Llamados Violencia Familiar

Año 2017/18/19

Martin Bellini, Joaquín Espona y Tamara Bertomeu
Universidad Tecnológica Nacional, FRBA
CABA, Buenos Aires, Argentina

Abstract
Este artículo va a tener como objetivo analizar datos obtenidos de las llamadas al #147 de violencia familiar, para poder encontrar patrones de comportamiento y utilizarlos para para futuras mejoras en el sistema Estatal.
Keywords
Violencia Familiar, Estado Nacional,

1 INTRODUCION

El objetivo del siguiente Paper es poder analizar los datos de las llamadas por violencia familiar en CABA. En base a estos datos se entenderán los patrones y comportamiento de los actores dentro del Set de Datos. A continuación

2 PREPARACION DATA SET

En cuanto al Data Set escogido para llevar a cabo el entrenamiento del algoritmo se analizaron los tipos de datos, features y cantidad de los mismos para asegurarnos resultados válidos.

El mismo contiene los llamados del año 2017, 2018 y hasta el tercer trimestre del año 2019.

El Data set completo posee originalmente 16 Features y 23.420 Líneas a analizar.

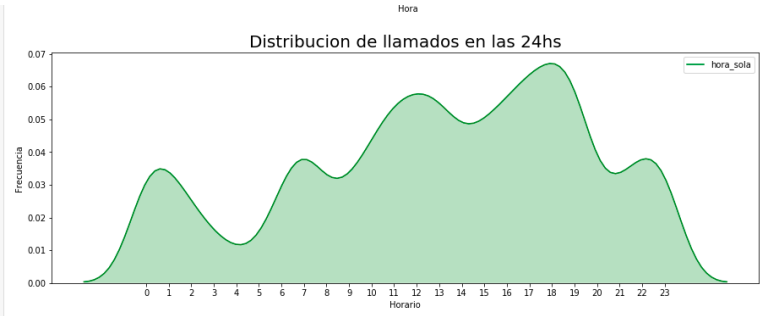
Los primeros pasos fueron concatenar los datos de los años que se habían colectado, limpiar de valores nulos, estandarizar el contenido y por último se analizó el contenido del Data Set mediante distintos modelos de regresión y clasificación.

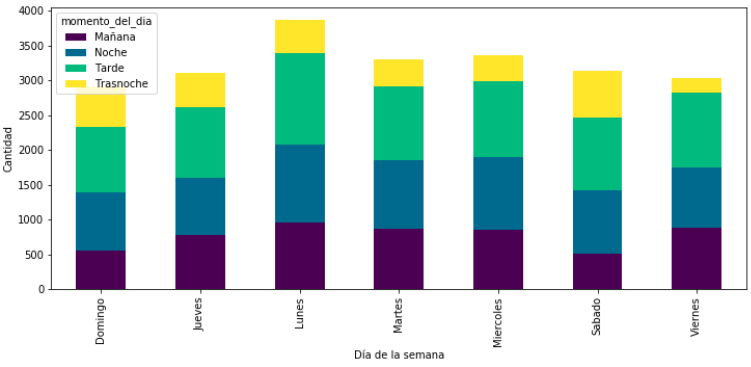
3 EXPLORACION Y ANALISIS DE DATOS

Para llevar a cabo el análisis de datos se analizó y estudio en detalle cada Feature presente en el data set. Adicionalmente, Las principales features de interés fueron analizadas mediante gráficos para comprender la situación de forma visual.

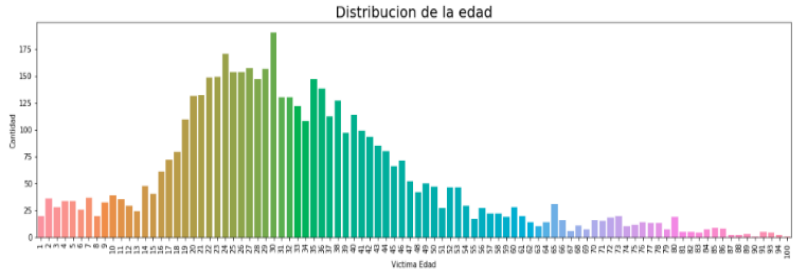
A continuación, se listan dichas Features con sus respectivos gráficos pre limpieza del data set:

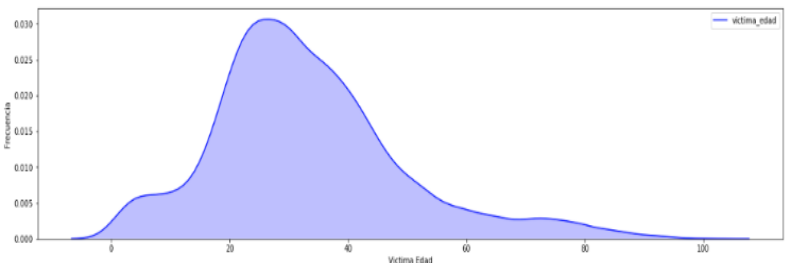
- *Tiempo y Momento del día.*
  Esta Feature se refiere en que momento del día sucede el llamado. Para esto se normaliza el Data Set teniendo en cuenta rangos de tiempos para clasificar los momentos en el día.





- *Edad*
  Esta Feature se refiere la edad de las victimas por las cuales se hizo el llamado.





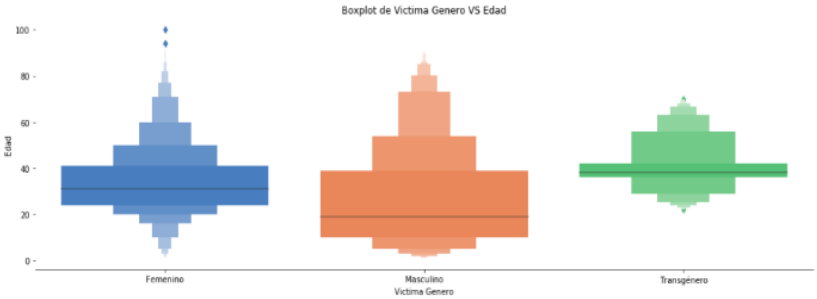
En base a estos datos pudimos entender que el Data Set necesitaba ser normalizado para seguir adelante.

4 NORMALIZACION

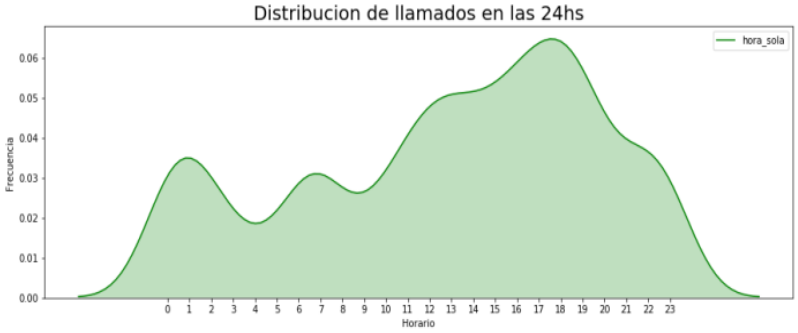
En el análisis anterior se entendió que había datos en las Features que se referían a los mismos casos, pero era ligeramente diferente entre sí en cuanto a la forma en que estaban escritos.

Una vez ya normalizado el Data Set se ve de la siguiente forma:

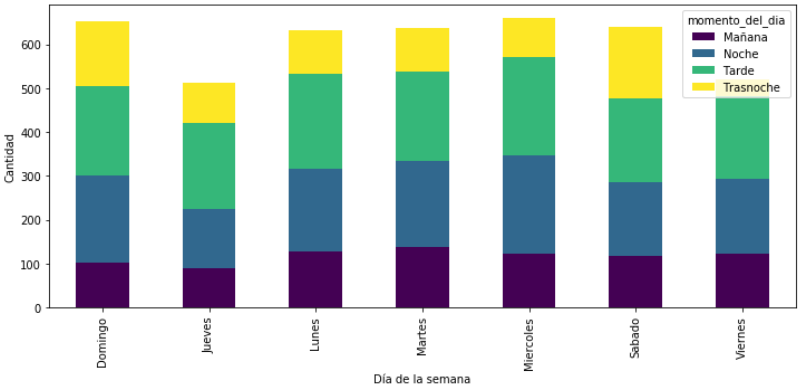
a) *Relación Edad de la víctima – Genero de la victima*



b) *Frecuencia de llamados:*

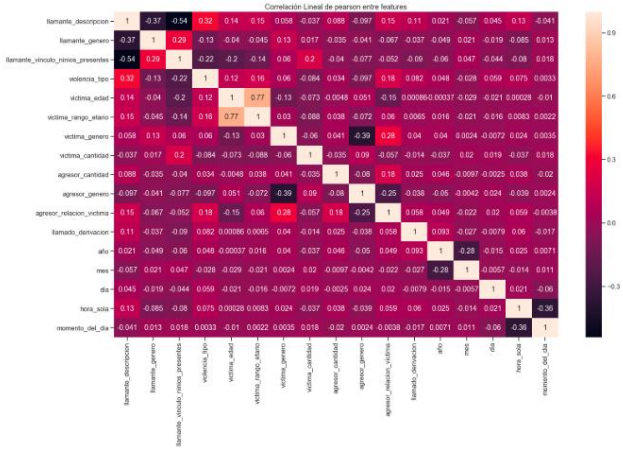


c) *Distribucion según dia de la semana*



5 CORRELACION ENTRE FEATURES

Con el Data set ya limpio se realiza un Heat Map para poder determinar si existe algún tipo de correlación lineal entre las Features:



Donde se obtuvo como conclusión de que existe una relación inversamente proporcional entre los géneros de la víctima y genero del agresor con un valor de -0.39.

En base a todo lo hecho hasta el momento, procedemos a aplicar los distintos métodos de Feature Selection para utilizar en los modelos de regresión con el fin de poder estimar la edad de la víctima. Los métodos utilizados son los Threshold y Lasso. Las features seleccionadas fueron las siguientes:

```
Features de THRESHOLD [8]
Index(['llamante_descripcion', 'llamante_vinculo_ninios_presentes',
       'victima_rango_etario', 'agresor_relacion_victima',
       'llamado_derivacion', 'mes', 'día', 'hora_sola'],
      dtype='object')

Features de LASSO [12]
Index(['llamante_descripcion', 'llamante_genero',
       'llamante_vinculo_ninios_presentes', 'victima_rango_etario',
       'victima_genero', 'agresor_cantidad', 'agresor_genero',
       'agresor_relacion_victima', 'año', 'mes', 'día', 'momento_del_dia'],
      dtype='object')
```

6 REGRESION

En este caso se llevarán a cabo los modelos de regresión con distintos tipos de métodos de selección de Features o sin Feature Selection.

En base a esto se llevaron a cabo distintas iteraciones por los diferentes métodos:

-Linear Regression [LR]

-KNN Regression [KNN]

-Support Vector Regression Linear [SVR-L]

-Support Vector Regression Gaussiano [SVR-G]

Para medir la performance del modelo, se midieron distintos errores:

SIN FEATURE SELECTION [SF]

TABLA COMPARATIVA SIN FEATURE SELECTION

	Model	MAE	MSE	RMSE
0	LR-SF	5.544114	96.565527	9.826776
1	KNN-SF	7.201988	116.512367	10.794089
2	SVR-L-SF	4.938397	117.785760	10.852915
3	SVR-G-SF	5.214431	84.497585	9.192257

Llamadas Violencia Familiar

16 de Noviembre de 2019, CABA, Buenos Aires, Argentina

CON THRESHOLD [TH]

TABLA COMPARATIVA CON THRESHOLD

	Model	MAE	MSE	RMSE
0	LR-TH	5.517641	99.126198	9.956214
1	KNN-TH	6.950942	118.486057	10.885130
2	SVR-L-TH	4.934553	121.197110	11.008956
3	SVR-G-TH	4.484192	77.481729	8.802371

CON LASSO [LA]

TABLA COMPARATIVA CON LASSO

	Model	MAE	MSE	RMSE
0	LR-LA	5.543652	96.573017	9.827157
1	KNN-LA	6.211749	96.955604	9.846604
2	SVR-L-LA	4.939136	117.689975	10.848501
3	SVR-G-LA	4.710262	78.795152	8.876663

COMPARACION COMPLETA

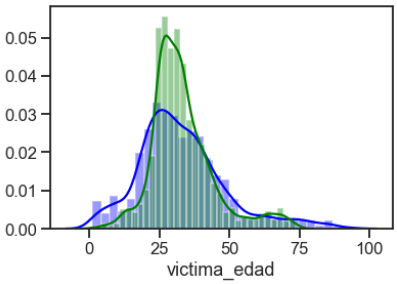
TABLA COMPARATIVA COMPLETA

	Model	MAE	MSE	RMSE
0	LR-SF	5.544114	96.565527	9.826776
1	KNN-SF	7.201988	116.512367	10.794089
2	SVR-L-SF	4.938397	117.785760	10.852915
3	SVR-G-SF	5.214431	84.497585	9.192257
4	LR-TH	5.517641	99.126198	9.956214
5	KNN-TH	6.950942	118.486057	10.885130
6	SVR-L-TH	4.934553	121.197110	11.008956
7	SVR-G-TH	4.484192	77.481729	8.802371
8	LR-LA	5.543652	96.573017	9.827157
9	KNN-LA	6.211749	96.955604	9.846604
10	SVR-L-LA	4.939136	117.689975	10.848501
11	SVR-G-LA	4.710262	78.795152	8.876663

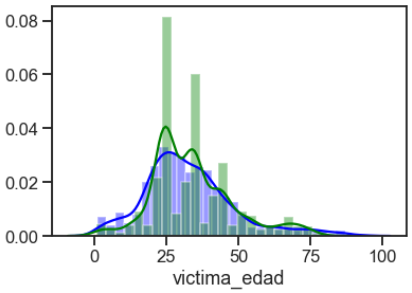
Los modelos seleccionados para avanzar con la Predicción de la edad son:

- KNN con Lasso
- SVR Gaussiano con Threshold

Se eligen estos modelos, para comparar la performance al momento de hacer la retroalimentación con datos Descartados previamente durante la limpieza del data Set.
El KNN con Lasso debido a que su histograma y curva de distribución es la más fiel al set de testing.

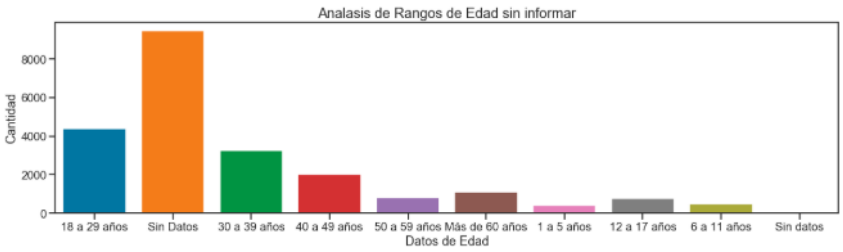


EL SVR Gaussiano con Threshold debido a que es el de Menor margen de error en todos los aspectos.



7 PREDICCION DE LOS VICTIMAS SIN EDAD

Tras entrenar los modelos y seleccionar los 2 con los cuales se desea continuar, procedimos a obtener todas aquellas samples que se descartaron previamente Debido a que estaban "Sin Datos" en la feature edad de la víctima.



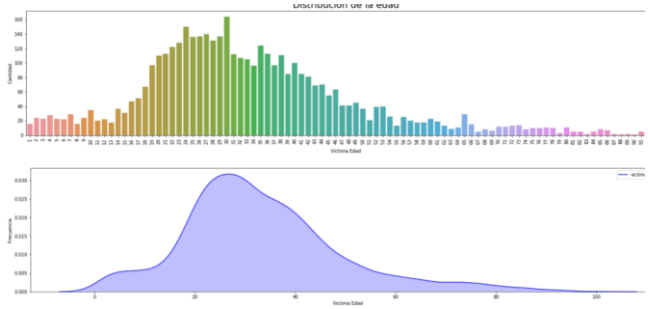
Tras aplicar la misma metodología de limpieza aplicada anteriormente al data set, obtenemos aproximadamente 1827 samples adicionales de un total de 9.600 para ser utilizadas en los modelos entrenados previamente. Estos, para poder ser aplicados a los modelos de regresión seleccionados, tuvieron que pasar por el Lasso y Threshold a modo de llegar a la cantidad de dimensiones necesarias.

Hechas las predicciones se integraron los datos obtenidos con los originales, ampliando la cantidad de samples de 4200 a 6000 [Aumento del 42%], lo cual significo que pudimos rescatar un 20% de los datos descartados inicialmente debido a que no contenían la edad de la víctima.

A continuación, se muestra los nuevos gráficos, los cuales cambiaron su forma en base a los datos predichos por los modelos de regresión KNN y SVR Gaus siano.

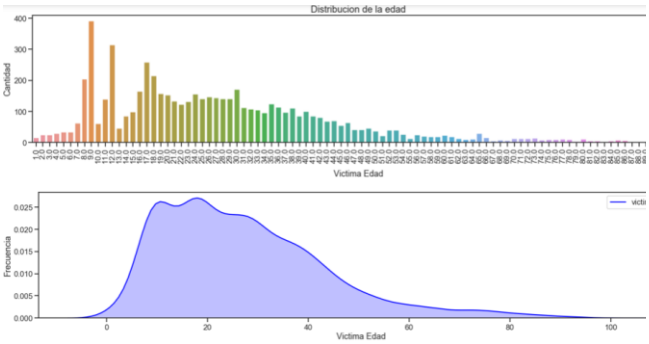
DATA SET ORIGINAL

```
count    4256.000000
mean       33.311325
std        15.811422
min         1.000000
25%        23.000000
50%        31.000000
75%        41.000000
max       100.000000
Name: victima_edad, dtype: float64
```



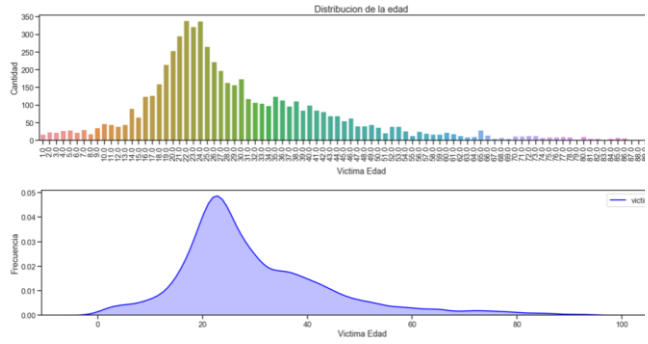
SVR-GAUSSIANO CON THRESHOLD

```
count    6083.000000
mean       27.235081
std        16.338202
min         1.000000
25%        15.000000
50%        24.000000
75%        36.000000
max       100.000000
Name: victima_edad, dtype: float64
```



KNN CON LASSO

```
count    6083.000000
mean       29.620746
std        14.570278
min         1.000000
25%        21.000000
50%        26.000000
75%        36.000000
max       100.000000
Name: victima_edad, dtype: float64
```



8 CONCLUSIONES DE LA REGRESION

Tras haber aplicado los modelos de regresión, hecho un reaprovechamiento de la información descartada y visualizado sus respectivos resultados.
Se llegó a las siguientes conclusiones:

a) *Que un modelo tenga menor error no garantiza que tenga la mejor performance.*
b) *Algunos modelos tendrán mayor dificultad par a imitar la naturaleza de los datos, es decir, de la situación que se está analizando. Esto se puede deber a la matemática que hay detrás d e cada modelo de regresión.*
c) *Es importante hacer una comparativa grafica e ntre los valores predichos por un modelo vs los valores de testing a la hora de entrenar un modelo para ver la distribución de los datos.*

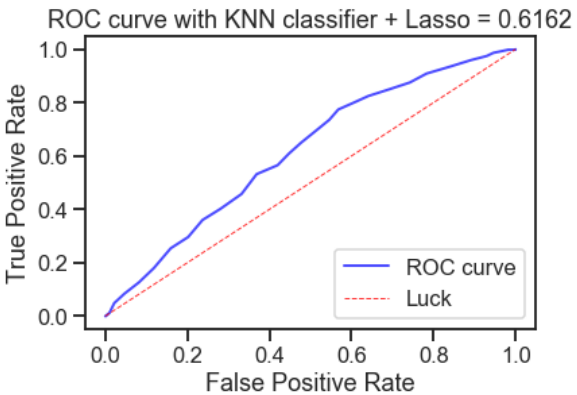
9 PREDICCION DE COMO FINALIZA UN LLAMADO

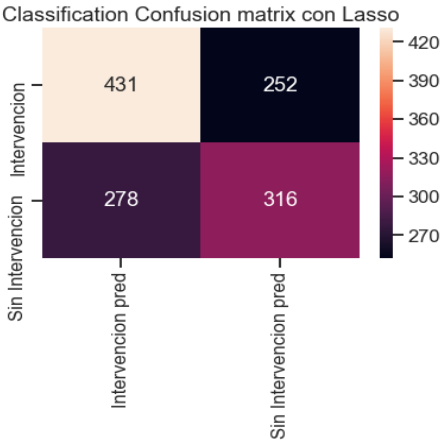
Por otro lado, se trabajó sobre un modelo de clasificación que tenía como finalidad predecir como terminaría un llamado al #147. Se puede ver en que deriva un llamado al 147 en la feature "Llamado_derivacion".
Debido a que el mismo consta de múltiples finales posibles, se optó tras probar una clasificación multi etiquetas que daba un accuracy del 40%, reducirlo a uno binario donde los finales pueden ser "Intervención" o "Sin Intervención".

Para la limpieza del data set, se aplicó la misma lógica aplicada anteriormente. Adicionalmente, se utilizó como método de selección de Features a Lasso y como modelo de clasificación a KNN y a SVM Lineal.

RESULTADOS KNN CON LASSO:

```
El Accuracy Test con Lasso es 0.5849647611589663
```

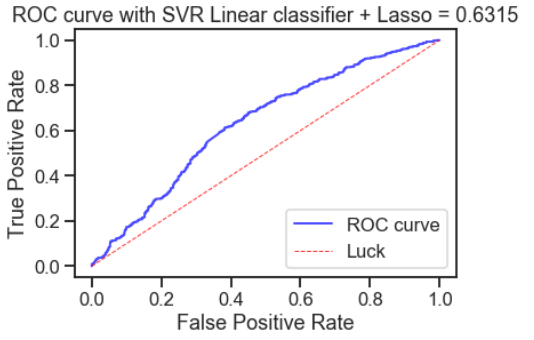


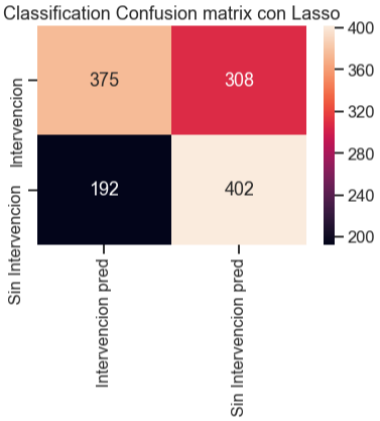


Llamadas Violencia Familiar
16 de Noviembre de 2019, CABA, Buenos Aires, Argentina

RESULTADOS SVM LINEAL CON LASSO:

```
El Accuracy Test sin Feature Selection es 0.6084573218480
```





10 CONCLUSIONES DE LA CLASIFICACION

Tras ver los resultados de la clasificación, sus respectivos accuracy y la curva AUC-ROC.

Determinamos que no es posible, al menos por el momento y bajo la metodología implementada, aplicar modelos de clasificación Debido a su gran margen de error, apenas superior al 0.5 según AUC ROC. En otras palabras, es por muy poco más confiable que arrojar una moneda al aire al momento de clasificar en que derivo el llamado al #147.

11 CONCLUSIONES DEL PROYECTO

Tras haber recorrido todo este camino, hemos podido aprender y llegar a las siguientes conclusiones sobre los llamados por violencia familiar del #147:

a) *Gran parte de los llamados ocurren durante la tarde/noche, especialmente en el intervalo de 12hs a las 20hs.*
b) *Las edades de las víctimas se concentran principalmente en el rango de 20 a 36 años, disminuyendo hacia ambos extremos.*
c) *La mediana para las mujeres [como victimas] se encuentra entorno a los 30 años. Y en el caso de los hombres, alrededor de los 18-20 años.*
d) *Una notable cantidad de información se encontraba incompleta u no se contestaba. Esto se puede comprender debido a la naturaleza del problema.*

Llamadas Violencia Familiar

16 de Noviembre de 2019, CABA, Buenos Aires, Argentina

Finalmente, fuimos capaces hasta cierto punto aprovechar data descartada en el filtro y limpieza para nutrir al proyecto llegando a un resultado aceptable y demostrando que no siempre el modelo de menor error es la mejor opción. Ya que hay que comprender como funcionan matemáticamente los modelos y entender la naturaleza del problema que uno está enfrentando.

DATOS EXTRAS:

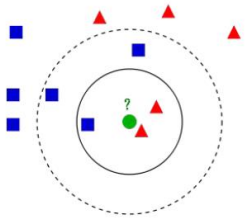
a) *1827 samples rescatados.*
b) *incremento del 42% del dataset con edades predichas vs el dataset original.*
c) *se logró re-utilizar un 20% del dataset "Sin Datos".*
d) *Pasamos del 19% utilizado al 27% del total de samples [de 4.265 a 6.083 de un total de 22.722].*

ANEXOS DE LOS MODELOS:

Los modelos utilizados funcionan de la siguiente manera:

KNN:

Es un modelo de tipo lineal que el cual utiliza una serie de "k" vecinos más cercanos para determinar el valor de una feature. Calcula la distancia del elemento nuevo a cada uno de los existentes y ordena esas distancias para Seleccionar a qué grupo al que pertenece.



$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

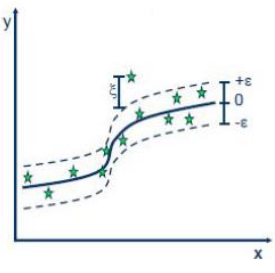
Para realizar la regresión, se determinan los "k" vecinos más cercanos por distancia euclídea (distancia par-a-par). Para ello, realiza la interpolación de los Y en los "k" vecinos (ej por promedio).
Estos últimos pesos puede ser: Uniformes (todos por igual) o por distancia.

$$d(x_a, x_b) = \sqrt{(x_{a1} - x_{b1})^2 + (x_{a2} - x_{b2})^2 + \ldots + (x_{ap} - x_{bp})^2}$$

Los pesos y el número de vecinos "k" son híper parámetros definidos por el usuario.

SVR:

Es otro método utilizado para realizar regresiones, es un modelo de tipo lineal aunque la utilización de kernels permite trabajar sobre valores que originalmente no son linealmente separables.
Buscar el híper plano que maximice el margen.
Determina el margen/radio (épsilon) como función de costo y trata de que todas las muestras estén dentro del Margen (o "tubo").



$$C\sum_{n=1}^{N}\xi_n + 1/2\,\|w\|^2$$