UNIVERSIDAD DE CASTILLA-LA MANCHA

MINERÍA DE DATOS

ENTREGABLE 2: PREPROCESO, TRANSFORMACIÓN E HIPÓTESIS

Análisis de violencia con armas en EE. UU.

DARÍO ANDRÉS FALLAVOLLITA FIGUEROA

FERNANDO POTENCIANO SANTIAGO

IGNACIO ROZAS LÓPEZ

ISRAEL MATEOS APARICIO RUIZ SANTA QUITERIA

ADRÍAN JULIÁN RAMOS ROMERO

LAURENTIU GHEORGHE ZLATAR

27 de noviembre de 2023

Índice

1. Porcentajes de participación

El reparto de la participación entre integrantes ha sido equitativo, y se muestra en la tabla 1.

Apellidos y nombre	Correo	Participación
Fallavollita Figueroa, Darío Andrés	DarioAndres.Fallavollita@alu.uclm.es	16,6 %
Mateos Aparicio Ruiz Santa Quiteria, Israel	Israel.Mateos@alu.uclm.es	16,6 %
Potenciano Santiago, Fernando	Fernando.Potenciano@alu.uclm.es	16,6 %
Ramos Romero, Adrián Julián	AdrianJulian.Ramos@alu.uclm.es	16,6 %
Rozas López, Ignacio	Ignacio.Rozas@alu.uclm.es	16,6 %
Zlatar, Laurentiu Gheorghe	LaurentiuGheorghe.Zlatar@alu.uclm.es	16,6 %

Cuadro 1: Porcentaje de participación de los integrantes

2. Introducción

El presente documento tiene como fin comentar la etapa de preprocesamiento de nuestros conjuntos de datos. Nuestro objetivo es el análisis de los incidentes violentos con armas en EE. UU.

Para ello hemos recopilado diferentes *datasets*, siendo el principal datos sobre incidentes violentos con armas en EE.UU. Los demás serán secundarios, y contienen información sobre datos de pobreza y de leyes para la regulación de las armas. Su función será enriquecer nuestro conjunto de datos principal y probar (o refutar) nuestras hipótesis iniciales, las cuales son:

- Se producen más incidentes con armas los fines de semana.
- Se producen más incidentes con armas en los meses de verano.
- Los incidentes con armas son más frecuentes en los estados con mayor pobreza.
- Cuantas más leyes sobre el uso de armas existen en un estado, menos incidentes con armas se producen en él.

Las fuentes de nuestros conjuntos de datos son las siguientes:

Incidentes violentos con armas en EE. UU: Obtenidos de `https://www.kaggle.com/datasets/jameslko/gun-violence-data` y cuya fuente original es `https://www.gunviolencearchive.org`.

Datos de pobreza en EE.UU: Obtenidos mediante técnicas de *web scraping* de `https://www.povertyusa.org/`

Leyes para armas de fuego en EE. UU: Obtenidos directamente de `https://mail.statefirearmlaws.org/`.

Para poder extraer conocimiento útil de los datos, primero debemos hacer unas series de operaciones de preprocesamiento y transformación, con el fin de llegar a una o varias tarjetas de datos sobre las que aplicar algoritmos de minería de datos.

Estas operaciones consistirán en una limpieza de los datos manejando valores nulos y *outliers* (o valores extremos), la resolución problemas de integración como la codificación y representación (por ejemplo, de fechas), la selección de características y su transformación.

Finalmente, plantearemos unas líneas de trabajo que nos servirán como guía para nuestros próximos pasos.

Aclaración El apartado de selección e integración de varias fuentes pedido en el enunciado del entregable se explica en la sección del preprocesado, ya que es una parte de este, y hemos visto conveniente integrarlo en él para una mayor claridad.

3. Preproceso y transformación

3.1. Incidentes con armas

Nuestro conjunto de datos de incidentes con armas viene dado en un único fichero, en el que cada registro corresponde a un incidente en un estado y fecha concretos. Ya que las columnas que seleccionaremos no contienen nulos, y *a priori* estos están relativamente limpios, hemos decidido limpiar el conjunto de datos al completo, aunque después solo seleccionaremos algunas columnas. Esto es con el fin de mostrar que somos capaces de limpiar un conjunto de datos de un tamaño considerable.

En el primer paso de nuestra limpieza, nos encontramos con nulos en varias columnas. La estrategia general a seguir ha sido la de imputar estos valores con la media de la variable en caso de ser numérica o la moda en caso de ser categórica. Sin embargo, nos hemos encontrado con algunos casos particulares en los que hemos seguido otra estrategia:

- En las variables con más de un 50 % de nulos se ha imputado un valor *Unknown*, por no ser el estadístico una medida representativa de todo el conjunto de datos.
- Varias columnas presentan un formato de texto propio. Por ejemplo, la variable *gun_type* presenta una única cadena de texto, en la que se especifica el tipo de cada arma involucrada (presentando así una dependencia con la variable *n_guns involved*.

  En estas variables se han extraído los valores individuales del formato que usa, se ha aplicado la estrategia de media o moda según el tipo de variable, y se han imputado estos valores siguiendo de nuevo el formato concreto.

  Esto ocurre con las variables relacionadas con las armas involucradas y con datos de los participantes del incidente.

- En variables en las que no tiene sentido utilizar la media o la moda, y que presentan un gran porcentaje de valores únicos (por ejemplo, *notes* o *address*, se ha imputado el valor *Unknown*.

- Para las variables de latitud y longitud, por contener información espacial muy concreta, se ha usado la media para el estado en concreto en lugar de la media para toda la variable.

- Para las variables de distrito, se ha usado la moda para el estado en concreto, por tener este valor un significado distinto dependiendo del distrito.

Respecto al tratamiento de *outliers*, se han controlado los de las variables numéricas exceptuando el identificador del incidente por ser una variable única, y de la latitud y longitud por estar en un contexto geográfica muy amplio (la totalidad de los Estados Unidos).

Tras analizarlos, se han encontrado valores extremos como edades de 300 años (claramente erróneos) e incidentes con más de 300 armas involucradas. Gracias a este análisis, hemos observado que el conjunto de datos incluye como incidentes ciertas acciones del gobierno tales como confiscaciones masivas y operaciones de *gun buyback*. Este tipo de incidentes no tiene relevancia para nuestro objetivo y, de hecho, introduciría ruido en el mismo. Por tanto, hemos decidido eliminar los valores que distan más de 3 desviaciones estándar de la media de la variable.

Respecto a la integración, hemos separado las fechas en las variables *year*, *month* y *day* para poder trabajar con otros conjuntos de datos que presentan una granularidad de año. También hemos convertido las variables que presentaban el formato propio anteriormente mencionado en listas con los valores individuales, que es lo que realmente nos interesa.

Finalmente, hemos seleccionado las variables de fecha y estado, ya que para nuestro análisis lo único que nos interesa es el número de incidentes. También hemos descartado los datos de 2018, pues solo presentan incidentes hasta marzo. Además, para las hipótesis en las que haya que usar también datos de otros *datasets* descartaremos los datos previos al 2015, ya que los demás conjuntos de datos comienzan en ese año. A su vez, cuando se usen en conjunto con los datos de leyes, se descartaran los datos del estado "District of Columbia", ya que no tenemos datos de leyes para ese estado.

3.2. Datos de pobreza

Respecto a la limpieza del conjunto de datos de pobreza, nos hemos encontrado con casos puntuales de nulos, que hemos reemplazado por la media (por ser numéricas) de la variable en el estado en concreto. Ya que son datos demográficos, dependen de la situación socioeconómica concreta de cada estado, por lo que no hemos usado la media de la variable completa (que correspondería a los Estados Unidos).

En lo que respecta a los valores extremos, la mayoría de nuestras variables consisten de porcentajes, por lo que están escaladas del 0 al 1. Por esta razón, ninguna de estas variables

presenta valores extremos. Por el contrario, otras variables como la población o el número de habitantes viviendo en estado de pobreza si que presentan algunos. Sin embargo, para nuestro análisis usaremos porcentajes respecto a la población (por las diferencias en esta de un estado a otro, e.g. de Alaska a California), por lo que hemos decidido no eliminar los *outliers* de esas variables del conjunto de datos.

Los datos de este conjunto presentan un formato adecuado para su integración con los demás, por lo que se ha pasado a la selección directamente.

Aunque a primera vista pensabamos que todos los indicadores de pobreza presentes en el conjunto de datos eran interesantes para nuestro análisis, tras analizar la correlación entre las distintas variables nos hemos dado cuenta de que existe una fuerte correlación entre todos los indicadores. Por tanto, hemos decidido quedarnos con sólo uno de ellos. Este ha sido el porcentaje de la población viviendo en estado de pobreza, por ser el más general.

Además, con el fin de poder escalar el número de incidentes por población al crear nuestra tarjeta de datos, nos hemos quedado también con la población.

3.3. Leyes de regulación de armas

Respecto al conjunto de datos acerca de leyes reguladoras de las armas, nos encontramos ante unos datos algo distintos. Cada registro del conjunto de datos se refiere a un estado y año concreto, y presenta una variable por cada ley, con un valor de 1 si está activa o de 0 en caso contrario, además de una variable con la cantidad de leyes activas. Además, hacemos uso de un *codebook* (diccionario de datos) en el que cada ley está categorizada.

Este *codebook* presenta valores nulos en variables referentes a anotaciones sobre la ley que no nos son de interés. En estas, se ha imputado un valor *Unknown*. Respecto a los valores extremos, el *codebook* solo presenta variables categóricas, mientras que el conjunto de datos en sí solo presenta variables categóricas o booleanas, a excepción de la cantidad total de leyes en cada estado y año. Por tanto, no se ha realizado ninguna operación respecto a *outliers*.

Para facilitar el manejo del conjunto de datos en su posterior integración con los demás, se han cambiado los nombres de las columnas al mismo formato que las demás, i.e. en minúscula y usando guiones bajos como separados.

Se han seleccionado sólo el código de la categoría y los nombres de las variables de cada ley para el *codebook*, y todas las variables para el conjunto de datos en sí, ya que posteriormente agruparemos las leyes por su categoría. Además, se han descartado los datos previos a 2015 para coincidir con los demás conjuntos de datos.

3.4. Transformación

Una vez hemos limpiado los conjuntos de datos, hemos dado un formato común y que nos es útil, y hemos seleccionado los datos que necesitamos, hemos pasado a la transformación de los datos para la generación de las tarjetas de datos.

Ya que nuestras hipótesis necesitan de distintos conjuntos y distintas granularidades a nivel temporal, hemos decidido crear una tarjeta de datos para las dos primeras hipótesis y otra por cada una de las restantes.

Para la tarjeta de datos de las dos primeras hipótesis, que tratan del análisis del número de incidentes en fines de semana y en meses de verano, hemos creado dos nuevas variables en función de la información que tenemos: una que indica si el incidente en concreto se produce o no en fin de semana, y otra que indica si se produce o no en meses de verano (junio, julio y agosto). Mediante estas variables, después hemos agrupado el conjunto por estado, año, si es fin de semana y si es mes de verano. Ya que hay más días fuera del fin de semana que en ellos, y más días fuera del verano que en él, hemos escalado el número de incidentes dividiéndolo por el número de días en el que se dan.

Para la tarjeta de datos de la tercera hipótesis, que trata del análisis de la relación entre incidentes y datos de pobreza, hemos agregado los incidentes a nivel de estado y año. Además, hemos dividido este número de incidentes por cada 100.000 habitantes, con el fin de escalar nuestros datos.

Para la tarjeta de datos de la cuarta y última hipótesis, que trata del análisis de la relación entre incidentes y leyes reguladoras de armas, hemos agregado de nuevo los incidentes por estado y año. Hemos creado también una nueva variable por cada categoría de ley, y hemos asignado a cada una de ellas el número de leyes activas de esa categoría. También hemos añadido la variable con el número total de leyes activas, y hemos escalado el número de incidentes de la misma manera que en la segunda tarjeta de datos.

Estas tarjetas de datos se muestran con más claridad en la siguiente sección del documento.

4. Características de la tarjeta de datos

Como se ha explicado en la sección anterior, hemos decidido crear una tarjeta de datos para las dos primeras hipótesis y otra por cada una de las restantes.

A continuación, mostramos los diccionarios de datos para cada una de las tarjetas de datos.

4.1. Hipótesis 1 y 2: fines de semana y meses de verano

Para las dos primeras hipótesis, la tarjeta de datos presenta las variables mostradas en la tabla 2.

Columna	Tipo de dato	Descripción
state	String	Estado al que se refiere el registro
year	int	Año al que se refiere el registro
is_weekend	int	1 si se refiere a incidentes producidos en fin de semana, 0 en caso contrario
is_summer	int	1 si se refiere a incidentes producidos en junio, julio o agosto, 0 en caso contrario
n_incidents_per_day	float	Número de incidentes con armas al día

Cuadro 2: Diccionario de la tarjeta de datos para las hipótesis 1 y 2

4.2. Hipótesis 3: datos de pobreza

Para la tercera hipótesis, la tarjeta de datos presenta las variables mostradas en la tabla 3.

Columna	Tipo de dato	Descripción
state	String	Estado al que se refiere el registro
year	int	Año al que se refiere el registro
n_incidents	float	Número de incidentes por 100.000 habitantes
poverty_rate	float	La tasa de pobreza, que representa el porcentaje de la población viviendo por debajo del umbral de pobreza (para una familia de 4 miembros, unos 27.500 dólares)

Cuadro 3: Diccionario de la tarjeta de datos para la hipótesis 3

4.3. Hipótesis 4: leyes de regulación de armas

Para la cuarta hipótesis, la tarjeta de datos presenta las variables mostradas en la tabla 4.

Columna	Tipo de dato	Descripción
state	String	Estado al que se refiere el registro
year	int	Año al que se refiere el registro
n_incidents	float	Número de incidentes por 100.000 habitantes
lawtotal	int	Cantidad total de leyes vigentes
laws_1	int	Cantidad de leyes de categoría 1 vigentes
laws_2	int	Cantidad de leyes de categoría 2 vigentes
...		
laws_14	int	Cantidad de leyes de categoría 14 vigentes

Cuadro 4: Diccionario de la tarjeta de datos para la hipótesis 4

5. Líneas de trabajo

De acuerdo con nuestras hipótesis y las tarjetas de datos obtenidas, hemos planteado las siguientes líneas de trabajo:

1. Regresión para modelar la relación entre el día de la semana (laboral o fin de semana) y la incidencia de crímenes con armas, determinando si hay una asociación significativa entre estos factores.

2. Por su parecido con la anterior, plantearemos de nuevo una regresión para modelar la relación entre el mes (junio, julio o agosto, es decir, meses de verano, u otro) y la incidencia de crímenes con armas, determinando si hay una asociación significativa entre estos factores.

3. Predicción mediante una regresión del número de incidentes con armas en un estado y año dado una tasa de pobreza.

4. Reducción del número de características de la tarjeta de datos referente a leyes mediante PCA, y clusterización (e.g. mediante *k-means*) en grupos dependiendo de las leyes vigentes. Esto nos proporcionará una visión más detallada de cómo diferentes enfoques legislativos pueden estar asociados con patrones específicos de crímenes con armas.

Además, se ha planteado la posibilidad de explorar una línea de trabajo adicional que no se corresponde con ninguna hipótesis en concreto, sino que consistiría en la combinación de todas. Por la gran variedad de factores socioeconómicos que pueden influir en nuestro tema, sería interesante unir todos ellos.

Aclaración Estas líneas de trabajo son sólo una primera aproximación a los algoritmos a utilizar, de modo que nos sirva de guía en los próximos pasos a seguir. Sin embargo, es muy probable que cambien respecto a los algoritmos utilizados finalmente.