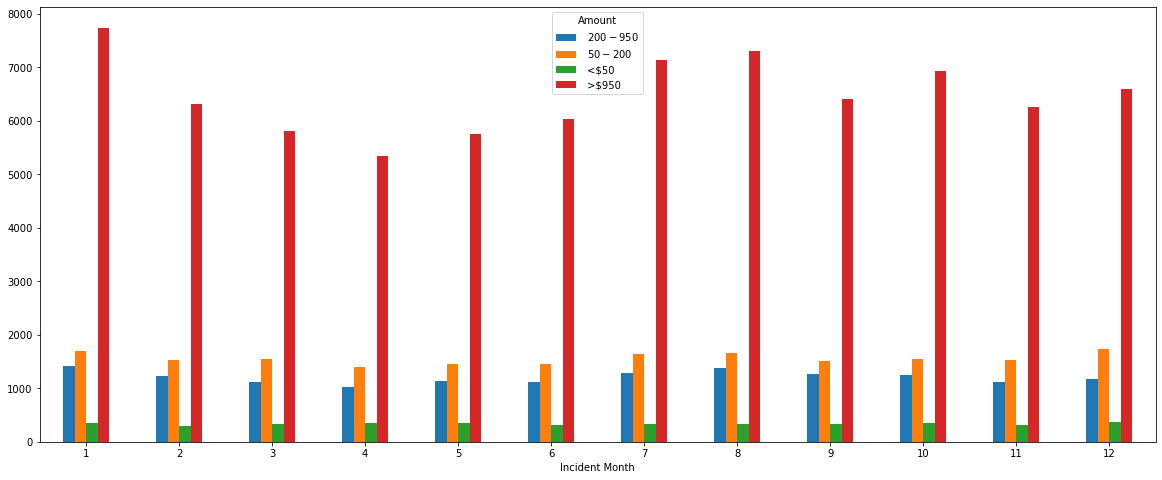
Hình 2.3 Số ca trộm cắp theo từng giờ và từng ngày trong tuần từ 2018 đến 2020

### Những thuộc tính ảnh hưởng đến biến mục tiêu.

Trong đồ án này, nhóm chúng em sẽ xây dựng mô hình dự đoán mức độ thiệt hại từ các vụ trộm cắp tại San Francisco (biến mục tiêu là 'Amount’ như đã đề cập ở trên)

Trước tiên, bỏ các cột:

* Incident Category: chọn sample của Incident Category mặc định là ‘Larceny theft’.
* Row ID, Incident ID, Incident Number, Incident Category: các feature này sẽ không được đưa vào mô hình vì là mã định danh, không cần thiết để trong phát triển mô hình.
* Report Datetime, Report Type Code, Report Type Description, Filed Online, Incident Description, Resolution: vô nghĩa khi sử dụng các biến này để dự đoán khi vụ án chưa xảy ra.



Hình 2.4 Grouped bar chart giữa hai biến ‘Incident Month’ và ‘Amount’

Từ biểu đồ thấy được tỉ lệ các vụ trộm cướp của bốn mức độ thiệt hại là tương đối giống nhau giữa các tháng, có thể nói thuộc tính ‘Incident Month’ không có mối tương quan với biến ‘Amount’ - biến đầu ra của mô hình. Tương tự với các biến ‘Incident Day of Week’, ‘Incident Year’.

Kết hợp sử dụng hàm SelectKBest trong module sklearn.feature\_selection để tính giá trị Chi2 score. Từ đó chọn được các biến có ảnh hưởng đến mô hình là ‘time\_hour’, ‘Incident Subcategory’, ‘CNN’, ‘Police District’, ‘Supervisor District’, ‘Incident Datetime’, ‘Neighborhoods’.

## Phát triển mô hình

### Tiền xử lý dữ liệu

Thực hiện encoding các biến vừa tìm được, chuẩn hóa lại bằng hàm StandardScaler, chia thành hai tập train và test với tỉ lệ là 9:1.

### Huấn luyện mô hình

Xây dựng ba mô hình phân loại: Logistic Regression, RandomForest, KNN trên tập train với bộ tham số mô hình là bộ tham số tối ưu nhất được tìm ra từ các tham số cho trước khi sử dụng hàm GridSearchCV.

Bảng 2.4 Bộ tham số tương ứng của mỗi mô hình

|  |  |
| --- | --- |
| **Mô hình** | **Bộ tham số** |
| Logistic Regression | C: 100, penalty: none, solver: sag |
| RandomForest | n\_neighbors: 25, weights: uniform |
| KNN | bootstrap: True, max\_depth: 10, max\_features: log2, n\_estimators: 13 |

### Đánh giá mô hình

Đánh giá mô hình dựa trên thang đo: Accuracy, F1-score trên tập test

Bảng 2.5 Accuracy và F1-score trên tập test của mỗi mô hình

|  |  |  |
| --- | --- | --- |
| **Mô hình** | **Accuracy** | **F1-score** |
| Logistic Regression | 0.67 | 0.54 |
| RandomForest | 0.70 | 0.64 |
| KNN | 0.66 | 0.56 |

Mô hình có khả năng phân lớp tốt nhất trong cả ba mô hình là RandomForest với Accuracy score và F1- score lần lượt là 0.70 và 0.64

# KẾT LUẬN

Trong đồ án này, nhóm chúng em đã phân tích về bộ dữ liệu **The San Francisco Police Department’s (SFPD) Incident Report Dataset,** đồng thời xây dựng được một model dự đoán mức độ thiệt hại của nạn nhân khi vướng vào một vụ trộm cướp tại San Francisco. Thông qua các bước như tiền xử lý dữ liệu, phân tích thăm dò, chọn ra các biến quan trọng để xây dựng mô hình, điều chỉnh tham số và đánh giá mô hình, nhóm chúng em có được ba mô hình khá tương đồng nhau và đều cho ra một mô hình phân lớp có khả năng dự đoán tương đối. Trong đó, mô hình RandomForest cho kết quả khả quan nhất. Tuy nhiên, độ chính xác của mô hình vẫn còn có thể cải thiện bằng cách sử dụng các mô hình khác như CART hay sử dụng các kỹ thuật oversampling như SMOTE.

**TÀI LIỆU THAM KHẢO**

1. Nitish Yadav, Ashish Kumar, …, City Crime Mapping using Machines Learning Techniques, 2020.
2. scikit-learn.org Link: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectKBest.html (06/12/2021).

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Bùi Hồng Thiên | Tiền xử lý dữ liệu, Huấn luyện mô hình, Tổng hợp file code, Làm demo |
| 2 | Phạm Thảo Nhi | Phân tích thăm dò, Huấn luyện mô hình, Tìm tham số tối ưu, Soạn slide |
| 3 |  |  |

**PHỤ LỤC HÌNH ẢNH**

Hình 1: Top 10 subcategory của các vụ trộm cắp

A picture containing graphical user interface

Description automatically generated