UNIVERSIDAD TECNOLÓGICA METROPOLITANA
del Estado de Chile

Escuela de Informática
Facultad de Ingeniería
Minería de datos INFB8104

“Modelo Predictivo de Crimen con Arma de Fuego en EE.UU.”

Proyecto Semestral

Integrantes: Shu-yi Wong Baxter
Sebastian Perez Berrios
Mathias Cáceres Bustamante

Profesor(a): Pablo Figueroa Plaza

Fecha de entrega: 23-07-2021

Índice de Contenidos

1. Introducción

Estados Unidos concentra una de las mayores tasas de posesión de armas, puesto que las barreras para adquirir estas son bajas. Para corroborar lo anterior, se puede mencionar que: dependiendo del estado, la edad mínima para adquirirlas se encuentra en un rango de 18 a 21 años, además con los recientes hechos de inestabilidad en EE.UU. las ventas de armas y munición se han disparado desde que comenzó la pandemia del Covid-19. Por otra parte, es común ofrecer descuentos en restaurantes, sólo por mostrar las armas que se posean. Es por ello, que ante la creciente adquisición centraremos nuestro análisis, en las cifras que guardan relación con el crimen a mano armada.

Por otro parte, un modelo predictivo se define como "*un sistema que emplea datos y estadísticas para predecir resultados, a partir de unos modelos de datos*". Dichos modelos pueden ser usados en diversas áreas ya sea para predecir resultados deportivos, flujos empresariales, avances tecnológicos, entre otros. La finalidad que brindan los modelos predictivos, radica en que proporciona información precisa para responder cualquier tipo de pregunta y otorga a los usuarios la oportunidad de prevenir o anticiparse a un hecho. Este último punto, cobra mucha relevancia cuando se trata a nivel empresarial para obtener ventajas comparativas tras obtener información detallada. Para aplicar un análisis de minería de datos es común extraer información esencial de diversas fuentes para ir confeccionando los modelos predictivos donde se destaca las siguiente información claves: datos demográficos, datos económicos, datos sobre transacciones, datos geográficos, datos transaccionales, datos sobre encuestas, entre otros.

Para lograr predecir datos, se debe previamente realizar un análisis predictivo de estos, es por ello que se hace necesario describir cuales son los datos utilizados, posteriormente realizar un diagnósticos para comprender estadísticamente las relaciones existentes entre estos. Es por ello, que a lo largo de esta contribución se requerirá el uso de la minería de datos, un análisis exhaustivo y una comprensión de la información para predecir finalmente "*cómo distribuir recursos de emergencia para atender muertos y lesionados producto de un crimen con arma de fuego en E.E.U.U*". Cabe destacar, que dependerá de varias características, por lo que predecir conlleva a tomar decisiones.

A lo largo de esta contribución, se darán a conocer diversos modelos predictivos que son utilizados en la actualidad, estos se diferencian entre paramétricos y no paramétricos. La diferencia principal radica en que el primero hace más suposiciones específicas, sobre la población utilizada para confeccionar el modelo. Dentro de estos modelos predictivos se

pueden destacar los siguientes: Regresión logística, Bosques aleatorios, Árboles de decisión, Modelos lineales generalizados, entre otros. Cada modelo tiene un uso en particular y responden a preguntas específicas utilizando determinados tipos de conjuntos de datos. A pesar de las diferencias matemáticas y metodológicas entre los diferentes modelos, el objetivo en común es predecir basándose en datos pasados y obteniendo resultados futuros. Finalmente, a lo largo de esta contribución se aplicará análisis PCA, aplicación de clustering árboles de decisión, redes bayesianas y máquinas de vector de soporte. Con la incorporación de estos modelos brindara tener distintas visiones y acercamiento de precisión para el objetivo general, el cual se abordará posteriormente.

2. Aspectos Generales del Proyecto

2.1. Delimitación del tema del proyecto

Se busca predecir cómo distribuir los distintos recursos de emergencia(policía, ambulancias), considerando las muertes y lesionados en los diversos estados del país Norteamericano, tras haber ocurrido un crimen a mano armado, esto en base a datos adquiridos desde una corporación sin fines de lucro formada en el 2013, para entregar toda la información correspondiente a la violencia con arma de fuego. Los datos obtenidos corresponden a incidentes con armas de fuego entre el año 2013 y 2018.[1]

2.2. Formulación de la problemática del proyecto

Para contextualizar, el dataset otorgado por kaggle considera los diversos crímenes cometidos acorde a las características que estaban envueltas. Dando como resultado un .CSV que registra principalmente variables cualitativas, en su mayoría datos tipos textos y además presenta redundancia de datos. Esto se debe principalmente, a que la información contiene delimitadores, es muy detallada y no es generalizada. Lo anterior, descrito se puede apreciar en la siguiente figura. Mediante Power BI se decidió solucionar este problema.

participant_age	participant_age_group	participant_gender
0::20	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+	0::Male||1::Male||3::Male||4::Female
0::20	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+	0::Male
0::25||1::31||2::33||3::34||4::33	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+	0::Male||1::Male||2::Male||3::Male||4::Male
0::29||1::33||2::56||3::33	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+	0::Female||1::Male||2::Male||3::Male
0::18||1::46||2::14||3::47	0::Adult 18+||1::Adult 18+||2::Teen 12-17||3::Adult 18+	0::Female||1::Male||2::Male||3::Female
0::23||1::23||2::33||3::55	0::Adult 18+||1::Adult 18+||2::Adult 18+||3::Adult 18+||4::Adult 18+...	0::Female||1::Female||2::Female||3::Female||4::Male||5::Male
0::51||1::40||2::9||3::5||4::2||5::15	0::Adult 18+||1::Adult 18+||2::Child 0-11||3::Child 0-11||4::Child 0-1...	0::Male||1::Female||2::Male||3::Female||4::Female||5::Male

Imagen 1: Datos cualitativos originales (Fuente: Elaboración Propia)

Los datos recopilados respecto a los crímenes son: identificación , fecha, estado, ciudad o comuna, dirección, número de muertes, números de lesionados, incidente de la url, origen del url, pérdida de url del incidente, distrito del congreso, armas robadas, tipos de armas, características del incidente, latitud, descripción de la ubicación, longitud, numeros de armas envueltas, notas, edad de los participantes, grupo de edad de los participantes, género de los participantes, nombre de los participantes, relación de los participantes, estado de los participantes, tipo de participantes, fuentes, distrito de la casa del estado, distrito del senado estatal.

Resumen del CSV: 29 columnas, 16 columnas de tipo texto, 6 columnas de tipo entero, 2 columnas de tipo url, 5 columnas de otros tipos de datos.

2.3. Objetivos

Objetivo General:

- Desarrollar un modelo predictivo para distribuir ambulancias y policías, cuando ocurra un crimen a mano armada, en los diversos estados de EE.UU.

Objetivo Específico:

- Recopilación de información respecto a crímenes a manos armada en EE.UU.
- Preparación de los datos para generar modelos predictivos.
- Elaboración de modelos predictivos con las herramientas de Minería de datos.
- Comparar diferentes modelos predictivos y definir el más preciso.
- Basado en los resultados de los modelos, evaluar e interpretar resultados y contrastarlos con la hipótesis.

2.4. Hipótesis

¿Es posible predecir cómo distribuir los recursos ambulancias y policías, cuando ocurra un crimen a mano armada para los diversos estados de EE.UU ?

Con este modelo se busca distribuir de una mejor manera los recursos puesto que estos son limitados. Poniendo como ejemplo, una ambulancia o patrulla policial tienen dos llamadas simultáneas en estados diferentes, ¿Dónde deberían ir?. Es así, que con la predicción del modelo dará respuesta a tal pregunta. Además, se pueden atender lesionados de manera oportunidad, actuar rápido en caso de ser un crimen sea reciente e intentar atrapar a los sospechosos o el simple hecho de asumir la muerte.

2.5. Definición de los datos a utilizar

Toda la información utilizada para la confección del proyecto tuvo como base el dataset de Kaggle. El CSV contiene toda la información, solo que tras el uso de Power BI se procedió a categorizar los distintos valores, algunas columnas reducirlas, y otras columnas fueron creadas a partir de otras. A continuación se especificarán las columnas que contendrá este archivo el cual llamaremos "*Avance 5.5*":

- **date:** Este dato estaba en formato texto, cuyo formato era dd-mm-aaaa.
- **month:** De la columna "*Date*" fue posible obtener el "*Month*", lo cual requirió una conversión previa a valor numérico y además, extraerlo de la fecha. El conjunto abarca los 12 meses del año definidos como 1: January, 2: February,..., 12: December.
- **day:** De la columna "*Date*" fue posible obtener el "*Day*", lo cual requirió una conversión previa a valor numérico y además, extraerlo de la fecha. El conjunto abarca los 7 días de la semana definidos como 1: Monday, 2: Tuesday,...,7: Sunday.
- **state:** Columna tipo texto la cual fue convertida a numérica. El conjunto abarca los diferentes estados de EE.UU. definidos como 0: Albama, 1: Alaska,…, 49: Wyoming.
- **city or county:** Esta columna contiene diversa información de las ciudades y está en tipo texto.
- **address:** Esta columna contiene diversa información como direcciones especificas, se encuentra en formato texto.
- **n° killed:** Esta columna contiene valores numéricos, corresponde al número de muertos en el incidente.
- **n° lesionado:** Esta columna contiene valores numéricos, corresponde al número de lesionados en el incidente.
- **congressional district:** Esta columna contiene valores numéricos, específica en que distrito se encuentra el incidente.
- **gun stolen:** Dicha columna contenía redundancia, delimitadores y además contenía muchos detalles. Es por ello que se decidió trabajar con 3 conjuntos una vez resuelto lo anterior descrito. El conjunto abarca de 0: Not-stolen, 1: Stolen,…, 2: Unknown. En otras palabras, se generaliza por la mayor concordancia de los datos en caso de no ser robada el arma el valor es cero, si hay armas robadas el valor es 1 y en caso de no tener información el valor es 2.
- **gun type:** Dicha columna contenía redundancia, delimitadores y además contenía muchos detalles. Para especificar, cómo se agruparon todas aquellos modelos de armas ya sean 9mm, 40 S&W, Handgun, entre otros modelos fueron ajustadas a la categoría más indicada. El conjunto abarca de 0: Gal, 1: Handgun, 2: Shotgun, 3:

Rifle, 4: Other, 5: Unknown, 6: Conjunto de categorías. En otras palabras, cuando se encontraban diversos modelos de pistolas(siempre dependiendo del calibre), estas se categorizaban en un conjunto mayor. En caso de encontrar revolver(gal) el valor al que corresponde es cero, posteriormente si la categoría estaba asociado a pistola(Handgun) el valor es 1. Para el caso, que los modelos involucrados corresponden al conjunto de escopetas(Shotgun) el valor es 2. Enseguida, si existía un arma que se asocia a los Rifles el valor asignado es 3. Cuando existían armas que no podían ser categorizadas en las anteriores descritas se clasificaban en otras cuyo valor es 4. Por otra parte, si no se otorgaba información se clasificaba en desconocido(Unknown) cuyo valor es 5. Finalmente, cuando los crímenes involucran diversos tipos de armas y abarcan dos categorías el valor es 6.

- **incident characteristics:** Esta columna contiene diversa información específica del incidente, se encuentra en formato texto.
- **location description:** Esta columna contiene diversa información específica del incidente, se encuentra en formato texto.
- **latitud:** Esta columna contiene valores numéricos de la latitud geográfica del lugar del accidente.
- **longitude:** Esta columna contiene valores numéricos de la longitud geográfica del lugar del accidente.
- **n° guns involved:** Esta columna contiene valores numéricos y especifica la cantidad de armas que estuvieron involucradas.
- **notes:** Esta columna contiene diversa información específica y además está en tipo texto.
- **participant age group:** Dicha columna contenía redundancia, delimitadores y además contenía muchos detalles. Principalmente, existen a lo largo de estas 5 conjuntos en primer lugar ”*Adult +18*”, seguido de “*Child 0-11*”,“*Teen 12-17*”, “*Unknown*” y finalmente la combinación de estos grupos. En otras palabras, se encuentran las personas mayores de 18 años, seguidos de niños, adolescentes y otros rangos que comprende a más de una categoría.
- **participant gender:** Dicha columna contenía redundancia, delimitadores y además contenía muchos detalles. Principalmente, existen a lo largo de estas 3 conjuntos en donde se identificaron “*Male*”, “*Female*” y “*Unknown*”. De forma más específica se pueden apreciar el grupo de hombres, mujeres y en ciertos casos no se otorga el sexo.
- **avg_age:** Columna confeccionada, del dataset original cuyo datos son de tipos numéricos y saca el promedio de edad considerando la edad de las personas entre la cantidad de involucrados.

2.6. Normalización de Datos

La normalización de los datos desde Colaborario de Google(Colab), cuya base es el lenguaje de programación Python se lleva de la siguiente manera: En primer lugar, se debe leer el CSV, para así generar el "*Dataframe*", esto se logra mediante el uso de la librería "*Pandas*".

Posteriormente, se debe tratar de datos donde en primer lugar se procedió a identificar si el dataset contiene valores nulos, como muestra la siguiente imagen.

```
datos.isnull().sum()

incident_id                  6
date                         8
month                        8
day                          8
state                        7
city_or_county               6
address                  16497
n_killed                     8
n_injured                    8
congressional_district   11951
gun_stolen                   0
gun_type                     0
incident_characteristics   334
location_description    197482
latitude                  7931
longitude                 7931
n_guns_involved              0
notes                        0
participant_age_group        0
participant_gender           0
avg_age                   2063
dtype: int64
```

Imagen 2: Datos nulos del dataset (Fuente: Elaboración Propia)

Como es apreciable, el dataset contiene una gran cantidad de datos nulos ya sea por error del csv o inexistencia de los datos. Junto al grupo de trabajo se decidió cuáles serían las variables que usamos para la predicción y ante esto se decidió eliminar los nulos que se visualizan en la columna date, esto conllevo a que se eliminaran aquellas filas en el dataset que también afectan y agregaba nulos a las otras columnas necesarias, así dejando los datos llenos de información que si se utiliza.

Consiguiente a lo anterior, se eliminaron las columnas que no se utilizarían porque los datos que entrega son cualitativos, es decir, no aportan información necesaria. Estos datos corresponden a: "incident_id", "date", "address", "incident_characteristics", "location_description", "latitude", "longitude", "notes", "city_or_county", y "congressional_district"

Para el caso de date, se decidió eliminarla puesto que se utilizarán los valores separados de aquella columna es decir "month" y "day".

Luego de esto, se procede a detectar los “Outliers”, una vez realizado esto se procede dividir el “*Dataset*” en dos variables las cuales llamaremos “*variables_objetivo*” y “*variables_independientes*”. Donde la primera, contiene la resultante entre la cantidad de muertes y lesionados(“*heridos*”) y la segunda la columnas restantes.

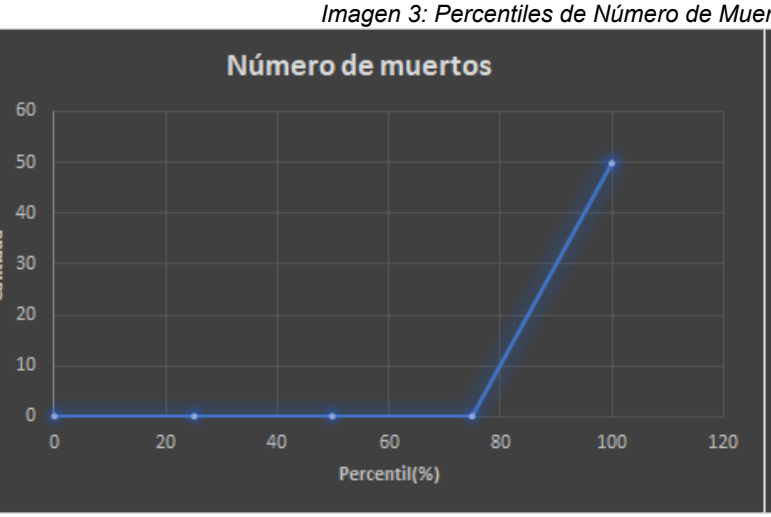
Enseguida se procede a realizar las distintas normalizaciones sobre las “*variables_independientes*”, esto es posible mediante la liberia “*sklearn*” mediante sus funciones como: “*scale*”(preprocessing.scale()), “*normalize*”(preprocessing.normalize()), “*minmaxscale*”(preprocessing.minmax_scale()). En esta contribución se utilizaron estas tres normalizaciones y adicionalmente uno sin normalizar.

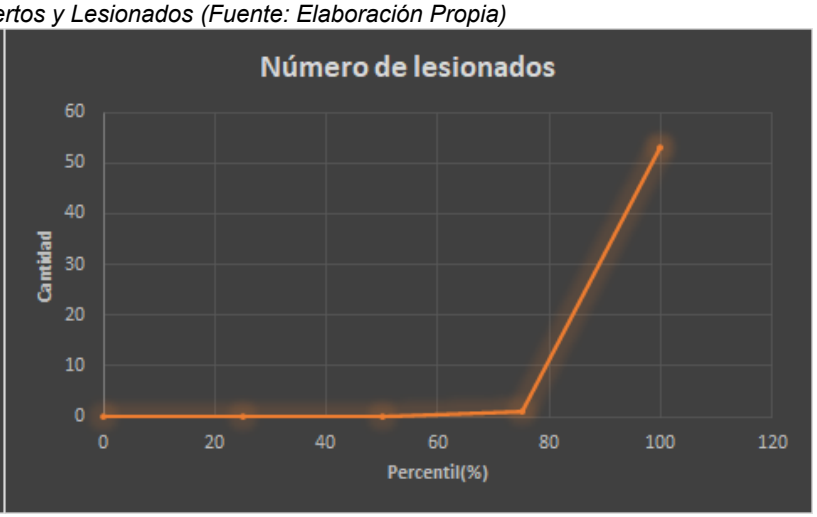
3. Análisis de Datos y Estadística Descriptiva

A partir de esta tabla, se podrá obtener los valores de promedio, desviación estándar, valores mínimo (percentil 0%), el primer cuartil (percentil de 25%), la mediana (percentil 50%), el tercer cuartil (percentil 75%) y valores máximos (percentil de 100%), para así poder describir las estadísticas presentes en la base de datos. Además se ha confeccionado un gráfico de percentil vs cantidad, para número de muertos(n_killed) y número de lesionados(n_injured). De los mismos gráficos, se logra visualizar que en el punto del valor máximo, fue lo que tiene mayor variación de datos en comparación a los puntos del valor mínimo, primer cuartil, mediana y tercer cuartil, lo cual implica la existencia de outliers. Todo lo anterior se visualiza en las siguientes figuras.

La gráfica de la izquierda alberga la cantidad de muertes respecto a los percentiles y la gráfica de la derecha la cantidad lesionados respecto a los percentiles. Cabe mencionar que en ambos casos las cantidades son bajas. Es por ello que más adelante se toma la decisión con estos “*target*” trabajar juntos y no separados.

Imagen 3: Percentiles de Número de Muertos y Lesionados (Fuente: Elaboración Propia)





Por otra parte, en la siguiente figura se contemplan por cada columnas dichas estadísticas respecto a los percentiles, el promedio e incluso la suma de los datos. Se espera que después de detectar los “outliers” se corrijan y los datos sean menos dispersos entre sí.

	month	day	n_killed	n_injured	gun_stolen	gun_type	n_guns_involved	participant_age_group	participant_gender
count	239560.000000	239560.000000	239560.000000	239560.000000	239560.000000	239560.000000	239560.000000	239560.000000	239560.000000
mean	6.351866	4.059050	0.252308	0.494027	1.956616	4.562364	0.803289	1.545354	1.384129
std	3.447823	2.023221	0.521819	0.730013	0.232755	1.803090	3.642508	2.737423	0.785232
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	2.000000	0.000000	0.000000	2.000000	5.000000	0.000000	0.000000	1.000000
50%	6.000000	4.000000	0.000000	0.000000	2.000000	5.000000	1.000000	0.000000	1.000000
75%	9.000000	6.000000	0.000000	1.000000	2.000000	5.000000	1.000000	2.000000	2.000000
max	12.000000	7.000000	50.000000	53.000000	2.000000	16.000000	400.000000	7.000000	3.000000

Imagen 4: Estadística Descriptiva del dataset (Fuente: Elaboración Propia)

Por último, la correlación entre las variables se puede visualizar en la siguiente figura, donde por orden se encuentran: df_index, mes(month), dia(day), n_killed(número de muertos), n_injured (número de lesionados), gun_stolen (armas robadas), gun_type (tipo de armas), n_guns_involved(numero de armas envuletas), participant_age_group (grupo de edad del participante), participant_gender (género del participante).

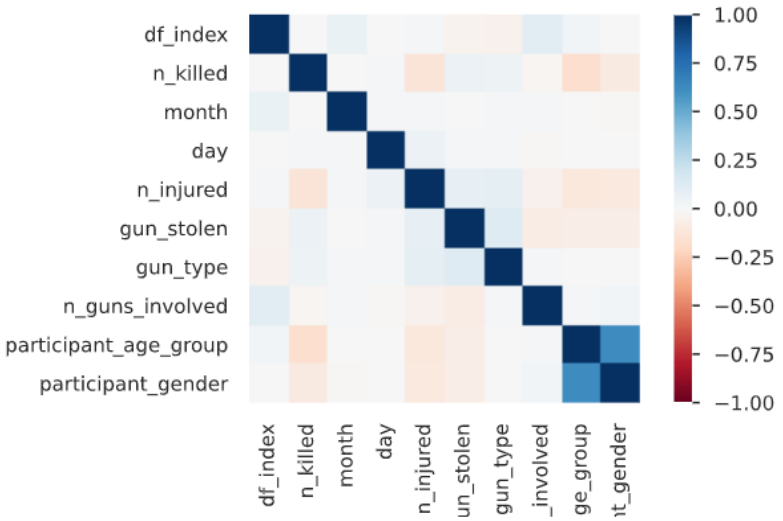


Imagen 5: Correlación de los datos del dataset (Fuente: Elaboración Propia)

Con el coeficiente de Person se puede identificar que tanto para el número de muertos y el número de lesionados, la correlación es “*negativa*” en su gran mayoría y en algunos casos existe correlación “*positiva*”. También se procedió a graficar los histogramas para cada campo, los cuales se podrán ver en el anexo.

3.1. Detección de outliers

Cuando se refiere a un "*outlier*", este se define como una observación o conjunto de observaciones que resultan inconsistentes respecto al conjunto de datos(Barnett y Lewis, 1994). En otras palabras, aquella observación errónea o atípica cuyo comportamiento es muy diferente a los datos restantes.

Existen métodos para corregir estos tipos de observaciones, puesto que su incidencia puede "desviar" los resultados. Por consiguiente, a lo largo de esta contribución se ocuparan los diagramas de Box-Plot para la detención de estos. También, se procedió a realizar los box-plot para cada campo, los cuales se podrán ver en el anexo.

Para el caso de método basado en el recorrido intercuartílico, se obtuvieron las cotas superior e inferior, el valor máximo y mínimo, cuartil inferior y superior, la mediana, rango inter-cuartil y los Outliers ya sean severos o no, esto para cada sección de datos agrupados lo cual se podrá ver con mayor detalles a continuación.

1) Para "month", "day", "participant_gender" no se encontraron outliers.
2) Para "gun_type", en esta columnas se perdió el 25% de datos puesto que se comportan de forma "anormal" considerando al resto. Es por ello que, se procedió a desprenderse de estos datos. Y logrando así, minimizar la cantidad de categorías.
3) Para "n_guns_involved", es esta columna se presentan datos que elevaban considerablemente la medidas de tendencia central de los datos, es por ello que se decidió eliminar estos datos puesto que no correspondian a un porcentaje significante(Menor a 5%).
4) Para "participant_age_group" se decidió desprenderse de aquellos datos que no aportan información respecto a la edad de los grupos, se considera un porcentaje menor. Además, dicho grupo también decidió englobarse en aquellos grupos que pueden ser imputados(mayores de 18) respecto a quienes no pueden ser imputados(menores de 18).
5) Para "avg_age", si bien en esta columna no aparece en el coeficiente de Person esto debe que la esta variable era tipo flotante, por lo tanto para se procedió a englobar en cuatro conjuntos: el primero abarca a todas edades menor a 18 años, el segundo aquellos mayores o igual 18 y menores de 25 años, enseguida los mayores o igual a 25 años y menores de 45 años, el tercero aquellos con 45 años o mayores y menores de 70 años. Finalmente, en el último conjunto aquellos mayores o iguales a 70 años.

6) Respecto a “n_killed” y “n_injured” se desprendieron de ambos un 25% de los datos puesto que estos no aportan información y tenían un comportamiento distinto al resto. Quedando como un variable binaria llamada “heridos” la cual se confeccionó de la siguiente manera: en caso que no exista ningún muerto y ningún lesionado, el valor de herido el valor es cero. Por otra parte, si al menos existe un lesionado y un lesionado(y en viceversa) el valor es uno. Al igual, si existe un muerto y un lesionado el valor es uno.
7) Con la columna “gun_stolen”, se decidió desprenderse de aquellas armas que no se sabía si fueron robadas o no, ya que casi el 75% de los datos no aportan información y solo desviaba el comportamiento de la variable, por lo que se dejó así una variable binaria.
8) Para el caso de “state”, se considera su comportamiento lineal en cuanto a sus cuartiles. Para intentar minimizar el impacto de los diversos estados ya que son 50 se decidió agrupar mediante sus superficies, creando una nueva columna llamada “state2” convirtiendo sus datos en solo 4 conjuntos el primero abarca aquellos estado de 4.000 a 100.000 km^2. El segundo conjunto, desde 1.000.000 a 2.000.000 km^2. El tercero, desde 2.000.000 a 3.000.0000. Y finalmente, el último conjunto de 3.000.000 o mayores, esto permite tanto disminuir la cardinalidad de los datos como convertir la variable cualitativa “state” en una variable ordinal concordante con la superficie del estado.

3.2. Analisis post-Outliers

Los problemas de aquellos datos llamados “atípicos” pueden incidir en la precisión de los resultados obtenidos para predecir. Para ello es necesario, definir la estrategia para reducir su impacto, puesto que si no se detecta un valor extremo, su consecuencia en la estimaciones de medidas de tendencia central ocasionando así una desviación de la zona donde existe mayor densidad. En el ámbito inferencial, todas las pruebas de hipótesis resultan ser sensibles al incumplimiento de supuesto en los modelos y a la presencia de outliers. Sin embargo, es necesario recalcar que el hecho de ser un dato outlier, no significa que sea un dato erróneo. A continuación, se dará a conocer como quedó la relación de los datos post-detección y eliminación de “*outliers*” destacándose la medidas estadística de los datos y la relación de Pearson existente entre los datos.

En la gráfica que se aprecia a continuación, se puede visualizar la confección del “target” y su comportamiento.

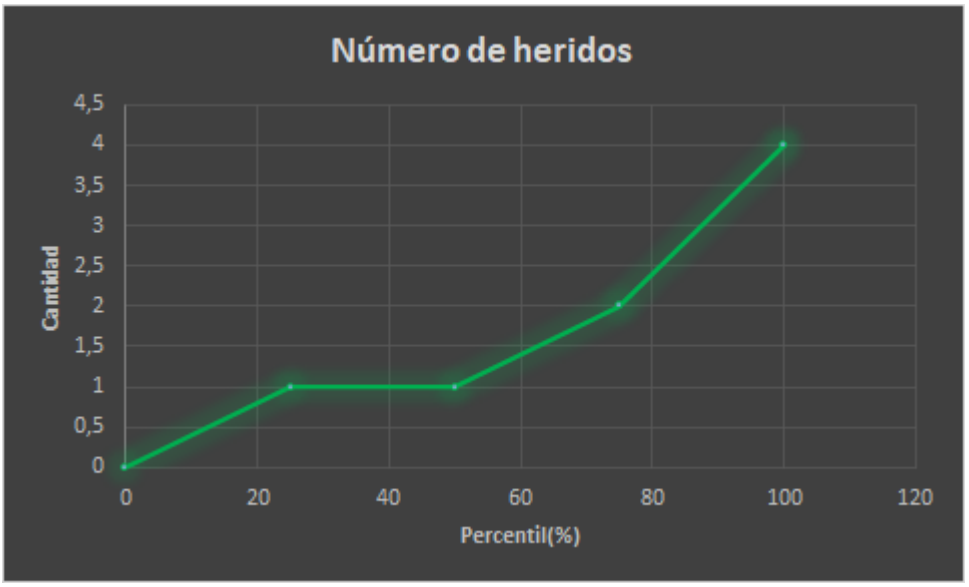


Imagen 6: Percentiles de los heridos (Fuente: Elaboración Propia)

Por otra parte, en la siguiente figura se contemplan por cada columnas dichas estadísticas respecto a los percentiles, el promedio e incluso la suma de los datos. Se espera que los datos sean menos dispersos entre sí puesto que los "*outliers*" no afectan.

	month	day	gun_stolen	gun_type	n_guns_involved	participant_age_group	participant_gender	avg_age	heridos	state2
count	4813.000000	4813.000000	4813.000000	4813.000000	4813.000000	4813.000000	4813.000000	4807.000000	4813.000000	4813.000000
mean	6.192396	3.930605	0.829628	1.735924	1.487014	0.077706	1.112819	1.738090	0.196967	1.168918
std	3.602260	1.940556	0.375998	1.178866	0.752447	0.267737	0.449840	1.318208	0.397748	0.997677
min	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	2.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000
50%	6.000000	4.000000	1.000000	1.000000	1.000000	0.000000	1.000000	2.000000	0.000000	1.000000
75%	9.000000	6.000000	1.000000	3.000000	2.000000	0.000000	1.000000	2.000000	0.000000	2.000000
max	12.000000	7.000000	1.000000	4.000000	3.000000	1.000000	3.000000	4.000000	1.000000	3.000000

Imagen 7: Estadística Descriptiva de los datos luego del análisis de outliers (Fuente: Elaboración Propia)

Al contemplar la figura anterior, uno a simple vista puede notar que las muestras se mueven en intervalos más acotados en cuanto a su distribución y además, es más "fácil" identificar a qué categoría pertenece algún dato.

Por otra parte, la relación existente entre las variables mediante la correlación de Pearson se acentuó en algunos parámetros que previamente no estaban tan esclarecidas. Existe una mayor cantidad al menos 3 variables (gun_stolen, gun_type, n_guns_involved) cuya relación es inversa. Por otro lado, 4 variables(month, day, participant_age_group, state2) cuya relación es no lineal y finalmente 2 variables (participant_gender, avg_age) cuya relación es directamente.

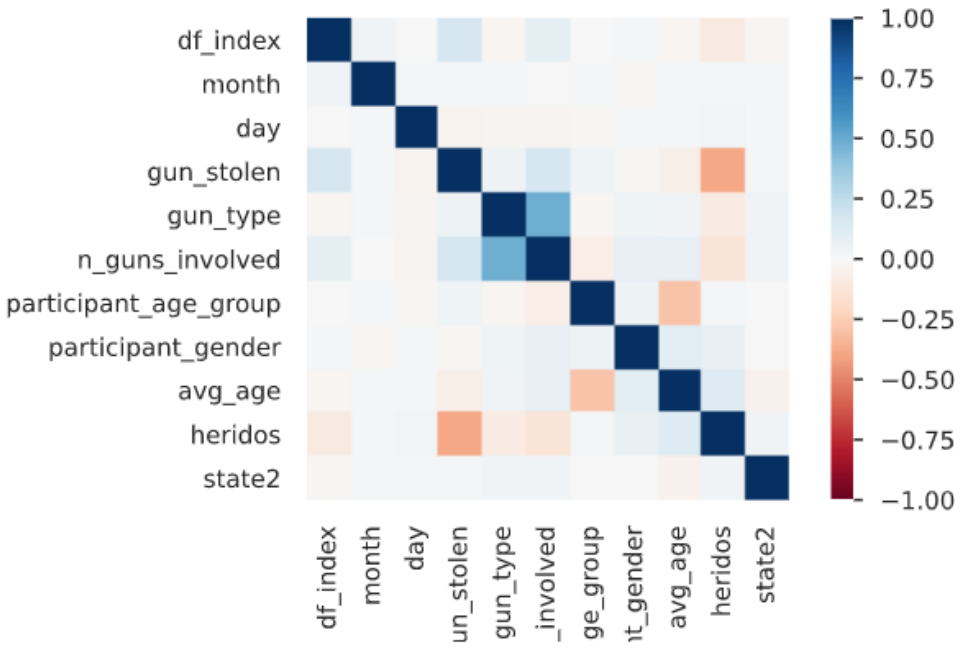


Imagen 8: Correlación de los datos luego del análisis de outliers (Fuente: Elaboración Propia)

4. Aplicación de Clustering

4.1. Aplicación de KNN

En el siguiente apartado se procedió a aplicar el algoritmo “K-means” y lo pondremos a prueba para que establezca un agrupamiento particional de la variable “heridos”. Como se mencionó anteriormente, dicho valor es un binario, por lo tanto mediante el método del codo se espera que la curva indique un punto de inflexión en el valor 2.

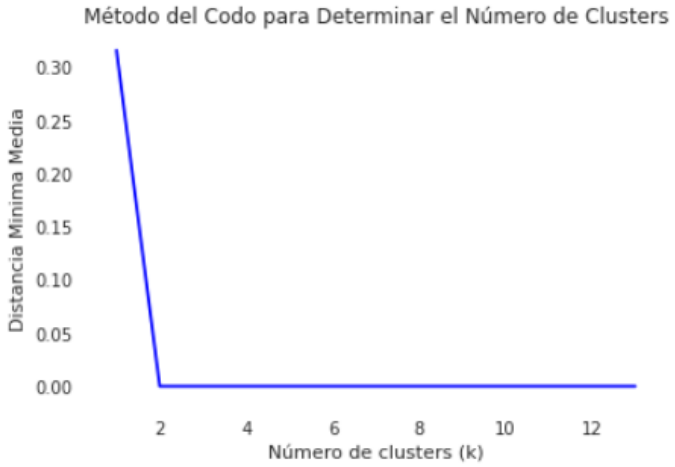


Imagen 9: Método del Codo para Determinar el Número de Clusters (Fuente: Elaboración Propia)

Este comportamiento era esperado, pero su impacto hubiera sido notable al considerar diversas variables cuantitativas sin un “patrón lógico” y este método fuera capaz de indicar la cantidad indicada para “dividir” la información mediante clusters.

4.1.1. KNN para 3 Grupos

En la siguiente figura, se visualiza el gráfico respecto al clustering con 3 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides(color verde), destacando que mayor distancia implica mayor diferencia de estos. Nuevamente, este análisis se pondrá a prueba, esperando que el algoritmo sea capaz de agrupar la cantidad de heridos teniendo en consideración el estado.

Imagen 10: Distribución de los datos para KNN 3 Grupos (Fuente: Elaboración Propia)

4.1.2. KNN para 5 Grupos

En la siguiente figura, se visualiza el gráfico respecto al clustering con 5 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides(color verde), destacando que mayor distancia implica mayor diferencia de estos. Nuevamente, este análisis se pondrá a prueba, esperando que el algoritmo sea capaz de agrupar la cantidad de heridos teniendo en consideración el estado.

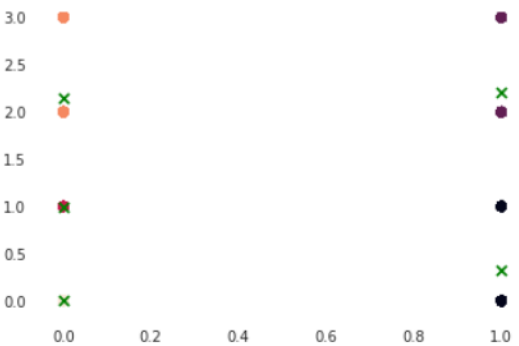


Imagen 11: Distribución de los datos para KNN 5 Grupos (Fuente: Elaboración Propia)

4.1.3. KNN para 7 Grupos

En la siguiente figura, se visualiza el gráfico respecto al clustering con 7 grupos, generados en el programa de Notebook Python, lo cual se detalla los colores que se diferencian en los grupos (clusters) y sus respectivos centroides(color verde), destacando que mayor distancia implica mayor diferencia de estos. Nuevamente, este análisis se pondrá a prueba, esperando que el algoritmo sea capaz de agrupar la cantidad de heridos teniendo en consideración el estado.

Imagen 12: Distribución de los datos para KNN 7 Grupos (Fuente: Elaboración Propia)

4.2. Agrupación Jerárquica

Para poder organizar los datos jerárquicamente, se utiliza la estructura de árbol o dendrograma, lo cual este tipo de estructura permite determinar el número adecuado de agrupamientos, como también para la detección de “outliers”. Para el caso de agrupación respecto las variables que componen el modelo , a partir de la siguiente figura se visualiza el dendograma, lo cual al principio solo se visualiza un cluster superior el cual contiene todos los datos, y por cada paso se van segmentando los clusters hasta formar uno solo como parte de la totalidad, siendo que desde el segundo paso de la división los valores de otros clusters se vuelven muy pequeños como resultado. Cabe destacar que, esta visualización se puede aplicar a datos, los cuales tendrán el mismo comportamiento. Pero como mayoritariamente el proyecto abordado tiene en su mayoría variables cualitativas se hace un poco “innecesario” la aplicación de este modelo.

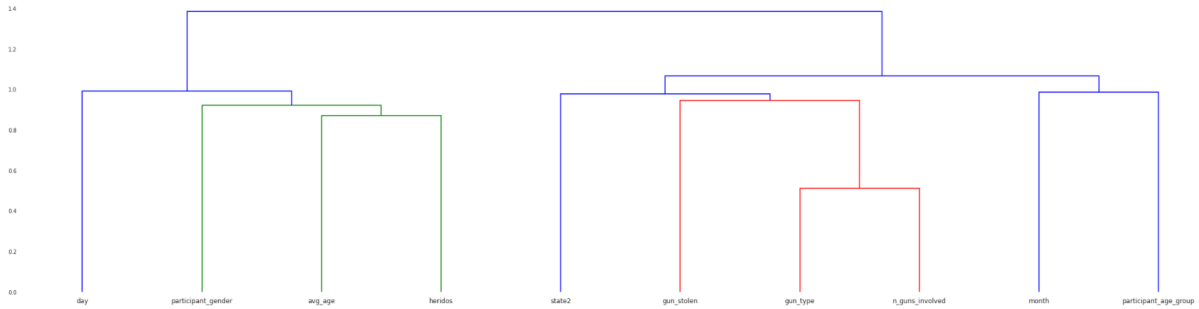


Imagen 13: Agrupación Jerárquica de los datos (Fuente: Elaboración Propia)

5. Análisis de Componentes Principales (PCA) y Factorial

5.1 Análisis de PCA

PCA, cuyas siglas en inglés dicta "*Principal Components Analysis*", y su traducción es "*Análisis de Componente Principales*", es un método cuya finalidad consiste en simplificar la complejidad de espacios muestrales, conservando la información y si es posible perder la menor cantidad de datos. En otras palabras, este método permite "comprimir" la información aportada por diversas variables, en unas pocas componentes.
Con el dataset que se ha trabajado a lo largo de toda la contribución, es necesario aplicar los siguientes pasos para realizar un correcto análisis de PCA:

1. Normalización de los datos: Dicho apartado es necesario aplicar un distribución normal en los datos, es por ello que se hace necesario que los datos sean de tipo número. A lo largo de esta contribución se dará a conocer las diversas normalización que se pueden ajustar en el código.

Rama	Procurar tener descomentar las siguientes líneas
1. Sin normalizar	`X_std = variables_independientes.values` `X_stdo = variables_objetivo.values`
2. Normalizado	`X_std = preprocessing.normalize(variables_independientes)` `X_stdo = preprocessing.normalize(variable_objetivo)`
3. MinMaxScaler	`X_std = preprocessing.minmax_scale(variables_independientes)` `X_stdo = preprocessing.minmax_scale(variable_objetivo)`
4. Scale	`X_std = preprocessing.scale(variables_independientes)` `X_stdo = preprocessing.scale(variable_objetivo)`

Tabla 1: Normalizaciones de los datos (Fuente: Elaboración Propia)

2. Calcular la matriz de Covarianza: Luego de normalizar los datos a utilizar, se debe calcular la matriz de covarianza para obtener los valores de ella, ver la distancias y además verificar que estos valores sean numéricos y viables para ser utilizados. En la siguiente ilustración se pueden observar los valores obtenidos al no aplicar normalización versus un MinMaxScaler

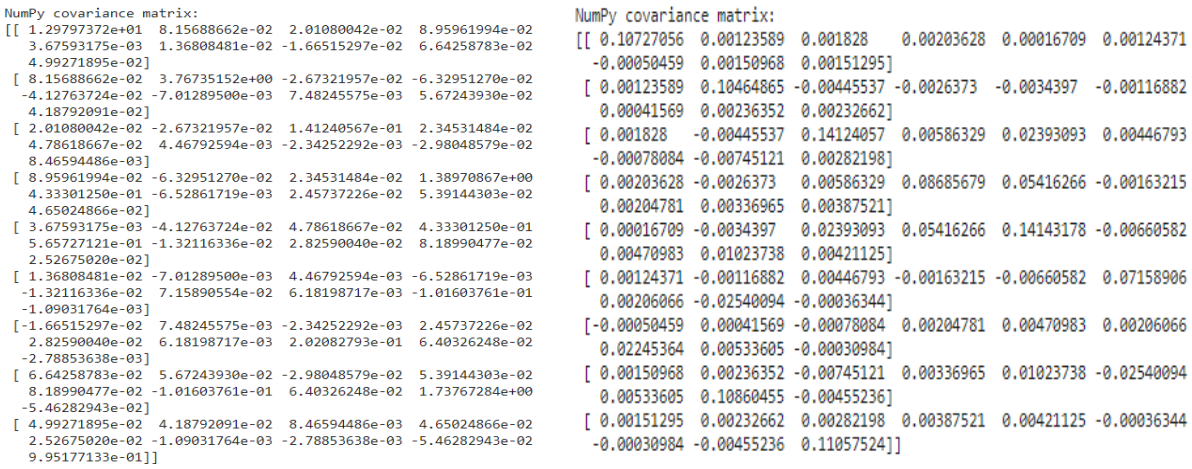


Imagen 14: Matriz de Covarianza: No normalizado(Izquierda) y MinMaxScaler(Derecha)
(Fuente: Elaboración Propia)

3. Calcular el vector con los valores eigen: Estos corresponden a números (eigenvalores) y vectores (eigenvectores) asociados a matrices cuadradas, los cuales proporcionan la información sobre cuando son transformados por el operador, dando el lugar a un múltiplo escalar de sí mismos, con lo que no cambian su dirección. Cabe mencionar que una transformación queda completamente determinada por sus vectores propios y valores propios, por lo que, para poder finalizar esta transformación de datos y este análisis PCA, se deben obtener estos valores y vectores para finalizar este proceso. En la siguiente figura, se aprecia los valores correspondientes a “eigenvalues” considerando la rama 1(No normalizar en comparación de un MaxMinScaler.

```
Eigenvectors
[[-9.99901669e-01  8.79768309e-03 -7.70587345e-03 -5.63493440e-03
  -3.69237861e-03 -2.63785183e-03  1.38365931e-03  1.46501691e-03
   1.86208983e-03]
 [-8.84499384e-03 -9.98937817e-01 -1.18211005e-02  3.83148381e-02
  -1.91264339e-02 -5.45292032e-03 -9.46687433e-04 -2.60148149e-03
  -5.83269394e-03]
 [-1.55327183e-03  7.99128642e-03 -7.37613898e-03  3.37019286e-02
   3.28863701e-03 -1.43959417e-01  5.31367417e-02 -6.60439543e-02
  -9.85303534e-01]
 [-7.74640308e-03  2.89971356e-02  3.21463082e-01  8.55064083e-01
  -1.00129566e-01  3.91916945e-01  4.96580810e-03  6.10069747e-03
  -3.06489948e-02]
 [-5.86847085e-04  1.61053521e-02  1.77675265e-01  3.50234076e-01
  -2.52335376e-02 -9.04726611e-01 -2.98467735e-02 -6.51120099e-02
   1.45638213e-01]
 [-1.00300140e-03  2.54404893e-03 -5.74279488e-02  1.74355563e-02
  -9.18675841e-03  1.31751915e-02 -9.93080977e-01  7.96638709e-02
  -5.98031071e-02]
 [ 1.25235524e-03 -2.26636534e-03  4.55720044e-02  6.02298484e-03
  -1.02212714e-03 -7.10981721e-02  7.92282687e-02  9.91889508e-01
  -5.19836752e-02]
 [-5.95232726e-03 -2.61035516e-02  9.26173613e-01 -3.56928515e-01
   8.25141354e-02  4.12058047e-02 -6.12473510e-02 -3.38756565e-02
  -2.61218254e-02]
 [-4.20118584e-03 -1.37389256e-02 -4.08304173e-02  1.25808054e-01
   9.90985896e-01  1.35427970e-02 -4.47241696e-03  3.71509950e-03
   5.34276883e-03]]

Eigenvalues
[12.98182604  3.7714429   1.7829169   1.54906309  0.98433327  0.3837086
  0.06419539  0.19878982  0.13401088]
```

```
Eigenvectors
[[-2.35591196e-02 -8.36012918e-03  3.00125759e-02  4.05987173e-02
  -1.49132759e-02  6.54623283e-02 -8.25684407e-01 -5.56390464e-01
   3.15159473e-02]
 [ 6.28111415e-02  4.97161120e-03 -1.37037176e-02  1.61886043e-03
   9.37987053e-02 -8.45703129e-01  2.39072466e-01 -4.55287023e-01
   8.59402934e-02]
 [-4.35202075e-01 -1.24962025e-02 -9.83365236e-02 -2.57859168e-03
  -7.36233784e-01  1.58804934e-02  1.63676443e-01 -2.32057867e-01
  -4.21702586e-01]
 [-4.43788469e-01 -3.40433021e-03 -8.27920936e-01 -8.97849563e-02
   1.95522576e-01 -7.37506680e-02 -1.06112528e-01  1.22457109e-01
   1.99002359e-01]
 [-7.69774685e-01  4.11682803e-02  5.36604646e-01  5.39061347e-02
   2.39716274e-01 -1.15786983e-01 -5.83151527e-02  1.39583228e-01
   1.45420210e-01]
 [ 5.12198280e-02  8.88616940e-02  1.18424578e-01 -8.78119711e-01
  -2.70686654e-01 -1.46437030e-01 -1.58845230e-01  1.77587451e-01
   2.29950740e-01]
 [-2.70980976e-02 -9.91758304e-01  3.58573979e-02 -1.13078877e-01
   3.75090534e-02 -2.54186145e-03  6.77579665e-03 -2.30705399e-03
  -1.19978794e-02]
 [-8.91175603e-02  8.10047542e-02 -9.55290023e-03 -4.48822561e-01
   5.21346897e-01  2.48138019e-01  2.09367206e-01 -3.56157088e-01
  -5.29061225e-01]
 [-7.27866353e-02 -3.37925261e-04  2.10325159e-02 -4.53046778e-02
  -8.22524261e-02  4.22350782e-01  3.87977791e-01 -4.86233285e-01
   6.48160391e-01]]

Eigenvalues
[0.18866167 0.02162841 0.05241622 0.0590492  0.13977039 0.10187918
 0.10597352 0.11044369 0.11484855]
```

Imagen 15: Vectores con valores eigen: No normalizado(Izquierda) y MinMaxScaler(Derecha)

(Fuente: Elaboración Propia)

El método PCA, relaciona cada una de las componentes con un "*eigevector*" cuyo orden se establece de manera creciente, logrando así relacionar la primera componente de "*eigenvector*" con el "*eigenvalores*" más alto. Una vez realizado los pasos anteriores, se deben listar los valores y vectores proporción obtenidos, para calcular la varianza la cual estará representada en un drama de barras la varianza correspondiente a cada autovalor y las acumulada de estos. Posteriormente, se genera una matriz que toma a los pares de autovalor- autovector y se representa gráficamente para así obtener el análisis de PCA.

5.2 Resultados del análisis de PCA

En la siguiente figura se puede visualizar la distribución por cada varianza existente en los datos recaudos del dataset, donde se destaca que existe una componente que tiene más peso que el resto, y alrededor de cuatro que su peso es menor. Se debe hacer hincapié en que si una componente tiene mayor varianza individual, implica que dicha componente puede aportar mucha información respecto al resto. A la izquierda se aprecia los resultados "no normalizados" y la derecha con "MinMaxScaler".

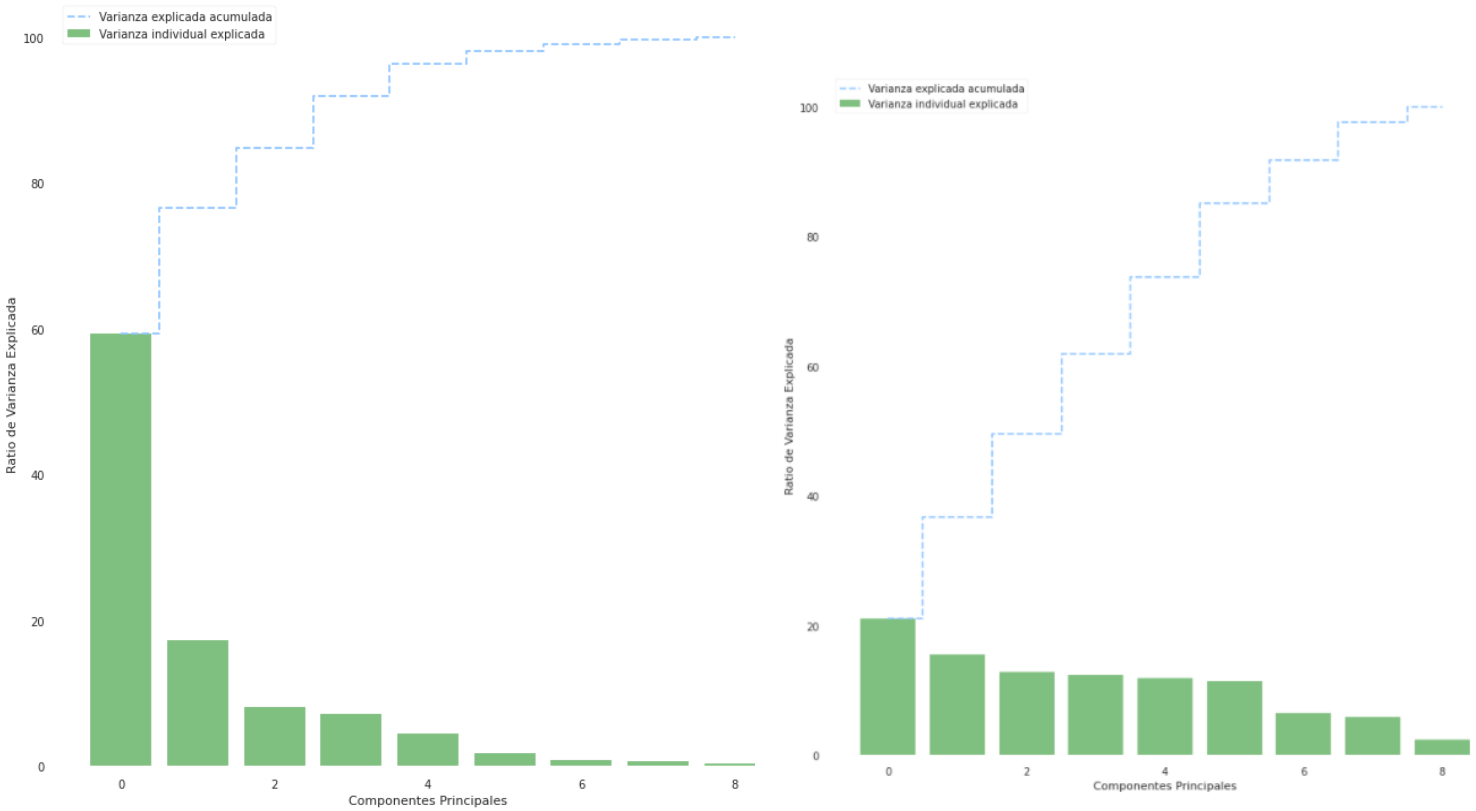


Imagen 16: Distribución de la varianza: No normalizado(Izquierda) y MinMaxScaler(Derecha)
(Fuente: Elaboración Propia)

5.3 Análisis Factorial

El análisis factorial, considera la agrupación de una serie de procedimientos análisis multivariable, donde analiza la relación entre las variables. De tal forma, dicho análisis brinda la posibilidad de estudiar la interdependencia entre un conjunto de variables. Su finalidad es descubrir algún "patrón" o algo no observable. Con la reducción de la información que son proporcionadas por "p"(variables observadas), con la menor pérdida de información en un número inferior a "k" (variables no observadas). Algunas de las características de la reducción o agrupación de las variables son las siguientes:

- Juntar bajo un factor variables que estén muy correlacionadas entre sí.
- Garantizar que las variables una vez agrupadas en los distintos factores, guarden relación.

En tal caso, que la correlación entre los factores el valor e igual a cero, informa que dicho factor representa una dimensión distinta en los datos.

5.4 Resultado de Análisis Factorial

El análisis de componentes principales (PCA) y el análisis factorial guardan similitud, ya que ambos utilizan la simplificación de la estructura de un conjunto de variables. Sin embargo, los análisis difieren de varias maneras importantes como en los siguientes puntos:

- En el análisis de componentes principales, estos componentes se calculan como combinaciones lineales de las variables originales. Mientras que, en el análisis factorial, las variables originales se definen como combinaciones lineales de los factores.

- En el análisis de componentes principales, su finalidad es explicar la proporción de la varianza total en las variables como sea posible. Por otra parte, el objetivo en el análisis factorial es explicar las covarianzas o correlaciones entre las variables.

- El análisis de componentes principales busca reducir los datos a un número más pequeño de componentes. Por otra parte, el análisis factorial pretende entender los constructos que subyacen a los datos.

Ambos métodos permiten la simplificación y proporcionan estabilidad al transformar los datos, es por ello que la incorporación de ambos debe ser considerado para abordar problemas de predicción. En la siguiente figura se visualiza el gráfico, el cual mediante el método del codo, indica que con 2 o 7 factores se produce un quiebre en la curva del gráfico.

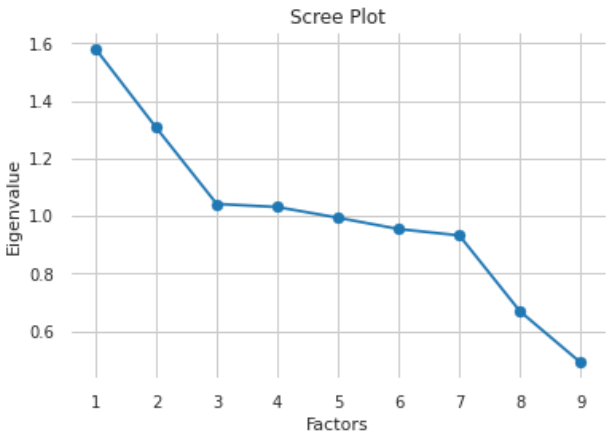


Imagen 17: Método del Codo para conocer los factores (Fuente: Elaboración Propia)

Una vez seleccionado la reducción de 3 factores, en la figura se aprecia cómo cada componente índice en dicho factor. En tal caso al identificar "*month*" esta variable en los factores tiene poca incidencia, por lo que se puede decir que "representa" en baja a escala la composición de tal factor. En caso contrario, a "*n_guns_inolved*" en el factor 1, donde su significancia es considerable.

	Factor1	Factor2	Factor3
month	0.010316	0.013525	0.009615
day	-0.038089	-0.009022	0.046034
gun_stolen	0.164527	0.034993	-0.121154
gun_type	0.486081	0.015432	0.025063
n_guns_involved	0.997454	0.006512	0.046502
participant_age_group	-0.064715	0.982228	-0.162060
participant_gender	0.074251	0.093049	0.220219
avg_age	0.057227	-0.198040	0.552002
state2	0.047068	-0.011283	-0.075211

Imagen 18: Valores de los factores para cada dato para Análisis Factorial (Fuente: Elaboración Propia)

Finalmente, se obtiene información respecto a la proporción de la varianza, la acumulación de esta y la carga de SS por cada factor obtenido.

	Factor1	Factor2	Factor3
SS Loadings	1.275008	1.014547	0.404803
Proportion Var	0.141668	0.112727	0.044978
Cumulative Var	0.141668	0.254395	0.299373

Imagen 19: Estadística de los factores para Análisis Factorial (Fuente: Elaboración Propia)

6. Métodos de Clasificación

Con la finalidad de encontrar el mejor modelo predictivo para cumplir la hipótesis planteada, se procede a modelar los datos en los diferentes modelos predictivos, en primera instancia se evalúan los árboles de decisiones, para este caso se consideran las variables Índice de Gini y Entropía Cruzada. También se evalúa con Random Forest y Adaboost con la finalidad de poder realizar una comparación y análisis de manera más precisa en el comportamiento de los modelos con los datos entregados.

Para realizar este análisis, se definen las variables X e Y cómo las variables independientes y dependientes, además es necesario definir las variables que se utilizaran para entrenamiento y evaluación de los modelos, es necesario definir que se trabajó con el 20% de los datos.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import plot_confusion_matrix

# Separamos en dos variables (X e Y)
X = datos3
Y = datos4['heridos']

# Separando los datos en sets de entrenamiento y evaluación
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2)
```

Imagen 20: Código de definición de los Algoritmos para Métodos de Clasificación (Fuente: Elaboración Propia)

6.1 Árbol de Decisión: Índice de Gini

El índice de gini mide el grado de pureza del nodo, esto mide la probabilidad de no sacar dos registros del mismo nodo. Se selecciona la variable con menor gini ponderado. Para esto se trabajó con la librería sklearn que contiene predefinidas funciones para modelos de árboles de decisiones, como se presenta en la siguiente imagen:

```
# Creando el modelo con criterio 'gini'
arbolGini = DecisionTreeClassifier(criterion='gini')

# Ajustando el modelo
arbolGini.fit(X_train, y_train)
```

Imagen 21: Código del Método Árbol de Decisión con Índice de Gini (Fuente: Elaboración Propia)

6.2 Árbol de Decisión: Entropía Cruzada

La variable objetivo categórica definida como la entropía corresponde a una donde se escoge el nodo que tiene la entropía más baja en comparación a su padre y el resto de los nodos, mientras menor sea la entropía, es mejor. Ajustando la librería de sklearn a criterio de entropía se realiza el mismo ejercicio.

```
# Creando el modelo con criterio 'Entropy'
arbolEntropy = DecisionTreeClassifier(criterion='entropy')
# Ajustando el modelo
arbolEntropy.fit(X_train, y_train)
```

Imagen 22: Código del Método Árbol de Decisión con Entropía Cruzada (Fuente: Elaboración Propia)

6.3 Random Forest

Un random forest es una técnica de aprendizaje automático supervisada que a diferencia de los árboles de decisión, tienen la capacidad de generalizar los problemas en cuanto al error, consistiendo en un conjunto de árboles de decisiones combinados a través de un bagging, la cual hace que diferentes árboles sean capaces de ver distintas porciones de los datos, si bien, se puede creer que esto crea una confusión entre los árboles al ver los datos, en realidad cada uno realiza los entrenamientos por separados, de forma que termine combinando los resultados, compensando los errores entre ellos y obteniendo un mejor resultado. En caso de este modelo, se continuó utilizando la librería ya mencionada, ajustando sus valores como corresponden al modelo.

```
# Importando el random forest
from sklearn.ensemble import RandomForestClassifier

randomForest = RandomForestClassifier() # Creando el modelo
randomForest.fit(X_train, y_train) # Ajustando el modelo
```

Imagen 23: Código del Método Random Forest (Fuente: Elaboración Propia)

6.4 Adaboost

Este algoritmo, nace del término "*Boosting*", el cual es una meta algoritmo cuya base se encuentra en el aprendizaje automático, donde se reduce el sesgo y varianza en un contexto de aprendizaje supervisado. El algoritmo mencionado anteriormente, es considerado como un clasificador débil, puesto que la correlación sucede con la clasificación correcta. Por otra parte, "Adaboost" históricamente es considerado como uno de los más importante, debido a que fue una de las primeras incursiones algorítmicas que logró aprender de clasificadores débiles. Se debe destacar que hoy, existen muchas variaciones y modificaciones que cumplen la misma trea donde encontramos a: "*LPBoost*", "*BrownBoost*", "*LogitBoost*","*XGBoost*", entre otros.

Existen diversos algoritmo de "*Boosting*" el cual pueden ser englobados en "*AnyBoost*". Esta última categoría, actúa mediante el descenso del gradiente en el espacio funcional utilizando una función de costo convexa. El funcionamiento a grandes rasgos es clasificar conjuntos de datos originales y estos son ajustados, haciendo una copia adicional del clasificador en el mismo conjunto, señalando donde los pesos de aquella instancias clasificadas erróneamente se ajustan de tal que los clasificadores posteriores, solo se centran en aquellos casos críticos.

```
# Utilizando AdaBoost
adaBoost = AdaBoostClassifier(n_estimators=500,
                              learning_rate=1.5)
# Ajustando los datos
adaBoost = adaBoost.fit(X_train, y_train)
```

Imagen 24: Código del Método Adaboost (Fuente: Elaboración Propia)

6.5 Naive Bayes

El Naive Bayes o también conocido como el Clasificador bayesiano ingenuo corresponde a un clasificador probabilístico fundamentado en el teorema de bayes, el cual relaciona conjuntos de variables aleatorias, mediante un grafo dirigido. Su característica principal es la composición de una red gráfica sin ciclos donde se representan las variables aleatorias y las relaciones probabilísticas que existen entre ellas, permitiendo así soluciones a problemas de decisiones cuando exista la incertidumbre.

Se debe hacer hincapié en que esta red, corresponde a una representación de dependencias para el razonamiento de probabilidades, puesto que los nodos son variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables.

```
gNaiveBayes = GaussianNB()
gNaiveBayes.fit(X_train, y_train)
```

Imagen 25: Código del Método Naive Bayes (Fuente: Elaboración Propia)

6.6 Resultados de Modelos de Clasificación

Una vez de haber evaluado los distintos modelos de clasificación, se realizó una evaluación final de todos, donde el parámetro que se ajusta es la rama para normalizar entregando diversos resultados en los modelos de clasificación los cuales se aprecian en la siguiente tabla:

Con la rama 2 “Normalizado”			
Modelo	Precisión modelo inicial entrenamiento	Precisión modelo inicial prueba	Evaluación Precisión(%)
Árbol Gini	0.965	0.750	73.767
Árbol Entropy	0.965	0.753	73.582
Random Forest	0.965	0.794	78.783
AdaBoost	0.829	0.823	81.634
Clasificador Bayesiano Ingenuo	0.811	0.816	81.116

Tabla 2: Resultados de cada Modelo con datos Normalizados (Fuente: Elaboración Propia)

Con la rama 3 “MaxMinScaler”			
Modelo	Precisión modelo inicial entrenamiento	Precisión modelo inicial prueba	Evaluación Precisión(%)
Árbol Gini	0.967	0.748	74.057
Árbol Entropy	0.967	0.747	73.394
Random Forest	0.967	0.793	78.596
AdaBoost	0.831	0.825	81.634
Clasificador Bayesiano Ingenuo	0.805	0.83	81.116

Tabla 3: Resultados de cada Modelo con datos en MixManScaler (Fuente: Elaboración Propia)

Con la rama 4 “Scale”			
Modelo	Precisión modelo inicial entrenamiento	Precisión modelo inicial prueba	Evaluación Precisión(%)
Árbol Gini	0.962	0.769	73.810
Árbol Entropy	0.962	0.755	73.627
Random Forest	0.962	0.802	78.761
AdaBoost	0.832	0.830	81.634
Clasificador Bayesiano Ingenuo	0.808	0.825	81.116

Tabla 4: Resultados de cada Modelo con datos en Scale (Fuente: Elaboración Propia)

Con la rama 1 “Sin normalizar”			
Modelo	Precisión modelo inicial entrenamiento	Precisión modelo inicial prueba	Evaluación Precisión(%)
Árbol Gini	0.967	0.740	73.662
Árbol Entropy	0.967	0.744	73.207
Random Forest	0.967	0.771	78.638
AdaBoost	0.836	0.804	81.634
Clasificador Bayesiano Ingenuo	0.809	0.814	81.116

Tabla 5: Resultados de cada Modelo con datos Sin Normalizar (Fuente: Elaboración Propia)

Por lo que, en este apartado se puede apreciar que “*AdaBoost*” y el “*Clasificador Bayesiano Ingenuo*” en ese orden son los modelos que presentan mejores resultados cuando los datos se encuentran Normalizados. Mientras que, el Árbol tanto de Gini, como Entropy tienen los porcentajes menores en la evaluación.

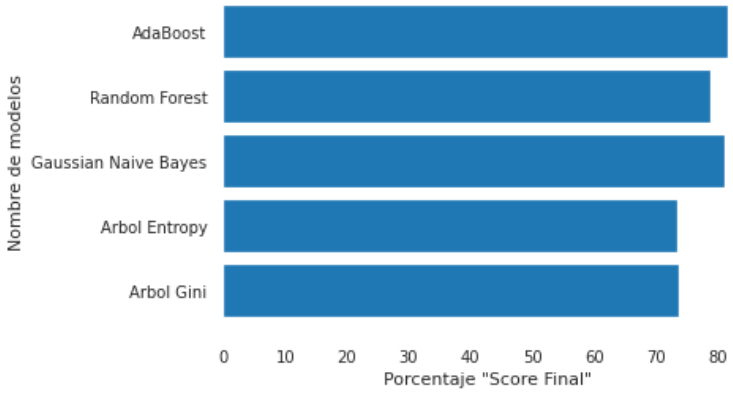


Imagen 26: Gráfico de porcentaje de los Modelos con datos Normalizados (Fuente: Elaboración Propia)

Al cambiar usar el “*MinMaxScaler*”, se nota un pequeña mejora en líneas generales pero la tendencia sigue conservandose. Donde “*Adaboost*” y el “*Clasificador Bayesiano Ingenuo*” presentan los mejores resultados. Adicionalmente, el “*Árbol Entropy*” y “*Árbol Gini*” no sobrepasan el 75%. Esto se puede observar en la siguiente figura.

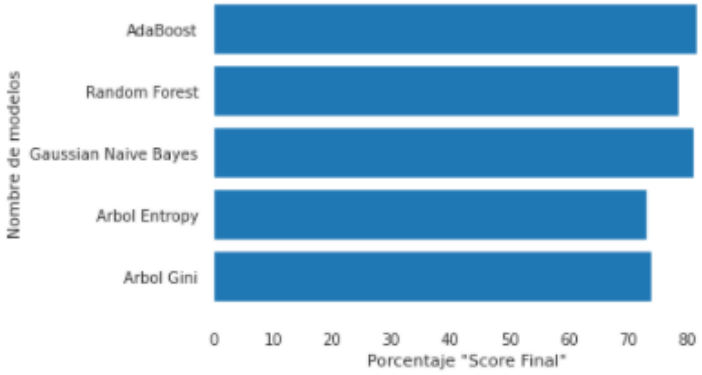


Imagen 27: Gráfico de porcentaje de los Modelos con datos en MinMaxScaler (Fuente: Elaboración Propia)

Al ajustar la normalización y usar "*Scale*", se nota muy poca mejora en líneas generales pero la tendencia sigue conservandose. Donde "*Adaboost*" y el "*Clasificador Bayesiano Ingenuo*" presentan los mejores resultados. Adicionalmente, el "*Árbol Entropy*" y "*Árbol Gini*" no sobrepasan el 75%. Esto se puede observar en la siguiente figura.

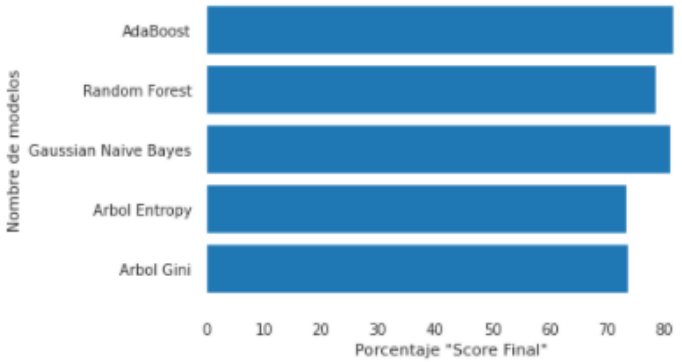


Imagen 28: Gráfico de porcentaje de los Modelos con datos en Scale(Fuente: Elaboración Propia)

Al no aplicar la normalización, la tendencia sigue conservandose. Donde "*Adaboost*" y el "*Clasificador Bayesiano Ingenuo*" presentan los mejores resultados. Adicionalmente, el "*Árbol Entropy*" y "*Árbol Gini*" no sobrepasan el 75%. Esto se puede observar en la siguiente figura.

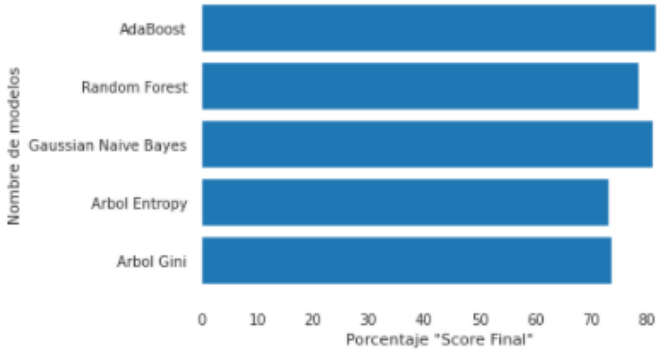


Imagen 29: Gráfico de porcentaje de los Modelos con datos sin normalizar (Fuente: Elaboración Propia)

Por otra parte, al centrarse en la profundidad de los árboles se obtienen los siguientes valores:

Segunda rama (normalizar)			
Profundidad	Resultados	Árbol de Decisión	AdaBoost
3	Error cuadrático medio	0.356	0.350
	Precisión prueba	18.86%	21.02%
	Precisión entrenamiento	19.22%	17.50%
4	Error cuadrático medio	0.356	0.362
	Precisión prueba	18.96%	15.60%
	Precisión entrenamiento	20.36%	15.75%
5	Error cuadrático medio	0.362	0.379
	Precisión prueba	16.22%	7.83%
	Precisión entrenamiento	21.78%	14.29%

Tabla 6: Resultados de las profundidades con datos Normalizados (Fuente: Elaboración Propia)

Adicionalmente, al centrarse en la profundidad de los árboles se obtienen los siguientes valores:

Tercera rama (“MaxMinScaler”).			
Profundidad	Resultados	Árbol de Decisión	AdaBoost
3	Error cuadrático medio	0.353	0.357
	Precisión prueba	22.58%	20.43%
	Precisión entrenamiento	18.23%	17.38%
4	Error cuadrático medio	0.357	0.369
	Precisión prueba	20.71%	15.31%
	Precisión entrenamiento	19.67%	15.70%
5	Error cuadrático medio	0.361	0.374
	Precisión prueba	18.99%	12.93%
	Precisión entrenamiento	21.51%	15.67%

Tabla 7: Resultados de las profundidades con datos en MaxMinScaler (Fuente: Elaboración Propia)

Por otra parte, al enfocarse en la profundidad de los árboles se obtienen los siguientes valores:

Cuarta rama (“Scale”).			
Profundidad	Resultados	Árbol de Decisión	AdaBoost
3	Error cuadrático medio	0.365	0.362
	Precisión prueba	19.32%	20.42%
	Precisión entrenamiento	18.74%	18.25%
4	Error cuadrático medio	0.367	0.369
	Precisión prueba	18.51%	17.35%
	Precisión entrenamiento	20.13%	18.44%
5	Error cuadrático medio	0.369	0.385
	Precisión prueba	17.22%	10.03%
	Precisión entrenamiento	21.63%	12.96%

Tabla 8: Resultados de las profundidades con datos en Scale (Fuente: Elaboración Propia)

Finalmente, cuando no se realiza la normalización se refleja en la profundidad de los árboles con los siguientes valores:

Primera rama(“Sin normalizar”)			
Profundidad	Resultados	Árbol de Decisión	AdaBoost
3	Error cuadrático medio	0.376	0.374
	Precisión prueba	18.65%	19.28%
	Precisión entrenamiento	18.76%	18.32%
4	Error cuadrático medio	0.377	0.383
	Precisión prueba	18.15%	15.43%
	Precisión entrenamiento	20.06%	14.84%
5	Error cuadrático medio	0.377	0.378
	Precisión prueba	18.08%	17.87%
	Precisión entrenamiento	21.77%	16.6%

Tabla 9: Resultados de las profundidades con datos sin normalizar (Fuente: Elaboración Propia)

7. Redes Bayesianas, Máquinas de Soporte Vectorial y Validación Cruzada

7.1 Redes Bayesianas

Las redes bayesianas o red de bayes son un modelo probabilístico que relaciona un conjunto de variables aleatorias mediante un grafo dirigido el cual se caracteriza por ser una red gráfica acíclica en la cual se respetan diferentes variables de manera aleatoria y sus relaciones probabilísticas entre ellas, lo cual, permite conseguir soluciones a problemas de decisión en caso de incertidumbre.

Aplicando las librerías existentes para la evaluación de Modelos de Redes Bayesianas, se procedió a evaluar en cada caso entregando como resultado su Error Cuadrático Medio y la Precisión del Modelo mostrando los resultados en las siguientes tablas:

Rama 2 “Normalizado”	
Error Cuadrático Medio	0.184
Precisión Modelo	0.815

Tabla 10: Resultados de la red bayesiana con datos Normalizados (Fuente: Elaboración Propia)

Rama 3 “MaxMinScaler”	
Error Cuadrático Medio	0.203
Precisión Modelo	0.796

Tabla 11: Resultados de la red bayesiana con datos en MaxMinScaler (Fuente: Elaboración Propia)

Rama 4 “Scale”	
Error Cuadrático Medio	0.176
Precisión Modelo	0.823

Tabla 12: Resultados de la red bayesiana con datos en Scale (Fuente: Elaboración Propia)

Rama 1 “Sin Normalizar”	
Error Cuadrático Medio	0.174
Precisión Modelo	0.829

Tabla 13: Resultados de la red bayesiana con datos sin normalizar (Fuente: Elaboración Propia)

7.2 Máquinas de Soporte Vectorial

Support Vector Machines (SVM) o como en su traducción al español Máquinas de Soporte Vectorial son un tipo de algoritmo de aprendizaje supervisado utilizado como clasificador en problemáticas de clasificación y regresión, el cual como función realiza un hiperplano óptimo donde clasifica en dos espacios dimensionales un conjunto de datos. Para realizar esta separación entre las dos clases que identifica el modelo, es necesario algo llamado “kernel” el cual se le podría determinar como la línea que crea la separación, no obstante, es conocido que no siempre es posible hacer una separación lineal, es por ello que existen diferentes kernel para poder solucionar aquellas problemáticas.

Utilizando las herramientas entregadas por las librerías importadas, se evaluaron los datos a estudiar en el modelo SVM con 3 diferentes kernel las cuales son: Poly (Polinominal-homogénea), RBF (Base Radial Gaussiana) y Linear (Lineal), con el fin de encontrar cual es el más preciso. También esta evaluación se realizó con cada una de las ramas (Sin normalizar, normalizado, minmaxscaler y scale), las siguientes tablas muestran los resultados de cada experimento.

Con Rama 2(“Normalizado”)				
Tipo de Kernel	X_test	y_test	Precisión	Error cuadrático medio test
Poly	0.121	0.075	0.084	0.372
RBF	0.028	0.112	0.088	0.380
Linear	-0.024	0.087	-0.057	0.389

Tabla 14: Resultados de SVM con datos Normalizados (Fuente: Elaboración Propia)

Con Rama 3(“MinMaxScaler”)				
Tipo de Kernel	X_test	y_test	Precisión	Error cuadrático medio test
Poly	0.134	0.026	0.101	0.378
RBF	0.203	0.132	0.197	0.373
Linear	0.114	0.071	0.138	0.372

Tabla 15: Resultados de SVM con datos en MinMaxScaler (Fuente: Elaboración Propia)

Con Rama 4(“Scale”)				
Tipo de Kernel	X_test	y_test	Precisión	Error cuadrático medio test
Poly	0.153	0.027	0.111	0.379
RBF	0.169	0.100	0.137	0.378
Linear	0.097	0.018	0.095	0.385

Tabla 16: Resultados de SVM con datos en Scale (Fuente: Elaboración Propia)

Con Rama 1(“Sin normalizar”)				
Tipo de Kernel	X_test	y_test	Precisión	Error cuadrático medio test
Poly	0.037	0.035	0.077	0.396
RBF	0.077	0.124	0.138	0.393
Linear	0.008	0.055	0.106	0.393

Tabla 17: Resultados de SVM con datos sin normalizar (Fuente: Elaboración Propia)

7.2 Resultados Redes Bayesianas y SVM

Analizando cada resultado obtenido ejecutando el Modelo de Redes Bayesianas con los datos trabajados, es posible identificar que el modelo tiene una mejor precisión de asertividad cuando los datos están en rama 4 “scale”, así como también es posible identificar que su error cuadrático medio es menor, es decir, los datos no están tan diferentes o alejados de lo ideal.

Respecto al análisis de los resultados con SVM, es posible identificar que cuando se evaluó con la Rama 3 de MinMaxScaler, éste entregó mejores resultados en cuanto a su precisión y en cuanto al Error Cuadrático medio, es decir tiene menos datos demasiado separados de la muestra predecida. En cuanto a los diferentes kernel, es posible identificar que el uso de kernel RBF (Base Radial Gaussiana) entrega mejores resultados, por lo tanto podríamos definir que los datos utilizados no tienen una separación tan notoria y es necesario utilizar un método que viaje entre lo lineal y lo complejo.

7.3 Validación Cruzada

La Validación Cruzada o cross-validation es una técnica que se utiliza bastante para evaluar los resultados de un análisis estadístico y crear una garantía de que los datos de entrenamiento y de prueba son independientes entre sí. Este proceso consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre las diferentes particiones, se utiliza en casos de predicción para estimar la precisión del modelo y así llevarlo a la práctica.

Para nuestro caso, se realizó validación cruzada con K Fold = 20 para Árbol de decisión con Índice de Gini, Árbol de decisión con Entropía Cruzada, Random Forest, Adaboost, Bayesiano Ingenuo y Redes Bayesianas, la siguiente tabla muestra los resultados de acierto tanto en general como en su estado de testeo.

Modelos	Acierto para no heridos		Acierto para heridos		Acierto del modelo	
	All	Test	All	Test	All	Test
Árbol de Decisión con índice de gini	99.92%	83.91%	86.67%	34.8%	97.31%	74.23%
Árbol de Decisión con Entropía Cruzada	99.92%	84.36%	86.45%	34.74%	97.27%	74.55%
Random Forest	99.34%	91.30%	88.82%	30.88%	97.27%	79.38%
Adaboost	95.58%	95.31%	32.10%	30.48%	83.26%	82.54%
Bayesiano Ingenuo	90.17%	90.34%	45.47%	45.31%	83.18%	82.48%
Redes Bayesianas	90.37%	90.08%	45.67%	45.34%	81.51%	81.32%

Tabla 18: Resultados de Validación Cruzada con K=20. (Fuente: Elaboración Propia)

Cómo es posible inferir y entender de los datos obtenidos, entre los 6 modelos ejecutados, Random Forest tiene una mejor predicción de los datos en cuanto a nuestra hipótesis, puesto que si se revisa, el modelo predice casi en su totalidad los casos sin heridos, no obstante, nosotros deseamos saber cuando hay heridos, por ello el 88.82% que muestra en el proceso completo sería el más correcto, no obstante, hay que destacar que al momento de realizarse el test del modelo predictivo, su porcentaje de acierto disminuye hasta un 30.88%, esto se puede deber a la baja cantidad de datos que utiliza el modelo para el proceso de pruebas.

8. Conclusiones

Un modelo predictivo para análisis del Crimen con Arma de Fuego en EE.UU, proporciona una toma de decisiones para tiempos futuros y no solo, sino que en instantes cortos de tiempo. Cabe destacar que cualquier persona que porte un arma de fuego (legal o no), tiene una tremenda responsabilidad de su uso ya que puede afectar y poner en peligro a otras personas, ya sea los delincuentes (defensa propia) o cualquier individuo inocente.
Particularmente el enfoque de la contribución se destina a cómo distribuir recursos de emergencia para así tener una mayor probabilidad de “no perder tiempo” e intentar atender a personas heridas de gravedad en estados donde los crímenes sean esporádicos . A su vez, se recogieron datos confiables y se trataron para determinar este objetivo. Por otra parte, se generaron varios modelos predictivos, donde el más acertado después realizar la validación cruzada corresponde a Random Forest.

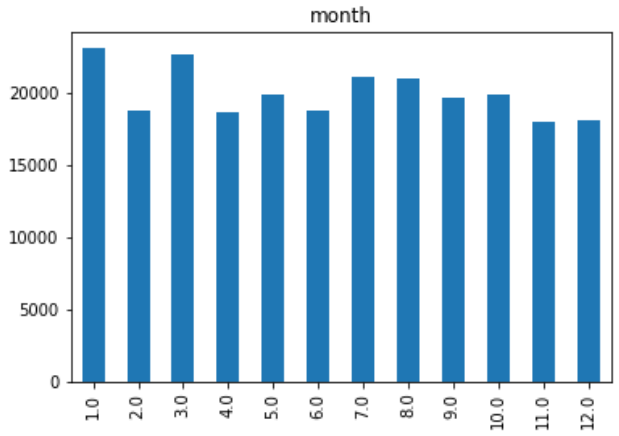
Para el modelo predictivo en general, es importante utilizar las herramientas y realizar estructuración de codificación correctamente, ya que este tipo de contexto corresponde a problemas de regresión, por lo que si se utiliza la codificación para clasificación puede ocurrir problemas en cuanto a las sintaxis y generación de resultados. Cabe destacar que en la predicción, dentro de este proyecto, en ningún caso puede estimar con total exactitud, pero si se deben considerar lo probable que esto puede ocurrir y así anticipar con un plan de acción para mitigar estos casos, previniendo cantidad de heridos y muertes.

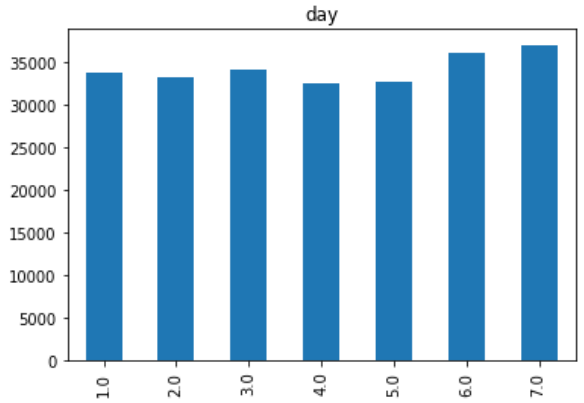
9. Referencias

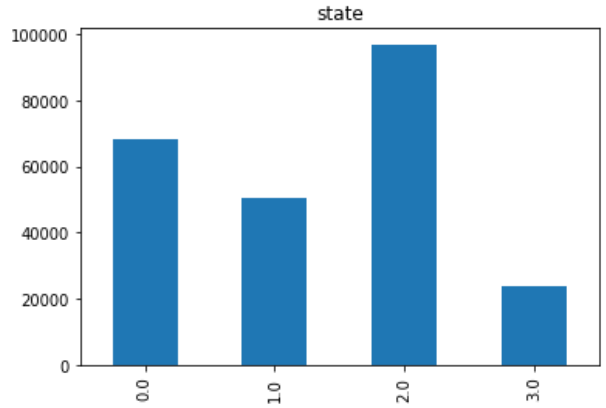
[1] *Gun Violence Data*. (2018, 15 abril). Kaggle. https://www.kaggle.com/jameslko/gun-violence-data

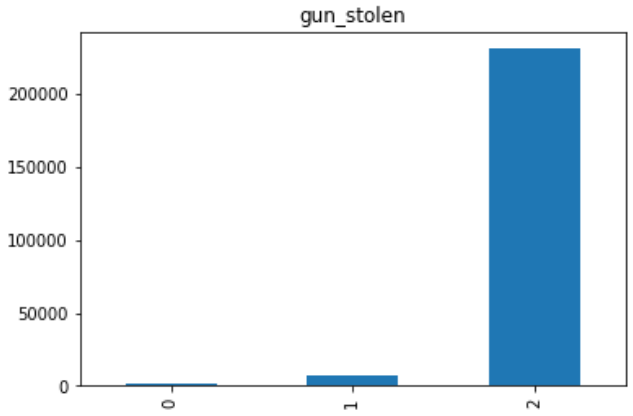
10. Anexos

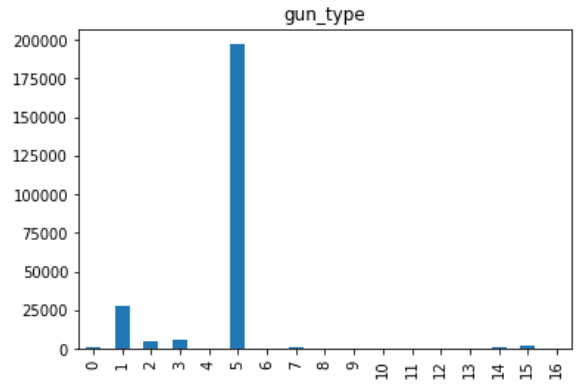
Histogramas de cada uno de los datos sin análisis de outliers

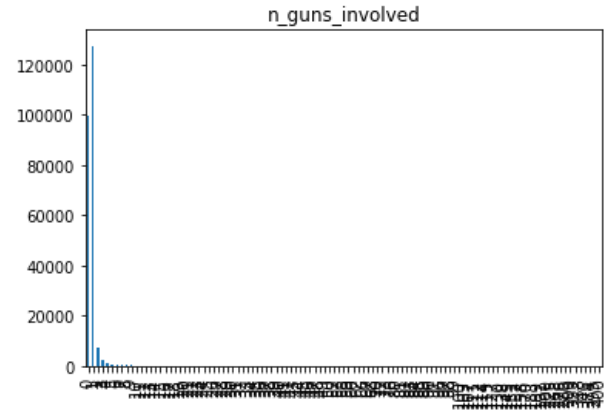


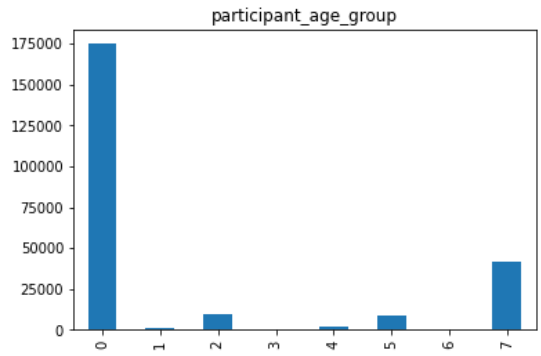


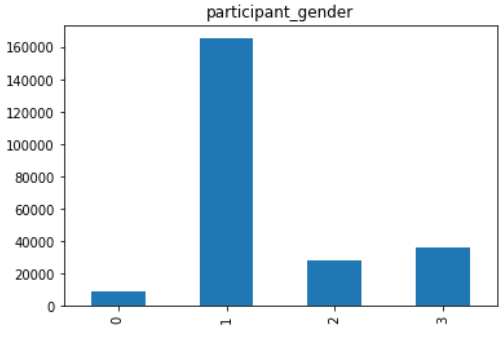


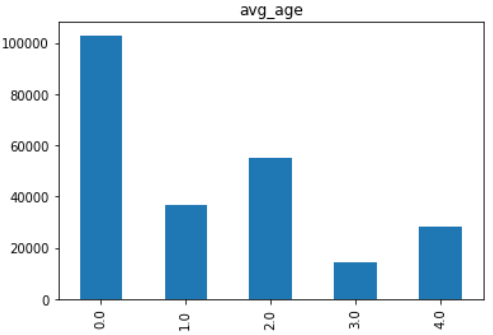






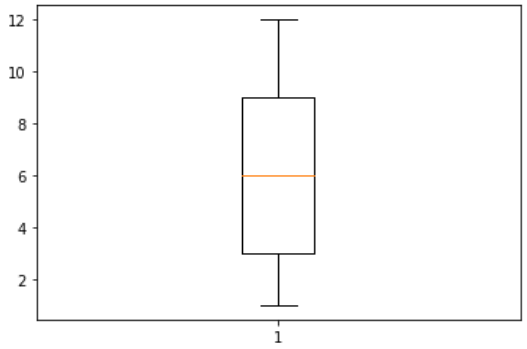




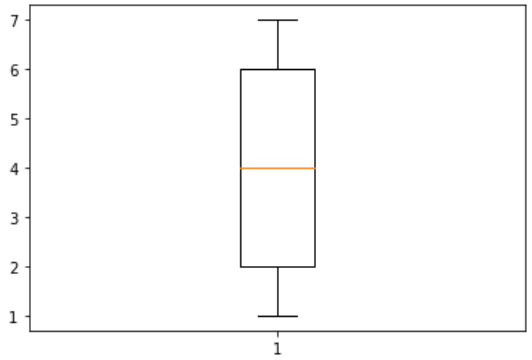


Boxplots de cada uno de los datos

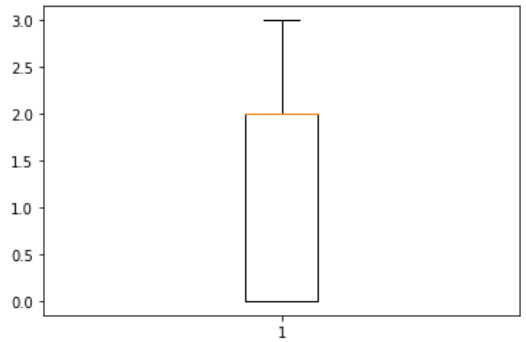
Month



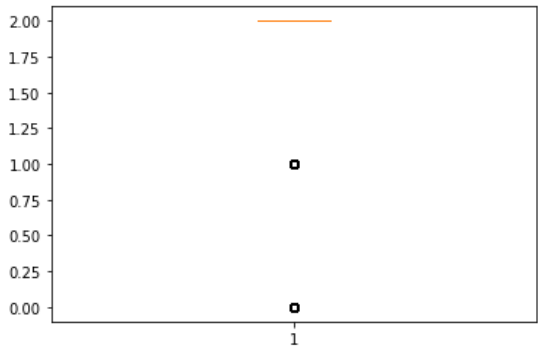
Day



State2

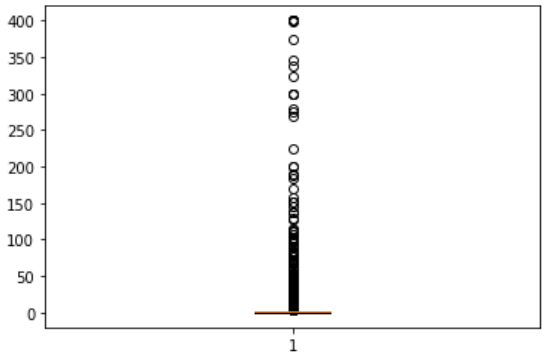


Gun_Stolen

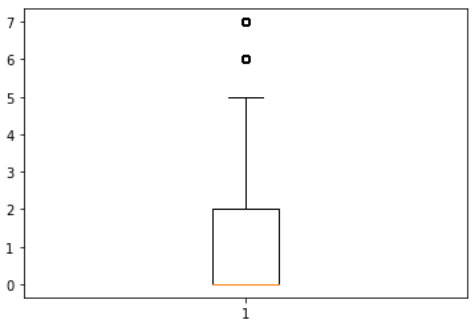


Gun_Type

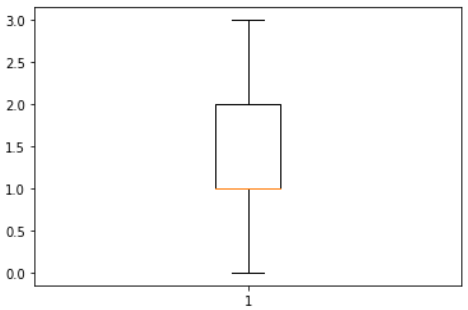
n_guns_involved



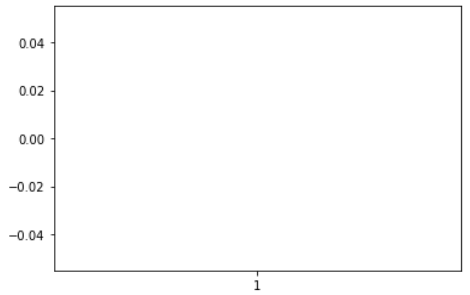
participant_age_group



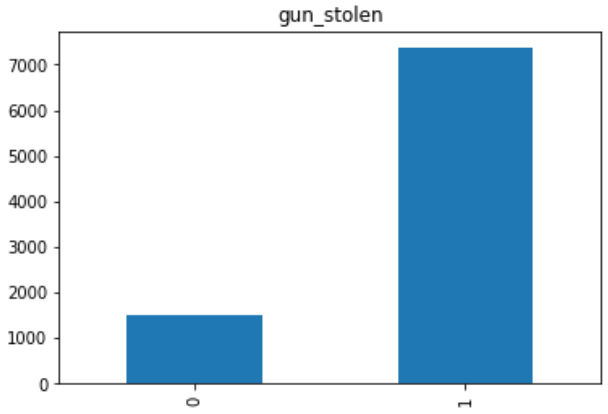
participant_gender

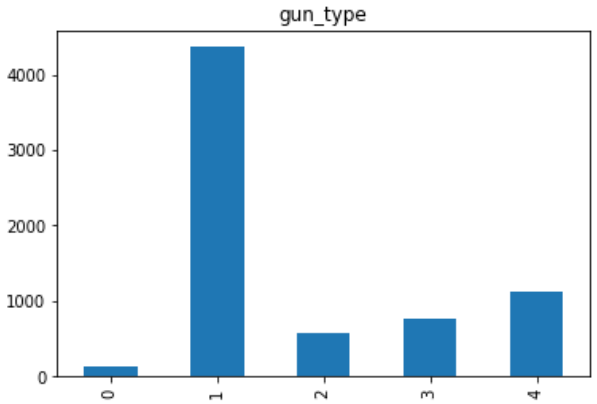


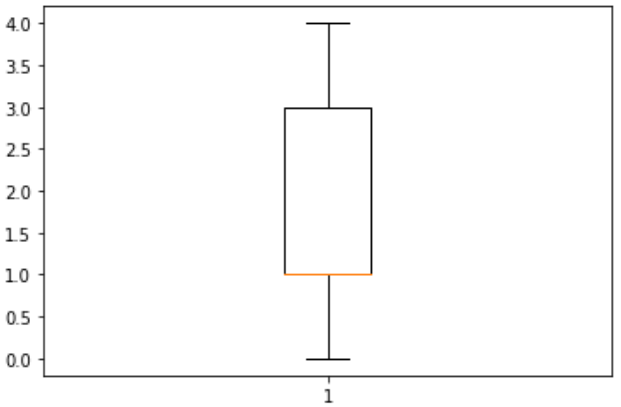
avg_age

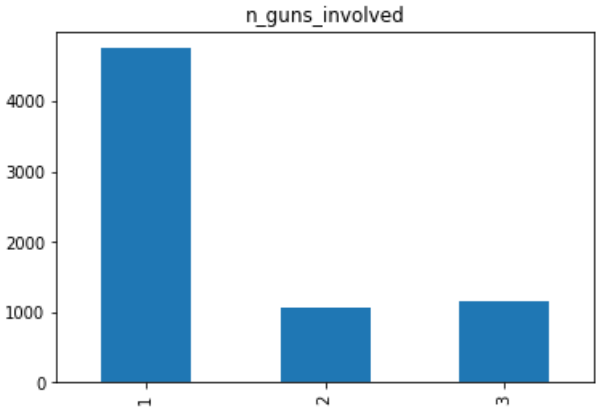


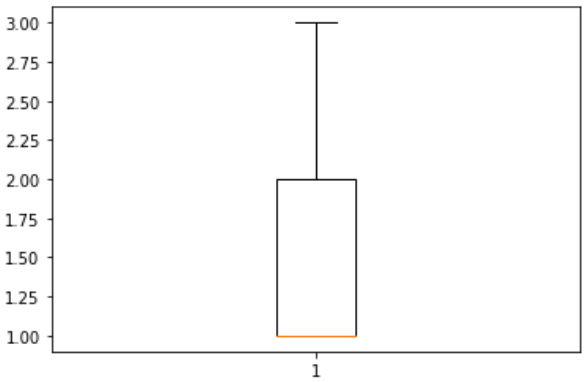
Histogramas luego de la identificación de outliers y normalización de datos

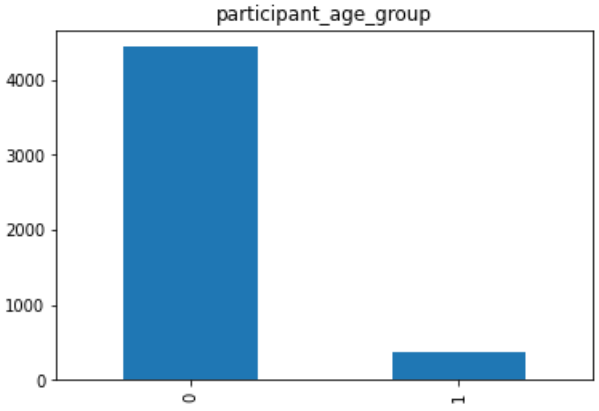


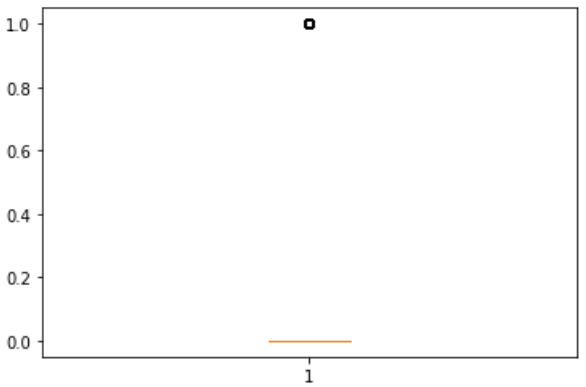








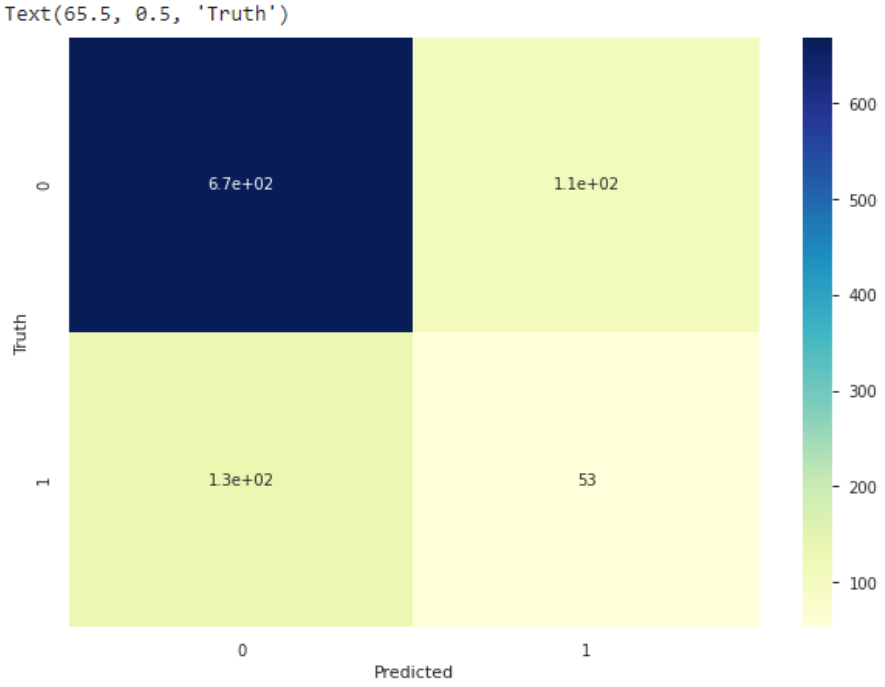




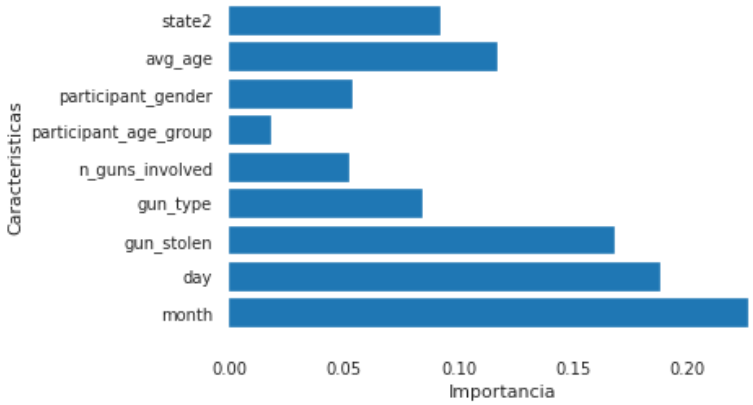
Datos de Gini con normalizar:

Matriz de confusión

```
array([[669, 106],
       [134,  53]])
```



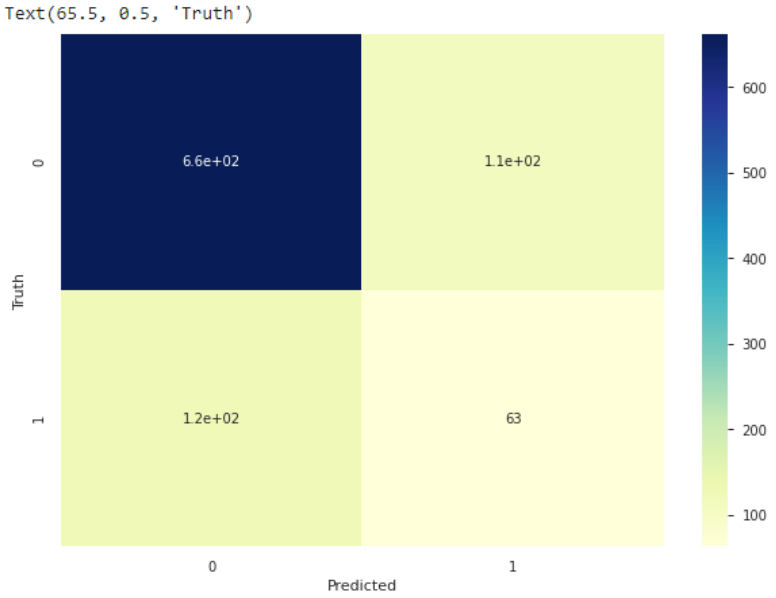
Importancia de las características



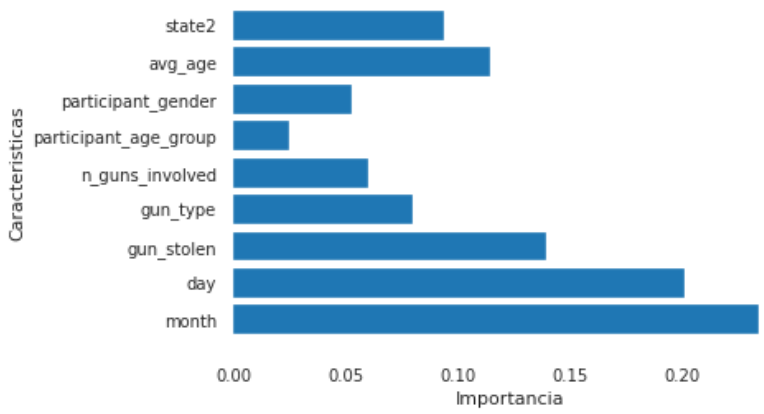
Datos de Entropía Cruzada con normalizar

Matriz de confusión

```
array([[662, 113],
       [124,  63]])
```



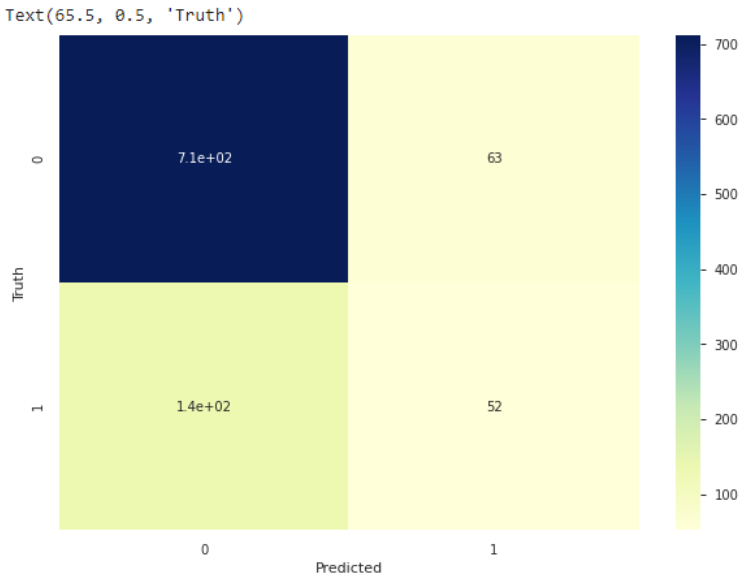
Importancia de las características



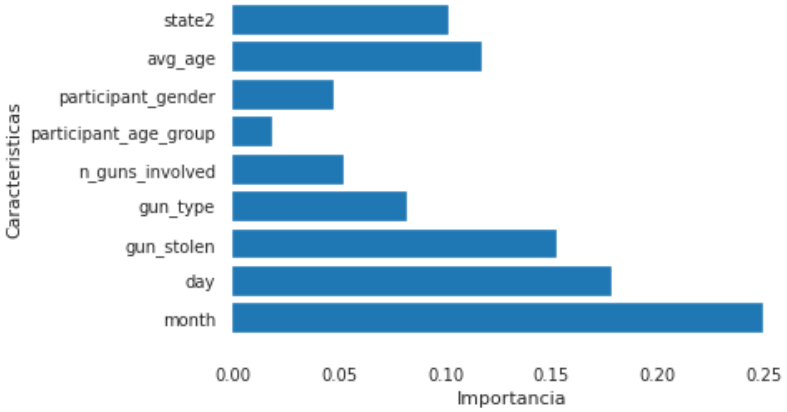
Datos de Random Forest con normalizar

Matriz de confusión

```
array([[712,  63],
       [135,  52]])
```



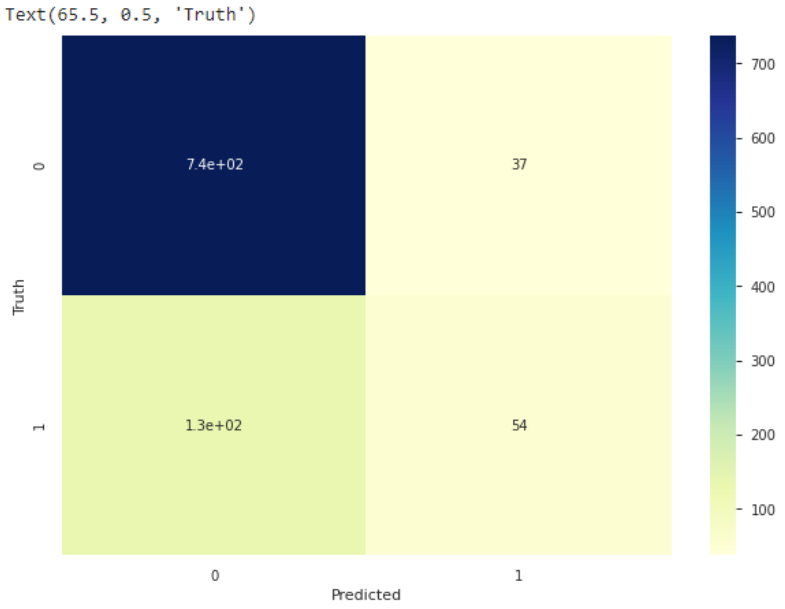
Importancia de las características



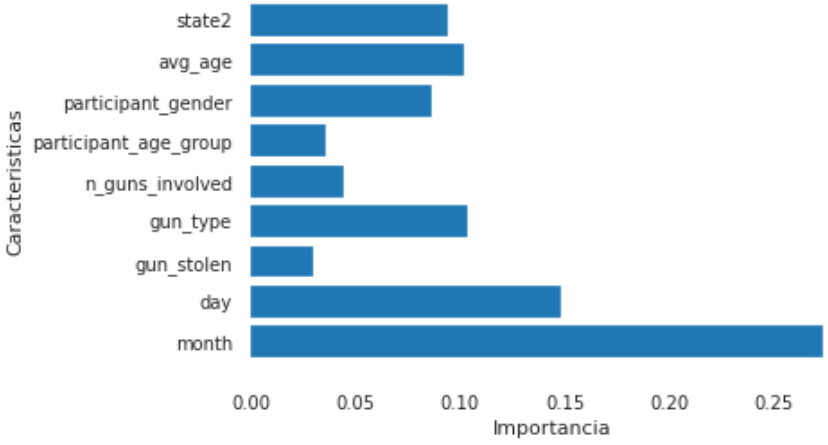
Datos de AdaBoost con normalizar

Matriz de confusión

```
array([[738,  37],
       [133,  54]])
```



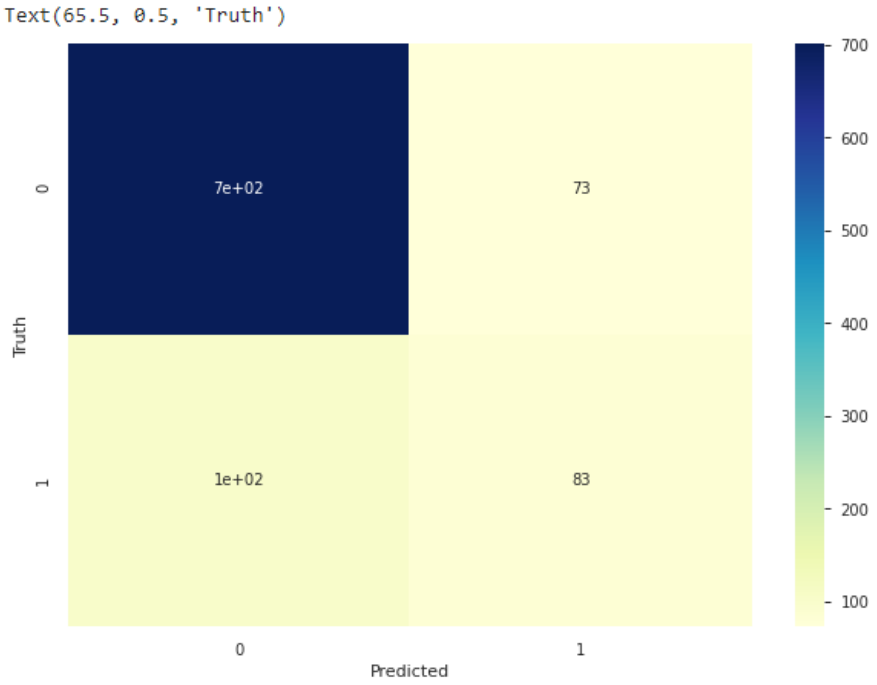
Importancia de las características



Datos de Clasificador Bayesiano Ingenuo con normalizar

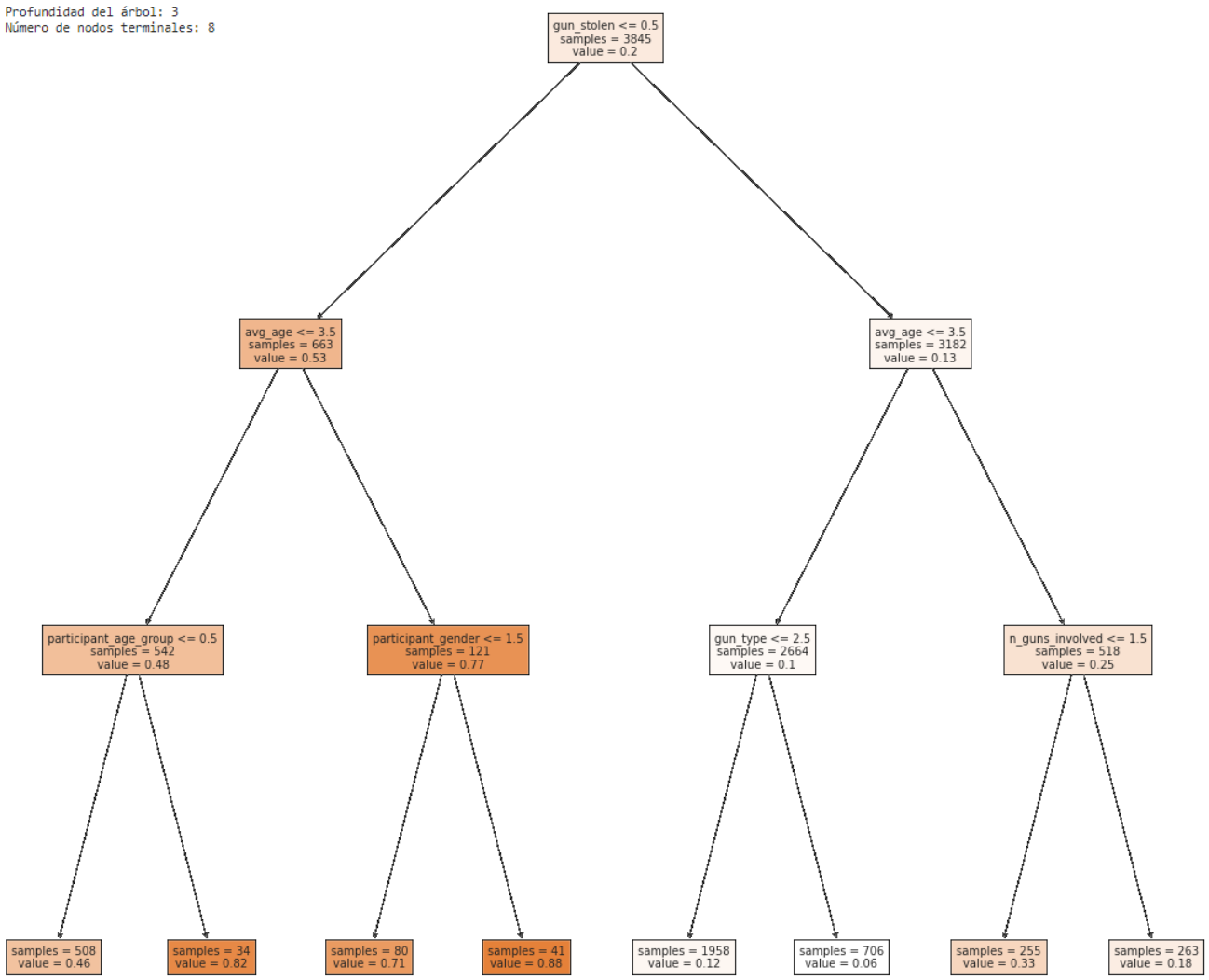
Matriz de confusión

```
array([[702,  73],
       [104,  83]])
```



Importancia de las características

Datos de árbol de decisión profundidad 3

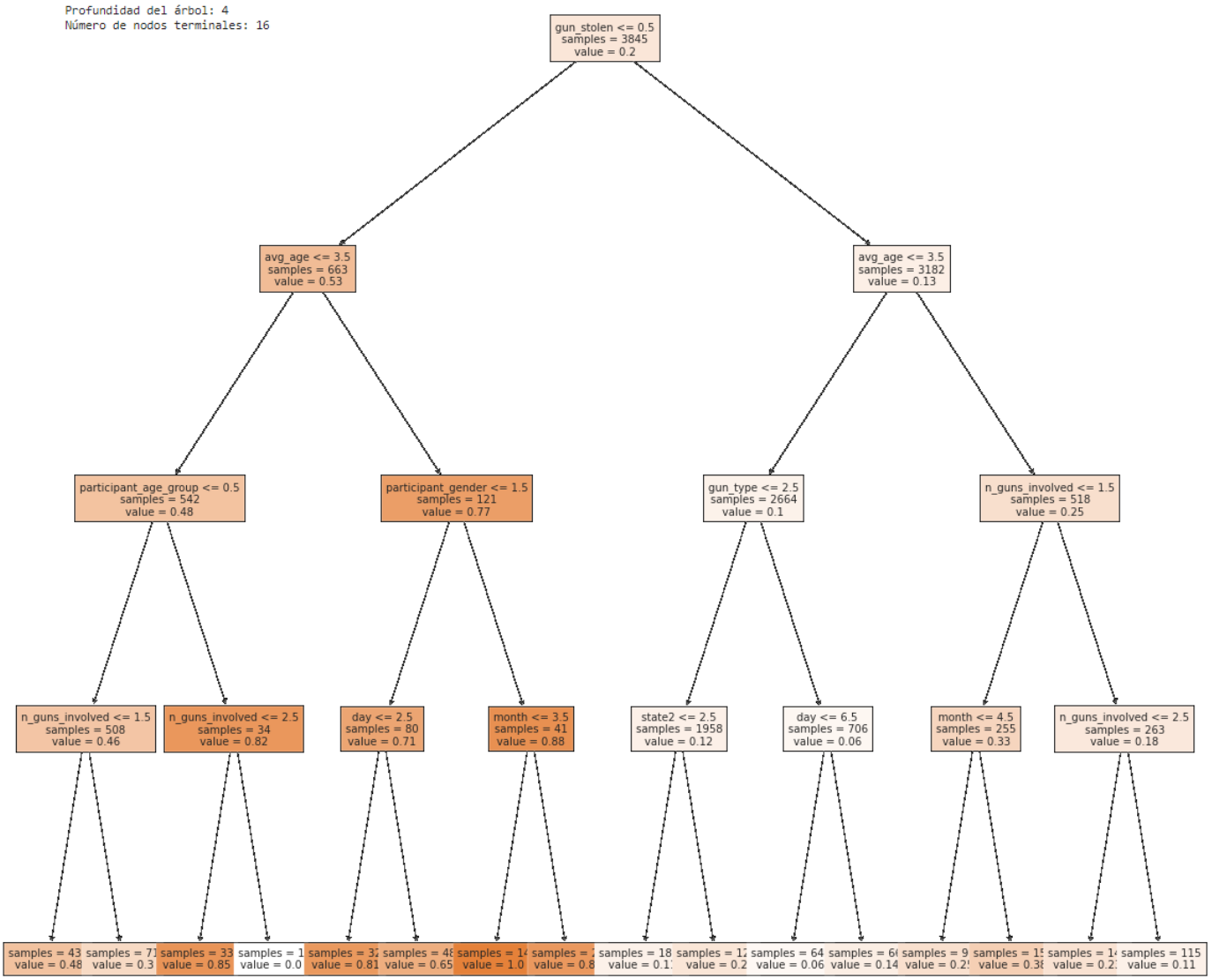



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- value: [0.46]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- value: [0.82]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- value: [0.71]
|   |   |--- participant_gender >  1.50
|   |   |   |--- value: [0.88]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- value: [0.12]
|   |   |--- gun_type >  2.50
|   |   |   |--- value: [0.06]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- value: [0.33]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- value: [0.18]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.766053
7	avg_age	0.152906
5	participant_age_group	0.036589
4	n_guns_involved	0.025089
3	gun_type	0.013024
6	participant_gender	0.006339
0	month	0.000000
1	day	0.000000
8	state2	0.000000

Datos de árbol de decisión profundidad 4

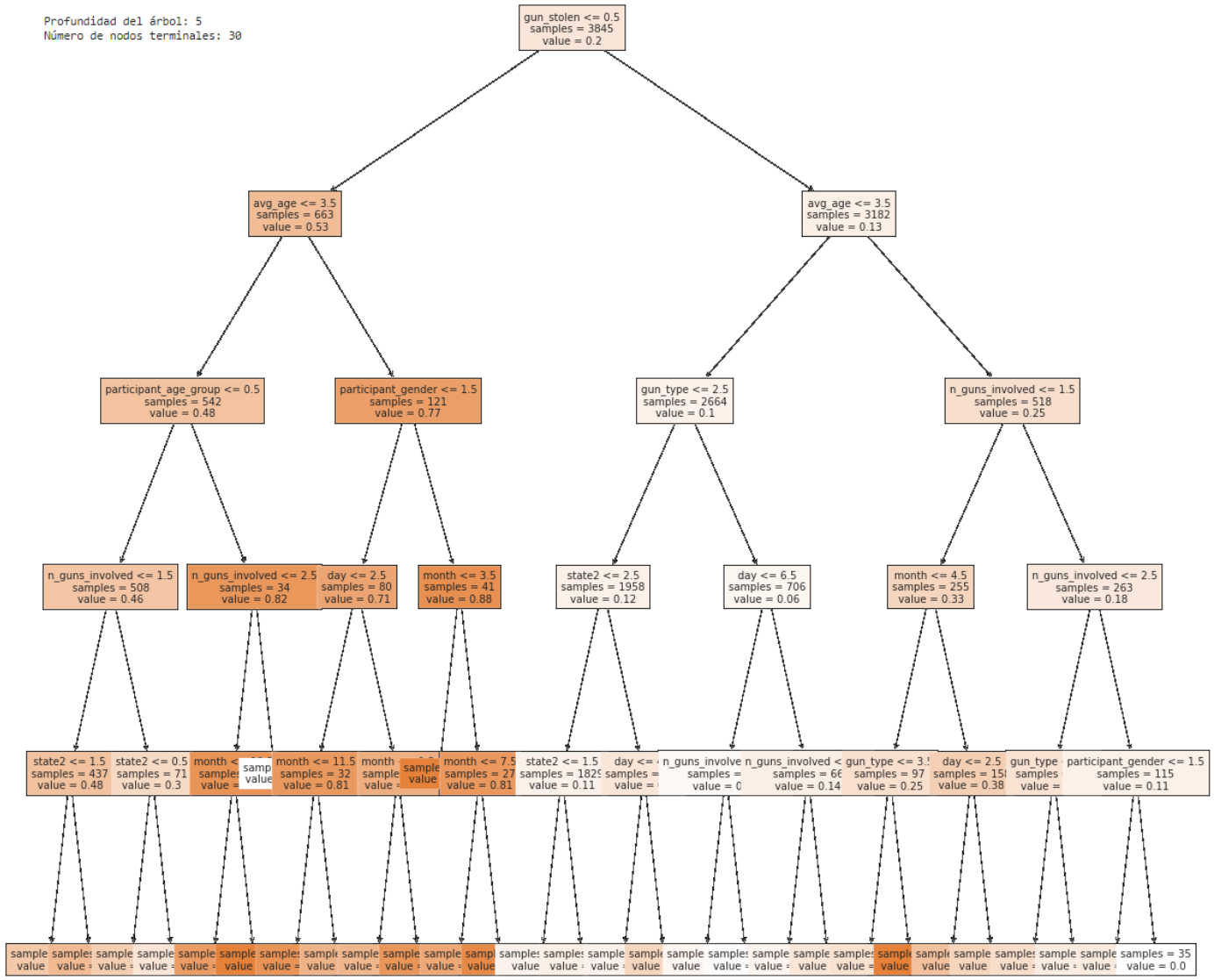



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- n_guns_involved <= 1.50
|   |   |   |   |--- value: [0.48]
|   |   |   |--- n_guns_involved >  1.50
|   |   |   |   |--- value: [0.30]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- value: [0.85]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- day <= 2.50
|   |   |   |   |--- value: [0.81]
|   |   |   |--- day >  2.50
|   |   |   |   |--- value: [0.65]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- value: [0.81]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- state2 <= 2.50
|   |   |   |   |--- value: [0.11]
|   |   |   |--- state2 >  2.50
|   |   |   |   |--- value: [0.20]
|   |   |--- gun_type >  2.50
|   |   |   |--- day <= 6.50
|   |   |   |   |--- value: [0.06]
|   |   |   |--- day >  6.50
|   |   |   |   |--- value: [0.14]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- month <= 4.50
|   |   |   |   |--- value: [0.25]
|   |   |   |--- month >  4.50
|   |   |   |   |--- value: [0.38]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- value: [0.23]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- value: [0.11]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.723068
7	avg_age	0.144326
4	n_guns_involved	0.053616
5	participant_age_group	0.034536
3	gun_type	0.012293
0	month	0.011021
8	state2	0.007768
1	day	0.007388
6	participant_gender	0.005983

Datos de árbol de decisión profundidad 5




```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- n_guns_involved <= 1.50
|   |   |   |   |--- state2 <= 1.50
|   |   |   |   |   |--- value: [0.43]
|   |   |   |   |--- state2 >  1.50
|   |   |   |   |   |--- value: [0.55]
|   |   |   |--- n_guns_involved >  1.50
|   |   |   |   |--- state2 <= 0.50
|   |   |   |   |   |--- value: [0.39]
|   |   |   |   |--- state2 >  0.50
|   |   |   |   |   |--- value: [0.21]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- month <= 10.50
|   |   |   |   |   |--- value: [0.78]
|   |   |   |   |--- month >  10.50
|   |   |   |   |   |--- value: [1.00]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- day <= 2.50
|   |   |   |   |--- month <= 11.50
|   |   |   |   |   |--- value: [0.85]
|   |   |   |   |--- month >  11.50
|   |   |   |   |   |--- value: [0.60]
|   |   |   |--- day >  2.50
|   |   |   |   |--- month <= 9.50
|   |   |   |   |   |--- value: [0.58]
|   |   |   |   |--- month >  9.50
|   |   |   |   |   |--- value: [0.83]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- month <= 7.50
|   |   |   |   |   |--- value: [0.69]
|   |   |   |   |--- month >  7.50
|   |   |   |   |   |--- value: [0.93]
```

```
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- state2 <= 2.50
|   |   |   |   |--- state2 <= 1.50
|   |   |   |   |   |--- value: [0.10]
|   |   |   |   |--- state2 >  1.50
|   |   |   |   |   |--- value: [0.13]
|   |   |   |--- state2 >  2.50
|   |   |   |   |--- day <= 4.50
|   |   |   |   |   |--- value: [0.11]
|   |   |   |   |--- day >  4.50
|   |   |   |   |   |--- value: [0.33]
|   |   |--- gun_type >  2.50
|   |   |   |--- day <= 6.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.08]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.03]
|   |   |   |--- day >  6.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.08]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.20]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- month <= 4.50
|   |   |   |   |--- gun_type <= 3.50
|   |   |   |   |   |--- value: [0.24]
|   |   |   |   |--- gun_type >  3.50
|   |   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  4.50
|   |   |   |   |--- day <= 2.50
|   |   |   |   |   |--- value: [0.50]
|   |   |   |   |--- day >  2.50
|   |   |   |   |   |--- value: [0.34]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- gun_type >  3.50
|   |   |   |   |   |--- value: [0.15]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- participant_gender <= 1.50
|   |   |   |   |   |--- value: [0.16]
|   |   |   |   |--- participant_gender >  1.50
|   |   |   |   |   |--- value: [0.00]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.676180
7	avg_age	0.134967
4	n_guns_involved	0.055010
5	participant_age_group	0.032296
8	state2	0.025492
1	day	0.024552
0	month	0.021873
3	gun_type	0.019192
6	participant_gender	0.010437

Datos de Adaboost profundidad 3, 4, 5 (Importancia de los predictores en el modelo).

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.676830
7	avg_age	0.135146
6	participant_gender	0.061673
4	n_guns_involved	0.051130
3	gun_type	0.045670
5	participant_age_group	0.027278
1	day	0.002273
0	month	0.000000
8	state2	0.000000

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.541014
7	avg_age	0.133566
6	participant_gender	0.090740
4	n_guns_involved	0.058121
3	gun_type	0.056397
1	day	0.041551
0	month	0.036736
8	state2	0.024309
5	participant_age_group	0.017565

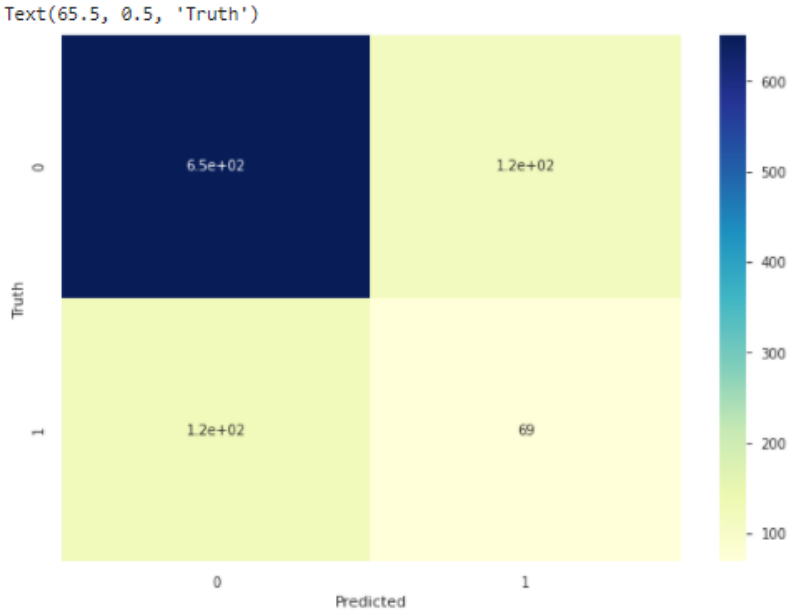
Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.373705
7	avg_age	0.123015
0	month	0.101418
6	participant_gender	0.095719
1	day	0.084811
3	gun_type	0.081013
4	n_guns_involved	0.065991
8	state2	0.055311
5	participant_age_group	0.019016

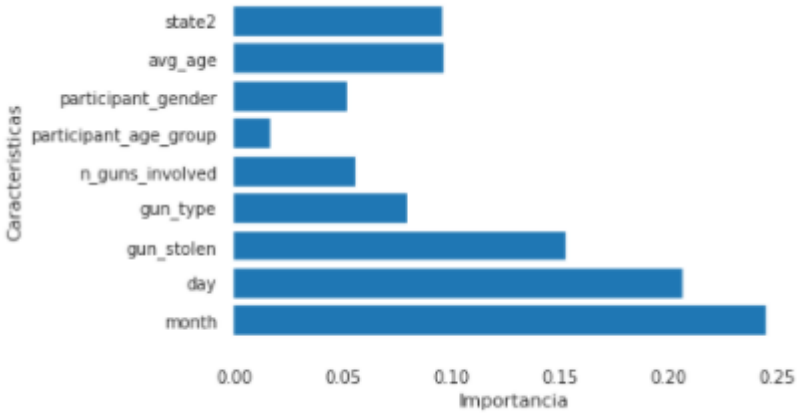
Datos de Gini con “*MaxMinScale*”:

Matriz de confusión

```
array([[651, 117],
       [125,  69]])
```



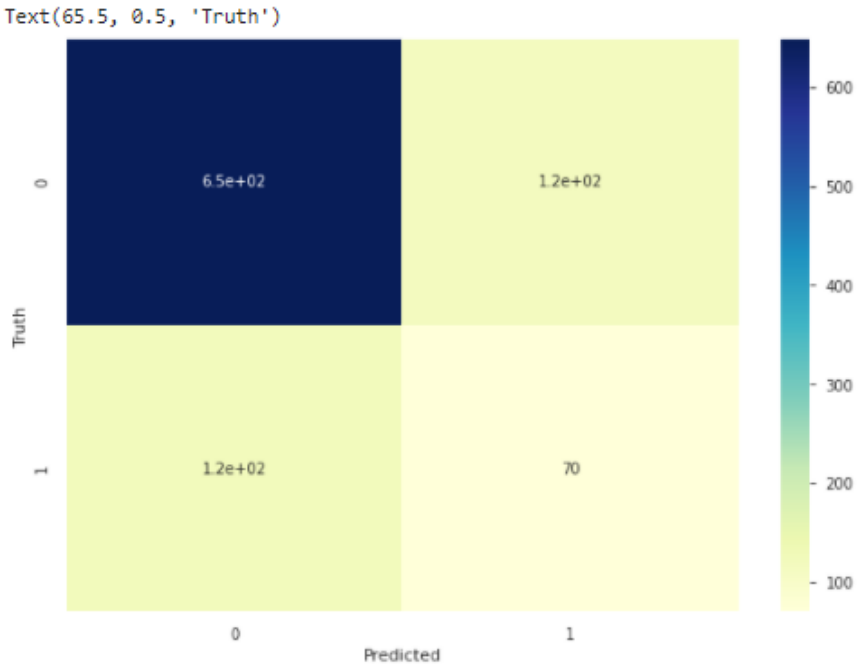
Importancia de las características



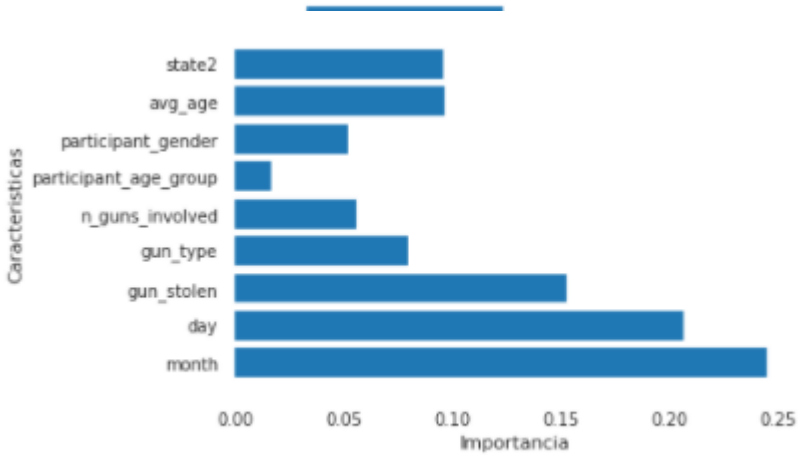
Datos de Entropía Cruzada con “*MaxMinScale*”:

Matriz de confusión

```
array([[649, 119],
       [124,  70]])
```



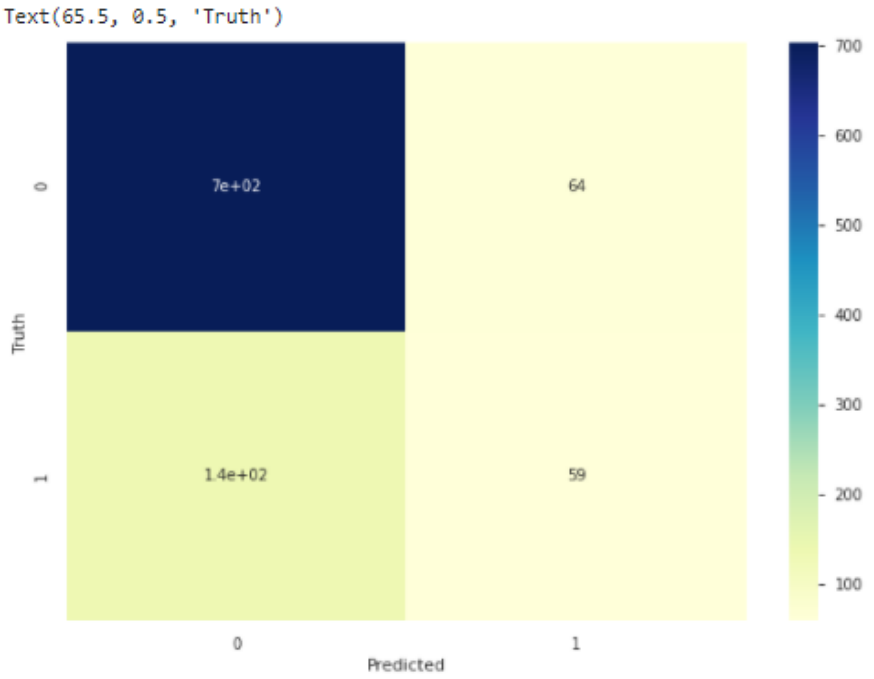
Importancia de las características



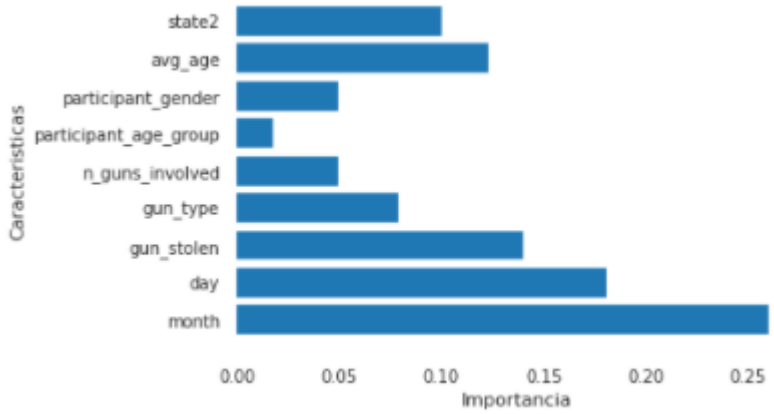
Datos de Random Forest con “*MaxMinScale*”:

Matriz de confusión

```
array([[704,  64],
       [135,  59]])
```



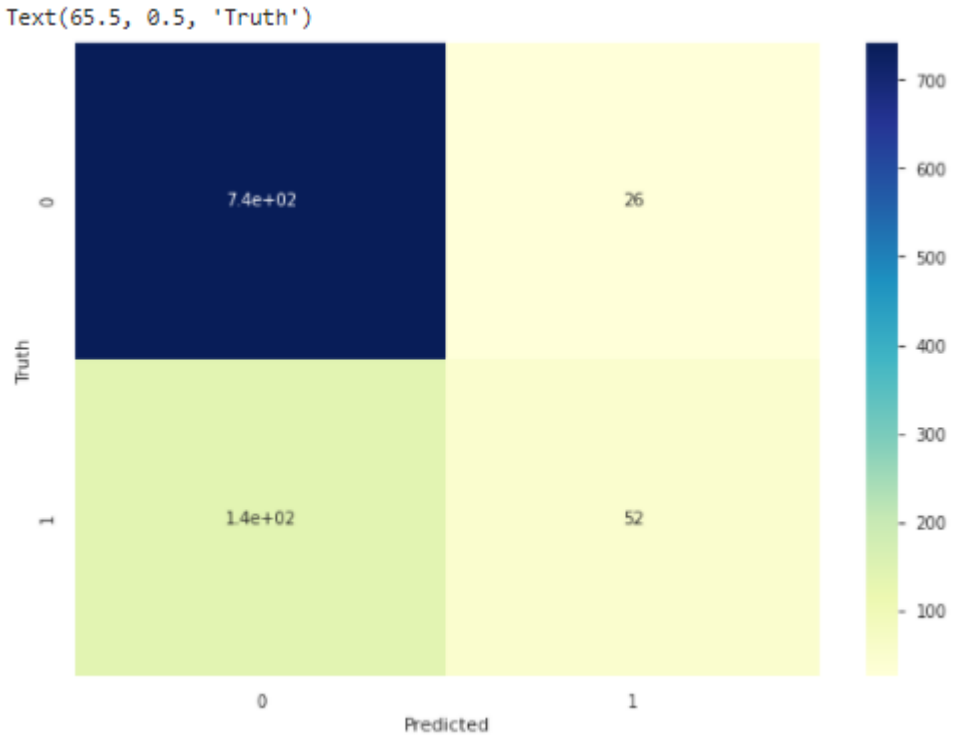
Importancia de las características



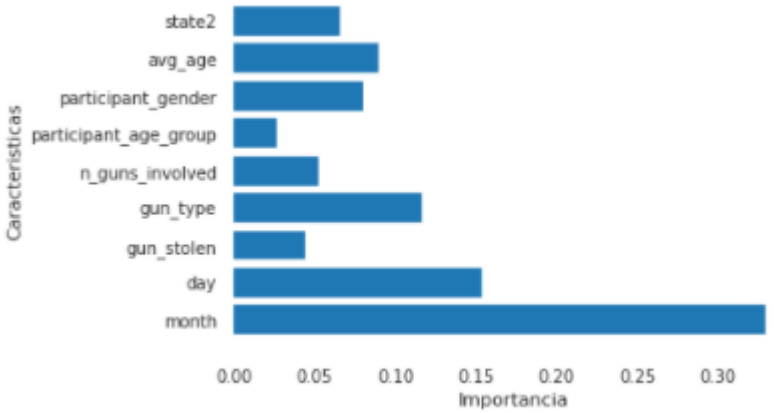
Datos de AdaBoost con “*MaxMinScale*”:

Matriz de confusión

```
array([[742,  26],
       [142,  52]])
```



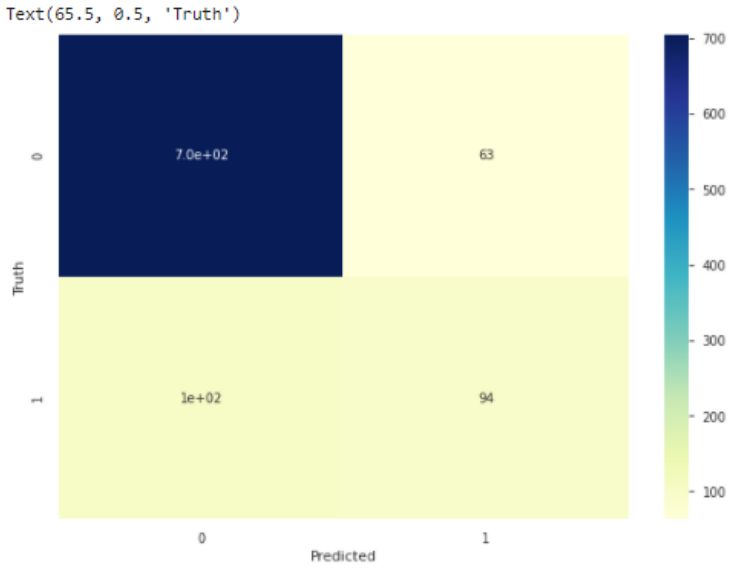
Importancia de las características



Datos de Clasificador Bayesiano Ingenuo con “*MaxMinScale*”:

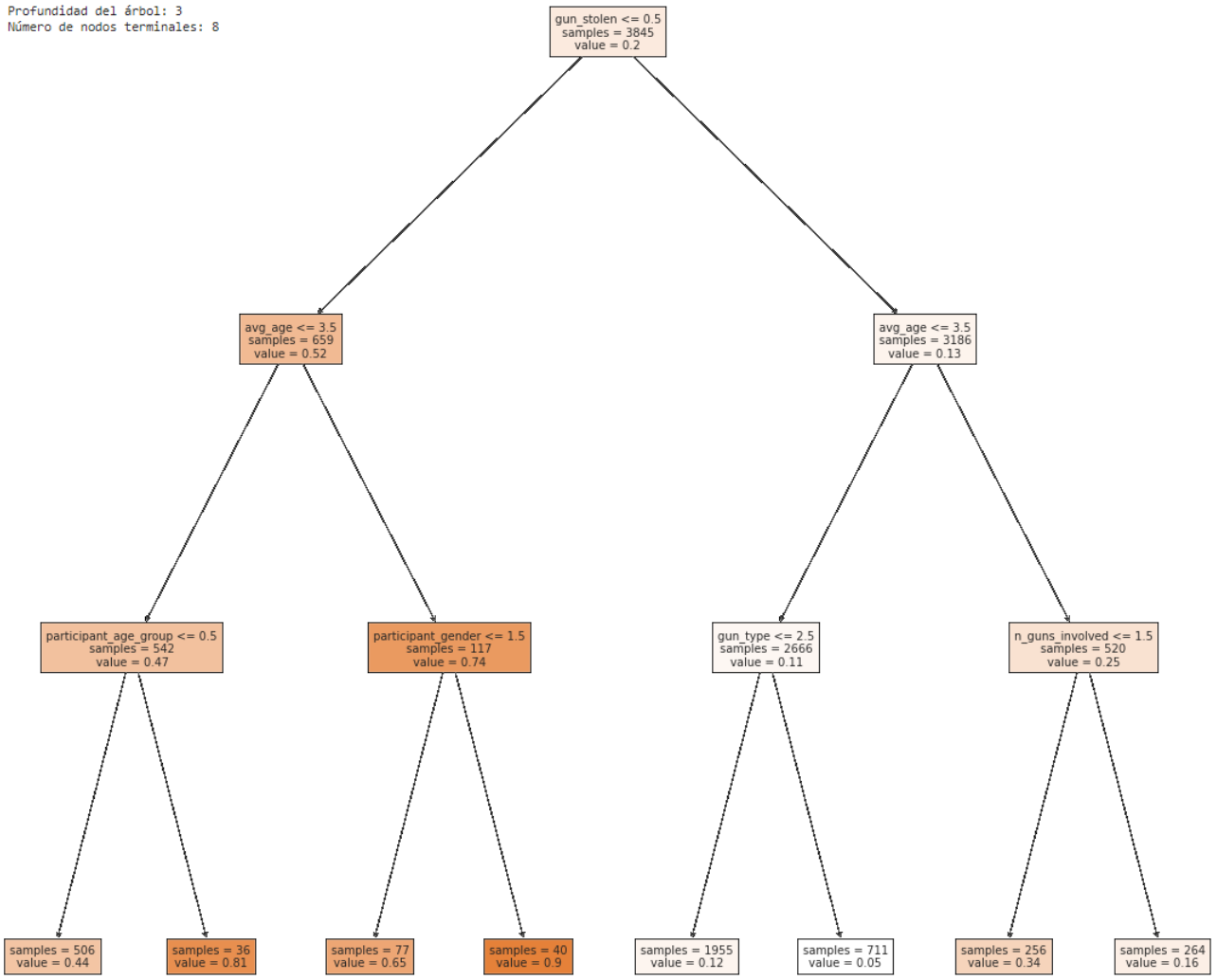
Matriz de confusión

```
array([[705,  63],
       [100,  94]])
```



Importancia de las características

Datos de árbol de decisión profundidad 3

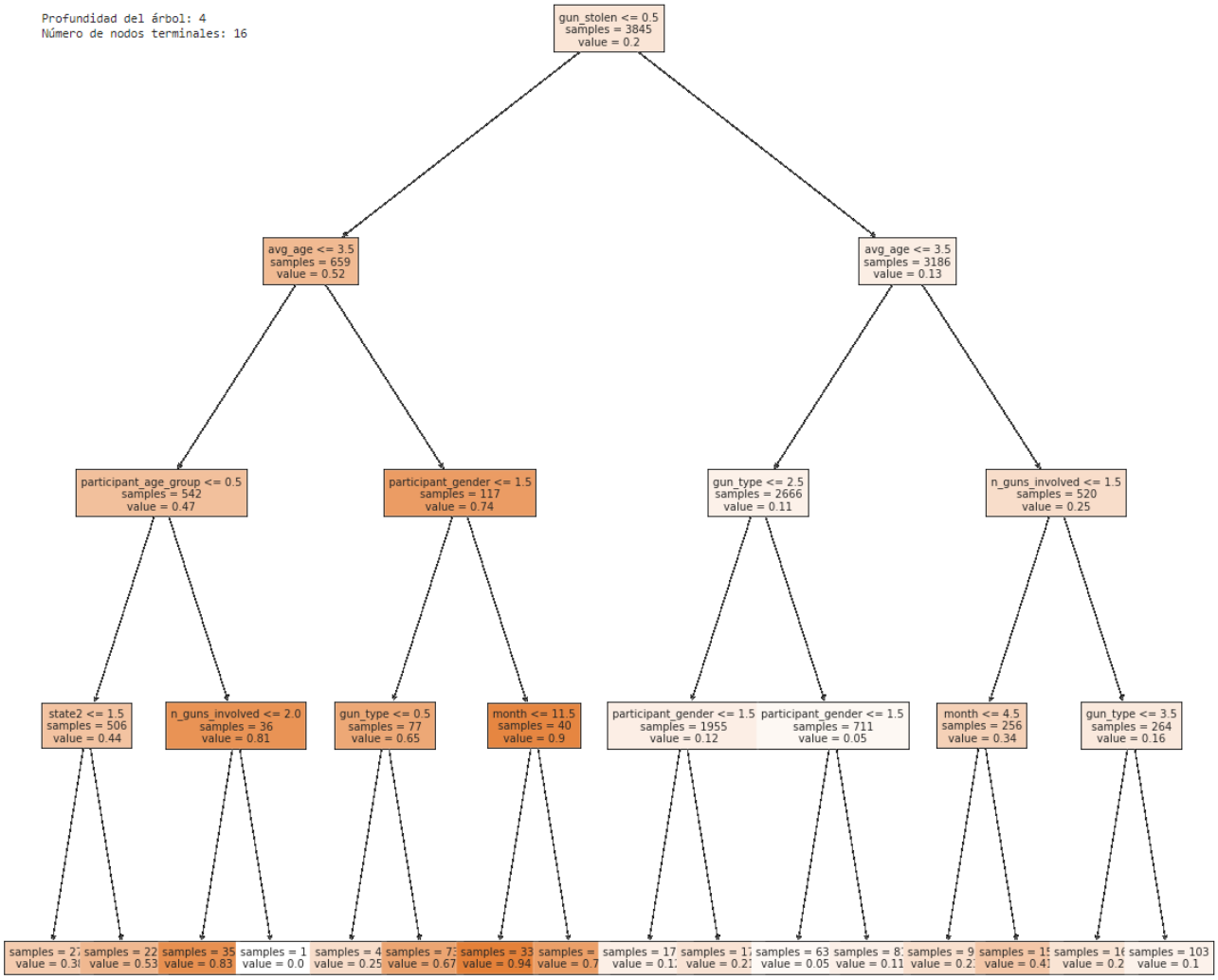



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- value: [0.44]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- value: [0.81]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- value: [0.65]
|   |   |--- participant_gender >  1.50
|   |   |   |--- value: [0.90]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- value: [0.12]
|   |   |--- gun_type >  2.50
|   |   |   |--- value: [0.05]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- value: [0.34]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- value: [0.16]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.738056
7	avg_age	0.146034
5	participant_age_group	0.039646
4	n_guns_involved	0.038510
3	gun_type	0.022773
6	participant_gender	0.014979
0	month	0.000000
1	day	0.000000
8	state2	0.000000

Datos de árbol de decisión profundidad 4

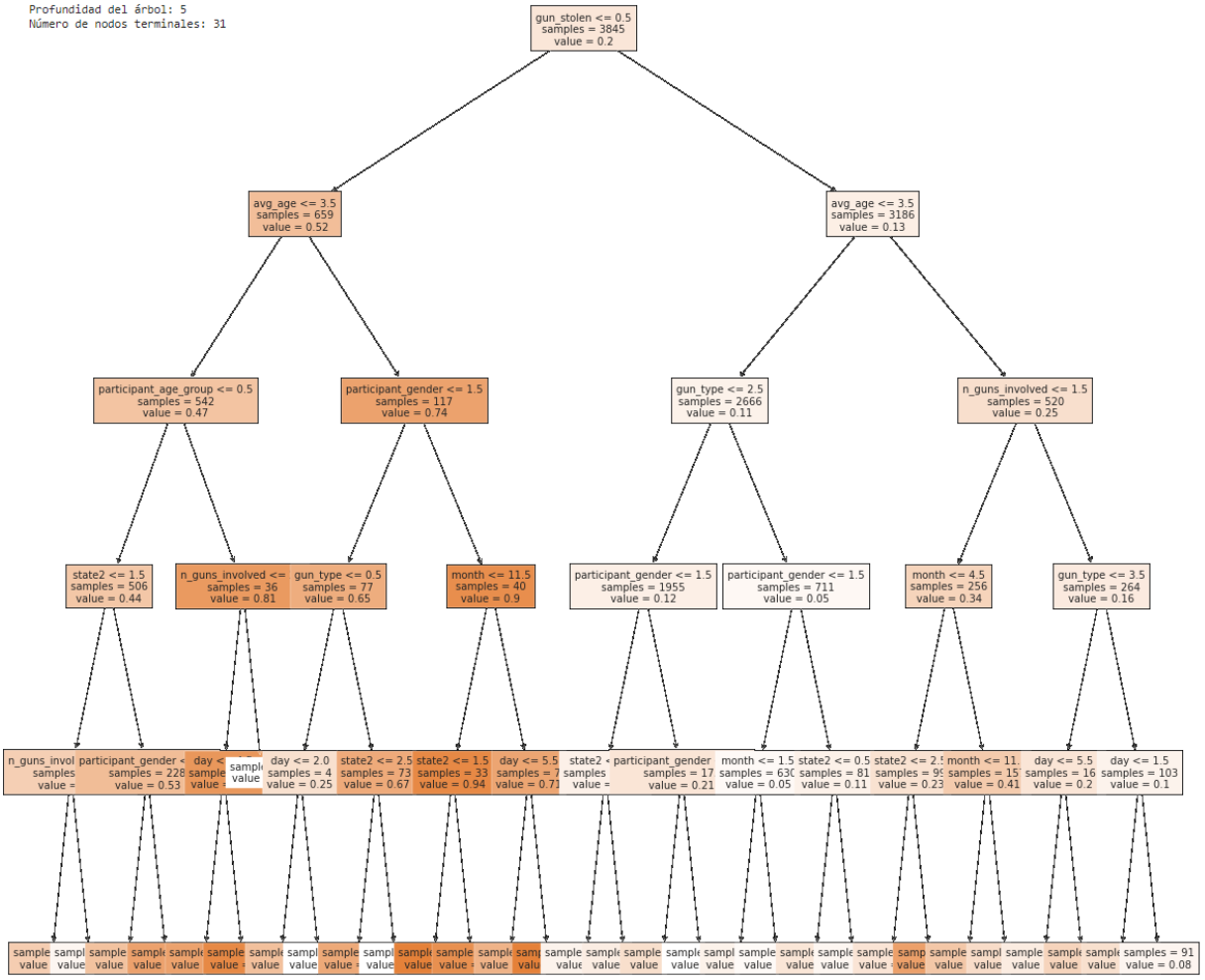



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- state2 <= 1.50
|   |   |   |   |--- value: [0.38]
|   |   |   |--- state2 >  1.50
|   |   |   |   |--- value: [0.53]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- value: [0.83]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- gun_type <= 0.50
|   |   |   |   |--- value: [0.25]
|   |   |   |--- gun_type >  0.50
|   |   |   |   |--- value: [0.67]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 11.50
|   |   |   |   |--- value: [0.94]
|   |   |   |--- month >  11.50
|   |   |   |   |--- value: [0.71]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.12]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.21]
|   |   |--- gun_type >  2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.05]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.11]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- month <= 4.50
|   |   |   |   |--- value: [0.23]
|   |   |   |--- month >  4.50
|   |   |   |   |--- value: [0.41]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- gun_type <= 3.50
|   |   |   |   |--- value: [0.20]
|   |   |   |--- gun_type >  3.50
|   |   |   |   |--- value: [0.10]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.684155
7	avg_age	0.135369
4	n_guns_involved	0.041302
5	participant_age_group	0.036751
3	gun_type	0.032897
6	participant_gender	0.027012
8	state2	0.023229
0	month	0.019284
1	day	0.000000

Datos de árbol de decisión profundidad 5




```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- state2 <= 1.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.39]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.07]
|   |   |   |--- state2 >  1.50
|   |   |   |   |--- participant_gender <= 1.50
|   |   |   |   |   |--- value: [0.49]
|   |   |   |   |--- participant_gender >  1.50
|   |   |   |   |   |--- value: [0.74]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- day <= 4.50
|   |   |   |   |   |--- value: [0.74]
|   |   |   |   |--- day >  4.50
|   |   |   |   |   |--- value: [0.94]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- gun_type <= 0.50
|   |   |   |   |--- day <= 2.00
|   |   |   |   |   |--- value: [0.50]
|   |   |   |   |--- day >  2.00
|   |   |   |   |   |--- value: [0.00]
|   |   |   |--- gun_type >  0.50
|   |   |   |   |--- state2 <= 2.50
|   |   |   |   |   |--- value: [0.68]
|   |   |   |   |--- state2 >  2.50
|   |   |   |   |   |--- value: [0.00]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 11.50
|   |   |   |   |--- state2 <= 1.50
|   |   |   |   |   |--- value: [1.00]
|   |   |   |   |--- state2 >  1.50
|   |   |   |   |   |--- value: [0.88]
|   |   |   |--- month >  11.50
|   |   |   |   |--- day <= 5.50
|   |   |   |   |   |--- value: [0.60]
|   |   |   |   |--- day >  5.50
|   |   |   |   |   |--- value: [1.00]
|--- gun_stolen >  0.50
```

```
|--- avg_age <= 3.50
|   |--- gun_type <= 2.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- state2 <= 1.50
|   |   |   |   |--- value: [0.10]
|   |   |   |--- state2 >  1.50
|   |   |   |   |--- value: [0.14]
|   |   |--- participant_gender >  1.50
|   |   |   |--- participant_gender <= 2.50
|   |   |   |   |--- value: [0.26]
|   |   |   |--- participant_gender >  2.50
|   |   |   |   |--- value: [0.00]
|   |--- gun_type >  2.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- month <= 1.50
|   |   |   |   |--- value: [0.11]
|   |   |   |--- month >  1.50
|   |   |   |   |--- value: [0.04]
|   |   |--- participant_gender >  1.50
|   |   |   |--- state2 <= 0.50
|   |   |   |   |--- value: [0.20]
|   |   |   |--- state2 >  0.50
|   |   |   |   |--- value: [0.07]
|--- avg_age >  3.50
|   |--- n_guns_involved <= 1.50
|   |   |--- month <= 4.50
|   |   |   |--- state2 <= 2.50
|   |   |   |   |--- value: [0.21]
|   |   |   |--- state2 >  2.50
|   |   |   |   |--- value: [0.75]
|   |   |--- month >  4.50
|   |   |   |--- month <= 11.50
|   |   |   |   |--- value: [0.45]
|   |   |   |--- month >  11.50
|   |   |   |   |--- value: [0.28]
|   |--- n_guns_involved >  1.50
|   |   |--- gun_type <= 3.50
|   |   |   |--- day <= 5.50
|   |   |   |   |--- value: [0.15]
|   |   |   |--- day >  5.50
|   |   |   |   |--- value: [0.34]
|   |   |--- gun_type >  3.50
|   |   |   |--- day <= 1.50
|   |   |   |   |--- value: [0.25]
|   |   |   |--- day >  1.50
|   |   |   |   |--- value: [0.08]
```

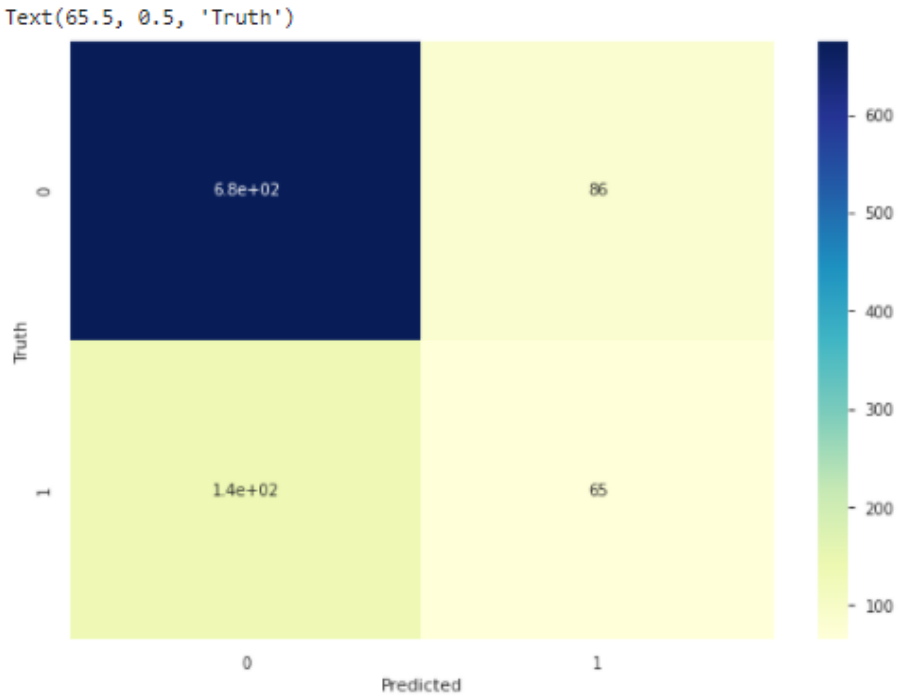
Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.625699
7	avg_age	0.123803
6	participant_gender	0.053302
4	n_guns_involved	0.048391
8	state2	0.042589
5	participant_age_group	0.033611
3	gun_type	0.030087
0	month	0.025129
1	day	0.017390

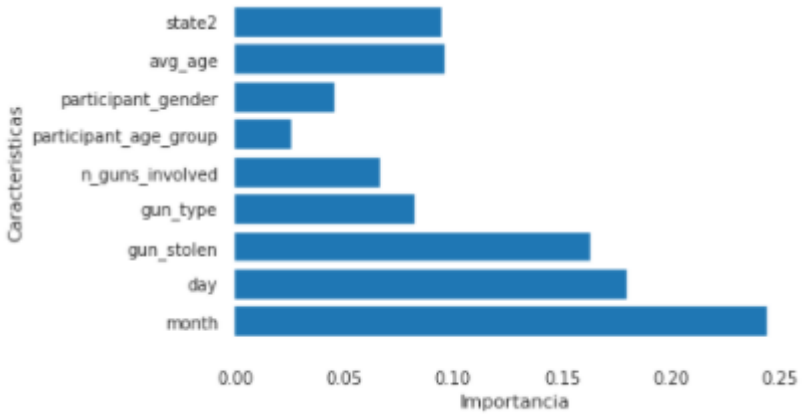
Datos de Gini con “*Scale*”:

Matriz de confusión

```
array([[675,  86],
       [136,  65]])
```



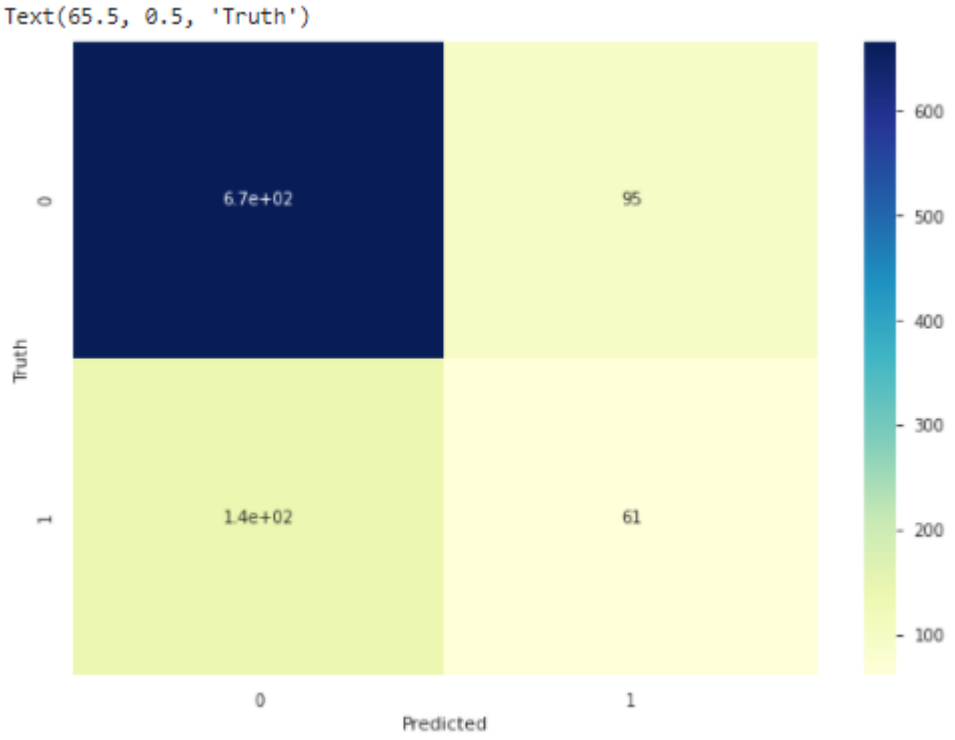
Importancia de las características



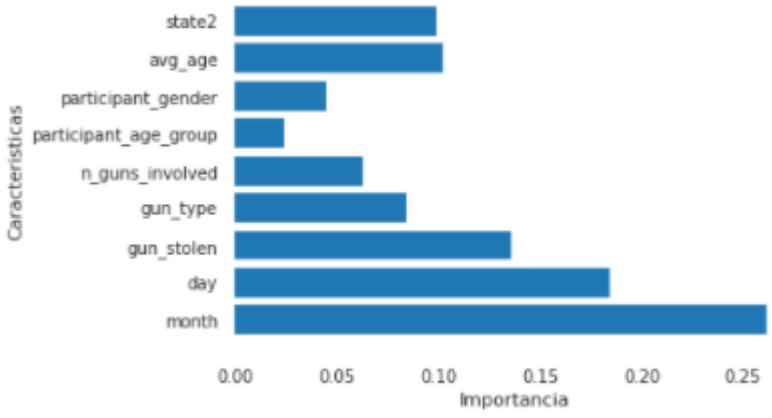
Datos de Entropía Cruzada con “*Scale*”:

Matriz de confusión

```
array([[666,  95],
       [140,  61]])
```



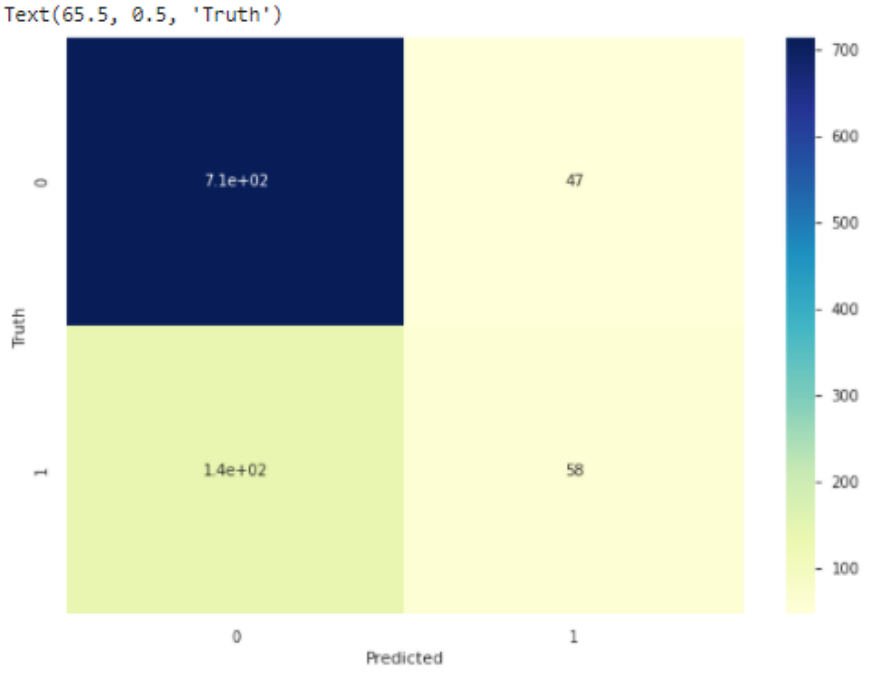
Importancia de las características



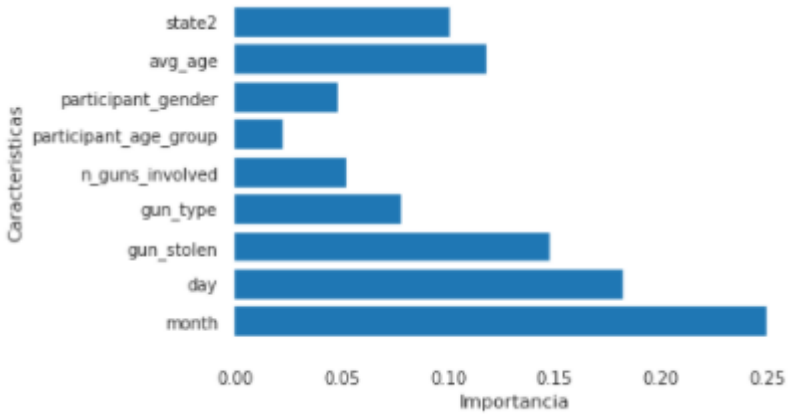
Datos de Random Forest con “*Scale*”:

Matriz de confusión

```
array([[714,  47],
       [143,  58]])
```



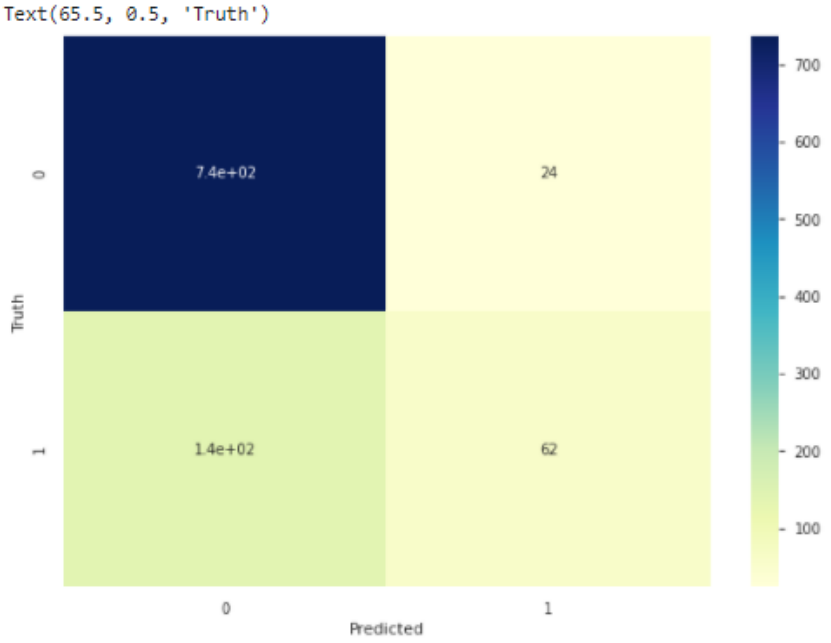
Importancia de las características



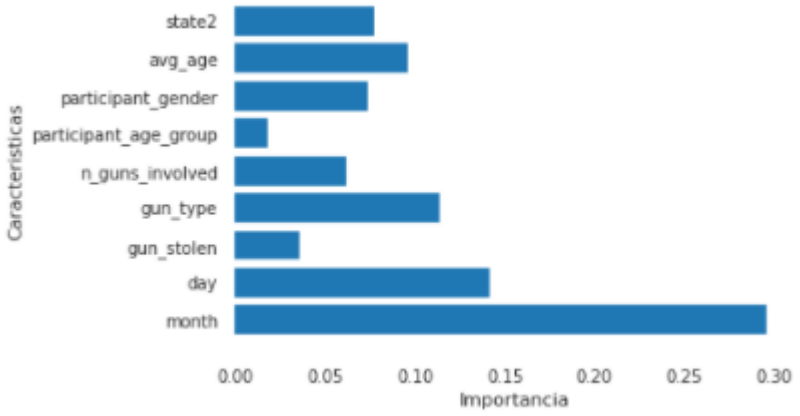
Datos de AdaBoost con “*Scale*”:

Matriz de confusión

```
array([[737,  24],
       [139,  62]])
```



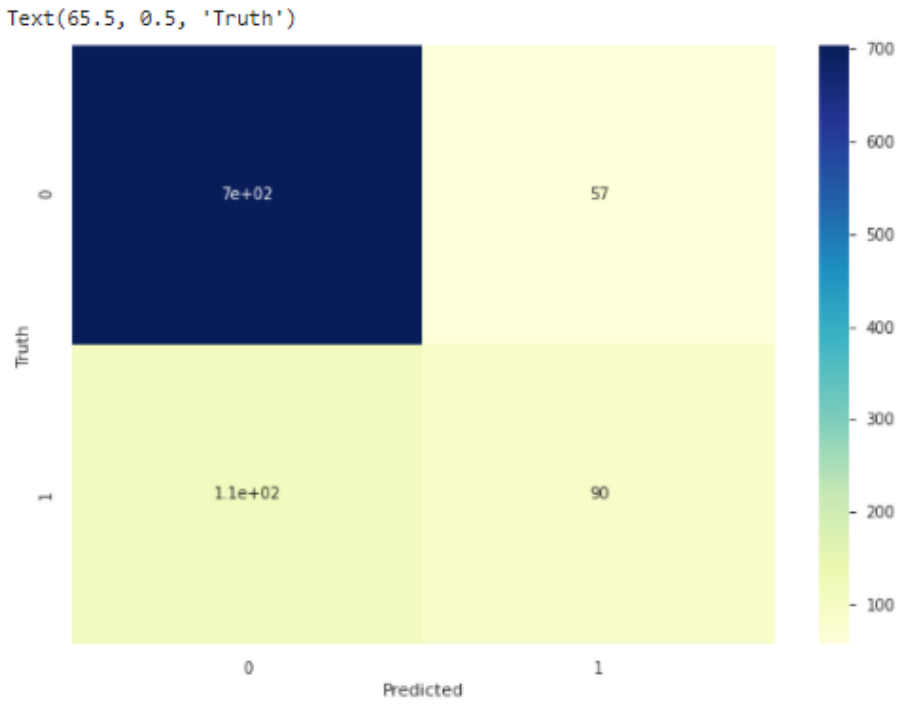
Importancia de las características



Datos de Clasificador Bayesiano Ingenuo con “*Scale*”:

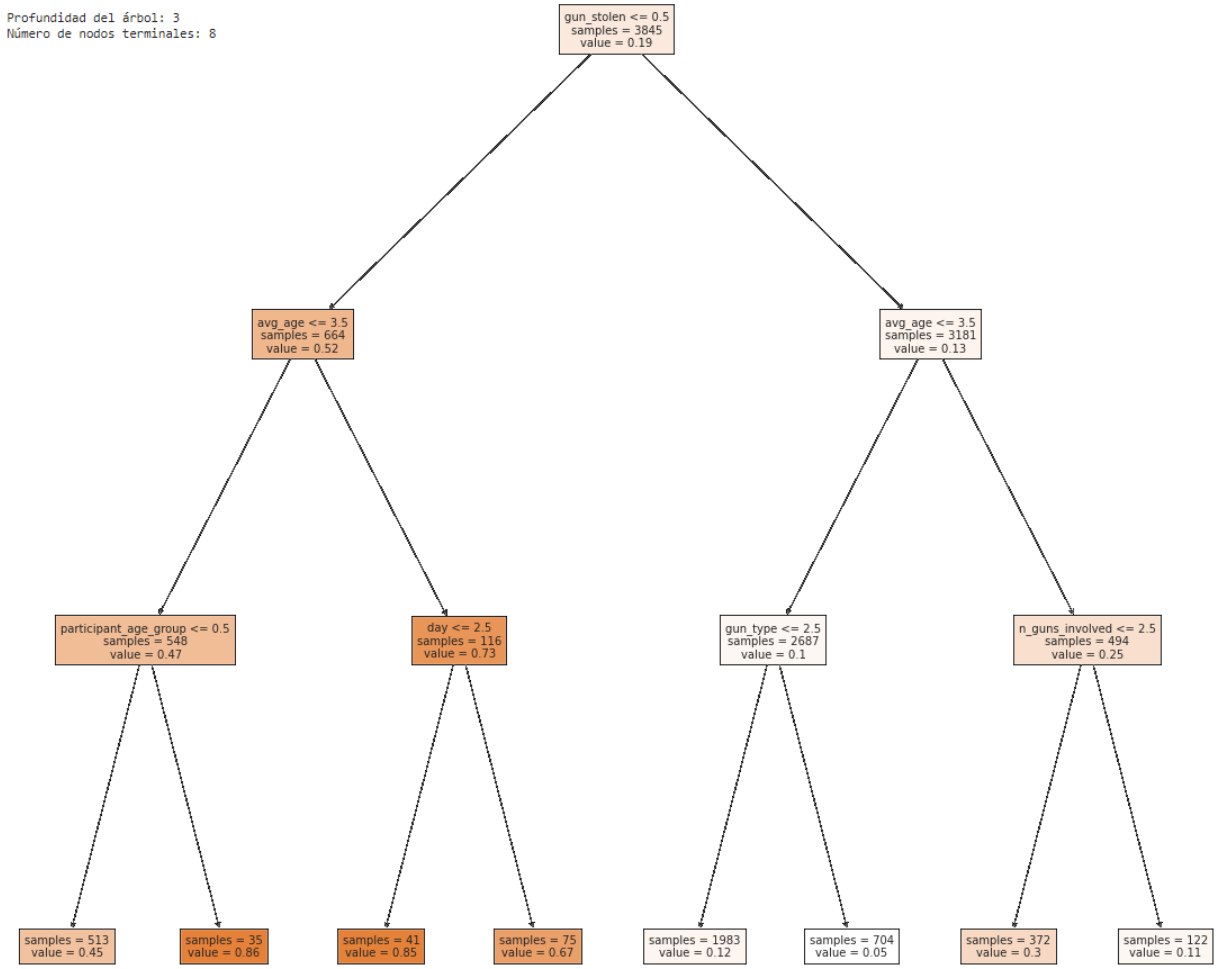
Matriz de confusión

```
array([[704,  57],
       [111,  90]])
```



Importancia de las características

Datos de árbol de decisión profundidad 3

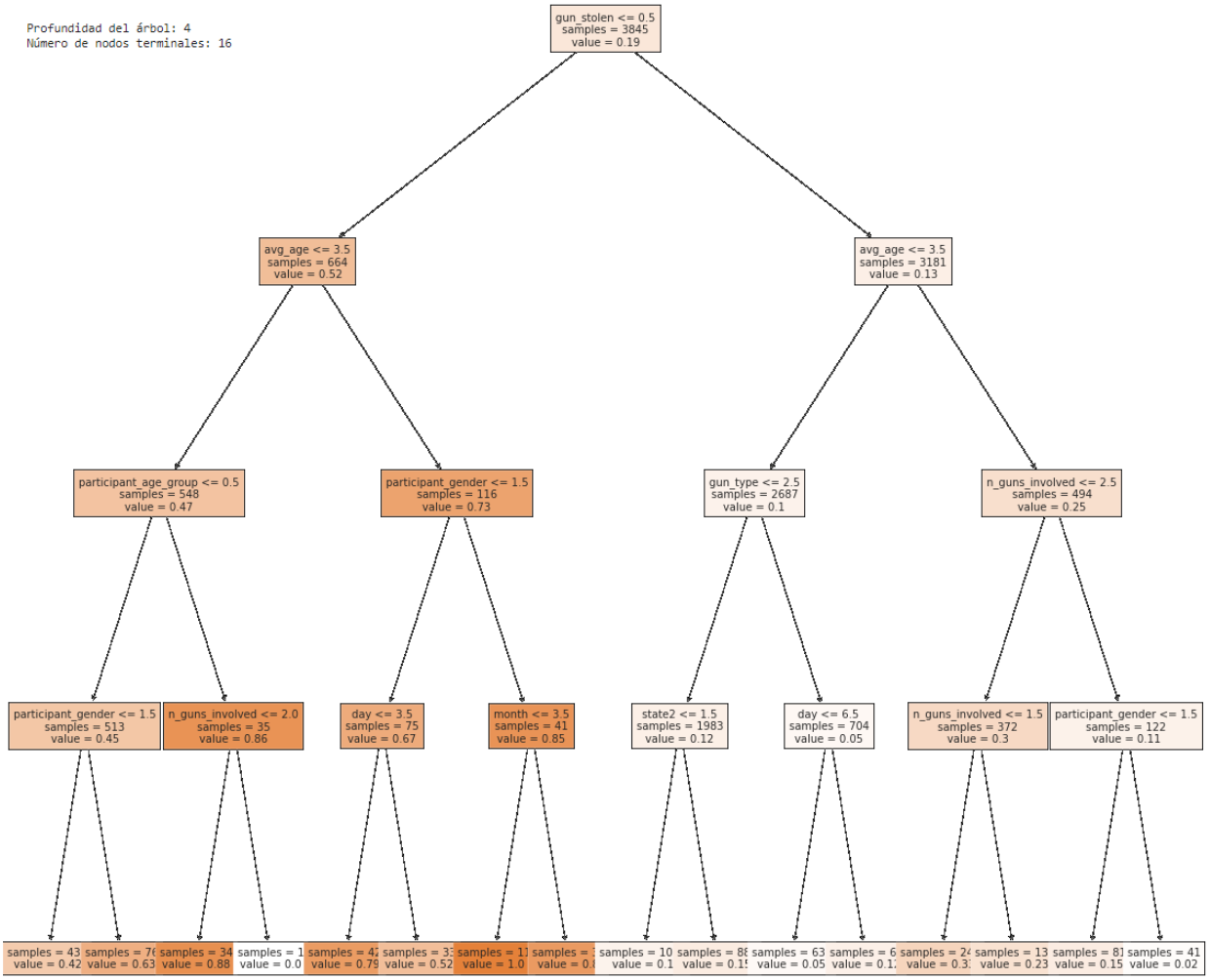



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- value: [0.45]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- value: [0.86]
|   |--- avg_age >  3.50
|   |   |--- day <= 2.50
|   |   |   |--- value: [0.85]
|   |   |--- day >  2.50
|   |   |   |--- value: [0.67]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- value: [0.12]
|   |   |--- gun_type >  2.50
|   |   |   |--- value: [0.05]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 2.50
|   |   |   |--- value: [0.30]
|   |   |--- n_guns_involved >  2.50
|   |   |   |--- value: [0.11]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.755016
7	avg_age	0.137733
5	participant_age_group	0.048598
4	n_guns_involved	0.030005
3	gun_type	0.020421
1	day	0.008227
0	month	0.000000
6	participant_gender	0.000000
8	state2	0.000000

Datos de árbol de decisión profundidad 4

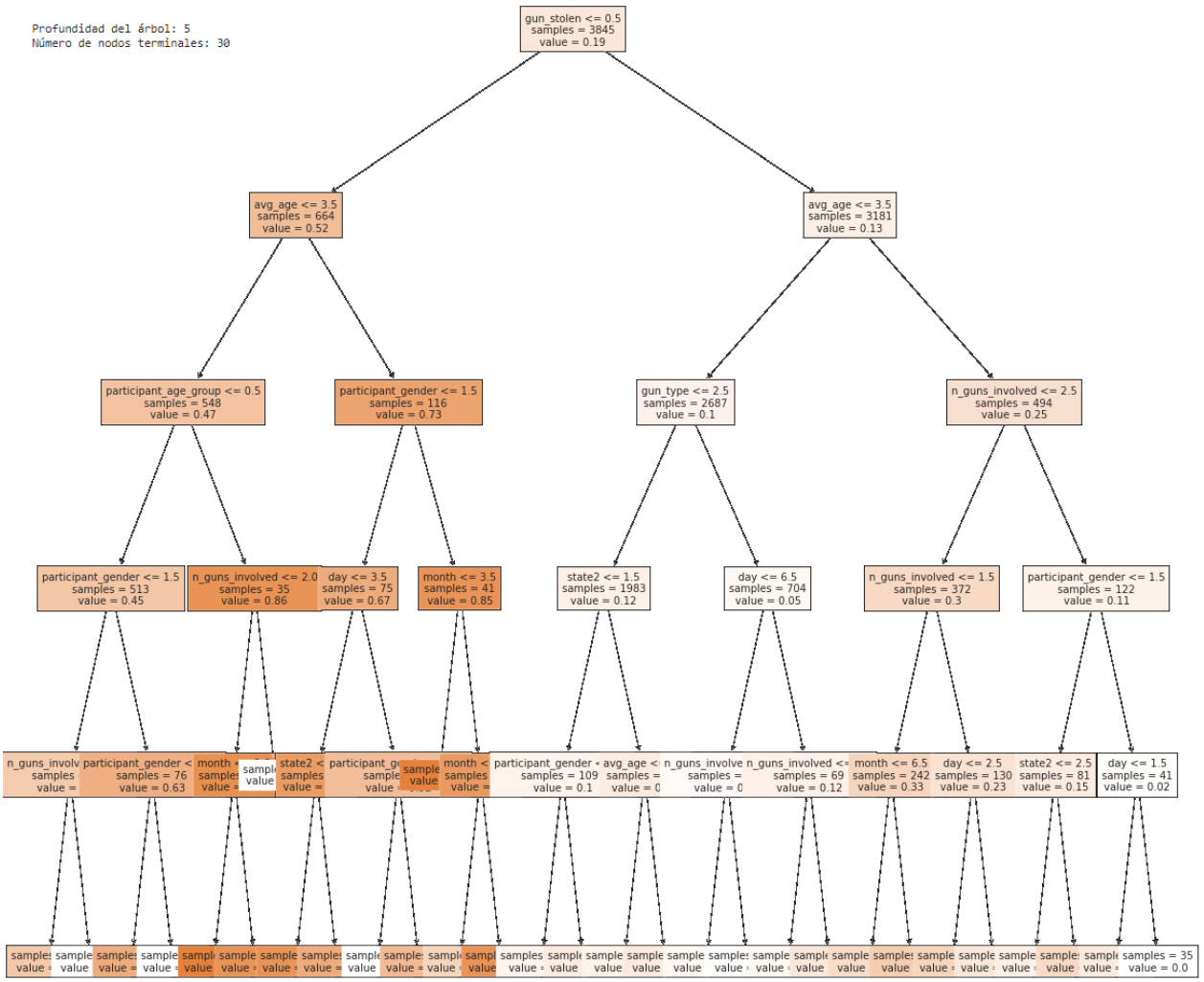



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.42]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.63]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- value: [0.88]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- day <= 3.50
|   |   |   |   |--- value: [0.79]
|   |   |   |--- day >  3.50
|   |   |   |   |--- value: [0.52]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- value: [0.80]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- state2 <= 1.50
|   |   |   |   |--- value: [0.10]
|   |   |   |--- state2 >  1.50
|   |   |   |   |--- value: [0.15]
|   |   |--- gun_type >  2.50
|   |   |   |--- day <= 6.50
|   |   |   |   |--- value: [0.05]
|   |   |   |--- day >  6.50
|   |   |   |   |--- value: [0.12]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 2.50
|   |   |   |--- n_guns_involved <= 1.50
|   |   |   |   |--- value: [0.33]
|   |   |   |--- n_guns_involved >  1.50
|   |   |   |   |--- value: [0.23]
|   |   |--- n_guns_involved >  2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.15]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.02]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.702785
7	avg_age	0.128205
5	participant_age_group	0.045236
4	n_guns_involved	0.041726
6	participant_gender	0.035849
3	gun_type	0.019008
1	day	0.013603
8	state2	0.010928
0	month	0.002660

Datos de árbol de decisión profundidad 5




```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.43]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.00]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- participant_gender <= 2.50
|   |   |   |   |   |--- value: [0.66]
|   |   |   |   |--- participant_gender >  2.50
|   |   |   |   |   |--- value: [0.00]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- month <= 2.50
|   |   |   |   |   |--- value: [1.00]
|   |   |   |   |--- month >  2.50
|   |   |   |   |   |--- value: [0.85]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- participant_gender <= 1.50
|   |   |   |--- day <= 3.50
|   |   |   |   |--- state2 <= 1.50
|   |   |   |   |   |--- value: [0.85]
|   |   |   |   |--- state2 >  1.50
|   |   |   |   |   |--- value: [0.67]
|   |   |   |--- day >  3.50
|   |   |   |   |--- participant_gender <= 0.50
|   |   |   |   |   |--- value: [0.00]
|   |   |   |   |--- participant_gender >  0.50
|   |   |   |   |   |--- value: [0.55]
|   |   |--- participant_gender >  1.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- month <= 5.50
|   |   |   |   |   |--- value: [0.33]
|   |   |   |   |--- month >  5.50
|   |   |   |   |   |--- value: [0.85]
|--- gun_stolen >  0.50
```

```
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- state2 <= 1.50
|   |   |   |   |--- participant_gender <= 1.50
|   |   |   |   |   |--- value: [0.09]
|   |   |   |   |--- participant_gender >  1.50
|   |   |   |   |   |--- value: [0.16]
|   |   |   |--- state2 >  1.50
|   |   |   |   |--- avg_age <= 0.50
|   |   |   |   |   |--- value: [0.10]
|   |   |   |   |--- avg_age >  0.50
|   |   |   |   |   |--- value: [0.17]
|   |   |--- gun_type >  2.50
|   |   |   |--- day <= 6.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.06]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.02]
|   |   |   |--- day >  6.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.05]
|   |   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |   |--- value: [0.19]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 2.50
|   |   |   |--- n_guns_involved <= 1.50
|   |   |   |   |--- month <= 6.50
|   |   |   |   |   |--- value: [0.28]
|   |   |   |   |--- month >  6.50
|   |   |   |   |   |--- value: [0.40]
|   |   |   |--- n_guns_involved >  1.50
|   |   |   |   |--- day <= 2.50
|   |   |   |   |   |--- value: [0.33]
|   |   |   |   |--- day >  2.50
|   |   |   |   |   |--- value: [0.19]
|   |   |--- n_guns_involved >  2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- state2 <= 2.50
|   |   |   |   |   |--- value: [0.13]
|   |   |   |   |--- state2 >  2.50
|   |   |   |   |   |--- value: [0.30]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- day <= 1.50
|   |   |   |   |   |--- value: [0.17]
|   |   |   |   |--- day >  1.50
|   |   |   |   |   |--- value: [0.00]
```

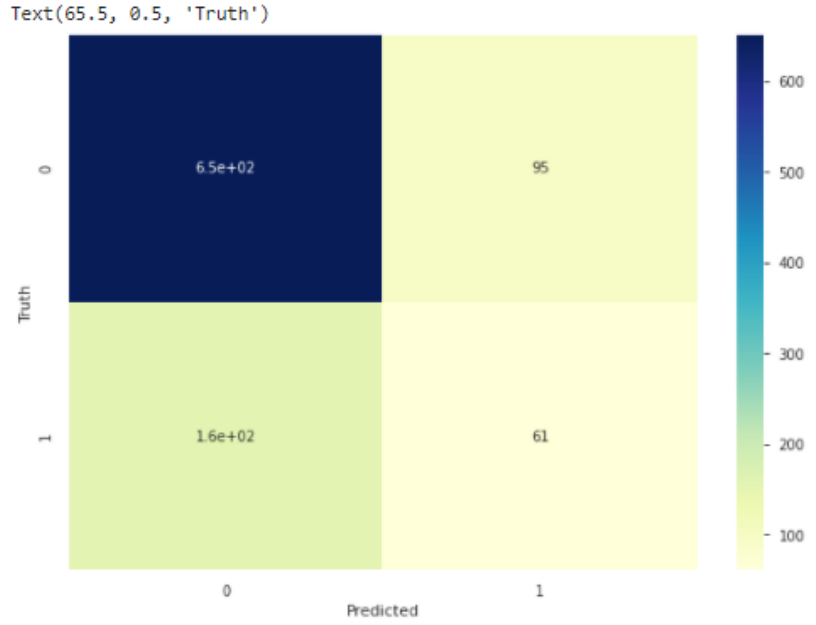
Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.654222
7	avg_age	0.126299
4	n_guns_involved	0.061391
6	participant_gender	0.049975
5	participant_age_group	0.042110
1	day	0.017785
3	gun_type	0.017695
0	month	0.015784
8	state2	0.014739

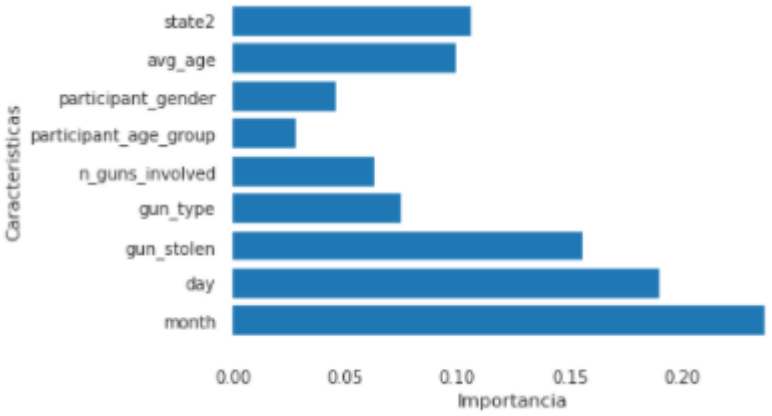
Datos de Gini “sin normalizar”:

Matriz de confusión

```
array([[651,  95],
       [155,  61]])
```



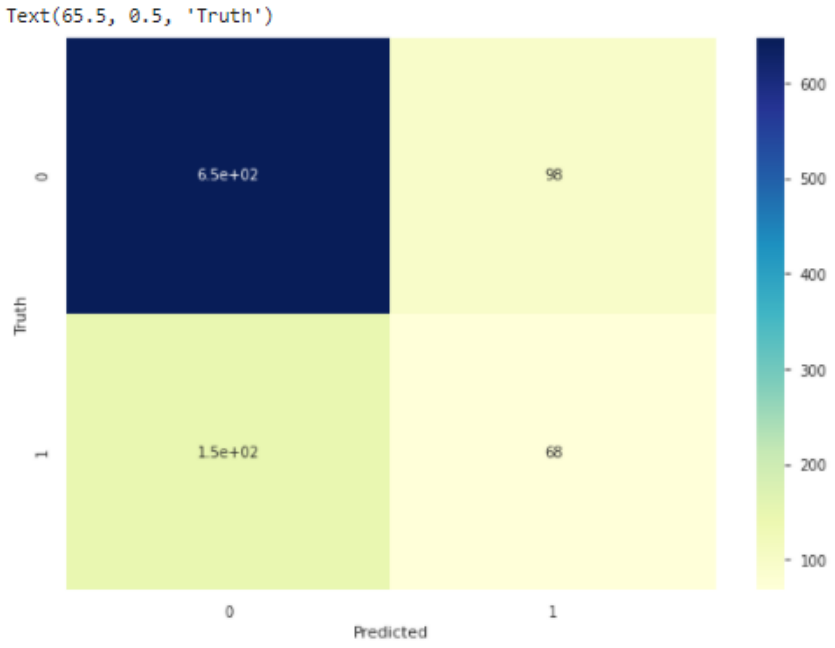
Importancia de las características



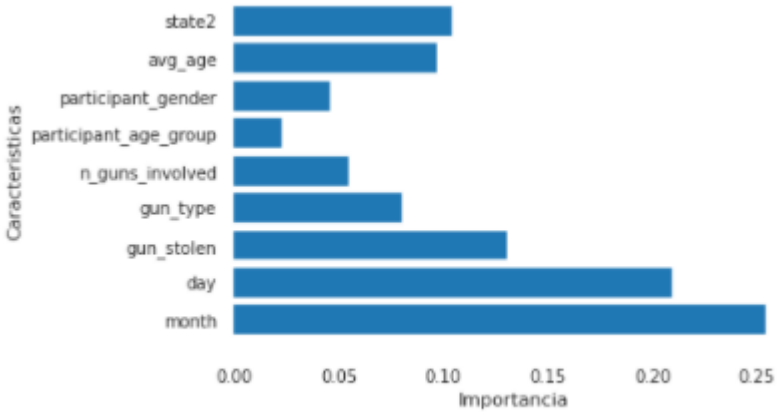
Datos de Entropía Cruzada “sin normalizar”:

Matriz de confusión

```
array([[648,  98],
       [148,  68]])
```

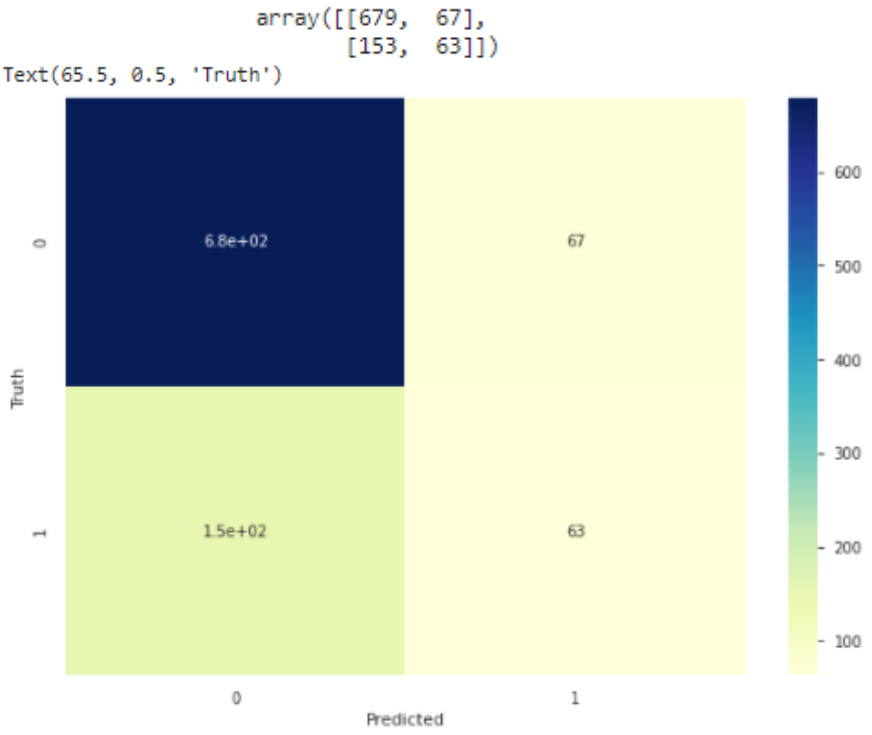


Importancia de las características

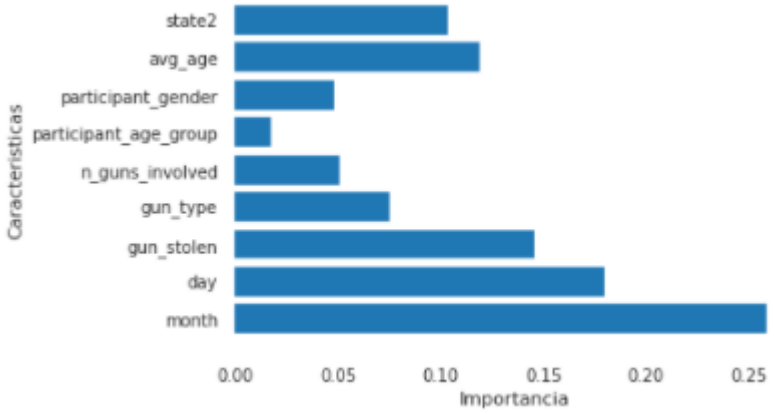


Datos de Random Forest “sin normalizar”:

Matriz de confusión



Importancia de las características

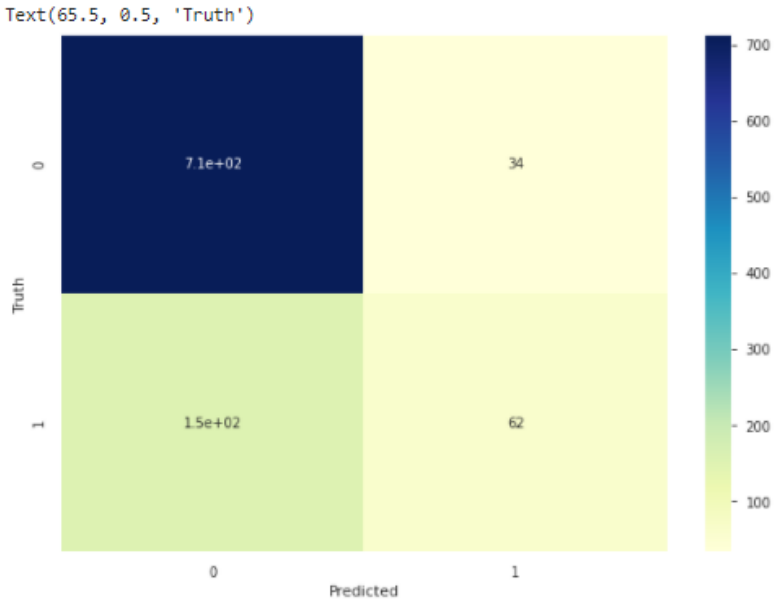


Datos de AdaBoost “sin normalizar”

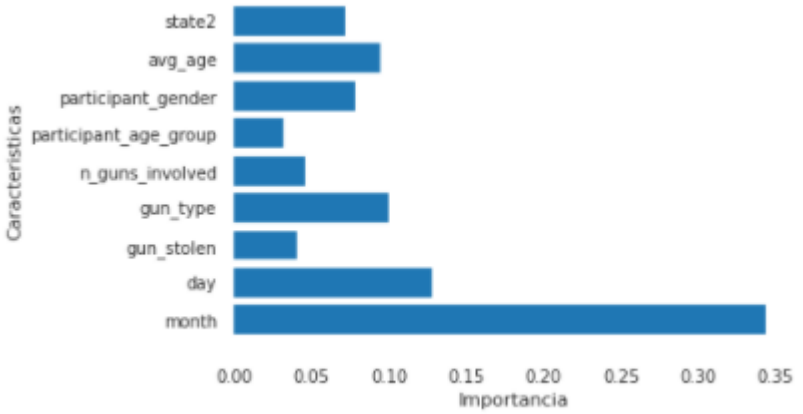
Matriz de confusión

```
array([[712,  34],
       [154,  62]])
```

:



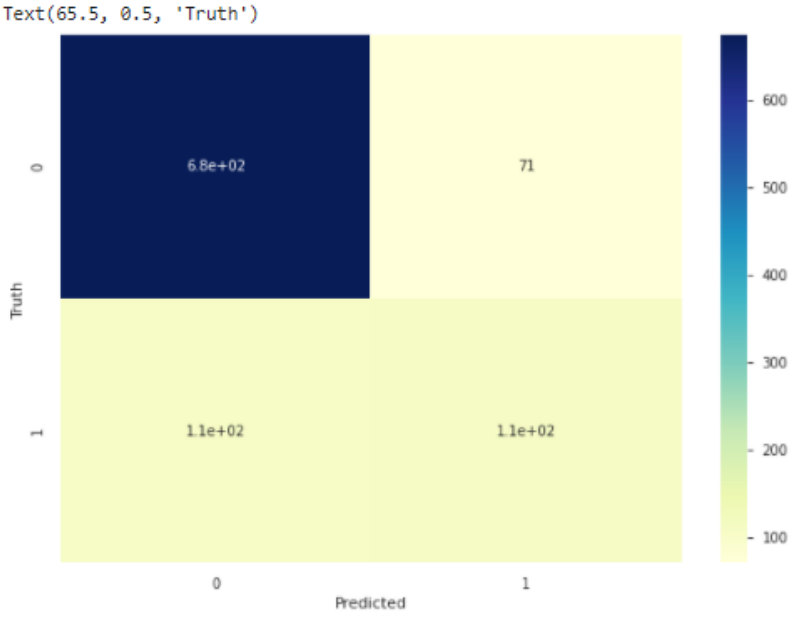
Importancia de las características



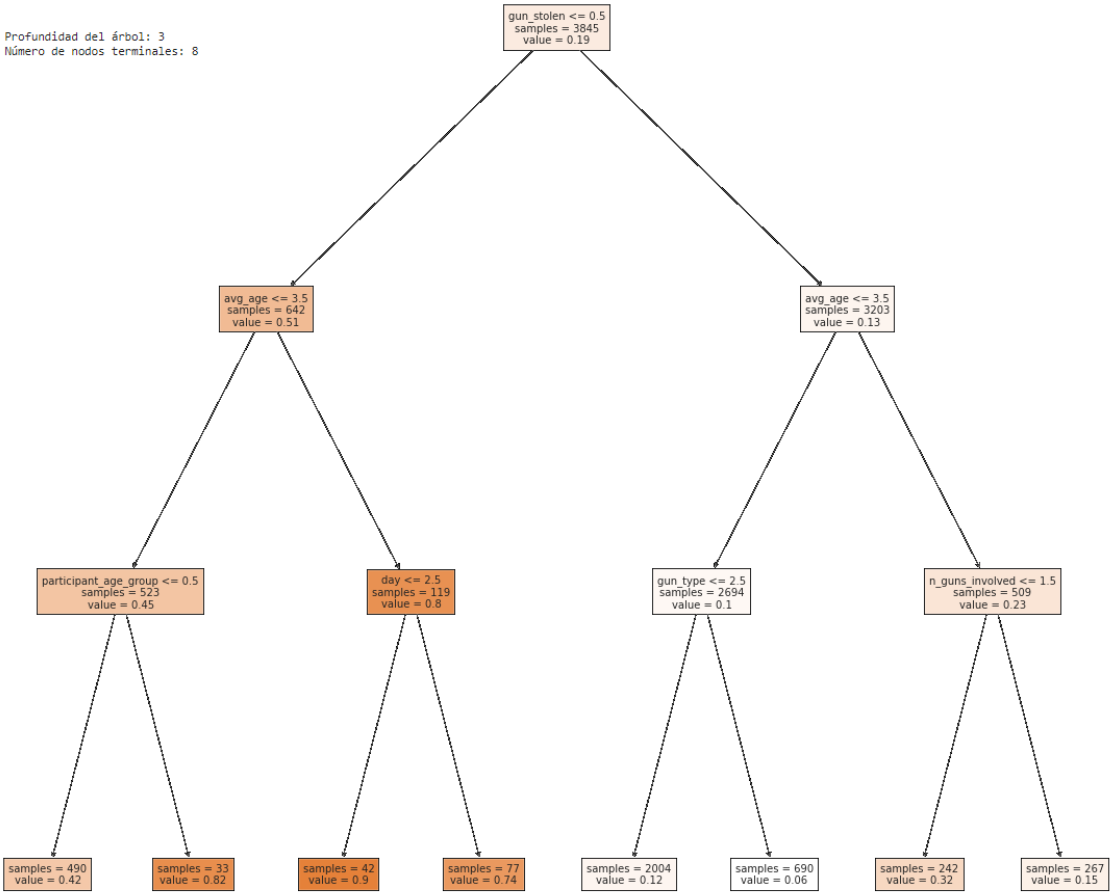
Datos de Clasificador Bayesiano Ingenuo “Sin normalizar”:

Matriz de confusión

```
array([[675,  71],
       [107, 109]])
```



Datos de árbol de decisión profundidad 3

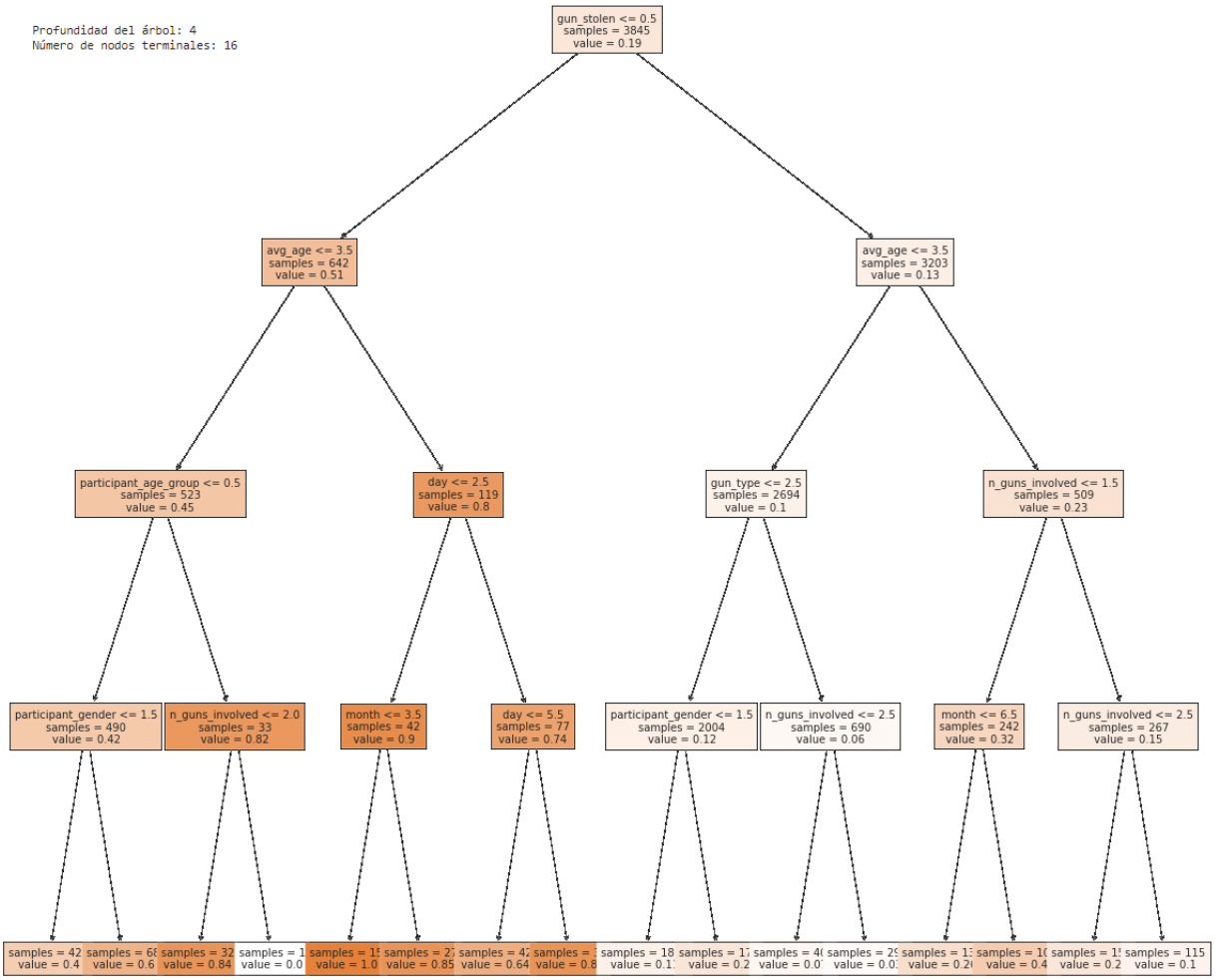



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- value: [0.42]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- value: [0.82]
|   |--- avg_age >  3.50
|   |   |--- day <= 2.50
|   |   |   |--- value: [0.90]
|   |   |--- day >  2.50
|   |   |   |--- value: [0.74]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- value: [0.12]
|   |   |--- gun_type >  2.50
|   |   |   |--- value: [0.06]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- value: [0.32]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- value: [0.15]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.727815
7	avg_age	0.170525
5	participant_age_group	0.043136
4	n_guns_involved	0.032542
3	gun_type	0.019362
1	day	0.006620
0	month	0.000000
6	participant_gender	0.000000
8	state2	0.000000

Datos de árbol de decisión profundidad 4

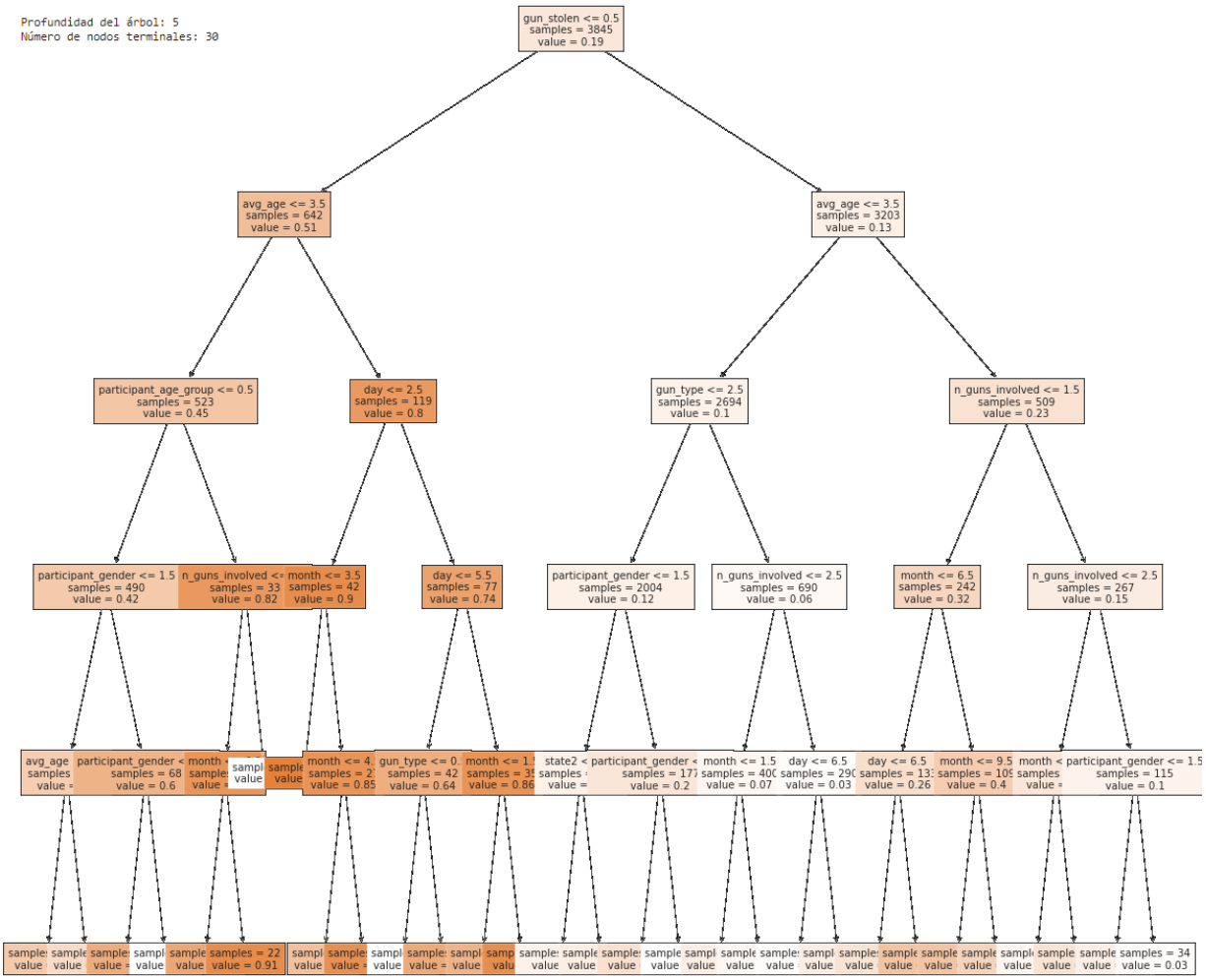



```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.40]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.60]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- value: [0.84]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- day <= 2.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- value: [0.85]
|   |   |--- day >  2.50
|   |   |   |--- day <= 5.50
|   |   |   |   |--- value: [0.64]
|   |   |   |--- day >  5.50
|   |   |   |   |--- value: [0.86]
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- value: [0.11]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- value: [0.20]
|   |   |--- gun_type >  2.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- value: [0.07]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- value: [0.03]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- month <= 6.50
|   |   |   |   |--- value: [0.26]
|   |   |   |--- month >  6.50
|   |   |   |   |--- value: [0.40]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- value: [0.20]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- value: [0.10]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.680857
7	avg_age	0.159523
4	n_guns_involved	0.044006
5	participant_age_group	0.040353
6	participant_gender	0.030737
3	gun_type	0.018113
1	day	0.013575
0	month	0.012836
8	state2	0.000000

Datos de árbol de decisión profundidad 5




```
|--- gun_stolen <= 0.50
|   |--- avg_age <= 3.50
|   |   |--- participant_age_group <= 0.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- avg_age <= 2.50
|   |   |   |   |   |--- value: [0.42]
|   |   |   |   |--- avg_age >  2.50
|   |   |   |   |   |--- value: [0.28]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- participant_gender <= 2.50
|   |   |   |   |   |--- value: [0.66]
|   |   |   |   |--- participant_gender >  2.50
|   |   |   |   |   |--- value: [0.00]
|   |   |--- participant_age_group >  0.50
|   |   |   |--- n_guns_involved <= 2.00
|   |   |   |   |--- month <= 6.00
|   |   |   |   |   |--- value: [0.70]
|   |   |   |   |--- month >  6.00
|   |   |   |   |   |--- value: [0.91]
|   |   |   |--- n_guns_involved >  2.00
|   |   |   |   |--- value: [0.00]
|   |--- avg_age >  3.50
|   |   |--- day <= 2.50
|   |   |   |--- month <= 3.50
|   |   |   |   |--- value: [1.00]
|   |   |   |--- month >  3.50
|   |   |   |   |--- month <= 4.50
|   |   |   |   |   |--- value: [0.50]
|   |   |   |   |--- month >  4.50
|   |   |   |   |   |--- value: [0.88]
|   |   |--- day >  2.50
|   |   |   |--- day <= 5.50
|   |   |   |   |--- gun_type <= 0.50
|   |   |   |   |   |--- value: [0.00]
|   |   |   |   |--- gun_type >  0.50
|   |   |   |   |   |--- value: [0.68]
|   |   |   |--- day >  5.50
|   |   |   |   |--- month <= 1.50
|   |   |   |   |   |--- value: [0.60]
|   |   |   |   |--- month >  1.50
|   |   |   |   |   |--- value: [0.90]
|--- gun_stolen >  0.50
```

```
|--- gun_stolen >  0.50
|   |--- avg_age <= 3.50
|   |   |--- gun_type <= 2.50
|   |   |   |--- participant_gender <= 1.50
|   |   |   |   |--- state2 <= 1.50
|   |   |   |   |   |--- value: [0.10]
|   |   |   |   |--- state2 >  1.50
|   |   |   |   |   |--- value: [0.13]
|   |   |   |--- participant_gender >  1.50
|   |   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |   |--- value: [0.24]
|   |   |   |   |--- participant_gender >  2.50
|   |   |   |   |   |--- value: [0.00]
|   |   |--- gun_type >  2.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- month <= 1.50
|   |   |   |   |   |--- value: [0.13]
|   |   |   |   |--- month >  1.50
|   |   |   |   |   |--- value: [0.06]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- day <= 6.50
|   |   |   |   |   |--- value: [0.02]
|   |   |   |   |--- day >  6.50
|   |   |   |   |   |--- value: [0.14]
|   |--- avg_age >  3.50
|   |   |--- n_guns_involved <= 1.50
|   |   |   |--- month <= 6.50
|   |   |   |   |--- day <= 6.50
|   |   |   |   |   |--- value: [0.22]
|   |   |   |   |--- day >  6.50
|   |   |   |   |   |--- value: [0.47]
|   |   |   |--- month >  6.50
|   |   |   |   |--- month <= 9.50
|   |   |   |   |   |--- value: [0.46]
|   |   |   |   |--- month >  9.50
|   |   |   |   |   |--- value: [0.34]
|   |   |--- n_guns_involved >  1.50
|   |   |   |--- n_guns_involved <= 2.50
|   |   |   |   |--- month <= 2.50
|   |   |   |   |   |--- value: [0.08]
|   |   |   |   |--- month >  2.50
|   |   |   |   |   |--- value: [0.22]
|   |   |   |--- n_guns_involved >  2.50
|   |   |   |   |--- participant_gender <= 1.50
|   |   |   |   |   |--- value: [0.12]
|   |   |   |   |--- participant_gender >  1.50
|   |   |   |   |   |--- value: [0.03]
```

Importancia de los predictores en el modelo

	predictor	importancia
2	gun_stolen	0.627212
7	avg_age	0.156670
6	participant_gender	0.061281
4	n_guns_involved	0.040539
5	participant_age_group	0.037174
0	month	0.027491
3	gun_type	0.023418
1	day	0.022315
8	state2	0.003901