Tipología y ciclo de vida de los datos

Autores: Sara Belén Ramos González | Daniel Jesús Cruz Garzón

Junio 2022

INTRODUCCIÓN

Presentación

Esta práctica corresponde a la segunda de la asignatura ***Tipología y ciclo de vida de los datos.***

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.

- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

RESOLUCIÓN DE LA PRÁCTICA

En primer lugar, implementaremos las librerías que nos harán falta llevar a cabo nuestro análisis.

A continuación, resolveremos cada una de las preguntas que se formulan en el enunciado de la práctica.

---

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Elegimos analizar en nuestro estudio el dataset 'heart.cvs' del repositorio que encontramos en la siguiente ruta: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/download

Importancia del dataset y problema que pretende resolver

Se trata de un conjunto de datos donde se recoge información necesaria para poder predecir ataques del corazón a partir de un análisis de datos ya que contiene los datos referidos a personas que han sufrido un infarto y sujetos de control con parámetros sanos.

Los motivos por los cuales hemos elegido este dataset es porque conocer los factores de riesgo y los valores por los cuales podemos anticipar un posible infarto puede ayudar a algo tan importante como lo es el hecho de salvar vidas.

Además, sabemos por los datos estudiados de los últimos años que las enfermedades cardíacas causan la mayoría de las muertes en todo el mundo, tanto en el caso de hombres como de mujeres. Así que tendremos que seguir avanzando en el análisis de este tipo de información para mejorar la salud de los humanos.

Descripción Dataset

A continuación definiremos cada una de las variables que contiene nuestro dataset:

1. Age: edad en años del paciente, es de tipo numérica y tenemos un registro que varía desde los 29 a los 77 años.

2. sex: género del paciente siendo (1 = hombre, 0 = mujer), variable categórica.

3. cp: tipo de dolor torácico, siendo cuatro los valors que puede tomar este campo (variable categórica: 0 = angina típica, 1 = angina atípica, 2 = dolor no anginoso, 3 = asintomático).
4. trtbps: presión arterial en reposo (en mm Hg), variable numérica.
5. chol: colesterol en mg/dl, variable numérica.
6. fbs: (azúcar en sangre en ayunas > 120 mg/dl) (1 = si; 0 = no), variable categórica.
7. restecg : resultados electrocardiográficos en reposo (0 = normal, 1 = anomalías, 2 = hipertrofia ventricular).
8. thalach: frecuencia cardíaca máxima alcanzada, variable numérica.
9. exng: si se ha producido una angina de pecho por realizar ejercicio (1 = si, 0 = no), variable categórica.
10. olpeak: depresión del ST inducida por el ejercicio en relación con el reposo, variable numérica.
11. slp: la pendiente del segmento ST del ejercicio máximo (0: pendiente descendente; 1: plano; 2: ascendente), variable categórica.
12. caa: numero de vasos (1-3)
13. thall: trastorno de la sangre llamado talasemia (0 = Nulo; 1 = defecto fijo, es decir, no hay flujo sanguíneo en alguna parte del corazón; 2 = flujo sanguíneo normal; 3 = defecto reversible)
14. output: el atributo predicho - diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico) Sus posibles valores serán:

    0 =\<50% estrechamiento del diámetro (Ataque al corazón = No).

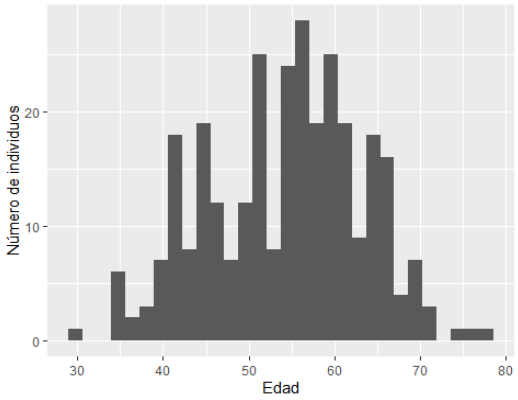
    1 =\> 50% de estrechamiento del diámetro (Ataque al corazón = Sí).

Análisis exploratorio (para descripción)

Una vez que ya hemos decidido los datos que vamos a analizar, hemos realizado la limpieza necesaria y tenemos una primera descripción de los mismos, vamos a profundizar para conocer mejor su distribución. Esto servirá incluso para arrojar conclusiones más rigurosas desde un punto de vista de un claro entendimiento de la información que estamos tratando.

Por lo tanto, iremos analizando cada uno de los campos, dibujando un gráfico de barras o histograma según convenga según el tipo de variable del que se trate.

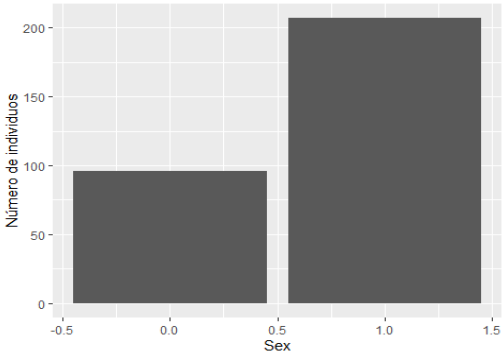
1. Campo edad.



Parece que hay una mayor concentración de la muestra en torno a los 60 años, concretamente en el intervalo de los 55 a 62 años.

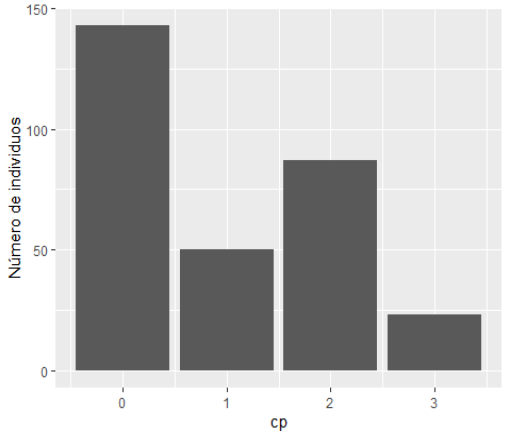
2. Campo sexo.

Ahora estudiaremos cuál es el sexo más frecuente en nuestro conjunto.



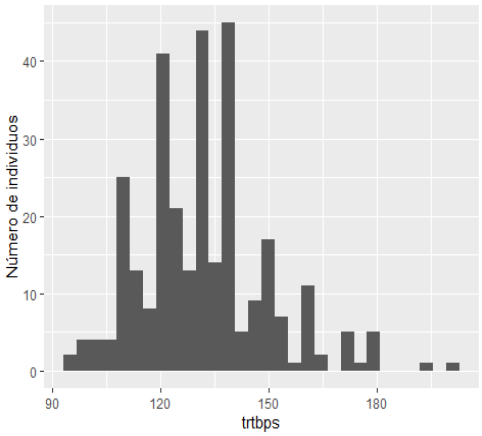
Según este gráfico vemos que con este análisis hay claramente un sesgo de los datos hacia el género masculino, ya que son el doble el número de individuos hombres que el número de mujeres que se rperesentan .

3. Campo cp.



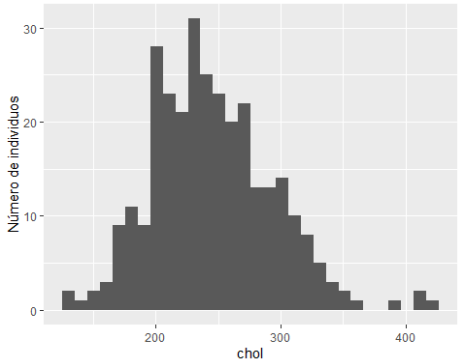
La mayoría de personas son asintomáticas, aunque también hay un número elevado de personas con angina atípica.

4. Campo trtbps.



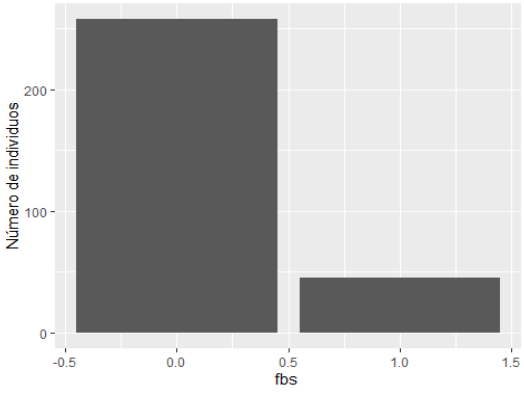
Lo más normal en nuestros pacientes estudiados es que tengan una presión arterial de entre 105 mm Hg y 140 mm Hg.

5. Campo chol.



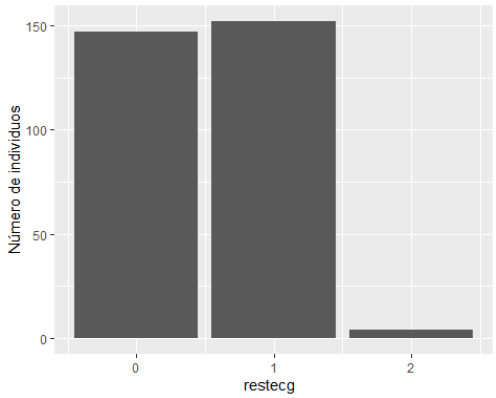
En este caso, nuestros pacientes presentan en su mayoría un colesterol de entre 210 y 280 mg/dl.

6. Campo fbs .



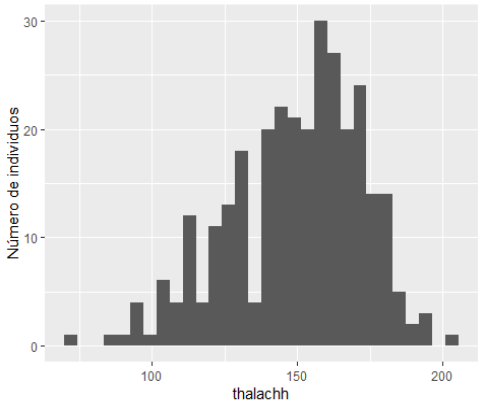
La tasa de azúcar en sangre en ayunas, en su mayoría, es <=120 mg/dl.

7. Campo restecg .



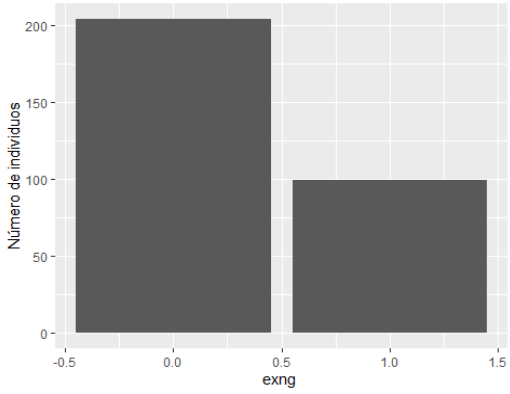
No tendremos en nuestra muestra casi ningún paciente con un electrocardiograma cuyo resultado sea having ST-T wave abnormality. Todos serán, bien normales o bien con hipertrofia.

8. Campo thalachh, .



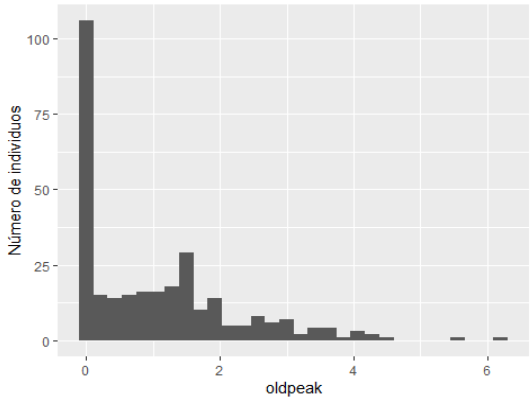
Para este campo observamos que la frecuencia cardíaca máxima alcanzada se encuentra sobre todo entre 150y 175.

9. Campo exng .



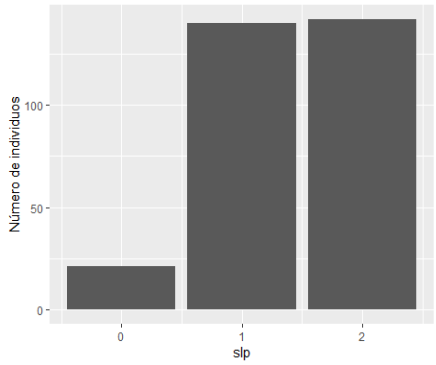
La mayoría de nuestros pacientes no ha tenido antes una angina, aunque los que si la han tenido, representan un tercio de nuestro dataset.

10. Campo oldpeak .



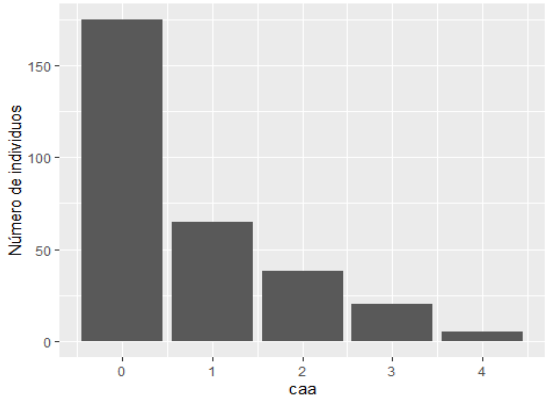
La depresión del ST inducida por el ejercicio en relación con el reposo se encuentra sobre todo en valores en torno al 0.

11. Campo slp.



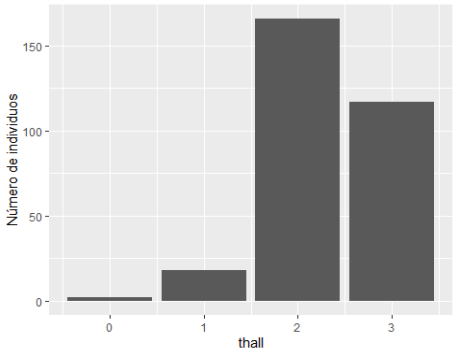
La pendiente del segmento ST del ejercicio máximo será casi siempre bien plana, o bien ascendente.

12. Campo caa.



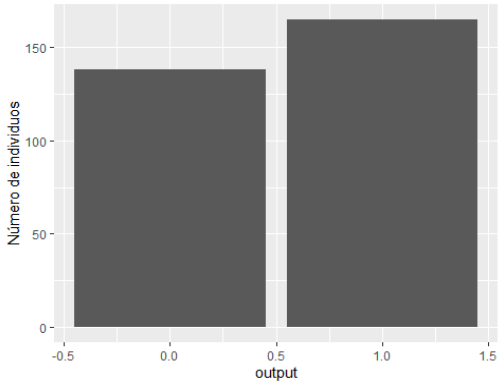
El número de vasos principales coloreados por fluoroscopia suele ser ninguno.

13. Campo thall.

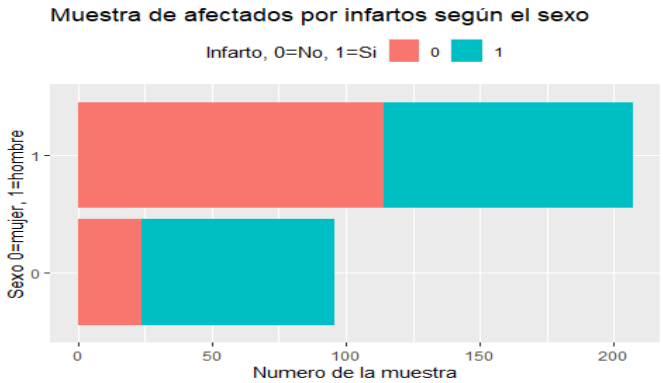


El resultado de la prueba de esfuerzo con talio muy rara vez arrojará ningún defecto o un defecto fijo. Por lo que casi siempre será o bien defecto normal, o bien reversible.

14. Campo output.

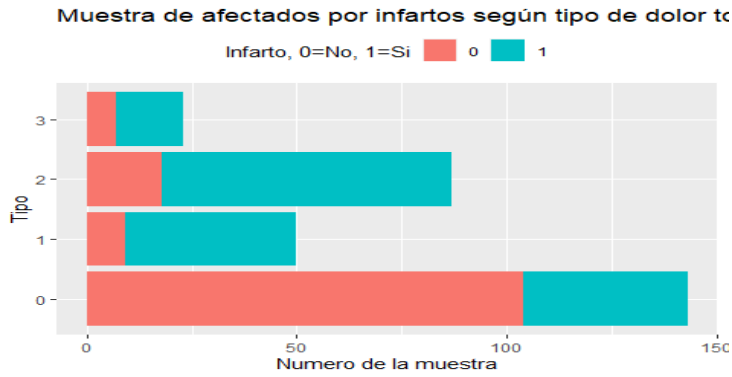


El atributo predicho que diagnostica enfermedad cardíaca está igualmente repartido, con lo que casi la mitad de nuestros pacientes tendrán ataque al corazón, y casi la otra mitad, no lo tendrá. Comprobamos con nuestros cálculos que tenemos 165 registros de pacientes que han sufrido un infarto y 138 que no, es decir, de grupo de control. Por indagar un poco más en nuestra variable output, si la segmentamos según el sexo, podemos observar lo siguiente:



Podemos decir que las mujeres tienen mayor probabilidad de sufrir un ataque al corazón, ya que en su mayoría, adoptan el valor de infarto en el atributo output de nuestro dataset. En cambio, los hombres, tienen casi las mismas probabilidades de sufrir un ataque al corazón, que de no sufrirlo, ya que se reparte de manera equitativa el valor dicotómico de nuestra variable predicha.

Si segmentamos según la variable cp (tipo de dolor) tendremos lo siguiente:



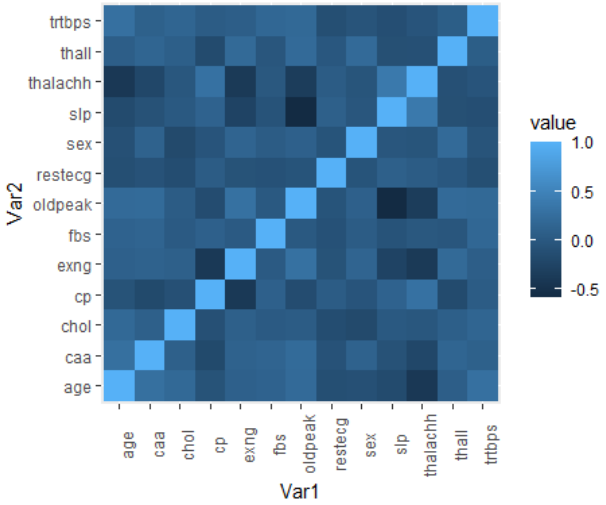
Podemos observar que el dolor no anginoso es la que en proporción está mas relacionada con ataques al corazón.

Estudio de correlaciones

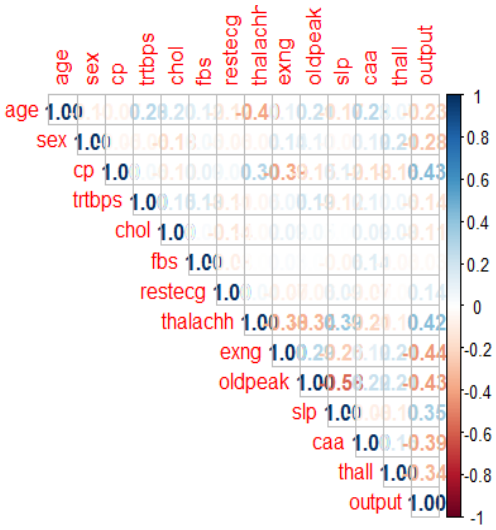
A continuación, haremos un análisis de las distintas correlacciones que existen entre las variables. Para ello usaremos las función "qplot", que a través de la función cor, nos dará una visualización de la fuerza de relación entre nuestras variables.

En este caso, la visualización vendrá dibujada a modo de mapa de calor, e interpretaremos los resultados obtenidos gracias a la escala de colores.

También podemos extraer los valores concretos de nuestra correlación, para asegurarnos de que estamos en lo cierto y por si queremos analizar más en profundidad cada caso.



Incluso podemos dibujar un gráfico muy intuitivo y “friendly” que es el que obtenemos a continuación:



Al interpretar los resultados obtenidos observamos que no hay ningún campo que tenga una correlación alta entre ellos, tanto si nos fijamos en los valores numéricos como si nos fijamos en los colores de los gráficos. Si acaso, podemos nombrar que las variables que mas peso tienen en cuanto a una relación entre ellas son oldpeak y slp.

---

2. Integración.

Tal y como observamos en nuestro anterior resumen de la estructura del dataset, el fichero “heart.csv” contiene 303 registros y 14 variables.

Este es el conjunto de datos con el que trabajaremos para hacer nuestro análisis, y usaremos todas las variables contenidas en el mismo para ver cómo se distribuyen los datos.

De hecho, con estos datos tendremos suficiente información para llevar a cabo nuestro estudio y lograr el objetivo de extraer conclusiones asociadas a si un individuo sufrirá de un ataque al corazón o no, que es el principal foco de nuestro estudio.

---

3. Limpieza de datos.

3.1. Elementos nulos

En este caso no encontramos nada anómalo con estas comprobaciones, y todos los registros están informados.

Si hubiéramos encontrado, por ejemplo, presencia de NAs en algún campo deberíamos de tratarlo.

Inicialmente cuantificar el valor de esos NAs y dependiendo de esta dimensión, tomaríamos 2 alternativas diferentes.

La primera sería eliminar las filas/registros donde aparece el NA siempre y cuando sean pocos.

La segunda se aplicaría si el valor de NAs es alto, esto nos conduce a eliminar toda la columna/campo ya que una alta presencia de NAs distorsionaría el modelo por no poder evaluar fiablemente el campo.

Existe una tercera alternativa, que siempre y cuando los datos nos lo permitan, podríamos transformar el campo de NA a "Desconocido", "No informado", … Pero como en el caso primero, esto sólo es válido si la muestra de NAs es pequeña.

3.2. Valores extremos

Un método bastante eficiente para detectar outliers es estudiar las medidas estadísticas de nuestros campos del dataset, y para ello, nos fijaremos en los valores estadísticos obtenidos mediante el comando summary para cada uno de los campos de este conjunto de datos.

Lo primero que tendremos que observar será la distancia entre los valores de media y mediana de cada una de las dimensiones, que nos dará información sobre si ocurre algo que tengamos que analizar (en el caso de que sean éstas muy distantes).

Luego, calcularemos el rango intercuantílico (IQR = Q3 - Q1), donde tomaremos como valor atípico extremos aquél que dista 3 veces del mismo por debajo de Q1 o por encima de Q3 ($q < Q1 - 3 * IQR$ o bien $q > Q3 + 3* IQR$).

Además, iremos lanzando el diagrama de cajas para cada caso, dibujando también los puntos, ya que es una forma bastante gráfica que nos reforzará, de manera visual, las conclusiones que obtengamos de nuestro análisis estadístico.

Ahora, estudiaremos entonces cada una de las variables, realizando este análisis aquí explicado en los casos donde aplique realizarlo.

1. Campo age.

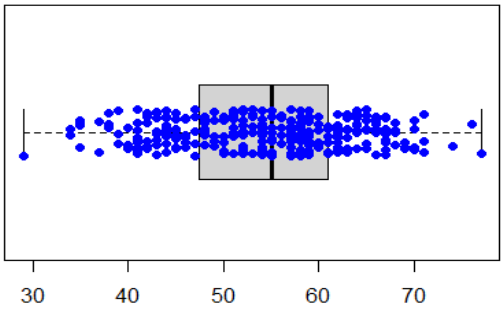
La media y la mediana son muy cercanas, lo cual denota que no tendremos ningún valor atípico en este campo.

IQR = 61 - 47.50 = 13.5

Umbral inferior = 47.50 - 3 *13.5 = 7

Umbral superior = 61+ 3 *13.5 = 101.5

Podemos decir además que nuestros valores mínimo y máximo están dentro de nuestro umbral.



Si observamos el gráfico, vemos que todos los puntos caen dentro de nuestros bigotes, tal y como intuíamos de nuestro análisis de medidas.

No tendremos valores extremos en este campo.

2. Campo sex.

Se trata de una variable dicotómica, en la que observando sus valores estadísticos concluimos que su mínimo es 0 y su máximo es 1, con lo cual, todo está según lo esperado.

3. Campo cp.

En este caso, la variable es discreta, pudiendo tomar 4 posibles valores. Si observamos su resumen estadístico, todo parece normal, ya que el máximo es 3 (el mayor valor que puede tomar) y su mínimo cero (coincidiendo con el mínimo valor que puede tomar).

4. Campo trtbps.

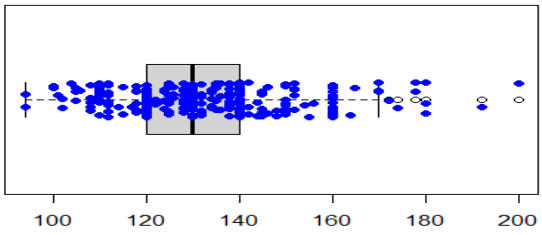
La media y la mediana son muy cercanas, lo cual hace parecer que no tendremos valores muy alejados de nuestro rango mayoritario.

IQR = 140- 120= 20

Umbral inferior = 120- 3 *20=60

Umbral superior = 140+ 3 *20=200

Podemos decir además que nuestros valores mínimo y máximo están dentro de nuestro umbral.



En este caso visualizamos algunos “posibles outliers”, pero no los considerarmos valores extremos por no estar más alejados que nuestros umbrales marcados según los cálculos estadísticos.

No tendremos valores extremos en este campo.

5. Campo chol

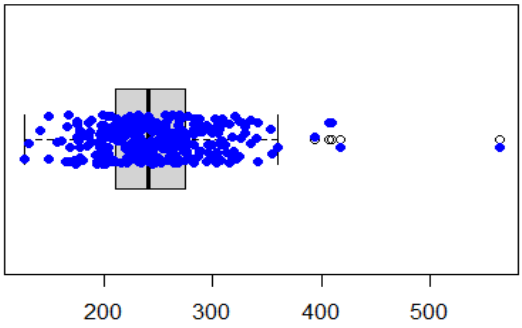
La media y la mediana están ligeramente alejadas.

IQR = 274.5 - 211= 63.5

Umbral inferior = 211- 3 *63.5= 20.5

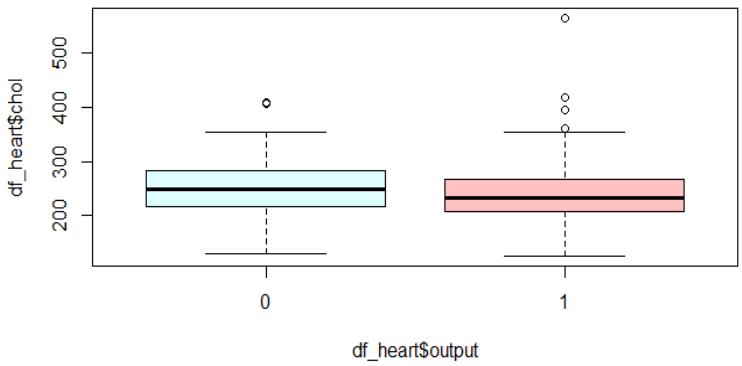
Umbral superior = 274.5+ 3 *63.5= 465

Podemos decir entonces que nuestro valores máximo se sitúa fuera de nuestro umbral.

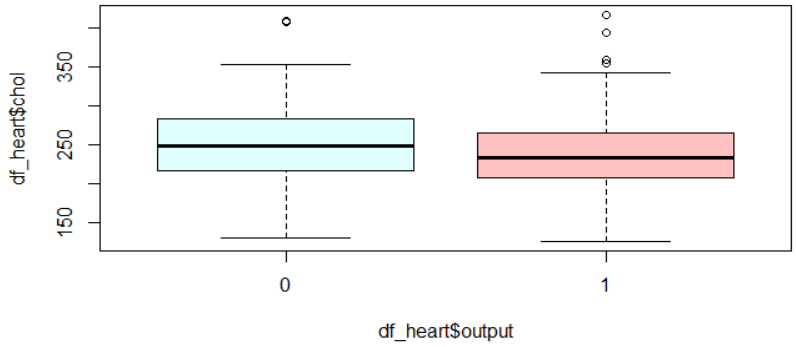


En colesterol tenemos un outlier que nos puede llevar a confusión a la hora de evaluar esta varaible con respecto a la probabilidad de provocar infartos.

Detallemos un poco más este resultado comparándolo con nuestro atributo predicho.



Por lo tanto, vamos a asignarle a este valor extremo la media de los del grupo control.



6. Campo fbs.

Se trata de una variable dicotómica, donde coherentemente, tenemos 1 como valor máximo y 0 como valor mínimo.

No tendremos valores extremos en este campo.

7. Campo restecg.

En este caso, los valores posibles serán un total de 3 distintos, y concuerda con nuestras medidas estadísticas obtenidas.

No tendremos valores extremos en este campo.

8. Campo thalach .

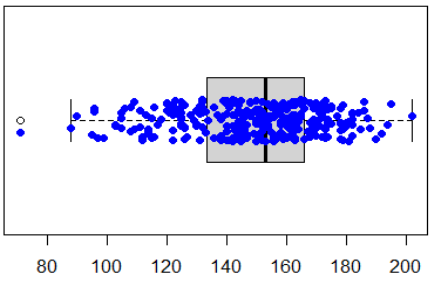
La media y la mediana son muy cercanas, por lo que no hay indicio de valores extremos a simple vista.

IQR = 166- 133.5= 32.5

Umbral inferior = 133.5- 3 *32.5=36

Umbral superior = 166+ 3 *32.5=263.5

Podemos decir además que nuestros valores mínimo y máximo están dentro de nuestro umbral calculado.



Si nos fijamos en el gráfico, y tratándose de una variable que mide valores “máximos” de frecuencia cardíaca, decidimos no eliminar los dos valores que se salen del extremo izquierdo de nuestro bigote del diagrama, ya que caen dentro de nuestro umbral.

No tendremos valores extremos en este campo entonces.

9. Campo exng.

Se trata de una variable dicotómica que queda correctamente definida con sus valores estadísticos calculados.

No tendremos valores extremos en este campo.

10. Campo oldpeak.

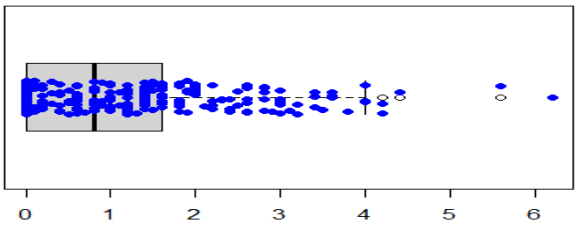
La media y la mediana son muy cercanas, aunque al tratarse de valores contenidos en un rango no muy grande, tendremos que analizarlo con especial cuidado por si acaso.

IQR = 1.60- 0= 1.6

Umbral inferior = 0- 3 *1.6= -4.8

Umbral superior = 1.60+ 3 *1.60= 6.4

Podemos decir que nuestros valores mínimo y máximo están dentro de nuestro umbral calculado.



Lo mismo apreciamos observando el gráfico.

Así que no tendremos valores extremos en este campo.

11. Campo slp

En este caso, este campo solo puede tomar 3 valores, y parece todo correcto en las medidas estadísticas.

No tendremos valores extremos en este campo.

12. Campo caa.

En este caso, este campo solo puede tomar 5 valores, y parece todo correcto en las medidas estadísticas.

No tendremos valores extremos en este campo.

13. Campo thall

En este caso, este campo solo puede tomar 4 valores, y parece todo correcto en las medidas estadísticas.

No tendremos valores extremos en este campo.

14. Campo output

En este caso, este campo sólo puede tomar 2 valores, y parece todo correcto en las medidas estadísticas.

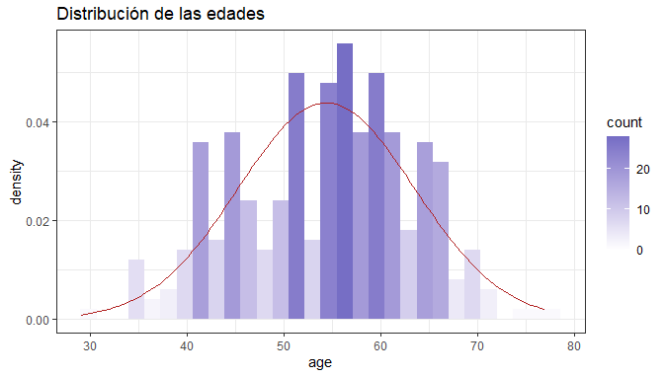
No tendremos valores extremos en este campo.

---

4. Análisis de datos.

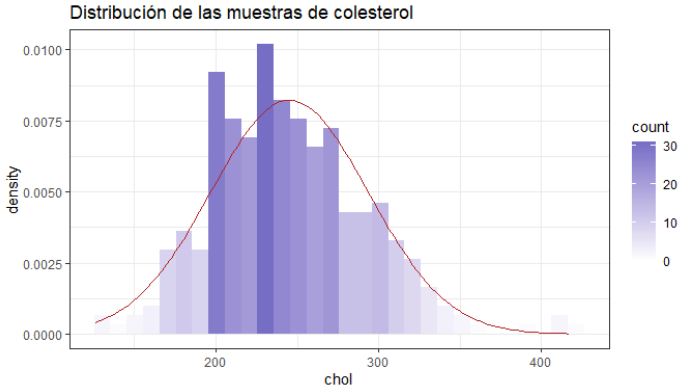
4.1. Contraste de Hipótesis

Primero para tener un estudio sólido debemos comprobar la normalidad de nuestra muestra antes de la obtención de resultados.



La distribución de las edades se aproxima a una distribución normal.

Distribución del colesterol de la muestra, que usaremos en el posterior contraste de hipótesis:



Podemos comprobar que las muestras de las edades y de colesterol se distribuyen normalmente.

En este apartado nos vamos a preguntar si la media del colesterol en sangre es mayor en los que han sufrido un infarto, con respecto a los que no.

Realizamos un contraste de hipótesis. La formulación de las hipótesis es:

$H_0 : \mu_{ha} \leq \mu_{ctrl}$

$H_1 : \mu_{ha} > \mu_{crtl}$

Siendo μ_{ha} la media del colesterol en sangre de los pacientes que han sufrido infoarto H_0 es la hipótesis nula que indica que las medias son menores o iguales. H_1 es la hipótesis alternativa que es la que planteamos en la pregunta de investigación, que la media del colesterol de los pacientes que han sufrido infarto μ_{ha} es mayot a la media del colesterol en el grupo control μ_{ctrl}.

Ya que se trata de una hipótesis referente a la media, vamos a realizar un test sobre esta. Como disponemos de una muestra grande podemos asumir la normalidad de la muestra por el teorema del límite central.

No se conoce la varianza de la población. Primero debemos comprobar si las varianzas son iguales

Obtenemos `p-value = 0.5746`
Ya que tenemos un valor de p superior al nivel de significación ($\alpha = 0.05$) estableciado debemos aceptar la igualdad de varianzas en las dos poblaciones.

Por lo tanto aplicamos un test unilateral por la derecha sobre la media, tenemos dos muestras independientes con varianza igual.

```
## t = -1.9409, df = 301, p-value = 0.9734
```

Obtetnemos el estadístico p mayor que $\alpha = 0.05$ por tanto no podemos rechazar la hipótesis nula y concluimos que la media del colesterol en personas afectadas con un infarto es igual o menor que en las que no han sido afectadas con una confianza del 95%. Aunque en base a los estudios consultados el colesterol es un factor de riesgo para poder desarrollar una enfermedad cardiovascular, en nuestro contraste de hipótesis no ha salido tal cosa, esto puede deberse al tipo de colesterol medido o al número de la muestra, además de poder considerar otro tipo de variables como el desarrollo de otras patologías que influyen en el resultado independientemente de la cantidad de colesterol.

En base a esto vamos a desarrollar otro contraste de hipótesis, pero en este caso con la población que no ha desarrollado trastorno de la sangre llamado talasemia (variable thall igual a 2), comparando sus niveles de colesterol.

```
## t = 2.6004, df = 301, p-value = 0.004886
```

Obtenemos el estadístico p menor que $\alpha = 0.05$ por tanto se puede rechazar la hipótesis nula y concluimos que la media del colesterol en personas afectadas con un infarto es mayor que en las que no han sido afectadas con una confianza del 95%, en personas no afectadas por talasemia. El colesterol es un factor de riesgo en las enfermedades coronarias en personas aparentemente sanas.

4.2 Modelo de Regresión logística

Ya que la variable dependiente de nuestro estudio, output, es una variable categórica y dicotómica podemos crear un modelo de regresión logística con el objetivo de predecir futuros casos.

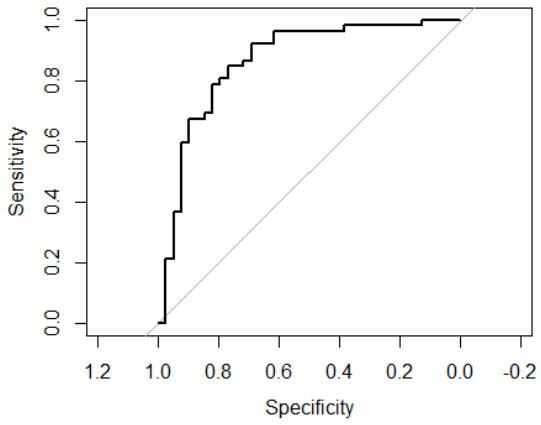
Para realizar la regresión logística y comprobar su efectividad vamos a dividir el data frame en dos, uno de train y otro de test de manera aleatoria, representando el 70% y el 30% respectivamente.

Creamos el modelo y probamos con el test y creamos la matriz de confusión.

```
## ytest  0  1
##     0 27 12
##     1  7 45
```

El modelo tiene una exactitud del 79.12%.

Otra medida que nos puede ser útil para medir la precisión del modelo es medir el área bajo la curva ROC:



El área debajo de la curva es de 0.861 por lo que el modelo de regresión logística podemos decir que tiene un buen ajuste.

Por último, otro dato que nos puede interesar es la sensibilidad del modelo, ya que es preferible tener falsos positivos (pacientes que podemos catalogar de riesgo de ataque

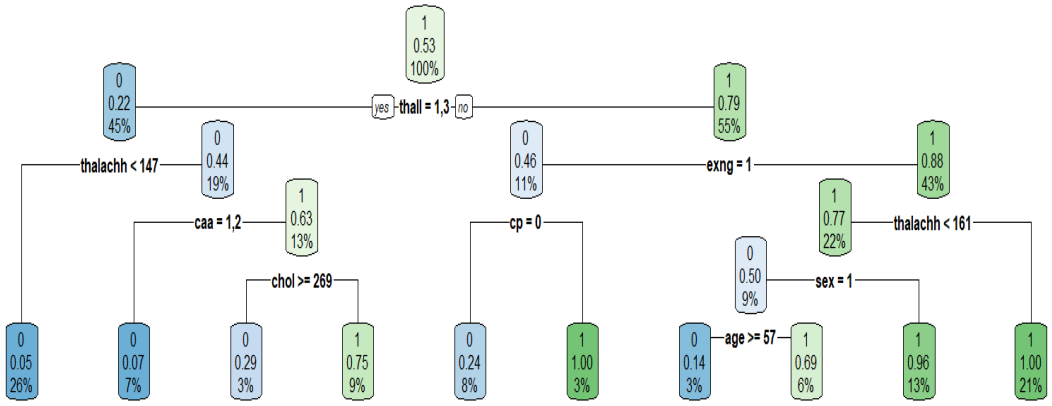
al corazón sin ser realmente así) que falsos negativos (pacientes catalogados como sanos cuando son propensos a sufrir un infarto). La sensibilidad se define como :

$$Sensibilidad = \frac{VP}{P}$$

La sensibilidad del modelo es del 20.6%, sería necesario entrenar un modelo más robusto estudiando la influencia de las variables o con muestras más grandes.

4.3 Árbol de decisión

Otro modelo que podemos realizar es un árbol de decisión, donde según los valores que tomen las variables explicativas podamos predecir la variable dependiente output. Para ello creamos el modelo con el conjunto de entrenamiento.



Usando todas las variables explicativas obtenemos unos valores de precisión del 72.5%, y una sensibilidad del 74.3%, en nuestro caso el modelo de regresión logística tiene mayor poder de predicción.

Conclusiones

A través del estudio de los datos hemos podido concluir que la distribución de la muestra de los sujetos cumple una distribución normal, lo que nos ha permitido realizar distintos modelos de predicción teniendo un mejor ajuste en el modelo de regresión logística.

También hemos comprobado como en pacientes aparentemente sanos la presencia de mayor cantidad de colesterol puede influir en el desarrollo de un infarto.

Bibliografía

Cachofeiro, V. «Alteraciones del colesterol y enfermedad cardiovascular». https://www.fbbva.es/microsites/salud_cardio/mult/fbbva_libroCorazon_cap13.pdf. (Fecha de consulta: 01/06/22.)

Contribuciones al trabajo desarrollado en esta práctica

CONTRIBUCIONES	FIRMA
Investigación previa	SBRG, DJCG
Redacción de las Respuestas	SBRG, DJCG
Desarrollo del código	SBRG, DJCG