



Almacenes y minería de datos

**Proyecto Final: 'Base de datos sobre Terrorismo '**

**Equipo Los + aplicados:**

Cortés Macias Gretel Penélope 317312184

Velázquez Barrón Marilú Yatzael 318353492

Peña Nuñez Axel Yael 318279754

Escalante Castañeda Lenin Alberto 420003193

Noviembre-Diciembre 2023

# Índice

**1. Introducción** **3**

1.1. Descripción del problema . . . . . 3

**2. Análisis Exploratorio** **4**

2.1. Variables numéricas . . . . . 5

2.2. Relaciones entre variables . . . . . 5

2.3. Correlaciones . . . . . 6

**3. Preprocesamiento de datos** **7**

**4. Clasificacion** **7**

4.1. Arbol de Clasificación . . . . . 7

4.2. Red Neuronal . . . . . 10

**5. Evaluación de modelos de clasificación** **12**

5.1. Evaluación de calidad del árbol . . . . . 12

5.2. Tabla de diferencias de medidas . . . . . 13

**6. Asociación** **14**

6.1. Variable Objetivo . . . . . 15

6.2. Algoritmo apriori . . . . . 15

**7. Agrupación** **17**

7.1. Análisis de Conglomerados . . . . . 17

7.2. Características de Cada Cluster . . . . . 20

**8. Conclusiones** **22**

# 1. Introducción

## 1.1. Descripción del problema

El objetivo de este proyecto es desarrollar un sistema de minería de datos, utilizando la metodología CRISP.
El problema consiste en que de acuerdo con los resultados que obtengas, puedas proponer algunas medidas y/o mejores prácticas sobre el conjunto de datos que se indica en la presente especificación.
Los terroristas explotan los puntos débiles de los países -tanto en desarrollo como desarrollados- para financiar, organizar, equipar y adiestrar a los nuevos miembros, cometer sus atentados y ocultarse para no ser detenidos. Así pues, la creación de capacidad en todos los Estados debe ser la piedra angular de la lucha mundial contra el terrorismo. Estudiar datos sobre terrorismo es de suma importancia por varias razones:
- Seguridad nacional: El terrorismo representa una amenaza significativa para la seguridad nacional en muchos países. Estudiar datos sobre terrorismo permite a los gobiernos y las agencias de seguridad comprender mejor las amenazas y tomar medidas para prevenirlas y responder a ellas de manera efectiva.
Prevención de ataques: El análisis de datos sobre terrorismo puede ayudar a identificar patrones y tendencias que podrían indicar la planificación de un ataque terrorista. Esta información puede utilizarse para prevenir atentados y salvar vidas.
- Toma de decisiones políticas: Los datos sobre terrorismo pueden informar a los tomadores de decisiones políticas sobre la formulación de políticas y estrategias para abordar el terrorismo. Comprender las causas subyacentes y las motivaciones detrás del terrorismo es esencial para desarrollar respuestas efectivas.
- Protección de derechos humanos: El estudio de datos sobre terrorismo también puede ser útil para garantizar que las respuestas gubernamentales al terrorismo respeten los derechos humanos y las libertades civiles. Esto es esencial para equilibrar la seguridad nacional con la protección de las libertades individuales.
- Sensibilización pública: La divulgación de datos sobre terrorismo puede aumentar la conciencia pública sobre la amenaza del terrorismo y la necesidad de tomar medidas para prevenirla. Esto puede fomentar la colaboración entre la sociedad y las autoridades en la lucha contra el terrorismo.
-Cooperación internacional: Dado que el terrorismo es una amenaza transnacional, el intercambio de datos y la cooperación internacional son esenciales. El estudio de datos sobre terrorismo permite a los países colaborar en la lucha contra el terrorismo y compartir información relevante.

El conjunto de datos «Global Terrorism Database» (GTD) es una base de datos de código abierto que incluye información sobre ataques terroristas en todo el mundo desde 1970 hasta 2017. La GTD incluye datos sistemáticos sobre incidentes terroristasnacionales e internacionales que han ocurrido durante este período y ahora incluye más más de 180,000 ataques. La base de datos está mantenida por investigadores del Consorcio Nacional para el Estudio del Terrorismo y las Respuestas al Terrorismo (START), con sede en la Universidad de Maryland. Este conjunto de datos se puede utilizar para abordar una de las mayores amenazas a la seguridad y la estabilidad en el mundo actual. Proporciona información valiosa que puede utilizarse para prevenir ataques, tomar decisiones políticas informadas y proteger los derechos humanos, entre otros objetivos importantes.

# 2. Análisis Exploratorio

El conjunto de datos cuenta con 135 columnas, cada columna representa una variable, y 181691 renglones, cada renglón representa un ataque terrorista.

La tabla completa de cada variable se encuentra en el archivo **tablavariable.csv**, dicha tabla contiene:

- Tipo de atributo
- Valores permitidos
- Porcentaje de valores nulos
- Medidas de tendencia central
- Información sobre variables categóricas
- Outliers

A continuación se muestran algunos de estos atributos para algunas de las variables del conjunto de datos:

## 2.1. Variables numéricas

| Variable | Tipo de Dato | Valores Permitidos | Máximo | Mínimo | Media | Desv. Est. | Distribución | Valores Perdidos |
|---|---|---|---|---|---|---|---|---|
| eventid | Numérico (Fecha) | | 201712310032 | 197000000001 | 200300000000 | 1,325,957,000 | Normal | |
| year | Numérico | | 2017 | 1970 | 2003 | 13.26 | Normal | |
| Month | Numérico | 0-12 | 12 | 0 | 6.47 | 3.39 | No aplica | |
| Day | Numérico | 0-31 | 31 | 0 | 15 | 8.81 | No aplica | |
| Country | Numérica nominal | 4-160 | 1004 | 4 | | | | |
| Latitude | Numérico | | 74.63 | -53.16 | 74.63 | | Exponencial | 4556 |
| longitude | Numérico | | 179 | -86785896 | -459 | | Exponencial | 4557 |
| specificity | Numérico | 1-5 | 5 | 1 | 1.451 | | Exponencial | 6 |
| crit1 | Binomial | 0-1 | | | | | | |
| crit2 | Binomial | 0-1 | | | | | | |
| crit3 | Binomial | 0-1 | | | | | | |
| attacktype1 | Categórica | 1-9 | | | | | | |
| attacktype2 | Categórica | 1-9 | | | | | | 175377 |
| targtype1 | Categórica | 1-22 | | | | | | |
| natly1 | Categórica | 4-1004 | | | | | | 1559 |

## 2.2. Relaciones entre variables

En el conjunto de datos existen variables fuertemente ligadas entre sí:

- eventid y approxdate representan la fecha en la que ocurrió un ataque, sin embargo approxdate contiene 95 % de valores perdidos, por lo que sería adecuado eliminar approxdate del conjunto de datos.
- Country y countrycode representan el país de incidencia del ataque, sin embargo country code es una variable categórica numérica, donde cada país contiene un número que representa al país; por lo que remover a la variable country sería adecuado al ser de tipo string. Esto sucede igualmente con las variables regioncode y region_txt; alternative y alternative_txt; attacktype_1 y attacktype1_txt; natlty1_txt y natlty1, etc.
- Location es una variable redundante, debido a que la información contenida en esta variable está contenida en las variables: **country, provstate, city, vicinity y specificity**; y es una variable tipo string con 60 % de valores perdidos, por lo que sería ideal removerla del conjunto de datos.

- Algo similar a lo anterior ocurre en la variable summary, que contiene una descripción textual del suceso; sin embargo las variables: **attacktype_1, weapon_information, targtype_1** entre otras; contienen esta misma información, por lo que sería ideal removerla.

- Muchas variables expanden información de variables anteriores, un ejemplo de esto es la variable attacktype_1, de ella se desprenden las variables attacktype_2 y attacktype_3; estas variables secundarias contienen más de 90 % de valores perdidos, por lo que realizar imputación de datos no sería muy viable; por lo que eliminarlas sería lo más adecuado.

## 2.3. Correlaciones

A continuación se muestran las correlaciones más relevantes:

- Existe una correlación alta entre INT_LOG e INT_IDEO de 0.95; esto indica que los ataques perpetrados en países diferentes a los de origen de los perpetradores están motivados por motivos ideológicos
- crit3 y alternative tienen una correlación fuerte de 0.88; esto indica que los atáques realizados en el contexto de conflictos armados no siempre pueden ser atribuidos a organizaciones terroristas, e incluso pueden ser perpetrados por fuerzas armadas de los bandos involucrados.
- claimode2 y ransompaid tienen una alta correlación de 0.81, esto indica que grupos que atribuyen responsabilidad de un ataque vía llamada suelen extorsionar a las víctimas; y estas ceden a sus demandas y realizan algún tipo de pago a los perpetradores
- claimmode3 y nreleased tienen una alta correlación de 0.8; esto indica que aquellos grupos que se atribuyen un ataque antes de realizarlo suelen liberar rehenes tras negociaciones.
- isthostkid y extended tienen una correlación notable de 0.64; esto indica que ataques en los que se toman niños como rehenes suelen extenderse por varios días.
- ndays y hostkidoutcome tienen una fuerte correlación negativa de -0.79; esto indica que mientras más diás trascurren de una toma de rehenes con niños involucrados, más probable es que tome lugar un intento de rescate.

67 variables totales tienen una correlación mayor en valor absoluto a 0.5; esto indica que las variables del conjunto de datos tienen correlaciones altas, aunado a esto;

muchas de estas variables son categóricas que emplean números para distinguir a las categorías, por lo que emplear PCA para reducción de dimensiones no es adecuado, sin embargo pueden emplearse otras técnicas como escalamiento multimensional o análisis de correspondencias múltiples. Emplear una técnida de reducción de dimensionalidad negativa sería conveniente debido al alto número de variables del conjunto de datos.

# 3. Preprocesamiento de datos

# 4. Clasificacion

## 4.1. Arbol de Clasificación

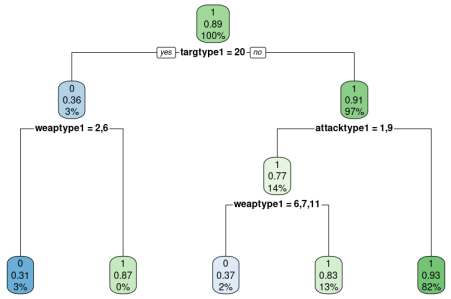
Aqui esta el codigo, en el que hicimos nuestro Arbol decisión, utilzamos las variables targType1, attackType1, weapType1, country, region. Posteriormente mostramos que son variables categoricas.

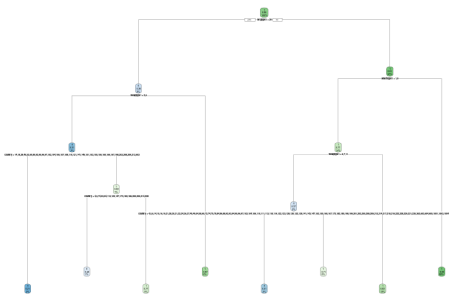
Utilizamos diferentes valores para parámetros tales como podado y/o cantidad mínima de registros en las hojas.

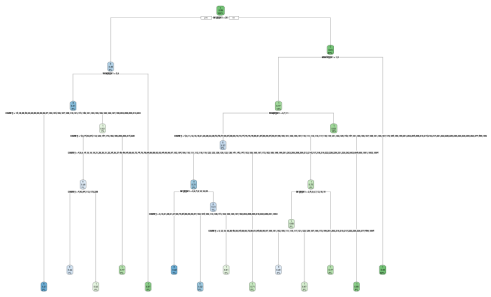
Notamos como nuestro árbol cambia al utilizar diferentes valores para la poda o la cantidad minima de registros a las hojas, dentro de la función rpart.control() que se pasa al modelo rpart(). Encontramos a minsplit: Este parámetro determina el número mínimo de observaciones que debe tener una hoja antes de que se considere para una división. Un valor más alto hará que el árbol sea más pequeño y menos complejo. cp (parámetro de complejidad): Este parámetro controla el tamaño del árbol mediante el podado. Un valor más grande permitirá árboles más complejos; un valor más pequeño podará más agresivamente, resultando en un árbol más simple.

```r
library(rpart.plot)
library(ggplot2)
# Convertir las variables a factores
datos$attacktype1 <- as.factor(datos$attacktype1)
datos$targtype1 <- as.factor(datos$targtype1)
datos$weaptype1 <- as.factor(datos$weaptype1)
datos$country <- as.factor(datos$country)
datos$region <- as.factor(datos$region)
# attacktype1 grafica
ggplot(datos, aes(x = attacktype1, fill = success)) +
  geom_bar(position = "fill") +
  ylab("Proporción") +
  ggtitle("Relación entre Tipo de Ataque y Éxito")
# Gráfico para 'attacktype1'
ggplot(datos, aes(x = attacktype1, fill = factor(success))) +
  geom_bar(position = "fill") +
  ylab("Proporción") +
  xlab("Tipo de Ataque") +
  ggtitle("Proporción de Éxito por Tipo de Ataque") +
  scale_fill_discrete(name = "Éxito")
# targtype
ggplot(datos, aes(x = targtype1, fill = factor(success))) +
  geom_bar(position = "fill") +
  ylab("Proporción") +
  xlab("Tipo de Objetivo") +
  ggtitle("Proporción de Éxito por Tipo de Objetivo")
#  modelo de árbol de decisión con un subconjunto de variables
#minsplit: Este parámetro determina el número mínimo de observaciones que debe tener un nodo antes de que se considere para una división.
#Un valor más alto hará que el árbol sea más pequeño y menos complejo.
#cp (parámetro de complejidad): Este parámetro controla el tamaño del árbol mediante el podado.
#Un valor más grande permitirá árboles más complejos;
#un valor más pequeño podará más agresivamente, resultando en un árbol más simple.
modelo_cart <- rpart(success ~attacktype1+ targtype1 + weaptype1 +
                       country + region,
                     data = datos,
                     method = "class",
                     control = rpart.control(minsplit = 1000, cp = 0.001, xval = 10))

# Calculamos la importancia de las variables
importancia <- varImp(modelo_cart)

# Graficamos la importancia de las variables
plot(importancia)

# Muestra un resumen del modelo
summary(modelo_cart)

# Visualiza el árbol
rpart.plot(modelo_cart)
```

Y a continuación tenemos como se ve el arbol con al utilizar diferetes valores.





## 4.2. Red Neuronal

Para hacer la Red Neuronal se deicidio tomar como entradas los valores .ªttacktype1", çountry", ïyear", ïmonthƶ apartir de Clasficar en base a la "succesçuando un ataque terrorista sera exitoso y cuando no segun la fecha, el pais y el tipo de ataque que se realice.

Para la contruccion de la red e cargo el dataset de los pasos anteriores, despues se selecciona los valores que queremos como entrada y como salida, y con esto podemos dividir nuestro conjunto de entrenamiento para posteriormente crear nuestra red y entrenarla devolviendonos una grafica de la red y una tabla comparativa con los valores reales de "succesƶ los valores predichos.

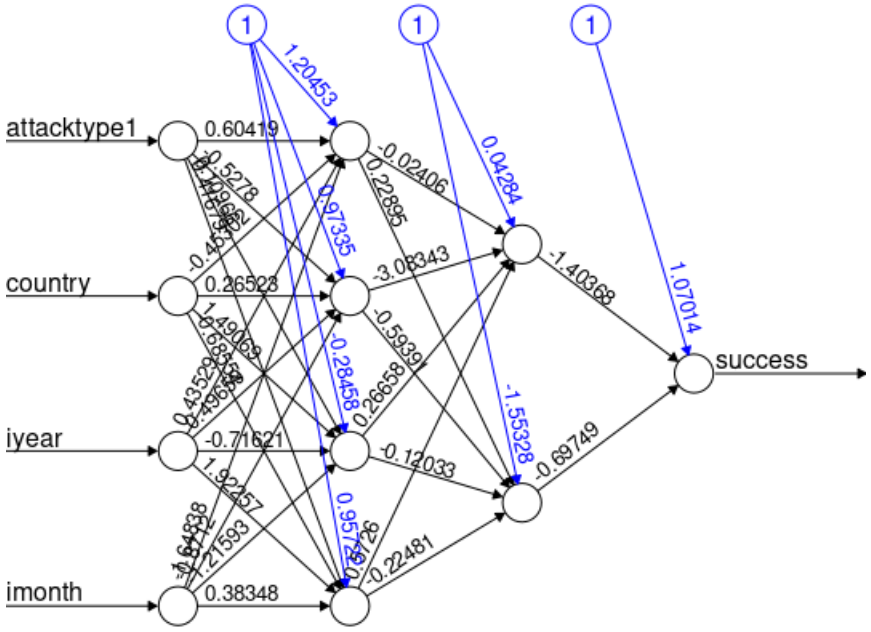
Notamos que no obtuvimos un buen resultado ya que nuestro error es demasiado grande, lo que hace que no podamos confiar en nuestra predicciones. Intenteamos solucionar el problema modificando el codigo y los parametros de Parámetros de Entrenamiento posibles. Ademas nos vimos limitados al realizar pruebas ya que la cantidad de datos a analizar es demasiado grande lo que no hace posible agregar mas entradas, capas ocultas, etc, sin que el proceso empiece a tardar horas de procesamiento.

Pensamos que las características que se eligieron para entrenar el modelo (en este caso, attacktype1, country, iyear, imonth) podrían no ser suficientes o relevantes para predecir con precisión el éxito de un ataque terrorista, pero probando otras y realizando un analisis PCA llegamos al mismo resultado.

Lo que nos lleva a que el error podria estar en la arquitectura de la Red: La elección del número de capas ocultas y el número de neuronas en cada capa puede influir significativamente en el rendimiento del modelo asi como los parametros que se eligieron, la función de Activación y Salida; pero no pudimos encontrar la configuracion correcta.

```r
# ================================================================================
# REDES NEURONALES, Proyecto
# ================================================================================


# Cargar las librerías necesarias
library(caTools)
library(neuralnet)

# Cargar los datos
datosTerrorismo <- read.csv("datosPrepTerrorismo2.csv", header = TRUE, sep = ",")

# Seleccionar las columnas de entrada y salida
columnas_entrada <- c("attacktype1", "country", "iyear", "imonth")
columna_salida <- "success"

# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
muestra <- sample.split(datosTerrorismo$success, SplitRatio = 0.66)
datos_entrenamiento <- datosTerrorismo[muestra, ]
datos_prueba <- datosTerrorismo[!muestra, ]

# Crear y entrenar la red neuronal
red_neuronal <- neuralnet("success ~ attacktype1  + country + iyear + imonth", data = datos_entrenamiento[, c(columnas_entrada, columna_salida)], hidden = c(4, 2), linear.output = TRUE)

# Realizar predicciones en el conjunto de prueba
predicciones <- predict(red_neuronal, newdata = datos_prueba[, columnas_entrada])

# Convertir las predicciones a valores binarios (0 o 1)
umbral <- 0.50
predicciones_binarias <- ifelse(predicciones >= umbral, 1, 0)

# Crear una tabla comparativa
tabla_comparativa <- data.frame(Real = datos_prueba$success, Predicho = predicciones_binarias)

# Calcular la precisión
precision <- mean(datos_prueba$success == predicciones_binarias)
cat("Precisión en el conjunto de prueba:", precision, "\n")

# Mostrar la red neuronal
plot(red_neuronal)
```

Y a continuación tenemos como se ve la red neuronal




Error: 5888.622464 Steps: 57

# 5. Evaluación de modelos de clasificación

## 5.1. Evaluación de calidad del árbol

Para esto nosotros sacamos la matriz de confusión de nuestro árbol y esto fue lo que nos dio:

```
> confusionMatrix(matriz_confusion)
Confusion Matrix and Statistics

          Realidad
Prediccion      0      1
         0   5184   2383
         1  14875 159249

               Accuracy : 0.905
                 95% CI : (0.9037, 0.9064)
    No Information Rate : 0.8896
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3351

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.25844
            Specificity : 0.98526
         Pos Pred Value : 0.68508
         Neg Pred Value : 0.91457
             Prevalence : 0.11040
         Detection Rate : 0.02853
   Detection Prevalence : 0.04165
      Balanced Accuracy : 0.62185

       'Positive' Class : 0
```

Precisión (Accuracy): El modelo tiene una precisión del 90.5 %, lo que indica que el 90.5 % de las predicciones hechas por el modelo son correctas.

Intervalo de Confianza al 95 % para la precisión: El intervalo de confianza va del 90.37 % al 90.64 %, lo que sugiere que podemos estar bastante seguros de la precisión del modelo.

Tasa de No Información (No Information Rate): Es la proporción de la clase más frecuente en los datos, en este caso es 88.96 %. La precisión del modelo es significativamente mejor que esta tasa, como lo indica el valor P muy cercano a cero.

Valor Kappa: El Kappa es de 0.3351, lo cual puede considerarse como una concordancia justa. El Kappa mide la concordancia entre las predicciones y los valores reales, teniendo en cuenta el acuerdo que podría ocurrir por casualidad.

Valor P de la prueba de McNemar: Este valor P también es cercano a cero, lo que indica que hay una diferencia significativa entre los errores de clasificación para las dos

clases (falsos positivos versus falsos negativos).

Sensibilidad (Recall o True Positive Rate): La sensibilidad es del 25.844 %, lo que significa que el modelo identifica correctamente el 25.844 % de los casos positivos reales.

Especificidad (True Negative Rate): La especificidad es del 98.526 %, lo que indica que el modelo es muy bueno identificando los casos negativos reales.

Valor Predictivo Positivo (Positive Predictive Value o Precision): Es del 68.508 %, mostrando la proporción de predicciones positivas que fueron correctas.

Valor Predictivo Negativo (Negative Predictive Value): Es del 91.457 %, indicando la proporción de predicciones negativas que fueron correctas.

Prevalencia: La prevalencia del valor positivo en tu conjunto de datos es del 11.04 %.

Tasa de Detección (Detection Rate): Es el porcentaje de verdaderos positivos detectados por el modelo, en este caso, 2.853 %.

Prevalencia de Detección (Detection Prevalence): Es el porcentaje del total de predicciones positivas, 4.165 % en este caso.

Precisión Balanceada (Balanced Accuracy): Es el promedio entre sensibilidad y especificidad, y es del 62.185 %, lo cual puede considerarse moderado

## 5.2. Tabla de diferencias de medidas

| **Métrica** | **Valor** |
|---|---|
| Precisión (Accuracy) | 90.5 % |
| Intervalo de Confianza al 95 % | (90.37 %, 90.64 %) |
| Tasa de No Información (No Information Rate) | 88.96 % |
| Valor P [Acc ¿NIR] | 2.2e-16 |
| Kappa | 0.3351 |
| Valor P de McNemar | 2.2e-16 |
| Sensibilidad (Recall) | 25.844 % |
| Especificidad | 98.526 % |
| Valor Predictivo Positivo (Precision) | 68.508 % |
| Valor Predictivo Negativo | 91.457 % |
| Prevalencia | 11.04 % |
| Tasa de Detección | 2.853 % |
| Prevalencia de Detección | 4.165 % |
| Equilibrio de Precisión (Balanced Accuracy) | 62.185 % |

Cuadro 1: Resumen de la matriz de confusión y estadísticas del modelo

# 6. Asociación

Para aplicar reglas de asociación al conjunto de datos proporcionado, primero necesitamos asegurarnos de que las variables sean adecuadas para este tipo de análisis. Las reglas de asociación son más efectivas con variables categóricas que representan la presencia o ausencia de un atributo, como podría ser el caso de las compras en un supermercado donde se analiza si un producto fue comprado o no.

En el conjunto de datos proporcionado, hay una mezcla de variables numéricas y categóricas. Variables como 'eventid', 'iyear', 'imonth', 'day', 'latitude', 'longitude' no son adecuadas para las reglas de asociación ya que son únicas para cada registro o representan medidas continuas. Sin embargo, variables como 'extended', 'country', 'region',

'specificity', 'vicinity', 'crit1', 'crit2', 'crit3', 'doubtterr', 'multiple', 'success', 'suicide', 'attacktype1', 'targtype1', 'natlty1' 'weaptype1' , entre otras, podrían ser binarizadas (es decir, convertidas en formato de presencia/ausencia) y utilizadas para el análisis de reglas de asociación.

## 6.1. Variable Objetivo

Si decidimos que la variable objetivo es, por ejemplo, 'success' (éxito del ataque terrorista), entonces podríamos aplicar reglas de asociación para descubrir qué combinaciones de las otras variables categóricas están frecuentemente asociadas con ataques exitosos o no exitosos. O también podriamos ver que tipos de ataques estan fuertemente relacionados a que tipo de armas.

## 6.2. Algoritmo apriori

Aqui utilizamos nuestro algoritmo a priori, en donde para construir reglas de asociación de alta confianza que nos permitan predecir la variable objetivo success, podemos ajustar los parámetros de este algoritmo Apriori para que se enfoque en reglas con un alto nivel de confianza.

Asi que: el valor de confidence a 0.8 para enfocarnos en reglas con al menos un 80 % de confianza

```r
#Asociacion
#cargar datos
datos <- read.csv("Documentos/Almacenes y Mineria de Datos/Proyecto/ClasificacionArboles/datosPrepTerrorismo2.csv", stringsAsFactors = TRUE)
# Preparar los datos
# Cargar las librerias necesarias
library(arules)
library(dplyr)

# Seleccionar variables relevantes en un nuevo dataset
trules <- datos %>%
  select(country, attacktype1, targtype1, weaptype1, nkill, nwound,success)

# Cambiar nkill y nwound a factores con una codificación basada en rangos
trules <- trules %>%
  mutate(nkill = case_when(
    nkill == 0 ~ 0,
    nkill < 2 ~ 1,
    nkill < 6 ~ 2,
    nkill < 16 ~ 3,
    TRUE ~ 4
  )) %>%
  mutate(nwound = case_when(
    nwound == 0 ~ 0,
    nwound < 2 ~ 1,
    nwound < 6 ~ 2,
    nwound < 16 ~ 3,
    TRUE ~ 4
  ))

# Cambiar todo a factores
trules <- trules %>%
  mutate(across(everything(), as.factor))

# Aplicar el algoritmo Apriori con los parámetros especificados
terror_rules <- apriori(trules, parameter = list(support = 0.01, confidence = 0.5, minlen = 2, maxlen = 5))

#variable objetivo success
# Inspeccionar las reglas que contienen la variable 'success' como consecuente
rules_with_success <- subset(terror_rules, rhs %pin% "success=1" | rhs %pin% "success=0")

# Inspeccionar las reglas encontradas que incluyen la variable objetivo 'success'
inspect(rules_with_success)

# Inspeccionar las reglas encontradas
inspect(terror_rules)
```

```
[73]  {country=205, nwound=0}       => {success=1} 0.01172320 0.9184994  0.01276343 1.0324878   2130
[74]  {country=147, weaptype1=5}    => {success=1} 0.01067747 0.9463415  0.01128289 1.0637852   1940
[75]  {country=147, nwound=0}       => {success=1} 0.01435954 0.9052741  0.01586210 1.0176213   2609
[76]  {country=182, nwound=0}       => {success=1} 0.01431001 0.9197029  0.01555938 1.0338407   2600
[77]  {country=209, nkill=0}        => {success=1} 0.01042429 0.8581785  0.01214700 0.9646810   1894
[78]  {country=209, weaptype1=6}    => {success=1} 0.01016011 0.8974234  0.01132142 1.0087963   1846
[79]  {country=209, nwound=0}       => {success=1} 0.01396327 0.8945698  0.01560892 1.0055885   2537
[80]  {targtype1=8, nkill=0}        => {success=1} 0.01427148 0.8766058  0.01628039 0.9853951   2593
[81]  {attacktype1=3, targtype1=8}  => {success=1} 0.01083158 0.8817204  0.01228459 0.9911445   1968
[82]  {targtype1=8, weaptype1=6}    => {success=1} 0.01093065 0.8764342  0.01247172 0.9852023   1986
[83]  {targtype1=8, nwound=0}       => {success=1} 0.01658860 0.8983607  0.01846542 1.0098498   3014
[84]  {targtype1=15, nkill=0}       => {success=1} 0.01079855 0.8605263  0.01254878 0.9673201   1962
[85]  {attacktype1=3, targtype1=15} => {success=1} 0.01026468 0.8949136  0.01147002 1.0059750   1865
[86]  {targtype1=15, weaptype1=6}   => {success=1} 0.01036375 0.8907285  0.01163514 1.0012705   1883
[87]  {targtype1=15, nwound=0}      => {success=1} 0.01473381 0.9034762  0.01630791 1.0156002   2677
[88]  {attacktype1=2, nkill=4}      => {success=1} 0.01101320 0.9852290  0.01117832 1.1074987   2001
[89]  {weaptype1=5, nkill=4}        => {success=1} 0.01063344 0.9837067  0.01080956 1.1057876   1932
[90]  {attacktype1=3, nkill=4}      => {success=1} 0.01002801 0.9717333  0.01031972 1.0923283   1822
[91]  {weaptype1=6, nkill=4}        => {success=1} 0.01098568 0.9679922  0.01134894 1.0881229   1996
[92]  {nkill=4, nwound=0}           => {success=1} 0.01259281 0.9529363  0.01321474 1.0711984   2288
[93]  {country=603, nkill=1}        => {success=1} 0.01033623 0.9665466  0.01069398 1.0864978   1878
[94]  {country=603, weaptype1=5}    => {success=1} 0.01014910 0.9215392  0.01101320 1.0359049   1844
[95]  {country=61, nkill=2}         => {success=1} 0.01069398 0.9923391  0.01077654 1.1154913   1943
[96]  {country=61, attacktype1=2}   => {success=1} 0.01213599 0.9950361  0.01219653 1.1185230   2205
[97]  {country=61, weaptype1=5}     => {success=1} 0.01474481 0.9831193  0.01499799 1.1051272   2679
[98]  {country=61, nkill=0}         => {success=1} 0.01252676 0.9738982  0.01286250 1.0947617   2276
[99]  {country=61, attacktype1=3}   => {success=1} 0.01116731 0.9873479  0.01131041 1.1098807   2029
[100] {country=61, weaptype1=6}     => {success=1} 0.01080956 0.9849549  0.01097468 1.1071906   1964
[101] {country=61, nwound=0}        => {success=1} 0.02477283 0.9872779  0.02509205 1.1098020   4501
[102] {targtype1=21, nkill=0}       => {success=1} 0.02655608 0.9442270  0.02812467 1.0614083   4825
[103] {attacktype1=3, targtype1=21} => {success=1} 0.02842188 0.9471753  0.03000699 1.0647226   5164
[104] {targtype1=21, weaptype1=6}   => {success=1} 0.02829529 0.9476498  0.02985839 1.0652559   5141
[105] {targtype1=21, nwound=0}      => {success=1} 0.03017211 0.9468048  0.03186729 1.0643061   5482
[106] {country=159, weaptype1=5}    => {success=1} 0.01218002 0.9642702  0.01263134 1.0839389   2213
[107] {country=159, nkill=0}        => {success=1} 0.01625287 0.9119827  0.01782147 1.0251624   2953
[108] {country=159, attacktype1=3}  => {success=1} 0.01601070 0.9338684  0.01714449 1.0497642   2909
[109] {country=159, weaptype1=6}    => {success=1} 0.01563644 0.9293425  0.01682527 1.0446766   2841
[110] {country=159, nwound=0}       => {success=1} 0.02738165 0.9501528  0.02881816 1.0680695   4975
[111] {nkill=3, nwound=4}           => {success=1} 0.01193785 0.9850136  0.01211948 1.1072567   2169
[112] {targtype1=14, nwound=4}      => {success=1} 0.01213049 0.9923458  0.01222405 1.1154988   2204
[113] {attacktype1=3, nwound=4}     => {success=1} 0.02873560 0.9954242  0.02886769 1.1189592   5221
[114] {weaptype1=6, nwound=4}       => {success=1} 0.02937405 0.9861419  0.02978684 1.1085250   5337
[115] {targtype1=19, nkill=0}       => {success=1} 0.02022665 0.8522727  0.02373260 0.9580422   3675
[116] {attacktype1=3, targtype1=19} => {success=1} 0.02057889 0.8625144  0.02385919 0.9695550   3739
[117] {targtype1=19, weaptype1=6}   => {success=1} 0.02020463 0.8613327  0.02345741 0.9682266   3671
[118] {targtype1=19, nwound=0}      => {success=1} 0.02093665 0.8527236  0.02455267 0.9585491   3804
[119] {country=160, attacktype1=2}  => {success=1} 0.01070499 0.9012975  0.01187731 1.0131511   1945
[120] {country=160, weaptype1=5}    => {success=1} 0.01741418 0.8833054  0.01971479 0.9929262   3164
[121] {country=160, attacktype1=3}  => {success=1} 0.01007755 0.8273836  0.01218002 0.9300644   1831
[122] {country=160, weaptype1=6}    => {success=1} 0.01050135 0.8126065  0.01292304 0.9134533   1908
[123] {country=160, nwound=0}       => {success=1} 0.02165765 0.8427929  0.02569748 0.9473859   3935
```

Las reglas que más nos van a importar son aquellas con una alta confianza y un soporte razonable. La confianza alta indica que hay una fuerte relación entre el antecedente (lhs, por sus siglas en inglés, que significa ”lado izquierdo de la regla”) y el consecuente (rhs, ”lado derecho de la regla”), en este caso ‘success=1‘. El soporte indica la proporción de transacciones en los datos que contienen tanto el antecedente como el consecuente.

En las reglas que tenemos, todas tienen una confianza muy alta (más del 80 %), lo que sugiere que son buenos predictores de ‘success=1‘. Sin embargo, para considerar cuáles son las más importantes, también deberíamos fijarnos en el soporte y el incremento (lift).

Un incremento mayor a 1 sugiere que la relación entre el antecedente y el consecuente es más fuerte de lo que se esperaría si fueran independientes.

Por ejemplo, la regla [1] ‘country=145 =¿success=1‘ tiene una confianza del 98.42 % y un incremento de 1.106, lo cual es bastante significativo. Esto significa que los ataques terroristas en el país con el código 145 tienen una alta probabilidad de ser exitosos, y esta probabilidad es ligeramente mayor de lo que se esperaría al azar.

Después de observar analizamos todas las demás reglas y notamos que tipo de armas estan mas relacionados con ciertos tipos de ataque, un ejemplo bastante alto fue:

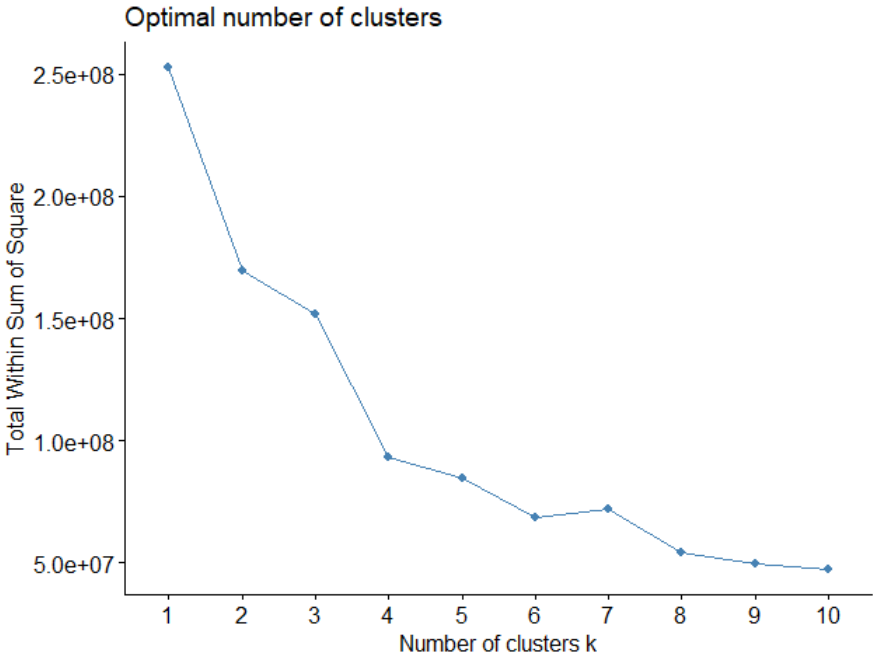
| | | | |
|---|---|---|---|
| 7 | Ataque a instalaciones/infraestructuras | 8 | Incendiario |

# 7. Agrupación

## 7.1. Análisis de Conglomerados

Se empleó el algoritmo KMeans para conocer la cantidad de grupos en el conjunto de datos; para esto, primero se removió la variable eventid del conjunto de datos, debido a que el algoritmo KMeans es sensible a variables con alto rango de valores posibles, como la variable antes mencionada.
Tras realizar esta medida, se seleccionó una muestra aleatoria sin reemplazo de 10,000 elementos para detectar el número de clusters empleando KMeans; primero se ejecutó este algoritmo para determinar el número optimo de grupos empleando la regla de suma de cuadrados.

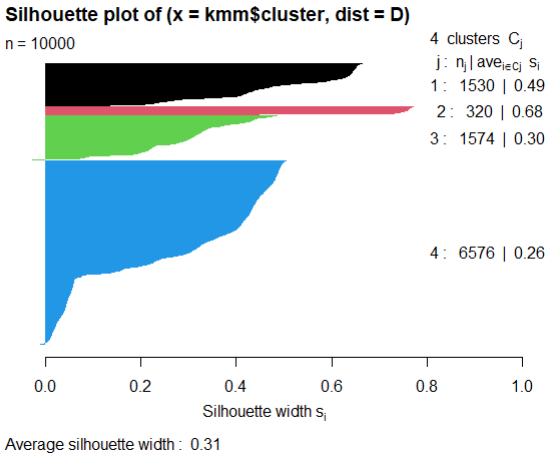
A continuación se muestra la gráfica de distancias entre cada punto y su cluster asignado, por regla general se elige como número ideal de grupos como el punto en el que existe una inflexión en esta gráfica.

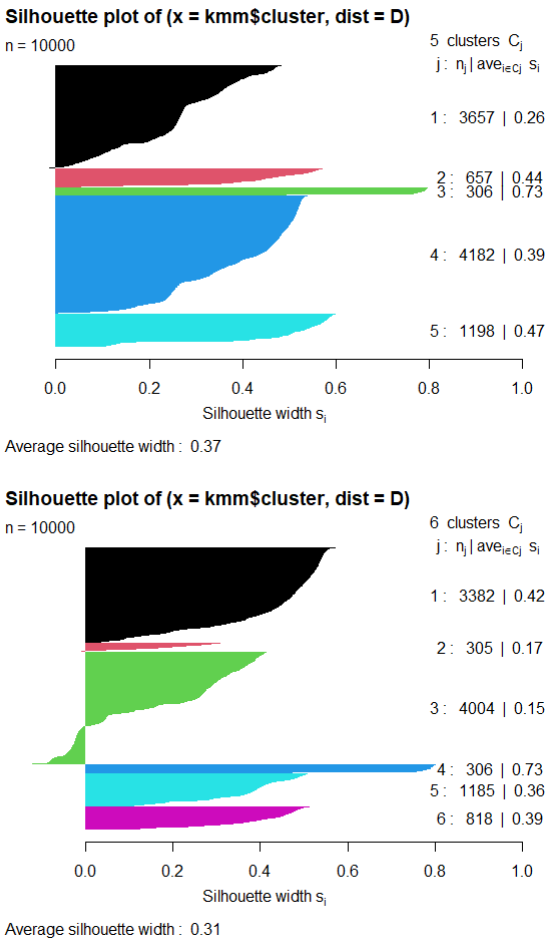



De donde se observa que el número óptimo de grupos en el conjunto de datos es de entre 4 y 6.

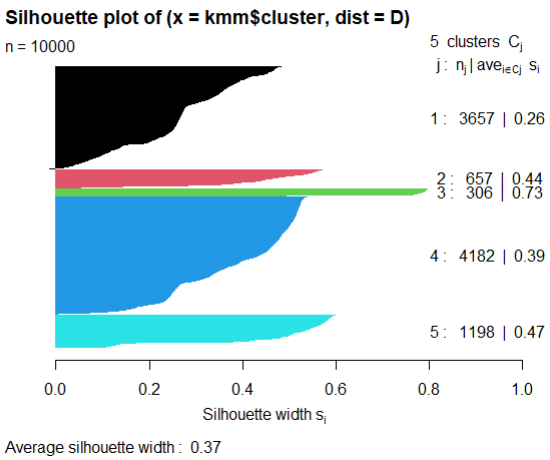
Para determinar con mayor certeza la cantidad de clusters a tomar en cuenta, se realizó una gráfica de siluetas tomando en cuenta 4, 5 y 6 clusters.

Esta gráfica es una representación de que tan bien separados está cada uno de los conglomerados, estas gráficas miden que tan similares son los elementos dentro de un cluster y que tan diferentes son en promedio los elementos de clusters diferentes.

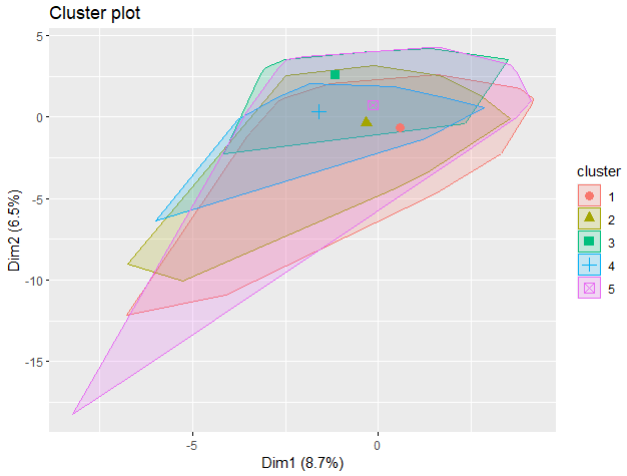





La gráfica grafica la silueta, que es una métrica que indica la separación entre cada uno de los clusters. Valores cercanos a 1 indican que los puntos tienen una distancia alta entre los miembros de su grupo y baja con los miembros de grupos diferentes, mientras que valores negativos o muy cercanos a -1 indican que estos elementos probablemente deben ser miembros de otro cluster, lo que indica que sería adecuado tomar una cantidad diferente de grupos para el algoritmo KMeans.
La gráfica además muestra el valor de silueta promedio de cada cluster.



Vemos de esta gráfica, que el número ideal de grupos es de 5, ya que tiene la silueta promedio por grupo más alta, y concentra la menor cantidad de elementos con silueta negativa. Mientras que la gráfica de siluetas con 6 clusters muestra muchos elementos con silueta negativa, lo que indica que no es una buena segmentación del conjunto de datos.
A continuación se muestra una representación en 2D de los grupos generados por el algoritmo KMeans, es importante notar que para realizar esta representación se aplican técnicas de reducción de dimensiones para reducir el conjunto a solo 2 variables, por lo que el hecho de que parezca que los 5 grupos están uno sobre el otro es un efecto inducido por reducción de dimensionalidad y no refleja con precisión exacta la apariencia de estos conjuntos en el dataset orignal.




## 7.2. Características de Cada Cluster

A continuación se muestran los tamaños de cada cluster:

- Grupo 1: 4129 elementos
- Grupo 2: 2408 elementos
- Grupo 3: 665 elementos
- Grupo 4: 306 elementos
- Grupo 5: 2427 elementos

Y a continuación se muestra una descripción de cada cluster tomando el promedio de cada una de las variables por cada conjunto de datos.
**Grupo 1:** Concentra una gran cantidad de ataques ocurridos entre 2006 y el 2015 (rango intercuartil), entre junio y octubre de estos años, que duraron como máximo un día y

ocurrieron en países del suroeste y sur de asia; como por ejemplo: Cambodia, Afganistan, India, Pakistan, Vietnam, etc.
Los objetivos de estos ataques eran: policiales, militares y gubernamentales a través de ataques armados o colocación de explosivos. La gran mayoría de estos ataques no fueron adjudicados a ningún grupo por admisión propia, y el número de muertes por ataque está en un rango de entre 1 y 2 fallecidos, pero un ataque dentro de este grupo mató a 140 personas y dejó heridas a 360.
**Grupo 2:** Ataques ocurridos entre 1992 y 2014 (rango intercuartil), ocurridos entre los meses de abril y noviembre de estos años, en Europa del Este, sur de Asia, Medio Oriente y el Norte de África.
Los ataques dentro de este grupo se caracterizan por ser del tipo armado o con bombas/explosivos y fueron hechos en su mayoría a objetivos militares y gubernamentales, el arma más común para realizar estos ataques fueron los explosivos, el ataque terrorista que provocó más víctimas mortales en este grupo cobró un saldo de 114 personas muertas y 386 heridos.
**Grupo 3:** Ataques producidos entre 1982 y 1991, en las regiónes de Centro América, Sudamérica y el Caribe, estos ataques ocurrieron en el contexto de las intervenciones norteaméricanas a países de estas regiones durante la Guerra Fría. Estos ataques fueron realizados con explosivos a objetivos militares y aeroportuarios tanto civiles como gubernamentales. El peor ataque dentro de este grupo dejó un saldo de 108 muertos y 78 heridos
**Grupo 4:** Ataques producidos entre 1978 y 1998 en Europa central, principalmente en el Reino Unido, ocurridos en el contexto del conflicto de Insurgencia Irlandesa.
Los principales objetivos de estos ataques eran militares, policiales y gubernamentales, especialmente a embajadas y sedes de organizaciones internacionales.
Los ataques fueron realizados en su mayoría por explosivos, tales como coches bomba, ataques suicidas, explosivos detonados por forma remota y proyectiles (morteros, RPG, misiles). La nacionalidad de las victimas era en su mayoría de Irlanda del Norte. El peor ataque dentro de este grupo mató a 48 personas.
**Grupo 5:** Ataques producidos entre 1987 y 2014, ocurridos en su gran mayoría en el sureste de asia (Afganistan, Bangladesh, India, Pakistan) empleando explosivos, principalmente proyectiles, coches bomba y explosivos remotos. El peor ataque de este grupo mató a 311 personas y dejó heridas a 220

# 8. Conclusiones

Este proyecto presenta la creación de una base de datos sobre terrorismo utilizando la metodología CRISP. El objetivo principal del proyecto es analizar los datos sobre terrorismo para proponer medidas y mejores prácticas en la prevención y respuesta a esta amenaza.

En el proyecto se explora el conjunto de datos utilizado, se identifican variables relevantes y se muestran correlaciones significativas entre algunas de ellas. Además, se consideran redundantes algunas variables y se eliminan para mejorar la calidad de los datos.

Entre las correlaciones más relevantes encontradas en el análisis exploratorio de los datos sobre terrorismo, se destaca la relación entre los ataques terroristas y la región geográfica, así como la relación entre el tipo de ataque y el objetivo del mismo. Estas correlaciones son importantes para entender mejor el fenómeno del terrorismo y para diseñar estrategias efectivas de prevención y respuesta.

En conclusión, el proyecto ha logrado su objetivo de crear una base de datos sobre terrorismo y de analizar los datos para proponer medidas y mejores prácticas en la prevención y respuesta a esta amenaza. Los resultados obtenidos son relevantes para entender mejor el fenómeno del terrorismo y para diseñar estrategias efectivas de prevención y respuesta.