Máximo Ripani
Juan Pedro Pompei
Agustín Silvio Andrés Rojas
Ignacio Colmeiro

Clasificación binaria de ataques cardíacos

Regresión Logística

Introducción

En el siguiente informe se realizará una regresión logística para predecir la probabilidad de un paciente clínico de sufrir un ataque cardíaco. Para realizar el siguiente, se utilizó un dataset de 462 observaciones, con 10 variables distintas. Los datos utilizados son de hombres de un distrito de la región de Western Cape, Sudáfrica, el cual tiene una alta tasa de muertes debidas a los ataques cardíacos. El fin de este, es poder ayudar a prevenir estos ataques cardíacos y poder diagnosticar los tratamientos acordes para poder disminuir la cantidad de ataques cardíacos que ocurren. Para ello se buscará encontrar los factores más influyentes en la enfermedad. Asimismo encontrar patrones que los distingan, a través de un análisis estadístico descriptivo e inferencial.

Para la realización de la Regresión Logística se utilizaran 10 variables explicativas distintas para poder predecir si el paciente obtuvo o no un ataque cardíaco, las cuales son: presión arterial, consumo acumulativo de tabaco (kg), colesterol, adiposidad, historial familiar, personalidad (en este caso tipo A), obesidad, consumo de alcohol, edad.

En el mismo, se realizó un análisis de Estadística Descriptiva para poder analizar la relación entre las variables explicativas y la explicada, así como también ver la distribución de estas mismas, buscando así mostrar diferencias explicativas comparando distribuciones entre aquellos que tienen o no la enfermedad. Al terminar de analizar las distintas relaciones, se continuó con la etapa de clasificación pertinente al estudio. Durante esta etapa se eligió un valor de corte para poder clasificar binariamente nuestra regresión. Es decir, para poder predecir si el paciente va a sufrir la enfermedad o no, en base a la probabilidad encontrada previamente.

Por otro lado se estudió de forma inferencial, con una regresión logística la relación entre la variable explicada y aquellas explicativas. De la misma manera se utilizo un modelo logístico para clasificar la pertenencia a cada una de las posibles clases de la variable explicada. Se evaluó el modelo con las métricas especificidad, sensibilidad y F1 ponderando aquella de mayor importancia para la razón de estudio del presente trabajo.

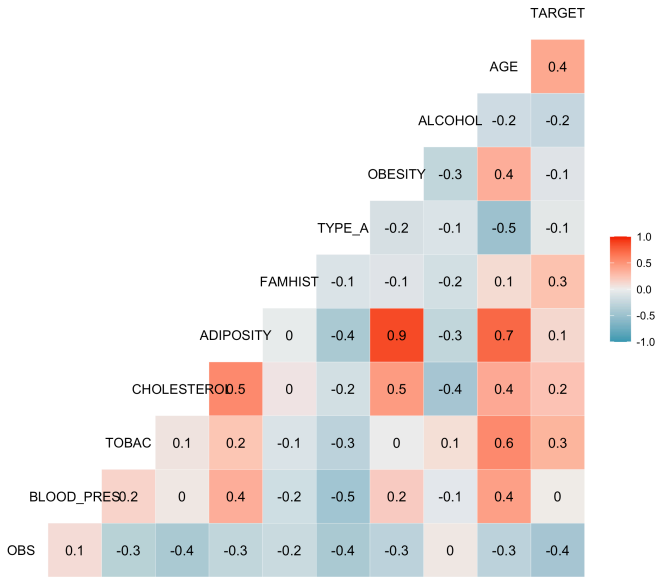
Análisis exploratorio

Como primer acercamiento al análisis, una matriz de correlaciones entre la variables explicativas:

$$Cov_{x,y} = E[x,y] - E[x]E[y]$$

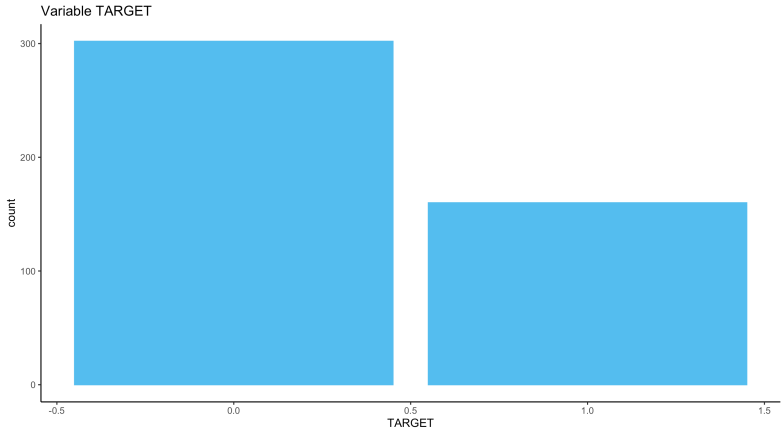
$$\rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \sigma_y} \qquad (1)$$

La función (1) representa la relación reciproca entre dos variables o fenómenos. En este caso, indicaría el grado de relación que tienen tanto las variables dependientes como la independiente entre sí (recíprocamente). Representando la matriz de correlaciones gráficamente:

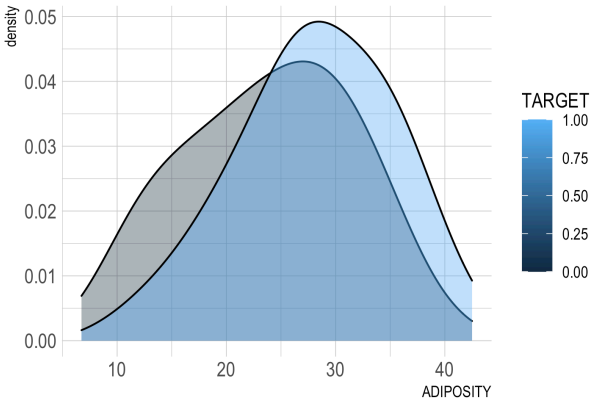


Es posible apreciar correlaciones altas entre algunas variables, por ejemplo aquella correlación entre *obesidad* y *adiposidad*, la cual tiene mucho sentido. Por otro lado, con respecto a la variable explicada sus correlaciones más altas son con: *historia familiar, edad, tabaco, alcohol* entre otras.

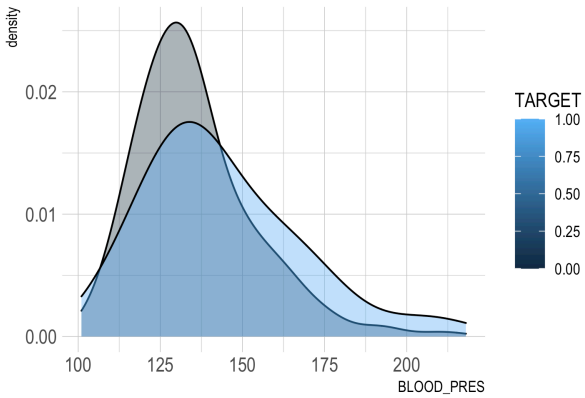
Las clases de la variable dependiente están desbalanceadas. Esto podría generar sesgos en el modelo que perjudiquen la etapa de clasificación. Existen técnicas de sobremuestreo (oversampling) que podrían ayudar a mejorar el modelo sin embargo no son pertinentes para el objeto de este estudio y serán consideradas en futuras investigaciones.



A continuación se presentan la función de distribuciones de densidad de las variables: adiposidad, presión sanguínea, y cantidad de tabaco acumulado respectivamente.



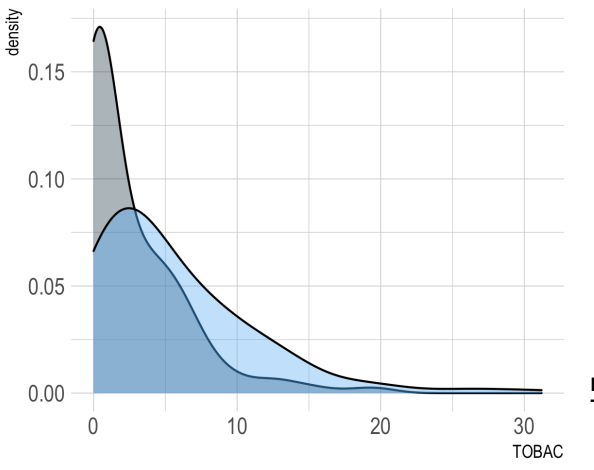
DISTRIBUCION ADIPOSIDAD



DISTRIBUCIÓN PRESIÓN SANGUINEA

En cada uno de los casos se muestra una diferencia significativa entre la observaciones pertenecientes a la clase positiva y la clase negativa. Esto puede dar indicios acerca de la influencia de estas variables sobre aquella explicada.

Con respecto al tabaco podemos notar que las muestras pertenecientes a la clase positiva tienen una distribución más densa en casos con alto índice de tabaquismo, a su vez que la clase negativa presenta una gran densidad en valores nulos de tabaco. En la morfología de la distribución de la variable adiposidad no hay gran diferencia, sin embargo es notorio el desplazamiento

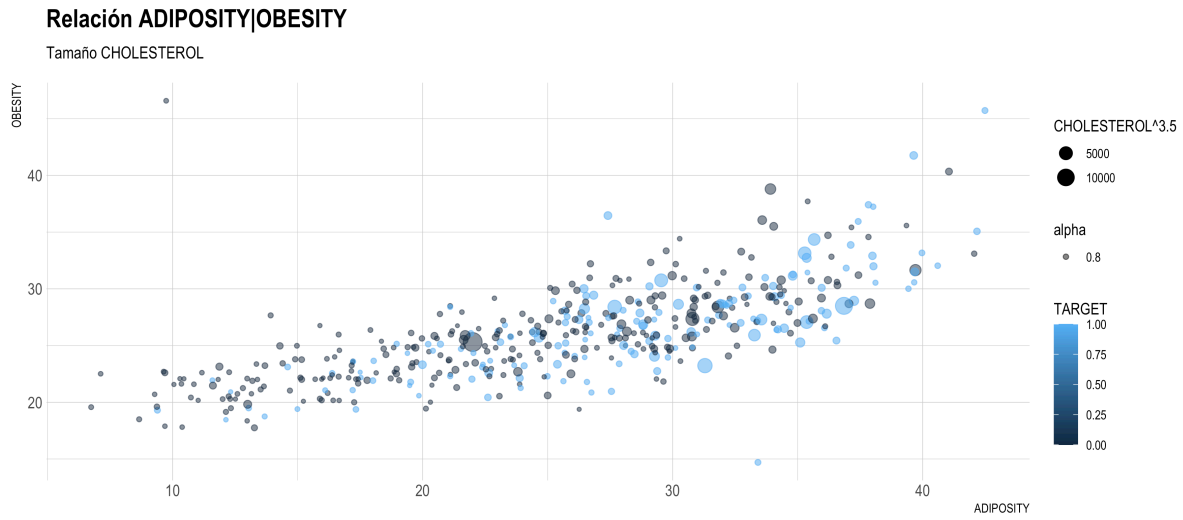


DISTRIBUCION TABACO

hacia valores más altos de la variable. La distribución de la presión sanguínea no sufre mayores desplazamientos entre las clases, sin embargo es notoria una brecha entre las clases que representa una mayor presión sanguínea para las muestras de los casos positivos. Se encuentran ulteriores representaciones gráficas en el anexo.

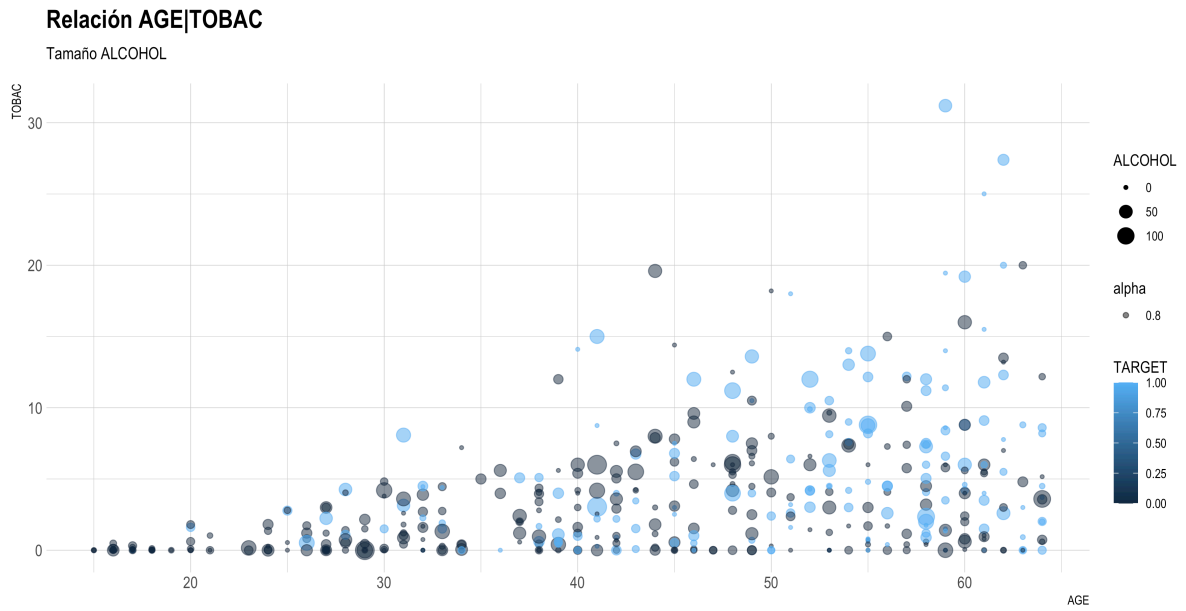
La obesidad, adiposidad, colesterol, son variables que por lógica cotidiana resultan relacionadas. El lector podría pensar que aquellos que tener niveles altos de esta variables serían indicadores de un peor estado de salud. Resulta aún más evidente cuando es representado gráficamente:

Es posible notar una relación positiva entre las variables obesidad y adiposidad. A su vez la variable colesterol es representada a través del tamaño de los puntos. De esta manera se puede ver la interacción entre las tres variables en simultáneo y en conjunto con la variable target. Aquellos puntos de color celeste representan los casos positivos, los cuales es posible apreciar una mayor concentración en puntos de mayor tamaño y más cercanos a la esquina superior derecha del gráfico.



Considerando la lógica cotidiana anteriormente mencionada, resulta también evidente una relación entre el tabaco y el alcohol. Históricamente este duo ha presentado un impacto negativo en la salud de las personas. Además suelen tener un comportamiento similar dado que muchas veces sus consumidores son los mismos.

En este caso se nota el incremento del tabaco acumulado a lo largo de los años, el tamaño de los puntos es representativo del valor de la variable alcohol. Se aprecia, una vez más, una mayor concentración de los puntos celestes (positivos) en el extremo superior derecho.



Regresión Logística

Dado que el problema del estudio es una clasificación del tipo binaria, la recta de regresión lineal no representa de la mejor manera posible la división entre las clases positiva y negativa. Es decir existen determinadas observaciones que escapan de aquellas relaciones que la regresión lineal puede explicar. En estos casos, donde la variable explicada se distribuye de forma binomial, es posible lograr una mejor aproximación al problema a través de un modelo de **regresión logística**. Ambos son métodos lineales y guardan similitudes. El modelo logístico se caracteriza por la implementación de la función logística.

Dada la variable dependiente:

$$Y_i \sim Binomial(n_i, p_i)$$

$$p_i = E(\frac{Y_i}{n_i} | X_i)$$

Donde X_i representa el vector de n dimensiones de las variables explicativas. A dichas probabilidades, se les aplica una transformación logarítmica de la siguiente forma denominándose dada su transformación, *logit*:

$$logit(p_i) = ln(\frac{p_i}{1-p_i})$$

$$= \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_n X_{n,i}$$

En este punto dada la transformación logarítmica no es posible ajustar la recta utilizando el método de mínimos cuadrados ordinarios. En cambio se utiliza el *método de máxima verosimilitud*. Este proceso se realiza para varias rectas hasta encontrar aquella de *máxima verosimilitud.*

Finalmente se realiza otra transformación, esta vez con la *función logística*. La misma transforma el input en un valor entre 0 y 1, es decir una probabilidad.

$$p_i = \frac{1}{1 - e^{-(logit(p_i))}} \qquad (2)$$

A travez de la función logística (2) se obtiene la predicción final de la probabilidad de dicho ensayo binomial.

Primer acercamiento: regresión logística

Al realizar la primera regresión logística, la Prueba de Wald, permite comprobar que hay relaciones entre las variables explicativas. Esta prueba se puede utilizar para decidir que factores o variables independientes son importantes para describir la probabilidad de ocurrencia de nuestro suceso de interés. Pues se va a establecer bajo hipótesis nula que el β_i (siendo i las distintas variables explicativas) es igual a cero. Si el *p-valor* es inferior a 0.05 se rechaza esta hipótesis nula con un nivel de significancia: $\alpha = 5\,\%$.

Es posible observar que hay ciertas variables que tienen un *p-valor* muy pequeño, por lo que para aquellas variables se puede inferir tienen relación con la probabilidad de sufrir un ataque cardiaco.

Dichas variables explicativas y significantes, son las siguientes: consumo de tabaco, colesterol, historial familiar, su tipo de personalidad (type-A) y la edad. En la siguiente se muestran los resultados del primer análisis de regresión:

	Estimate	Std. Error	Z value	Pr(>|z|)
BLOOD_PRES	0.0065040	0.0057304	1.135	0.256374
TOBAC	0.0793764	0.0266028	2.984	0.002847
CHOLESTEROL	0.1739239	0.0596617	2.915	0.003555
ADIPOSITY	0.0185866	0.0292894	635	0.525700
FAMHIST	0.9253704	0.2278940	4.061	4.90e-05
TYPE_A	0.0395950	0.0123202	3.214	0.001310
OBESITY	-0.0629099	0.0442477	-1.422	0.155095
ALCOHOL	0.0001217	0.0044832	27	0.978350
AGE	0.0452253	0.0121298	3.728	0.000193

En la última columna se observa que para estas variables comentadas previamente el *p-valor* está realmente por debajo del valor que se estableció como el nivel de significancia.

Clasificación

Esta etapa del estudio es dedicada a las clasificación binaria. Se buscó conseguir el mejor modelo que distinga si una observación pertenece o no a la clase positiva. Aquellas observaciones que arrojen una probabilidad alta serán más tendientes a ser consideradas pertenecientes al grupo en cuestión. Para ello se elige un valor de corte a partir del cual se pertenece o no a cierto grupo. Es decir si $p_i > \theta$ la observación i -esima pertenece al grupo positivo, de lo contrario al negativo. Sin embargo como se verá más adelante el *valor de corte* de dichas probabilidades será objeto de estudio y determinante en el mismo.

Para la clasificación los datos fueron estandarizados, es decir que se transformaron, manteniendo sus relaciones, para que tengan $\mu = 0$ y $\sigma = 1$. Esta práctica favorece al correcto funcionamiento del modelo logístico y muchas veces aumenta su performance. Luego de la estandarización, los datos fueron divididos en dos sets distintos: unos de *entrenamiento*, otro de *testeo.* De esta forma sería posible evaluar el modelo con datos que no haya visto nunca y comprobar si efectivamente es capas de generalizar. Si la performance del modelo empeora mucho en los datos de testeo, implica que el modelo se esta ajustando demasiado a los datos de entrenamiento y no sería capas de generalizar correctamente.

Para la evaluación del modelo se utiliza una matriz de confusión. La misma nos permite ver las predicciones de la forma: *verdaderos-positivos, falsos-positivos, verdaderos-negativos, falsos-negativos.* Una vez que se cuenta con la matriz de confusión es posible calcular a su vez la especificidad, sensibilidad y la métrica F1 de nuestro modelo. Estas últimas tres, son de especial interés ya que nos darán información acerca de distintos puntos de vista del modelo.

La *especificidad* indica que porcentaje de aquellos valores predichos como positivos, realmente lo eran:

$$especificidad = \frac{VP}{VP + FP}$$

La *sensibilidad* indica que porcentaje de aquellos que realmente era casos positivos, se lograron clasificar como tales:

$$sensibilidad = \frac{VP}{VP + FN}$$

Finalmente la métrica *F1* representa un promedio de las dos anteriores, por lo que puede ser utilizada para una evaluación más integral del modelo.

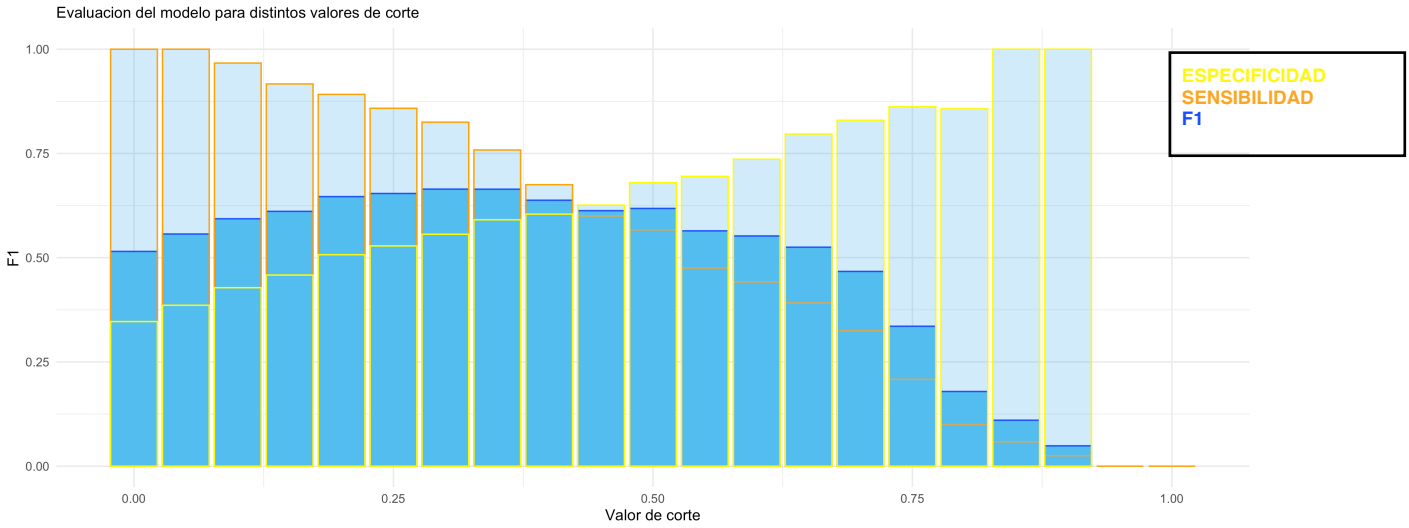
$$F1 = \frac{2 \cdot especificidad \cdot sensibilidad}{especificidad + sensibilidad}$$

Para el caso del presente estudio la métrica de principal relevancia es la sensibilidad, dado que al tratarse de ataques cardiacos los *falsos negativos* pueden ser muy costosos (sin ir muy lejos, pueden costar una vida) por lo que al evaluar el modelo es importante ponerle especial atención.

El primer modelo arrojó dio los siguientes resultados en la evaluación sobre datos de testeo:

$F1 = 0.55; \quad especificidad = 0.55; \quad sensibilidad = 0.55;$

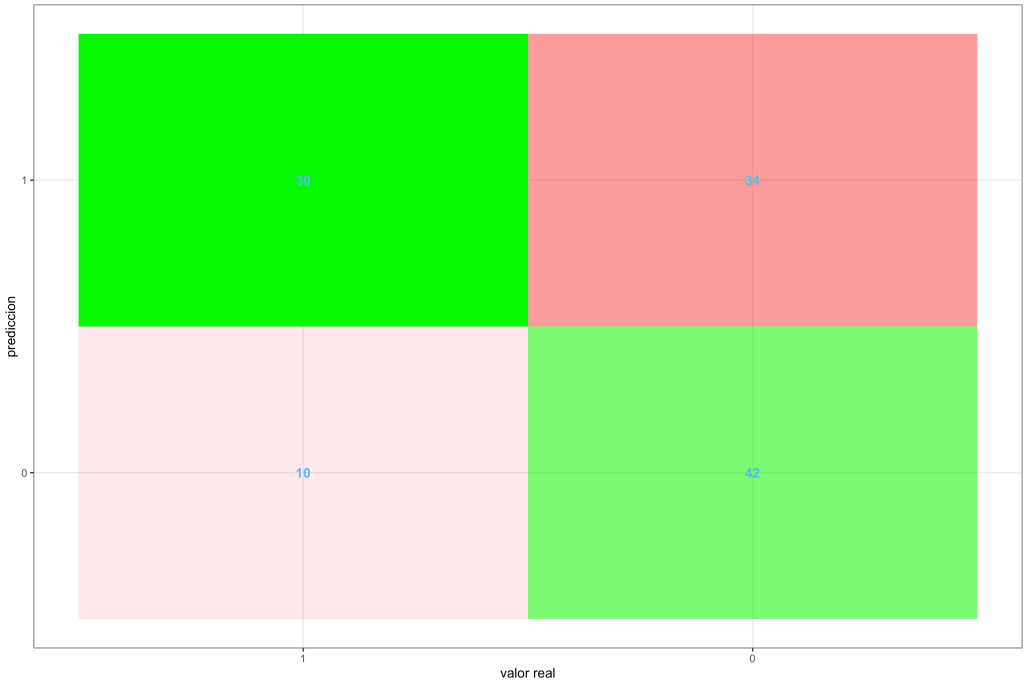
Si bien los resultados superan lo que sería un modelo totalmente azaroso, por ejemplo tirar una moneda para decidir, es posible mejorar el mismo. Para ellos es posible modificar el valor de corte para optimizar la métricas de nuestro modelo. Por default se utilizó un valor de corte $\theta = 0.5$ es decir que si la probabilidad es mayor a 0.5, dicha observación se la considera positiva. Calculando la *especificidad*, *sensibilidad* y el *valor F1* para distintos valores de corte θ, podemos encontrar el θ óptimo que maximice las métricas de nuestro modelo.



Vemos en este caso el valor de la sensibilidad, especificidad y F1 para los distintos valores de θ. Finalmente se seleccionó el valor de corte $\theta = 0.3$. Por últimos se realizo una nueva evaluación del modelo sobre los datos de testeo pero esta vez con el nuevo valor de corte, los resultados muestran mejoras en el modelo:

$F1 = 0.66; \quad especificidad = 0.556; \quad sensibilidad = 0.825;$

La métrica de interés, *especificidad* tuvo una mejora notable con el nuevo valor de corte $\theta = 0.3$. Finalmente se presenta la matriz de confusión del último modelo:



Conclusiones y futuras investigaciones

Se puede concluir que este modelo es realmente útil, ya que gracias a las medidas de especificidad, sensibilidad y en consecuencia, F1, son superiores a estas métricas obtenidas si este modelo hubiese sido totalmente azaroso. Es decir, si para este modelo se tiraría una moneda al aire para determinar que ocurre con el paciente en cuestión. Por otro lado más allá de la clasificación, el modelo sirve para explicar las relaciones entre las distintas variables explicativas y la variable target, pudiendo así sacar conclusiones acerca de que y como influyen las variables regresoras en la variable respuesta.

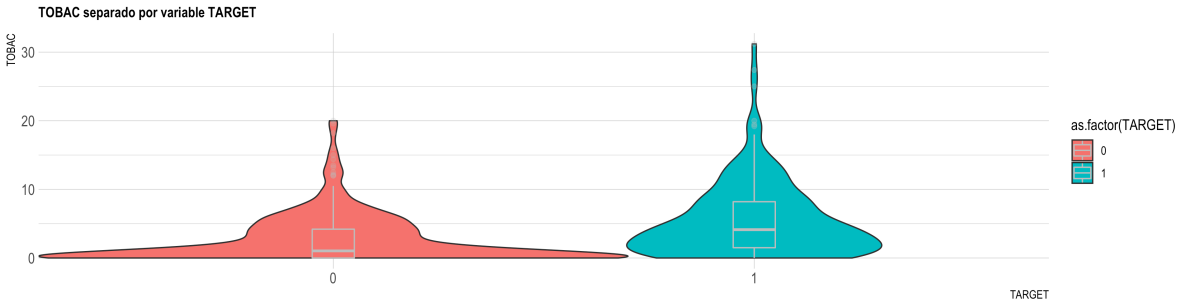
Ademas, se concluye que hay ciertas variables que tienen una relación directa, con respecto a la variable explicada. Como esta nombrado previamente las variables como el consumo de tabaco, el colesterol, el historial familiar, la edad, entre otros tienen una relevancia significante, para este conjunto de datos, con respecto a la explicación de si un paciente va a sufrir el ataque cardiaco o no.

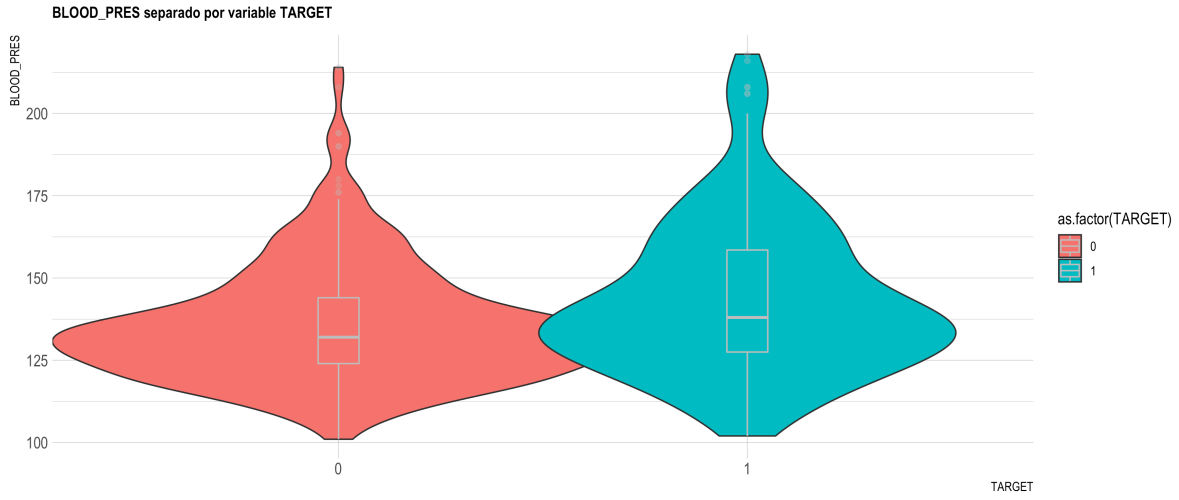
Con respecto a las futuras investigaciones que se podrían llevar a cabo, se puede destacar las distintas técnicas que se podrían utilizar para ayudar al modelo. Como se menciono previamente, el sobremuestreo (oversampling) es un buen ejemplo de como se podría mejorar el modelo. Ademas, se podría agregar más datos al data-set para que este modelo sea todavía mas certero y menos acotado, ya que en este caso de estudio, se refiere solo a hombres. Esto le abriría las puertas a una investigación todavía mas profunda en el futuro, así como también ser de mayor utilidad, si los pacientes no son solo de sexo masculino. Así como agregar mas observaciones, se podría agregar más variables explicativas que ayuden a explicar la variable explicada como por ejemplo: si el paciente sufre de diabetes, si el paciente esta falto de actividad física, consumo de drogas, antecedentes médicos, entre otros.

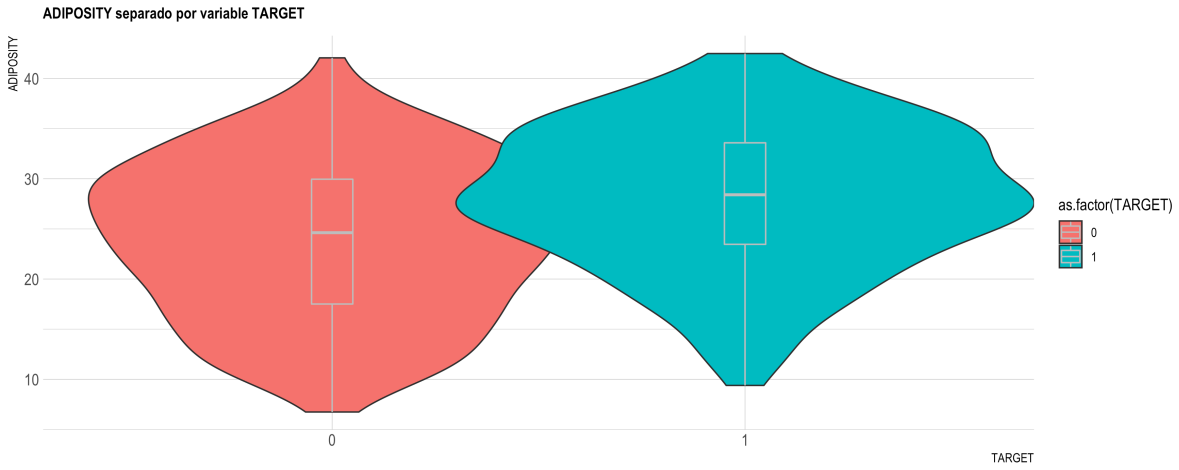
En otro aspecto de las futuras investigaciones, se podría ampliar el área de muestreo. Es decir, en este caso, los datos provienen de un solo distrito de Sudáfrica. En este caso, es realmente útil ya que se trata de un distrito con una alta tasa de mortalidad debido a ataques cardiacos. Si quisiese podría obtener datos de distintos distritos como para poder obtener ulterior información acerca de las relaciones dependiendo el distrito en el que se encuentre. Esto también serviría para comparar los efectos dependiendo en relación al área geográfica.

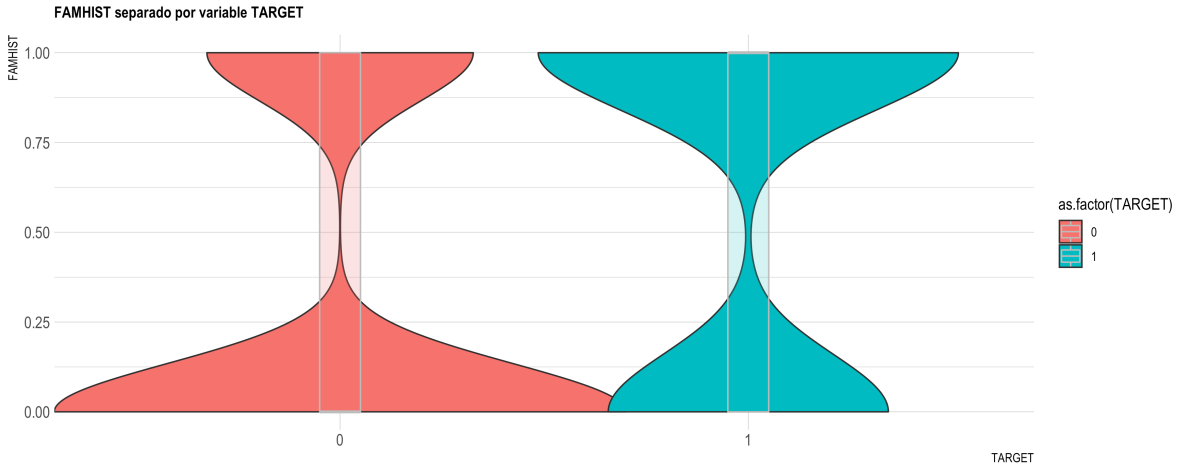
Se le podría agregar también, la comparación de este modelo con distintos modelos estadísticos para así elegir el que mejor se adecue al problema en cuestión.

Anexo






ADIPOSITY separado por variable TARGET
ADIPOSITY
40
30
20
10
0
1
TARGET
as.factor(TARGET)
0
1


FAMHIST separado por variable TARGET
FAMHIST
1.00
0.75
0.50
0.25
0.00
0
1
TARGET
as.factor(TARGET)
0
1