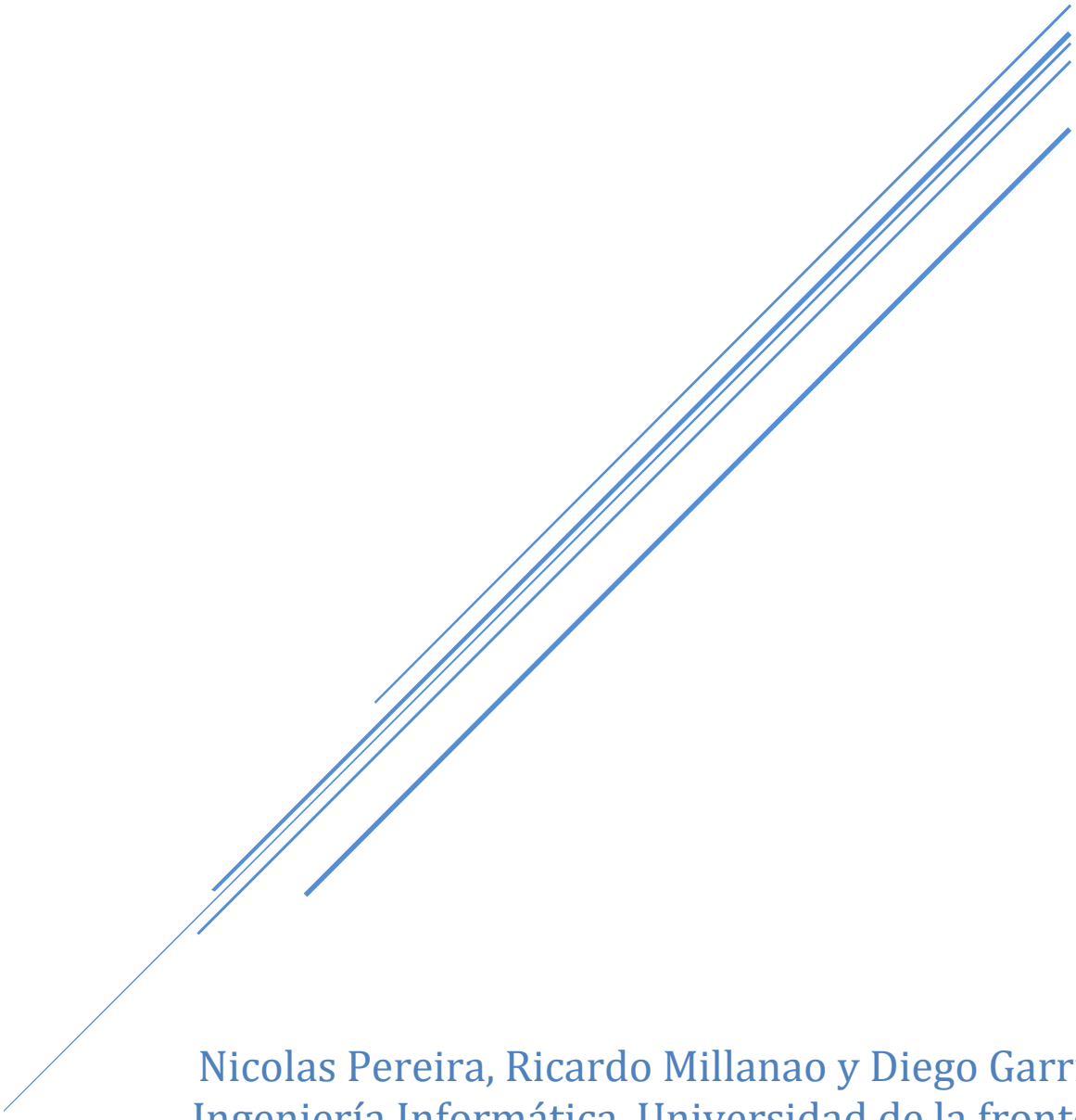
HITO 3 GRUPO 3: CRIMES US



Nicolas Pereira, Ricardo Millanao y Diego Garrido
Ingeniería Informática, Universidad de la frontera
ICC 732-1: Ingeniería de datos

HITO 3 Grupo 3: Crime Crimes US

Diego Garrido, Ricardo Millanao y Nicolás Pereira
Ingeniería Informática, Universidad de la frontera
d.garrido07, r.millanao02, n.pereira01 como @ufromail.cl
Temuco, Chile 15 noviembre, 2022

***Resumen*- El siguiente paper refleja la realización paso a paso de un proyecto de Ingeniería de Datos.**

Este proyecto se basa en un dataset, el cual representa alrededor de 300000 llamados de emergencias por crímenes dentro de Estados Unidos.

En el proyecto se busca responder preguntas de clasificación y de clustering surgidas por un EDA previo mediante métodos como Árbol de decisión, KNN y K-means

I. INTRODUCCIÓN

La criminalidad en el mundo ha sido un tema a tratar desde el inicio de la sociedad. Los estados y países continuamente se han propuesto bajar los índices de delitos. En la actualidad se poseen tecnologías que permiten predecir la ocurrencia de ciertas eventualidad mediante el machine learning.

Para este documento se tiene en posesión un dataset que representa las características de algunos de los llamados de emergencia por causalidad de delitos en algunas ciudades de Estados Unidos en el transcurso de los años 2016 y 2022.

Tomaremos como labor principal la inspección de los datos recopilados en este dataset, el análisis y la detección de problemáticas relacionadas a estos llamados. En base a esto se realizará un modelo de machine learning a modo de saber si es posible responder a las problemáticas detectadas.

II. METODOLOGÍAS

Resumen:
La metodología principal utilizada para realizar este proyecto, se efectuó en tres partes fundamentales:

- Análisis preliminar: Se analiza brevemente el dataset
- Limpieza de datos: Se eliminan los valores NA del dataset y columnas que no se utilizarán.
- Tratamiento de datos: Se obtienen nuevas columnas que pueden ser relevantes para el Análisis de datos y Experimentos
- Análisis de datos: Se analizan los datos en busca de relaciones entre estos.
- Preguntas-Problemas: Se formulan preguntas y problemas en base al análisis de datos.
- Experimentos: Se formulan los experimentos a realizar
- Resultado de Experimentos: Resultados obtenidos de los experimentos
- Análisis de Experimentos: Se analizan y responden a las preguntas y problemas planteados en base a los resultados obtenidos.

II.I. Análisis preliminar

El Análisis exploratorio de datos está compuesto de la elección de un conjunto de datos óptimo para la realización de este proyecto, de este modo se eligió un dataset compuesto por llamadas de emergencia dentro de Estados Unidos, el cual consta de 306094 filas y 36 columnas originalmente.

II.II. Limpieza de datos

Dentro del análisis se implementaron técnicas de limpieza que permitió evaluar si nuestro dataset cumplía con las características suficientes para realizar los experimentos posteriores. Se identificaron las siguientes columnas que poseían valores nulos:

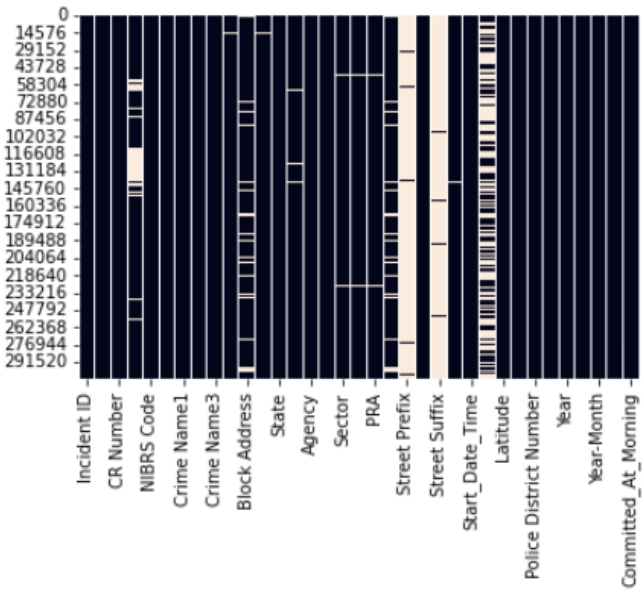


Figura 1: Datos nulos

Luego de la limpieza de los datos, el conjunto de datos resultante queda con 253.852 filas y 17 columnas, de esta

forma comprobaremos si estas nuevas variables nos dan un mejor resultado.
Las columnas a utilizar serán las siguientes:

1. Victims: número de víctimas en el llamado de emergencia.
2. Dispatch Date Time: fecha completa en la cual ocurrió el delito
3. Committed_At_Morning: si fue cometido en la mañana o en la tarde
4. Crime name 1: Categoría de crimen general
5. Crime name 2: Categoria de delito
6. Crime name 3:Especificación del delito
7. City: Ciudad en la cual aconteció el delito
8. Place: Direccion del delito
9. Sector:Sector de la patrulla designada para acontecer al suceso
10. Beat:Policías designados al sector
11. Street name: nombre de la calle del suceso
12. Street type: Categoria de calle
13. Latitude: latitud del lugar exacto del suceso
14. Longitude: longitud del lugar exacto del suceso
15. Year: Año del delito
16. Month: Mes que se realizó del delito
17. Day: Día que se realizó el delito.

II.III. Tratamiento de datos

II.III.I. Fechas

Se realiza una transformación en la variable *Dispach Date Time*, la cual corresponde al momento exacto en el cual se acudió a la emergencia. De esto obtenemos una serie de variable que creemos que puede tener impacto en la ocurrencia de delitos:

- date_complete: Momento en el que se enviaron a los agentes policiales en formato datetime.
- Year: Año cuando se realizó el delito.
- Month: Mes cuando se realizó el delito.
- Day: Dia cuando se realizó el delito.
- Hour: Hora cuando se realizó el delito.
- DayOfYear: Día del año cuando se comete el delito.
- Week: Número de la semana cuando se comete el delito.
- Quarter: Cuarto del día cuando se comete el delito
- DayOfWeek: Número de la semana cuando se comete el delito.
- month_name: Nombre del mes cuando se comete el delito.
- day_name: Nombre del día cuando se comete el delito.

II.III.II. Datos Meteorologicos

Dado los tipos de datos que se tienen dentro del dataset, se ha decidido como equipo que es necesario buscar otras fuentes de datos para complementar nuestro dataset como lo es el caso del siguiente ejemplo del análisis de la ciudad de Montgomery County

Creemos que las variables meteorológicas pueden ser útiles para determinar los comportamientos de los delitos en nuestro dataset

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: Temperatura en grados celsius
- windspeedKmph: velocidad del viento en km/h
- humidity:Humedad en el ambiente
- precipMM:Precipitación en milímetros
- cloud cover :La proporción de recubrimiento de nubosidad

Además de agregar una columna para saber si es una fecha especial, la cual se denomina “Is conmemoration day”

II.III.III Festividades

Creemos que las fechas conmemorativas o pueden tener alguna incidencia en la ocurrencia de crímenes, es por eso que se establece una nueva variable binaria que indique si el delito ocurrió un día festivo:

- 'Is_Commemoration_Day': Indica si el delito se comete en un día conmemorativo

II.III. Análisis de datos

Posterior a la adición de estos datos se incorpora una matriz de correlación para determinar las relaciones entre variables cuantitativas

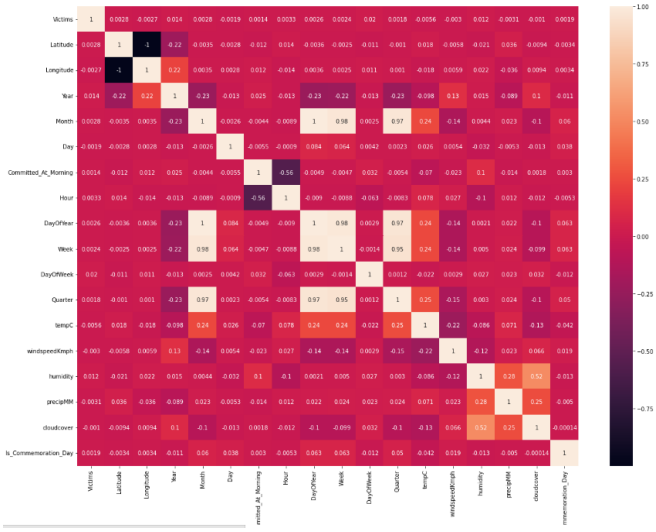


Figura 2: Matriz de correlación

En ella no se pudo observar de manera significativa relaciones entre variables importantes, más que entre las de meteorología, por eso en la etapa posterior, para aclarar dudas

del equipo y demostrar las tendencias delictuales se procede a implementar una serie de gráficos, los cuales nos sirven para detectar ciertos estándares entre el comportamiento delictual.

En primera instancia se procede a implementar un gráfico de barras, el cual nos muestra la cantidad de delitos cometidos dentro de un mes, según el año

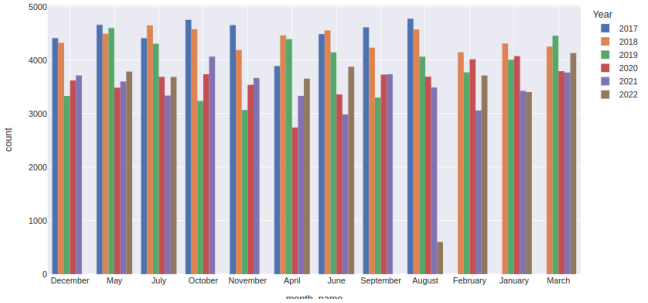


Figura 3: comparativa de los meses del año

Con el pudimos darnos cuenta que el año 2017 y el 2022 están incompletos, el año 2017 está comprendido entre el mes de abril hacia delante y el año 2022 se comprende entre los meses de Enero a Junio, ya que al momento de comenzar este proyecto estábamos a mediados de agosto, por ende este mes se encuentra disminuido.

En general este gráfico tiene un carácter descendente entre la cantidad de delitos cometidos, sin embargo se aprecia que los meses de primavera y verano del hemisferio norte poseen una mayor conducta delictiva en comparación a los de invierno.

El siguiente gráfico representa la conducta delictual en los días de la semana, según la jornada (mañana-tarde).

Figura 4: Comparativa de jornadas

En este se puede apreciar que hay una variación importante entre jornadas, dado que la tarde representa una mayor cantidad de delitos registrados en comparación con la mañana.

El gráfico que se muestra en esta oportunidad tenemos un registro de la conducta delictual comprendidas en las 24 horas de un día.

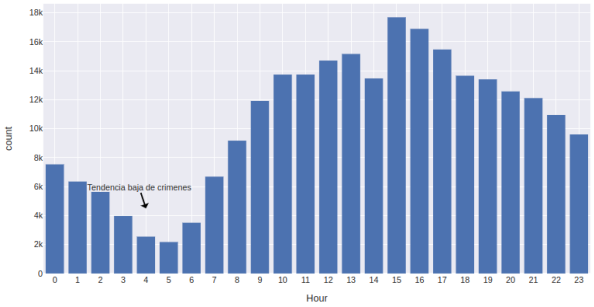


Figura 5: Comparativa horas del dia

En él se puede observar que la gran mayoría de los delitos se comprende entre las horas comprendidas desde las 13hrs a las 18 hrs.
Y existe una disminución de los delitos entre las horas comprendidas en el rango de 3 a 6 hrs.

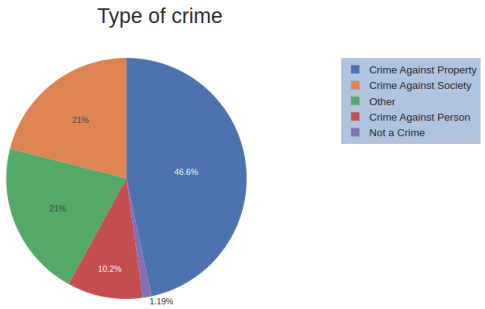


Figura 6: Tipos de crimen recurrentes

En este gráfico se puede observar la distribución entre los tipos de crímenes, se puede ver que la proporción de los crímenes contra la propiedad representan una parte importante en nuestro universo de datos siendo este aproximadamente un 47% del total de crímenes cometidos, también se aprecia que las llamadas que no son crímenes son mínimas representando solo un 1.19%.

El gráfico a continuación representa un gráfico el cual nos deja saber la tendencia en la cantidad de crímenes, esto por su tipo de crimen.

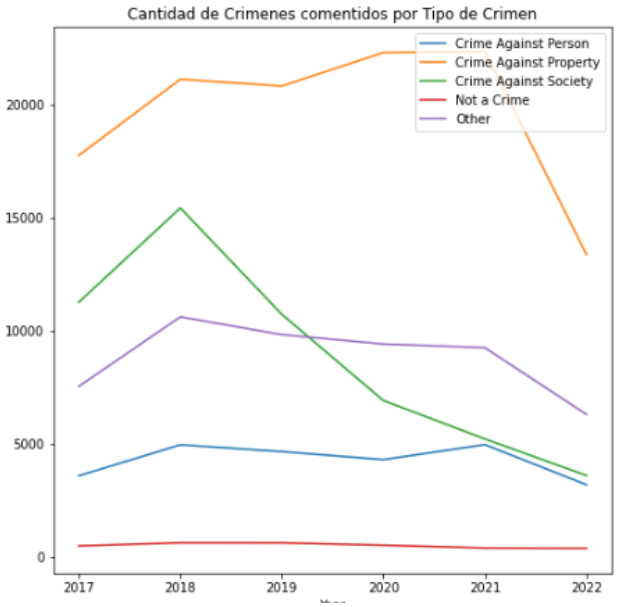


Figura 7: Cantidad de crímenes cometidos

En él se puede ver que existe una tendencia decreciente en todos los tipos de crímenes, pero con mayor caída en los crímenes contra la propiedad en el transcurso del año 2021.

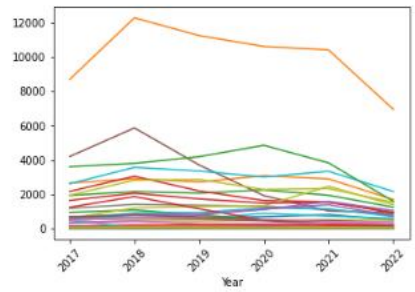


Figura 8: Detalle de los crímenes cometidos

En este gráfico se aprecia un detalle del gráfico anterior, esto tomando en consideración el detalle del tipo de crimen ejecutado. Sin embargo se puede observar la misma tendencia descendente en los últimos años.

El gráfico que se muestra en esta ocasión, trata de la delictualidad en los sectores

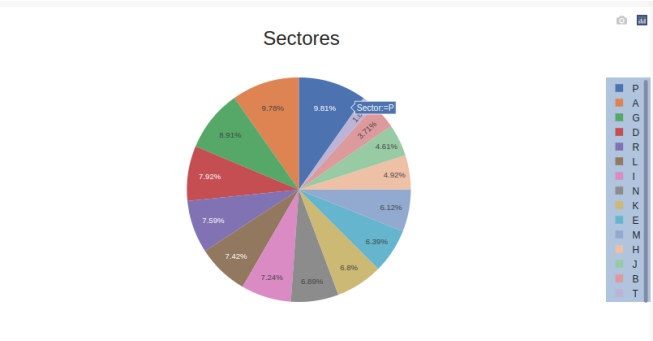


Figura 9: comparativa porcentajes de crímenes cometidos en sectores

En él se puede apreciar que todos los sectores tienen una cantidad equitativa, los sectores que más se distancian del resto serían el P, A y G, siendo los que tienen mayor diferencia.

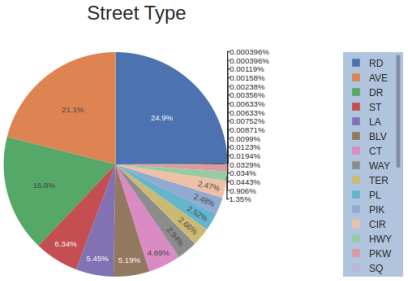


Figura 10: comparativa porcentaje de crímenes cometidos en el tipo de calle

En el gráfico que podemos apreciar anteriormente, podemos ver la conducta delictual según el tipo de calle en el que se comete, en el podemos ver que las calles de tipo carretera o “Road ” son las que presentan mayor proporción de crímenes junto con las avenidas.

Por último en el siguiente gráfico se visualiza la proporción de delictualidad, esto según la temperatura ambiente del momento de cometer el delito

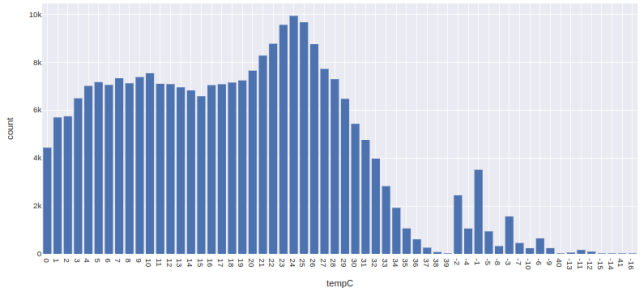


Figura 11: Comparativa cantidad de crímenes por temperatura ambiental

Se puede apreciar que la mayor cantidad de delitos se concentran en las temperaturas más templadas(entre 21 y 27 grados celcius), mientras que la menor proporción se encuentra en temperaturas extremas 39, 40 y 41 grados celsius, en caso de temperaturas altas y -16, -15 y -14 en las temperaturas bajas.

II.IV Preguntas

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.

III. EXPERIMENTOS

Luego de haber realizado el análisis de los datos entregados por el EDA, es posible establecer que tipo de problemas queremos resolver y qué preguntas surgen. De esta forma damos a conocer los dos experimentos abordados para este proyecto.

A. Experimento 1

El experimento 1 aborda la pregunta 1 realizada en la sección II.IV, donde definimos esta pregunta como un problema de clasificación. Este problema se realizará con el algoritmo de Árbol de decisión. Para evaluar el modelo se tomará como métrica principal precisión.

La variable target es “Crime Name2” donde se utiliza “Crime Name3” para filtrar los crímenes del tipo robo. Las variables o columnas predictoras son:

- 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

Para la creación del modelo se define el set de prueba y el de entrenamiento con un 33% de los datos como datos para realizar el testeo de la predicción .Estos se encuentran supervisados mediante las métricas establecidas por el método.

Posteriormente a la creación del modelo se realiza un evaluación mediante `cross validation`, o con la combinación de splits de 10 y de 15.

B. Experimento 2

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

Para el modelo de árbol de decisión se toma las variables X (features) y (corresponde a la clase) para generar el árbol de decisión. Con el método fit entrenamos el clasificador con los datos de X y la clase asociada a cada uno. Para ver qué tan bien fue el entrenamiento, podemos evaluar el clasificador ejecutándose sobre instancias y verificando que la clase sea la correcta. Pero para no ejecutar el modelo sobre instancias que utilizamos para entrenarlo, es que tenemos el método *train_test_split* [5] 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.

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

Los resultados arrojados por la matriz de confusión son

```
[[ 1457  4787  2245     0     0]
 [    0 32986  5704     0     0]
 [    0  8547  9106     0     0]
 [    0   650   347     0     0]
 [  136 13399  3985     0     0]]
```

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

Para la evaluación de los modelos se utilizó la función *cross_val_score* con los datos iniciales X (valores predictores) e y (variable target), dando el resultado la siguiente matriz:

```
Cross Validation Scores are [0.49694162 0.40188859 0.39173298 0.62875638 0.65649127 0.5
 0.62008552 0.4221404  0.56384369 0.53112009 0.45504217 0.5603991
 0.38555648 0.5823732  0.54543295]
```

Además el Average Cross Validation score da como resultado el valor de :0.5185315083278949

Finalmente el resultado obtenemos una lista que refleja los valores por cada validación de los datos de prueba, donde los valores rondan entre 36% y como máximo 63%

Para el modelo KNN, utilizamos el también el método train_test_split para dividir el conjunto de datos en 4 features y utilizamos la clase KNeighborsClassifier para entrenar el algoritmo a partir de los features X_train e y_train. En primera instancia tenemos los resultados preliminares a partir del método *score* de la misma clase anterior, el cual refleja el Accuracy de los set de entrenamiento y de test. Los valores resultantes son; Accuracy of K-NN classifier on training set: 0.57
Accuracy of K-NN classifier on test set: 0.49

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	0.20	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

Para la evaluación de este modelo, también utilizaremos cross validation, ya que nos permitirá testear el modelo a partir de la cantidad de segmentaciones que utilicemos, esto nos da la siguiente matriz

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

esta lista que refleja los valores por cada validación de los datos de prueba, donde los valores como mínimo son de 5% y como máximo un 46%

C. Experimento 3

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**.** Dentro de este algoritmo debemos resaltar las siguientes variables [6].

- labels_: proporciona etiquetas de clase predichas (clúster) para cada punto de datos.

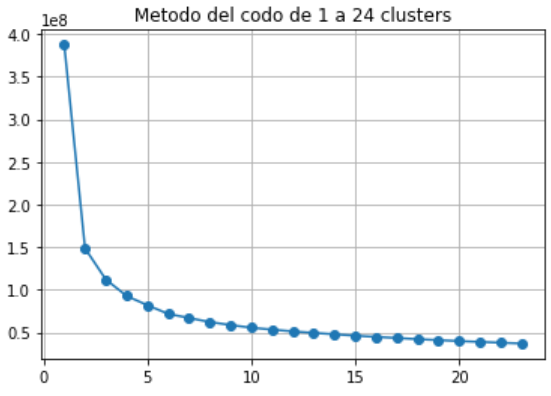
- inertia_: da la suma de cuadrados dentro del grupo. Este es un total de la suma de cuadrados dentro del clúster para todos los grupos.
- n_iter_: número de iteraciones que ejecuta el algoritmo k-means para obtener una suma mínima de cuadrados dentro del clúster

Estos atributos son importantes para poder realizar el algoritmo de Clustering.
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

Luego de la selección de los atributos se inició la construcción del algoritmo bajo la prueba del **método del codo**, la cual consiste en la implementación de un patrón de comportamiento el cual permite establecer una métrica inicial fundamental para entrenar el modelo inicial y finalmente el entrenamiento del modelo.

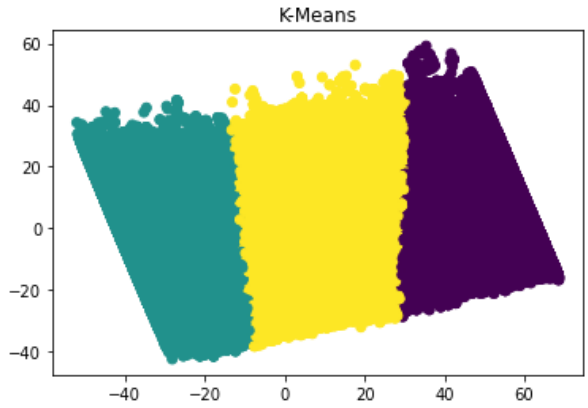
FIGURE II
MÉTODO DEL CODO



El gráfico resultante nos muestra el error de K-Means usando diferentes números de clusters. En nuestro caso podemos notar que un valor óptimo es 3 (mirar donde se forma el codo o el punto tras el cual el error decrece de manera menos significativa). Si eligiéramos 4 o más, veríamos más particiones, pero posiblemente estaríamos separando clusters ya existentes en clusters más pequeños

Lo siguiente es visualizar los cluster obtenidos a partir de la métrica obtenida anteriormente:

FIGURE III
VISUALIZACIÓN DE CLUSTERS



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.

IV. RESULTADOS

Luego de realizar los correspondientes experimentos, procedemos a realizar el análisis de los resultados y verificar si aciertan o no con la hipótesis o los resultados esperados de cada experimento bajo reglas y métricas establecidas.

IV.I. Experimento 1

El método de classification_report es un resumen de las métricas del modelo de clasificación. En este caso lista la precisión, recall, f1score y support :

	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

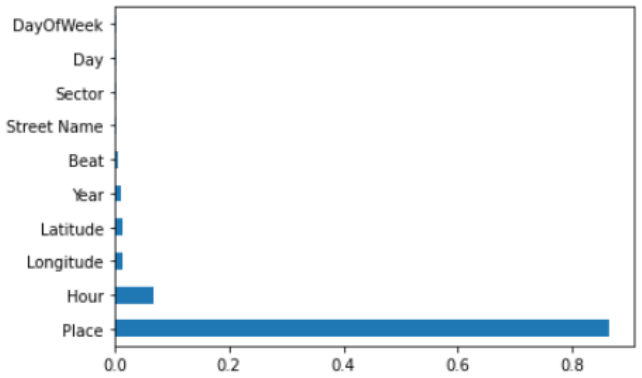
El valor obtenido de Accuracy en test set: 0.4963

```
Cross Validation Scores are [0.40532052 0.45741733 0.4801404  0.48087936 0.53500831 0.53002032
 0.58470349 0.4876201  0.49186992 0.5218034  0.49094605 0.50166297
 0.48484848 0.50831486 0.52679231]
Average Cross Validation score :0.49915652266879884
```

Luego de realizar la clasificación se evalúa el modelo mediante Cross Validation. Como resultado obtenemos una lista que refleja los valores por cada validación de los datos

de prueba, donde los valores rondan entre 40% y como máximo 58%

En base a la clasificación realizada es posible obtener las 10 variable con mayor incidencia en la clasificación de 'Crime Name2' con crímenes de tipo Robo:



Las siete variables con mayor incidencias son (ordenadas de mayor a menor incidencia):

1. Place
2. Hour
3. Longitude
4. Latitude
5. Year
6. Beat
7. Street Name

IV.II. Experimento 2

En el caso del primer experimento propuesto, los resultados de ambos modelos están supervisados por bajo las métricas establecidas por el método *classification_report*

A. Árbol de decisión: Para este modelo los resultados obtenidos están bajo features de entrenamiento con el método *train_test_split*, de esta forma evitamos el *overfitting* y evaluamos el modelo bajo datos de test. Se define el set de prueba y el de entrenamiento con un 33% de los datos como datos para realizar el testeo de la predicción y que siempre ocupe el mismo *random* para la evaluación del resultado.
Para ver qué tan bien fue el entrenamiento, podemos evaluar el clasificador ejecutándose sobre instancias y verificando que la clase sea la correcta. El valor obtenido de Accuracy en test set: 0.5224897719228785

B. KNN: Para este otro modelo los resultados obtenidos están también bajo features de entrenamiento con el método *train_test_split*, y evaluamos el modelo bajo datos de test. En este caso también se define el set de prueba y el de entrenamiento con un 33% de los datos como datos para realizar el testeo de la predicción. Para ver qué tan bien fue el entrenamiento, podemos evaluar el clasificador ejecutándose sobre instancias y verificando que la clase sea la correcta. Los resultados obtenidos tanto para el clasificador de entrenamiento como para el de test son los siguientes;

Accuracy of K-NN classifier on training set: 0.57
Accuracy of K-NN classifier on test set: 0.50

IV.I. Experimento 3

En el caso del segundo experimento propuesto el resultado del modelo está supervisado de una manera no gráfica, es decir bajo una métrica llamada coeficiente de Silhouette la cual establece la distancia promedio entre el resto de los puntos de su misma clase y además la distancia promedio a todos los puntos del cluster más cercano, lo cual significa que si la métrica resultante debiese estar en un rango entre -1 y 1, donde 1 significa que algo está bien asignado, -1 significa que algo está mal asignado y 0 significa que hay solapamiento de clusters.

A. K-Means: Al aplicar el método del codo, el cual nos establece que el valor óptimo de clusters es de 3, se aplica dentro de K-Means, aplicando el método *fit*, para definir a qué cluster pertenece cada punto y nos devuelve los valores correspondientes a las 3 clases esperadas.
El resultado obtenido refleja bajo la función *silhouette_score* el resultado de 0.19659388483593335

V. ANÁLISIS DE LOS RESULTADOS

V.I Experimento 1

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 las métricas obtenidas, que son inferiores al 50% en la mayoría de delito de robo, podemos afirmar que **no es posible predecir los delitos de tipo robo.**

V.II. Experimento 2

Para el modelo de Árbol de decisión el resultado obtenido refleja que existe un 52% de acierto en la predicción del tipo de crimen (Crime Name 1). A partir del resultado podemos afirmar que el modelo no cumple con un porcentaje muy elevado de acierto, estas condiciones pueden explicarse a que, en primer lugar nuestro dataset no está muy bien balanceado, esto se puede explicar ya que más del 50% de los llamados de emergencia están categorizados como asaltos a la propiedad y las demás tipos de crímenes poseen una baja proporción dentro del conjunto de datos.

Para el modelo con KNN el resultado obtenido refleja que hay un 57% de acierto sólo en los test de entrenamiento pero el porcentaje disminuye a un 50% cuando el clasificador utiliza los set de test, de igual manera que el anterior este posee un valor similar, si bien tenemos un mejor resultado con el set de entrenamiento no hay una gran diferencia cuando usamos el set de test, al igual que el anterior la explicación es que muchos filas del dataset no están debidamente equilibrados al momento de estar categorizados por el tipo de crimen, pero podemos implementar técnicas más avanzadas en este tipo de algoritmo para evidenciar mejores resultados al momento de clasificar los crímenes.

V.III. Experimento 3

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. Pero, si los datos no están bien separados, sería difícil encontrar el número óptimo de clusters. Por otra parte en una primera instancia el resultado obtenido sin los datos agregados al dataset el resultado era cercano 0.167, esto podría significar que las variables de tiempo y clima si influyen en la ocurrencia de un crimen.

CONCLUSIONES

Bajo los resultados de los experimentos realizados, podemos concluir que la aplicación de las técnicas utilizadas en los experimentos no permite establecer un valor óptimo que sea rescatable, pero si implementaremos técnicas aún más profesionales podríamos tener un resultado mejor.

La importancia de elegir un conjunto de datos óptimo para realización de un proyecto de estas características es esencial, esto quiere decir que la compresión de la estructura, variables y contexto son relevantes para tener una base sólida para el desarrollo y construcción de los algoritmos, si bien la realidad es que no siempre es posible conseguir más datos de las clases minoritarias los mejor que podemos hacer es buscar y evaluar cada conjunto de datos de acuerdo a nuestras necesidades.

REFERENCIAS

[1] User Guide — pandas 1.5.1 documentation. (s. f.). https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

[2] Donohue, Susan K. and Richards, Larry G. October 2011. “P-12 Engineering Education: Using Engineering Teaching Kits to Address Student Misconceptions in Science.” *Proceedings* of the 41st Frontiers in Education Conference, Rapid City, SD, pp. F2A-1 – F2A-3.

[3] 3.1. Cross-validation: evaluating estimator performance. (s. f.).scikit-learn. https://scikit-learn.org/stable/modules/cross_validation.html

[4] Kaplan, Avi and Maehr, Martin L. June 2007. “The Contributions and Prospects of Goal Orientation Theory.” *Educational Psychology Review* 19(2), pp. 141 – 184.

[5] sklearn.model_selection.train_test_split. (s. f.). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[6] Bedre, R. (2022c, abril 10). k-means clustering in Python [with example]. Data science blog. https://www.reneshbedre.com/blog/kmeans-clustering-python.html

[7] La función confusion_matrix | Interactive Chaos. (s. f.). https://interactivechaos.com/es/manual/tutorial-de-machine-learning/la-funcion-confusionmatrix

INFORMACIÓN DE AUTORES

Diego Garrido, Estudiante, Ingeniería Informática, Universidad de la Frontera.

Ricardo Millanao, Estudiante, Ingeniería Informática, Universidad de la Frontera

Nicolás Pereira, Estudiante de Ingeniería Informática, Universidad de la Frontera.