Final Machine Learning Project Murders by gender analysis in México 2015-2018

Carlos Baltazar

22/10/2020

Contents

1 Introduction

In the most recent years, homicides in Mexico have increased due various factors. War against drug cartels has influenced heavily in the rise of violence in all the country, but femicide has been in the public eye in the past few years.

A femicide can be defined as an "intentional murder of women because they are women, by a man motivated by hatred, contempt, pleasure or a sense of ownership of women." For many years, this crimes were labeled as "passion crimes", diminishing the systematic abuse over women.

Women collectives have been fighting to address this problem and change laws to guarantee their safety and punish this hate crime. Nevertheless, people in charge of making this changes keeps devaluating or diverting the problem. Even public opinion has sayings like "more men are killed than women", here one example.

In this context, this work is conducted to explore the various variables in the murders on Mexican population to explore if there is any difference between genders. Using Data Analysis techniques and Machine Learning algorithms to define if any variables have more weight to predict the sex of the victim.

Notice that the source of the data is in spanish, but through the document, the key words will be translated.

2 Data Analysis

The data used is the official reported by the National Institute of Statistics and Geography (INEGI by its name in Spanish, Instituto Nacional de Estadística y Geografía). These data sets can be consulted here.

With a quick glance, we notice that there are 59 variables in the data:

```
names(data)
```

```
##  [1] "ent_regis"  "mun_regis"  "ent_resid"  "mun_resid"  "tloc_resid"
##  [6] "loc_resid"  "ent_ocurr"  "mun_ocurr"  "tloc_ocurr" "loc_ocurr"
## [11] "causa_def"  "lista_mex"  "sexo"       "edad"       "dia_ocurr"
## [16] "mes_ocurr"  "anio_ocur"  "dia_regis"  "mes_regis"  "anio_regis"
## [21] "dia_nacim"  "mes_nacim"  "anio_nacim" "ocupacion"  "escolarida"
## [26] "edo_civil"  "presunto"   "ocurr_trab" "lugar_ocur" "necropsia"
## [31] "asist_medi" "sitio_ocur" "cond_cert"  "nacionalid" "derechohab"
## [36] "embarazo"   "rel_emba"   "horas"      "minutos"    "capitulo"
## [41] "grupo"      "lista1"     "gr_lismex"  "vio_fami"   "area_ur"
## [46] "edad_agru"  "complicaro" "dia_cert"   "mes_cert"   "anio_cert"
## [51] "maternas"   "lengua"     "cond_act"   "par_agre"   "ent_ocules"
## [56] "mun_ocules" "loc_ocules" "razon_m"    "dis_re_oax"
```

The first step is to select only the data that we will use:

- State where the murder occurred
- Cause of death
- Classification of death
- Victim's sex
- Age
- Day occurred
- Month occurred
- Year occurred
- Civil status
- Place where it happened

- Family violence
- Scholarship
- Relationship with aggressor

```
cols <- c("ent_ocurr", "causa_def", "lista_mex", "sexo", "edad", "dia_ocurr", "mes_ocurr",
          "anio_ocur", "edo_civil",  "lugar_ocur", "vio_fami", "escolarida", "par_agre")
```

The downloaded the data contains all the deaths in Mexico from years 2015-2018, which is the latest year available, binded into one data frame.

2.1 Data Preparation

It has to be considered that this data contains all the deaths confirmed in Mexico, natural and murders; it is needed to filter just the murder cases. The column "lista_mex" has the type of death, being 55 the equal to murder. The number of murders can be obtained the number of rows after the selection of cases:

```
data <- data %>% filter(lista_mex==55)
nrow(data)
```

```
## [1] 114085
```

However, there is a lot of data that has unknown values for relation between killer and victim, to have an accurate analysis, it is only taken into consideration values different from 88, the code for unknown.

The number of rows is displayed as well.

```
data <- data %>% filter(par_agre<88)
nrow(data)
```

```
## [1] 1850
```

Notice that "complete data" is only about 1% of the whole data set, which indicates a lack of systematic follow up to murders.

At this point, the data set, only has "coded" information, each column has a file where the data frame has been coded. It is need a dictionary to open the different files to have the data in one single data frame. The dictionary is in the same zip file downloaded.

```
dictionary <- read.csv(unzip(dl, files = "diccionario_de_datos/diccionario_datos_defunciones_registradas

values<-sapply(cols, function(col_name){
  values<-dictionary$CATÁLOGO[which(dictionary$NEMÓNICO==col_name)]
  values<-paste(values,".csv", sep="")
  values<-str_replace_all(values, fixed(" "), "")
  return(values)
})

values<-as.vector(values)
```

Once the files that are needed to fill the table are known, each variable has different considerations.

2.2 Data Cleaning

In this section, it is shown how each variable is manipulated to fit our data frame. Almost all files with the coded information have 2 columns:

- CVE = The code of the variable in the main data set.
- Description = The values to fit in the main data set.

The Description column is joined to the data set and then renamed to the original column name.

In the case of the state where it happened, we have 3 columns, as in the same table is also the community in that state. It is taken only the rows that contain a state information and move it to the data frame.

```
new_column<-read.csv(unzip(dl, files = paste("catalogos/",values[1], sep="")))
new_column<-new_column %>% rename(ent_ocurr=cve_ent) %>% mutate_if(is.factor, as.character) %>% filter(

data <- left_join(data, new_column) %>% select(-ent_ocurr, -cve_mun, -cve_loc)
```

Cause of death has the place where it happened in the same string, we use REGEX to filter different places into just the cause.

Here is one example:

```
data$causa_def<-str_replace_all(data$causa_def, "^Agresióncondisparodearmacorta\\D+" , "Disparo de arma
```

Age has special coding, with a number predeceasing the actual value, being:

- 1 - hours
- 2 - days
- 3 - months
- 4 - years

For computation, we take all 1, 2, and 3 as less than one year and remove all unknown values for year

```
data$edad<-str_replace_all(data$edad, "^[1-3]\\d*", "0")
data$edad<-str_replace_all(data$edad, "^[4]", "")
data$edad<-as.numeric(data$edad)
data <- data %>% filter(edad < 998)
```

For date, remove unknown values

```
data <- data %>% filter(dia_ocurr < 99)

new_column<-read.csv(unzip(dl, files = paste("catalogos/",values[7], sep="")))
new_column<-new_column %>% rename(mes_ocurr=CVE)

data <- left_join(data, new_column) %>% select(-mes_ocurr) %>% rename(mes_ocurr=DESCRIP)

data <- data %>% filter(anio_ocur < 9999)
```

For relationship, we add a column based on the sex of the aggressor, and for unknown or unrelated we label as “Not applies”. The original data has 2 similar values, does not applies and unrelated, because of its similarity, they are merged into one single variable. Notice that “unknown” is not modified.

```
new_column<-read.csv(unzip(dl, files = paste("catalogos/",values[13], sep="")))
new_column<-new_column %>% rename(par_agre=CVE)
new_column$DESCRIP<-as.character(new_column$DESCRIP)
new_column$DESCRIP<-str_replace_all(new_column$DESCRIP, fixed(" "), "")

sexo_agresor<-c("Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "
                "Mujer",  "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "
                "Hombre", "Mujer",  "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre",
                "Mujer",  "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "
                "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "Mujer", "Hombre", "
                "No especificado", "No especificado", "No especificado", "No especificado", "Hombre", "
                "No especificado", "No especificado", "No aplica", "No aplica")

new_column <- cbind(new_column, sexo_agresor)

data$par_agre<-replace(data$par_agre, data$par_agre==72, 71)
data <- left_join(data, new_column) %>% select(-par_agre) %>% rename(par_agre=DESCRIP)
```

Finally all columns are converted into factors, as the caret package will be used.

2.3 Exploratory Data Analysis

First, the evolution of available data for murders per sex with the parameters we need is presented. (Men = Hombre, Women = Mujer in spanish):

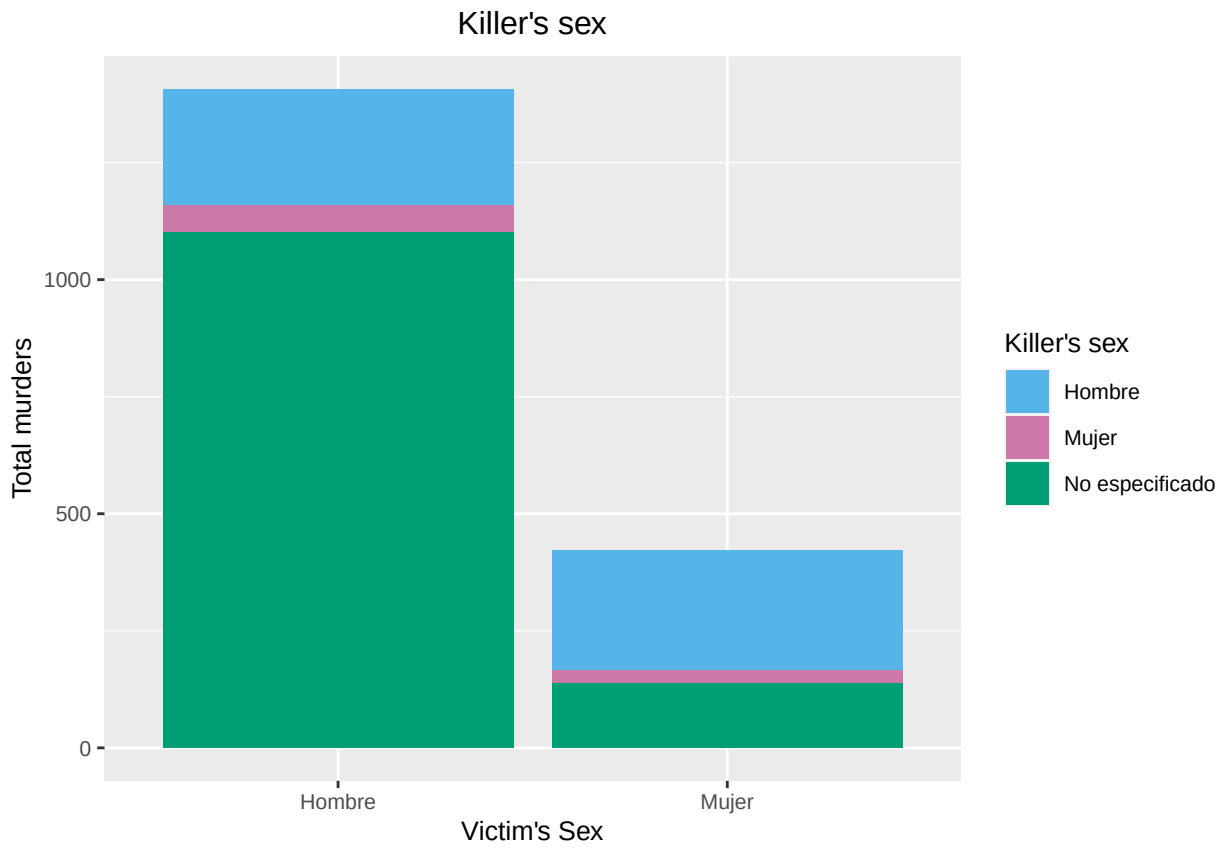
Table 1: Murders per year for men

Year	Number
2015	384
2016	480
2017	339
2018	190

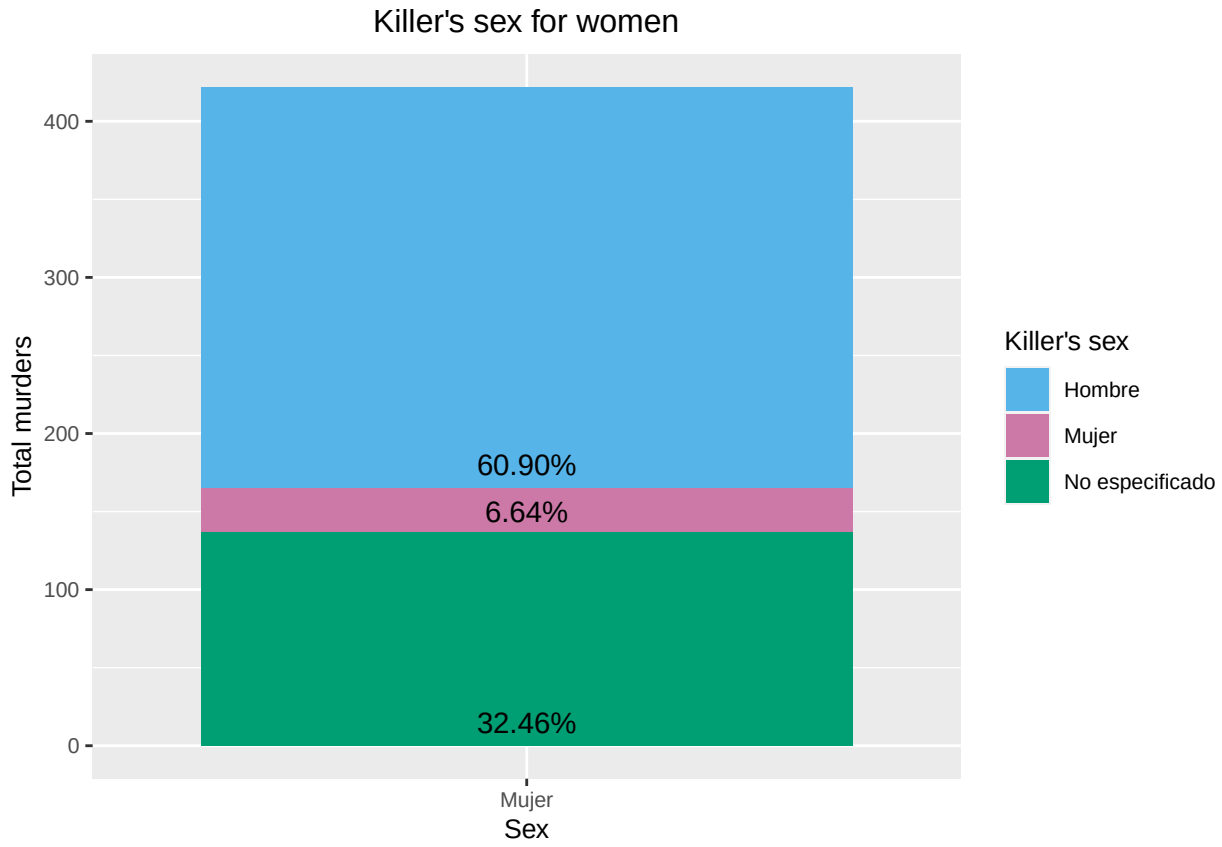
Table 2: Murders per year for women

Year	Number
2015	133
2016	145
2017	76
2018	65

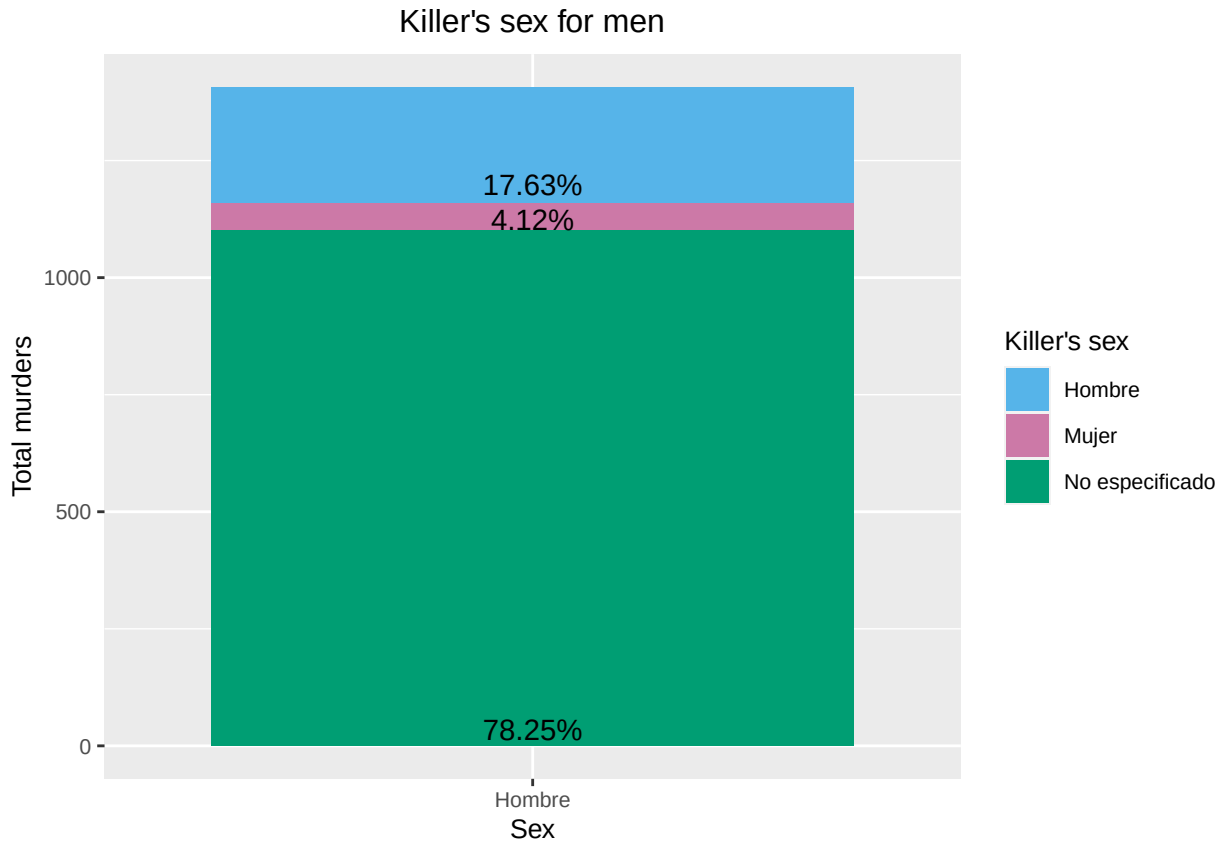
Next, the difference between the gender of the aggressor for each gender:



Let's separate the graph only for women. To have a better understanding of the data, the percentage is added manually:

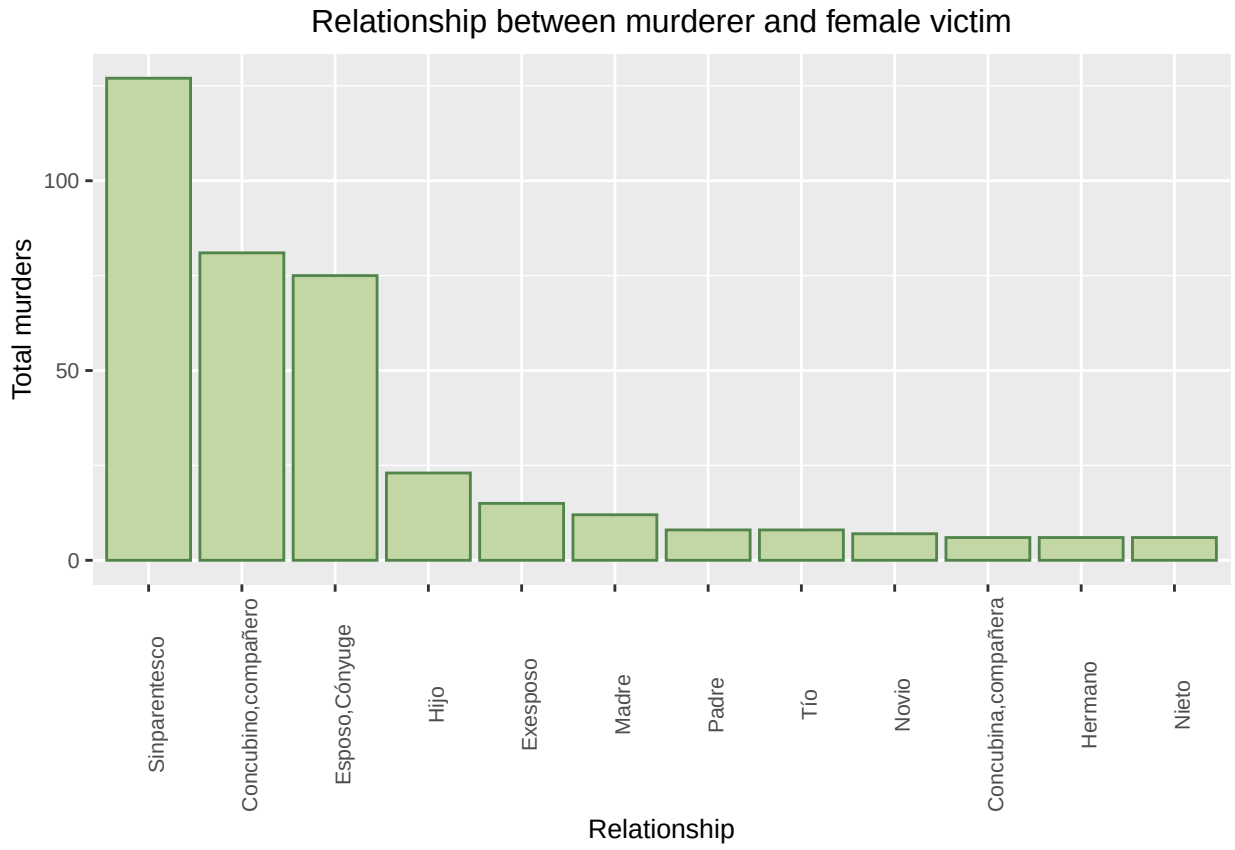


The same procedure is applied to the graph only for men, adding in the percentage in the graph:

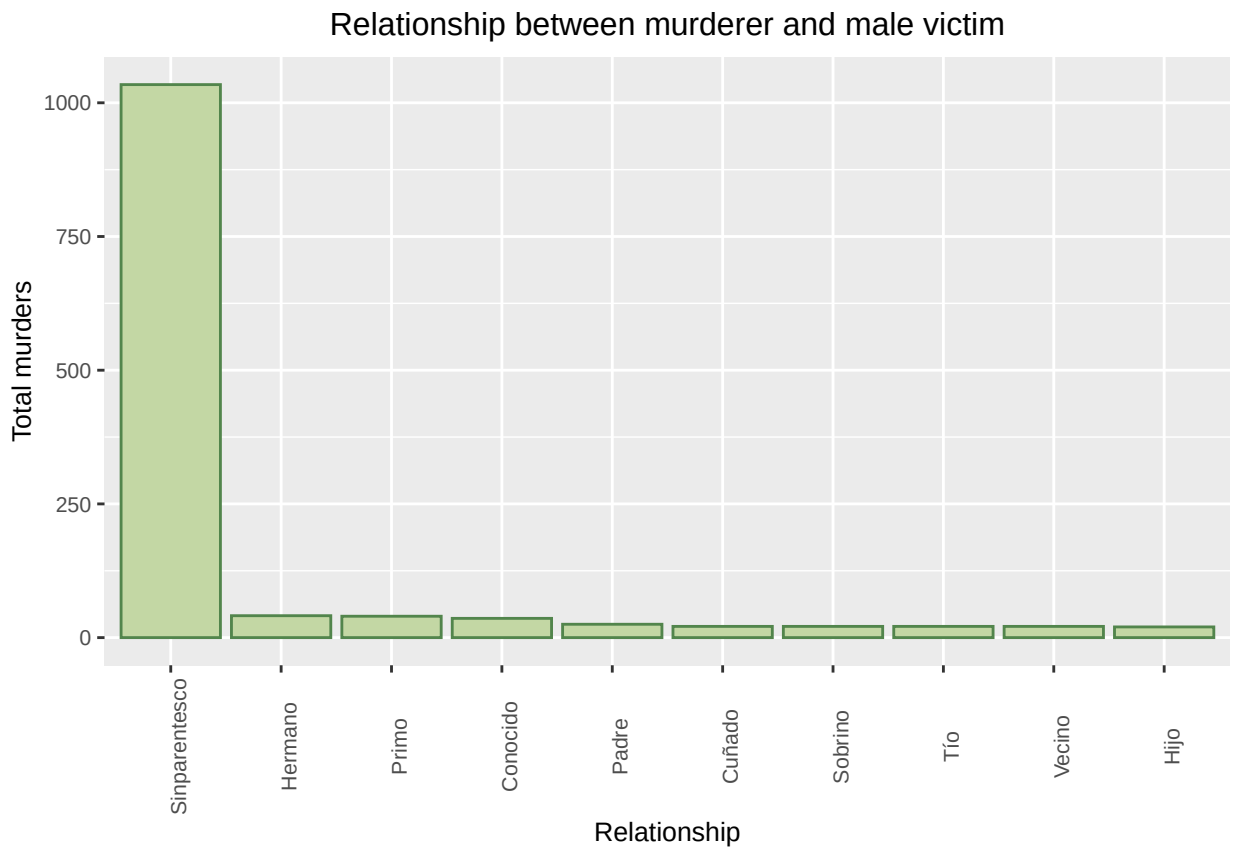


From these graphs, a significant difference between both, regarding the proportions of the attacker's sex, is evidenced.

Now, let's explore the relation between the killer with the victim, taking a look first only for women:

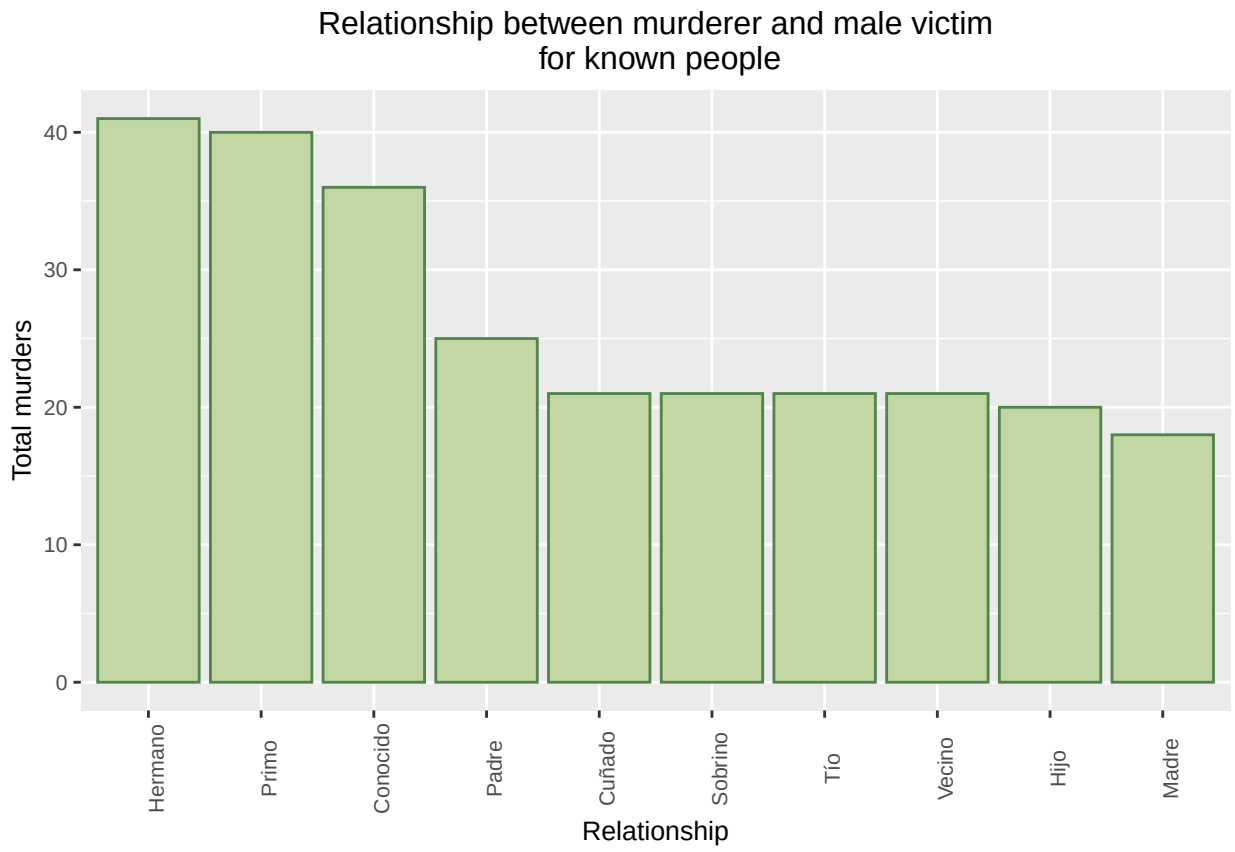


And now for male victim:



Notice that for both genders, the killer had no relationship with the victim (Sin parentesco).

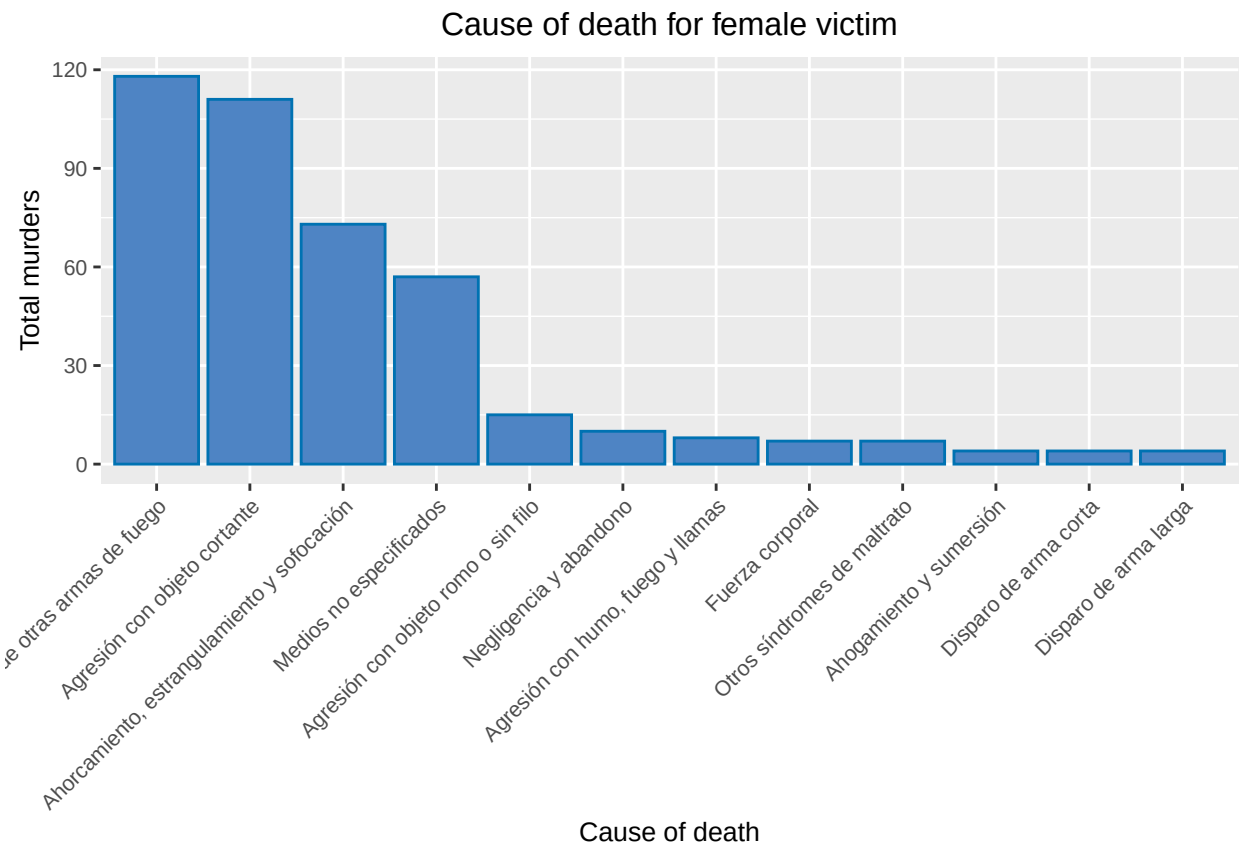
To have a better understanding murders between known individuals, let's filter the unrelated murderers for men:



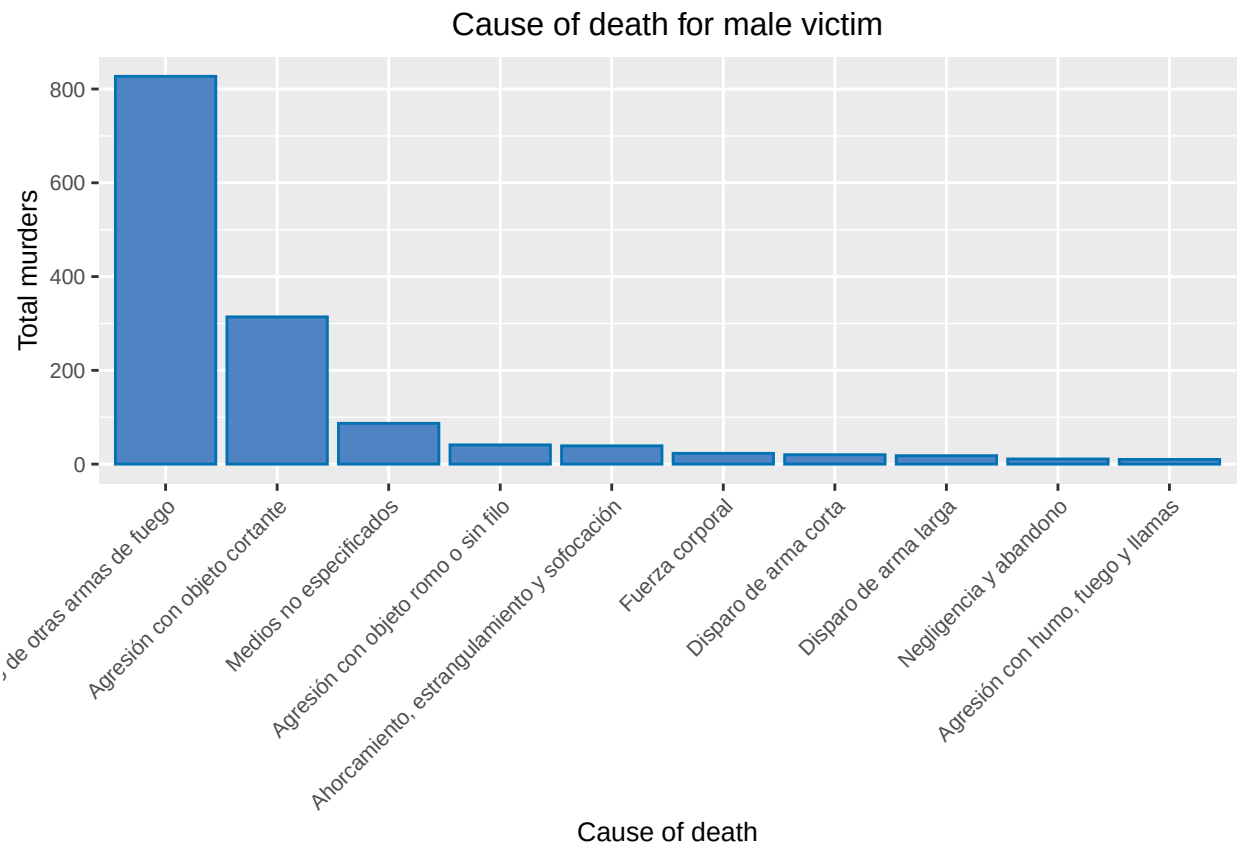
Observe that almost all killers for men in this graph, can be considered to have a "blood" relationship (Brother, Cousin, Acquaintance, Father, Brother-in-law, Nephew, Uncle, Neighbor, Son, Mother in the same order of the graph), rather than through partnership as shown in the women's graph (Partner, Husband, Son, Ex-husband, Mother, Father, Uncle, Boyfriend, Girlfriend, Brother, Grandson).

Finally, the exploration of how they died, making the plots, again, for women and men:

Causes of death for women:



Causes of death for men:



It can be observed that in the top 2 causes, both, woman and men, have been killed by a gun (armas de fuego) and sharp object aggressions (agresión con objetos cortantes). But the difference in the amount of murders by choking (Ahorcamiento, estrangulación y sofocación) between both genders is remarkable.

Finally, the last 2 graphs are wrapped in one chart, keeping only the cases where the number is bigger than 10:

For women:

Table 3: Relation between murderer and cause of death

Relationship	Cause	Number
Sinparentesco	Disparo de otras armas de fuego	59
Concubino,compañero	Agresión con objeto cortante	30
Sinparentesco	Agresión con objeto cortante	27
Esposo,Cónyuge	Agresión con objeto cortante	24
Esposo,Cónyuge	Disparo de otras armas de fuego	23
Concubino,compañero	Ahorcamiento, estrangulamiento y sofocación	19
Sinparentesco	Ahorcamiento, estrangulamiento y sofocación	18
Concubino,compañero	Disparo de otras armas de fuego	11
Concubino,compañero	Medios no especificados	11

For men:

Table 4: Relation between murderer and cause of death

Relationship	Cause	Number
Sinparentesco	Disparo de otras armas de fuego	675
Sinparentesco	Agresión con objeto cortante	184
Sinparentesco	Medios no especificados	52
Sinparentesco	Agresión con objeto romo o sin filo	28
Sinparentesco	Ahorcamiento, estrangulamiento y sofocación	26
Primo	Disparo de otras armas de fuego	20
Sinparentesco	Fuerza corporal	17
Conocido	Disparo de otras armas de fuego	16
Hermano	Agresión con objeto cortante	16
Hermano	Disparo de otras armas de fuego	16
Primo	Agresión con objeto cortante	16
Sinparentesco	Disparo de arma corta	16
Conocido	Agresión con objeto cortante	13
Tío	Disparo de otras armas de fuego	12

It has been pointed in the data exploration that there are similarities that both victim's genders share, but also some significant differences that, knowing the conditions of the death, we can use Machine Learning algorithms to predict the victim's sex.

3 Models

The caret package will be used to test different Machine Learning models, compare them and get the best approach.

First, it is need to separate the data in 80% train set and 20% test set

```
set.seed(1, sample.kind="Rounding")
index <- createDataPartition(y = data$sexo, times = 1, p = 0.2, list = FALSE)
train_set <- data[-index,]
temp <- data[index,]

validation <- temp %>%
  semi_join(train_set, by = "edad") %>%
  semi_join(train_set, by = "anio_ocur") %>%
  semi_join(train_set, by = "causa_def")

removed <- anti_join(temp, validation)
train_set <- rbind(train_set, removed)

rm(temp, removed, index)
```

Here is presented again the predictors on which the sex of the victim will be calculated, the columns that we selected in the beginning, which included:

- State where the murder occurred
- Cause of death

- Classification of death
- Victim's sex
- Age
- Day occurred
- Month occurred
- Year occurred
- Civil status
- Place where it happened
- Family violence
- Scholarship
- Relationship with aggressor

The first and most basic model, is to predict predict Male for all, as we saw the number of men killed are greater than women. This value will be used as baseline to compare the results of each model.

```
base<-mean("Hombre"==validation$sexo)
```

A table to store and compare the results the results is constructed:

Table 5: Model used and result

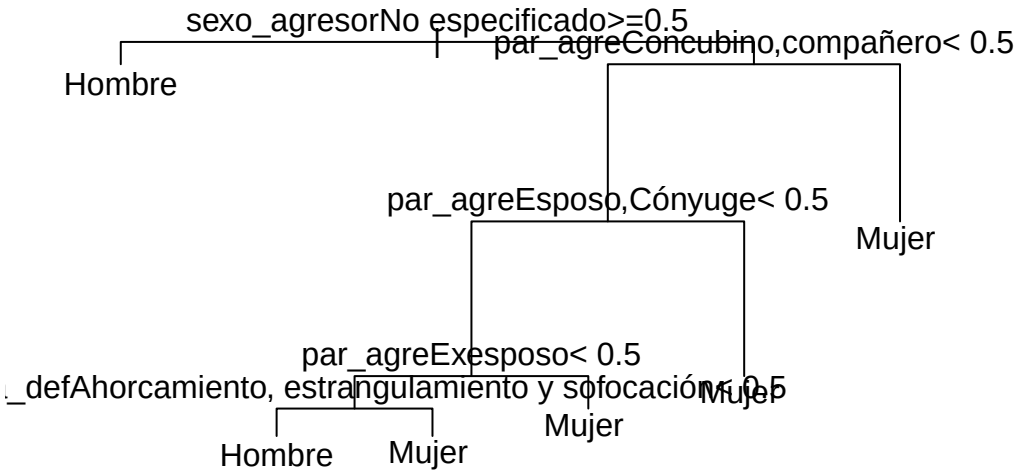
Method	Accuracy
All Men	0.775

Classification And Regression Trees (CART) model:

```
train_CART <- train(sexo ~ .,
                    method = "rpart",
                    tuneGrid = data.frame(cp = seq(0, 0.05, 0.002)),
                    data = train_set)

rpart_preds <- predict(train_CART, validation)
CART <- mean(rpart_preds == validation$sexo)
```

The decision tree for this model:



Observe that the significant differences found in the data exploratory section, agree with the results of the CART model, being aggressor's gender and relation with the victim the variables used in the model.

The results binded to the comparative table.

Table 6: Model used and result

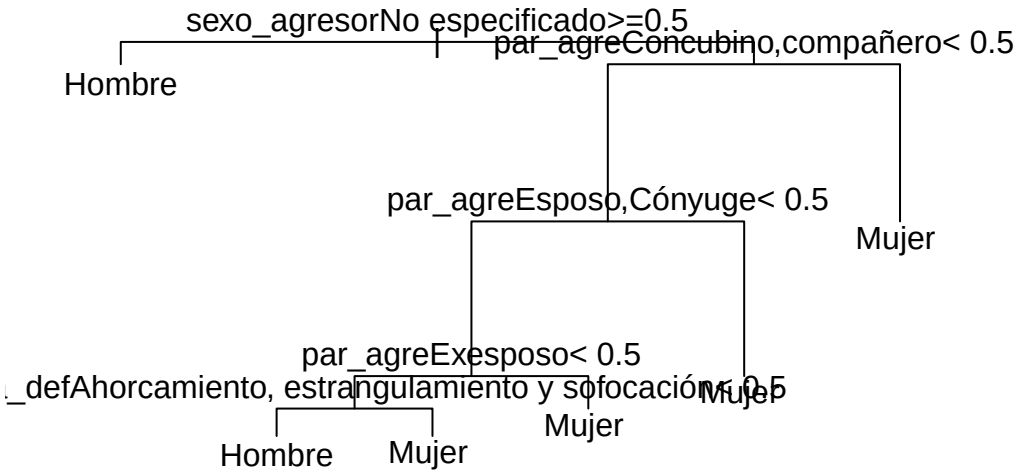
Method	Accuracy
All Men	0.7750000
Decisions Tree	0.8583333

The caret package has other method that has not previously tested, CART 1SE:

```
train_rpart <- train(sexo ~ .,
                     data = train_set,
                     method = "rpart1SE")

rpart_preds <- predict(train_rpart, validation)
CART1SE <- mean(rpart_preds == validation$sexo)
```

The decision tree:



The results table updated:

Table 7: Model used and result

Method	Accuracy
All Men	0.7750000
Decision Tree	0.8583333
Decision Tree 1SE	0.8583333

Notice that this method does not need a tuning parameter and has the same result for the first method. This can be helpful in case our choices for tuning does not include the best fit.

The KNN model:

```
train_knn <- train(sexo ~ .,
                   method = "knn",
                   data = train_set,
                   tuneGrid = data.frame(k = seq(3, 51, 2)),
                   trControl = trainControl(method = "cv", number = 10, p = 0.9))

knn_preds <- predict(train_knn, validation)
knn <- mean(knn_preds == validation$sexo)
```

The updated chart:

Table 8: Model used and result

Method	Accuracy
All Men	0.7750000
Decision Tree	0.8583333
Decision Tree 1SE	0.8583333
KNN	0.8361111

For the GLM model:

```
train_glm <- train(sexo ~ .,
                   method = "glm",
                   data = train_set)

glm_preds <- predict(train_glm, validation)
GLM<-mean(glm_preds == validation$sexo)
```

The updated table:

Table 9: Model used and result

Method	Accuracy
All Men	0.7750000
Decision Tree	0.8583333
Decision Tree 1SE	0.8583333
KNN	0.8361111
GLM	0.8416667

Finally, Random Forest model:

```
train_rf <- train(sexo ~ .,
                  data = train_set,
                  method = "rf",
                  ntree = 100,
                  tuneGrid = data.frame(mtry = seq(1:7)))

rf_preds <- predict(train_rf, validation)
RF<-mean(rf_preds == validation$sexo)
```

The table with all results:

Table 10: Model used and result

Method	Accuracy
All Men	0.7750000
Decision Tree	0.8583333
Decision Tree 1SE	0.8583333
KNN	0.8361111
GLM	0.8416667

Method	Accuracy
Random Forest	0.8611111

Let's see the Variable Importance for the model with greater accuracy, the Random Forest Model:

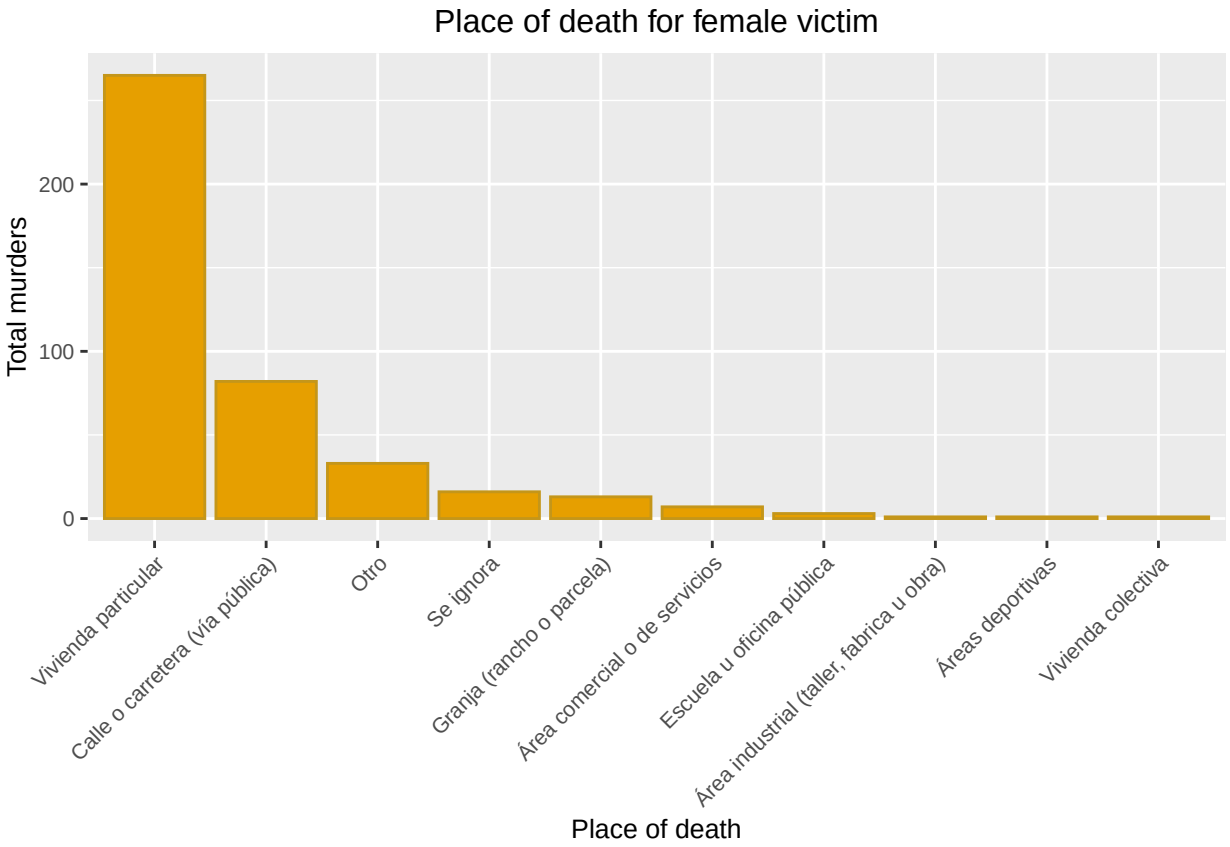
```
varImp(train_rf)
```

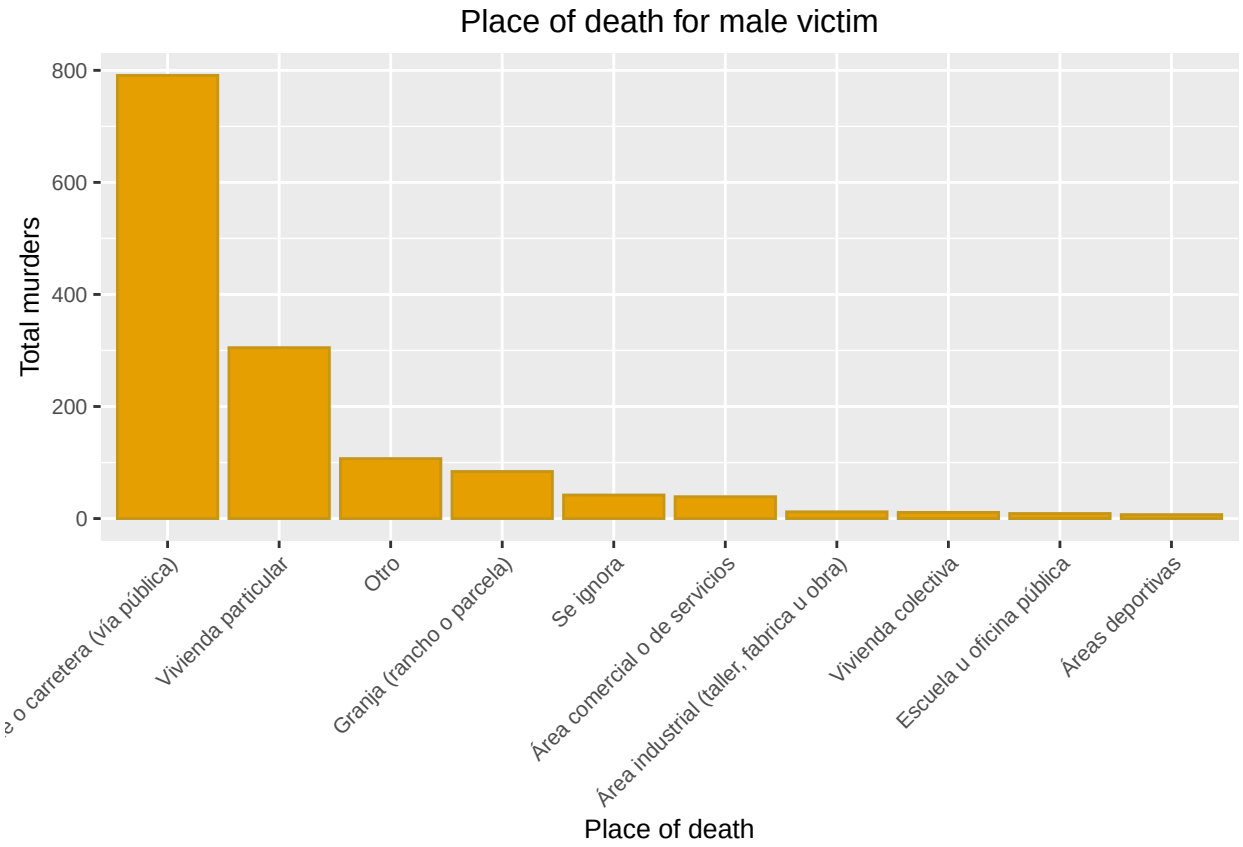
```
## rf variable importance
##
##   only 20 most important variables shown (out of 220)
##
##                                                 Overall
## par_agreConcubino,compañero                     100.000
## par_agreEsposo,Cónyuge                           91.534
## sexo_agresorNo especificado                      71.877
## lugar_ocurCalle o carretera (vía pública)        51.234
## lugar_ocurVivienda particular                    50.467
## causa_defAhorcamiento, estrangulamiento y sofocación  41.792
## par_agreSinparentesco                            38.235
## vio_famiHubo violencia no familiar               34.162
## causa_defDisparo de otras armas de fuego         28.433
## vio_famiNo especificado                          16.091
## edo_civilViudo (a)                               15.632
## par_agreExesposo                                 14.662
## nom_locCiudad de México                          12.672
## causa_defAgresión con objeto cortante            11.541
## edo_civilUnión libre                             10.238
## anio_ocur2016                                    10.198
## anio_ocur2015                                     9.308
## nom_locMéxico                                     9.222
## anio_ocur2017                                     8.723
## causa_defMedios no especificados                  8.623
```

Once again, observe that the main differences found in the graphs of exploratory data section, are the main variables used in the models.

It can be noticed that a variable that hadn't been explored is part of the final model, the place where the murder occurred.

The following plots show the data separated per gender.


Place of death for female victim
Total murders
200
100
0
Vivienda particular
Calle o carretera (vía pública)
Otro
Se ignora
Granja (rancho o parcela)
Área comercial o de servicios
Escuela u oficina pública
Área industrial (taller, fabrica u obra)
Áreas deportivas
Vivienda colectiva
Place of death



There is a remarkable difference in the place where the murders happen. Women are more likely to be killed in their own home (Vivienda particular), rather than men, that die more frequently on the street (Vía pública).

4 Results

The final result chart presented:

Table 11: Model used and result

Method	Accuracy
All Men	0.7750000
Decision Tree	0.8583333
Decision Tree 1SE	0.8583333
KNN	0.8361111
GLM	0.8416667
Random Forest	0.8611111

5 Conclusion

The caret package can handle different Machine Learning methods, one of the many advantages is that the output will be consistent across all of them. In this way, it can be used to test various and then pick the best without knowing which one is it, just coding. It has also improved previous methods included in the

same package.

Limitations are also presented by the data itself. Almost 1% from all the data has the complete information to correctly classify it and use it in the analysis.

It has been proved that there are significant differences on how men and women are killed. While the number of killed men is higher than killed women, the circumstances are different. Each case has to be appointed with their pertinent solutions. While some need attention of public defense and systematic intervention to insecurity, others need different kind of solutions.

5.1 Future work

The train set was built separating the data obtained from last 4 years. Once the 2019 results are available, the models can be tested using the data obtained in the present work as the train set and, after performing data cleaning, use the coming years as test set.

6 References

[1] https://rafalab.github.io/dsbook/

[2] https://topepo.github.io/caret/train-models-by-tag.html#tree-based-model

[3] https://www.inegi.org.mx/temas/mortalidad/

[4] https://www.theguardian.com/world/2018/jul/23/mexico-crime-homicides-violence-up-report

[5] https://www.csis.org/analysis/femicides-mexico-impunity-and-protests

[6] https://www.eluniversal.com.mx/english/what-femicide