ESCUELA DE INGENIERÍA Y CIENCIAS
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Predicción de un tiroteo en el estado de Texas el próximo mes

Integrantes: Vanessa González, José Miguel Yuseff
Profesores: Carlos Flores G, Francisco Santibáñez P.
Fecha de entrega: Sábado 15 de Agosto 2020

Índice

Descripción del problema

Un tiroteo masivo se entiende como un episodio en el que mueren cuatro o más víctimas a manos de un individuo con un arma de fuego. Estos episodios no incluyen matanzas entre pandillas o asesinatos que involucren a varios miembros de una misma familia.

De acuerdo a diversos estudios, Estados Unidos corresponde al país en que se han registrado más tiroteos masivos que en cualquier otro país del mundo, teniendo el 31% de los tiroteos ocurridos entre 1966 y 2012.

El estado con más tiroteos en EEUU es California, donde se produjo el ataque más mortífero en la historia del país, cuando 59 personas murieron en un festival de Las Vegas. Luego de California, en el tercer lugar, se encuentra Texas. Este estado ha tenido cuatro de los tiroteos más letales en los últimos 20 años, entre los cuales se encuentra el tiroteo de El Paso ocurrido en 2019, que fue un ataque terrorista a manos de un joven supremacista blanco, quien asesinó a 23 personas.

Bajo este contexto y problemática, lo que queremos predecir a partir de los datos es si ocurrirá un tiroteo en el estado de Texas durante el próximo mes.

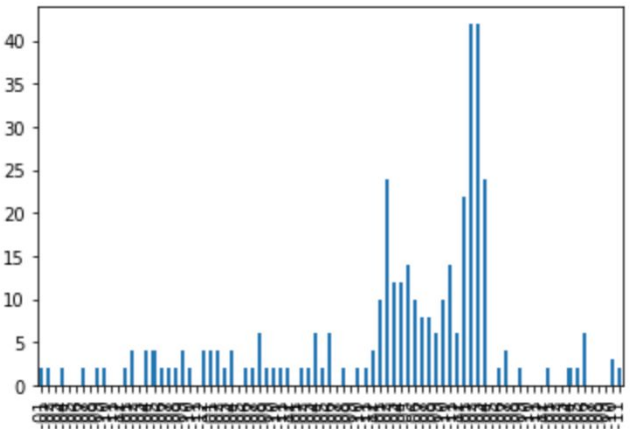
Análisis de Datos

Los datos iniciales son los tiroteos en los últimos 50 años (1966 hasta 2017), acompañados de la fecha en que ocurrió cada uno y otros datos como: Localización, Edad, Sexo, Raza, Muertos, Heridos.

Se analizaron únicamente los datos referidos a *Fecha, Muertos, Heridos, Raza* y *Sexo*, ya que estos son a nuestro parecer los que podrían tener un mayor impacto en la predicción de un tiroteo.

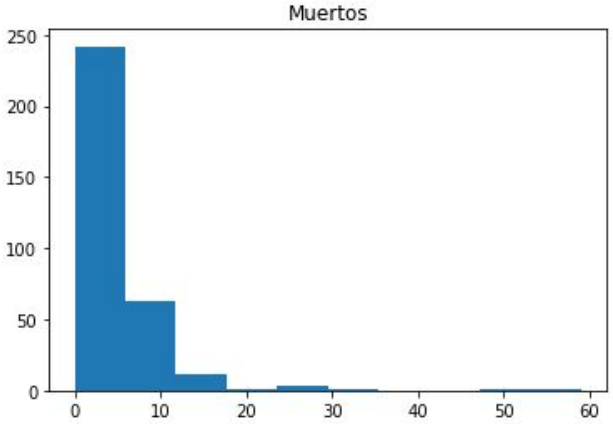
1) Fecha:

   El gráfico realizado para analizar las fechas en que ocurrieron los tiroteos de la base de datos corresponde a la cantidad de tiroteos (eje y) por mes (eje x). De acuerdo a dicho gráfico, existen meses en que no ocurre ningún tiroteo, por tanto es factible tratar de predecir si ocurrirá un tiroteo o no en un estado particular el próximo mes.

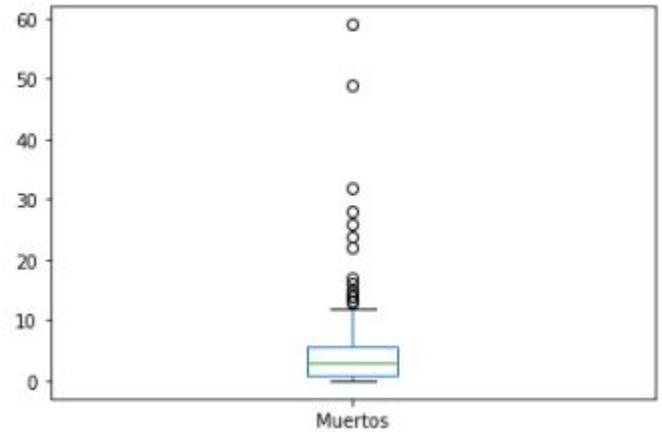


2) Muertos:

   El gráfico realizado corresponde a la la cantidad de muertos (eje x) y su frecuencia (eje y). Se puede observar del gráfico que en más del 75% de los tiroteos mueren menos de 6 personas, con un promedio de 4 personas por tiroteo. Existen algunos datos extremos en los sectores de 30 a 35 y 50 a 60 muertos que se mantuvieron para la predicción ya que son valores que sí podrían ocurrir en un tiroteo.

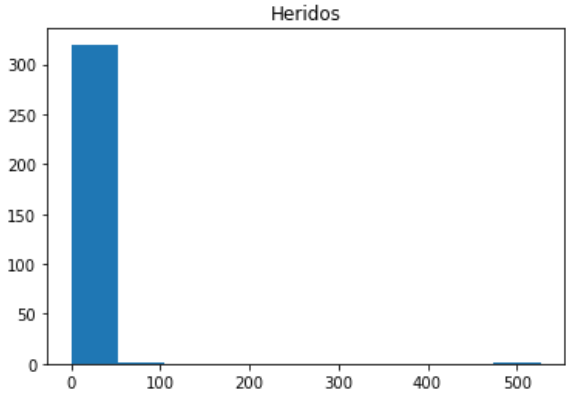


count	323.000000
mean	4.436533
std	5.783208
min	0.000000
25%	1.000000
50%	3.000000
75%	5.500000
max	59.000000

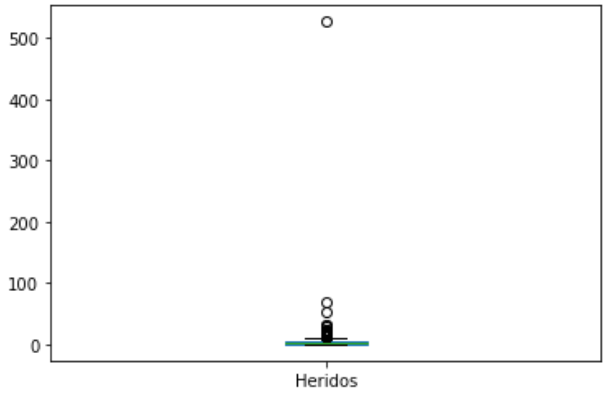


3) Heridos:

El gráfico realizado corresponde a la cantidad de heridos (eje x) y su frecuencia (eje y). Se puede observar del gráfico que en más del 75% de los tiroteos son heridas menos de 5 personas, con un promedio de 6 personas por tiroteo. Existe un único dato extremo alrededor de los 520 heridos, que al igual que antes se mantuvo por la misma razón.



count	323.000000
mean	6.176471
std	29.889182
min	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
max	527.000000

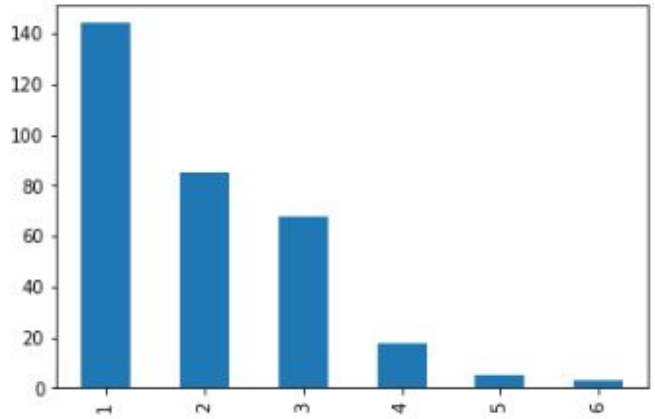


4) Raza:

Para analizar la raza de los tiradores y luego añadirla en la matriz de características, se le otorgó un número a cada una de las razas presentes en la base de datos. La asignación fue la siguiente:

1: *Blanco o Caucásico*
2: *Negro o Afroamericano*
3: *Otro*
4: *Asiático*
5: *Latino*
6: *Nativo Americano*

Se puede observar a partir del gráfico que la raza caucásica es la predominante cuando ocurre un tiroteo, pues constituye el 44.5% de los tiradores. Esta raza es seguida por los afroamericanos, con un 26.3%. Por tanto, ambas razas constituyen alrededor del 70%.



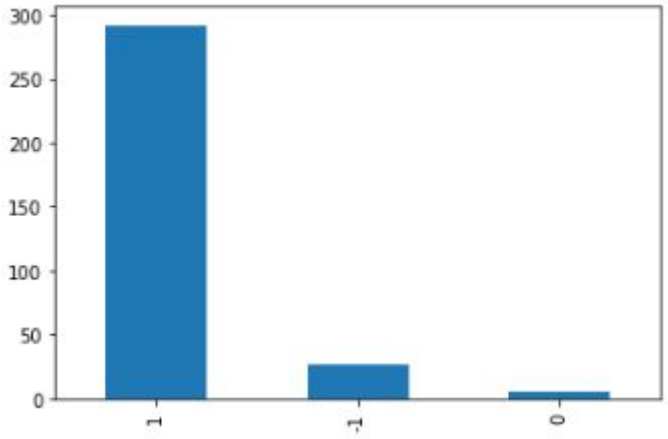

```
1    144
2     85
3     68
4     18
5      5
6      3
Name: Raza, dtype: int64
```

5) Sexo

Al igual que con la variable anterior, para analizar el sexo de los tiradores y añadirlo en la matriz de características, se asignó un número a cada sexo de la siguiente forma:

0: *Mujer*
1: *Hombre*
-1: *Otro/desconocido*

Luego, a partir del gráfico y el conteo de datos, se observa que los hombres son quienes usualmente comienzan un tiroteo, pues constituyen el alrededor del 90% de los tiradores.




```
 1    292
-1     26
 0      5
Name: Sexo, dtype: int64
```

Matriz de Características

De las variables que teníamos inicialmente, descartamos las siguientes:

- Resumen
- Empleado
- Latitud
- Longitud
- Empleado en
- Área de incidente
- Espacio abierto o cerrado
- Objetivo
- Causa
- Problemas mentales
- Policías muertos
- Edad

Dejando solo las variables *Lugar, Fecha, Muertos, Heridos, Víctimas totales, Raza* y *Sexo*. Cabe señalar que la variable *Lugar* en cada caso fue reemplazada solo por el estado al que correspondía la ubicación, sin considerar ciudades y/o poblaciones.

Algunas variables, como *Empleado en* y *Latitud*, fueron descartadas por no parecernos útiles. Por otro lado, variables como *Edad* y *Problemas mentales*, fueron descartadas porque una gran cantidad de tiroteos no presentaba dicho dato, y además, los datos presentes no mostraron ninguna tendencia. Y por último, variables como *Área de incidente* y *Objetivo*, se descartaron por presentar demasiados datos diferentes entre sí, que no entregaban información como conjunto.

A partir de las variables que se mantuvieron, se decidió crear un conjunto de variables nuevas, presentadas en la siguiente tabla:

N_tiroteo_ultimo_mes	Número de tiroteos ocurridos en EEUU durante el último mes
N_tiroteo_ultimos_6mes	Número de tiroteos ocurridos en EEUU durante los últimos 6 meses
N_muertos_ultimo_mes	Promedio del número de muertos por tiroteo en EEUU durante el último mes
N_muertos_ultimos_6mes	Promedio del número de muertos por tiroteo en EEUU durante los últimos 6 meses
N_heridos_ultimo_mes	Promedio del número de heridos por tiroteo en EEUU durante el último mes

N_heridos_ultimos_6mes	Promedio del número de heridos por tiroteo en EEUU durante los últimos 6 meses
raza_pred_ultimos_12meses	Raza de tiradores que más se repite en los últimos 12 meses
sexo_pred_ultimos_12meses	Sexo de tiradores que más se repite en los últimos 12 meses

Cada una de las variables nuevas considera un lapso de tiempo previo al mes en que se quiere predecir si ocurrirá un tiroteo en Texas o no, pues de esta forma se toman en cuenta los efectos de lo ocurrido en dichos lapsos de tiempo y las tendencias que presentan algunos datos.

En los casos de *raza_pred_ultimos_12meses* y *sexo_pred_ultimos_12meses*, se considera la tendencia de ambos datos durante el último año, es decir, cuáles son la raza y el sexo que más se repiten.

Por otro lado, en los casos de *N_tiroteo_ultimo_mes, N_muertos_ultimo_mes* y *N_heridos_ultimo_mes,* lo que se considera es la gravedad de los tiroteos ocurridos en el último mes, y por tanto el impacto que esto genera en las autoridades y las medidas que estas aplican. Por ejemplo, si los números de estas tres variables son muy altos, se podría asumir que el mes siguiente se tomarán mayores resguardos en lugares públicos con el fin de prevenir los tiroteos, lo cual reduce la posibilidad de que estos ocurran. De la misma manera, si los números fueran muy bajos se esperaría que durante el mes siguiente no existan mayores resguardos.

Asimismo, en los casos de *N_tiroteo_ultimos_6mes, N_muertos_ultimos_6mes* y *N_heridos_ultimos_6mes*, se toma en cuenta la gravedad de los tiroteos ocurridos en los últimos 6 meses y el impacto de estos en las medidas tomadas por las autoridades. Sin embargo, esta vez se considera la toma de medidas de resguardo a largo plazo que puedan existir, a diferencia de los casos anteriores, donde se considera solo el corto plazo.

A continuación, veamos un ejemplo de la matriz obtenida, donde se presentan solo las fechas de los cinco tiroteos más recientes.

	N_tiroteo_ultimo_mes	N_tiroteo_ultimos_6mes	N_muertos_ultimo_mes	N_muertos_ultimos_6mes	N_heridos_ultimo_mes	N_heridos_ultimos_6mes	raza_pred_ultimos_12meses	sexo_pred_ultimos_12meses
2017-11-05	3.0	8.0	10.67	13.13	7.67	69.0	1.0	1.0
2017-11-01	2.0	7.0	3.0	11.29	1.5	76.0	1.0	1.0
2017-10-18	2.0	6.0	31.0	12.67	265.0	88.67	1.0	1.0
2017-10-01	1.0	6.0	59.0	12.67	527.0	88.17	1.0	1.0
2017-06-14	3.0	6.0	3.67	3.67	0.67	1.33	2.0	1.0

Por último, cabe señalar que al momento de crear nuevas variables se quiso agregar la regulación de armas, los porcentajes de hombres y mujeres, y los porcentajes de cada raza por estado. Sin embargo, esto no fue posible debido a que no se encontraron fuentes confiables de dicha información. Si por el contrario, se hubiese encontrado la información, la predicción mejoraría.

Modelos predictivos generados

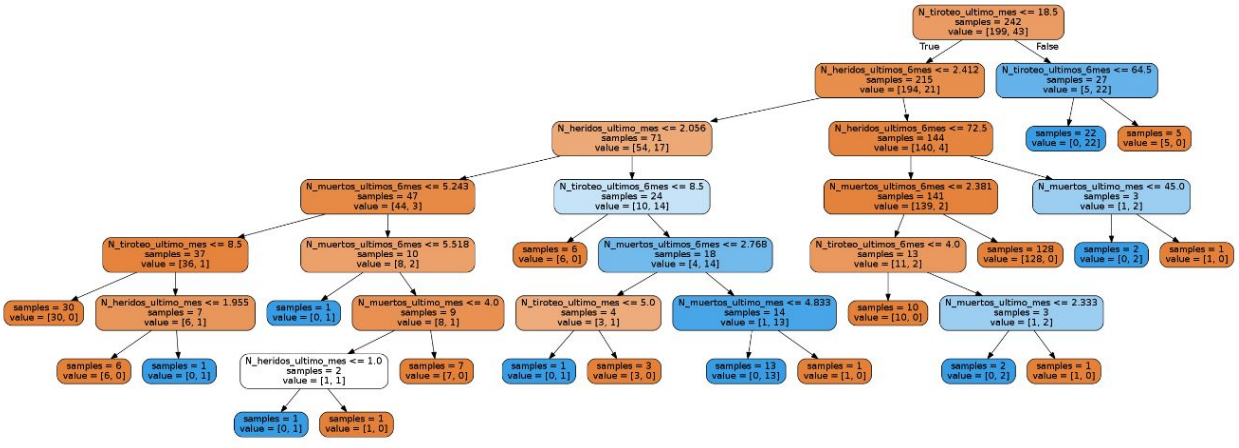
Para poder realizar la predicción deseada, se implementaron dos modelos predictivos de aprendizaje supervisado: Árbol de decisiones y Random Forest.

El Árbol de decisiones es un algoritmo de aprendizaje supervisado que se puede utilizar para problemas de clasificación y regresión, el cual consiste en una serie de decisiones secuenciales que se toman para llegar a una decisión final. Este algoritmo evalúa todas las variables de entrada, que en este caso corresponden a las características de la matriz presentada anteriormente.

Por su parte, el modelo Random Forest funciona con un conjunto de árboles de decisión, donde cada uno de estos realiza una predicción. Luego, el modelo presenta como predicción final la que más se repita en dichos árboles.

A continuación, se utilizará el modelo de árbol de decisiones para explorar en mayor detalle la exactitud del modelo al momento de predecir si habrá un tiroteo masivo en Texas el próximo mes.

A partir del modelo de árbol de decisiones, utilizando el 75% de los datos para entrenar dicho modelo, y el 25% para probarlo, se obtuvo el siguiente árbol :



De este árbol se desprendieron los siguientes indicadores:

Correctitud	Sensibilidad	Especificidad	Precisión	Tasa Real	F1 Score
79.01%	50.0%	89.83%	64.71%	27.16%	56.41%

A primera vista se pueden ver buenos indicadores de correctitud y especificidad, lo que indicaría que el modelo posee una buena base para predecir un tiroteo. Sin embargo, esto no es del todo preciso, ya que la tasa real es de apenas un 27%, lo que implicaría que por lo general no ocurren tiroteos en Texas. Esto afecta la viabilidad del modelo, pues a partir de los porcentajes de correctitud y especificidad, este logra acertar seguidamente cuando no va a ver un tiroteo en Texas, lo cual no es nuestro objetivo. Esto último se puede deducir entendiendo que la correctitud es el porcentaje de acierto del modelo sobre las veces que el modelo predijo correctamente cuando habría y no habría un tiroteo (los verdaderos positivos y los verdaderos negativos), mientras que la especificidad es el acierto solo cuando no habría un tiroteo (verdaderos negativos).

Por su parte, la sensibilidad puede darnos un verdadero acercamiento a si nuestro modelo es viable, ya que muestra el porcentaje de acierto sobre el éxito, es decir, cuántas veces el modelo acertó que ocurriría un tiroteo en Texas sobre las veces que sí hubieron tiroteos. Por tanto, que la sensibilidad sea de un 50% indica que nuestro modelo solo pudo predecir correctamente la mitad de las veces en las que ocurrió un tiroteo en ese estado.

Por otro lado, la precisión muestra que el modelo acertó un 64% de la veces en que predijo que sí habría un tiroteo, lo cual en función de la base de datos es un buen porcentaje, o bien, nos indica que el modelo podría ser útil.

Para validar mejor los indicadores obtenidos se realizó una validación cruzada de los datos, en la que se separó los datos en 10 segmentos distintos, cambiando los sets de entrenamiento y obteniendo un indicador global para los distintos modelos generados. De esta manera, se logra descartar la posibilidad de que una ordenación precisa de los datos haya alterado los porcentajes obtenidos.

Para estos 10 modelos se midió el F1 Score, ya que dicho valor es una combinación entre la sensibilidad y precisión, y por tanto que nos dice qué tan exacto fue nuestro modelo en general.

El promedio obtenido para los 10 modelos fue el siguiente:

F1 Score promedio (Decision Tree)
42.6 %

Este F1 Score global es menor al obtenido anteriormente, por lo tanto, nuestro modelo es un poco menos exacto de lo que creímos.

Luego, se comparó este F1 Score global con el que se obtendría usando un modelo predictivo de Random Forest, para así observar qué modelo predictivo es mejor.

Nuevamente, realizando una validación cruzada con 10 segmentos, se obtuvo el siguiente F1 Score:

F1 Score promedio (Random Forest)
41.9%

Este F1 Score es menor al obtenido con el árbol de decisiones. Lo que indicaría que el modelo predictivo de árbol de decisión es un mejor modelo para la predicción de nuestro problema. Esto debido a que el F1 Score mide tanto la sensibilidad como la precisión, por lo cual a mayor F1 Score se tendrá un modelo más certero y viable a la hora de predecir cuando va a ocurrir un tiroteo, cosa que no ocurre necesariamente con los demás parámetros, ya que estos últimos también consideran la predicción de cuando no ocurre un tiroteo.

Conclusión

Como se dijo en la sección de métodos predictivos generados, el F1 Score fue más alto al aplicar el Árbol de decisiones, aunque no de manera significativa. Por lo tanto, dicho método es levemente mejor que Random Forest en este caso.

Este resultado no es lo que esperaríamos concluir, pues entendemos que Random Forest es un modelo más robusto que el Árbol de decisiones, por lo cual debiera ser más preciso. Esto debido a que el Árbol de decisiones tiende a “sobre-ajustar”, es decir, tiende a aprender muy bien los datos de entrenamiento, pero su generalización no es tan buena. En cambio, cuando se utiliza Random Forest, distintos árboles ven distintas porciones de datos. De esta forma, ningún árbol ve todos los datos de entrenamiento y cada árbol se entrena con distintas muestras de datos para un mismo problema. Así, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

A pesar de lo mencionado anteriormente respecto al Árbol de decisiones, este presenta una gran ventaja: es mucho más fácil de visualizar y, por medio de la visualización, se hace más simple entender el algoritmo que utiliza.

En ambos modelos, para mejorar la predicción convendría aumentar el número de datos o tiroteos en la lista, y también el número de características consideradas. Por ejemplo, se podrían agregar las características que quisimos agregar: regulación de armas por estado, porcentaje de hombres y mujeres en cada estado,y porcentaje de cada raza por estado. De esta forma, todos los indicadores del modelo aumentarían, es decir, serían más exactos.

Asimismo, para el Árbol de decisiones convendría separar por temporalidad los datos con que se entrena y prueba el modelo. Sabemos que el modelo se entrena con el 75% de los datos y se prueba con el 25%, siendo este último un conjunto de datos aleatorio. Sin embargo, si ese 25% correspondiera a los últimos datos la predicción sería mejor ya que se estarían usando los tiroteos pasados para predecir uno futuro, lo que haría del modelo predictivo más fiable a la hora de querer predecir un tiroteo que aún no ha ocurrido.

Si estos modelos fueran aplicados en la vida real, considerando tanto las mejoras mencionadas como otras que puedan existir, serían muy útiles en Texas, Estados Unidos. Utilizando el modelo predictivo, las autoridades podrían anticiparse a los tiroteos que ocurrirán en dicho estado y tomar medidas de resguardo, y más aún, podrían salvar vidas.

Asimismo, el modelo podría ser aplicado en otros estados de EEUU, como California, donde las cifras de tiroteos masivos son excepcionales y las vidas que han cobrado estos atentados son muy numerosas.