UNIVERSIDAD
DE LA FRONTERA

www.ufro.cl

Hito 3: USCrimesDataset

Diego Garrido - Ricardo Millanao - Nicolas Pereira

Noviembre - 2022

Ingeniería de Datos

Introducción

El conjunto de datos…

Contiene información acerca de delitos que ocurrieron en ciertas ciudades de Estados Unidos. Estos datos entregan indican cuándo ocurrieron, qué delito se cometió, una descripción general al respecto y el número de víctimas.

Lugar

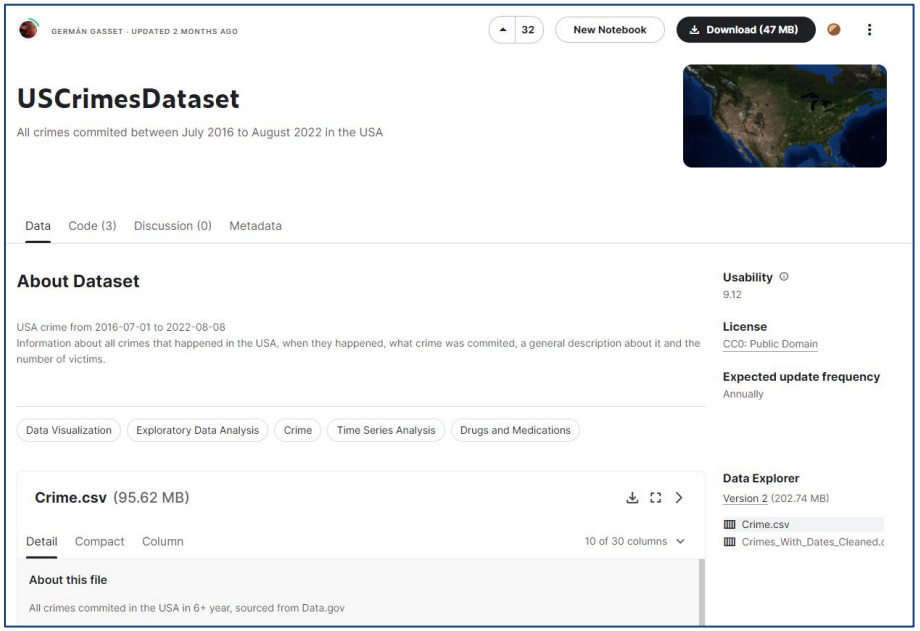
Estos registros se realizan sobre los 54 ciudades de Estados Unidos

Periodo

Los registros comprenden desde julio de 2016 hasta agosto de 2022

Motivación

Nuestra motivación es predecir delitos en base a una dimensión espacial y temporal donde eventualmente evaluar si se puede conseguir la predicción con el dataset USCrimesDataset.



Recurso
https://www.kaggle.com/datasets/jgiigii/uscrimesdataset

Dataset

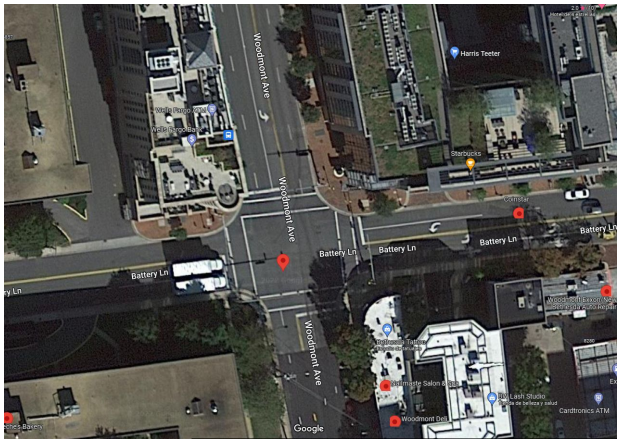
X	Incident ID	Offence Code	CR Number	Dispatch Date / Time	NIBRS Code	Victims	Crime Name1	Crime Name2	Crime Name3	Police District Name	Block Address
1	201181293	3522	180015424	03/30/2018 01:00:55 AM	35A	1	Crime Against Society	Drug/Narcotic Violations	DRUGS - OPIUM OR DERIVATIVE - POSSESS	BETHESDA	8300 BLK WOODMONT AVE

City	State	Zip Code	Agency	Place	Sector	Beat	PRA	Address Number	Street Prefix	Street Name	Street Suffix
BETHESDA	MD	20814.0	MCPD	Street - In vehicle	E	2,00E+02	54	8300.0		WOODMONT	

Street Type	Start_Date_Time	End_Date_Time	Latitude	Longitude	Police District Number	Location	Year	Month	Year-Month	Day	Committed_At_Morning
AVE	03/30/2018 01:01:00 AM		38.992.692.631	-77.097.062.905	2D	(38.9927, -77.0971)	2018	3	2018-03	30	FALSE

Recreación de la fila X=1

En la tarde del día 30 de marzo de 2018 a las 1:00:55 AM, en la ciudad de Bethesda del estado MD, se notifica un delito un delito hacia lo sociedad, infracción por la posesión de Opio en la calle con vehículo en la calle Woodmont



Fuente: Google Maps

Limpieza de datos

El dataset preliminarmente posee **306.094 filas** en conjunto de **36 columnas**.
Hemos elegido las variables en base a la predicción de delitos, y también variables que aportan datos que nos ayuden a entender los delitos. Por conceptos de este documentos la hemos agrupado en tres categorías:

- Tiempo: Entregan datos de tipo Tiempo
- Ubicación: Datos espaciales
- Delito: Datos del delito

Hemos omitido y eliminado variables que poseían datos nulos y vacíos. Luego de eso poseemos **229.289 filas** y **22 columnas** en el dataset

Columnas Eliminadas

Incident ID	Offence Code
Agency	CR Number
Street Suffix	Block Address
Address Number	Street Prefix
NIBRS Code	Location
Police District Name	Start_Date_Time
State	Zip Code
End_Date_Time	PRA
Year-Month	Police District Number

Tratamiento de datos - Ciudades

SILVER SPRING	88084
GAITHERSBURG	36592
ROCKVILLE	35966
GERMANTOWN	25320
BETHESDA	17992
MONTGOMERY VILLAGE	7825
TAKOMA PARK	5881
CHEVY CHASE	5396
POTOMAC	5372
DERWOOD	4632
KENSINGTON	4106
OLNEY	3902
BURTONSVILLE	3134
CLARKSBURG	2725
DAMASCUS	2111
BOYDS	1675
BROOKEVILLE	779
POOLESVILLE	736
SANDY SPRING	343
ASHTON	300
DICKERSON	253
CABIN JOHN	197
SPENCERVILLE	113
GLEN ECHO	74
BRINKLOW	61
MOUNT AIRY	58
BARNESVILLE	43
BEALLSVILLE	39
LAUREL	36
MT AIRY	19
GARRETT PARK	16
BELTSVILLE	12
HIGHLAND	9
WASHINGTON GROVE	8
WHEATON	6
WASHINGTON	6
WOODBINE	6
LANHAM	5
BETHEDA	2
ADELPHI	2
ROCKVILLLE	2
GAIHTERSBURG	1
BOWIE	1
CLAEKSBURG	1
FREDERICK	1
FRIENDSHIP HEIGHTS	1
SILVERS SPRING	1
GAITHERSBUG	1
NORTH BETHESDA	1
ROCKVIILE	1
GREENBELT	1
HAGERSTOWN	1
HYATTSVILLE	1
NORTH POTOMAC	1
0	1

Tenemos en el dataset registro de delitos de **54** ciudades, en donde:

- **16** ciudades registran más de 1000 víctimas
- **7** ciudades entre 100 a 1000 víctimas.
- **9** ciudades entre 10 y 100 víctimas.
- **23** ciudades con menos de 10 víctimas.

Por lo tanto se decide utilizar las ciudades con más de 300 víctimas por crímenes.

Tratamiento de datos - Fechas y Clima

Variables de Fechas

En base a la variable "*Dispatch Date / Time*" se Obtienen las siguientes variables:

Year	Integer	Año cuando se realizó el delito
Month	Integer	Mes cuando se realizó el delito
Day	Integer	Dia cuando se realizó el delito
Committed_At_Morning	Boolean	Índica si el delito se cometió de mañana o tarde
date_complete	String	Momento en el que se enviaron a los agentes policiales.
Hour	Integer	Hora cuando se realizó el delito
DayOfYear	Integer	Dia del año cuando se comete el delito
Week	Integer	Número de la semana cuando se comete el delito
DayOfWeek	Integer	Dia de la semana cuando se comete el delito
Quarter	Integer	Cuarto del día cuando se comete el delito
month_name	String	Nombre del mes cuando se comete el delito
day_name	String	Nombre del día cuando se comete el delito
Is_Commemoration_Day	Boolean	Índica si el delito se cometió en un día festivo

Variables de Clima

Creemos que las variables meteorológicas pueden explicar el comportamiento de ciertos crímenes

En base a la variable "*City*" se obtienen una serie de variables meteorológicas mediante un script de python utilizando la librería y api de world weather online. Estas variables están asociadas a cada crimen por la ciudad y hora de crimen.

tempC	Intger	Temperatura en grados celsius a la hora cuando se comete el delito
windspeedKmph	Integer	Velocidad del viento en Kilómetros por hora a la hora cuando se comete el delito
humidity	Integer	Humedad del ambiente a la hora cuando se comete el delito
precipMM	Float	Milímetros de precipitación a la hora cuando se comete el delito
cloudcover	Integer	Bloqueo de nubes a la hora cuando se comete el delito

Fuente de datos climaticos: https://www.worldweatheronline.com/

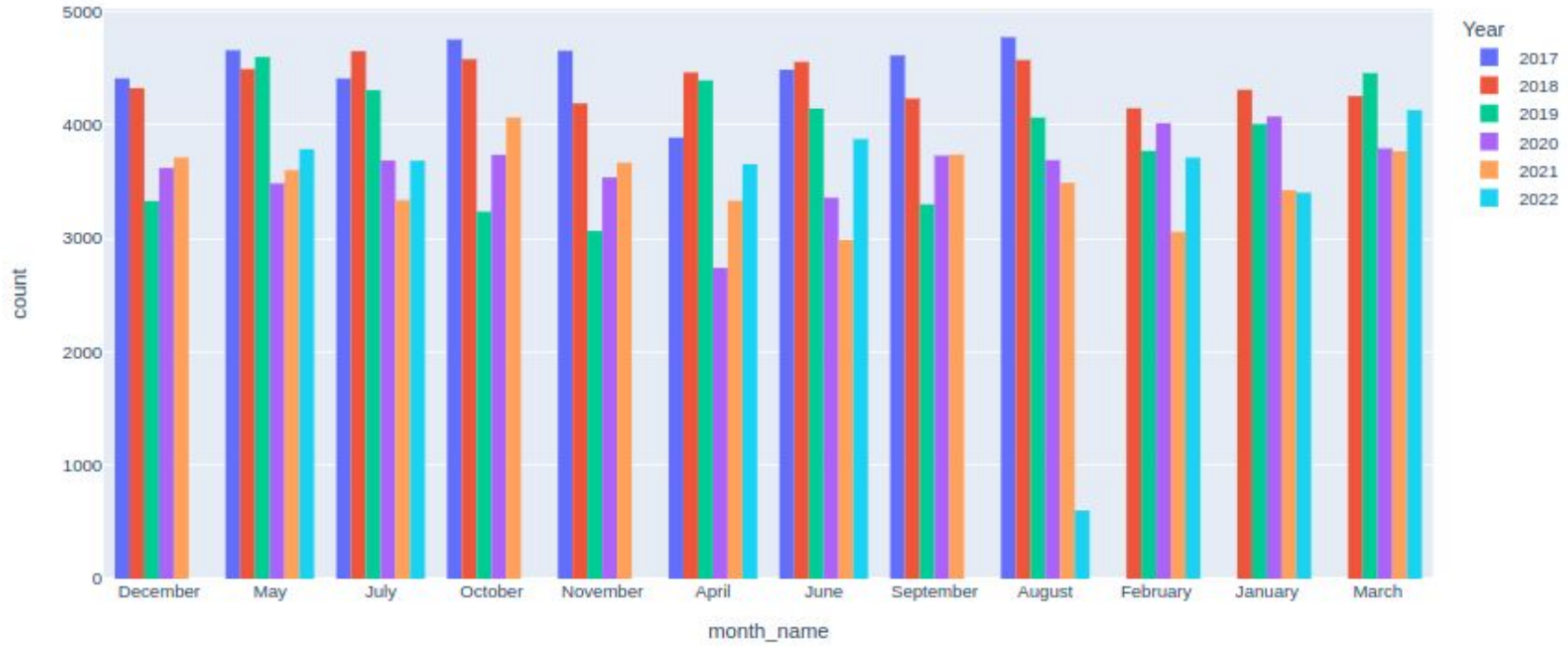
Exploración - Dataset

Delito		
Crime Name1	String	Tipo De crimen cometido.
Crime Name2	String	Categoría del crimen cometido.
Crime Name3	String	Información del crimen cometido.
Victims	Integer	Victimas afectadas
Place	String	Lugar donde se llevó a cabo el delito
Beat	String	Policías designados al sector anterior.
Ubicación		
City	String	Ciudad donde ocurrió el crimen.
Street Name	String	Nombre de la calle donde se cometió el delito
Street Type	String	Tipo de calle donde se cometió el delito
Sector	String	Lugar designado para ciertas patrullas
Latitude	float	Coordenada de latitud de la ubicación del delito
Longitude	float	Coordenada de longitud de la ubicación del delito

Tiempo		
Year	Integer	Año cuando se realizó el delito
Month	Integer	Mes cuando se realizó el delito
Day	Integer	Dia cuando se realizó el delito
Committed_At_Morning	Boolean	Índica si el delito se cometió de mañana o tarde
date_complete	String	Momento en el que se enviaron a los agentes policiales.
Hour	Integer	Hora cuando se realizó el delito
DayOfYear	Integer	Dia del año cuando se comete el delito
Week	Integer	Número de la semana cuando se comete el delito
DayOfWeek	Integer	Dia de la semana cuando se comete el delito
Quarter	Integer	Cuarto del día cuando se comete el delito
month_name	String	Nombre del mes cuando se comete el delito
day_name	String	Nombre del día cuando se comete el delito
Is_Commemoration_Day	Boolean	Índica si el delito se cometió en un día festivo
Clima		
tempC	Intger	Temperatura en grados celsius a la hora cuando se comete el delito
windspeedKmph	Integer	Velocidad del viento en Kilómetros por hora a la hora cuando se comete el delito
humidity	Integer	Humedad del ambiente a la hora cuando se comete el delito
precipMM	Float	Milímetros de precipitación a la hora cuando se comete el delito
cloudcover	Integer	Bloqueo de nubes a la hora cuando se comete el delito

Exploración

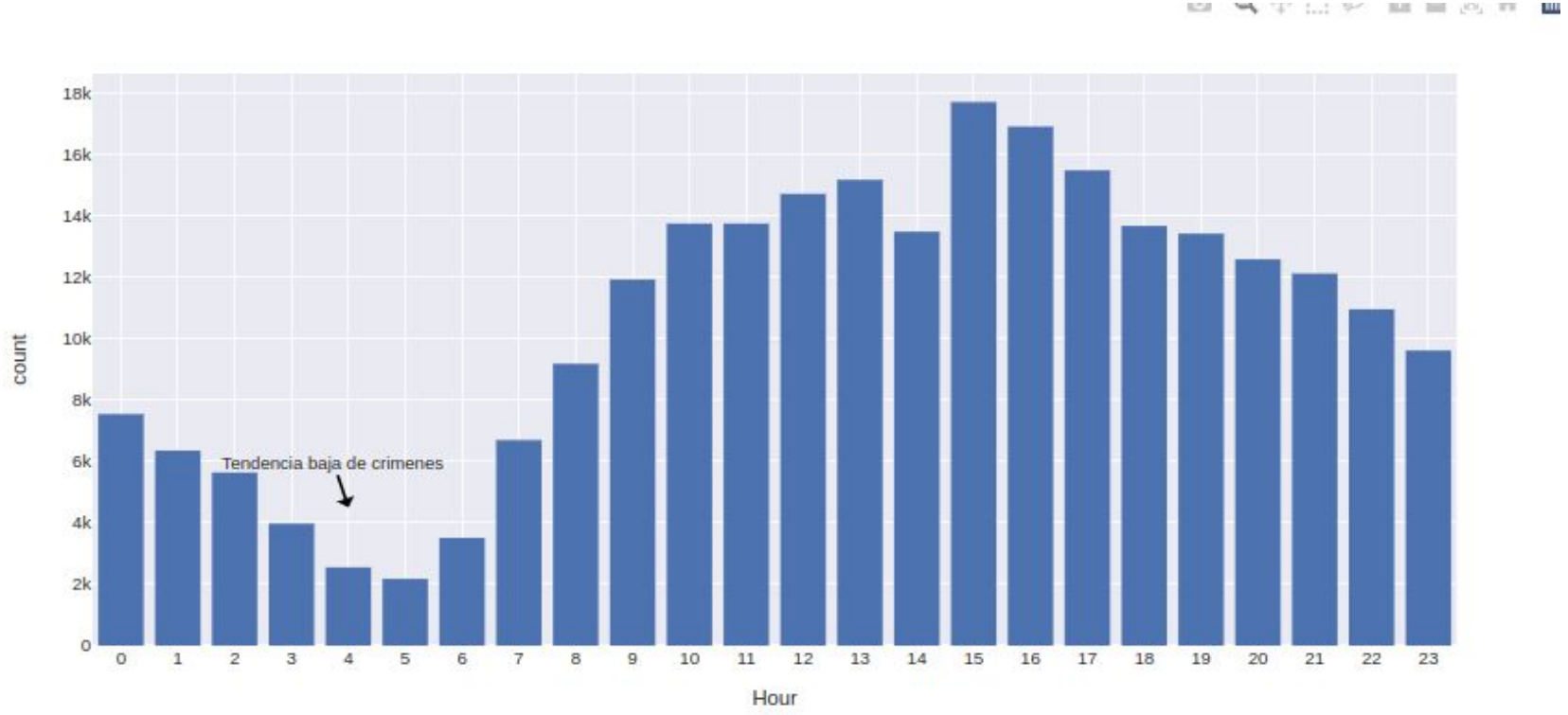
1. Delitos cometidos por cada mes y año



A partir del año 2017 existe una disminución de los delitos cometidos, sin embargo, el año 2017 y 2022 poseen una diferencia considerable con los demás años. Los delitos cometidos durante el año 2016 son registrados desde el mes de julio hasta el mes de diciembre, mientras que el año 2022 los delitos son registrados desde el mes de enero hasta una parte del mes de agosto. Debido a lo anterior estos años poseen una diferencia en sus registros respecto a los otros años.

Exploración

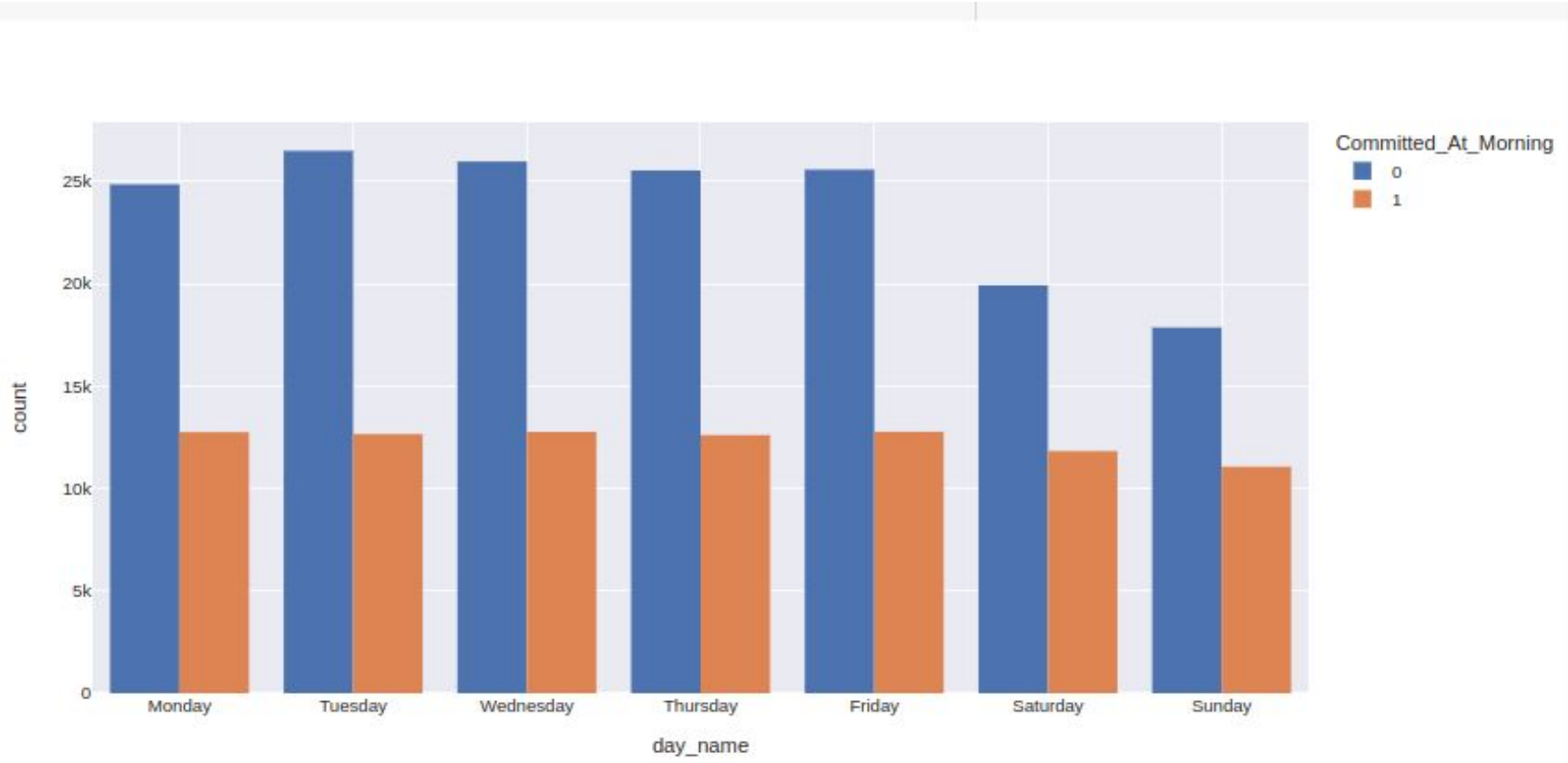
2. Delitos cometidos por hora



Existe un horario punta de las alertas entre las 15 pm a 16 pm y una baja en las alertas correspondientes al lapso de 4 am a 5 am

Exploración

3. Delitos cometidos mañana - tarde por día

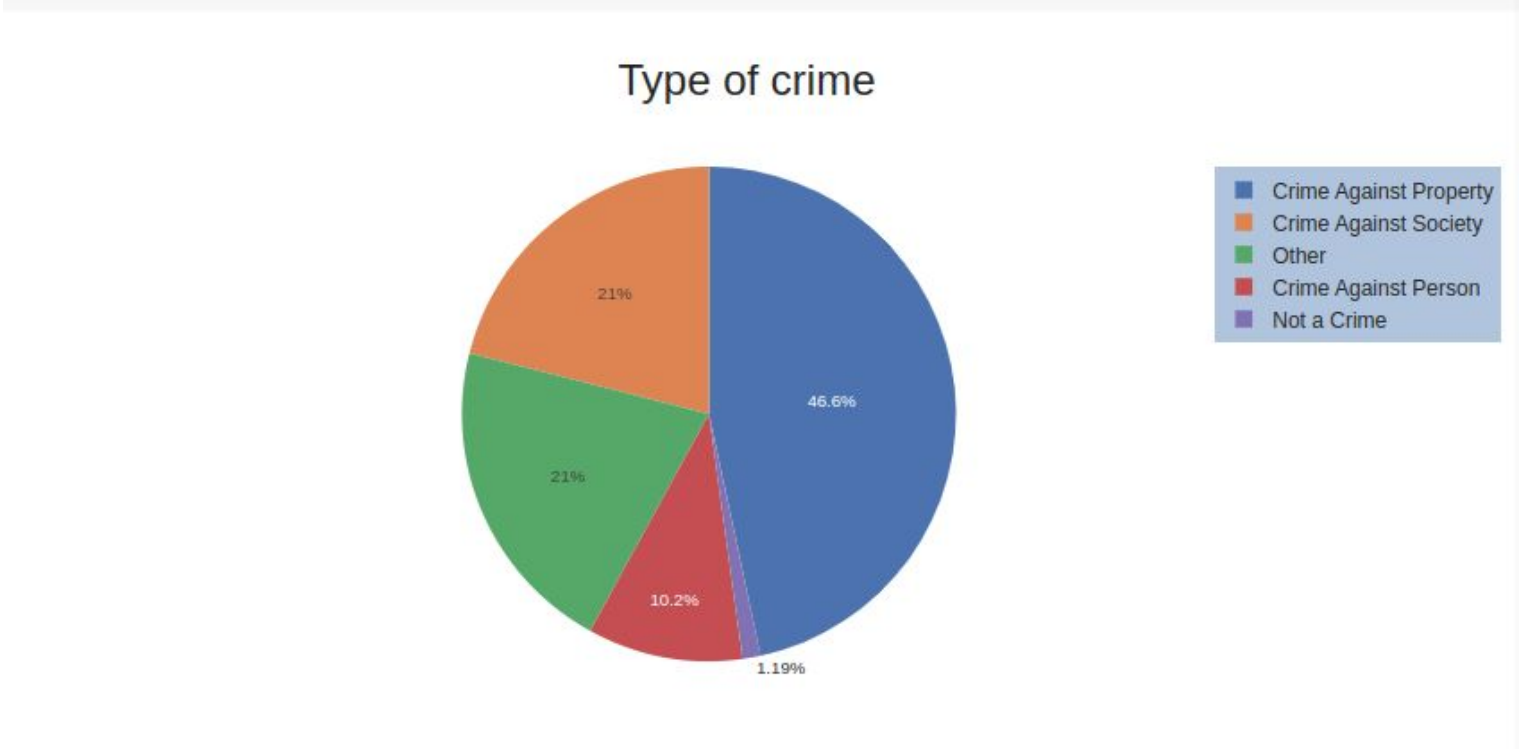


3. Desde el día lunes a viernes se puede visualizar una ocurrencia homóloga de los delitos, con una tendencia los días martes. Por otro lado la ocurrencia de delitos disminuye significativamente para el fin de semana, ósea los días Sábado y Domingo.

Exploración

4. Tipo de crímenes cometidos por año

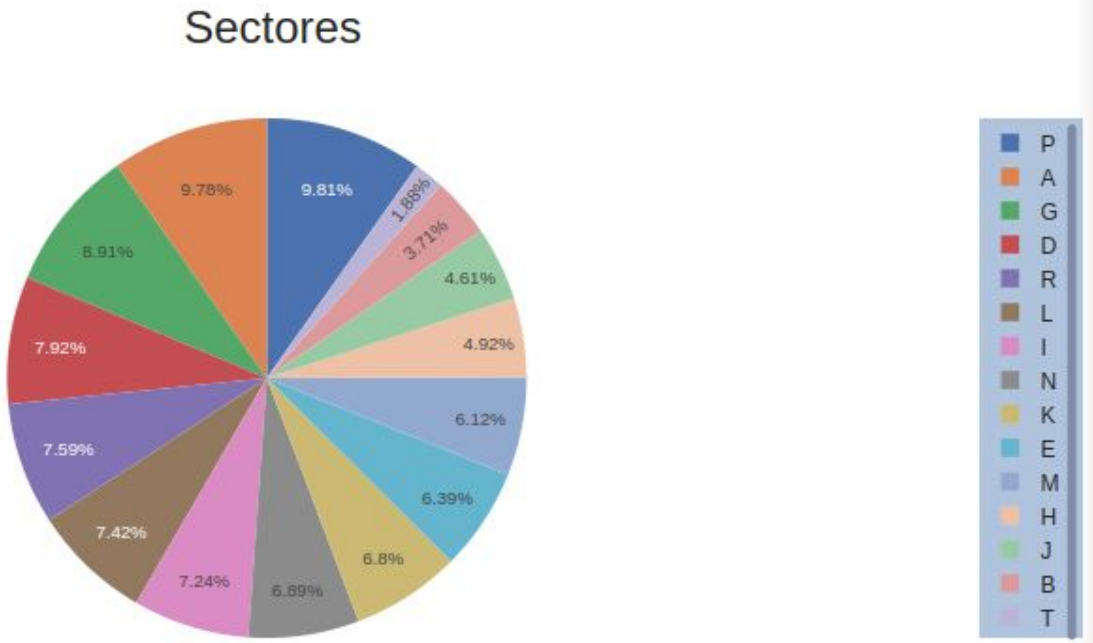
La variable **Crime Name 1,** de tipo Categórica, es candidata a ser el target de nuestro modelo, debido a no posee categoría extensas como **Crime Name 2** y **3**, ahora veremos su comportamiento:



4. Durante 2016 hubo un crecimiento lineal hasta el año 2017, posterior a este año se puede ver una variación de los delitos a excepción de los crímenes cometidos hacía la propiedad que se mantuvo en alza hasta el año 2021. Los tipos de crímenes representados en el gráfico de torta, se observa que el mayor porcentaje de crímenes es contra la propiedad.

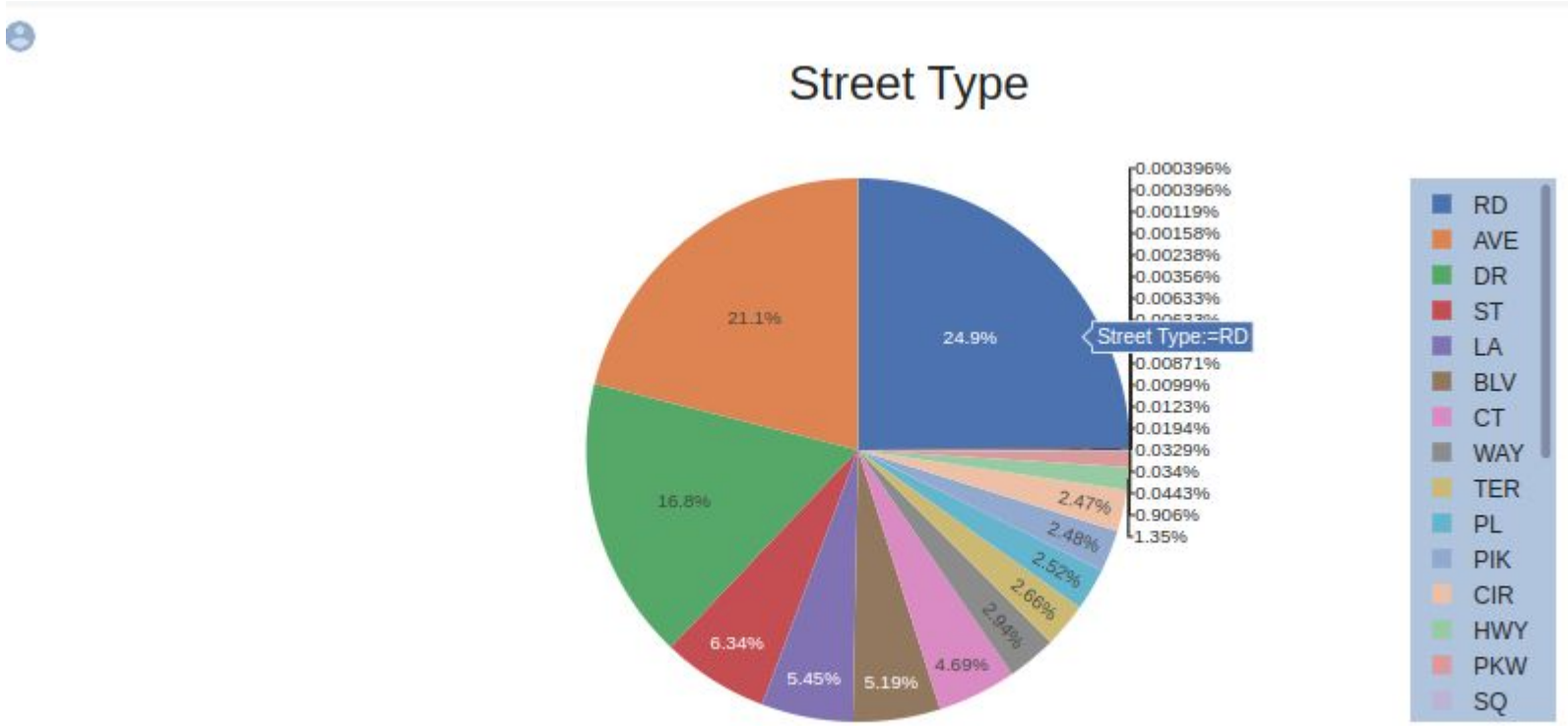
4.- Delictualidad en sectores

```
.show()
```



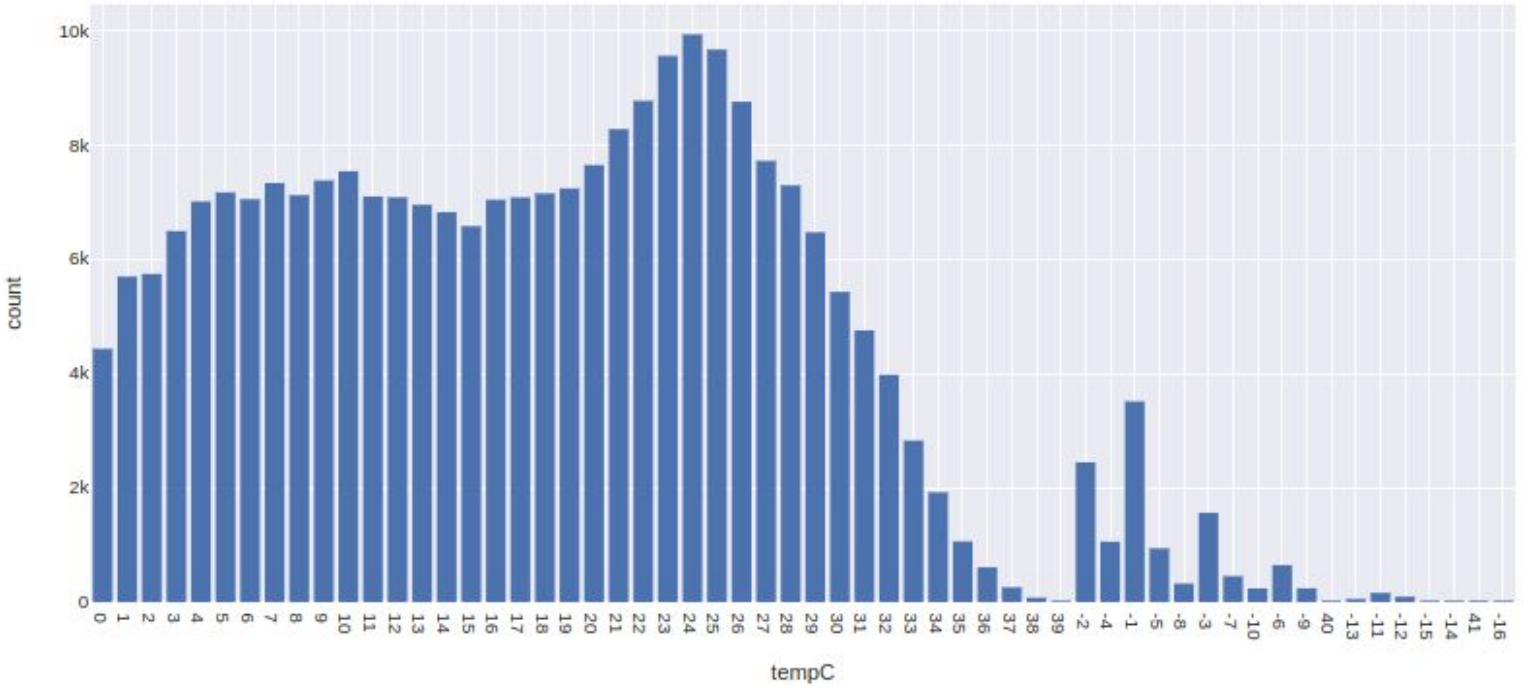
En el gráfico se aprecia los sectores con mayor tasa de delictualidad siendo relevantes los sectores P, A y G

5.- Delictualidad en tipos de calle



En el gráfico se puede apreciar que la tasa de delictualidad sube si se trata de carreteras (roads) y avenidas (Avenue)

6.- Delictualidad en relación a la temperatura ambiental



En el gráfico se muestra que en temperaturas extremas como 39,40 y 41 grados celsius, en caso de temperaturas altas y -16, -15 y -14 en las temperaturas bajas, se producen menos delitos que en las temperaturas más templadas como 24 y 25 grados

Preguntas y Propuesta

Dada la información obtenida mediante los análisis previos hemos encontrado la siguiente problemática, hemos visto que existen diferencias de la cantidad de delitos cometidos entre día y noche, entre los meses e incluso entre los años, pero nos gustaría saber cómo podemos predecir estos crímenes. Para esto debemos tener en cuenta los siguientes pasos

1. Debido a que el delito de robo es el que se produce con una mayor frecuencia, definimos el siguiente problema: ¿Qué variables son las que tienen mayor incidencia en clasificar Crime Name 2 como un delito de robo?. ¿Es posible predecir los delitos de robo en base a esto?
2. Clasificar los delitos de acuerdo a "Crime Name 1"
3. Evaluaremos si existen grupos con características similares en los delitos para de esta forma agruparlos en una nueva categoría (leve-medio-grave)

Con esto buscamos detectar patrones dentro los delitos y así, de esta forma, tomar medidas con respecto a estos patrones.

Experimento I - Variables

Variables predictors a utilizar

- Year
- Month
- Day
- Hour
- DayOfYear
- Week
- DayOfWeek
- Quarter
- Sector
- Place
- Beat
- Street Name
- Street Type
- Is_Commemoration_Day
- humidity
- precipMM
- cloudcover
- tempC
- windspeedKmph
- Latitude
- Longitude

Variable Target a utilizar

- Crime Name2
  a. Crime Name3

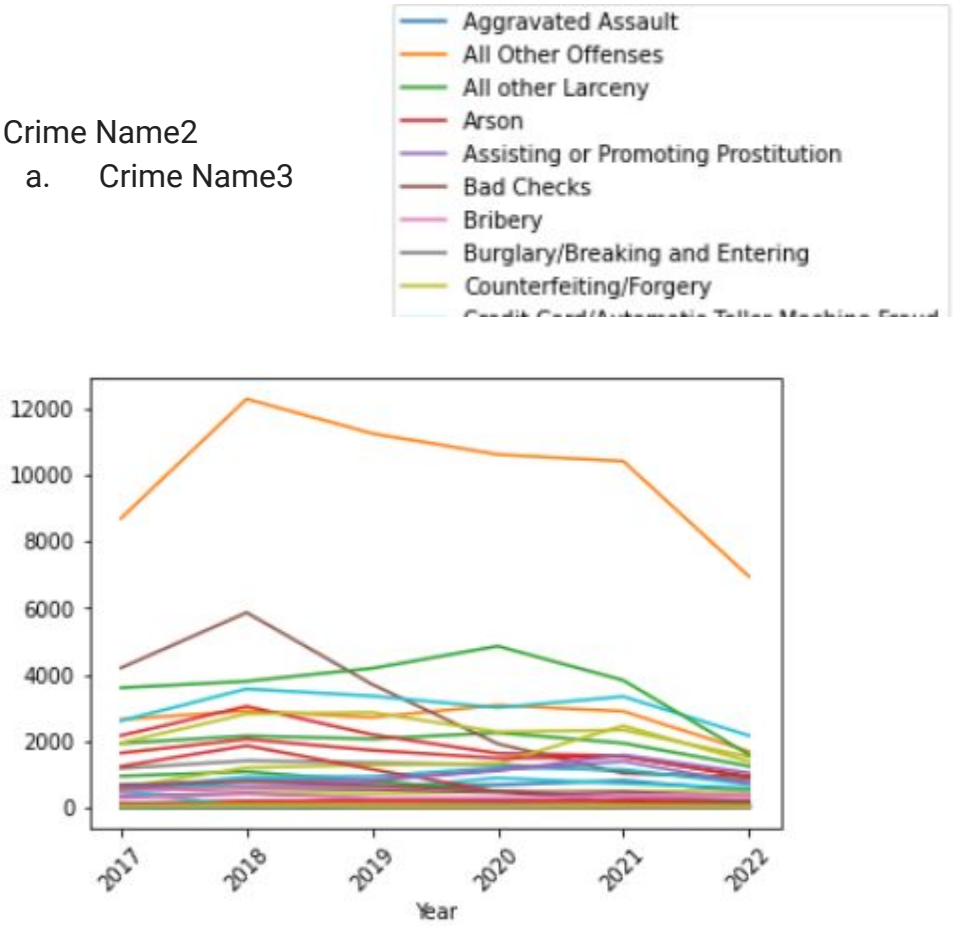


Gráfico: Categoría de Crímenes por años

Experimento I - Arbol de desición

Accuracy score (max_depth:8)

	precision	recall	f1-score	support
Aggravated Assault	0.12	0.00	0.00	1197
All Other Offenses	0.00	0.00	0.00	53
All other Larceny	0.56	0.33	0.42	3840
Arson	0.00	0.00	0.00	40
Burglary/Breaking and Entering	0.54	0.10	0.17	1702
From Coin/Operated Machine or Device	0.00	0.00	0.00	14
Intimidation	0.00	0.00	0.00	103
Motor Vehicle Theft	0.00	0.00	0.00	82
Pocket/picking	0.25	0.01	0.01	180
Purse-snatching	0.00	0.00	0.00	117
Robbery	0.30	0.05	0.09	926
Simple Assault	0.40	0.63	0.49	5956
Stolen Property Offenses	0.00	0.00	0.00	27
Theft From Motor Vehicle	0.61	0.89	0.73	7190
Theft from Building	0.39	0.50	0.44	3078
Theft of Motor Vehicle Parts or Accessories	0.29	0.04	0.07	1948
Weapon Law Violations	0.32	0.09	0.14	339
accuracy			0.50	26792
macro avg	0.22	0.16	0.15	26792
weighted avg	0.46	0.50	0.44	26792

```
Accuracy en test set: 0.49630486712451477
```

Evaluación del modelo: Cross Validation (40% - 55%)

```
Cross Validation Scores are [0.40532052 0.45760207 0.47995566 0.48143359 0.53500831 0.53002032
 0.58451875 0.48743533 0.49205469 0.5218034  0.49094605 0.50166297
 0.48484848 0.50831486 0.52697709]
Average Cross Validation score :0.4991934730324245
```

Experimento I - Resultados

Place es una variable que nos indica la forma del lugar donde se cometió el delito. Siendo el delito de tipo robo y siendo las variables `Place`, `Hour`, `Longitude` y `Latitude` con mayor incidencia, logramos establecer que el delito de este tipo depende de la hora y el lugar para que pueda ser cometido.

Mediante el clasificador de Árbol de decisiones fué posible obtener una precisión sobre el 50% en ciertos delitos como: Weapon Law Violations, Theft From Motor Vehicle, Burglary/Breaking and Entering y All other Larceny. Sin embargo, otros delitos de tipo robo poseen una precisión menor al 50%.

Estamos aplicando este modelo sin la utilización de técnicas que ayuden al desbalance, por tanto, en esta instancia con los resultados obtenidos creemos que no es posible predecir los delitos de tipo robo.

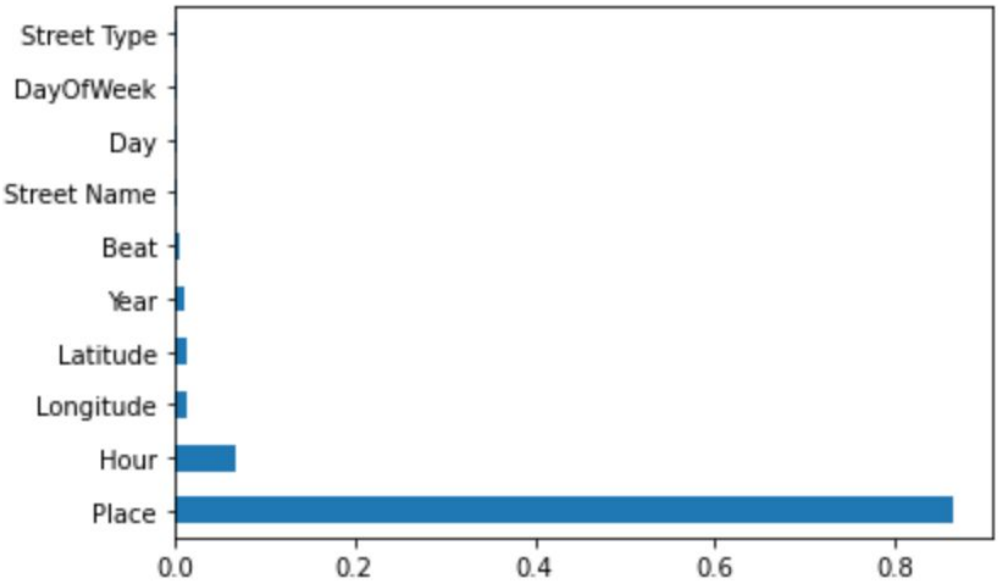


Gráfico: Variables con mayor incidencia en clasificar delitos de tipo robo

Desarrollo del modelo

Arbol de decisión

	precision	recall	f1-score	support
Crime Against Person	0.91	0.17	0.29	8019
Crime Against Property	0.53	0.98	0.68	37942
Crime Against Society	0.43	0.13	0.20	11939
Not a Crime	0.00	0.00	0.00	964
Other	0.00	0.00	0.00	16802

Cross Validation con 10 split.

[0.58145112 0.50732537 0.00300863 0.13443509 0.64234945 0.63096847

0.62983474 0.65477696 0.71416736 0.7722496]

Se observan los valores para el árbol de decisiones y su desempeño va entre el 0% aproximado y como máximo un 77.2%

Cross Validation con 15 split.

[0.6006279 0.53188567 0.50068677 0.00418602 0.00156976 0.20027471

0.6579894 0.64621623 0.62607103 0.6317614 0.61619465 0.67316371

0.68984237 0.76667975 0.77309001]

En cambio si se valida con 15 split se observan valores entre el 5% y como máximo un 77.3%

Desarrollo del modelo

KNN

	precision	recall	f1-score	support
Crime Against Person	0.29	0.10	0.15	6009
Crime Against Property	0.55	0.82	0.66	28758
Crime Against Society	0.38	0.28	0.32	9058
Not a Crime	0.00	0.00	0.00	744
Other	0.27	0.12	0.17	12765
accuracy			0.49	57334
macro avg	0.30	0.26	0.26	57334
weighted avg	0.43	0.49	0.43	57334

Cross Validation con 10 split.

[0.46441964, 0.45221069, 0.05132118, 0.17433393, 0.16862164, 0.18545328, 0.18017704, 0.16870885, 0.2268783, 0.30671957]

Se observan los valores para knn y su desempeño va entre el 5% y como máximo un 51%

Experimento II

El segundo consta de un problema de clasificación, bajo los algoritmos de **Árbol de decisión** y **KNN**, en el cual queremos predecir el tipo de crimen al que corresponde un llamado de emergencia, dando como variable *target* "Crime Name 1" y las variables o columnas predictoras a:

- Victims
- Year
- Month
- Hour
- Day
- Committed_At_Morning
- Is_Commemoration_Day
- humidity
- precipMM
- cloudcover
- Sector
- Street Type

Desarrollo del modelo

Arbol de decisión

Para el modelo de árbol de decisión se toma las variables X (features) y (corresponde a la clase) para generar el árbol de decisión.

Para ver qué tan bien fue el entrenamiento, podemos evaluar el clasificador ejecutándose sobre instancias y verificando que la clase sea la correcta.

El método train_test_split para generar cuatro listas, dos para entrenar el modelo y dos para testearlo. Nuevamente generamos el árbol de decisión a partir de los features de entrenamiento y finalmente con el método accuracy_score testeamos el algoritmo con los features de test.

Finalmente aplicando el método *accuracy_sore* el resultado sobre los features de y_test e y_pred es de 0.5224897719228785

Utilizando el método classification_report el cual nos permitió visualizar las variables precision, recall, f1-score y support, dando como resultado el siguiente gráfico:

	precision	recall	f1-score	support
Crime Against Person	0.91	0.17	0.29	8489
Crime Against Property	0.55	0.85	0.67	38690
Crime Against Society	0.43	0.52	0.47	17653
Not a Crime	0.00	0.00	0.00	997
Other	0.00	0.00	0.00	17520
accuracy			0.52	83349
macro avg	0.38	0.31	0.28	83349
weighted avg	0.44	0.52	0.44	83349

Desarrollo del modelo

KNN

- Creacion de features train_test_split
- Entrenamiento del dataset a partir del target y la data
- Creación del clasificador knn, a partir de los features de prueba
- Visualización de los resultado preliminares
- Evaluación de desempeño del modelo KNN

```
Accuracy of K-NN classifier on training set: 0.57
Accuracy of K-NN classifier on test set: 0.49
```

	precision	recall	f1-score	support
Crime Against Person	0.20	0.06	0.09	6438
Crime Against Property	0.54	0.79	0.65	29436
Crime Against Society	0.46	0.45	0.45	13265
Not a Crime	0.00	0.00	0.00	730
Other	0.29	0.12	0.17	13274
accuracy			0.49	63143
macro avg	0.30	0.28	0.27	63143
weighted avg	0.43	0.49	0.44	63143

```
scores = cross_val_score(knn, X, y, cv = k_folds)
print(scores)
```

```
[0.44433447 0.35954389 0.51233321 0.53007087 0.55647939 0.43093004
 0.42712911 0.47768935 0.43223661 0.47891674]
```

Experimento III

El tercero es un experimento realizado que consistió en la búsqueda de llamadas de emergencia que tuvieran características similares definiéndose como *leves, medios o graves,* mediante la construcción de un algoritmo no supervisado de **Clustering** bajo el método K-Means**.**

Para llevar a cabo este experimento, en primer lugar se necesito la selección de ciertas columnas de nuestro conjunto de datos, en este caso se eligieron las siguientes:

Victims
Year
Month
Hour
Day
Committed_At_Morning
Is_Commemoration_Day
humidity
precipMM
cloudcover

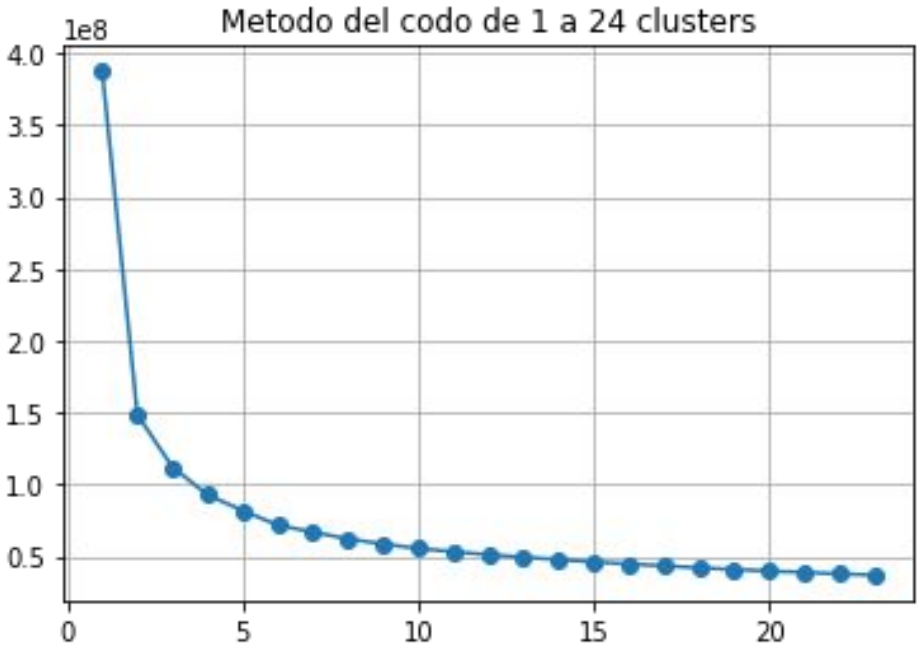


Grafico método del codo

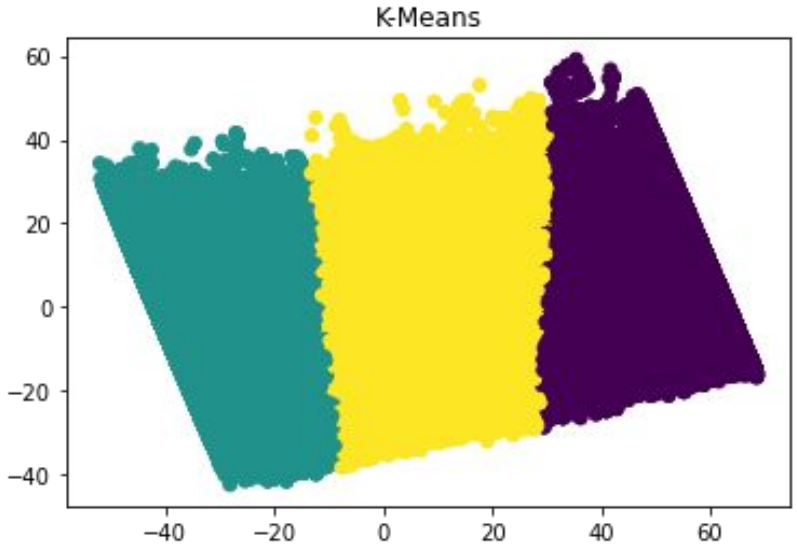
Desarrollo del modelo

K-Means

1. Cada cluster posee

   {1: 125011, 0: 61556, 2: 66004}

2.
3. Lo siguiente es visualizar los cluster obtenidos a partir de la métrica obtenida anteriormente:
4. En la imagen observamos que los 3 cluster obtenidos están distribuidos en 3 distintas partes del gráfico, lo que hace pensar si existen grupos de datos con características muy similares que permitirán categorizarlos en una nueva categoría.
5. Para el modelo de K-Means el resultado obtenido refleja que existe un score de 0.36638251589279375 esto se puede explicar a que el conjunto de datos no está bien separado, sería fácil identificar el número óptimo de grupos utilizando el método del codo.



CONCLUSIONES