Trabajo Final de Máster

Simulación y análisis de delitos sintéticos

Sergio Varela Lérida - Septiembre, 2021

1. Introducción.

Los estudios de simulación y el análisis de crímenes artificiales es un área en crecimiento, que comenzó entorno al año 2002 por parte de un pequeño grupo de criminólogos ambientales y geógrafos y estudia los mecanismos para general sucesos delictivos individuales (Liu & Eck, 2008).

Estudios recientes apuntan a la necesidad de este tipo de análisis y al uso de esta metodología como una herramienta útil para cuantificar los sesgos en las estadísticas policiales, generar conocimiento sobre patrones delictivos o evaluar intervenciones (Buil-Gil, Moretti & Langton, 2021; Pina-Sánchez, Buil-Gil, Brunton-Smith & Cernat, 2021). Así como arrojar algo de luz al conocimiento de la "cifra negra" y su distribución geoespacial (Buil-Gil, Medina y Shlomo, 2020). La cifra negra es la referida a la delincuencia oculta que no aparece registrada en las estadísticas oficiales. Son delitos que no han sido considerados como tales, por ausencia de denuncia, por ausencia de registro, porque no son detectados, o por otros motivos (Serrano, 2014).

En el presente estudio hemos utilizado diversas fuentes de datos. Por un lado hemos obtenido información demográfica de Barcelona, a partir de la cual hemos creado un dataset sintético, el cual simula dicha población con un total de 1.447.754 filas y con los parámetros obtenidos del censo.

Una vez creada la población con sus variables (sexo, edad, nacionalidad, lugar de residencia, nivel de educación) hemos simulado victimizaciones (hechos delictivos) a partir de los parámetros de la Encuesta de Victimización de Barcelona, añadiendo columnas a ese mismo dataset sintético con la información de los delitos simulados (si la persona ha sufrido o no un delito, si lo denunció en ese caso, qué tipo de delito sufrió,

etc.). Las encuesta de victimización es una técnica de investigación enfocadas al conocimiento de la criminalidad real. Es decir, ayuda a conocer aquellas victimizaciones que han tenido lugar pero que no aparecen en aquellas estadísticas que sólo tienen en cuenta casos en los que ha intervenido la policía, en los que hay denuncia, etc. Estos delitos que no aparecen en las estadísticas, pero que sin embargo tienen un impacto real, son los que componen la ya mencionada "cifra negra" (García-Pablos, 2013).

Por último, con ese dataset que almacena los delitos simulados, hemos creado tres modelos predictivos mediante diferentes técnicas estadísticas:

- Árboles de decisión: es una técnica que pertenece a los modelos de aprendizaje supervisado (necesitamos tener la variable dependiente en el conjunto de datos de entrenamiento) en el que se divide consecutivamente el conjunto de datos de entrenamiento en función de las variables de entrada hasta alcanzar un criterio de asignación con respecto a la variable objetivo. A veces pueden tener problemas de sobreajuste a la hora de generalizar.

- *Random forest*: es una técnica de machine learning que se utiliza para problemas de clasificación, se basa en árboles de decisión combinados con bagging. En el conjunto de datos lo hemos utilizado mediante h2o.

- *Gradient boosting*: Es una técnica de machine learning que al igual que el *random forest*, se basa en los árboles de decisión y sirve para clasificación. tanto en *random forest* como en el *gradient boosting*, existen mayores dificultades de interpretación que en los árboles de decisión. En el conjunto de datos lo hemos utilizado mediante h2o.

Todo el código y los datos utilizados se encuentran disponibles en esta misma carpeta para su consulta o para replicar los resultados obtenidos.

2. Creación de una población sintética de Barcelona.

En primer lugar, importamos los datos de la población de Barcelona por Barrios con los datos obtenidos del censo del ayuntamiento de Barcelona en 2020[1], excepto los datos socioeconómicos, ya que los más recientes son de 2018 (https://ajuntament.barcelona.cat/estadistica/castella/Estadistiques_per_temes/Poblacio_i_demografia/Poblacio/Padro_municipal_habitants/a2020/index.htm). También cargamos los datos de la Encuesta de Victimización de Barcelona 2020.

El primer paso es limpiar los datos que hemos importado para su posterior utilización, una vez hecho, vamos categorizando las variables para que podamos anexionar los datos de la encuesta posteriormente (por ejemplo, la edad por tramos). Vamos creando las columnas, empezando por un id, y después los barrios, ya que cada barrio tiene una determinada población. Una vez hecho, creamos las demás variables de la población sintética con los parámetros por barrios del censo.

```
syn <- syn %>%
  mutate(Hombre = rbinom(n = nrow(syn), size = 1, prob = a_sexo$meanage))
%>%      # Añadimos el sexo (H:1, M:0)
  mutate(Edad = rep.int(x = a_EDAD$Edad, times = a_EDAD$Total))%>%
# Añadimos la edad (Por inbtervalos)
  mutate(Superior = rbinom(n = nrow(syn), size = 1, prob = a_estudios$mea
ned))%>% # Añadimos la educación (Superior:1)
  mutate(Extranjero = rbinom(n = nrow(syn), size = 1, prob = a_extran$mea
nextra)) # Añadimos si la persona es extranjera (Extranjera:1)
```

3. Preparamos los datos de la Encuesta de Victimización de Barcelona.

Para preparar los datos de la encuesta, comenzamos por recategorizar los datos y eliminar columnas que no vamos a utilizar. Los delitos los recategorizamos en cuatro

[1] El muestreo de la Encuesta de Victimización de Barcelona 2020 se hizo sobre la población empadronada en Barcelona mayor de 16 años en diciembre de 2019. Con un error de ±1,6% para el conjunto de la muestra (IC= 95.5, 2σ).

tipologías: delitos contra la propiedad, delitos en residencias, delitos violentos y delitos que tienen que ver con vehículos. Seguimos el mismo proceso de limpieza de datos que en el anterior apartado, con la diferencia de que en este caso tenemos que tomar decisiones en cuanto a los datos missing, los cuales imputamos.

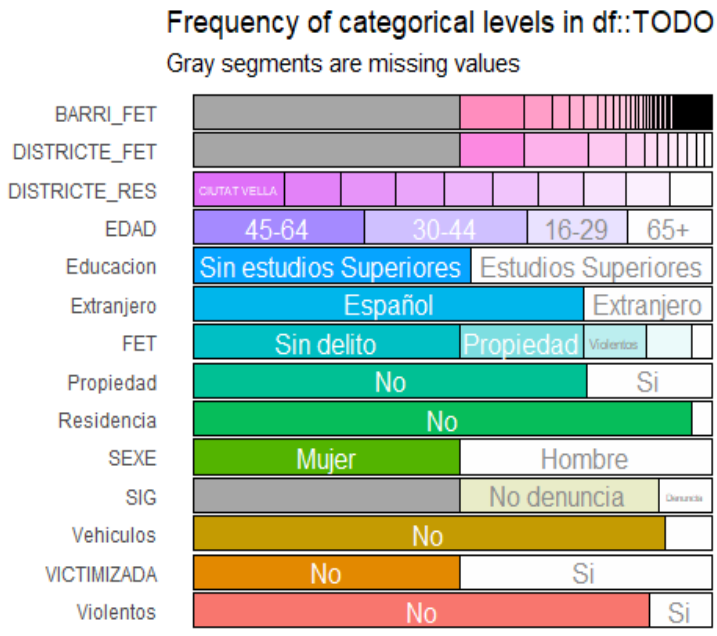
4. Descripción de los datos de la Encuesta de Victimización de Barcelona.

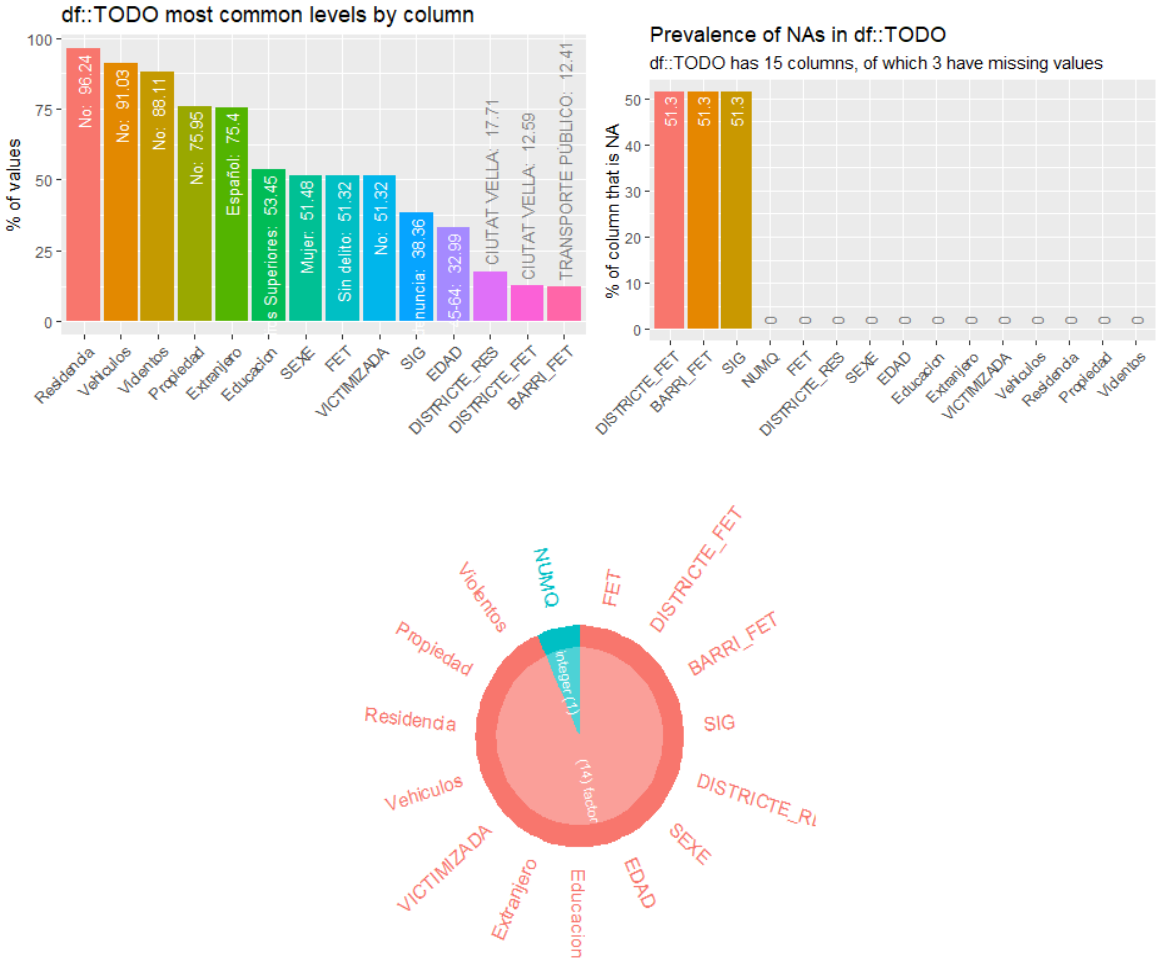
Hacemos un summary para ver cómo nos queda el dataframe. Vemos que queda bastante balanceado entre casos de delito y casos de no delito, lo cual es positivo de cara a realizar modelos.

```
##       NUMQ                FET                  DISTRICTE_FET
##  Min.   :100004   Vehículos : 577   CIUTAT VELLA      : 810
##  1st Qu.:200660   Residencia: 242   TRANSPORTE PÚBLICO: 798
##  Median :201928   Propiedad :1547   L'EIXAMPLE        : 457
##  Mean   :186085   Violentos : 765   SANT MARTÍ        : 228
##  3rd Qu.:203275   Sin delito:3301   SANTS-MONTJUÏC    : 163
##  Max.   :204705                     (Other)           : 675
##                                     NA's              :3301
##                                  BARRI_FET             SIG
##  TRANSPORTE PÚBLICO                  : 798   No denuncia:2467
##  EL RAVAL                            : 359   Denuncia   : 664
##  LA DRETA DE L'EIXAMPLE              : 211   NA's       :3301
##  SANT PERE, SANTA CATERINA I LA RIBERA: 182
##  EL BARRI GÒTIC                      : 173
##  (Other)                             :1408
##  NA's                                :3301
##        DISTRICTE_RES      SEXE         EDAD
##  CIUTAT VELLA  :1139   Mujer :3311   16-29:1261
##  NOU BARRIS    : 685   Hombre:3121   30-44:2015
##  SANT MARTÍ    : 683                 45-64:2122
##  L'EIXAMPLE    : 611                 65+  :1034
##  SANT ANDREU   : 595
```

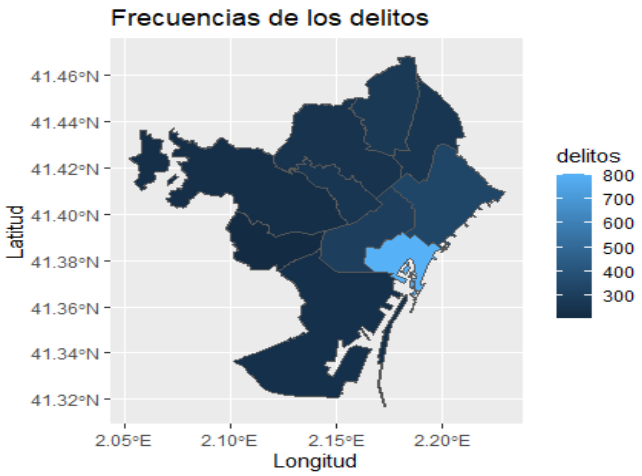
```
##  HORTA-GUINARDÓ: 565
##  (Other)       :2154
##                    Educacion          Extranjero   VICTIMIZADA Vehiculo
s
##  Estudios Superiores    :2994   Español    :4850   No:3301     No:5855
##  Sin estudios Superiores:3438   Extranjero:1582   Si:3131     Si: 577
##
##  Residencia Propiedad Violentos
##  No:6190    No:4885   No:5667
##  Si: 242    Si:1547   Si: 765
```

Nos quedan NAs que son los relativos a personas que en la encuesta dijeron no haber sufrido ninguna victimización. Los dejamos así de momento. A continuación se puede observar una descripción visual completa de los datos.





Podemos ver que el mayor número de delitos se encuentra entre los residentes del distrito de Ciutat Vella (el cual contiene 4 barrios: La Barceloneta, El Gòtic, El Raval y Sant Pere, Santa Caterina i la Ribera).



5. Modelos de Simulación de las victimizaciones.

5.1. Creación de las victimizaciones

En primer lugar simulamos las victimizaciones con las proporciones obtenidas de la Encuesta de Victimización de Barcelona. Para ello utilizamos un modelo de regresión logística, entrenado con el dataset de la Encuesta de Barcelona.

5.2. Creación del tipo de delito sufrido.

Añadimos el tipo de delito sufrido mediante la técnica del Gradient Boosting.

```
## H2OMultinomialMetrics: gbm
## ** Reported on training data. **
##
## Training Set Metrics:
## =====================
##
## Extract training frame with `h2o.getFrame("train_hex")`
## MSE: (Extract with `h2o.mse`) 0.3805246
## RMSE: (Extract with `h2o.rmse`) 0.6168667
## Logloss: (Extract with `h2o.logloss`) 1.015703
## Mean Per-Class Error: 0.6664227
## R^2: (Extract with `h2o.r2`) 0.7664027
## Confusion Matrix: Extract with `h2o.confusionMatrix(<model>,train = TR
UE)`)
```

5.3. Creación del lugar del delito.

Añadimos lugar del delito mediante la técnica del Gradient Boosting.

```
## H2OMultinomialMetrics: gbm
## ** Reported on training data. **
```

```
##
## Training Set Metrics:
## =====================
##
## Extract training frame with `h2o.getFrame("train_hex")`
## MSE: (Extract with `h2o.mse`) 0.3064844
## RMSE: (Extract with `h2o.rmse`) 0.5536103
## Logloss: (Extract with `h2o.logloss`) 0.8511026
## Mean Per-Class Error: 0.4110929
## R^2: (Extract with `h2o.r2`) 0.9836028
## Confusion Matrix: Extract with `h2o.confusionMatrix(<model>,train = TR
UE)`)
## ## Hit Ratio Table: Extract with `h2o.hit_ratio_table(<model>,train =
TRUE)`
```

5.4. Creación de la variable denuncia.

Continuamos añadiendo aquellos delitos que fueron conocidos por la policía (Denuncia), mediante un árbol de decisión. Además balanceamos los datos de entrenamiento mediante inframuestreo.

```
set.seed(13)
data_train <- TOVIC[,c(2,3,5,6:11)]
data_test <- TOVIC[,c(2,3,5,6:11)]

df_nos <- data_train %>%
  filter(SIG == 'No denuncia') %>%
  sample_frac(size = 0.3223)
df_sis <- data_train %>% filter(SIG == 'Denuncia')
data_train <- rbind(df_nos,df_sis)

arbol <- rpart(SIG~., data_train, method = 'class')
prediccion <- predict(arbol, synVIC, type = 'class')
```

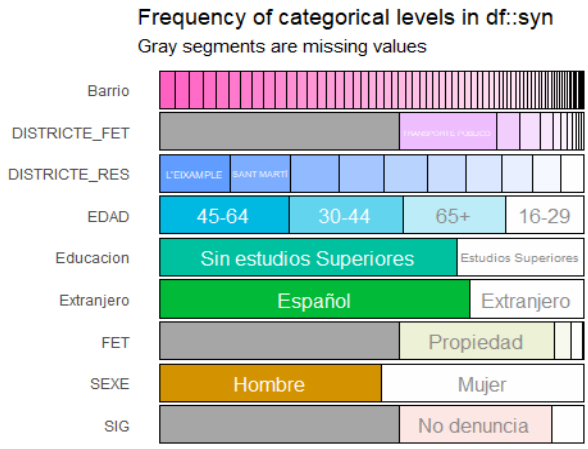
```
synVIC<-cbind(synVIC,prediccion)
synNOVIC$SIG<-NA
names(synVIC)[11] = "SIG"
syn<-rbind(synVIC, synNOVIC)
synVIC<-NULL
synNOVIC<-NULL
```

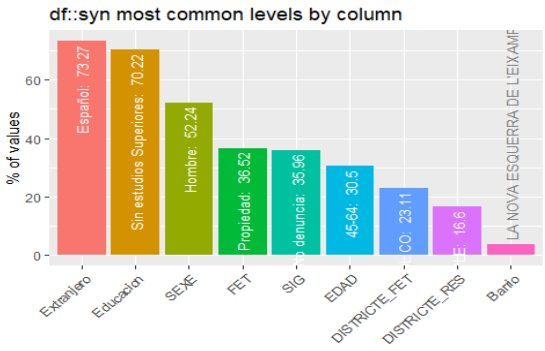
5.5. Descripción de los datos.

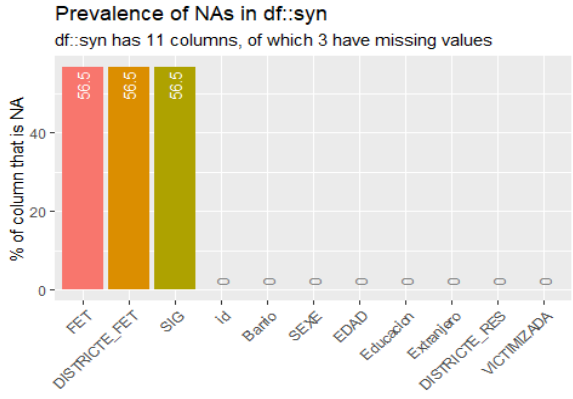
A continuación se expone un summary de los datos del dataset creado, así como algunos gráficos descriptivos.

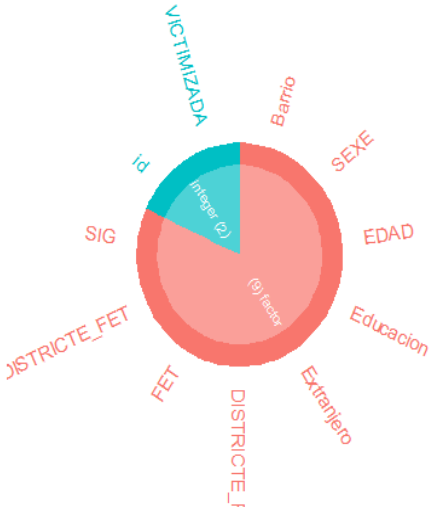
```
##        id                                      Barrio            SEXE
##  Min.   :      1   LA NOVA ESQUERRA DE L'EIXAMPLE:  52197   Mujer :691
438
##  1st Qu.: 361939   SANT ANDREU                   :  50290   Hombre:756
316
##  Median : 723878   LA SAGRADA FAMÍLIA            :  46616
##  Mean   : 723878   LA VILA DE GRÀCIA             :  44801
##  3rd Qu.:1085816   EL RAVAL                      :  41763
##  Max.   :1447754   LES CORTS                     :  40714
##                    (Other)                       :1171373
##     EDAD                          Educacion           Extranjero
##  16-29:263711   Estudios Superiores    : 431142   Español   :1060794
##  30-44:390172   Sin estudios Superiores:1016612   Extranjero: 386960
##  45-64:441572
##  65+  :352299
##
##
##         DISTRICTE_RES     VICTIMIZADA         FET            DISTRICTE
_FET
##  L'EIXAMPLE    :240393   Min.   :0.0000   Length:1447754     Length:14
47754
##  SANT MARTÍ    :207790   1st Qu.:0.0000   Class :character   Class :ch
```

```
aracter
##   SANTS-MONTJUÏC:164723   Median :0.0000   Mode   :character   Mode   :ch
aracter
##   HORTA-GUINARDÓ:152047   Mean   :0.4348
##   NOU BARRIS     :149585   3rd Qu.:1.0000
##   SANT ANDREU    :130564   Max.   :1.0000
##   (Other)        :402652
##            SIG
##   No denuncia:520589
##   Denuncia   :108953
##   NA's       :818212
```





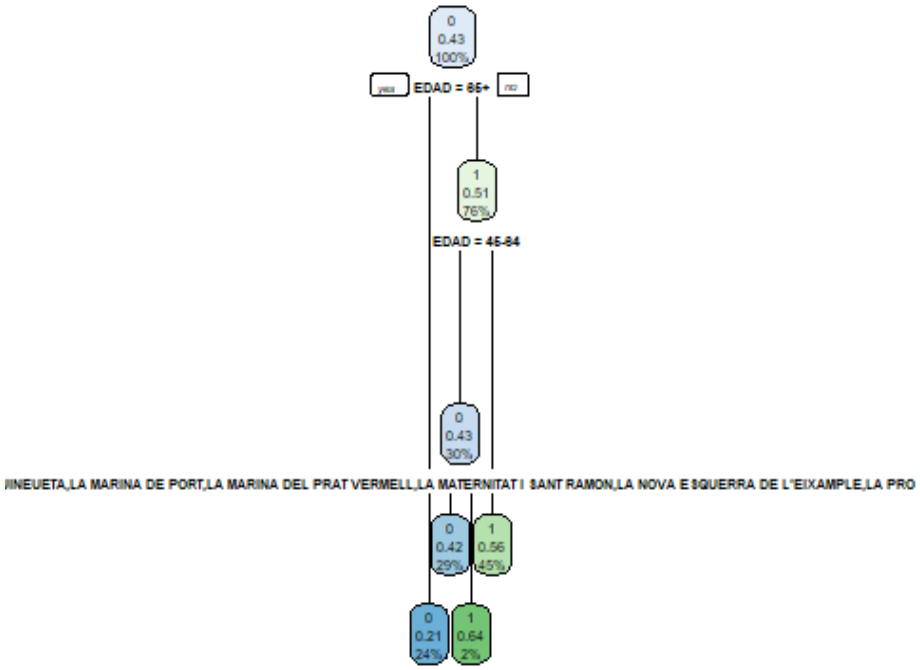




6. Modelos de Predicción de Victimizaciones.

6.1. Modelo para predecir victimizaciones.

Para predecir si una persona tiene más probabilidad de ser victimizada vamos a recurrir nuevamente a los árboles de decisión. Ya que nuestras variables predictoras son tipo factor.




```
##
## False- False+  True-  True+
##  49783  59415 104206  76146

## [1] "Sensibilidad"

## [1] 0.6046741

## [1] "Especificidad"

## [1] 0.6368742
```

A continuación, vamos a probar modelos de Random Forest y De Gradient Boost, para ver si conseguimos mejores predicciones. El modelo de Random Forest, es capaz de elevar la Sensibilidad hasta el 88% pero penaliza la especificidad hasta el 33%.

```
##
## False- False+  True-  True+
##  13964 108481  55140 111965

## [1] "Sensibilidad"

## [1] 0.8891121

## [1] "Especificidad"

## [1] 0.3369983
```

Con el modelo de Gradient Boosting, sucede algo similar, es decir, aumenta la sensibilidad, penalizando la especificidad.

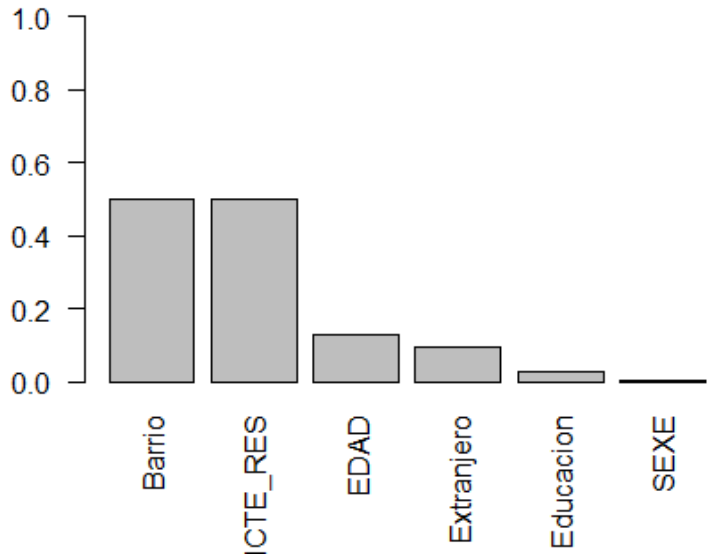
```
##
## False- False+  True-  True+
##  14018 108386  55235 111911

## [1] "Sensibilidad"

## [1] 0.8886833

## [1] "Especificidad"

## [1] 0.3375789
```

Tanto en el modelo de RF como en el de GB obtenemos resultados muy descompensados en cuanto a sensibilidad y especificidad. Por lo que la mejor opción sería el árbol de decisión, ya que nos compensa más estas dos métricas.

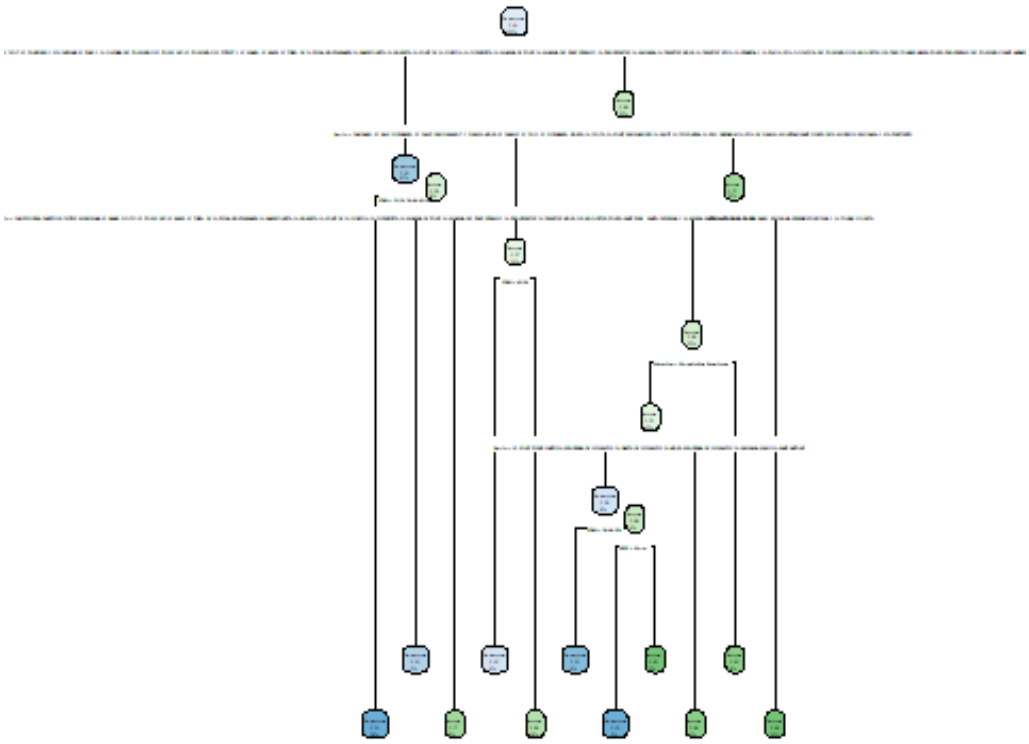
6.2. ¿Podemos predecir si la denuncia será firmada?

En este caso vamos a filtrar aquellos casos en los que efectivamente hubo un delito y trataremos de crear un modelo que nos indique quienes tienen más posibilidad de

firmar una denuncia. Vemos la importancia de las variables en términos de V de Cramer. Vamos a utilizar la técnica del inframuestreo para que así la variable SIG quede balanceada y podamos ajustar un buen modelo. Además vemos la importancia de las variables en términos de V de Cramer con respecto a SIG (si la denuncia fue interpuesta), viendo que el Distrito en el que se comete el delito es el que tiene una mayor asociación.




```
synv<-filter(syn,syn$VICTIMIZADA == 1)

SIG<-synv$SIG

trindex <- createDataPartition(synv$SIG, p = 0.8, list = F)

data_train <- synv[trindex,c(2:7,11)]

data_test <- synv[-trindex,c(2:7,11)]
```



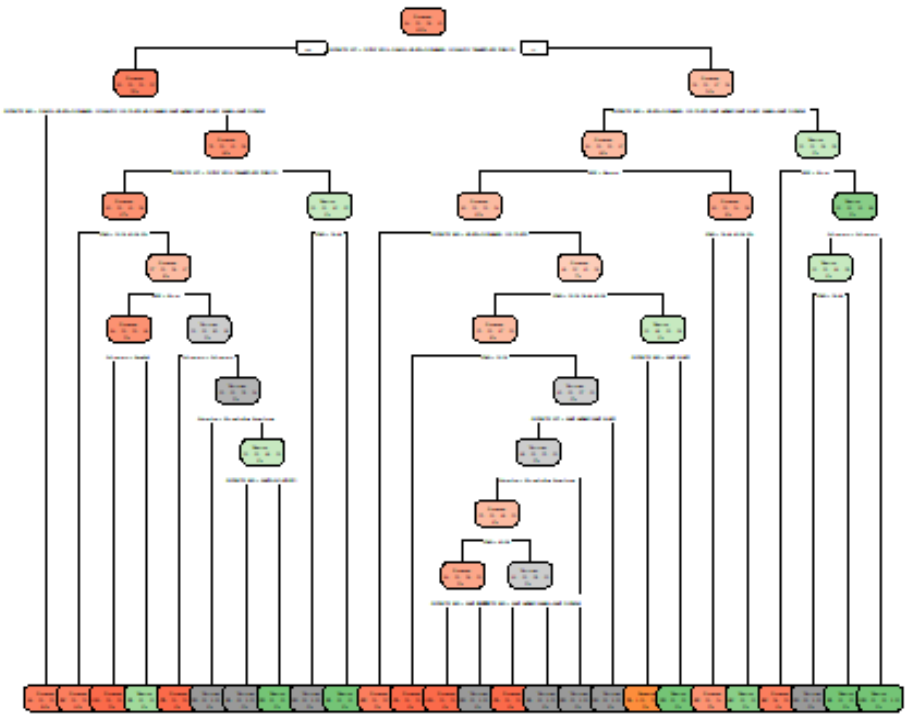
En este caso el modelo del árbol de decisión es muy bueno ya que nos clasifica muy bien los delitos que serán denunciados, especialmente destaca la Sensibilidad del 95%.

```
##
## False-  False+   True-   True+
##    874   17895   86222   20916

## [1] "Sensibilidad"

## [1] 0.9598899

## [1] "Especificidad"

## [1] 0.8281261
```

6.3. ¿Podemos predecir el tipo de victimización que ha sufrido una persona?

Para resolver esta pregunta vamos a recurrir a un modelo de árbol de decisión.

```
trindex <- createDataPartition(synv$FET, p = 0.3, list = F)

data_train <- synv[trindex,c(3:7,9:10)]

data_test <- synv[-trindex,c(3:7,9:10)]

arbol <- rpart(FET~., data_train, method = 'class')

arbol_podado <- prune(arbol, cp =
arbol$cptable[which.min(arbol$cptable[,"xerror"]),"CP"])

rpart.plot(arbol)

prediccion <- predict(arbol, data_test, type = 'class')

prediccionpod <- predict(arbol_podado, data_test, type = 'class')
```



```
##
##  False   True
##  18378 422301

## [1] "% Error"
```

```
## [1] 4.170383
```

En este caso conseguimos un modelo con un 4.17% de error.

7. Conclusiones.

Sólo la limpieza, el proceso y la estructuración de los datos nos ha llevado unas 800 líneas de código, en las cuales han surgido diferentes problemas que hemos ido resolviendo. A veces falta de datos, a veces exceso de variables, ya que pasar los datos de la encuesta a un formato estructurado es complejo por la multiplicidad de tablas que debemos combinar, o a las preguntas multirespuesta. La preparación, limpieza y anexión de los datos han sido gran parte de este TFM, por lo que los resultados siempre se podrán ampliar a partir del trabajo realizado, aprovechando el código ya creado.

Por otro lado, el acceso a los datos ha sido un poco tedioso en ocasiones debido a que las administraciones públicas no siempre nos dan la opción de descargar los datos directamente, si no que nos los dan en pdf o directamente nos los muestran en la web. Una de las limitaciones más importantes en cuanto a los modelos predictivos realizados ha sido el hecho de obtener como variables predictivas sólo variables categóricas o de tipo factor. De modo que los intentos realizados sobre los datos de la EVB no han sido totalmente satisfactorios, ya que podríamos haber utilizado más técnicas si hubiésemos dispuesto de más datos.

En cuanto a la creación de la población sintética hemos ido añadiendo variables con la información disponible y adaptándonos a la encuesta de victimización, por lo que nos ha podido limitar al tener que categorizar variables como la edad. Por otro lado, la fiabilidad de las victimizaciones simuladas, no es tan alta, ya que los modelos utilizados no tenían unos ajustes óptimos. Y, aunque, los tres modelos predictivos propuestos parecen tener un desempeño positivo, debemos tener en cuenta este detalle.

El primer modelo (predictor de victimización) ha sido el menos preciso, ya que obtuvimos una sensibilidad y especificidad (métricas escogidas para la evaluación de las predicciones cuando la variable objetivo era dicotómica) de entorno al 60 %, lo cual

puedo no parecer demasiado. Pero teniendo en cuenta la gran variedad de factores que influyen en una victimización (factores ambientales, situacionales, espacio-temporales, etc.) y que no controlamos en este estudio, no es desdeñable. Además no hemos podido mejorar las predicciones con otras técnicas como el GB o el RF, ya que la especificidad se reducía demasiado. Por lo que nos da un gran horizonte por delante para la mejora de estos modelos y de la anexión de nuevos datos que puedan ayudar a predecir.

El segundo modelo, que predice si un delito será denunciado o no, es bastante más preciso, con una sensibilidad del 95%, lo cual nos indica que las características demográficas son bastante importantes a la hora de firmar una denuncia.

Por último el modelo que predice el tipo de victimización que ha sufrido una persona, nos da un error de entorno a un 4% mediante un árbol de decisión, es decir, fue capaz de clasificar de forma correcta el 96% de los casos, aunque se puede ver en el apartado correspondiente que el árbol era bastante complejo, ya que utilizaba muchas ramificaciones (la poda no mejoró los resultados). Debemos tener en cuenta que al crear estas simulaciones, estamos simplificando la realidad, por lo que los resultados con datos reales serán previsiblemente peores.

A modo de visión prospectiva, sería enriquecedor contrastar los resultados con estadísticas policiales reales para poder medir el sesgo de éstas, así como testar los modelos entrenados con datos reales. Aunque esto último sea difícil, debido al carácter reservado de este tipo de datos.

8. Bibliografía.

Buil-Gil, D., Medina, J. y Shlomo, N. (2020). Measuring the dark figure of crime in geographic areas: small area estimation from the crime survey for england and wales. *The British Journal of Criminology, 61*, 364-388.

Buil-Gil, D., Moretti, A. & Langton, S.H. (2021). The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology.*

García-Pablos, A. (2013). *Criminología. Una introducción a sus fundamentos teóricos*. Tirant Lo Blanch, 7ª edición.

Liu, L. & Eck, J. (2008) An Overview of Crime Simulation. En Liu, L. & Eck, J. (Coords.), *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems* (pp. xiv-xxi).

Pina-Sánchez, J., Buil-Gil, D., brunton-smith, i., & Cernat, A. (2021). The impact of measurement error in models using police recorded crime rates. https://doi.org/10.31235/osf.io/ydf4b

Serrano, M.D. (2014). *El rol de la criminología para la seguridad en la sociedad contemporánea*. Dyckinson.