CODER HOUSE

Data Science
Comisión 19130
Profesor David Bustos Usta
Tutoa Corina Garrido

STOP POLICE BRUTALITY

MUERTES CIVILES CAUSADAS POR FUERZAS POLICIALES EN USA (2015-2020)

Magdalena Gambuli - Jese David Salazar García - Gonzalo Beloqui

ÍNDICE

MOTIVACIÓN

Debido a la creciente problemática mundial de muertes civiles en ocasión de encuentros con fuerzas policiales, hemos decidido analizar los datos relacionados a todos los civiles muertos por parte de la policía en Estados Unidos, para los años 2015 a 2020 ambos inclusive, con el relevamiento de distintas condiciones o variables relacionadas al evento (raza del civil, edad, género, si se encontraba armado, tipificación del encuentro con la policía, si el agente fue enjuiciado o no, entre otros.) y a la ciudad o estado donde ocurrió el hecho, para tratar de relacionar la incidencia de ciertas variables socioeconómicas de los estados.

STOP POLICE BRUTALITY

OBJETIVOS DE LA INVESTIGACIÓN

- ¿Cuál es la probabilidad de que una víctima civil tenga una enfermedad mental al momento del incidente con la policía?
- ¿Cuál es la distribución por raza de las víctimas?
- ¿Cual es la incidencia de los factores socioeconómicos y políticos de los estados en la cantidad de víctimas por millón de habitantes?
- ¿Cuáles son los estados con mayor cantidad de muertes por millón de habitantes?
- ¿Cuáles son las circunstancias mas comunes del encuentro entre el civil y la policía?, ¿los civiles generalmente están armados?

LINEA DE TIEMPO

20/03/2022
Elección personal de un dataset y/o temática posible

24/03/2022
Definición de 3 datasets y temáticas finalistas. Devolución del tutor sobre ventajas y desventajas de cada uno.

01/04/2022
Definición de variable target y aplicación de proceso de Data wrangling

06/04/2022
Desarrollo de análisis univariado

18/04/2022
Desarrollo de análisis bivariado y multivariado

18/05/2022
Desarrollo de algoritmo DECISION TREE y ajuste de modelo por 10 parámetros principales

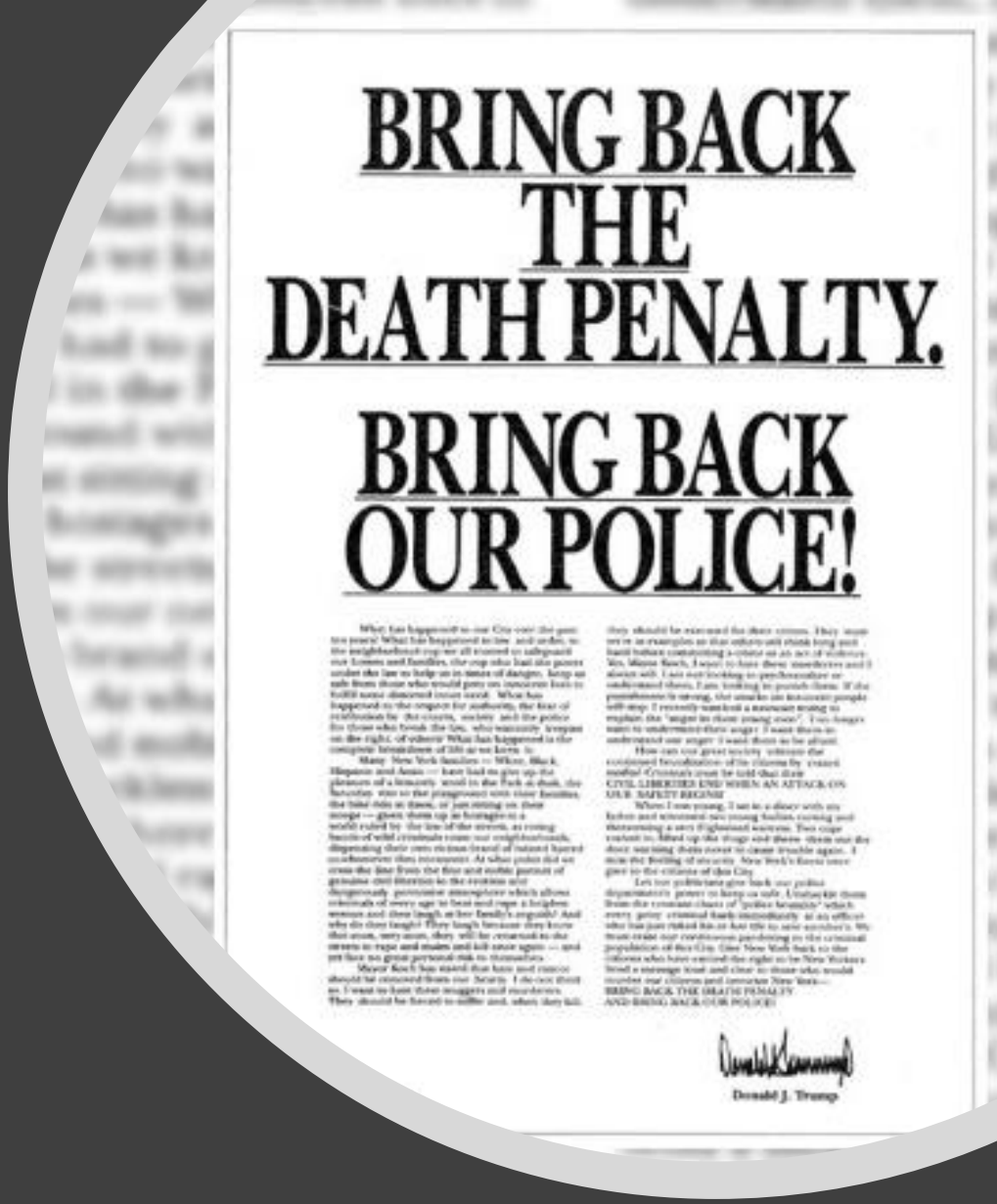
31/05/2022
Desarrollo de algoritmo XBOOSTER CLASSIFIER

06/06/2020
Desarrollo del documento ejecutivo

OBJETIVOS DEL MODELO

Se pretende estimar la probabilidad de ocurrencia de que una persona civil que muere a manos de la policía en EEUU posea una enfermedad mental.

Específicamente, y a los efectos de resaltar una situación poco considerada relacionada a esta problemática social, los análisis se centrarán en la condición de si la victima poseía una enfermedad mental al momento del encuentro con la policía, a los efectos de que la información y conclusiones obtenidas sean de utilidad para entes gubernamentales o asociaciones sin fines de lucro, con el objetivo de desarrollar herramientas de contención para evitar que personas con problemas mentales terminen abatidas por la policía.


BRING BACK THE DEATH PENALTY.

BRING BACK OUR POLICE!

Donald J. Trump

FUENTES DEL DATASET

- Datos de hechos y civiles

https://github.com/washingtonpost/data-police-shootings

- Datos socioeconómicos de los estados

https://data.ers.usda.gov/

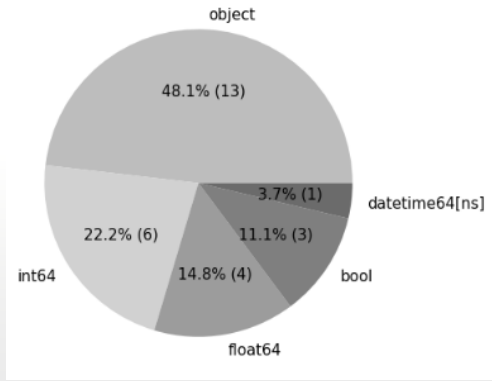
- Datos de partidos políticos y gobernador por Estado

https://www.openicpsr.org/openicpsr/project/102000/version/V3/view

CRITERIOS DE SELECCIÓN

- Trascendencia de la problemática
- Oportunidad y completitud de los datos

VARIABLES DEL DATASET



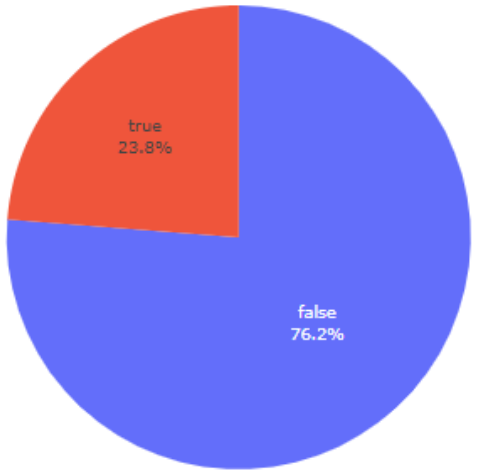
#	Nombre	#Datos	Tipo
0	Civil_Name	5793	object
1	Death_Date	5793	datetime64[ns]
2	Manner_of_death	5793	object
3	Armed	5793	object
4	Age	5793	int64
5	Gender	5793	object
6	Race	5793	object
7	City	5793	object
8	State_ID	5793	int64
9	Signs_of_mental_illness	5793	bool
10	Flee	5793	object
11	body_camera	5793	bool
12	longitude	5519	float64
13	latitude	5519	float64
14	is_geocoding_exact	5793	bool
15	Official_Disposition	5793	object
16	Encounter_Type	5793	object
17	Year	5793	int64
18	state_initial	5793	object
19	GDP_Millions	5793	int64
20	GDP_PerCapita	5793	int64
21	Unemployment_Rate	5793	float64
22	Poverty_Percent	5793	float64
23	Median_household_Income	5793	int64
24	Governor_Name	5793	object
25	Party	5793	object
26	State_Name	5793	object

DATA WRANGLING

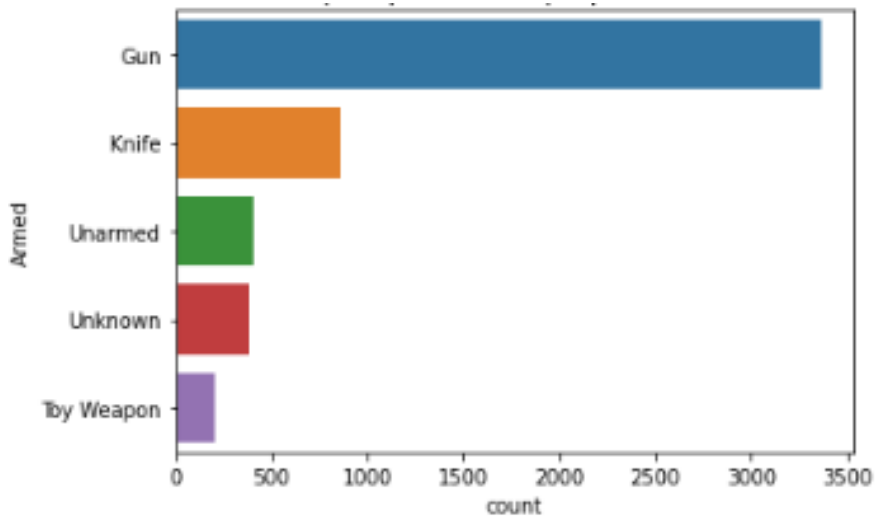
- Se considera que la completitud del dataset en cuanto a sus datos (5793 observaciones) es suficiente y sin errores significativos.
- Respecto a los datos faltantes de los campos “longitude” y “latitude”, los mismos se completaron a través de “fillna”, con la localización media de cada Estado, para que no se generen distorsiones significativas luego en visualizaciones espaciales.
- No se eliminó ningún campo, por considerarse todos relevantes, ya sea para el EDA o análisis descriptivo, tanto como para los modelos de predicción.
- A los efectos del modelado, se transformaron a “dummies” todos los campos de tipo “object”.
- Asimismo, no se consideraron para los modelos las columnas de nombre de la víctima, fecha de muerte y variables relacionadas a la locación, por presenter escasa o nula correlación.

ANALISIS UNIVARIADO

Proporción de civiles con y sin indicios de enfermedad mental



Muertes por tipo de arma que portaba la víctima



- Según estudios de la NAMI (National Alliance on Mental Illness) en EE.UU. 1 de cada 5 adultos sufre de una enfermedad mental, y 1 de cada 20 sufre una enfermedad mental grave.

- Hay un claro exceso de brutalidad policial en este caso comparando los números, ya que los casos más graves de enfermedad son 5,6% de la población, comparado al 23.8% que se observa que murió a manos de la policía.

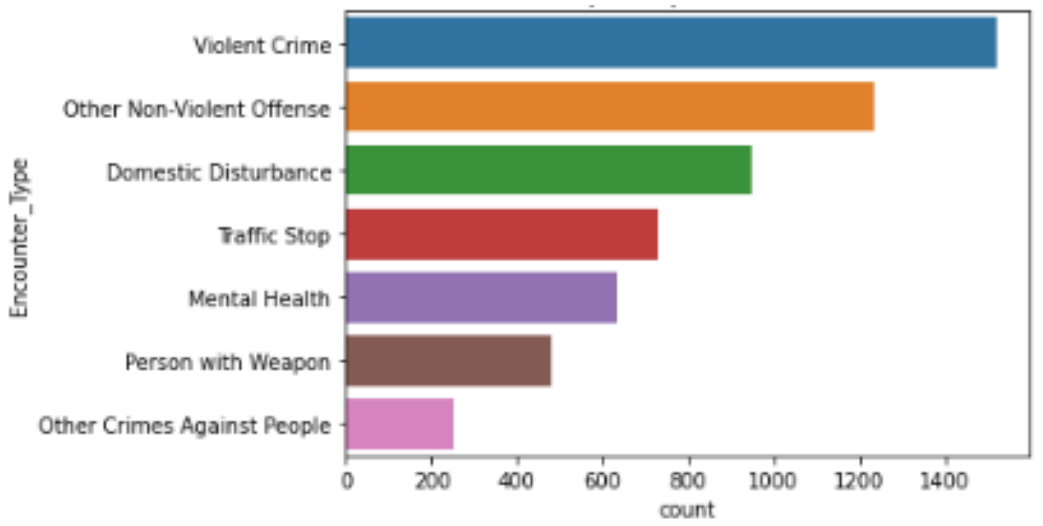
- En cuanto a nivel de amenaza de la víctima civil abatida, podemos observer que al menos en 500 casos se poseía arma de juguete o directamente se encontraba desarmada.

ANALISIS UNIVARIADO

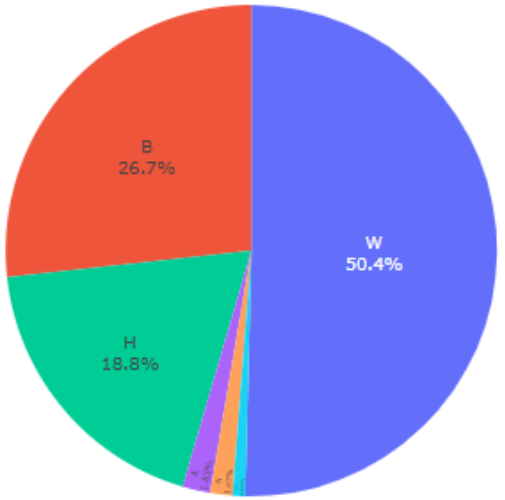
- En cuanto a tipo de incidente, puede observarse en el gráfico de la derecha que los casos de enfermedad mental sumados a los de disturbios domésticos o crímenes no violentos superan ampliamente a los crímenes violentos o civiles armados, lo que representaría, a priori un exceso policíaco el haber resultado muerto el civil involucrado.

- En cuanto al análisis por razas de las víctimas, según un studio de CNN, la etnia "blanca" promedio en EEUU representó en 2020 57,3%, teniendo 50,4% de víctimas civiles por parte de la policía, evidenciando una sub-representación. Por el contrario, la etnia "Negra", con una población promedio del 11,9%, presenta 27,5% de víctimas civiles, en una sobre-representación.

Cantidad de muertes civiles por tipo de encuentro con la policía

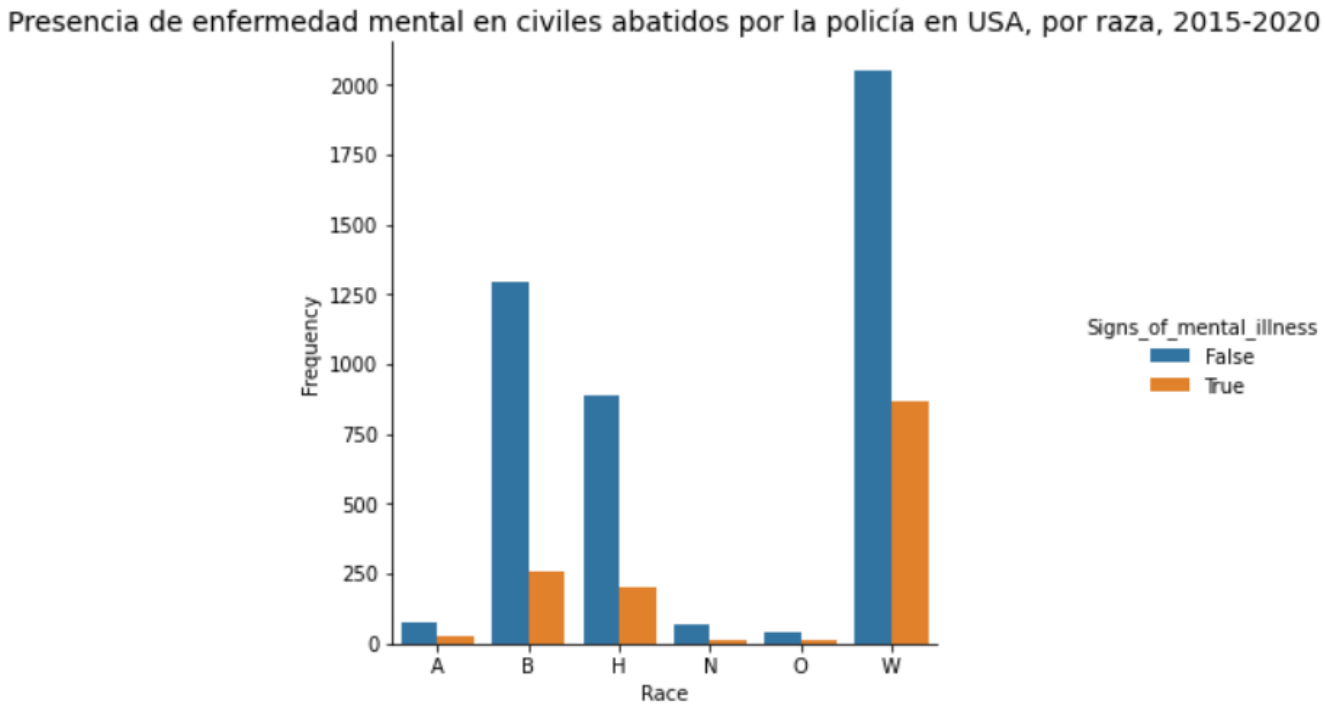


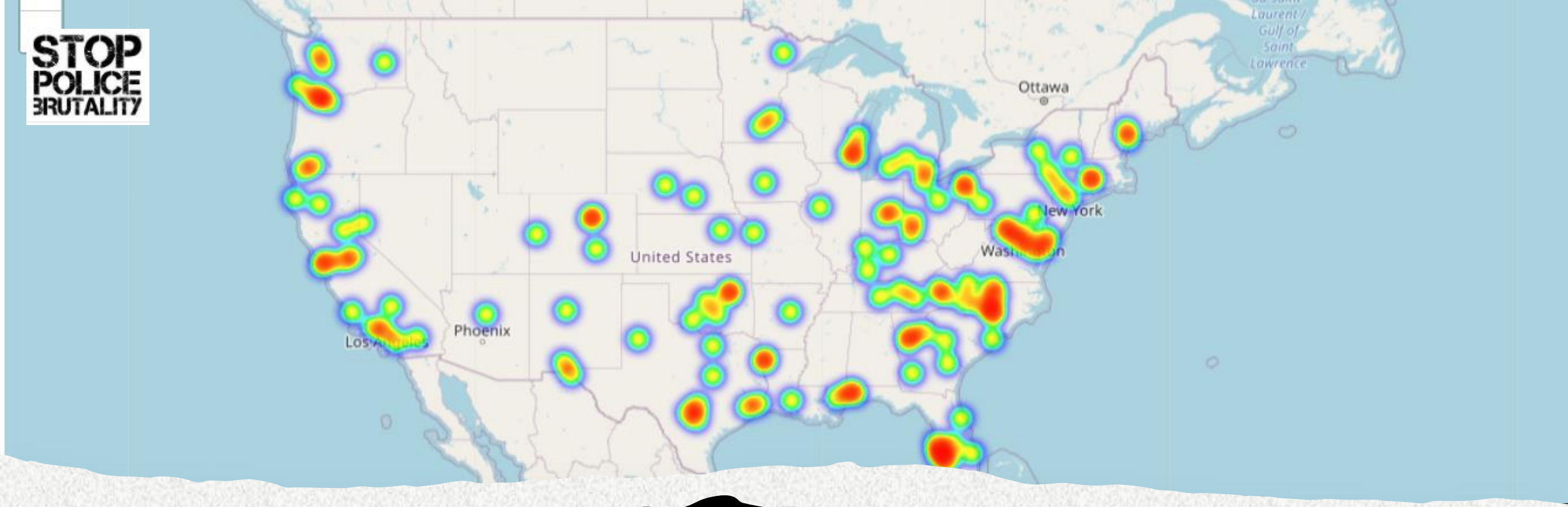
Proporción de muertes por etnia



ANALISIS BIVARIADO

- Comparando la cantidad de civiles muertos por raza y signo de enfermedad mental, podemos observar que la proporción de civiles con signos de enfermedad mental al momento de ser abatidos por la policía es mucho mas significativa para la raza blanca y asiática, con casi la mitad de civiles con dichas características.



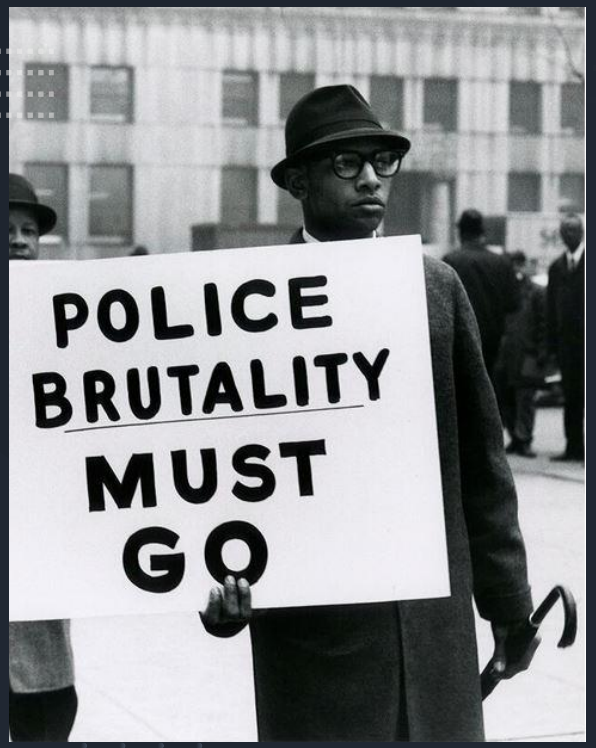


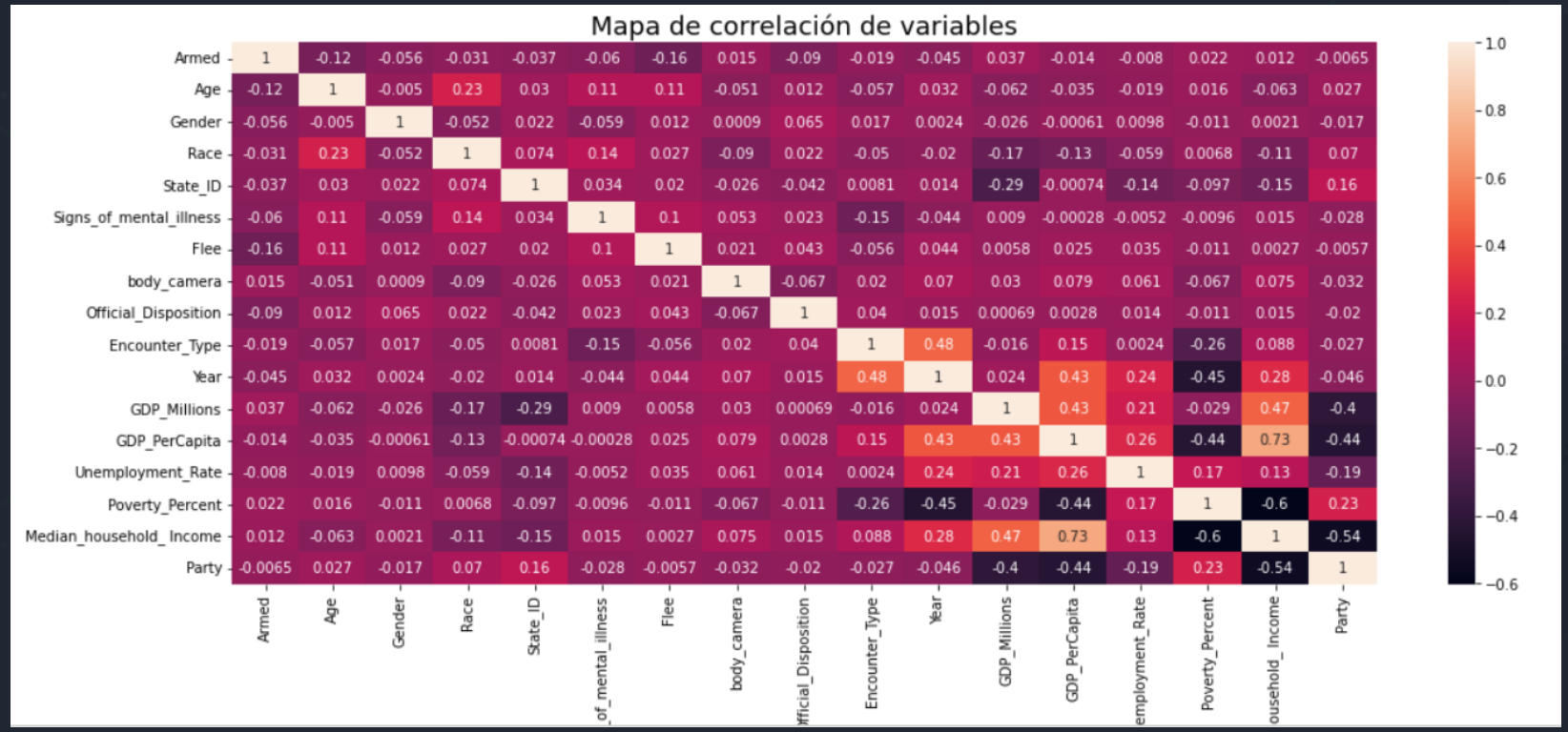
ANÁLISIS MULTIVARIADO

- **En la comparativa de muertes por millón por estado, pareciera estar más ligada a cuestiones demográficas de aglomeraciones urbanas, más que por condiciones socioeconómicas, dado que por mas que en las ciudades rurales del interior se da menor nivel económico, son las ciudades costeras de mayor poder adquisitivo las que tienen mayor incidencia de fallecimientos.**

ANALISIS MULTIVARIADO

- En cuanto a la correlación de la variable “sings_of_mental_illness”, puede observarse que no tiene una fuerte relación, ya sea directa o inversa con ninguna de las demás variables, a diferencia de las variables socioeconómicas, que si demuestran mayor correlación entre ellas.





ALGORITMOS ELEGIDOS

DECISION TREE

- ❑ Max Depth: 20
- ❑ Test = 0.3 / Train = 0.7
- ❑ Random State = 42
- ❑ Criterion = ‘Gini’

XGBOOST CLASSIFIER

- ❑ Learning Rate = 0.01
- ❑ Test = 0.3 / Train = 0.7
- ❑ Random State = 42
- ❑ n_estimators= 20
- ❑ Seed = 42

ALGORITMO: DECISION TREE

	precision	recall	f1-score	support
0	0.79	0.82	0.81	1316
1	0.38	0.34	0.36	422
accuracy			0.70	1738
macro avg	0.58	0.58	0.58	1738
weighted avg	0.69	0.70	0.70	1738

- Como puede observarse, el modelo presenta una significativamente mayor precisión para detectar casos “False” en cuanto a la variable de enfermedad mental de las victimas. Partiendo de métricas de infalibilidad (F1) de 0.81 para sin enfermedad mental y 0.36 con enfermedad mental. Las métricas de precisión y recall, en el mismo sentido, otorgan mayor predictibilidad cuando la variable en cuestión es false. De hecho, se dan mejores resultados para “false” partiendo de un caso real (0.82) que deduciendo a partir de las variables independientes (0.79). Para casos “true”, se da a la inversa.

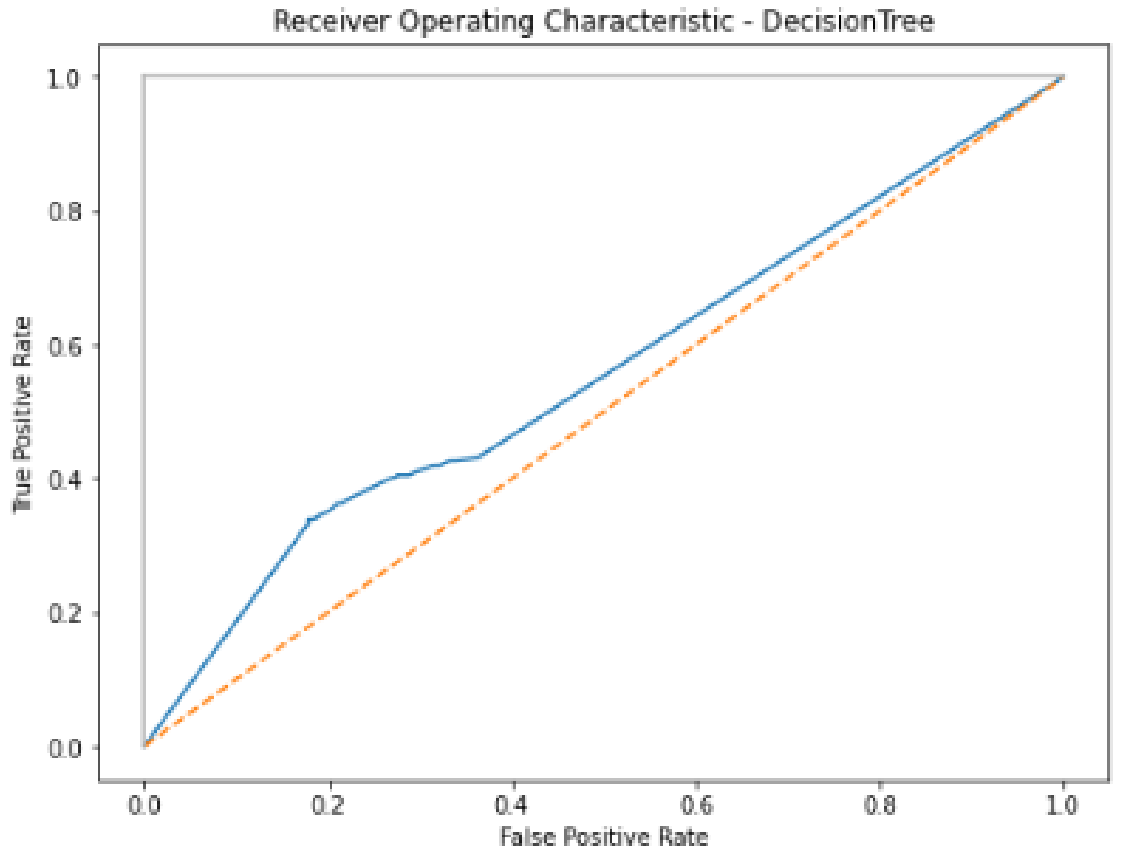
- En cuanto al accuracy, la precisión del modelo es del 70%.

- Esto pudiera significar que el modelo puede ser efectivo para determinar, una vez que haya fallecido un civil, si no se conoce el dato concreto, las probabilidades de que no haya sufrido una enfermedad mental en el momento de ser abatido por la policía, pero no para poder predecir precondiciones de enfermedades mentales a los efectos de diseñar políticas públicas.

ALGORITMO:

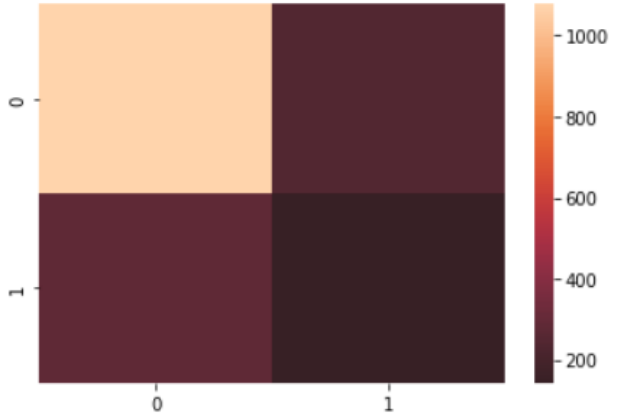
DECISION TREE – CURVA ROC

- **Calculando la Curva ROC y su AUC score, podemos determinar que el modelo tiene un 55% de probabilidad de que distinga entre presencia o ausencia de enfermedad mental de las víctimas civiles.**



ALGORITMO: DECISION TREE - AJUSTE 10 VARIABLES

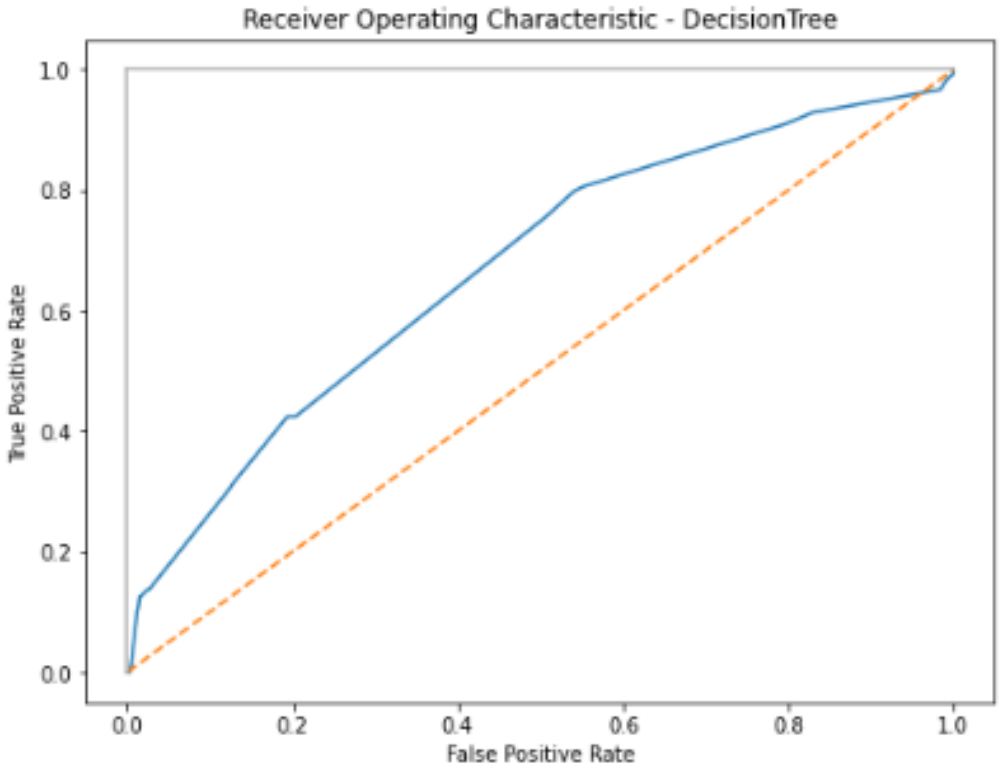
- Al ajustarse por las 10 principales variables en preponderancia, se puede obtener un “accuracy” mayor, del 77%, así como también una mejoría en la precisión de F1 score para los casos negativos (0.87), sin embargo, empeoran aún más la precisión para detectar casos verdaderos de presencia de enfermedad mental, reforzando lo mencionado anteriormente.



Age	0.214327
Year	0.086769
Unemployment_Rate	0.070584
GDP_Millions	0.069524
Poverty_Percent	0.055770
Flee_Not fleeing	0.053569
GDP_PerCapita	0.050817
Median_household_ Income	0.044086
State_ID	0.041067
Encounter_Type_Mental Health	0.037681

	precision	recall	f1-score	support
0	0.78	0.97	0.87	1316
1	0.63	0.14	0.23	422
accuracy			0.77	1738
macro avg	0.70	0.56	0.55	1738
weighted avg	0.74	0.77	0.71	1738

STOP POLICE BRUTALITY

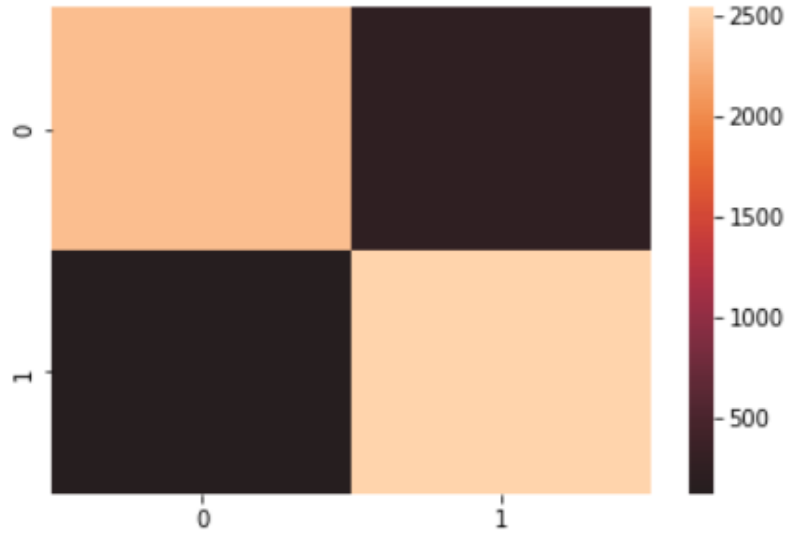


ALGORITMO: DECISION TREE – CURVA ROC - AJUSTADO

- **Calculando la Curva ROC y su AUC score, podemos determinar que el modelo tiene un 67% de probabilidad de que distinga entre presencia o ausencia de enfermedad mental de las víctimas civiles.**

ALGORITMO: XGBOOST CLASSIFIER

- Como puede observarse, las métricas del XBOOSTER presentan ratios más equilibrados, entre la predicción de positivos y negativos en cuanto a presencia de indicadores de enfermedad mental en víctimas civiles abatidas por la policía. Sobre todo, cabe destacar que la métrica de detección de casos positivos (recall = 0.95) es significativamente superior a la obtenida por el algoritmo decision tree, inclusive con parámetros ajustados de este último.



	precision	recall	f1-score	support
False	0.95	0.90	0.92	2634
True	0.91	0.95	0.93	2666
accuracy			0.93	5300
macro avg	0.93	0.93	0.93	5300
weighted avg	0.93	0.93	0.93	5300

MÉTRICAS FINALES DEL MODELO OPTIMIZADO

- En cuanto a la comparativa final de los dos modelos desarrollados, notamos significativamente más preciso el XGBOOST para predecir si un civil abatido por la policía presenta indicios de enfermedad mental en EEUU. Solamente se observa como más preciso al Decision tree en el indicador de "precisión", pero haciendo referencia a la positividad de predicción de casos negativos, cuando en realidad la motivación primigenia del análisis implica detectar positivos verdaderos para poder tomar acciones en formas de políticas públicas y prevenir la ocurrencia de estas muertes evitables.

Métricas DecisionTree:

- Accuracy: 0.7710
- Precisión: 0.9734
- Sensibilidad: 0.7792
- Especificidad: 0.6277
- F1 score: 0.8655

Métricas XGBoost con los mejores Hiperparámetros:

- Accuracy: 0.9226
- Precisión: 0.9055
- Sensibilidad: 0.9368
- Especificidad: 0.9096
- F1 score: 0.9208

FUTURAS LÍNEAS

Como posibilidad de mejora creemos que ahondar en mayor cantidad de modelos pudiera abordar a conclusions aún mas precisas, así como también incluir datos de otros años o de casos donde el encuentro entre el civil y la policía no terminó con la muerte del civil, ampliaría la posibilidad de identificar las situaciones y las características donde se la vida humana es salvada.