UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

Instituto de Física y Matemáticas

Licenciatura en Matemáticas Aplicadas

Análisis exploratorio de datos y un modelo ML para la identificación de patrones en ataques terroristas a nivel global

TESIS

para obtener el título de

Licenciado en Matemáticas Aplicadas

presenta

Adrián José Mendoza Chávez

Director de tesis:

Dr. Tomás Pérez Becerra

Co-director de tesis:

Dr. Alejandro Iván Aguirre Salado

Huajuapan de León, Oaxaca *Mayo de 2024*

Dedicatoria

A mi familia, especialmente a mis padres Pedro y Teresa, quienes hicieron posible este sueño.

Agradecimientos

A mi padre, Pedro Mendoza Miguel, le agradezco todo el apoyo incondicional que me ha brindado a lo largo de mi vida. Tu ejemplo de responsabilidad y compromiso con lo que amas hacer ha sido mi guía constante.

A mi madre, Teresa Chávez Castellanos, gracias por el amor con el que siempre me has cuidado. Tus enseñanzas sobre la lucha por nuestros sueños y la valentía para enfrentar los desafíos de la vida han sido fundamentales en mi desarrollo.

A mis hermanos, por su ejemplo de superación y cuidado, tanto directo como indirecto, ha sido invaluable para mí y gradezco cada consejo que me han ofrecido en mi camino.

A mis amigas, Lali, May y Bety, por los momentos inolvidables que compartimos juntos. Su compañía y apoyo fueron esenciales para hacer mi experiencia universitaria muy gratificante.

A Jessi, quiero agradecerte por tu constante apoyo y motivación durante todo este proceso. Tu presencia desde el inicio hasta el final ha sido fundamental para mantenerme enfocado en mis metas. Aprecio enormemente tu paciencia, disposición para escucharme y tus palabras de aliento, las cuales han sido un gran respaldo para mí.

A mi director de tesis, Dr. Tomás, le agradezco enormemente todo el respaldo brindado durante este proceso académico. Sus consejos y orientación fueron cruciales para mi desarrollo profesional. También quiero expresar mi gratitud al Dr. Alejandro, mi co-director de tesis, por su tiempo y dedicación en la revisión y corrección de mi trabajo.

A mis revisores, Dr. Jesús Alejandro Hernández Tello, Dr. Emmanuel A. Romano Castillo y Dr. Virgilio Vázquez Hipólito, les estoy muy agradecido por su tiempo y es-

fuerzo dedicado a la revisión de mi trabajo. Sus correcciones y consejos contribuyeron significativamente a mejorar la calidad de mi investigación.

A la Universidad Tecnológica de la Mixteca, quiero expresar mi profundo agradecimiento por todo el apoyo recibido durante mis años de estudio. Agradezco la excelencia académica de las clases y el invaluable conocimiento proporcionado por mis profesores, que han enriquecido mi formación académica de manera significativa.

Índice general

Introducción

El terrorismo es una forma inusual de manifestación, que puede tener un objetivo religioso, cultural, político o toma de poder con un proceder violento afectando a terceros. En [13] lo definen dependiendo de la perspectiva que se esté analizando (académica, jurídica, psicológica, criminológica, entre otras), por otra parte se menciona una clasificación desde cuatro planteamientos (por su origen, función, efectos y naturaleza). También se clasifica de una manera más simple como local o regional e internacional o global. El primero se concentra en un lugar específico y con un propósito claramente fijado, de tal manera que sus acciones van dirigidas sólo a la población de un espacio geográfico concreto, y el segundo está orientado a afectar la mayor cantidad de habitantes posible, no se limita a una región en particular. Su finalidad es afectar al público en general, pues sus actividades son enfocadas en aterrorizar a la comunidad mundial, de tal modo que no existen fronteras que limiten su actuación. Así, el terrorismo se define como un acto o serie de actos violentos contra individuos de un sector concreto con la intención de cambiar los resultados de algún proceso político.

Para que el terrorismo cumpla su objetivo se pueden presentar diversos cambios en la serie de sucesos violentos que se realizan; el sector de la población afectada; el número de muertes que provoca; los daños materiales que ocasiona; los procedimientos que se utilizan; entre otros. En [20] se menciona que el mejoramiento de los avances tecnológicos permitió al terrorismo expandirse por todo el mundo, aumentando su valor estratégico

en la organización, convirtiéndolo así en una amenaza mundial. Por ejemplo, durante la revisión bibliográfica se encontró que en la mayoría de los actos registrados a lo largo del tiempo, existe un gran vínculo entre los medios de comunicación con los atentados. Generalmente, los perpetradores de actos criminales tienen el objetivo que sus actos no sean anunciados en los medios de comunicación, mientras que los líderes terroristas al cometer un acto criminal prefieren que las consecuencias que provocan sus acciones violentas aparezcan en las primeras líneas de los noticieros. De aquí, surge el interés de poder identificar patrones que se encuentran en cada uno de los ataques terroristas con la finalidad de generar conclusiones o predicciones con la ayuda de los datos históricos obtenidos, y el uso de herramientas informáticas y estadísticas que faciliten la investigación.

Se puede apreciar en [3] que un instrumento relevante en la lucha antiterrorista es la información exacta y oportuna, la cual, acompañada de los análisis adecuados, constituye el alma de las operaciones antiterroristas. Cada tarea, desde aquellas diplomáticas, militares, financieras y políticas, hasta las advertencias ante futuros ataques, depende en gran medida de los servicios de inteligencia. La generación de conocimiento multidisciplinar resulta esencial para proporcionar indicios de un posible ataque, como puede ser la investigación de la cultura y mentalidad de las organizaciones terroristas, con el fin de señalar los principales puntos vulnerables sobre los cuales actuar y así poder prevenir, evitar y frustrar sus actividades. Se ha observado que la mayor parte de la información antiterrorista operacional suministrada a los responsables de tomar decisiones procede del análisis de comunicaciones.

Una de las ventajas que se generan con los avances tecnológicos es la facilidad de registrar y almacenar grandes cantidades de información, así como las diversas formas de manipularla de una manera ordenada sin mucha dificultad. En este sentido, se ha observado la gran relación que tiene el uso de la tecnología con el campo emergente conocido como ciencia de datos, en la cual se utilizan conocimientos en matemáticas y computación para generar aplicaciones en las variadas áreas de la ciencia, haciendo uso, por ejemplo, de pruebas de hipótesis, de análisis exploratorio, de modelos de aprendizaje máquina y de métodos de programación en general. Es en esta área donde se enmarca el

presente trabajo de tesis.

Una característica de la ciencia de datos se fundamenta en el análisis de la información, con el objetivo de aplicar métodos científicos, procesos y sistemas estadísticos e informáticos para extraer resultados que aporten conocimiento relevante (véase [16]). Esto muestra el camino a seguir haciendo uso de estas herramientas y de los registros relacionados con ataques terroristas ocurridos con el paso del tiempo.

La predicción de un posible ataque terrorista y la identificación de zonas de riesgo pueden ayudar en la toma de decisiones para su prevención, lo que contribuiría a salvar vidas humanas. En la revisión bibliográfica realizada, la mayoría de los análisis se enfocan en monitorear en tiempo real ciertas características, sin embargo, durante la recopilación de información, se observó que no existen modelos predictivos que alerten sobre la inminente amenaza. El problema que se desea resolver, y que a su vez se plantea como el objetivo general en este trabajo de tesis, es el de proporcionar un modelo que ejecute simulaciones de predicciones con un alto nivel de confianza fundamentado en técnicas de ciencia de datos. Para lograrlo, es esencial contar con una base de datos que contenga los registros necesarios. En esta investigación utilizaremos la compilación elaborada por investigadores del National Consortium for the Study of Terrorism and Responses to Terrorism ([17]), localizado en la Universidad de Maryland. Esta es una colección de código abierto que contiene datos registrados desde 1970 hasta 2017 acerca de ataques terroristas ocurridos a nivel global, llamada *Global Terrorism Database* (brevemente GTD), incluye incidentes nacionales e internacionales con más de 180,000 ataques y 100 variables sobre locación, tácticas, perpetradores, objetivos y víctimas. Estos registros son acordes a nuestro objetivo. Después de pasar por un necesario proceso de ajuste y limpieza, estos registros nos sirven para entrenar el programa de aprendizaje máquina diseñado con la finalidad de identificar patrones de grupos terroristas.

Para lograr el objetivo de la investigación, la tesis se estructuró de la siguiente manera:

Capítulo 1: Se presenta el análisis estructural de la base de datos, se muestra un contexto sobre su creación, la tipología de las variables y la visualización requerida para

determinar el procesamiento necesario.

Capítulo 2: Como consecuencia de la exploración realizada en el Capítulo 1, se determinó que se requería realizar una imputación a ciertas variables. En este capítulo se muestra el proceso matemático a detalle para completar la base.

Capítulo 3: Se ha demostrado que las variables redundantes o no relevantes afectan el desempeño de los modelos de aprendizaje. Como resultado del Capítulo 1 se observó la existencia de este tipo de variables, por lo que en este capítulo se muestra el método estadístico de reducción aplicado, conocido como *"Selección de las mejores k características"*.

Capítulo 4: Para finalizar la investigación, en este capítulo se muestra, entrena y se evalúa un modelo de k vecinos cercanos multiclase con métrica euclidiana, el cual determina a que grupo terrorista corresponde un nuevo registro.

A manera de justificación, el desarrollo de este trabajo permitirá predecir ataques futuros a corto o mediano plazo, así como regiones de riesgo susceptibles a ellos, basado en la identificación de causas y correlación de tales variables. Con lo que se logrará auxiliar en la toma de decisiones para proteger a la población vulnerable.

Capítulo 1

Análisis exploratorio y visualización de los datos

Este capítulo está dedicado a la exploración de la base, el reconocimiento de datos y variables y el análisis exploratorio necesario para solventar determinadas deficiencias. Además, se realizarán visualizaciones para obtener información relevante y se establecerán algunas conclusiones parciales.

1.1. Base de datos

Como se mencionó en la introducción, Global Terrorism Database (GTD) es un conjunto de datos de acceso abierto que contiene información sobre incidentes terroristas en todo el mundo desde 1970 y se actualizó recientemente con registros de 2019. Este estudio utilizará el conjunto de datos distribuidos inicialmente que incluye hasta antes de 2017.

Existen otras bases de datos, pero a diferencia de ellas, GTD contiene registros de incidentes terroristas nacionales e internacionales. Cada evento va acompañado de información como la fecha y el lugar. También se incluyen las armas y los objetivos utilizados, las víctimas y, en algunos casos, los grupos de perpetradores, lo que permite establecer más inferencias al analizar los datos como espaciales y temporales.

La GTD define un atentado terrorista como la amenaza o el uso real de la fuerza ilegal y la violencia por parte de un actor no estatal para alcanzar un objetivo político,

económico, religioso o social a través del miedo, la coacción o la intimidación. En la práctica, esto significa que para considerar un incidente para su inclusión en la GTD, deben estar presentes los tres atributos siguientes:

- **El incidente debe ser intencionado**. Es decir, el resultado de un cálculo consciente por parte del autor.
- **El incidente debe conllevar algún nivel de violencia o amenaza inmediata de violencia**. Incluida la violencia contra la propiedad, así como la violencia contra las personas.
- **La base de datos no incluye actos de terrorismo de Estado.**.

Además, deben darse al menos dos de los tres criterios siguientes para que un incidente se incluya en la GTD:

- **Criterio 1:** El acto debe tener un objetivo político, económico, religioso o social. En cuanto a los objetivos económicos, la búsqueda exclusiva de ganancias no satisface este criterio. Debe implicar la búsqueda de un cambio económico más profundo y sistémico.
- **Criterio 2:** Debe existir evidencia de la intención de coaccionar, intimidar o transmitir algún otro mensaje a una audiencia (o audiencias) más amplia que las víctimas inmediatas. Se considera el acto en su totalidad, independientemente de si cada individuo involucrado en la ejecución del acto era consciente de esta intención. Si alguno de los planificadores o tomadores de decisiones detrás del ataque tenía la intención de coaccionar, intimidar o publicar, se cumple el criterio de intencionalidad.
- **Criterio 3:** La acción debe estar fuera del contexto de actividades legítimas de guerra. Es decir, el acto debe estar fuera de los parámetros permitidos por el derecho internacional humanitario, en la medida en que se dirige a no combatientes.

Cada uno de estos tres últimos filtros de criterios puede aplicarse a la base de datos.

1.2. Tipos de datos

En la exploración inicial se observa que la base GTD cuenta con 181,691 registros o filas y con 135 variables o columnas (también denominadas "características"). En el cuadro 1.2.1 se muestran de una manera resumida las primeras cinco líneas.

eventid	iyear	imonth	iday	approxdate	...	country_txt
197000000001	1970	7	2	NaN	...	Dominican Republic
197000000002	1970	0	0	NaN	...	Mexico
197001000001	1970	1	0	NaN	...	Philippines
197001000002	1970	1	0	NaN	...	Greece
197001000003	1970	1	0	NaN	...	Japan

Cuadro 1.2.1: Primeros cinco registros de la base.

Algunas de las variables son:

- *eventid:* Identificador numérico del incidente.
- *iyear:* Año.
- *imonth:* Mes.
- *iday:* Día.
- *approxdate:* Cuando la fecha exacta del incidente no se conoce, este campo se utiliza para registrar la fecha aproximada del incidente.
- *country_txt:* País o lugar donde ocurrió el incidente.
- *region_txt:* Región de ocurrencia (12 regiones).
- *city:* Contiene el nombre de la ciudad, aldea o pueblo en el que ocurrió el incidente.
- *targtype1_txt:* Registra el tipo de objetivo.

- *gname:* Nombre del grupo que llevó a cabo el ataque.
- *weaptype1_txt:* Clase de arma utilizada.

Con la finalidad de iniciar el procesamiento de la base GTD, esta se considerará como en la siguiente definición.

Definición 1.2.1 (Datos multidimensionales)**.** *Un conjunto de datos multidimensionales* $\overline{D}$ *es un conjunto de n vectores,* $\overline{X_1}, \ldots, \overline{X_n}$ *tal que cada vector* $\overline{X_i}$ *contiene un conjunto de d características denotadas por* $(x_i^1, \ldots, x_i^d)$*.*

Para la base GTD, consideramos su dimensión de tamaño $n \times d$, donde $n = 181,691$ y $d = 135$. De ahora en adelante, será tratada como en la Definición 1.2.1.

Observemos en el Cuadro 1.2.1 que la base de datos $\overline{D}$ es mixta (también denominada como mezclada), ya que contiene registros con atributos mixtos, es decir, se encuentran los siguientes tipos de valores:

- Datos cuantitativos. Son de tipo numérico y tienen un orden natural, por ejemplo, la variable *iyear*.
- Datos categóricos. Contienen atributos que adquieren características discretas. Por ejemplo, la variable *region_txt*.
- Datos binarios. Es un caso especial de datos categóricos, en los que cada atributo puede tomar uno de dos valores discretos como máximo.
- Datos de texto. Son cadenas o son multidimensionales, dependiendo de su representación. Las variables con terminación *txt* son de este tipo.

La base de datos $\overline{D}$ en el entorno del lenguaje de programación Python contiene los tipos de datos del Cuadro 1.2.2, es decir, la base de datos $\overline{D}$ contiene 55 columnas de tipo float (números decimales) y 22 columnas de tipo int64 (números enteros), los cuales representan datos cuantitativos, además la base cuenta con 58 columnas de tipo object que contienen cadenas de texto.

Tipo de datos	Cantidad
float64	55
int64	22
object	58

Cuadro 1.2.2: Tipos de datos de la base $\overline{D}$.

Dentro de la base existe una variable que contiene los nombres de las organizaciones terroristas atacantes, denominada *gname*, y dado que se desea determinar el comportamiento de estos grupos, será considerada como la etiqueta de clase (también conocida como *"variable objetivo"*). Así que se utilizará un modelo supervisado para determinar la relación entre el resto de características y estos nombres, es decir, identificará la conducta de cada grupo perpetrador y, si se introduce una nueva entrada, la clasificará en alguno de ellos, por lo que el problema será planteado como uno de clasificación. En el caso de que el registro resulte ser atípico, es decir, aquel que no cae dentro de alguna clasificación, en otras palabras, la etiqueta de clase que se predice no pertenece a las definidas en la base, será interpretado como sin riesgo de ataque o agrupación desconocida. Por todo lo anterior, en este estudio, se asumirán las siguientes hipótesis:

- En caso de obtener un dato atípico se considerará como una situación sin riesgo de ataque, ya que si lo consideramos como un nuevo comportamiento de algún grupo terrorista, esto implica un modelo más complejo y será dejado para futuras investigaciones.

- El modelo aprenderá los comportamientos de cada grupo terrorista, la predicción se hará considerando un clasificador.

Ahora nos enfocaremos en la variable objetivo *gname*, algunos de los grupos terroristas contenidos en ella son:

- Unknown.

- Taliban.
- Islamic State of Iraq and the Levant (ISIL).
- Shining Path (SL).
- Farabundo Marti National Liberation Front (FMLN).
- Al-Shabaab.
- New People's Army (NPA).
- Irish Republican Army (IRA).
- Revolutionary Armed Forces of Colombia (FARC).
- Boko Haram.
- Kurdistan Workers' Party (PKK).
- Basque Fatherland and Freedom (ETA).
- Communist Party of India - Maoist (CPI-Maoist).
- Maoists.
- Liberation Tigers of Tamil Eelam (LTTE).
- National Liberation Army of Colombia (ELN).
- Tehrik-i-Taliban Pakistan (TTP).
- Palestinians.
- Houthi extremists (Ansar Allah).
- Al-Qaida in the Arabian Peninsula (AQAP).
- Nicaraguan Democratic Force (FDN).

Para considerar a estos grupos terroristas como las etiquetas de clase, debemos codificarlas mediante la asignación de un valor numérico a cada grupo, esto es, al grupo desconocido se le asignará el valor 0, al grupo Talibán el valor 1, y así sucesivamente hasta llegar al j-ésimo grupo. Por lo tanto, de ahora en adelante, el conjunto de etiquetas de clase será $\{1, 2, \ldots, j\}$, donde $j = 3537$, es decir, la base de datos $\overline{D}$ cuenta con 3537 diferentes grupos terroristas.

Antes de iniciar con el planteamiento del modelo, se requiere realizar una exploración a profundidad de la base en búsqueda de datos faltantes, datos atípicos, inconsistencias en los datos, etc. y, en caso de que existan, se solventen antes de someterla al aprendizaje; además, se extraerá información relevante. Todo ello se llevará a cabo en la sección siguiente.

1.3. Exploración y visualización

El objetivo de esta sección es extraer la mayor cantidad de información de la base $\overline{D}$, para ello, iniciaremos calculando las estadísticas mostradas en el Cuadro 1.3.1 para algunas variables numéricas. En ella se observa que en promedio los ataques son realizados a mediados de junio, ya que la media de *imonth* es de 6.46 y la de *iday* es de 15.5.

	iyear	imonth	iday	...	INT_ANY
Cantidad	181691	181691	181691	...	181691
Media	2002.63	6.46	15.5	...	-3.945952
Desv. Est.	13.25	3.38	8.81	...	4.691325
Mínimo	1970	0	0	...	-9
25 %	1991	4	8	...	-9
50 %	2009	6	15	...	0
75 %	2014	9	23	...	0
Máximo	2017	12	31	...	1

Cuadro 1.3.1: Estadísticas de algunas variables de la base.

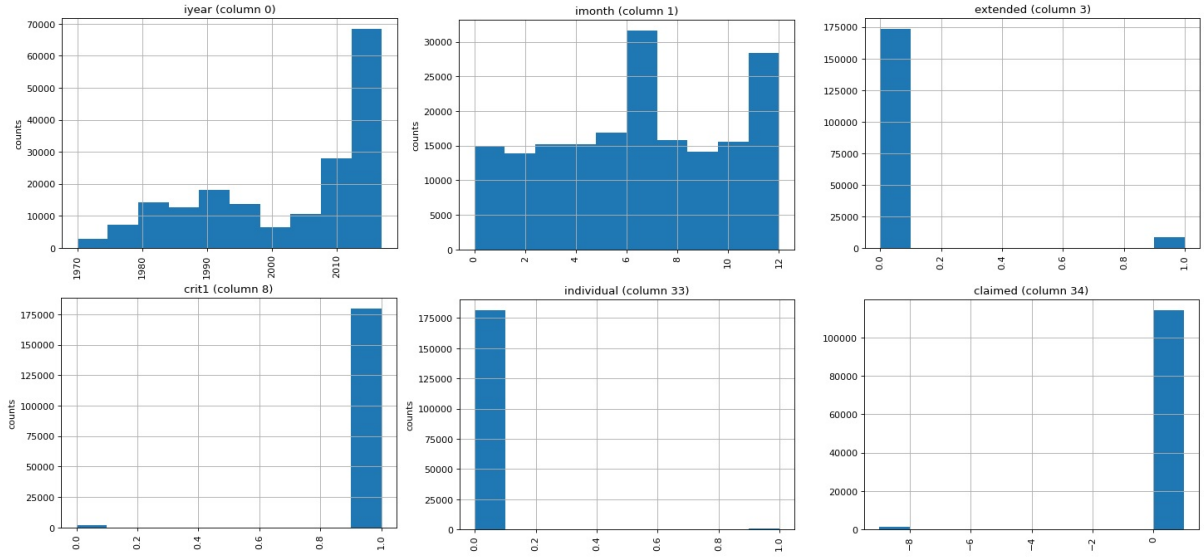

Figura 1.1: Distribución de algunas variables.

En la Figura 1.1 se muestra la distribución de algunas de las variables, obsérvese que en el gráfico superior izquierdo la variable se encuentra distribuida con un sesgo hacia la derecha, esta corresponde a la variable año, lo que muestra que a partir del año 2010 los ataques se intensificaron. La segunda corresponde al mes, se observa que los ataques son ocasionados con mayor frecuencia en junio y diciembre. También se observa que existen diversas variables binarias, tal como *individual* (inferior central), en ella se muestra que en su mayoría los valores son 0 y existe una cantidad relativamente pequeña de valores 1. Este comportamiento casi constante de estas variables nos indican que quizás su efecto con la variable de nuestro interés *gname* es mínimo, sin embargo, no se optará por eliminarlas y se realizará un análisis profundo en el Capítulo 3 sobre el efecto de estas variables en el modelo.

Private Citizens and Property	43511
Military	27984
Police	24506
Government (General)	21283
Business	20669
Transportation	6799
Utilities	6023
Unknown	5898
Religious Figures/Institutions	4440
Educational Institution	4322
Terrorists/Non-State Militia	3039
Journalists and Media	2948
Violent Political Party	1866
Airports and Aircraft	1343
Telecommunication	1009
NGO	970
Tourists	440
Maritime	351
Food or Water Supply	317
Abortion Related	263
Other	137

Cuadro 1.3.2: Distribución de objetivos de ataques globales.

Ahora, nos enfocaremos en el análisis individual de las variables. Iniciaremos con el Cuadro 1.3.2, en donde se muestran los principales objetivos de los ataques, en primer lugar, son dirigidos hacia ciudadanos y propiedades particulares, seguido de objetivos militares, policía y gobierno. Para tener una mejor visualización de estas cantidades se puede observar la Figura 1.2, donde los ataques a periodistas se encuentran en el lugar

número 13 y le siguen los aeropuertos y telecomunicaciones. En último lugar se encuentran los objetivos relacionados con el aborto.

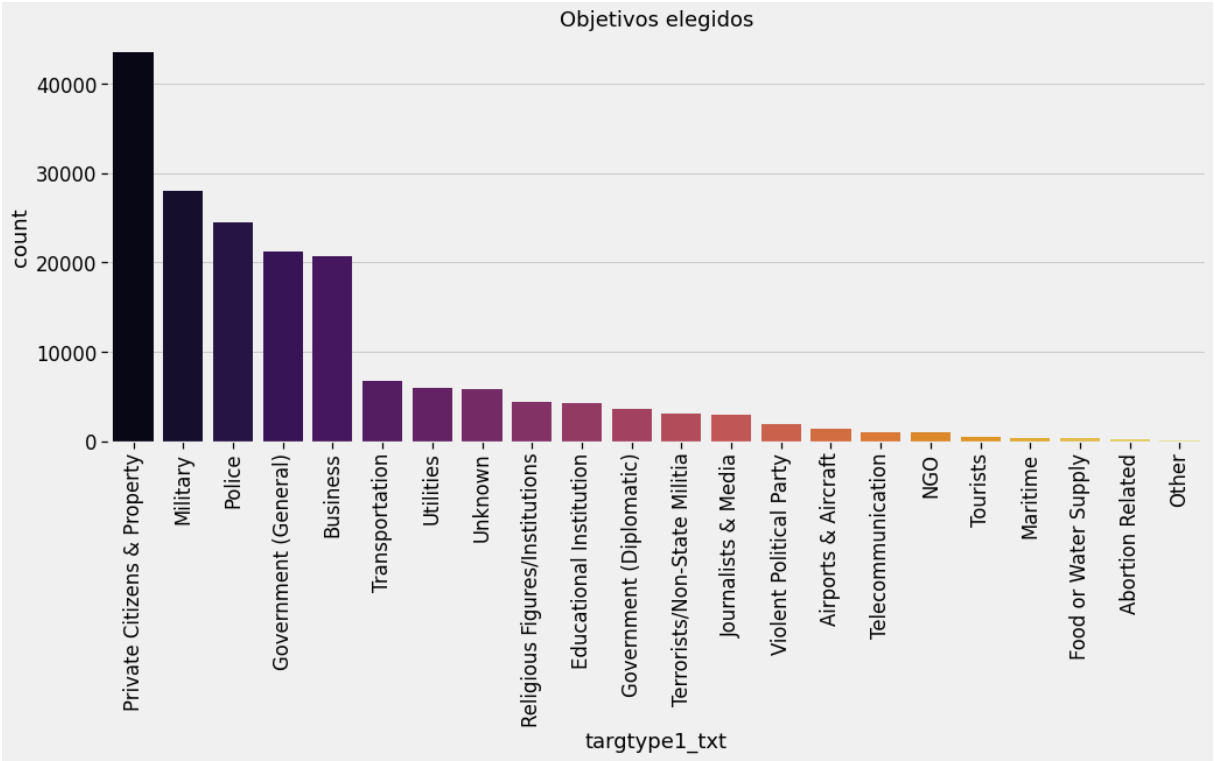

Figura 1.2: Frecuencia de objetivos elegidos por los grupos terroristas.

Unknown	82782
Taliban	7478
Islamic State of Iraq and the Levant (ISIL)	5613
Shining Path (SL)	4555
Farabundo Marti National Liberation Front (FMLN)	3351
Al-Shabaab	3288
New People's Army (NPA)	2772
Irish Republican Army (IRA)	2671
Revolutionary Armed Forces of Colombia (FARC)	2487
Boko Haram	2418
Kurdistan Workers' Party (PKK)	2310
Basque Fatherland and Freedom (ETA)	2024
Communist Party of India - Maoist (CPI-Maoist)	1878
Maoists	1630
Liberation Tigers of Tamil Eelam (LTTE)	1606
National Liberation Army of Colombia (ELN)	1561
Tehrik-i-Taliban Pakistan (TTP)	1351
Palestinians	1125
Houthi extremists (Ansar Allah)	1062
Al-Qaida in the Arabian Peninsula (AQAP)	1020
Nicaraguan Democratic Force (FDN)	895

Cuadro 1.3.3: Distribución de ataques por grupo terrorista a nivel global.

En el Cuadro 1.3.3 se muestra la distribución de ataques por los principales 21 grupos terroristas. En primer lugar, se tiene que el grupo Taliban es el que ha cometido una mayor cantidad de ataques, seguido por Islamic State of Iraq and the Levan (ISIL) y, en tercer lugar, el grupo Shining Path (SL). El grupo Al-Qaida in the Arabian Peninsula (AQAP) se encuentra en penúltimo sitio con 1020 ataques. En la Figura 1.3, se puede observar

una representación en forma de gráfica de barras de los principales grupos terroristas con mayor incidencia de ataques.

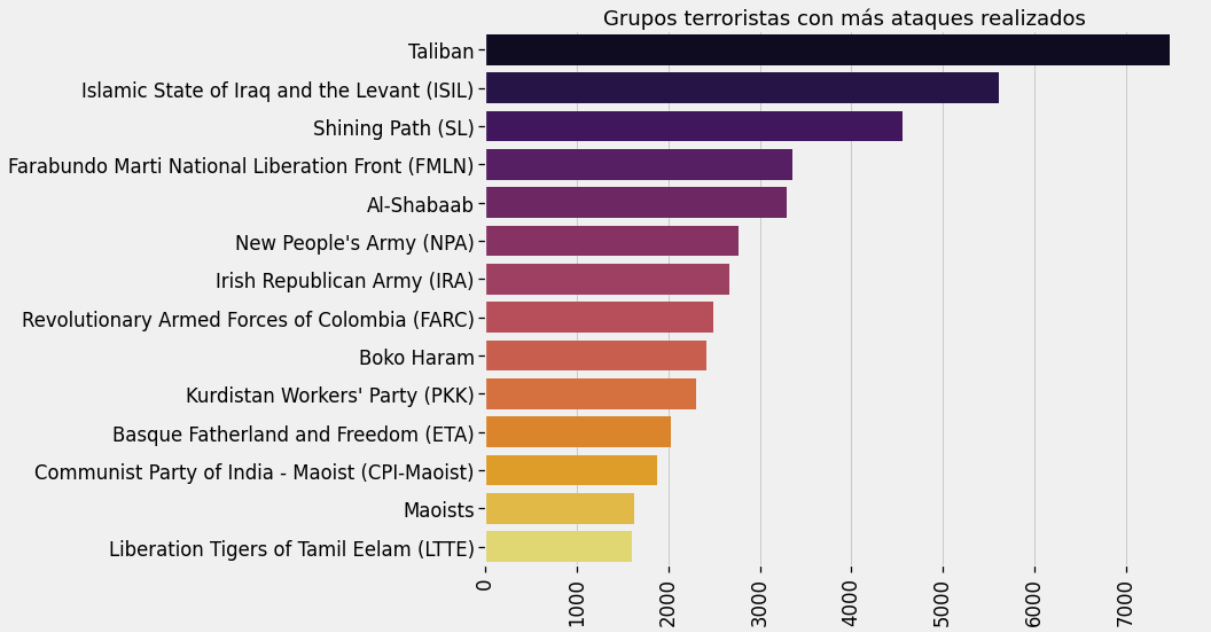

Figura 1.3: Grupos terroristas con mayor número de ataques realizados a nivel global.

En la base de datos $\overline{D}$, de los $n = 181,691$ registros de ataques terroristas se tiene que un 88.96 % fueron exitosos y un 11.04 % fracasaron, es decir, una gran parte de los ataques terroristas registrados cumplieron su objetivo.

23rd of September Communist League	44
Zapatista National Liberation Army	23
Popular Revolutionary Army (Mexico)	22
Union of the People (UDP)	21
Institutional Revolutionary Party (PRI)	14
Democratic Revolutionary Party	12
Pagan Sect of the Mountain	9
Independent Peasants Union	5
Gunmen	4
National Front for the Liberation of Cuba (FLNC)	4
Revolutionary Worker Clandestine Union of the People Party (PROCUP)	3
Animal Liberation Front (ALF)	3
Fuerzas Armadas Revolucionarias del Pueblo (FARP)	3
Revolutionary Student Front	3
Nat. Ind. Committee for Political Prisoners and Persecuted and Missing Persons	2
Francisco Villa People's Front	2
Southern Sierra Peasant Organization	2
Individuals Tending Toward Savagery	2
Zetas	2
Militant Peasants (NFI)	2

Cuadro 1.3.4: Distribución de ataques por grupo terrorista en México.

En el Cuadro 1.3.4 se observan los ataques realizados en México, en primer lugar se encuentra 23rd of September Communist League, seguido del Zapatista National Liberation Army y en penúltimo lugar el grupo de los Zetas y, finalmente, en último lugar el grupo de Militant Peasants (NFI) con 2 ataques. Cabe mencionar que dentro de la lista se encuentra el Partido Revolucionario Institucional (PRI) en quinto lugar. En la Figura 1.4 se muestra una gráfica con el número de ataques terroristas ocurridos en México en el

transcurso de los años 1970 hasta 2017, los resultados mostraron que el año en que ocurrió un mayor número de ataques fue en en el periodo de 1994 a 1997, en el surgimiento del Ejercito Zapatista de Liberación Nacional (EZLN).

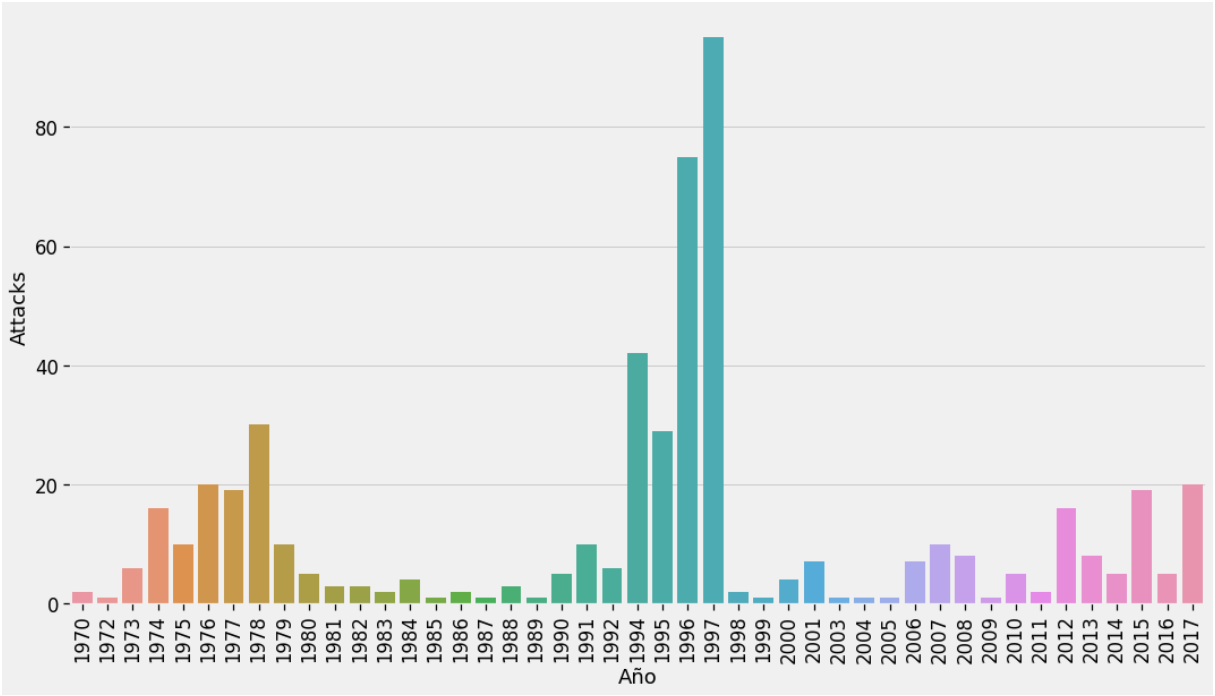

Figura 1.4: Número de ataques terroristas ocurridos en México durante los años 1970 y 2017.

También se puede obtener información similar de otro país en particular, por ejemplo, en el Cuadro1.3.5 se observan los ataques realizados en Estados Unidos, en primer lugar se encuentra Extremistas Antiaborto, seguido de Militantes de Izquierda y en penúltimo lugar el grupo Frente Unido por la Libertad (UFF), y finalmente, en último lugar el grupo de Extremistas Anti-Muslim con 27 ataques. En el Cuadro 1.3.6 se muestran los registros de los atentados ocurridos en Estados Unidos de Norte América el 11 de septiembre de 2001. Obsérvese que se llevaron a cabo en New York, Arlington y Snanksville, este último debido a que no llegó a su destino planteado originalmente, que al parecer era la Casa Blanca, se informa que los pasajeros lucharon contra los secuestradores y por ello el avión se estrella antes de llegar a su objetivo. También se puede observar el número de bajas

en cada evento, en el último caso, se trata del número de pasajeros y tripulación a bordo, además se cuenta con un estimado total de 1,384 y de 1,383 pérdidas humanas en cada uno de los ataques a las torres gemelas del World Trade Center. En el Cuadro 1.3.7 se presentan más detalles del primer evento. Estos ataques son muy conocidos en todo el mundo por el nivel de violencia utilizada, daños materiales, pérdidas humanas, entre otros.

Grupo	Cantidad
Anti-Abortion extremists	196
Left-Wing Militants	169
Fuerzas Armadas de Liberacion Nacional (FALN)	120
White extremists	87
New World Liberation Front (NWLF)	86
Black Nationalists	83
Animal Liberation Front (ALF)	76
Jewish Defense League (JDL)	74
Student Radicals	71
Earth Liberation Front (ELF)	66
Omega-7	54
Weather Underground, Weathermen	45
Macheteros	37
Anti-Government extremists	36
Black Liberation Army	36
Chicano Liberation Front	31
Armed Revolutionary Independence Movement (MIRA)	30
Jihadi-inspired extremists	30
United Freedom Front (UFF)	29
Anti-Muslim extremists	27

Cuadro 1.3.5: Distribución de ataques por grupo terrorista en Estados Unidos.

País	Ciudad	Número de bajas
New York City	Private Citizens y Property	1384
New York City	Private Citizens y Property	1383
Arlington	Government (General)	190
Shanksville	Private Citizens y Property	44

Cuadro 1.3.6: Ataques ocurridos en Estados Unidos el 11 de septiembre de 2001.

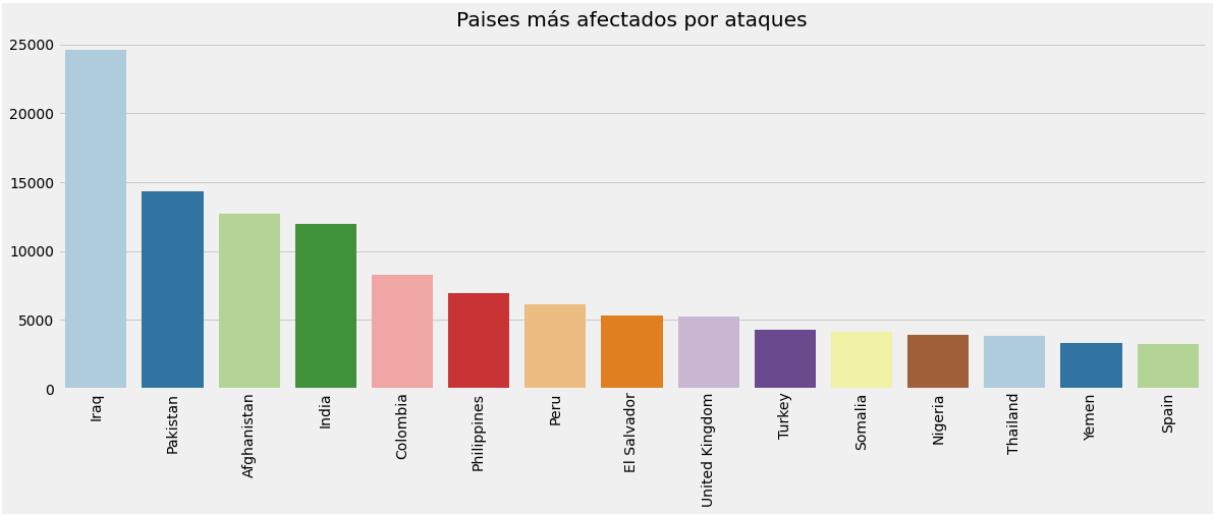

Figura 1.5: Quince principales países con mayor cantidad de ataques terroristas.

Year	2001
Month	9
Day	11
Country	United States
Region	North America
city	New York City
latitude	40.697132
longitude	-73.931351
AttackType	Hijacking
Killed	1384.0
Wounded	8190.0
Target	Passengers and crew members on American Airlines...
Summary	09/11/2001: This was one of four related attack...
Group	Al-Qaida
Target_type	Private Citizens y Property
Weapon_type	Vehicle (not to include vehicle-borne explosive...
Motive	Unknown
Property_Damage	Catastrophic (likely $\geq$ $ 1 billion)
success	1
casualties	9574.0

Cuadro 1.3.7: Registro del ataque a una de las torres gemelas en New York el 11 de septiembre de 2001.

Por otro lado, en la Figura 1.5 se brinda una visualización gráfica de los quince países más afectados por el terrorismo, en donde el país con el número más alto de ataques es Iraq, seguido de Pakistan, Afghanistan, India, Colombia, Filipinas, Perú, El Salvador, El Reino Unido, Turquía, Somalia, Nigeria, Tailandia, Yemen y España. Además, en el diagrama de la Figura 1.6 se muestra el número de muertes que han generado los ataques

ocurridos en cada uno de los principales países afectados, se observa que Iraq es el país con una mayor cantidad de muertes a causa de estos, provocando un deceso de más de 17500 personas. En la Figura 1.5 no se aprecia, pero en la exploración se obtuvo que Iraq es el país donde por un solo ataque se registró el mayor número de bajas civiles a nivel global con un total de 1570 decesos.

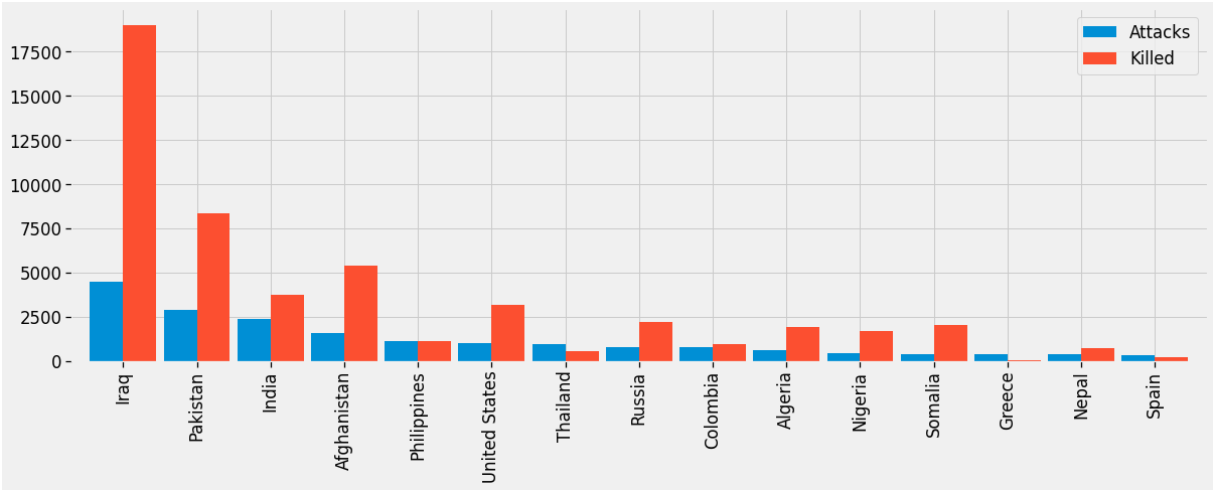

Figura 1.6: Comparación del número de muertes con el número de ataques en cada uno de los principales países.

En la Figura 1.7 se muestra el número de ataques ocurridos en Iraq durante los años de 1975 a 2017, estos empezaron a aumentar durante el año 2003, siendo el año 2014 donde se tuvo el mayor número de eventos en este país.

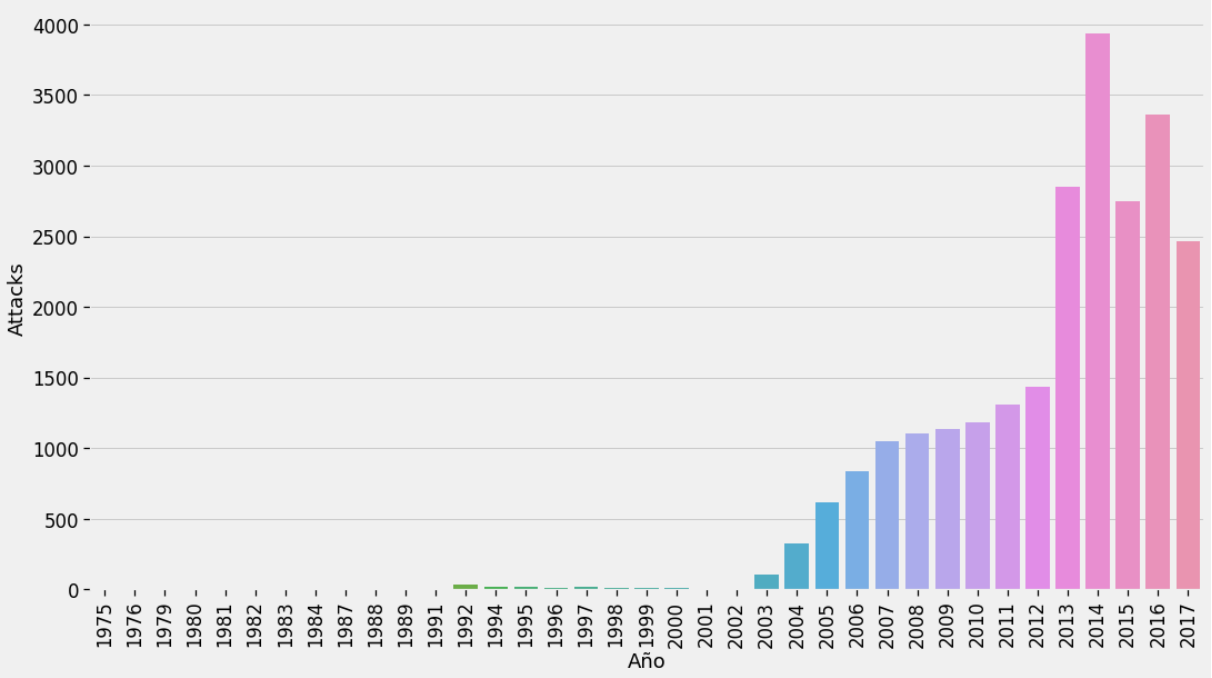

Figura 1.7: Número de ataques ocurridos en Iraq desde el año 1998 a 2017.

En un panorama más general, se puede extraer también información sobre las 12 regiones que tiene la base de datos $\overline{D}$, por ejemplo, en el Cuadro 1.3.8 se muestra una distribución de ataques que fueron realizados por los principales 20 grupos terroristas a nivel global en la región de América Central y el Caribe, en este caso se puede notar que, de los principales 20 grupos terroristas, 3 de ellos atacaron esta región, donde el grupo Frente Farabundo Martí para la Liberación Nacional (FMLN) ocupa el primer lugar con $3,351$ ataques seguido de Fuerza Democrática Nicaragüense (FDN) con 894 ataques. De manera similar, en el Cuadro 1.3.9 se puede observar que 7 de los 20 principales grupos terroristas atacaron en la región de Europa Oriental, siendo el Ejército Republicano Irlandés (IRA), junto con Patria Vasca y Libertad (ETA), los dos principales grupos que afectan esta región con $2,668$ y $2,022$ ataques, respectivamente.

Grupo	Cantidad
Farabundo Marti National Liberation Front (FMLN)	3351
Nicaraguan Democratic Force (FDN)	894
Revolutionary Armed Forces of Colombia (FARC)	3

Cuadro 1.3.8: Principales grupos terroristas a nivel global que atacaron América Central y el Caribe.

Grupo	Cantidad
Irish Republican Army (IRA)	2668
Basque Fatherland and Freedom (ETA)	2022
Kurdistan Workers' Party (PKK)	173
Islamic State of Iraq and the Levant (ISIL)	15
Palestinians	14
Liberation Tigers of Tamil Eelam (LTTE)	2
Al-Qaida in the Arabian Peninsula (AQAP)	2

Cuadro 1.3.9: Principales grupos terroristas a nivel global que atacaron Europa Oriental.

En la Figura 1.8 se muestra que la región con una mayor incidencia de ataques terroristas es el Medio Oriente y el norte de África, donde ocurrieron más de 50,000 ataques terroristas, mientras que la región del sur de Asia ocupa el segundo lugar con más de 40,000 ataques, seguidos a su vez de Sudamérica, África Sub-Sahara, Europa del Oeste, Sureste de Asia, América central y el Caribe, Europa del Este, Norteamérica, el Este de Asia, Asia Central y, finalmente, Australia y Oceanía son las regiones con menor incidencia.

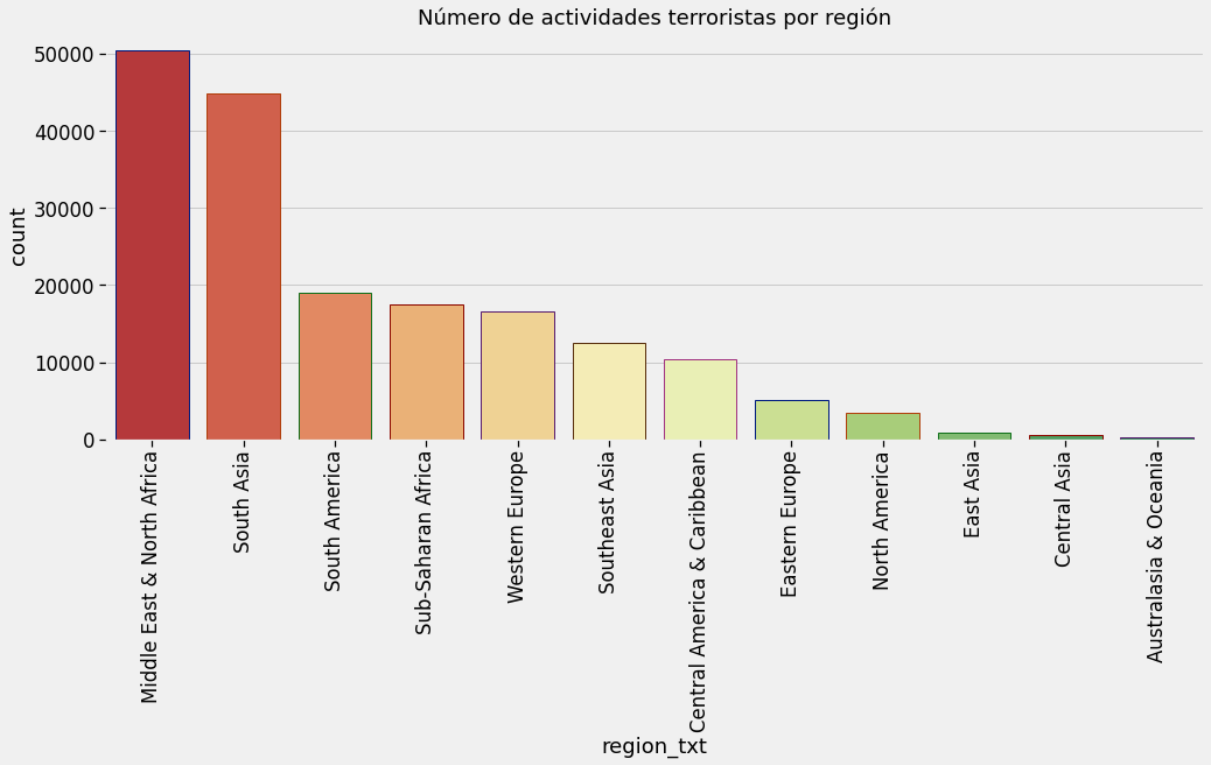

Figura 1.8: Cantidad de ataques terroristas ocurridos por región.

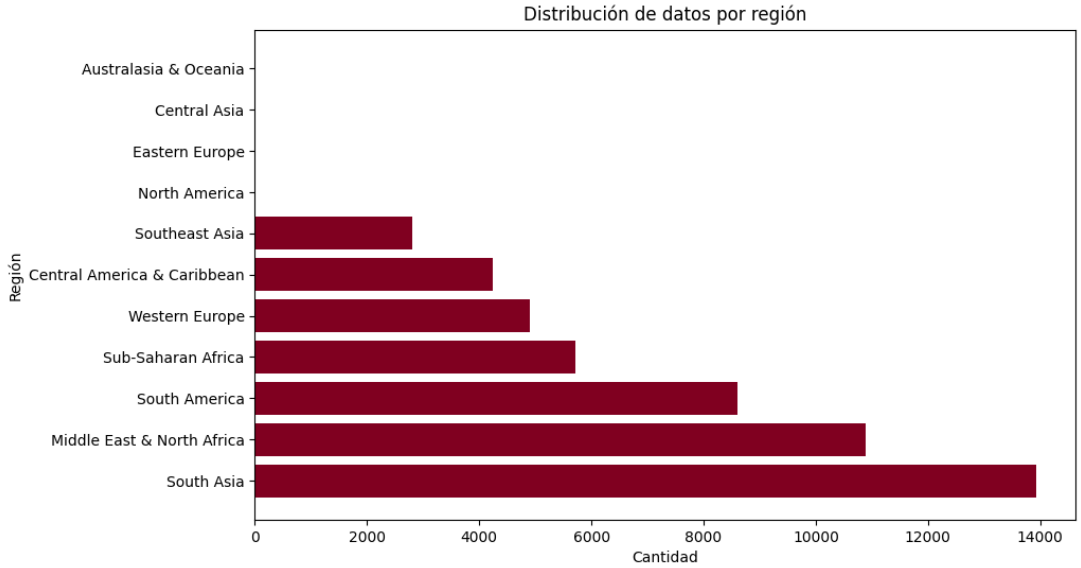

Figura 1.9: Distribución de ataques de los principales grupos terroristas a nivel global en cada una de las 12 regiones.

Por otro lado, en la Figura 1.9 se pueden observar las regiones ordenadas por número de ataques cometidos por los 20 principales grupos terroristas, donde se tiene que la región más afectada es Sur de África junto con Medio Oriente y África del Norte, mientras que la región menos afectada por estos grupos es Europa del Este, Asia Central, Australasia y Oceanía.

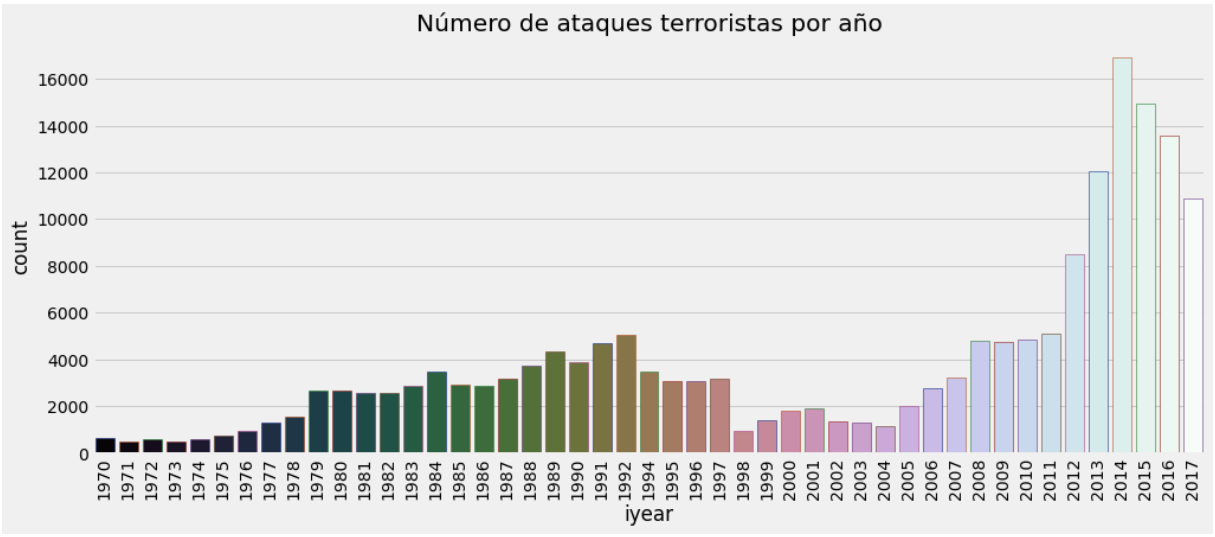

Figura 1.10: Número de ataques terroristas ocurridos a nivel global.

Así mismo, y de una manera más general, se puede ver en la Figura 1.10 el número de ataques ocurridos en todo el mundo desde el año 1970 al año 2017, donde se presenta un incremento a partir de 1977, alcanzando un máximo en 1992 y disminuyendo en el año 1998, sin embargo, a partir de este año, nuevamente se incrementan de forma considerable hasta alcanzar en el año 2014 un pico de más de 16,000 eventos, a partir de este año hasta el 2017 hubo un decrecimiento hasta llegar a los 10,000 ataques.

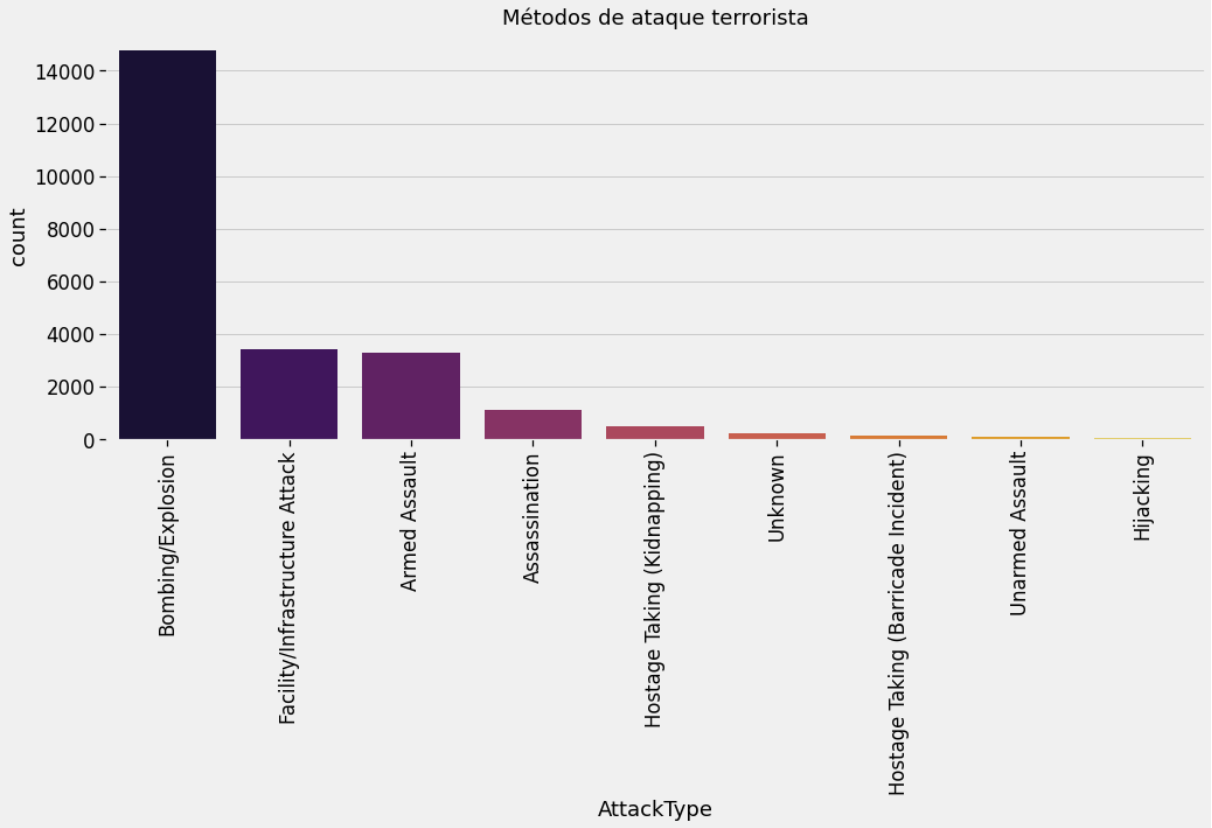

Figura 1.11: Frecuencia de los métodos utilizados en los ataques terroristas.

En el diagrama de la Figura 1.11 se presentan los distintos métodos utilizados y la cantidad de ataques terroristas que hicieron uso de cada método. Se identificó que el método más utilizado es el bombardeo/explosión, seguido de incursiones a instalaciones o ataques a infraestructura. En tercer lugar se encuentran los asaltos armados, seguidos de los asesinatos y raptos o toma de rehenes.

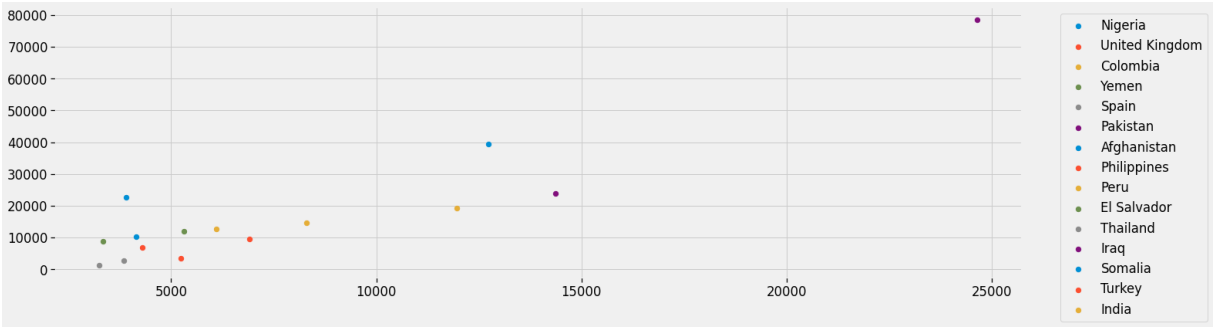

Figura 1.12: Relación entre el número de ataques con el número de bajas en algunos países.

En el diagrama de dispersión de la Figura 1.12 se presenta la relación entre el número de ataques comparado con el número de bajas en los países más azotados por el terrorismo, siguiendo una tendencia lineal siendo Iraq el que presenta una mayor cantidad de ambos, por lo que se puede concluir que en Nigeria han sido más efectivos los ataques, ya que una cantidad menor (3,907) ha provocado un número grande de bajas (22,682), lo mismo ocurre con Afganistán en comparación con Pakistan e India, finalmente, Iraq muestra un gran número de ataques junto con un gran número de bajas.

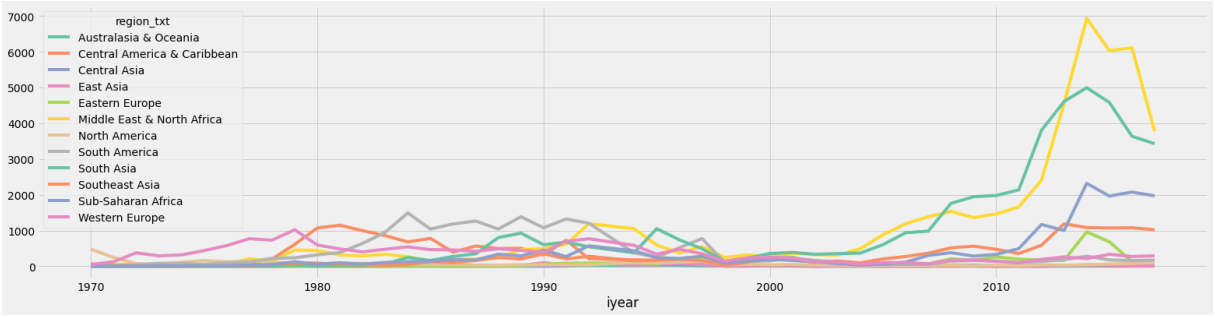

Figura 1.13: Evolución del fenómeno por regiones a lo largo del tiempo.

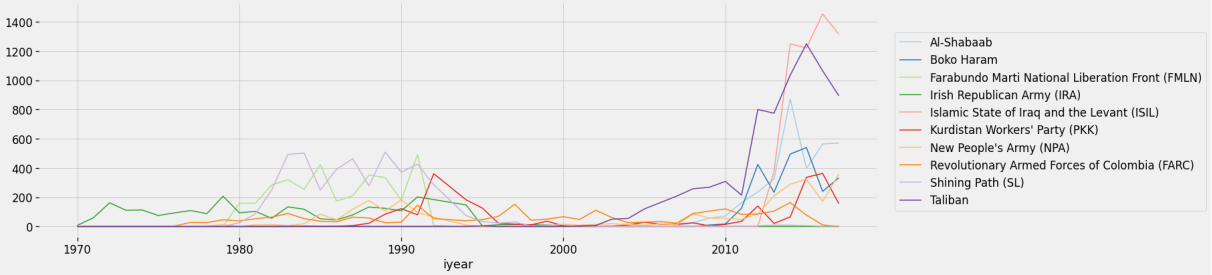

Figura 1.14: Evolución del fenómeno por grupos terroristas a lo largo del tiempo.

En la Figura 1.13 se puede apreciar la evolución que han tenido los ataques terroristas a lo largo de los años en cada una de las regiones mencionadas en la base de datos. Se tiene que entre los años 1970 y antes del año 2000 el número de ataques registrados era mucho menor en cada una de las regiones, comparado con el número de ataques que se registraron después del año 2000. Principalmente regiones como Australasia y Oceanía, Medio Oriente y África del Norte registraron un considerable aumento a partir del 2005. Así mismo, en la Figura 1.14 se muestra la evolución que han tenido los 10 principales grupos terroristas a través del tiempo, destacando los grupos Farabundo Marti National Liberation Front $FMLN$ y Shining Path SL con un mayor número de ataques antes del año 2000, mientras que a partir del año 2000 los grupos Islamic State of Iraq and the Levant $ISIL$ y Taliban superaron a los grupos antes mencionados con un incremento mayor de ataques.

En el mapa mundi de la Figura 1.15 se ilustra la distribución geográfica de los eventos a nivel global. Los puntos en color azul representan ataques en los cuales se dio un número de bajas menor a 75 civiles, mientras que los puntos en color rojo representan ataques donde ocurrió un número mayor a 75 bajas por ataque. A través de él se puede observar las zonas que presentan un mayor número de agresiones.

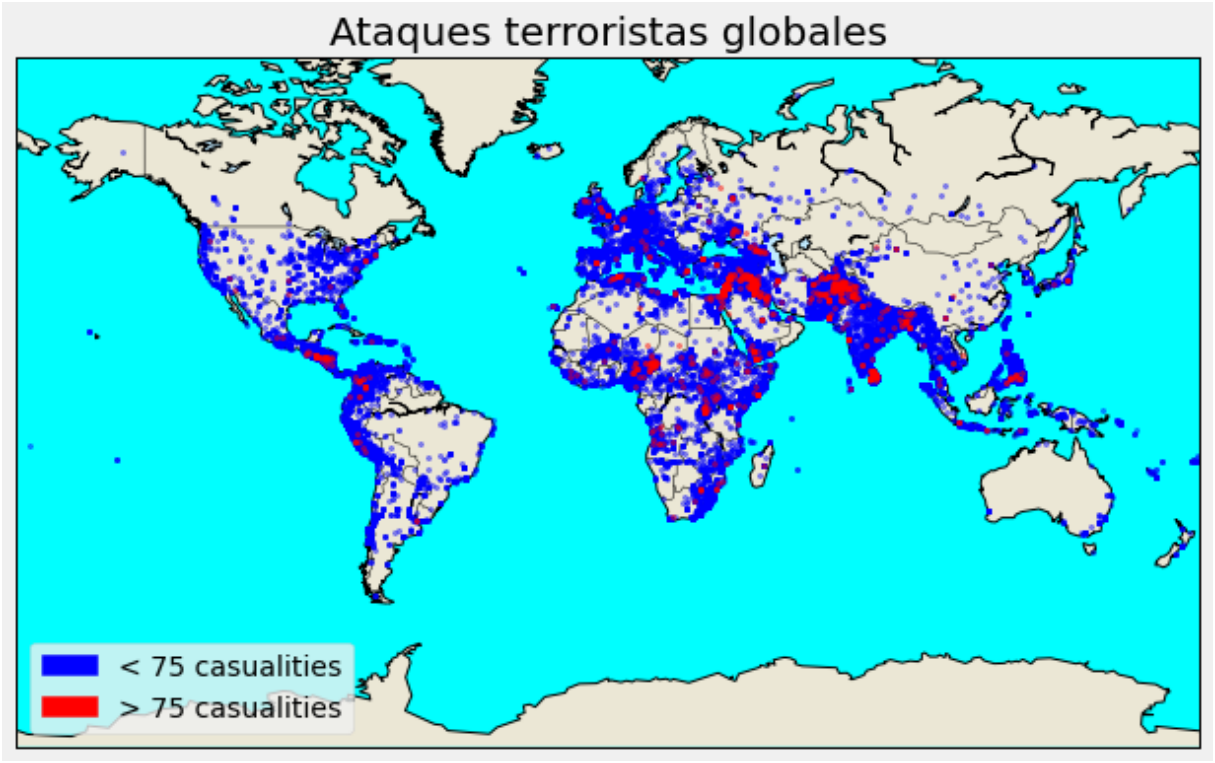

Figura 1.15: Visualización de ataques terroristas ocurridos a nivel global.

De una manera similar, se puede observar en el globo terráqueo de la Figura 1.16 una segunda visualización de los ataques ocurridos en todo el mundo, siguiendo la misma regla de colores antes mencionada. Obsérvese que en México la mayor parte de eventos se han realizado en el centro y sur del país.

Con todo lo anterior, se tiene un comprensión del fenómeno y se han realizado algunas conclusiones parciales. El siguiente paso es realizar un pre-procesamiento de la base, que inicia con el capítulo siguiente.

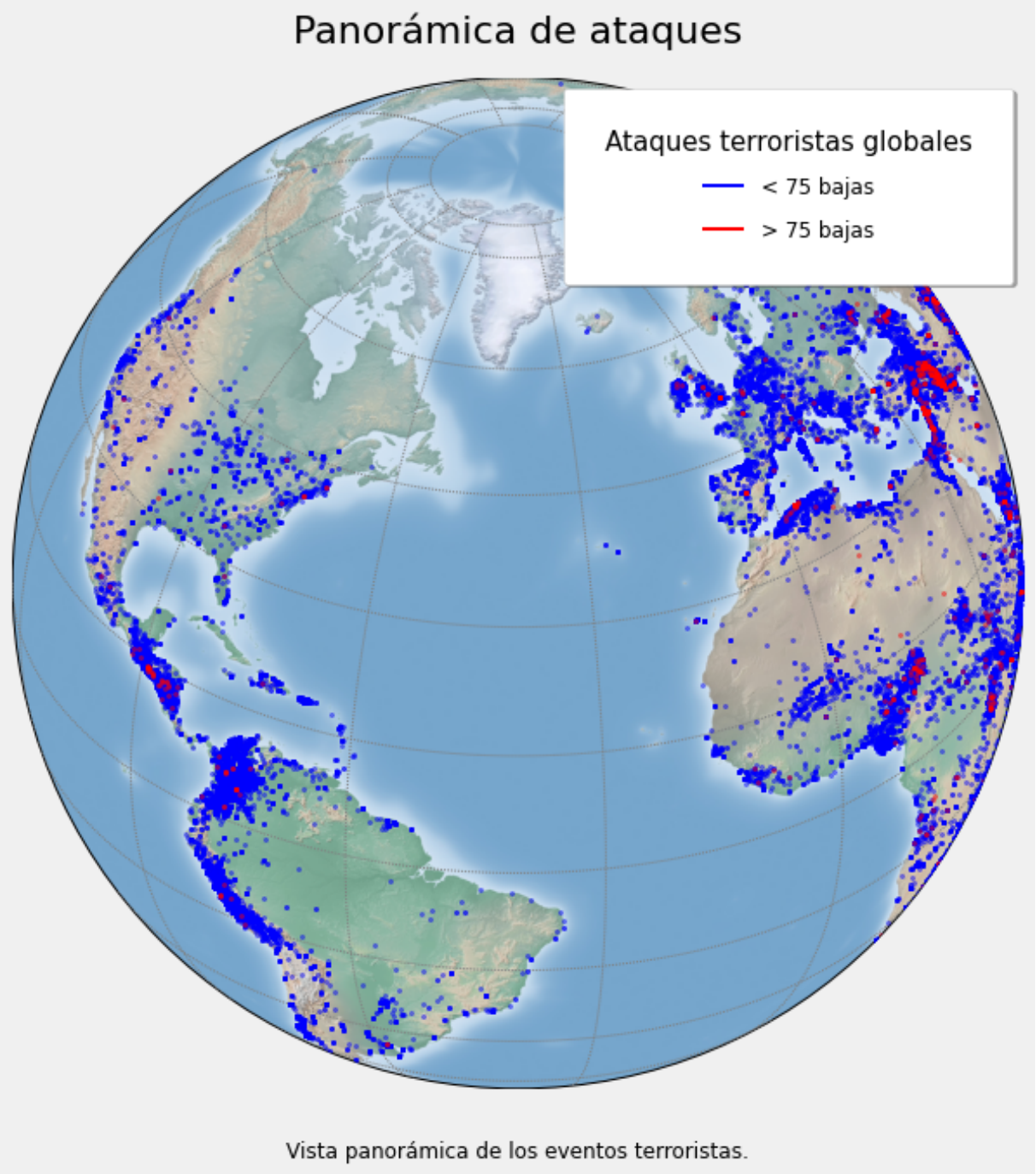

Figura 1.16: Globo terráqueo representando la distribución de ataques terroristas.

Capítulo 2

Imputación de datos

La base de datos $\overline{D}$ contiene un total de $n = 181,691$ registros de ataques terroristas, de los cuales 180,800 tienen fechas conocidas, mientras que 891 registros no cuentan con esa información, esto es, presentan fechas desconocidas, esta información faltante representa el 0.49 %. Esto motiva a verificar si estos datos influyen en los resultados del modelo en estudio y, de ser así, cómo se debe adecuar dicho modelo para manejar estos datos desconocidos. Por lo que uno de los objetivos de este capítulo es determinar qué registros cuentan con entradas desconocidas y brindar una solución a este problema mediante un método de imputación.

2.1. Datos perdidos

Es común en ciencia de datos encontrar bases con entradas perdidas por el propio proceso de recopilación o que incluso la misma naturaleza del problema no permita contar con ellos. Formalmente, un dato perdido se conceptualiza como en la Definición 2.1.1.

Definición 2.1.1. *Sea* $\overline{D} = \{x_{ij}\}$; $i = 1, \ldots, n$; $j = 1, \ldots, d$ *un conjunto de datos de dimensión* $n \times d$. *Denotamos la* i*-ésima fila de* $\overline{D}$ *por* $\overline{X}_i = (x_i^1, x_i^2, \ldots, x_i^d)$. *Se dice que* x_{ij} *es un dato perdido si este no es observado.*

La principal consecuencia de la existencia de datos faltantes es que pueden alterar la eficiencia en los métodos de aprendizaje máquina. En ocasiones, se deben ajustar para robustecerlos ante la ausencia de ellos. En el Cuadro 2.1.1 se muestra que existen variables con un alto porcentaje de datos perdidos para alguna de las variables, lo que dificulta o impide el diseño de un modelo que soporte tal grado de entradas vacías. Para solucionar este problema, se siguen tres estrategias:

1. Cualquier registro de datos que contenga una entrada ausente puede eliminarse por completo. Sin embargo, este enfoque puede no ser práctico cuando la mayoría de los registros contienen entradas que faltan, como lo es el presente caso. Por otro lado, si la variable tiene una alto porcentaje de valores faltantes, se puede considerar como no relevante y puede ser candidata a eliminarse.

2. Los valores que faltan pueden imputarse. Aunque existe la limitante de que los errores creados por el proceso de imputación puedan afectar a los resultados del algoritmo de extracción de características que se realizará en el Capítulo 3.

3. La fase analítica está diseñada de forma que pueda trabajar con valores perdidos. Muchos métodos de minería de datos están diseñados para trabajar con valores perdidos. Este enfoque suele ser el más deseable porque evita los sesgos adicionales inherentes al proceso de imputación. Sin embargo, como se mencionó anteriormente, se desconoce el umbral aceptable de entradas faltantes. Se plantea este estudio para posteriores investigaciones.

En la Figura 2.1 se muestra la distribución de los datos perdidos de cada variable. El color amarillo representa las entradas faltantes y el color morado representa los datos conocidos. Observemos que existen variables con relativamente poca información. Como se observa, se debe elegir alguna estrategia para tratar con este problema, por lo que, con el objetivo de evitar errores con el manejo de las variables con alto porcentaje de datos perdidos, haremos uso de la primera estrategia antes mencionada, es decir, la eliminación de las columnas que presenten un porcentaje mayor al 25 % de entradas vacías, esto

	Total	Porcentaje de datos faltantes
gsubname3	181671	99.988992
weapsubtype4_txt	181621	99.961473
weapsubtype4	181621	99.961473
weaptype4	181618	99.959822
weaptype4_txt	181618	99.959822
...	...	...
suicide	0	0.000000
success	0	0.000000
crit3	0	0.000000
property	0	0.000000
eventid	0	0.000000

Cuadro 2.1.1: Porcentaje de datos faltantes de algunas de las variables.

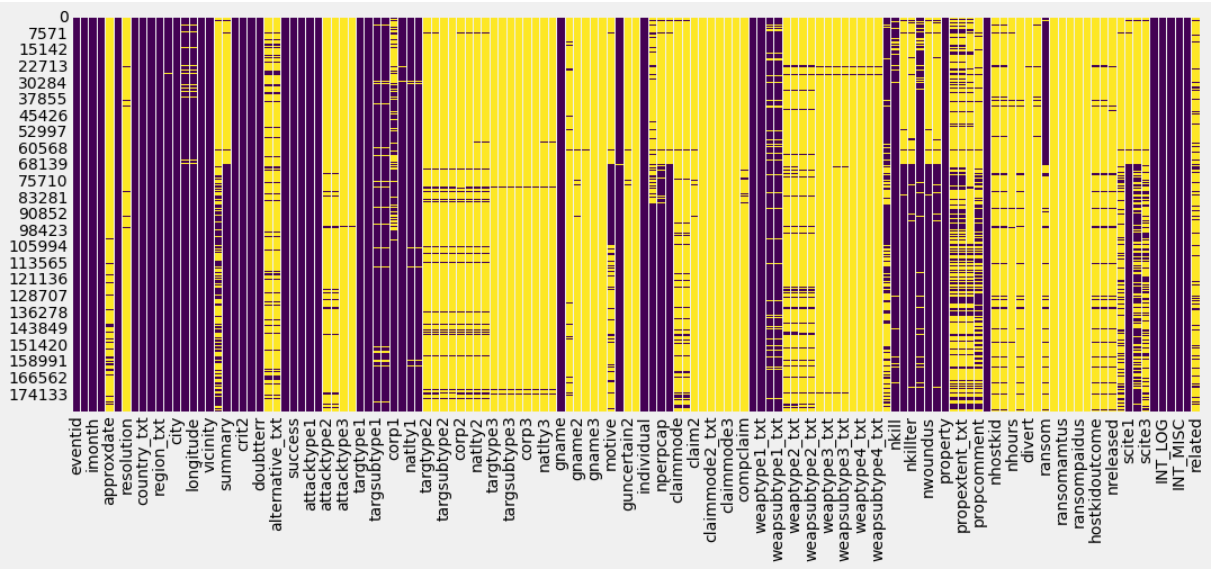

Figura 2.1: Visualización gráfica de los datos perdidos.

debido a que, si se opta por llenar estos espacios mediante alguna técnica que genere nuevas entradas para los datos faltantes (por ejemplo, completar con la media de los datos

conocidos), se pueden introducir errores en los siguientes procesos, incluso en el modelo; de hecho, tratar de hacer imputación nos llevaría a un análisis más profundizado en este tema y la literatura no recomienda imputar una cantidad relativamente grande con pocos datos. No existe un límite establecido en la literatura con respecto a un porcentaje aceptable de datos faltantes en un conjunto de datos para inferencias estadísticas válidas (véase [5]). Por ejemplo, Schafer ([15]) afirmó que una tasa faltante del 5 % o menos no tiene consecuencias. Bennett ([2]) sostuvo que es probable que el análisis estadístico esté sesgado cuando falta más del 10 % de los datos. Además, la cantidad de datos faltantes no es el único criterio por el cual un investigador evalúa el problema de los datos faltantes, en este sentido, Tabachnick y Fidell ([18]) postularon que los mecanismos de datos faltantes y los patrones de datos faltantes tienen un mayor impacto en los resultados de la investigación que la proporción de datos faltantes. En [10, pág. 12] se menciona que *"no se recomienda imputar datos en situaciones en que la omisión en una o más variables alcance porcentajes superiores al 20 %"*. Es un tema interesante que, en esta ocasión, se plantea para futuras investigaciones.

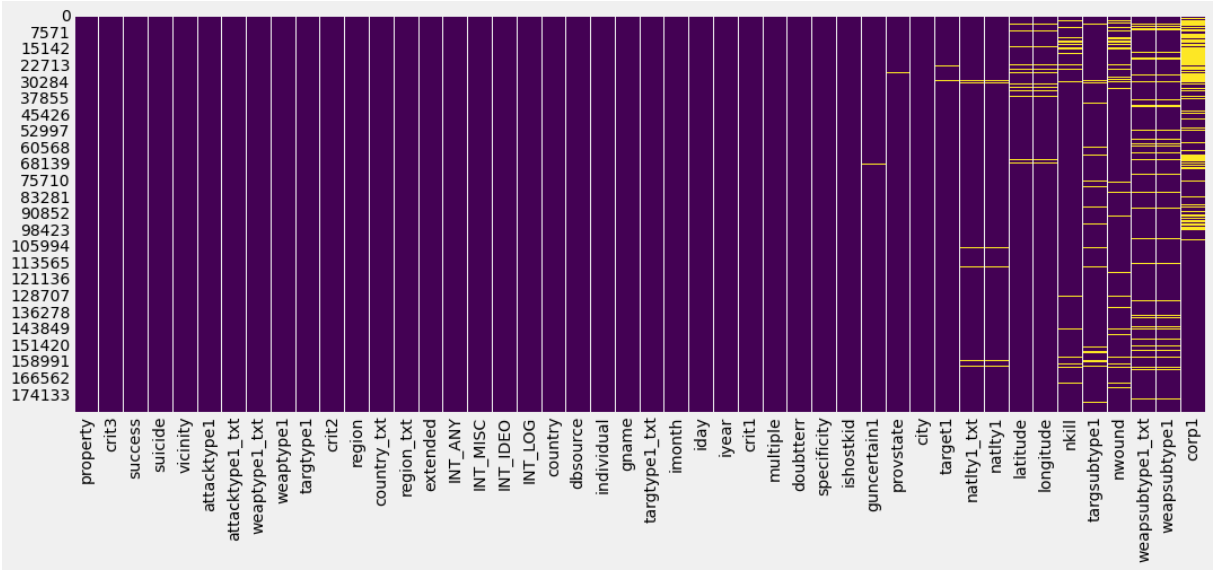

Figura 2.2: Visualización gráfica de la base después de la eliminación de variables con una presencia de más del 25 % de entradas faltantes.

Al eliminar las variables que presentan un porcentaje mayor al 25 % de entradas perdidas, directamente se conservan sólo aquellas que tienen un porcentaje menor o igual al 25 % de celdas vacías, estas variables se pueden observar mediante la representación gráfica mostrada en la figura 2.2.

Ahora, a las variables restantes les aplicaremos la segunda estrategia que se mencionó previamente, esto es, realizaremos una imputación a cada una de ellas, este proceso será realizado a partir de la siguiente sección mediante el método denominado *expectation maximization*.

2.2. Algoritmo EM

En esta sección se describe el algoritmo *expectation maximization*, conocido como el algoritmo EM. El algoritmo EM es un método iterativo cuyo propósito general es encontrar estimaciones de máxima verosimilitud en modelos paramétricos para datos incompletos. Dentro de cada iteración del algoritmo hay dos pasos, llamados el paso de expectativa o paso E y el paso de maximización o paso M. El nombre del algoritmo EM fue dado por Dempster, Laird y Rubin en 1977 (véase [4]), quienes proporcionaron una formulación general y unificada del algoritmo, mostraron sus propiedades básicas y proporcionaron diversos ejemplos y aplicaciones del mismo.

La idea fundamental es asociar al problema de datos incompletos un problema de datos completos, de tal manera que la estimación por máxima verosimilitud sea computacionalmente más manejable. La idea es estimar a partir de los datos conocidos la función de probabilidad para posteriormente generar un número aleatorio y suplir al dato desconocido (paso E), luego se aplica el proceso M para verificar que el dato es el adecuado y, en caso de que no lo sea, cambiarlo por otro aleatorio, finalizado este proceso, el nuevo dato pasa a ser parte de los conocidos y el proceso se repite hasta llenar todos los vacíos. Se parte de los valores iniciales de los parámetros para la función de distribución, que pueden ser la media y la desviación estándar, luego los pasos E y M se repiten hasta la convergencia. Específicamente, dado un conjunto de estimaciones de parámetros, como la

media y la matriz de covarianza para un entorno normal multivariante, el paso E calcula la expectativa condicional de la log-verosimilitud de los datos completos dados los datos observados y las estimaciones de los parámetros. Este paso a menudo se reduce al cálculo de estadísticas simples. Dada la log-verosimilitud de los datos completos, el paso M encuentra las estimaciones de los parámetros para maximizar la log-verosimilitud de los datos completos del paso E.

Formalmente, mostraremos el planteamiento matemático. Supongamos que se tiene una muestra de d vectores columna aleatorios idénticamente distribuidos de dimensión $n \times 1$, denotados por X_i (datos conocidos). El objetivo es estimar algún vector de parámetros θ de la función de distribución f de X_i, que se puede suponer normal (véase [5]). La imputación llena los datos vacíos, que denotamos por X^m, utilizando los datos observados, denotados por X^o, de diversas maneras, luego, los datos completos son utilizados para estimar θ.

Si conocemos la distribución de $X_i = (X_i^o, X_i^m)$, con vector de parámetros θ, entonces podremos imputar X_i^m a partir de un valor aleatorio generado de la distribución condicional

$$f(X_i^m | X_i^o, \theta).$$

El objetivo es muestrear valores a partir de la distribución a priori. Ya que no conocemos θ, debemos estimarlo de los datos, denotemos esta estimación como $\hat{\theta}$, y entonces utilizamos

$$f(X_i^m | X_i^o, \hat{\theta})$$

para imputar los datos perdidos.

El método se basa en la función de log-verosimilitud de los datos completos, que se define como

$$l(\theta, X_i) = K_i - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu),$$

donde X_i es el vector de datos observados para la variable i, K_i es una constante que se determina por el número de variables observadas, y μ y Σ son, respectivamente, el vector con entradas igual a la media y la matriz compuesta de la varianza en la diagonal

y ceros fuera de ella, que se deben estimar y son las entradas correspondientes del vector de parámetros θ.

Con lo anterior establecido, los pasos del algoritmo EM son los siguientes:

Paso Inicial: Sea $\theta^{(0)}$ un vector de parámetros inicial, que está compuesto por la media y la desviación estándar ya que, bajo nuestro supuesto, la distribución es normal. Sobre la base de esta estimación, posiblemente sesgada, el algoritmo puede comenzar.

Paso E: Dados los valores de los parámetros de la iteración t, denotados por el vector $\theta^{(t)}$, el paso E calcula la función objetivo, que en el caso del problema de datos faltantes es igual al valor esperado de la log-verosimilitud de los datos observados, con lo que se obtiene que

$$Q(\theta|\theta^{(t)}) = \int l(\theta, X) f(X^m|X^0, \theta^{(t)}) dX^m = E\left[l(\theta, X)|X^0, \theta^{(t)}\right]. \tag{2.2.1}$$

Ahora, se sustituye el valor esperado de X^m, dado X^0 y $\theta^{(t)}$. En algunos casos, esta sustitución se realiza en cada uno de los datos faltantes, pero es suficiente sustituir solo la función de X^m que aparece en la log-verosimilitud de los datos completos.

Paso M: El paso M determina $\theta^{(t+1)}$, el vector de parámetros que maximiza la log-verosimilitud de los datos imputados. Formalmente, $\theta^{(t+1)}$ satisface

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \text{ para todo } \theta. \tag{2.2.2}$$

Se puede demostrar que la verosimilitud de los datos observados aumenta en cada paso. Debido a que la log-verosimilitud está acotada superiormente, el método converge.

En síntesis, el proceso EM genera el vector de parámetros estimado $\hat{\theta}$, con el que se genera un valor aleatorio que sustituye a uno vacío y el proceso se repite hasta completar todos los valores faltantes.

La convergencia de este algoritmo se garantiza mediante el Teorema 2.2.3, antes de formularlo considérese las siguientes definiciones.

Definición 2.2.1. *Una familia de funciones de densidades de una variable aleatoria y, denotadas como $f(y \mid \theta)$, tal que $(\theta_1, \ldots, \theta_n) = \theta$ pertenece al espacio de parámetros denotado como Θ, pertenece a la familia exponencial si*

$$f(y \mid \theta) = h(y)g(\theta)e^{\sum_{i=1}^{n} w_i(\theta)T_i(y)}. \tag{2.2.3}$$

Donde $g(\theta) \geq 0$ y $w_i(\theta)$ son funciones que no dependen de y, $h(y) \geq 0$ y además, $T_i(y)$ no depende de θ. Esta forma corresponde a una familia exponencial con n-parámetros, donde n es el mínimo entero tal que 2.2.3 se cumple.

Definición 2.2.2. *Una familia exponencial de n-parámetros es regular si se cumplen las siguientes condiciones:*

1. $\Theta = \{\theta \in \Theta \mid \dfrac{1}{g(\theta)} = \displaystyle\int_{-\infty}^{\infty} h(y)e^{\sum_{i=1}^{n} w_i(\theta)T_i(y)}dy < \infty\}$.
2. Θ *es un conjunto abierto.*
3. *No existe dependencia lineal para el conjunto de $T_i(y)$ y $w_i(\theta)$ con $i = 1, \ldots, n$.*

La familia exponencial es ampliamente estudiada, pues este tipo de distribuciones tiene muchos beneficios al usarlos. Por ejemplo, en [8] se menciona la ventaja significativa cuando se trata de calcular cierto tipo de estadístico. Muchas distribuciones conocidas pertenecen a la familia exponencial regular (véase [12]), como la distribución normal, de Poisson, exponencial y Bernoulli pertenecen a la familia exponencial regular. El teorema 2.2.3 garantiza la convergencia del método EM bajo ciertas condiciones generales.

Teorema 2.2.3. *Si la distribución de los datos completos pertenece a la familia exponencial regular y la sucesión $l_y(\theta)$ generada por el calculo de la verosimilitud de los datos en cada iteración del algoritmo EM es acotada superiormente entonces $\theta^{(t)}$ converge a un valor estacionario $\theta^{(*)}$*

Bajo las premisas del teorema, basta con determinar la distribución de las variables a imputar, sin embargo, se observa que son de tipo normal, binomial, Poisson, algunas de ellas son mostradas en las Figuras 2.3, 2.4 y 2.5.

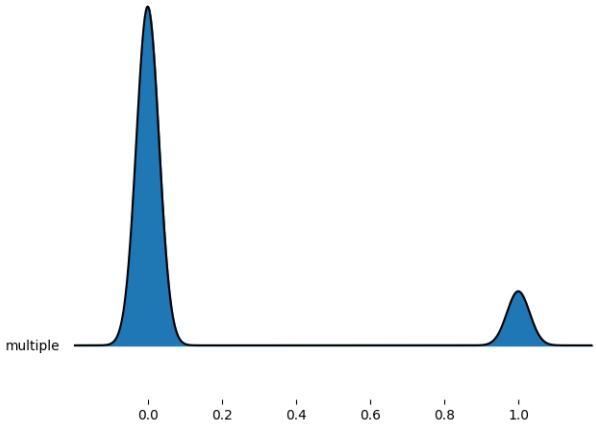

Figura 2.3: Distribución de frecuencias de la variable *múltiple*.

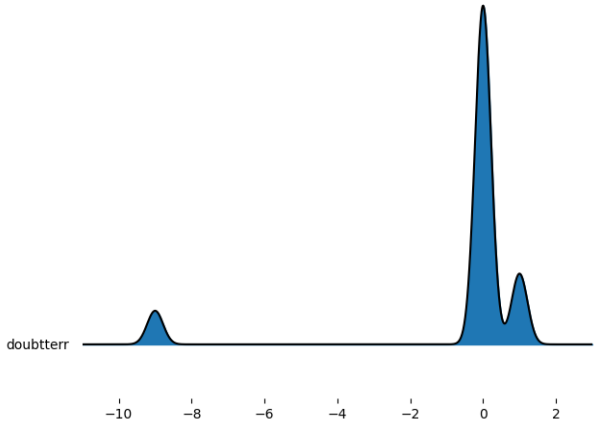

Figura 2.4: Distribución de frecuencias de la variable *doubtterr*.

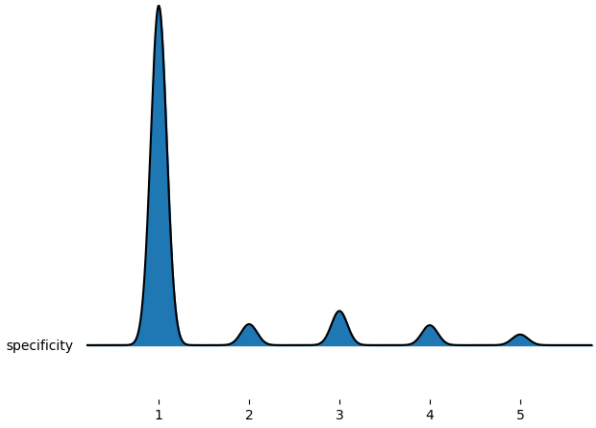

Figura 2.5: Distribución de frecuencias de la variable *specificity*.

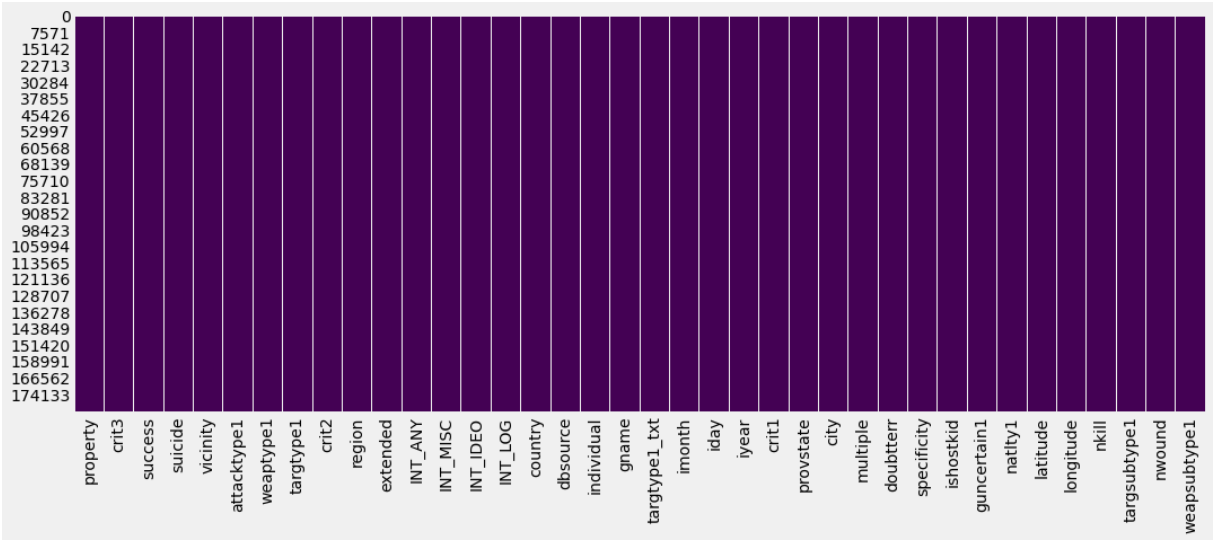

Figura 2.6: Resultado de la imputación a la base. No se observan entradas vacías de color amarillo.

De forma numérica, se calcula la función de verosimilitud para cada variable a imputar y se observa que para la variable *multiple*, la sucesión generada es no decreciente y acotada por el valor $-11{,}4$. Lo mismo ocurre para el resto de variables, por lo que la convergencia se garantiza para las variables con entradas faltantes.

Regresando al proceso EM, las variables con entradas vacías son imputadas y el resultado se muestra gráficamente en la Figura 2.6.

Ahora, la base se encuentra lista para ser sometida a un proceso de extracción de características, que será abordado en el siguiente capítulo. Obsérvese que la base de datos ha tenido un cambio en el número de columnas d pues pasó de tener 135 columnas a 38 columnas. Como la base de datos $\overline{D}$ cumple con la Definición 1.2.1, podemos verla de la siguiente manera:

$$\overline{D} = [\mathbf{X}; Y] \quad = [X_1, X_2, \ldots, X_j, \ldots, X_{d'}; Y], \tag{2.2.4}$$

donde $\mathbf{X} = X_1, X_2, \ldots, X_j, \ldots, X_{d'}$ es el conjunto de variables (columnas de la base) sin tomar en cuenta la variable objetivo *gname*, donde cada una de las variables $X_j = [x_1^j, \ldots, x_i^j, \ldots, x_n^j]^T$, para toda $j = 1, \ldots, d'$, con $d' = 37$ y la variable Y representa la

variable objetivo *gname*. Por lo que la base de datos esta conformada por $n = 181,691$ columnas y $d = 38$ columnas. Esta representación de la base será de ayuda durante el proceso de extracción de características.

Capítulo 3

Extracción de características

La siguiente etapa de clasificación es la extracción de características significativas, pues los datos de la base $\overline{D}$ pueden contener características que aporten una mayor cantidad de información para predecir las etiquetas de clase y otras que no lo hagan. Por ejemplo, la estatura de una persona es menos relevante para predecir algún tipo de raza, en comparación a su lengua, color de piel, entre otras. En este capítulo se mostrará a detalle este proceso.

3.1. Codificación de variables

Antes de realizar la extracción de características, para tener un mejor manejo y entendimiento de las variables o características, se realizó el renombramiento mostrado en el Cuadro 3.1.1.

Luego, para determinar las variables relevantes, se debe codificar a las variables categóricas para posteriormente procesarlas (véase [1, Sección 2.2.2.2, pág. 31]). En este caso, se transformarán a variables numéricas, pero se desea que se preserve la respectiva categoría. Por lo cual, realizaremos una codificación entera ordinal. Este método asigna un valor entero a cada clase dentro de una misma variable, por ejemplo, dentro de la variable *country_txt* se encuentran los países: Dominican Republic, México, y Philipi-

Nombre original	Renombramiento	Nombre original	Renombramiento
property	Propiedad	imonth	Mes
crit3	Criterio3	iday	Dia
success	Exito_de_ataque	iyear	Año
suicide	Acto_suicida	crit1	Criterio1
vicinity	vecindad	provstate	Provincia
attacktype1	Tipo_de_ataque	city	Ciudad
weaptype1	Tipo_de_arma	multiple	Incidente_multiple
targtype1	Tipo_de_objetivo	doubtterr	Duda
crit2	Criterio2	specificity	Especificidad de geocodificación
region	Region	ishostkid	Rehenes
extended	Evento_extendido	guncertain1	Certeza
INT_ANY	Internacional	natlty1	Nacionalidad de_victima
INT_MISC	Internacional_misc	latitude	Latitud
INT_IDEO	Internacional_ideologica	longitude	Longitud
INT_LOG	Internacional_logistica	nkill	Número_de_bajas
country	Pais	targsubtype1	Subtipo_de_objetivo
dbsource	Fuente	nwound	Heridos
individual	Individual	weapsubtype1	Subtipo_de_arma
gname	Grupo_terrorista	targtype1_txt	Tipo_de_objetivo

Cuadro 3.1.1: Ajuste en el nombre de cada variable.

nes; al realizar la codificación ahora se denominan: 58, 130 y 160, respectivamente. Este procedimiento se le realiza a las variables *Provincia*, *Ciudad*, *Fuente* y *tipo de objetivo*.

A continuación se expone de manera formal la metodología para codificar las variables, mediante algunos tipos de codificación existentes, y se justificará la elección del tipo de

codificación para aplicar a la base de datos GTD, para así posteriormente aplicar un método de selección de características adecuado.

Existen varios métodos de codificación entera que se pueden utilizar, dependiendo de las características de los datos y del contexto del análisis. Algunos de los métodos más comunes incluyen:

1. **Codificación ordinal:** En este enfoque, se asignan valores enteros a las categorías de una variable, según su orden o jerarquía. Por ejemplo, en una variable "Tamaño de vivienda", con categorías: "Pequeña", "Mediana" y "Grande", se pueden asignar los valores 1, 2 y 3, respectivamente. Esta codificación captura la relación de orden entre las categorías, pero no necesariamente refleja las diferencias de magnitud.

2. **Codificación One-Hot:** En este método, se crea una variable binaria (0 o 1) para cada categoría presente en la variable original. Si una observación pertenece a una categoría en particular, la variable correspondiente se establece en 1, mientras que las demás variables se establecen en 0. Esta codificación es útil cuando no existe una relación de orden entre las categorías y permite capturar la presencia o ausencia de una categoría en particular.

3. **Codificación basada en frecuencias:** En este enfoque, se asigna a cada categoría un valor entero según la frecuencia con la que aparece en el conjunto de datos. Las categorías más frecuentes pueden recibir valores más altos, lo que puede reflejar su importancia relativa en el análisis. Esta codificación es útil cuando se desea tener en cuenta la distribución de las categorías en los datos.

El tratamiento de las variables categóricas es fundamental y depende en gran medida de la naturaleza de estas variables. Las variables categóricas se clasifican en dos tipos principales: nominales y ordinales.

- Las **variables categóricas nominales** representan categorías sin un orden o jerarquía específica. Ejemplos de variables nominales podrían ser el color de los ojos

o la marca de un automóvil. En este caso, la codificación One-Hot es una opción común.

- Las **variables categóricas ordinales**, por otro lado, tienen una relación de orden o jerarquía entre las categorías. Por ejemplo, en una variable "Nivel educativo", con categorías "Primaria", "Secundaria", y "Universitaria", existe un orden específico. En este caso, la codificación ordinal puede ser apropiada, asignando valores enteros que reflejen el orden de las categorías.

En el caso de la base de datos $\overline{D}$, las variables categóricas que contiene son de tipo nominales, por lo que es recomendable aplicar la codificación One-Hot, ya que esta es recomendada cuando no se tiene un orden jerárquico en la variables. Sin embargo, al momento de realizar la codificación One-Hot sobre las variables categóricas de la base $\overline{D}$, el número de columnas de la base se incrementa significativamente, ya que de 38 columnas se incrementa a $2,936$, lo cual aumenta los tiempos de procesamiento para la ejecución del algoritmo KNN. Por tal motivo se optó por utilizar la codificación ordinal.

El tratamiento para realizar la codificación ordinal de las variables categóricas depende en esencia de la naturaleza de estas, se pueden distinguir dos casos:

- **Variables categóricas ordinales:** La codificación implica traducir el orden inherente de las categorías en una enumeración. Se asigna un número natural a cada categoría, respetando el orden implícito.

- **Variables categóricas nominales:** En este caso, la codificación implica una enumeración similar a la ordinal, aunque no hay un orden específico que se pueda seguir. La asignación de números a las categorías se realiza de manera prácticamente aleatoria debido a la falta de una pauta clara.

A continuación, se establecerá de manera formal el procedimiento matemático para realizar la codificación. Se tomará a la base de datos $\overline{D}$ con la siguiente estructura:

$$\overline{D} = [\mathbf{X}; Y] \quad = [X_1, X_2, \ldots, X_j, \ldots, X_{d'}; Y], \tag{3.1.1}$$

donde $\mathbf{X} = X_1, X_2, \ldots, X_j, \ldots, X_{d'}$ es el conjunto de variables de los datos (columnas de la base), donde cada una de las variables $X_j = [x_1^j, \ldots, x_i^j, \ldots, x_n^j]^T$, para toda $j = 1, \ldots, d'$, toma n valores, es decir, n es el número de observaciones del conjunto de datos $\overline{D}$, en esta investigación es de $181,691$ y el número de columnas d' es igual a 37.

La codificación se realiza a cada variable categórica X_j, formada por k_j categorías, se denotará por $[x_i^j] \in \{1, 2, \ldots, k_j\}$ a la categoría de la variable X_j a la que pertenece la observación x_i^j. De esta manera, se sustituye

$$x_i^j \to [x_i^j],$$

para todo $i = 1, \ldots, n$ y toda variable j-ésima.

Formalmente, la codificación se define de la siguiente manera: Por otro lado, $Y = [y_1, y_2, \ldots, y_n]^T$ contiene las etiquetas de cada atentado terrorista, ya que sus elementos son el nombre del grupo terrorista respectivo. La codificación realizada a Y es de tipo ordinal, es decir, se tiene que $y_i \to [y_i] \in \{0, 1, \ldots, 3537\}$, para todo $i = 1, \ldots, n$, ya que 3537 es el número de clases (grupos terroristas).

Definición 3.1.1 (Codificador de variable categórica)**.** *Un codificador es una aplicación donde*

$$\begin{aligned} \varphi : X_j &\to \mathbb{R}^p, \\ x_i^j &\mapsto \varphi(x_i^j), \ \textit{para todo } i = 1, \ldots, n. \end{aligned}$$

Observemos que la imagen de la aplicación φ no necesariamente tiene que estar en una dimensión, lo que representa que al aplicarle un método de codificación a una columna X_j puede generarse más de una variable numérica nueva.

Una vez establecido todo el marco teórico anterior se puede comenzar a definir y detallar el método de codificación ordinal para tratar las variables categóricas del conjunto de datos $\overline{D}$. Este método surge de forma natural y consiste en reemplazar cada una de las k_j categorías de X_j por un número entero. De esta forma la nueva variable se encontrará formada por números enteros pertenecientes al intervalo $[1, k]$. Para realizar esta codificación nos auxiliaremos de la Definición 3.1.2.

Definición 3.1.2 (Ordinal Encoder)**.** *El Ordinal Encoder es una aplicación, donde*

$$\varphi_{OE} : X_j \to \mathbb{N},$$
$$x_i^j \mapsto \varphi_{OE}(x_i^j) \in \{1, \ldots, k_j\}, \text{ para todo } i = 1, \ldots, n.$$

Cabe resaltar que este tipo de codificación aumenta la eficiencia de los métodos de minería sin generar un aumento de la dimensión. En esta investigación, la codificación realizada resultó ser la ideal dada la naturaleza del problema ya que permite reducir significativamente el número de características que se realizará en la sección siguiente.

3.2. Modelo SelectKBest

Las características irrelevantes generalmente dañan la precisión de los modelos de aprendizaje máquina (véase [1], [16] o [9, Sección 4.2, pag. 34]); además de afectar la eficiencia computacional. Por esta razón, el objetivo de los algoritmos de selección de características es elegir aquellas más informativas con respecto a la etiqueta de clase (en el presente caso es el grupo terrorista), con lo que se logra, además, una reducción de dimensionalidad. En esta investigación se utilizará un método denominado *Select K Best Features* (selección de las K mejores características) y abreviado como *SelectKBest*. Este es un modelo de filtrado que elige las K características con varianza más alta y elimina el resto. El proceso del algoritmo está basado en la selección de características planteado como sigue (véase [11, Capítulo 4, definición 4.1, pág. 64]).

Definición 3.2.1. *Para un conjunto dado $\mathcal{D}^{(1:d')}$ de d variables independientes e idénticamente distribuidas, un método de selección es un mapeo*

$$\Phi(\mathcal{D}^{(1:d')}) : \mathcal{Z}^{d'} \mapsto 2^{V_{d'}},$$

donde $\mathcal{Z}^{d'} = (X1, X_2, \ldots, X_{d'}, Y)$ es el conjunto de los d' vectores de características y la variable objetivo y $2^{V_{d'}} = \{S : S \subseteq V_{d'}\}$ es el conjunto potencia de $V_{d'} = \{1, 2, \ldots, d'\}$.

Una de las tareas de la selección de variables es determinar el *conjunto objetivo*, esto es, que el método sea consistente en el siguiente sentido.

Definición 3.2.2. *Un algoritmo de selección* $\Phi(\mathcal{D}^{(1:d')})$ *es consistente con respecto a un conjunto de características si*

$$\Phi(\mathcal{D}^{(1:d')}) \to S.$$

Esto es, converge al conjunto S*. El conjunto* S *es llamado el conjunto objetivo de* Φ*.*

El objetivo es garantizar que el conjunto S esté compuesto solo de las características más relevantes. En [11, pág. 63] se define formalmente a estas variables como se muestra a continuación.

Definición 3.2.3 (Relevancia)**.** *Una característica* X_i *es relevante a* Y*, si y solo si,*

$$\exists\ S \subseteq \{1, \ldots, d'\} : Y \not\perp X_i | X_S,$$

con $X_S = \{X_i | i \in S\}$*.*

La notación $\not\perp$ *significa que* $P(Y) \leq P(X_i|X_S)$*.*

Una característica X_i *es irrelevante a* Y*, si y solo si, no es relevante a* Y*.*

Para la presente, consideremos el caso univariado, en otras palabras, se desea determinar el conjunto S^A el cual es marginalmente dependiente de Y, cuyas componentes son las características X_i que satisfacen $Y \not\perp X_i|\emptyset$, o, equivalentemente, $P(Y|X_i) \neq P(Y)$.

El método *SelectKBest* está basado en las pruebas estadísticas de hipótesis, estas pruebas miden la dependencia de características individuales X_i con la variable objetivo *gname*, y por lo tanto se enfocan solo en la distribución marginal $f(x_i, y)$ (véase [11, Sección 4.1.1, pág. 71]). Esta prueba identificará a las características con dependencia significativa. Por lo tanto, la prueba de hipótesis está basada en los siguientes planteamientos:

- La hipótesis nula H_0^i afirma que la característica X_i es irrelevante a (o independiente de) Y.
- La hipótesis alternativa H_1^i afirma que X_i es relevante a (o dependiente de) Y.

De esta manera, cada característica cumplirá sólo una de las dos hipótesis. Para garantizar alguna de las dos opciones se realizará una prueba estadística, que puede definirse

como la función $\phi_i : \mathcal{Z}^{d'} \mapsto \{1, 0\}$, para cada observación X_i, donde se decide que H_0^i es cierta siempre y cuando $\phi_i(X_i, Y) = 0$; y H_1^i se elige cuando $\phi_i(X_i, Y) = 1$. Claro que esta decisión puede o no ser correcta debido a los dos tipos de errores que se pueden cometer (véase [11, pág. 72]), estos pueden medirse por dos tasas de error:

Tasa de error tipo I: $P(\phi_i = 1|H_0^i)$ mide la frecuencia con la que la característica X_i se clasifica como relevante cuando en realidad no lo es.

Tasa de error tipo II: $P(\phi_i = 0|H_1^i)$ es la medición en el caso opuesto al anterior.

La tasa del error tipo I es también llamada de falso positivo y la de tipo II es la de falso negativo.

Ahora, consideremos al parámetro α denominado como el nivel de la prueba. Este valor delimitará la probabilidad de que falsamente la hipótesis nula sea rechazada, esto es:

$$P(\phi_i = 1|H_0^i) \leq \alpha.$$

Cuando lo anterior se satisface, se dice que ϕ_i tiene una prueba de nivel (o de medida) α.

De un conjunto de pruebas de hipótesis ϕ_i, $i = 1, \ldots, d'$, en [11, pág. 72] se redefine el método de selección de características como

$$\Phi(D^{(1:d')}) = \{i : \phi_i = 1\}.$$

De esta manera, Φ depende solo de las distribuciones marginales $f(x_i, y)$.

Por otro lado, se tiene la prueba *z de Fisher*, que se basa en el cálculo del coeficiente de correlación de Pearson. Esta prueba calcula el estadístico

$$t = \frac{1}{2} \ln \frac{1 + r(X_i, Y)}{1 - r(X_i, Y)},$$

donde

$$r(X, Y) = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2}\sqrt{\sum_{j=1}^{n}(y_j - \overline{y})^2}},$$

el cual es asintóticamente Gaussiano, es decir, su distribución es

$$f(t) = N(\tanh^{-1} r(X_i, Y), (d' - 3)^{-1}),$$

para cualquier $f(x_i, y)$, esto debido al Teorema del Límite Central (véase [6]). Por lo tanto, la prueba es asintóticamente correcta.

Para alta dimensionalidad, como sucede en el presente caso, surge el siguiente *problema de múltiple testeo*. Supongamos que el método de selección de características realiza n pruebas $\phi_1, \ldots, \phi_n$, cada uno con un nivel α. En consecuencia, se tiene que el número esperado de errores es

$$\mathbb{E}[\Phi(D^{(1:d')}) \cap \{i : H_0^i\}] = \mathbb{E}\left[\sum_{i:H_0^i} \phi_i\right].$$

Si se supone en el peor de los casos que todos los H_0^i son ciertos, el valor esperado es igual a $n\alpha$, el cual puede ser relativamente grande, por ejemplo, si $n = 10^4$ y $\alpha = 0{,}05$, se tiene un error de 500, lo mismo se tiene aún si se considera α muy pequeño.

Para compensar el problema de multiplicidad, se controla la tasa de falsos positivos de cada una de las pruebas mediante el concepto de valor p.

Definición 3.2.4 (Valor p.)**.** *Un valor p para una hipótesis nula dada H_0 es una variable aleatoria $p \in [0, 1]$ que satisface*

$$P(p \leq \alpha | H_0) \leq \alpha,$$

para toda $\alpha \in [0, 1]$.

El valor p puede interpretarse como el más bajo nivel para el cual la hipótesis nula podría ser rechazada (véase [11, pág. 77]). Por lo tanto, el valor p mide la confianza en el rechazo. De hecho, se demuestra que para n pruebas independientes de nivel α, se tiene que

$$P\left(\exists i : p_i \leq \alpha | H_0^{(1:n)}\right) \leq \alpha,$$

con p_i los valores p de cada variable X_i. Lo que nos lleva a que las k variables más relevantes con la variable Y son aquellas con los valores p_i más pequeños. Por ejemplo, en la Figura 3.1 se muestran puntajes de las variables 15 y 14 a comparación con el resto, estas se representan por las líneas más altas y corresponden a *Internacional logística* e

Internacional ideológica. Si se eliminan las líneas de estas variables, se obtiene la Figura 3.2, en la cual se observa que la siguiente característica relevante es la 12, que corresponde a *Internacional*, seguida de la 16, *País*, luego la número 10, *Región*, la 32, *Nacionalidad de víctima*, luego la 22: *Año*, 18: *Individual*, 31: *Certeza*, finalmente, *Latitud* es la última característica. Obsérvese la Figura 3.3, donde se muestra el orden por relevancia.

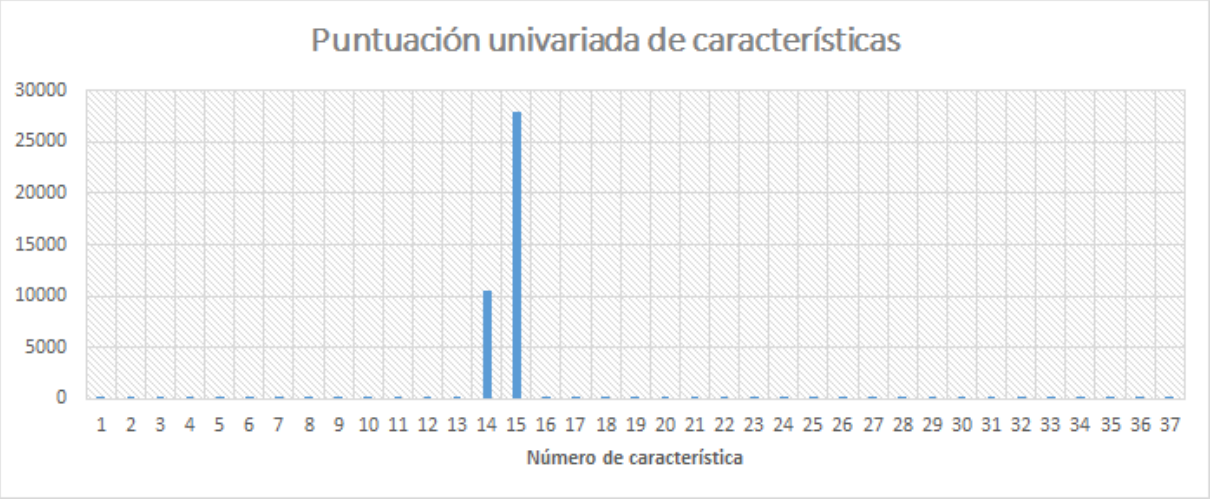

Figura 3.1: Diagrama de puntuación para cada variable, que muestra que las variables 14 y 15 son relevantes.

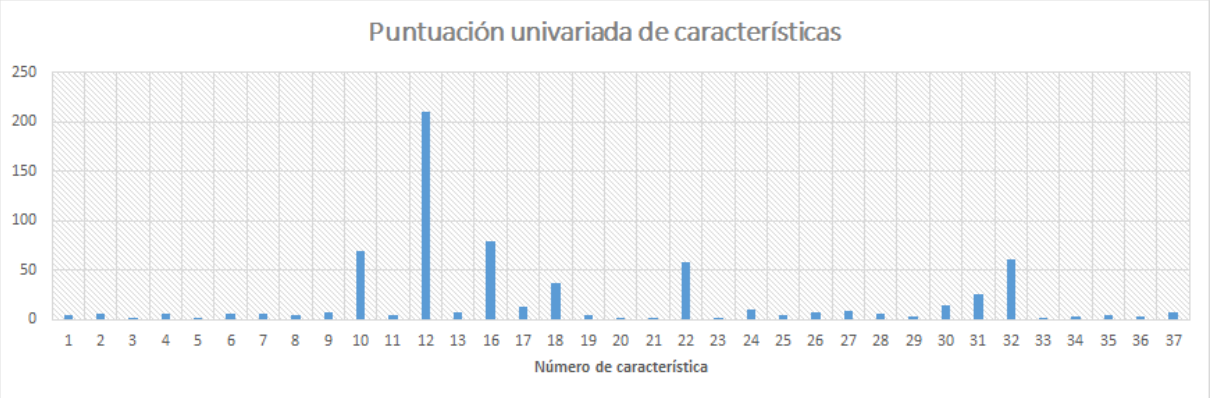

Figura 3.2: Diagrama de puntuación para cada variable, sin las variables 14 y 15.

Específicamente, las características de mayor relevancia con la variable objetivo, son:

1. *Internacional logística*: Esta variable se basa en una comparación entre la naciona-

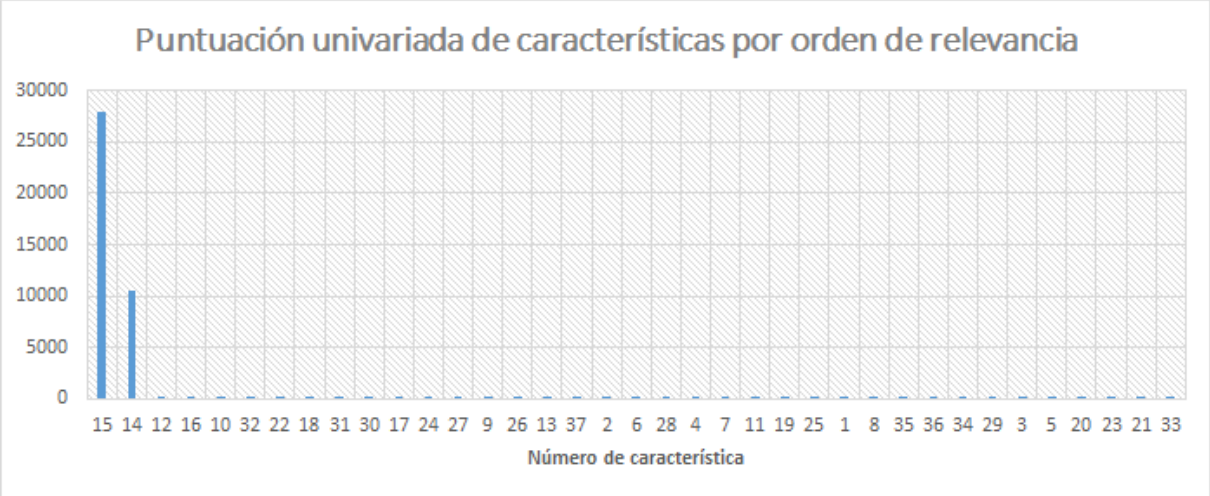

Figura 3.3: Diagrama de puntuación para cada variable por orden de relevancia.

lidad del grupo perpetrador y el lugar del ataque. Indica si un grupo perpetrador cruzó una frontera para realizar el ataque.

2. *Internacional ideológica*: Esta variable se basa en una comparación entre la nacionalidad del grupo perpetrador y la nacionalidad del objetivo(s)/víctima(s). Indica si un grupo perpetrador atacó a un objetivo de una nacionalidad diferente.

3. *Internacional*: Variable de tipo categórica que indica si el ataque fue de tipo internacional, doméstico o desconocido, de acuerdo a la definiciones tomadas por los creadores de la base.

4. *País*: Este campo identifica el país o lugar donde ocurrió el incidente.

5. *Región*: Este campo identifica la región en la que ocurrió el incidente. Las regiones se dividen en 12 categorías.

6. *Nacionalidad de víctima*: Esta es la nacionalidad del objetivo que fue atacado y no es necesariamente la misma que la del país en el que ocurrió el incidente, aunque en la mayoría de los casos lo es. Para incidentes de secuestro aéreo se registra la nacionalidad del avión y no la de los pasajeros.

7. *Año*: Este campo contiene el año en que ocurrió el incidente. En el caso de que ocurra(n) incidente(s) durante un período prolongado, el campo registrará el año en que se inició el incidente.

8. *Individual*: Esta variable indica si el ataque fue realizado o no por un sólo individuo.

9. *Certeza*: Esta variable indica si la información reportada por las fuentes sobre los nombres del grupo perpetrador se basan en especulaciones o afirmaciones dudosas.

10. *Latitud*: Este campo registra la latitud (basada en los estándares WGS1984) de la ciudad en la que ocurrió el evento.

Una vez realizado todo el proceso anterior, se obtienen las variables principales, las cuales son las que aportan una mayor información en el fenómeno y que tienen un mayor efecto en la variable objetivo. Además, en el capítulo anterior las variables antes mencionadas fueron imputadas, lo que ahora permite entrenar al modelo de aprendizaje máquina con la base limpia y completa, proceso que será descrito en el capítulo siguiente. Obsérvese que la base de datos $\overline{D}$ ha pasado por un proceso de exploración y limpieza, lo que ha permitido reducir el número de columnas, pues de tener 38 columnas, pasa a tener solamente 10, las cuales fueron mencionadas anteriormente. Ahora la dimensiones para la base de datos $\overline{D}$ son $n = 181,691$ y $d'' = 10$.

Capítulo 4

Modelo de aprendizaje máquina

En este punto, la base de datos $\overline{D}$ ha sido sometida a un proceso de limpieza, imputación y de extracción de características significativas. Este último con la finalidad de obtener las variables que aportan una mayor cantidad de información a la variable objetivo *gname*, la cual fue renombrada como *Grupo_ terrorista*.

Ahora, el objetivo es aplicar un modelo matemático de aprendizaje máquina (ML por sus siglas en inglés *Machine Learning*) para poder brindar una clasificación de los datos y, con base en esta, asignar una etiqueta de clase cuando un nuevo registro sea introducido. Este modelo será planteado en este capítulo y entrenado con la base de datos $\overline{D}$.

Para que el modelo logre predecir si un nuevo registro pertenece o no a algún grupo terrorista, se entrenará para detectar el comportamiento de cada atacante, posteriormente, clasifique la nueva entrada. Formalmente, este problema de clasificación se define como:

Definición 4.0.1 (Clasificación de datos)**.**

Dada una $n' \times d'$ matriz de datos de entrenamiento $\overline{D'}$, donde las entradas de $\overline{D'}$ son elementos de la matriz $\overline{D}$ y, por lo tanto, $n' < n$, y una etiqueta de clase que pertenece al conjunto $\{1, 2, \ldots, j\}$, asociado con cada una de las n filas en $\overline{D'}$ (registros en $\overline{D'}$). Se define la clasificación de datos como la creación de un modelo de entrenamiento M que se puede usar para predecir la etiqueta de clase de un registro d-dimensional $\overline{Y} \notin \overline{D'}$.

Es posible realizar la asignación de la etiqueta mediante dos formas, una es por agrupamiento y la segunda es por clasificación. En el caso del problema de agrupamiento, los datos se dividen en k grupos en función de la similitud, mientras que en el problema de clasificación, un registro de prueba también se clasifica en uno de los k grupos, pero esto se consigue en función de un modelo M aplicado a nuestra base de entrenamiento $\overline{D'}$, como se menciona en la Definición 4.0.1, sin tomar en cuenta la similitud de estos. En la presente investigación, el problema se abordará como clasificación y el modelo se planteará en la siguiente sección.

4.1. Modelo KNN

Algunos métodos de clasificación están basados en el principio de la coincidencia de un nuevo registro con el más cercano de los datos de entrenamiento. Un caso particular es el algoritmo KNN (por las siglas en inglés de *K-nearest neighbor*), el cual es un algoritmo que ha demostrado ser exitoso en la resolución de problemas de este tipo y que fue introducido por Evelyn Fix y Joseph Hodges en [7]. Este clasificador utiliza una métrica con cada X_i en $\overline{D}$ para construir las vecindades y determinar la clase a la que pertenece un nuevo registro, esto es, un objeto se clasifica por una mayoría de "votos" de sus vecinos y el objeto se asigna a la clase más común entre sus k vecinos más cercanos, donde k es un número entero positivo, regularmente pequeño. En el caso donde $k = 1$, se tiene que el nuevo registro simplemente tomará la etiqueta de clase de su único vecino más cercano. La idea general inicia considerando un punto formado por cada una de las características significativas de la base de entrenamiento, para que el algoritmo con estas entradas le asigne una etiqueta de clase, tomando en cuenta las etiquetas de su(s) vecino(s) más cercanos bajo una métrica definida y un correcto espacio de características. Se sabe que estas técnicas son sólidas en el caso de grandes conjuntos de datos y de dimensiones bajas, como hasta el momento se ha comportado la problemática.

Como se mencionó anteriormente, el problema será tratado como de clasificación y el objetivo será predecir etiquetas de clase discretas para nuevos registros cuya etiqueta de

clase se desconoce. Sean $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ el conjunto de observaciones de registros q-dimensionales $\overline{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^q$, y un correspondiente conjunto de etiquetas $Y = \{y_i\}_{i=1}^n \subset \mathbb{R}$, donde $q = 10$. El objetivo de la clasificación es proporcionar un modelo funcional y entrenado f que genere una predicción razonable de la etiqueta de clase $Y' \in \mathbb{R}$ para un patrón desconocido $X' \in \mathbb{R}^q$. Los registros sin etiquetas deben asignarse a vecindades de elementos similares previamente clasificados, en otras palabras, que bajo cierta métrica están cerca de él, que provienen de la misma distribución, o se encuentran en el mismo lado de una función de separación. Pero aprender de los registros ya observados presenta algunas dificultades, pues los conjuntos de entrenamiento pueden ser ruidosos, las características importantes pueden ser desconocidas, la similitud entre registros puede no ser fácil de definir, y las observaciones pueden no ser suficientemente descrito por distribuciones simples.

Con lo mencionado anteriormente, se debe tener definida una similitud en el espacio de datos $\mathbb{R}^q$, para este método se utilizará la métrica de Minkowski (también conocida como p-norma), definida de la siguiente manera:

$$\|X' - X_1\|^p = \left(\sum_{i=1}^{q} |(x_i') - (x_1)_i|^p \right)^{1/p}. \tag{4.1.1}$$

La elección de la métrica juega un papel fundamental en el rendimiento del algoritmo k-Nearest Neighbors (KNN). Para la presente investigación se elige la métrica de Minkowski con $p = 2$. La elección de esta métrica se fundamenta en su simplicidad y eficacia para problemas de clasificación y regresión. Es crucial tener en cuenta la naturaleza de los datos y ajustar la métrica según las características específicas del problema. En este caso, la métrica Minkowski con $p = 2$ fue adecuada dada la naturaleza de las variables en nuestro conjunto de datos.

La elección de k define la localidad de KNN. Se puede observar en la Figura 4.1 la diferencia de la clasificación KNN para los valores $k = 1$ y $k = 20$ aplicado a un conjunto de datos bidimensional formado por dos nubes de datos superpuestas de 50 puntos rojos y azules. La región del espacio de datos que se clasificarían como azules se muestran en

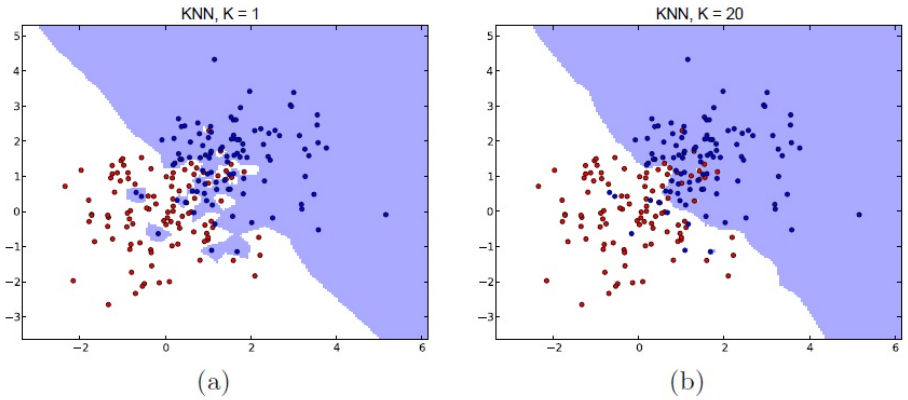

Figura 4.1: Comparación de las regiones para dos medidas de vecindad: (a) $k = 1$ y (b) $k = 20$. Imagen obtenida de [9].

morado, mientras que las áreas clasificadas como rojas se muestran en blanco. Obsérvese que para $k = 1$ la clasificación es local, mientras que para $k = 20$, el clasificador generaliza ignorando pequeñas aglomeraciones de patrones.

La cuestión que se plantea es cómo elegir el valor de k, es decir, qué tamaño de vecindades consigue el mejor resultado de clasificación. Este problema también se conoce como selección de modelo, y pueden emplearse varias técnicas, como la validación cruzada, para elegir el mejor modelo y los mejores parámetros. Este proceso se describe en la Sección 4.4

4.2. Multi-class K-Nearest Neighbors

En la sección anterior se describe el modelo KNN para clasificación de dos nubes de puntos, esto es conocido como clasificación binaria, sin embargo, el modelo KNN también puede aplicarse a problemas de clasificación donde la variable objetivo tiene más de dos clases, lo que se conoce como *clasificación multiclase*, debido a que la variable *Grupo_terrorista* cuenta con más de dos grupos de atacantes, el problema abordado en esta investigación es de este tipo. Por lo que, formalmente, se plantea el modelo como en la siguiente definición.

Definición 4.2.1. *Sea* $\overline{\mathbf{X}}$ *un registro desconocido, se fija el valor de* $k \in \mathbb{N}$*, se calculan las distancias entre* $\overline{\mathbf{X}}$ *y cada* $X_i \in \overline{D}$*, se forma el conjunto de distancias*

$$Dist(\overline{\mathbf{X}}) = \left\{ \|\overline{\mathbf{X}} - \overline{X_i}\|^p : \ \overline{X_i} \in \overline{D} \right\}.$$

Se toman los k *valores más pequeños del conjunto de distancias y se define como* $\mathcal{N}_K(\overline{\mathbf{X}})$ *al conjunto de índices de los registros correspondientes a los* k *valores más pequeños de* $Dist(\overline{\mathbf{X}})$*. El algoritmo* KNN *para la clasificación multiclase se auxilia de la función*

$$f_{KNN} : \mathbb{R}^q \to \{1, 2, \ldots, j\}$$

definida como

$$f_{KNN}(\overline{\mathbf{X}}) = \arg\max_{y \in Y} \sum_{i \in N_k(\overline{\mathbf{X}})} I(y_i = y), \tag{4.2.1}$$

donde $I(\cdot)$ *es la función indicadora que devuelve uno si su argumento es verdadero y cero en caso contrario.*

Para entender mejor el algoritmo KNN multiclase se exhibirá un ejemplo de su aplicación de manera gráfica aplicada a un conjunto de datos bidimensionales, los puntos representan observaciones de ciertos objetos o fenómenos, y nuestro objetivo es determinar a qué clase pertenecen nuevos puntos desconocidos.

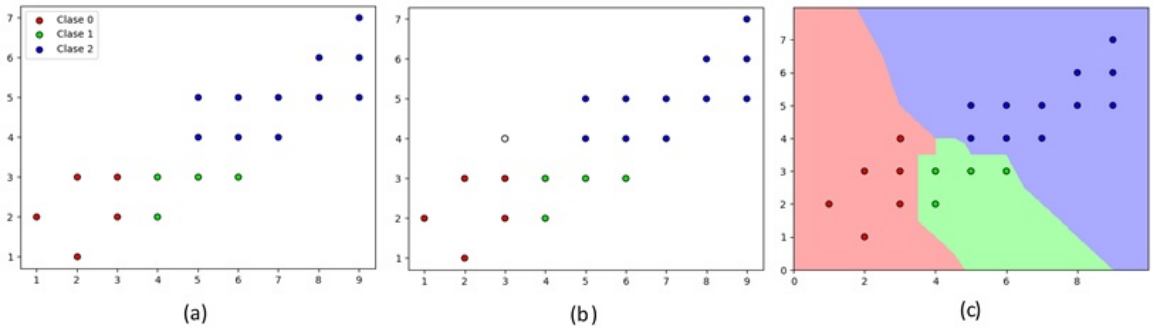

Figura 4.2: Proceso de clasificación de datos de manera gráfica.

Ejemplo 4.2.2. *En la Figura 4.2 se puede observar el proceso de clasificación de datos utilizando el algoritmo* KNN *mediante el uso de tres gráficas para un conjunto de puntos*

de entrenamiento previamente etiquetados en 3 diferentes clases. Los puntos de entrenamiento se muestran en la gráfica (a), donde cada punto está representado por su ubicación en el plano, se les ha asignado un color específico según la clase a la que pertenecen. Estas clases se basan en ciertas características o propiedades que se han medido o definido para cada objeto. En (b), se puede observar un nuevo punto que es $(3, 4)$ *que se presenta como un punto individual sin algún color específico de clase asignado y que no está presente en el conjunto de entrenamiento de la gráfica (a), este punto representa una nueva observación o dato que se desea clasificar en una de las clases definidas en (a). Finalmente, a través del análisis de las características y la ubicación del nuevo punto en relación con los puntos de entrenamiento el algoritmo* KNN *puede realizar una predicción que se basará en la clase mayoritaria de los* k *vecinos más cercanos al nuevo punto. La gráfica (c) muestra la representación gráfica del conjunto de datos junto con la clasificación realizada por el algoritmo* KNN *para el punto* $(3, 4)$*. Además, el algoritmo permite crear una malla de decisión que divide el espacio en regiones correspondientes a las diferentes clases, de esta manera, también se puede observar en (c) la representación gráfica de este mallado de decisión junto con los puntos de entrenamiento y el punto a clasificar, esto nos brinda una comprensión intuitiva del funcionamiento del algoritmo y su capacidad para clasificar nuevos puntos en función de sus vecinos más cercanos.*

4.3. Entrenamiento y validación

El modelo de aprendizaje KNN tiene la capacidad de aprender patrones o reglas a partir de un conjunto de datos de entrenamiento y luego utilizar este modelo para clasificar nuevas instancias o ejemplos no vistos.

Como se menciona al inicio del capítulo, trataremos nuestro problema como un problema de clasificación, para resolver este problema la base de datos $\overline{D}$ tiene que pasar por las siguientes dos fases:

Fase de Entrenamiento: Se seleccionará el modelo clasificador a utilizar, el cual será entrenado tomando como referencia el conjunto de entrenamiento denotado por $\overline{D}_{Train}$,

para que el modelo aprenda la estructura del conjunto de datos con base en las etiquetas correspondientes.

Fase de Prueba: En esta etapa, el modelo se empleará para asignar etiquetas de clase a cada instancia en el conjunto de prueba denotado por $\overline{D}_{Test}$, las cuales no están disponibles para el programa durante el entrenamiento. Se registrarán tanto el número de aciertos como el número de fallos, permitiendo el cálculo de métricas de desempeño del modelo, proporcionando una evaluación de su eficacia.

En esta sección se muestra el proceso de entrenamiento y validación del modelo de aprendizaje automático planteado mediante la función (4.2.1). Estos pasos nos permiten ajustar el modelo a los datos de entrenamiento y estimar su rendimiento en datos no vistos.

Debido a que la investigación se enfoca en el análisis de ataques terroristas y que la variable objetivo corresponde a los diferentes grupos terroristas, el algoritmo de clasificación KNN multiclase es el adecuado para resolver el problema que se presenta. Este modelo debe aprender a detectar las relaciones existentes entre los nuevos registros y sus respectivos grupos terroristas, utilizando las características intrínsecas de los registros conocidos. Para ello, utilizaremos un subconjunto de la base de datos $\overline{D}$ para el entrenamiento y evaluación del modelo, con el objetivo de predecir de manera precisa y acertada la pertenencia de nuevos registros a los distintos grupos terroristas.

El proceso que implica aprender las correspondencias entre las características de entrada y las etiquetas de clase se conoce como entrenamiento. Como se mencionó, los datos que son utilizados para este proceso se denominan datos de entrenamiento y los representaremos como $\overline{D}_{Train}$, este conjunto se forma del 70 % de los registros de la base $\overline{D}$ elegidos de forma aleatoria, donde cada registro contiene su respectiva etiqueta de clase (grupo terrorista). Durante el proceso de aprendizaje, el modelo aprende la estructura del conjunto de datos de entrenamiento $\overline{D}_{Train}$, posteriormente el modelo se considera entrenado.

Una vez que el modelo ha sido entrenado, se procede a validar su desempeño utilizando datos del conjunto de prueba. A este conjunto de prueba lo denominaremos $\overline{D}_{Test}$, que

corresponde al 30 % de registros restantes de la base $\overline{D}$. Es importante destacar que estos registros son sometidos al modelo sin etiquetas y que no los ha visto previamente. Cada registro en $\overline{D}_{Test}$ se denomina "registro de prueba". Durante la validación, se calcula la métrica de desempeño denominada *accuracy*, que se obtiene dividiendo el número de etiquetas correctas en $\overline{D}_{Test}$ entre el número total de registros. Una vez completada la validación, el modelo se utilizará para predecir las etiquetas de clase de nuevos registros que no cuentan con una etiqueta conocida.

Durante el proceso de entrenamiento el modelo usa la base de entrenamiento $\overline{D}_{Train}$, con el propósito de ajustar los parámetros de tal manera que la clasificación que se obtenga al someterlo a $\overline{D}_{Test}$ sea la óptima, es decir, que la etiqueta predicha coincida en su mayoría de veces con la etiqueta original.

Sin embargo, el algoritmo KNN tiene la particularidad de tener un *aprendizaje perezoso* (lazy learning) significa que el modelo no realiza un proceso explícito durante la fase de entrenamiento. En lugar de eso, aprende almacenando todos los datos de entrenamiento en la memoria, ya que el único parámetro con el que se cuenta es k, el cual representa el número de vecinos más cercanos a un registro desconocido que se tomaran en cuenta, para que con estos k registros se le asigne una etiqueta al registro desconocido. Una vez que seleccionamos el valor de k, este permanece siempre constante durante todo el proceso de entrenamiento y validación. La forma de elegir el valor de k se decide antes de aplicar el algoritmo KNN, y para elegir un valor adecuado para este, en la siguiente sección se abordará el proceso para la obtención de este valor.

Por lo tanto, el algoritmo no será sometido a un proceso explícito de entrenamiento, sino que comienza con el proceso de validación una vez que se ha seleccionado el valor de k con el que se va a ejecutar. Luego, es posible evaluar su funcionamiento mediante la métrica *accuracy*. El *accuracy* de entrenamiento refleja la capacidad del modelo para clasificar correctamente las etiquetas en el conjunto de datos con el que fue "entrenado", es decir, aquel conjunto que el modelo ya ha visto durante la fase de predicción. Por otro lado, el *accuracy* de los datos de prueba evalúa la habilidad del modelo para generalizar a datos no vistos previamente durante el entrenamiento.

Para realizar la fase de validación del modelo de clasificación KNN nos auxiliaremos de la función 4.2.1 y de los siguientes pasos para clasificar un registro $\overline{\mathbf{X}}$ con etiqueta de clase desconocida Y:

1. Se toma un registro $\overline{\mathbf{X}} \in \overline{D}_{Test}$.
2. Se calcula el conjunto $Dist(\overline{\mathbf{X}})$.
3. Se determinan los k valores más pequeños del conjunto obtenido en el punto anterior y se construye el conjunto $N_K(\overline{\mathbf{X}})$.
4. Se calcula $f_{KNN}(\overline{\mathbf{X}})$ y se asigna el vector a alguna de las clases según el resultado obtenido.

Al proceso anterior lo llamaremos algoritmo KNN multiclase. Una vez finalizado este proceso con cada uno de los puntos en $\overline{D}_{Test}$ se procede a trabajar con una métrica de validación con el objetivo de observar la efectividad de la clasificación que hizo el algoritmo KNN. Esta parte es fundamental para evaluar el rendimiento del modelo al igual que su capacidad para generalizar a datos no vistos.

El *accuracy* se define como la proporción de instancias correctamente clasificadas respecto al total de instancias, su ecuación es

$$Accuracy = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}.$$

La importancia del accuracy radica en su capacidad para proporcionar una medida global y fácilmente interpretable del éxito del modelo en la tarea de clasificación. Un alto accuracy indica que el modelo es efectivo en la mayoría de las predicciones, mientras que un bajo accuracy sugiere deficiencias en la capacidad de clasificación.

La elección de utilizar exclusivamente la métrica de accuracy en la evaluación del modelo K-Nearest Neighbors (KNN) se fundamenta en la complejidad inherente a la clasificación multiclase. Mientras métricas adicionales como la precisión, recall o la F1-score ofrecen una perspectiva detallada en contextos de clasificación binaria.

En esta investigación, se realizó un proceso para la obtención de un valor adecuado para k, este proceso se aborda en la sección siguiente. El valor de k que se encontró para ejecutar el algoritmo KNN es $k = 15$. En el Cuadro 4.3.1 se muestra el *accuracy* obtenido en la fase de entrenamiento (utilizando la base $\overline{D}_{Train}$) y validación (utilizando la base $\overline{D}_{Test}$), estos resultados indican un rendimiento satisfactorio del modelo en la clasificación multiclase.

Entrenamiento	**Validación**
0.9659	0.8582

Cuadro 4.3.1: *Accuracy* de entrenamiento y validación

4.4. Selección de modelo

Se observa que el accuracy obtenido en la fase de entrenamiento y validación del modelo KNN es bueno (Cuadro 4.3.1), la elección óptima del valor de k fue un componente esencial para lograr este resultado. Como se mencionó anteriormente es importante tener en cuenta que la precisión del algoritmo KNN puede verse afectada por varios factores, como el valor de k seleccionado y la distribución de los datos. En la Figura 4.3 se puede ver que un valor incorrecto de k puede resultar en un sobreajuste o subajuste del modelo. Además, si los datos no están bien separados o presentan solapamiento entre clases, el rendimiento del algoritmo puede disminuir. Por esto, se debe tener precaución al seleccionar el valor de k y considerar las características y distribución de los datos, a esta búsqueda del parámetro óptimo se le conoce como *selección de modelo*.

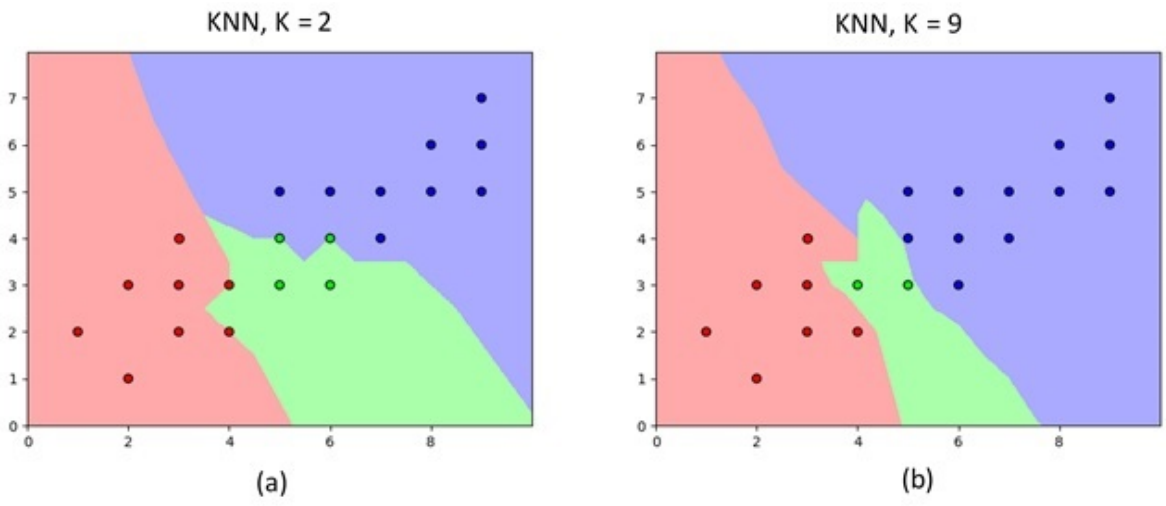

Figura 4.3: Representación gráfica para $k = 2$ y $k = 9$.

Considérese nuevamente el ejemplo 4.2.2, en él se observa que para $k = 3$ se obtiene una buena clasificación con respecto a los puntos de entrenamiento y la malla de decisión, ya que el color de etiqueta de cada punto de entrenamiento coincide con el color que se debe asignar con respecto a la malla de decisión. Por otro lado, en la gráfica (a) de la Figura 4.3 se puede observar que si tomamos un k más pequeño, como $k = 2$, se tiene una clasificación inexacta, debido a que el color de las etiquetas de algunos puntos de entrenamiento con respecto a la malla de decisión que ha creado el algoritmo KNN no coinciden con color de la etiqueta que tenía originalmente el punto de entrenamiento. De la misma manera, si elegimos un k más grande, como $k = 9$, en la gráfica (b) se puede observar el mismo comportamiento, notando que para este valor elegido de k el algoritmo generaliza la clasificación, tomando en cuenta las clases que tienen una mayor aparición en el conjunto de entrenamiento, en este caso los colores rojo y azul son los que predominan en la malla de decisión.

En la elección del valor de k para el algoritmo k-Nearest Neighbors, es esencial evaluar el rendimiento del modelo de manera robusta y evitar el sobreajuste o subajuste a un conjunto de datos específico. La validación cruzada es una técnica crucial para lograr esto, en particular, en esta investigación se utiliza el algoritmo denominado *Grid Search*

con Cross Validation (Búsqueda de cuadrícula con validación cruzada).

La validación cruzada implica dividir el conjunto de datos en m subbases o pliegues. En cada iteración del proceso de validación cruzada, se reserva un pliegue como conjunto de prueba, mientras que los otros $m-1$ pliegues se utilizan como conjunto de entrenamiento. Este proceso se repite m veces, asegurándose de que cada pliegue actúe como conjunto de prueba exactamente una vez.

A continuación se muestran los pasos realizados para la obtención de $k = 15$ por medio del algoritmo *Grid Search con Cross Validation*.

Inicio.

1. Mezclar aleatoriamente los datos.
2. Escoger el valor de m.
3. Dividir el conjunto de datos en m particiones.

Entrenamiento y validación.

4. Ocultar una partición.
5. Entrenar el modelo KNN con las $m-1$ particiones siguiendo los pasos dados en la sección anterior.
6. Validar con la partición oculta y almacenar el accuracy.

Repetir estos pasos m veces cambiando la partición oculta y en cada una de las particiones elegir un valor fijo de k distinto al resto. Una vez completado este proceso se procede a comparar los resultados obtenidos del accuracy en cada una de las iteraciones, eligiendo el valor de k con la métrica más alta.

Para ejecutar el algoritmo se eligió el valor de $m = 5$, y los valores candidatos para k son 5, 10, 15, 25, 30 y 50. En la Figura 4.4 se observa de manera gráfica este proceso, donde la base de datos $\overline{D}$ se divide en 6 subbases y en cada iteración se utiliza una diferente

como conjunto de prueba, mientras que las cinco restantes se utilizan como conjunto de entrenamiento.

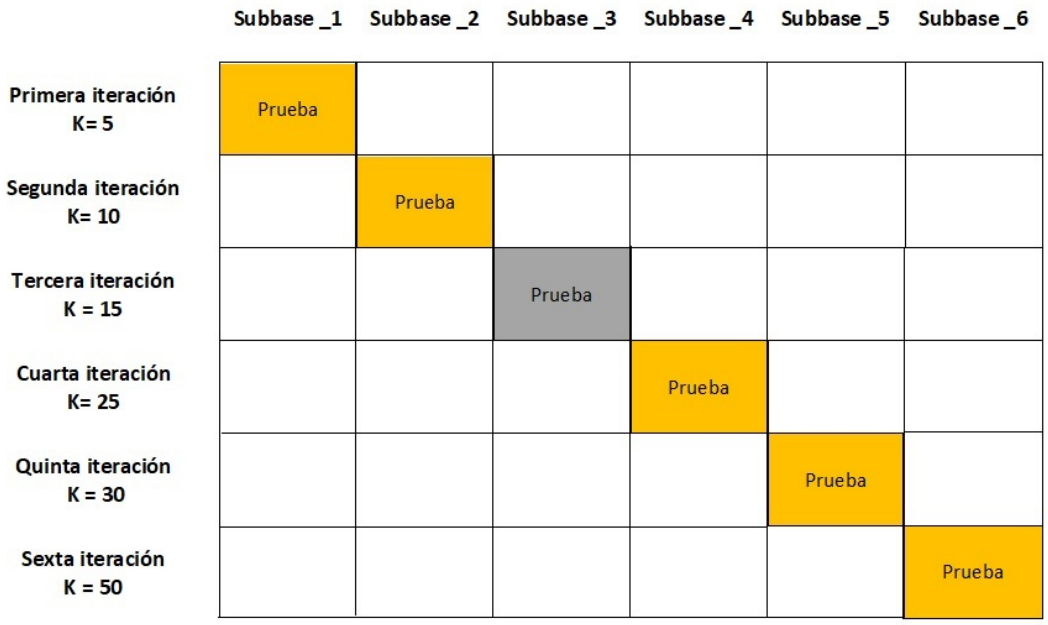

Figura 4.4: Representación gráfica del algoritmo *Grid Search con Cross Validation* para la obtención de k.

El resultado de esta búsqueda indicó que el valor óptimo para k fue 15. La elección de $k = 15$ se basa en un equilibrio entre la capacidad del modelo para ajustarse a los datos de entrenamiento y su capacidad para generalizar a nuevos datos.

4.5. Simulación

Como el modelo resulta tener un buen desempeño, se considera apto para realizar predicciones. Por consiguiente será sometido al registro cuyos valores son mostrados en el Cuadro 4.5.1. En este caso, el modelo entrenado arroja que los datos del registro corresponden a un posible ataque por parte del grupo terrorista Sudan People's Liberation Movement in Opposition (SPLM-IO).

Variable	**Dato**
Internacional logistica	0
Internacional ideologica	1
Internacional	1
Pais	58
Region	9
Nacionalidad de victima	58
Año	2032
Individual	1
Certeza	1
Latitud	18

Cuadro 4.5.1: Valores del registro sometido al algoritmo KNN.

Luego de un entrenamiento e implementación del modelo KNN para dar solución al problema de clasificación aplicado a la base de datos Global Terrorism Database (GTD), se ha obtenido un nivel de precisión, expresado como accuracy, que refleja la eficacia del modelo para la predicción de eventos terroristas. Para contextualizar estos resultados, se planea realizar una comparación directa con los hallazgos obtenidos en otras investigaciones que aplicaron algún procesamiento y modelos de aprendizaje automático a la misma base de datos utilizada en esta investigación. Al contrastar los resultados obtenidos con aquellos de otras investigaciones, se busca no sólo validar la solidez del modelo, sino también identificar posibles áreas de mejora que pueden ser abordados en futuros trabajos de investigación.

- En [14] se abordan modelos de aprendizaje automático basado en redes neuronales para predecir actividades terroristas futuras. Los autores utilizan datos del Global Terrorism Database y aplican técnicas de procesamiento de datos, como la codificación de variables y el manejo de datos faltantes. Luego, entrenan y evalúan los modelos para abordar clasificación binaria y multiclase. El modelo propuesto para la

clasificación binaria fue el híbrido CNN-LSTM, que utilizó una combinación de Convolutional Neural Network (CNN) y Long Short-Term Memory (LSTM). Por otro lado, el modelo propuesto para la clasificación multiclase fue un modelo DNN (Deep Neural Network). Los resultados mostraron que el modelo híbrido CNN-LSTM logró una precisión superior al 96 % en la clasificación binaria, mientras que el modelo DNN logró una precisión del 99.2 % en la clasificación multiclase.

- En [19] se enfoca en analizar grandes conjuntos de datos para descubrir patrones significativos y correlaciones ocultas en la base de datos GTD. Se exploran datos geoespaciales y de ubicación que permite generar información sobre la ubicación de lugares de moda alrededor de áreas afectadas por ataques terroristas, así como el uso de diversas herramientas de visualización de datos para proporcionar representaciones gráficas de los datos de la base.

En comparación con los resultados obtenidos en el primer artículo, donde se aborda la predicción de actividades terroristas futuras mediante modelos de aprendizaje automático basados en redes neuronales, nuestro enfoque difiere en la elección del algoritmo y el procesamiento de los datos. Aunque el modelo propuesto en [14] logra una precisión destacada, aborda el problema con diferentes técnicas al momento de procesar la información además de fusionar dos modelos de aprendizaje automático.

Por otro lado, el segundo artículo se centra en el análisis de grandes conjuntos de datos, descubriendo patrones significativos y correlaciones ocultas en la base de datos GTD, sin proporcionar un modelo específico entrenado que pueda dar una alerta de un posible ataque terrorista.

Conclusiones

El objetivo general de esta investigación fue identificar patrones en los ataques terroristas realizados a nivel mundial mediante un programa de aprendizaje máquina clasificatorio posterior a un análisis exploratorio de datos. Para lograrlo, se consideró a los grupos terroristas como la variable objetivo y se realizó el proceso siguiente: Se aplicaron técnicas de minería de datos a la base de datos *Global Terrorism Database* (Secciones 1.1 y 1.2) para identificar los tipos de datos, posterior a ello se realizó una exploración y visualización en la Sección 1.3 y se obtuvieron algunas de las siguientes conclusiones:

- A partir del año 2010 los ataques se intensificaron.
- Los ataques son ocasionados con mayor frecuencia en los meses de junio y diciembre.
- Los ataques son dirigidos principalmente hacia ciudadanos y propiedades particulares, seguidos de militares, policía y gobierno.
- El grupo Taliban es el que ha cometido una mayor cantidad de ataques seguido por el grupo armado ISIL.
- En México, la Liga Comunista 23 de Septiembre es el principal grupo terrorista, seguido por el EZLN. En Estados Unidos es el grupo Extremistas Antiaborto, también aparecen los eventos del 11 de septiembre de 2001.

- El país con más bajas respecto al número de ataques es Iraq. Estados Unidos de Norteamérica se encuentra en el sexto lugar, cabe mencionar que México no se encuentra entre los primeros 15.

Derivado del análisis anterior, que se encuentra en el Capítulo 1, se identificó que la base contenía datos faltantes, se puede observar en la Figura 2.1 de la Sección 2.1 que la situación era extrema, por lo que se realizó un procedimiento matemático para el "llenado" de las celdas vacías, este algoritmo se enmarca dentro del área conocida como *imputación de variables*. Consistió en aplicar el método *Expectation Maximization*, descrito a detalle en la Sección 2.2, aunque las columnas que tenían más del 75 % de entradas faltantes fueron eliminadas.

Ya con la base limpia y completa, en el Capítulo 3 se muestra el proceso de extracción de características significativas, iniciando con la codificación de variables en la Sección 3.1, luego, se aplicó el método de sección del tipo *SelectKBest*, el cual es un método estadístico basado en la comparación de los valores p obtenidos de pruebas ANOVA. Con lo que concluimos que las variables que más afectan al fenómeno son la nacionalidad extranjera del grupo perpetrador comparada con el lugar objetivo; la ideología opuesta entre el grupo y la población del lugar; la internacionalización del evento; el país y región, donde se prefiere a regiones de oriente como Iraq; la nacionalidad de la víctima; el año; si se realiza por un grupo o de forma individual y la latitud del evento, muy relacionada nuevamente con la ubicación y, por consecuencia, con el país.

Finalmente, en el Capítulo 4 se muestra el modelo KNN multiclase, se inicia describiendo su funcionamiento en la Sección 4.1, luego se plantea formalmente en la Sección 4.2. En la Sección 4.3 se muestra el proceso de entrenamiento y se combina con el algoritmo *Grid Search con Cross Validation* (Búsqueda de cuadrícula con validación cruzada), el cual realiza diversas pruebas para determinar el valor óptimo del número de vecindades k, en este caso, se determina que $k = 15$ es el de mejor eficiencia, al elegir este parámetro se obtiene un accuracy sobre el conjunto de entrenamiento de 0.9659 y de 0.8579 sobre el de prueba, con lo que se cataloga como un buen modelo. Por último, se realiza una simulación donde se muestra la predicción del modelo. Cabe mencionar que los artículos

mencionados al final del Capítulo 4. Utilizan la misma base, pero difieren del proceso realizado en esta investigación.

Con todo lo anterior, se logra el objetivo general del proyecto y el código de la programación se muestra en el repositorio que se encuentra en la dirección electrónica:

`https://acortar.link/AiTAhg`

o en el siguiente código QR:

Como investigación futura, se abre la posibilidad para la búsqueda de mejorar el accuracy del modelo sobre el conjunto de entrenamiento, ya que en este trabajo se realizó la prueba con diversos modelos combinados con una serie de métodos de selección de parámetros, obteniendo el presentado en este trabajo como el de mejor funcionamiento.

Bibliografía

[1] Charu C Aggarwal et al., *Data mining: the textbook*, vol. 1, Springer, 2015.

[2] Derrick A Bennett, *How can i deal with missing data in my study?*, Australian and New Zealand journal of public health **25** (2001), no. 5, 464–469.

[3] F. Cillufo and D. Frankin, *Combatir al terrorismo*, Revista de la OTAN **Invierno 2001/2002** (2001/2002), 12–15.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society. Series B (Methodological) **39** (1977), no. 1, 1–38.

[5] Y. Dong and C. Y. J. Peng, *Principled missing data methods for researchers*, SpringerPlus **2** (2013), no. 222, 1–17.

[6] R. A. Fisher, *Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population*, Biometrika **10** (1915), no. 4, 507–521.

[7] E. Fix and J. L. Hodges, *Discriminatory analysis. nonparametric discrimination: Consistency properties*, International Statistical Review / Revue Internationale de Statistique **57** (1989), no. 3, 238–247.

[8] Roberto Herrera, Adel Mendoza Mendoza, and Daniel Alfonso Mendoza Casseres, *Modelos de la familia exponencial*, Ingeniare (2012), no. 12, 89–98.

[9] O. Kramer, *Dimensionality reduction with unsupervised nearest neighbors*, Intelligent Systems Reference Library, vol. 51, Springer Berlin Heidelberg, Berlin, 2013.

[10] F. Medina and M. Galván, *Imputación de datos: Teoría y práctica*, Estudios estadísticos y prospectivos, Naciones Unidas, Santiago de Chile, julio de 2007.

[11] R. Nilsson, *Statistical feature selection: With applications in life science*, Linköping studies in science and technology: Dissertations, Department of Physics, Chemistry and Biology, Linköping University, 2007.

[12] María Fernanda Lerdo De Tejada Pavon, *Estimación de datos faltantes con el algoritmo em*, Ph.D. thesis, Universidad Nacional Autónoma de México, 2014.

[13] M. T. Rodriguez, *El terrorismo y nuevas formas de terrorismo*, Espacios Públicos **15** (2012), no. 33, 72–95.

[14] F. Saidi and Z. Trabelsi, *Predicting terrorist attacks using deep learning techniques*, Egyptian Informatics Journal **23** (2022), 437–446.

[15] J. L. Schafer, *Multiple imputation: a primer*, Statistical Methods in Medical Research **8** (1999), no. 1, 3–15, PMID: 10347857.

[16] S. Skiena, *Data science design manual*, Springer, 2017.

[17] N. C. (START), *The global terrorism database (gtd) [data file]*, `https://www.start.umd.edu/gtd`, 2019.

[18] B. G. Tabachnick and L. S. Fidell, *Using multivariate statistics*, 6th ed., Allyn & Bacon, Needham Heights, MA, 2012.

[19] S. S. Thakur, N. Saini and A. K. Pathak, *Data mining model framework for gtd (global terrorism database)*, International Conference on Cyber Resilience (ICCR) (2022), 1–5.

[20] L. Veres, *Prensa, poder y terrorismo*, Amnis. Revue d'études des sociétés et cultures contemporaines Europe/Amérique **4** (2004), 1–9.