DATA SCIENCE

Proyecto Final

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS

01

Presentación

PRESENTACIÓN

Mi nombre es Jessica Oshiro, actualmente me encuentro trabajando como analista de procesos en una reconocida empresa nacional con presencia en varias partes del mundo.

Introducción

02

INTRODUCCIÓN

Este proyecto está orientado a ayudar a la Policía Federal de Los Ángeles, Estados Unidos. El objetivo es poder brindarles un modelo de machine learning que les ayude a predecir incidentes en la ciudad para poder tomar las medidas necesitas y de esta forma, reducir la cantidad de incidentes.

03

Acerca de la información

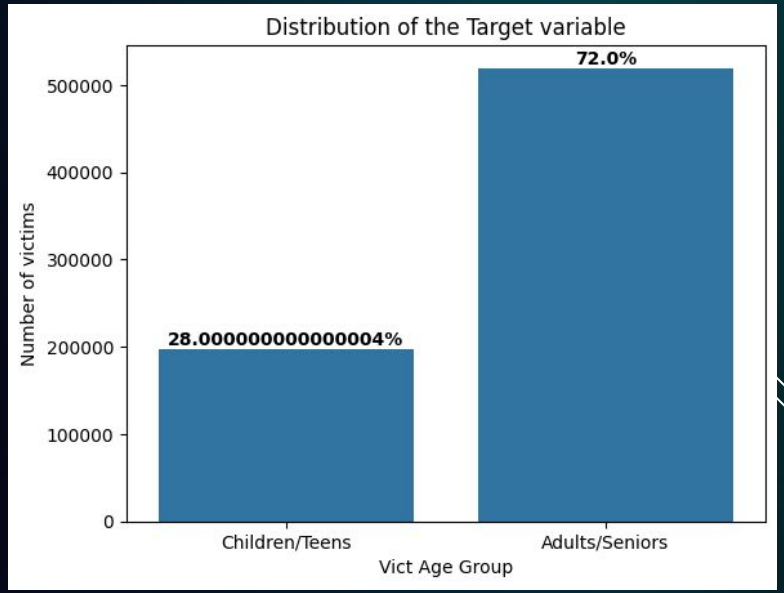
ACERCA DE LA INFORMACIÓN

El dataset utilizado está disponible en "Los Angeles Open Data". El mismo abarca los incidentes criminales en Los Ángeles entre los años 2020 y 2023. Cuenta con variables como la distribución geográfica de crímenes, perfiles demográficos de las víctimas y detalles sobre tipos de crímenes, mi motivación es revelar patrones y tendencias significativas para el alto mando de la policía federal de Los Ángeles.
El dataset cuenta con 717,699.

VARIABLE TARGET

Vict Age Group

Se realiza dos grupos de edades para luego poder aplicar un modelo de clasificación. Por un lado tenemos al grupo “Niños/Adolescentes” y por otro lado tenemos al grupo “Adultos/Ancianos”. La edad que define un grupo del otro es 18 años.



Hipótesis y Preguntas

04

HIPÓTESIS

Se plantea la hipótesis de que la edad y el género de las víctimas de crímenes en Los Ángeles están correlacionados con la naturaleza y la ubicación de los incidentes. Se espera que ciertos grupos demográficos sean más susceptibles a ciertos tipos de crímenes y que la distribución geográfica de los incidentes varíe en función de factores demográficos específicos. Además, se sugiere que la prevalencia de ciertos tipos de crímenes puede variar según la hora del día y la ubicación, afectando a diferentes grupos de edad de manera diferente. Esta hipótesis busca explorar las complejas interacciones entre la demografía de las víctimas y la naturaleza de los crímenes, proporcionando una comprensión más profunda de los patrones subyacentes en los datos de crímenes de Los Ángeles.

ALGUNAS PREGUNTAS

1

¿Existe una correlación entre la edad de las víctimas y la frecuencia de crímenes en áreas específicas de Los Ángeles?

2

¿Cómo varía la distribución de género de las víctimas según la edad?

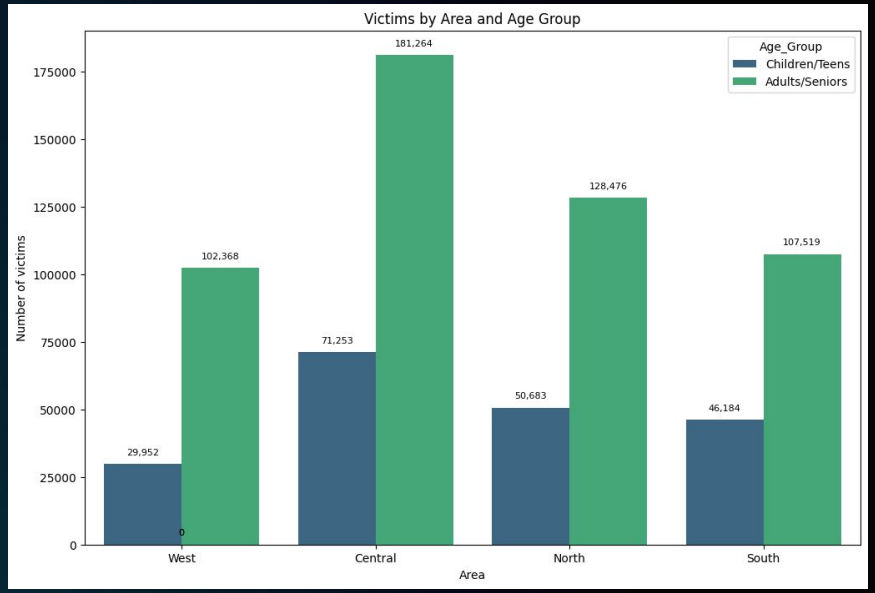
3

¿La hora del día afecta la naturaleza de los crímenes en función de la edad de las víctimas?

4

¿Cómo se distribuyen los crímenes en relación con la edad de las víctimas y el día de la semana de los incidentes?

¿Existe una correlación entre la edad de las víctimas y la frecuencia de crímenes en áreas específicas de Los Ángeles?

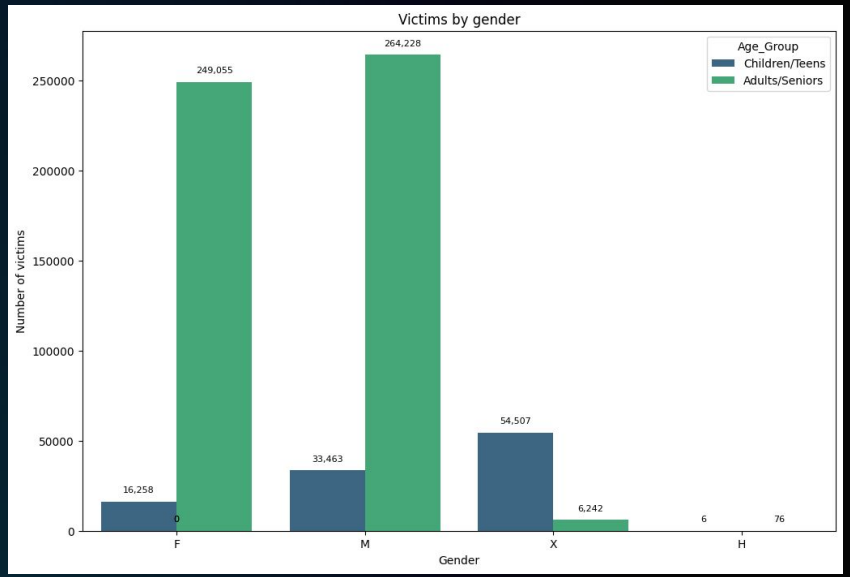


Respuesta:

Se sigue haciendo evidente que el mayor grupo afectado es el de adultos y ancianos, y en donde en "Central" es donde se encuentra más concentrado los crímenes.

Nota: Dado que la lista de áreas era extensa, se optó por agruparlos según regiones al Oeste, Centro, Norte y Sur.

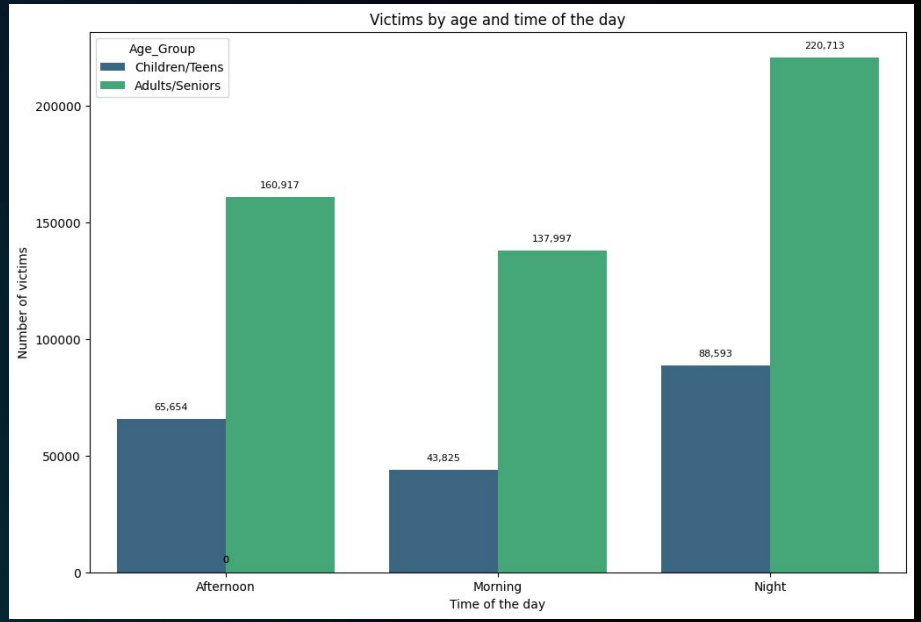
¿Cómo varía la distribución de género de las víctimas según la edad?



Respuesta:

Se sigue demostrando que el mayor grupo afectado es el de adultos y ancianos, y en cuanto al género, es poca la diferencia que existe entre víctimas femeninas y masculinas.

¿La hora del día afecta la naturaleza de los crímenes en función de la edad de las víctimas?

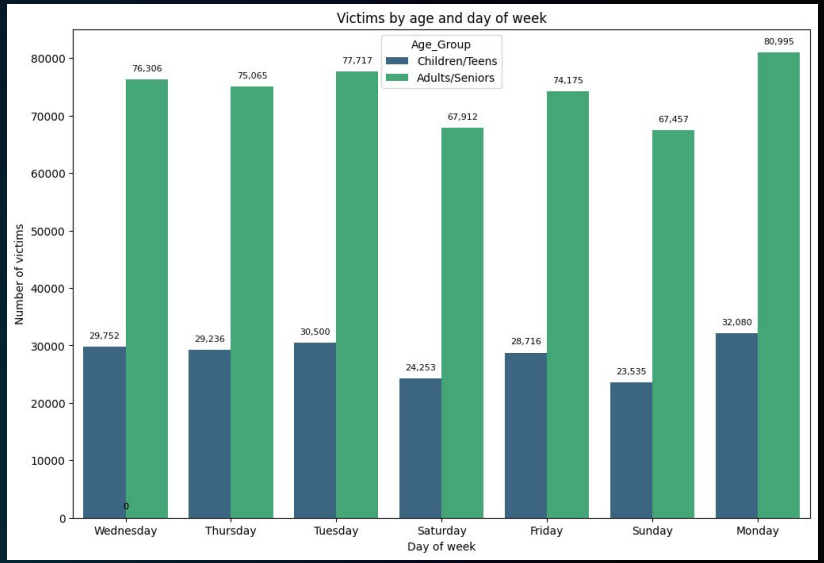


Respuesta:

Se reafirma que el grupo de víctimas destacado es el de adultos + ancianos, y se puede observar que el momento del día en donde pueden llegar a sufrir un accidente es durante la noche.

Nota: Dado que la lista de áreas era extensa, se optó por agruparlos por Mañana, Tarde y Noche.

¿Cómo se distribuyen los crímenes en relación con la edad de las víctimas y el día de la semana de los incidentes?



Respuesta:

El lunes es el día en donde hay más crímenes y los adultos + ancianos son los más afectados, entonces en base a esto, se podría recomendar que se tomen más medidas de seguridad en dicho día y, en base al gráfico anterior, más especificamente durante el turno noche.

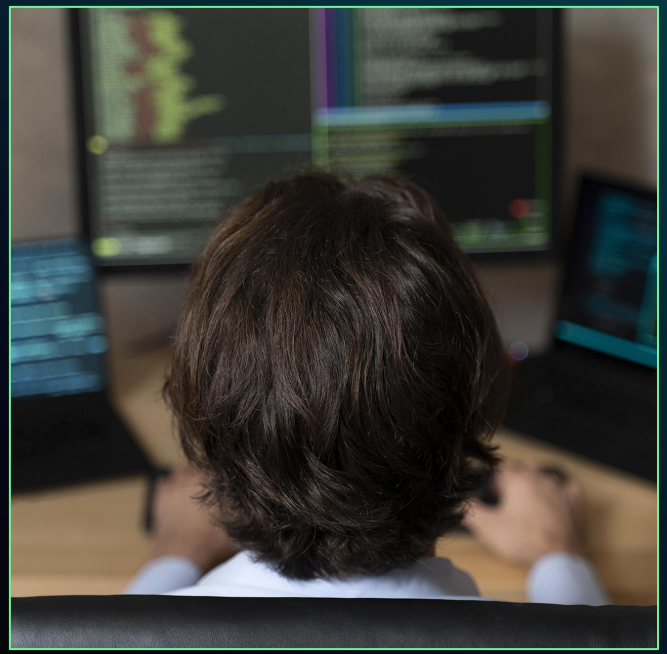
Nota: No contaba con una columna en el dataset que tuviese el día de la semana pero si contaba con una columna de fecha, por lo cual, se creó una columna nueva con la información que necesitaba.

05

Valores Vacíos

EXPLORACIÓN

```
Cantidad de columnas con valores vacíos:
(12,)

Variables con valores vacíos:
Crm Cd 4          717645
Crm Cd 3          715913
Crm Cd 2          664374
Cross Street      601766
Weapon Used Cd    468593
Weapon Desc       468593
Mocodes            98453
Vict Descent       93870
Vict Sex           93864
Premis Desc          398
Premis Cd              8
Crm Cd 1               8
```



TRATAMIENTO

De las 12 variables detectadas se han tratado solamente 3: “Vict Sex”, “Weapon Desc” y “Vict Descent”. El resto se descarta dado que no son relevantes para el análisis.

Se han rellenado los vacíos utilizando la moda. Adicionalmente, para “Weapon Desc” y “Vict Descent” se ha realizado un agrupamiento por categorías para reducir la cantidad opciones dentro de la variable. Por lo tanto, se han creado las columnas “Weapon_Grouped” y “Descent_Grouped”

```
print("Cantidad valores vacíos")
print("Vict Sex:", df['Vict Sex'].isnull().sum())
print("Weapon Desc:", df['Weapon Desc'].isnull().sum())
print("Vict Descent:", df['Vict Descent'].isnull().sum())
```

✓ 0.0s

```
Cantidad valores vacíos
Vict Sex: 0
Weapon Desc: 0
Vict Descent: 0
```

```
print("Valores únicos de Weapon_Grouped:", df['Weapon_Grouped'].unique())
print("Valores únicos de Descent_Grouped:", df['Descent_Grouped'].unique())
```

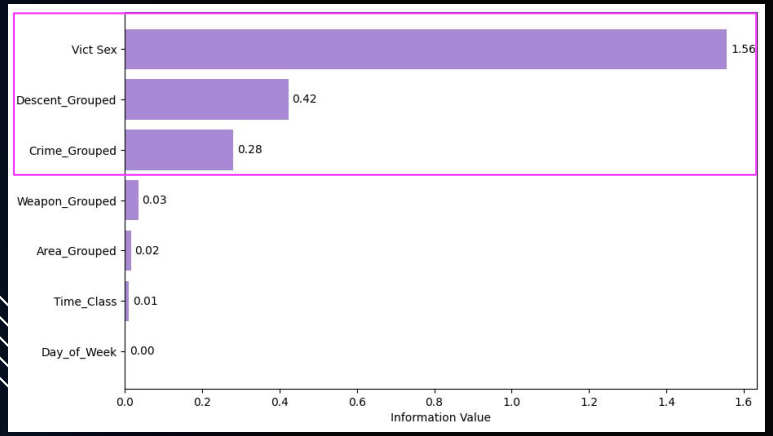
✓ 0.0s

```
Valores únicos de Weapon_Grouped: ['Others' 'Ropes/Sharp Weapons' 'Firearms']
Valores únicos de Descent_Grouped: ['Main Ethnic Groups' 'Unknown or Not Specified' 'Other Ethnic Minorities']
```

Information Value

06

Se ha utilizado la métrica Information Value para conocer las variables con mayor poder de predicción respecto a mi variable target, con el objetivo de poder luego utilizarlas en mis modelos.

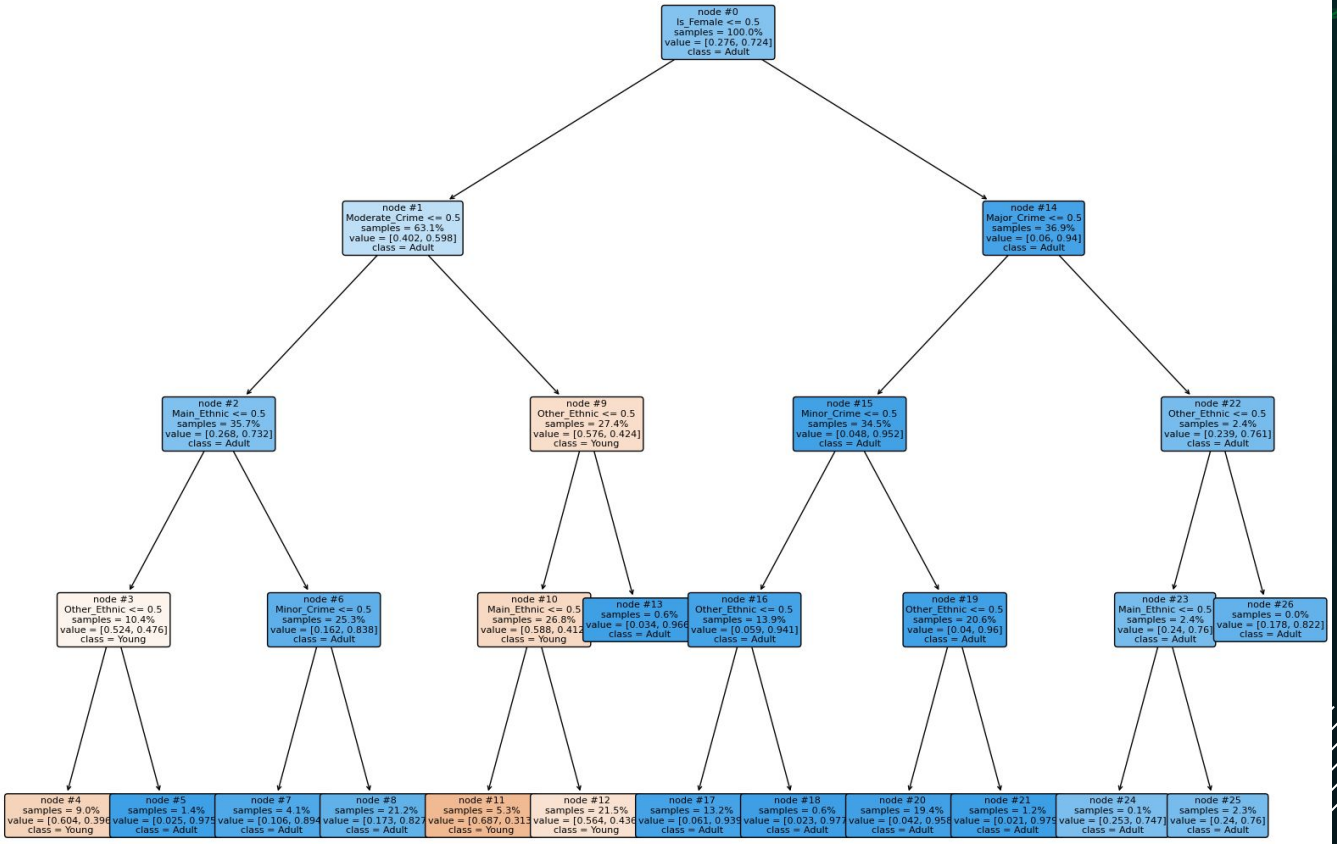


07

Modelo 1: Árbol de Decisión

Árbol de Decisión

Un modelo de árbol de decisión es una técnica de aprendizaje que divide repetidamente un conjunto de datos en subconjuntos más pequeños basados en características específicas. Esto crea un árbol de decisiones donde cada nodo representa una decisión basada en una característica, y cada hoja representa una predicción.

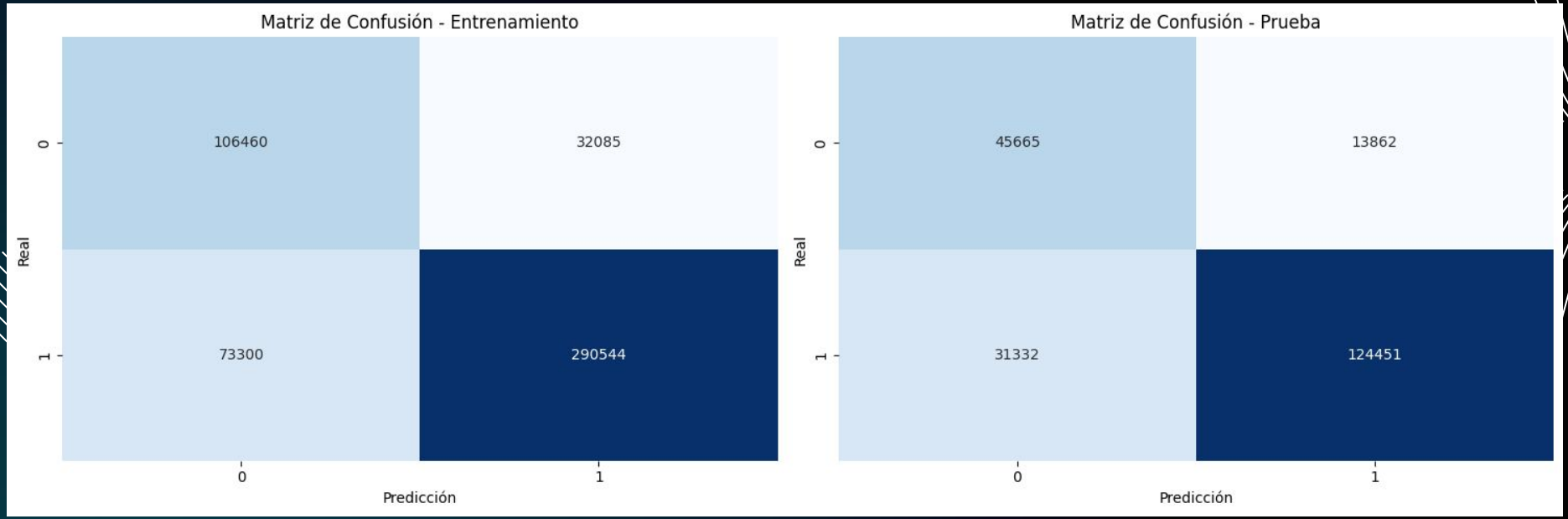

node #0
Is_Female <= 0.5
samples = 100.0%
value = [0.276, 0.724]
class = Adult
node #1
Moderate_Crime <= 0.5
samples = 63.1%
value = [0.402, 0.598]
class = Adult
node #14
Major_Crime <= 0.5
samples = 36.9%
value = [0.06, 0.94]
class = Adult
node #2
Main_Ethnic <= 0.5
samples = 35.7%
value = [0.268, 0.732]
class = Adult
node #9
Other_Ethnic <= 0.5
samples = 27.4%
value = [0.576, 0.424]
class = Young
node #15
Minor_Crime <= 0.5
samples = 34.5%
value = [0.048, 0.952]
class = Adult
node #22
Other_Ethnic <= 0.5
samples = 2.4%
value = [0.239, 0.761]
class = Adult
node #3
Other_Ethnic <= 0.5
samples = 10.4%
value = [0.524, 0.476]
class = Young
node #6
Minor_Crime <= 0.5
samples = 25.3%
value = [0.162, 0.838]
class = Adult
node #10
Main_Ethnic <= 0.5
samples = 26.8%
value = [0.588, 0.412]
class = Young
node #13
samples = 0.6%
value = [0.034, 0.966]
class = Adult
node #16
Other_Ethnic <= 0.5
samples = 13.9%
value = [0.059, 0.941]
class = Adult
node #19
Other_Ethnic <= 0.5
samples = 20.6%
value = [0.04, 0.96]
class = Adult
node #23
Main_Ethnic <= 0.5
samples = 2.4%
value = [0.24, 0.76]
class = Adult
node #26
samples = 0.0%
value = [0.178, 0.822]
class = Adult
node #4
samples = 9.0%
value = [0.604, 0.396]
class = Young
node #5
samples = 1.4%
value = [0.025, 0.975]
class = Adult
node #7
samples = 4.1%
value = [0.106, 0.894]
class = Adult
node #8
samples = 21.2%
value = [0.173, 0.827]
class = Adult
node #11
samples = 5.3%
value = [0.687, 0.313]
class = Young
node #12
samples = 21.5%
value = [0.564, 0.436]
class = Young
node #17
samples = 13.2%
value = [0.061, 0.939]
class = Adult
node #18
samples = 0.6%
value = [0.023, 0.977]
class = Adult
node #20
samples = 19.4%
value = [0.042, 0.958]
class = Adult
node #21
samples = 1.2%
value = [0.021, 0.979]
class = Adult
node #24
samples = 0.1%
value = [0.253, 0.747]
class = Adult
node #25
samples = 2.3%
value = [0.24, 0.76]
class = Adult

APLICACIÓN DEL MODELO

Generamos un conjunto para entrenamiento y otro para testeo. Aplicamos el algoritmo de árbol de decisión sobre la muestra de entrenamiento, y evaluamos su rendimiento contra el conjunto de testeo.

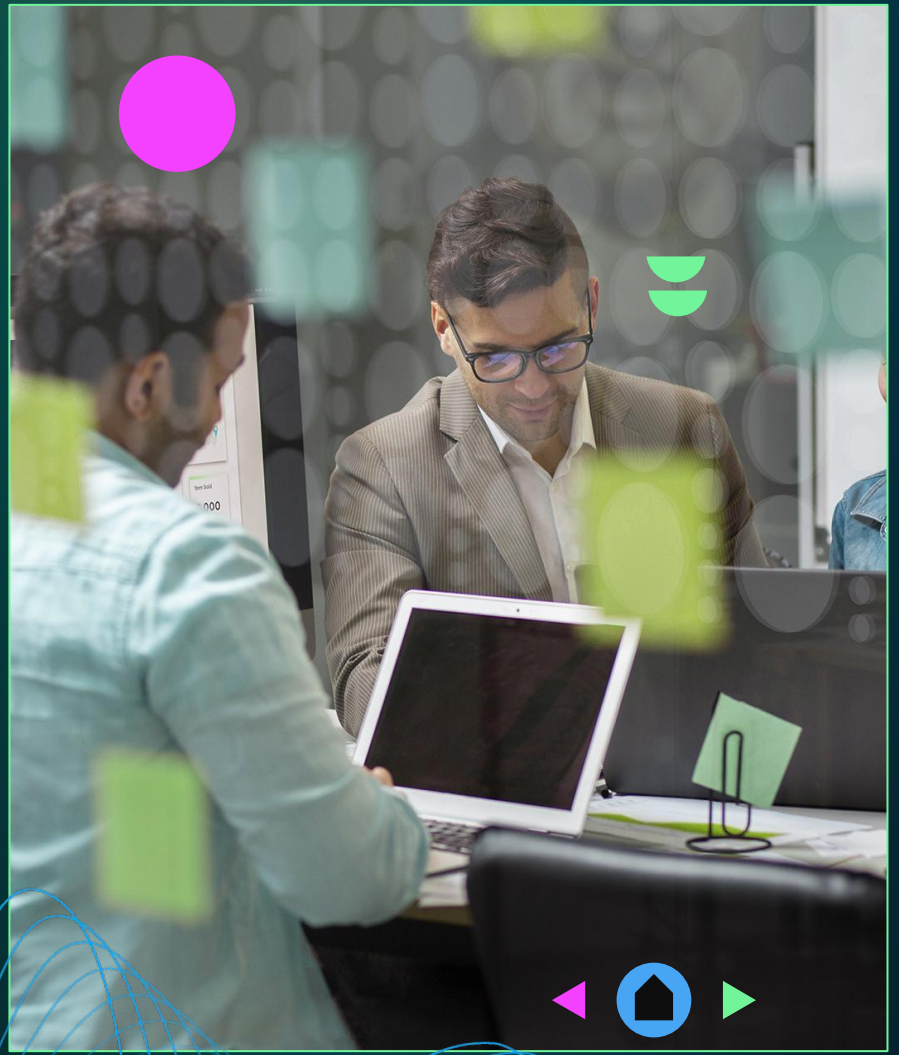
```
Métricas para el conjunto de entrenamiento:
Accuracy: 0.7902322702129226
Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.77      0.67    138545
           1       0.90      0.80      0.85    363844

    accuracy                           0.79    502389
   macro avg       0.75      0.78      0.76    502389
weighted avg       0.82      0.79      0.80    502389


Métricas para el conjunto de prueba:
Accuracy: 0.7900979982351028
Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.77      0.67     59527
           1       0.90      0.80      0.85    155783

    accuracy                           0.79    215310
   macro avg       0.75      0.78      0.76    215310
weighted avg       0.81      0.79      0.80    215310
```

MATRIZ DE CONFUSIÓN



CONCLUSIÓN:

El modelo de árbol de decisión muestra una precisión consistente tanto en el conjunto de entrenamiento como en el de prueba, con una precisión global cercana al 79%. Puede predecir eficazmente tanto crímenes menores como mayores, lo que sugiere una buena capacidad de generalización a datos nuevos. Este modelo puede ser útil en la planificación de estrategias policiales y de prevención del delito.



Modelo 2: Random Forest

08

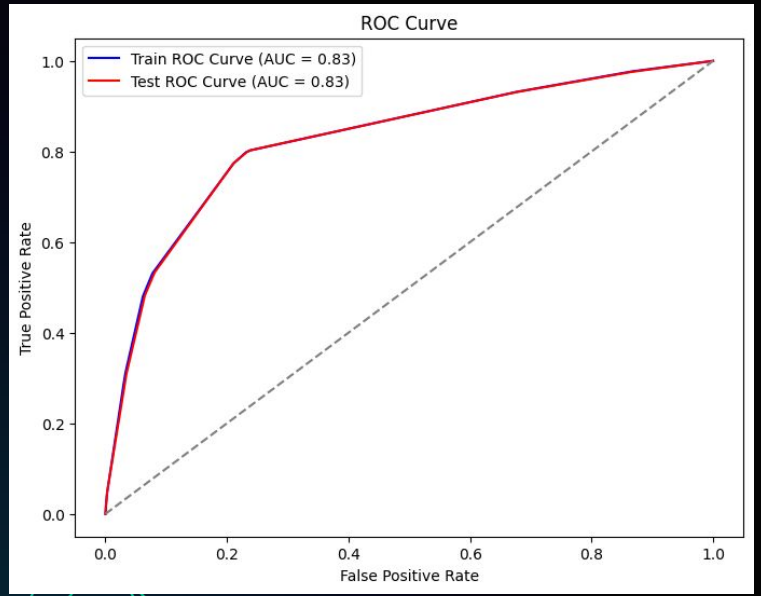
Random Forest

Random Forest es un modelo de aprendizaje automático que combina múltiples árboles de decisión para realizar predicciones más precisas y robustas. Cada árbol en el bosque se entrena de forma independiente utilizando una muestra aleatoria del conjunto de datos y características seleccionadas al azar en cada división del árbol. Luego, las predicciones de todos los árboles se promedian para obtener una predicción final.

APLICACIÓN DEL MODELO

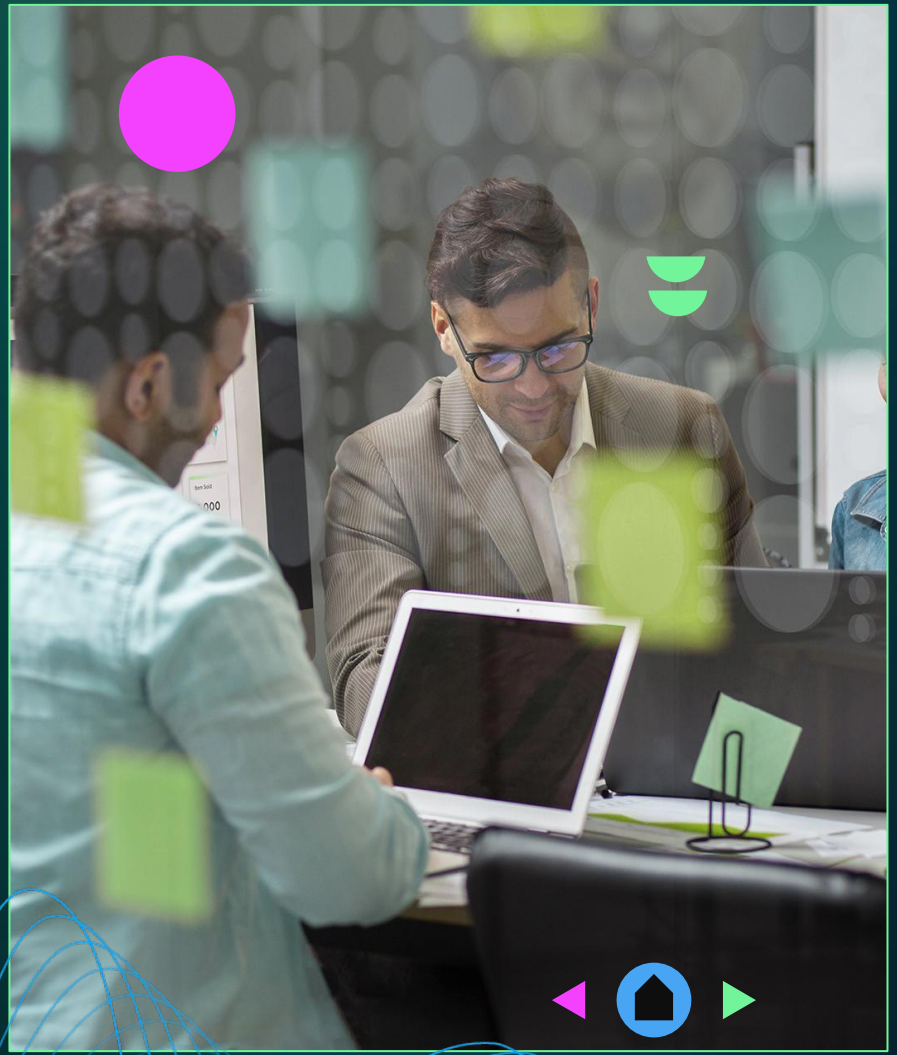
Resultados del modelo y gráfico de la curva ROC para comparar los rendimientos tanto para el grupo de test como train.

```
Accuracy: 0.7912544702986392
Precision por clase: [0.59595694 0.89775924]
Recall por clase: [0.76069683 0.802931  ]
F1-score por clase: [0.66832461 0.84770137]
AUC-ROC: 0.7818139134117047
```



CONCLUSIÓN:

El modelo de Random Forest mostró un rendimiento prometedor en la predicción del objetivo deseado, con una precisión global de aproximadamente 0.79 en el conjunto de prueba. Al evaluar por clase, se observa que la precisión para la clase positiva (Adulto) es alta, con un valor de alrededor de 0.90, mientras que para la clase negativa (Joven) es más moderada, alrededor de 0.60. Además, el modelo presenta un buen equilibrio entre la tasa de verdaderos positivos (Recall) y la tasa de falsos positivos, como lo indican los valores de AUC-ROC de aproximadamente 0.83. Estos resultados sugieren que el modelo es capaz de distinguir eficazmente entre las clases objetivo y puede ser útil en la clasificación de casos en función de las características proporcionadas.



09

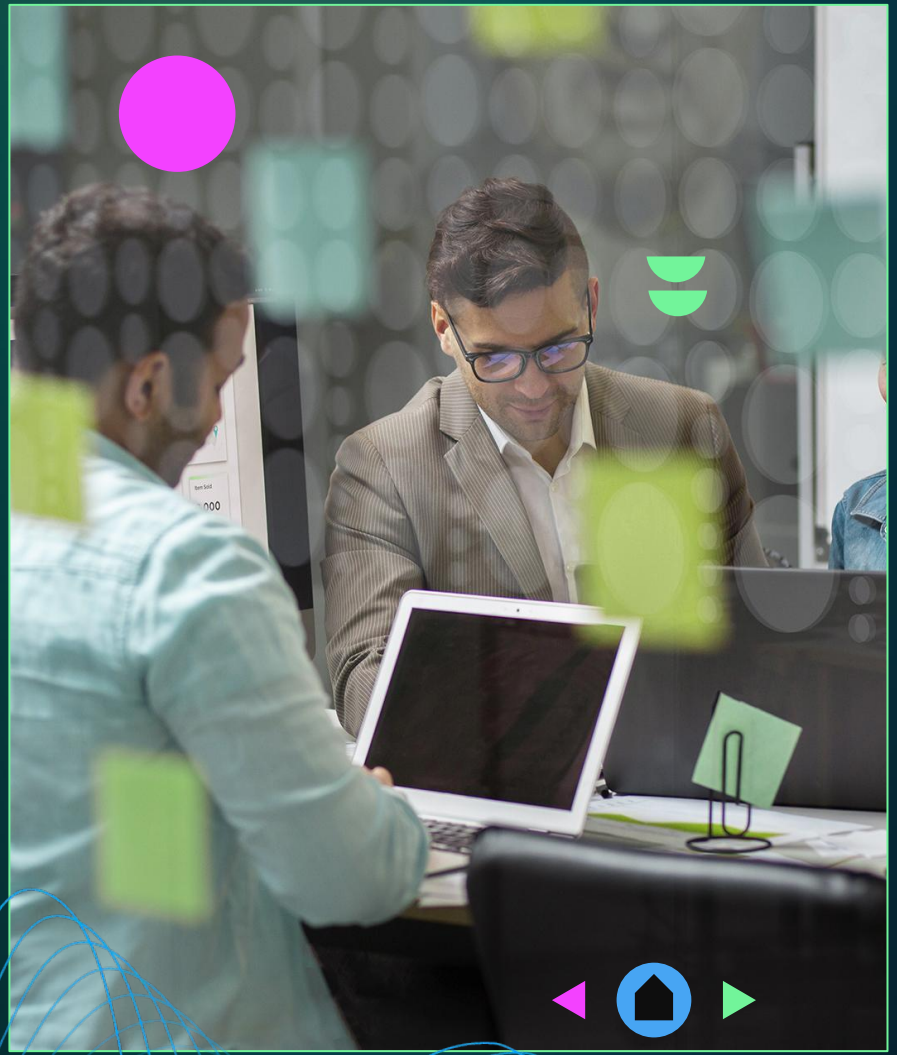
Optimización

GridSearch

El GridSearch es una técnica que busca la combinación óptima de hiperparámetros para un modelo de aprendizaje. Explora sistemáticamente todas las combinaciones posibles de valores de hiperparámetros y selecciona la mejor opción según el rendimiento del modelo en los datos de validación. Es una herramienta clave para maximizar la precisión y el rendimiento de los modelos de aprendizaje automático en entornos ejecutivos.

CONCLUSIÓN:

Utilizando el GridSearch se obtiene una combinación que nos da un accuracy de 0.79. Es más o menos similar a lo que se obtuvo con los modelos anteriores.



Conclusión Final

10

Basado en los resultados obtenidos, el modelo de Random Forest demostró ser el más efectivo en términos de precisión con un alto nivel de exactitud. Esta herramienta podría ser invaluable para la Policía de Los Ángeles en la lucha contra el crimen, ya que permite predecir con mayor precisión la ocurrencia de incidentes delictivos. Al utilizar este modelo, las autoridades podrían asignar recursos de manera más eficiente, priorizando áreas de mayor riesgo y adoptando medidas preventivas para reducir la incidencia delictiva. Además, al tener la capacidad de identificar patrones y tendencias delictivas, el modelo de Random Forest podría proporcionar información valiosa para la formulación de políticas de seguridad pública y la planificación estratégica de la aplicación de la ley en la ciudad

GRACIAS