UNIVERSIDAD
DE LA FRONTERA

www.ufro.cl

Hito 2: USCrimesDataset

Diego Garrido - Camilo Godoy - Ricardo Millanao - Nicolas Pereira

Octubre - 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

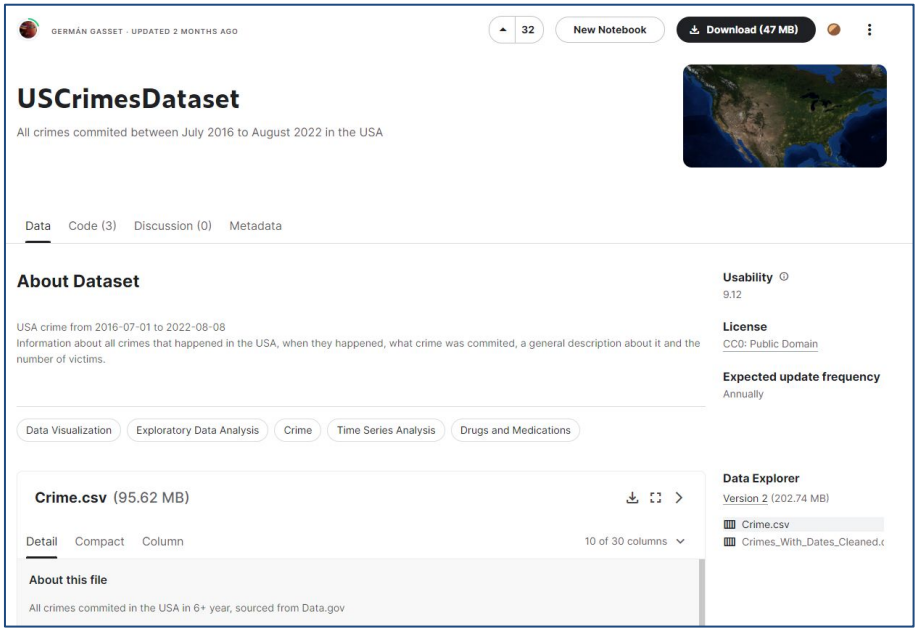
Estos registros se realizan sobre los 54 ciudades de Estados Unidos

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

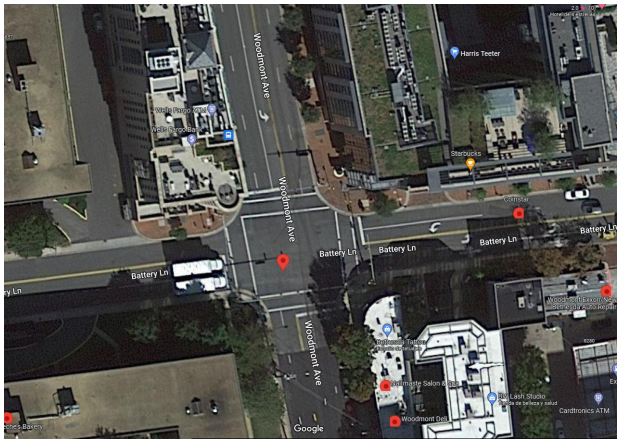
X	Incident ID	Offence Code	CR Number	Dispatch Date / Time	NIBRS Code	Victims	Crime Name1	Crime Name2	Crime Name3	Police District Name	Block Address
1	201181293	3522	180015424	03/30/2018 01:00:55 AM	35A	1	Crime Against Society	Drug/Narcotic Violations	DRUGS - OPIUM OR DERIVATIVE - POSSESS	BETHESDA	8300 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. Este delito fue delegado al distrito de policía 2D de esa ciudad



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

En la sección 2.1 se agregan 3 variables a partir de "**Dispatch Date / Time**":

- hour
- month
- day_name

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

Variable	Tipo	Descripción
Delito		
Crime Name1	Nóminal	Tipo de crimen cometido.
Crime Name2	Nóminal	Categoría del crimen cometido.
Crime Name3	Nóminal	Información del crimen cometido.
Police District Name	Nóminal	Nombre del distrito donde se cometió el delito
Offence Code	Nóminal	Código representativo del delito
Victims	Ordinal	Representa el número de personas afectadas en el delito.
Tiempo		
Year	Ordinal	Año cuando se realizó el delito
Month	Ordinal	Mes cuando se realizó el delito
Day	Ordinal	Dia cuando se realizó el delito
Committed_At_Morning	Ordinal	Índica si el delito se cometió de mañana o tarde
date	Intervalo	Momento en el que se enviaron a los agentes policiales.
hour	Intervalo	Hora cuando se realizó el delito
month	Intervalo	Mes cuando se realizó el delito
Ubicación		
Block Address	Nóminal	Dirección donde el crimen fue cometido
City	Nóminal	Ciudad donde ocurrió el crimen.
State	Nóminal	Estado donde se llevó a cabo el crimen.
Zip Code	Nóminal	Código de la ciudad donde se realizó el delito.
Sector	Nóminal	Lugar designado para ciertas patrullas
Address Number	Nóminal	Número de la calle donde se realizó el delito.
Location	Nóminal	Ubicación mediante coordenadas.

Exploración

City	Victims
SILVER SPRING	78228
ROCKVILLE	32927
GAITHERSBURG	32445
GERMANTOWN	22854
BETHESDA	16827
MONTGOMERY VILLAGE	7083
TAKOMA PARK	5437
POTOMAC	5085
CHEVY CHASE	4979
DERWOOD	4271
KENSINGTON	3692
OLNEY	3585
BURTONSVILLE	2806
CLARKSBURG	2604
DAMASCUS	1933
BOYDS	1607
BROOKEVILLE	730

City	Victims
POOLESVILLE	701
SANDY SPRING	327
ASHTON	278
DICKERSON	240
CABIN JOHN	176
SPENCERVILLE	111
GLEN ECHO	71
BRINKLOW	58
MOUNT AIRY	48
BEALLSVILLE	37
LAUREL	36
BARNESVILLE	30
MT AIRY	17
GARRETT PARK	15
BELTSVILLE	12
HIGHLAND	9
WASHINGTON GROVE	6
WASHINGTON	6
WOODBINE	6
LANHAM	5

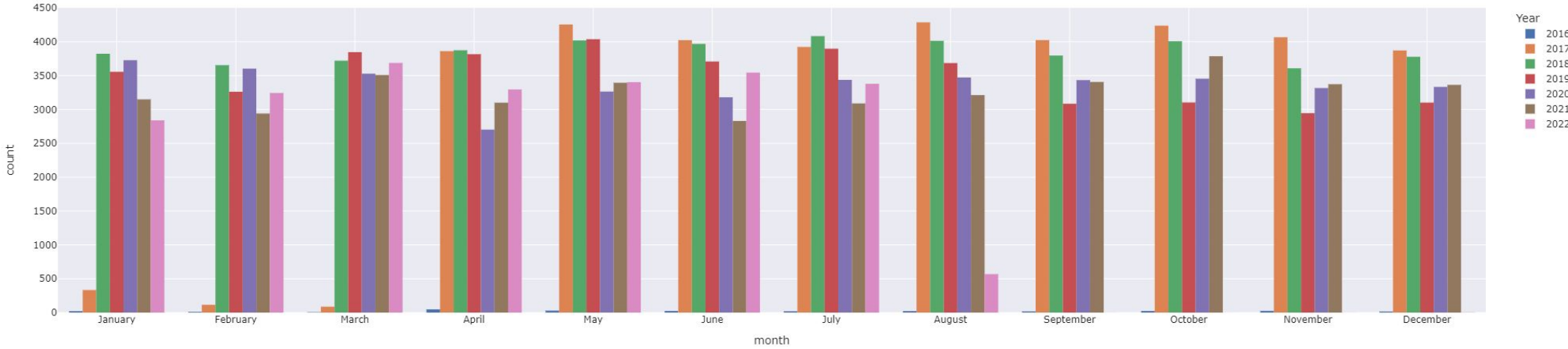
City	Victims
ADELPHI	2
NORTH POTOMAC	1
HAGERSTOWN	1
GREENBELT	1
FREDERICK	1
BOWIE	1

Tenemos en el dataset registro de delitos de **54** ciudades, de las cuales luego de la limpieza de datos se utilizan **43** ciudades, en donde:

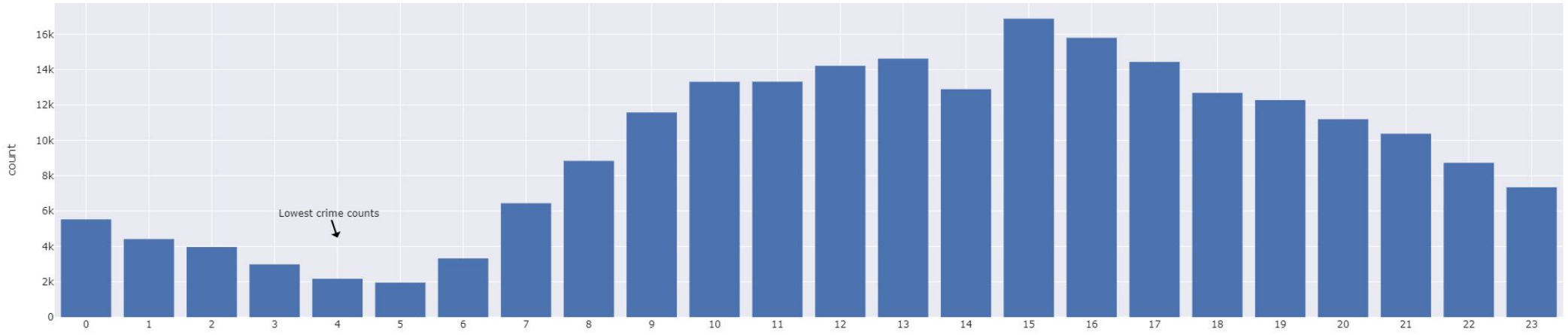
- **16** ciudades registran más de 1000 delitos
- **7** ciudades entre 100 a 1000 delitos.
- **9** ciudades entre 10 y 100 delitos.
- **11** ciudades con menos de 10 delitos.

Exploración

1. Delitos cometidos por cada mes y año



2. Delitos cometidos por hora

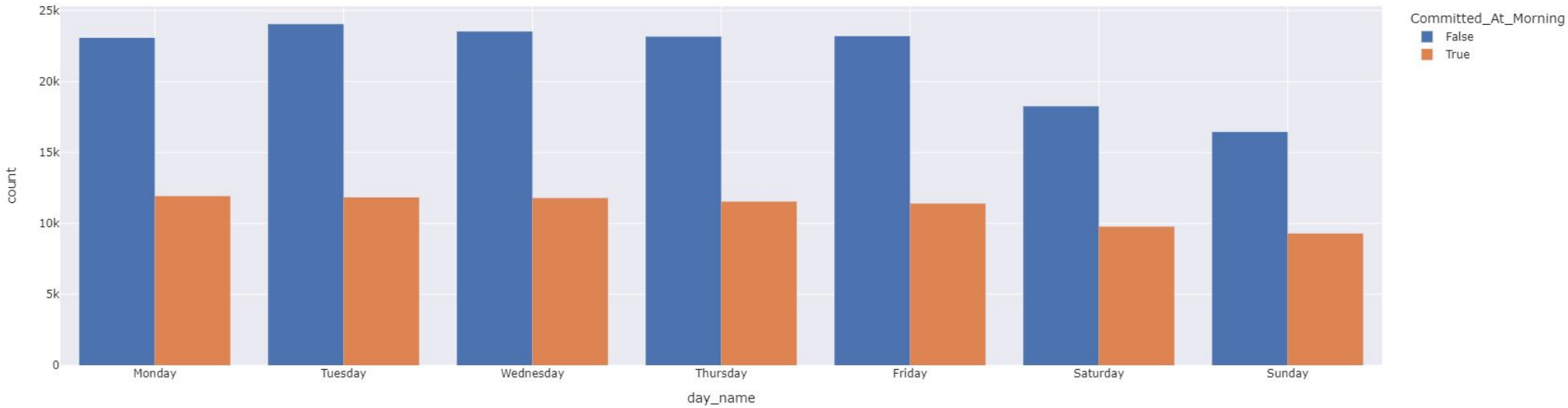


Exploración

1. 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.

2. 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

3. Delitos cometidos mañana - tarde por día

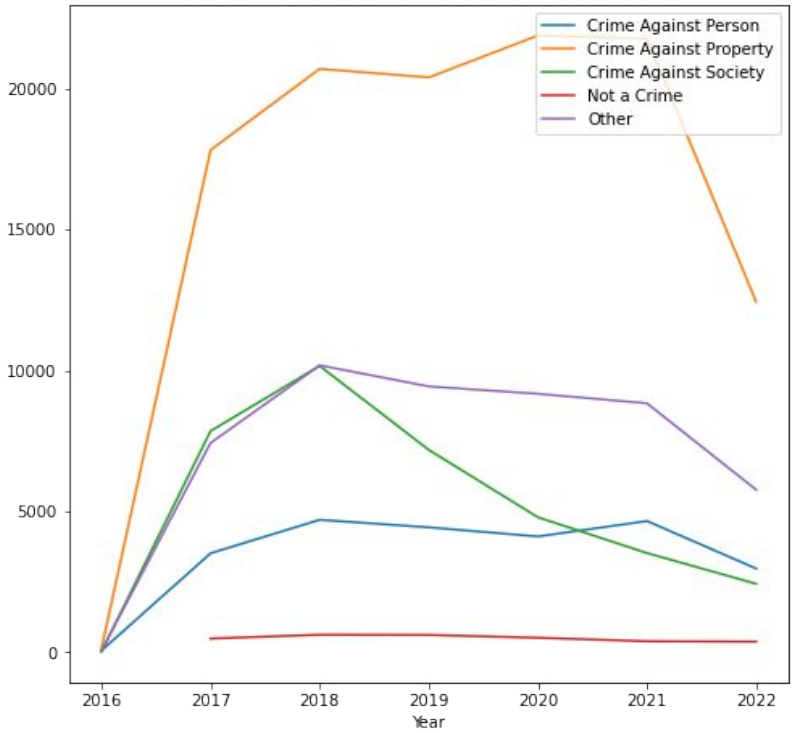


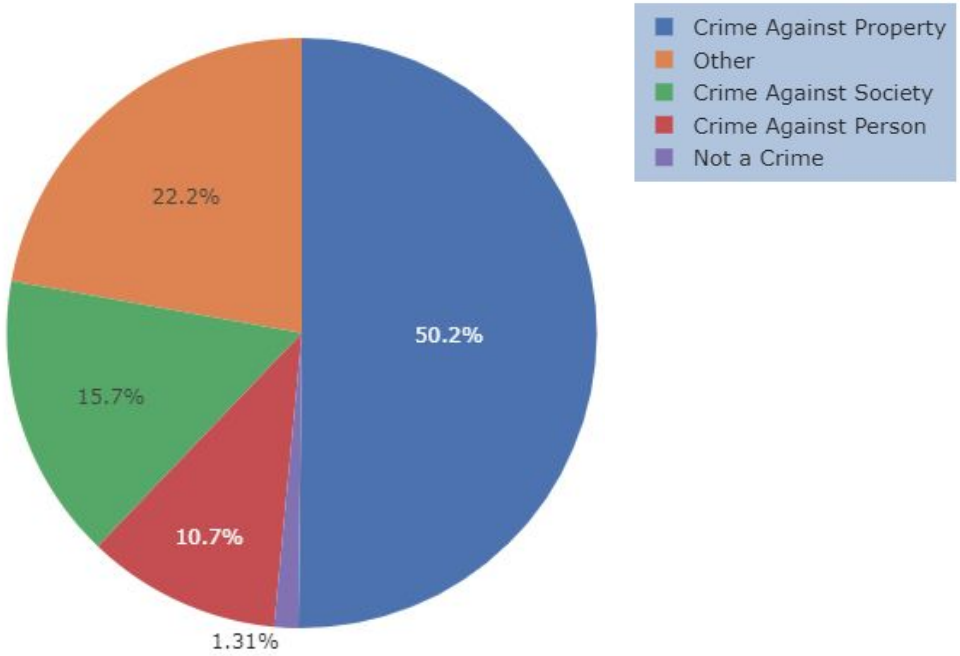
3. Desde el día lunes a viernes se puede visualizar una ocurrencia homóloga de los delitos, con una tendencia los días martes. Por otro lado la ocurrencia de delitos disminuye significativamente para el fin de semana, ósea los días Sábado y Domingo.

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.

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 2 preguntas:

1. ¿Qué delito deberíamos prestar atención en cierta fecha?
2. ¿Cuáles son los tipos de crímenes que debemos prestar atención durante cierto periodo de tiempo?
3. ¿Si tenemos una alerta en cierta hora y cierto día, que tipo de alerta de crimen será?

Luego del análisis efectuado podemos definir y plantear que nuestro experimento será propuesto como un problema de clasificación, tomando el atributo "Crime Name1" de nuestro dataset. Para esta propuesta se utilizaron los siguientes algoritmos;

1. Arbol de Decisión
2. KNN: Ya que al ser un modelo "supervisado" nos permitirá obtener la variable que deseamos predecir a partir de otras variables.

De igual forma las variables utilizadas para esta propuesta e implementadas en los modelos serán

1. Year
2. Month
3. Day
4. hour
5. Victims
6. Committed_At_Morning

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%