Tipología y ciclo de vida de datos

PRA2

Ricardo López Soria

44595500H

Universitat Oberta de Catalunya

ÍNDICE

1. Descripción del dataset

¿Por qué es importante y que pregunta/problema pretende responder?

En este dataset trata de relacionar, a través de una serie de atributos de una persona, si es mas probable que vaya a sufrir en un futuro un ataque al corazón o no.

Es importante este estudio ya que permitiría cuantificar de manera porcentual si una persona con esas condiciones estará mas cerca de sufrir este ataque, en base a eso se puede determinar que la persona intente solucionar esos parámetros médicos para que baje la probabilidad de sufrir un ataque al corazón.

Las variables que se encuentran en este dataset son los siguientes:

- **Age**: Edad del paciente
- **Sex**: Sexo del paciente (1 = ¿?, 0 = ¿?) En la información del conjunto de datos no se obtiene ninguna información acerca del significado de cada uno de los números. Ni si quiera en los comentarios propuestos por otras personas que le preguntan acerca de esta información.
- **Exang**: Si tiene angina producida por el ejercicio, valor comprendido entre (1 = si, 0 = no)
- **Caa**: Número de vasos principales en el corazón (de 0 a 3), números enteros, en principio los humanos tenemos 4, así que a falta de mas información en la web del dataset, se presupone que se tienen con el (0 = 1 vaso sanguíneo, con 3 = 4 vasos sanguíneos)**.**
- **Cp:** Dolor en el pecho, clasificados en dos niveles:
    - **0 =** Angina típica
    - **1 =** Angina atípica
    - **2 =** Dolor no anginoso
    - **3 =** Asintomático
- **Trtbps:** Presión arterial en reposo (en mmHg)
- **Chol:** Colesterol en ‘mg/dl’ obtenido a través del sensor IMC
- **FBS:** Azúcar en sangre en ayunas, si se tiene mas de 120 mg/dl (1 = verdadero, 0 = falso)
- **Rest_ecg:** Resultados electro cardíacos en reposo
    - **0 =** normal
    - **1 =** anomalía de la onda ST-T
    - **2 =** Hipertrofia ventricular izquierda probable o definida.
- **Thalach:** Frecuencia cardíaca máxima alcanzada.

- **Valor objetivo a aprender**
    - **0 =** menos probabilidad de ataque cardíaco.
    - **1 =** mas probabilidad de ataque cardíaco.

Con estos datos lo que se va a perseguir es averiguar cuales son los parámetros mas relevantes que hacen que se tenga mas probabilidades de tener un ataque cardiaco. En el caso de que alguno de estos parámetros sea excesivamente determinante, como puede ser el valor de ‘CA’, se eliminarían esos datos obtenidos creando dos grupos para analizar mejor los resultados.

En función de como se vaya dando el análisis de los datos se irá viendo como atacar el problema para obtener los datos mas interesantes de este dataset.

2. Integración y selección

De los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

En este caso el dataset de estudio no necesita fusionarse con otro dataset, puesto que el resultado de este dataset viene incorporado. Introducir otro dataset no tendría sentido.

Antes de entrar a analizar y retocar el dataset es necesario realizar un primer vistazo de lo que contiene. Para ello con 'R' se realiza el comando, 'summary' dándonos información de lo que se tiene. Al analizarlo nos encontramos con que se tienen variables de las que en el enlace web no se encuentra información. Estas variables son: *oldpeak*, *slp* y *thall*.

Dado que no se tiene información de lo que contiene, realizar un análisis de con estas variables no tendría sentido ya que no se conocerá el significado, por tanto estas columnas se eliminan y se seleccionan las otras.

Se ha realizado un estudio de duplicidad y se ha encontrado una usando la función 'duplicated', se ha eliminado este duplicado usando la función 'unique'.

Por tanto, el resumen de las variables con las que se parte son las siguientes:

```
> summary(datosHeartUniqueLess)
      age             sex               cp             trtbps           chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0   Min.   :126.0   Min.   :0.000  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.000  
 Median :55.50   Median :1.0000   Median :1.0000   Median :130.0   Median :240.5   Median :0.000  
 Mean   :54.42   Mean   :0.6821   Mean   :0.9636   Mean   :131.6   Mean   :246.5   Mean   :0.149  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0   3rd Qu.:274.8   3rd Qu.:0.000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0   Max.   :564.0   Max.   :1.000  
    restecg          thalachh         exng             caa             output     
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:133.2   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000  
 Median :1.0000   Median :152.5   Median :0.0000   Median :0.0000   Median :1.000  
 Mean   :0.5265   Mean   :149.6   Mean   :0.3278   Mean   :0.7185   Mean   :0.543  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :4.0000   Max.   :1.000  
```

Figura 1: Resumen de las variables.

La variable 'output' es el resultado que con el que se quiere relacionar las entradas.

3. Limpieza de datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Para encontrar elementos vacíos en un dataset con 'R' con el comando:
'colSums(is.na(datosHeart))'

Se obtiene una suma de los valores que están vacíos, en este caso el valor es cero, por lo tanto no hay ningún valor vacío o 'NA'.

Para la búsqueda de valores que son cero y que no deberían serlo, gracias a la Figura 1, se puede observar si hay algún valor fuera de lugar. Revisando variable por variable gracias al resumen que nos muestra 'R' y se puede observar en la Figura 1.

Analizando las variables podemos observar lo siguiente:

- **Age:** Tiene valores normales, un mínimo de 29 y un máximo de 77, valores razonables para ser la variable Edad.

- **Sex:** Solo hay ceros y unos, por tanto los valores que se ven en el resumen son los lógicos. Hay mas '1' que '0', pero como no sabemos a que sexo hace referencia, no se puede sacar mas conclusiones.

- **Cp:** Valores dentro de lo establecido por el enunciado.

- **Trtbps:** La presión arterial tiene valores extremos que habrá que analizar si tienen sentido, como el de 200mmHg. Según varias fuentes, con una tensión superior a 180mmHg habría que ir a sala de emergencias. Es un valor extremo que sin embargo hay mas valores parecidos a éste, por lo que no se considerará un error. En cambio, el valor extremo de 94mmHg no es tan raro de ver.

- **Chol:** Aunque el valor extremo máximo de 564mg/dl es muy raro, no es imposible y son valores que se puede llegar a tener. No se realizará ninguna acción sobre esta variable.

- **FBS:** Valor de 1 y 0, valores comprendidos entre ese rango.

- **Rest_ecg:** Valores comprendidos según enunciado, por lo que son correctos.

- **Thalach:** Valores normales, incluyendo los máximos y mínimos.

- **Exang:** Valor de 1 y 0, valores comprendidos entre ese rango.

- **Caa:** Encontramos que hay valores por encima de 3, según el enunciado esto no es posible, ya que el valor máximo es de 3. En este caso hay dos posibilidades, eliminar estas filas de datos o cambiar todos los que tengan valores de 4 por el del número 3. Se observa que hay 5 filas con el número 4. Dado que nos faltan datos sobre la recogida de datos para averiguar a que se debe el fallo. Se eliminarán estas filas que tengan el valor de 4.

- **Output:** Valor de 1 y 0, valores comprendidos entre ese rango.

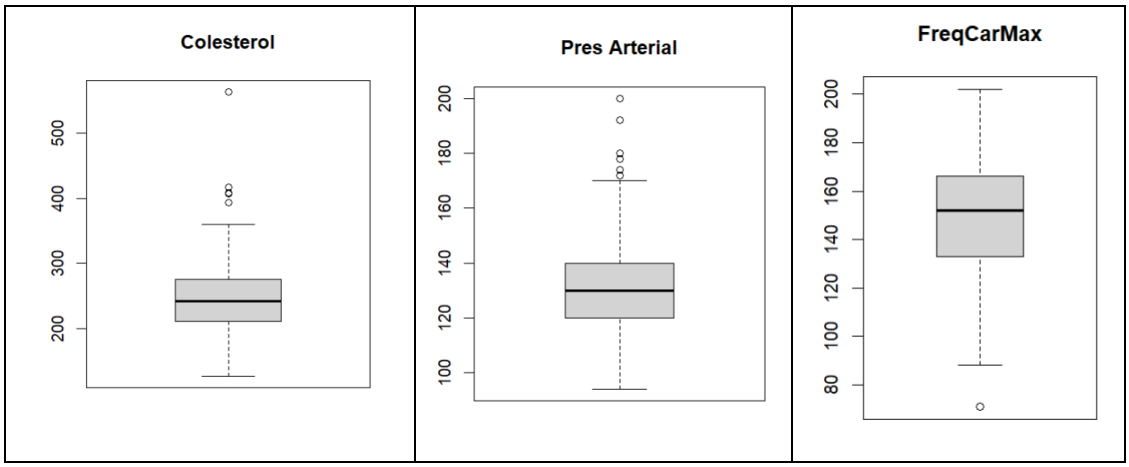
3.2. Identifica y gestiona los valores extremos

Eliminamos con 'R' las filas comentadas quedándonos en los siguientes datos.

```
> summary(datosHeartUniqueLess[datosHeartUniqueLess$caa < 4, ])
      age             sex               cp             trtbps      
 Min.   :29.00   Min.   :0.0000   Min.   :0.0000   Min.   : 94.0  
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:120.0  
 Median :56.00   Median :1.0000   Median :1.0000   Median :130.0  
 Mean   :54.51   Mean   :0.6779   Mean   :0.9597   Mean   :131.6  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.:140.0  
 Max.   :77.00   Max.   :1.0000   Max.   :3.0000   Max.   :200.0  
      chol            fbs             restecg          thalachh    
 Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
 1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.0  
 Median :241.5   Median :0.0000   Median :1.0000   Median :152.5  
 Mean   :246.9   Mean   :0.1477   Mean   :0.5235   Mean   :149.5  
 3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:165.8  
 Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
      exng             caa             output     
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000   Median :1.0000  
 Mean   :0.3289   Mean   :0.6745   Mean   :0.5403  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :3.0000   Max.   :1.0000  
> 
```

Figura 2: Eliminando filas que contengan caa > 3

Para asegurarnos de que los valores numéricos no contienen valores outlier se realizan gráficos de cada uno de ellos con 'R'. Conforme se vea alguno de los outlier se van eliminando esas filas y se vuelve a realizar el estudio con esa fila eliminada. El primero se realiza con el colesterol, que se observa que tiene uno muy alejado, por lo que se procede a su eliminación.



Valor extremo de colesterol: 564, se elimina, y el valor mínimo de frecuencia cardíaca máxima de 71 también se elimina. Los de la presión arterial se dejan ya que al ser un número de casos relevantes es probable que no sean valores extremos falsos.

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar

Se van a analizar todos los datos que quedan en el dataset una vez eliminadas las variables comentadas y eliminadas varias filas.

Se quiere realizar un modelo supervisado, es decir una clasificación, se comprobará según los datos de entrada si tendrá mas o menos posibilidades de tener un infarto. Una vez obtenido el modelo con los datos de aprendizaje, éste nos servirá para comprobar con nuevas entradas si estos pacientes tienen probabilidad de tener infarto.

Antes de realizar cualquier análisis se ha realizado una normalización de las variables: 'trtbps', 'chol' y 'thalachh'.

Por otro lado, merece la pena ver algunos gráficos que lo relacionen con el resultado a obtener. Por ejemplo, se puede ver la relación de sexo para ver si es mas probable o no que se tenga un infarto (se ha tomado como que el valor 1 = mujer y 0 = hombre), Figura 3.

Se muestra también los gráficos de otros valores para tener una visual de las variable. En principio solo se verán las que están discretizadas.

En la Figura 5 se observa como obtener un valor de 0 en Cp hace que tengas menos probabilidades de tener un ataque cardíaco (es decir un dolor de angina típica, según el enunciado). Para entender estos gráficos hay que saber que viene del gráfico mostrado en la Figura 4, donde hay un recuento de que valores tiene cada análisis, en los porcentajes lo que se hace es extenderlo para poder compararlos.

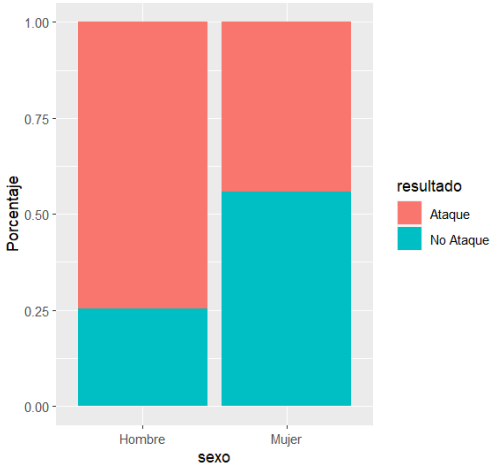


Figura 3: Porcentaje de ataques cardíacos en función del sexo

Ahora se muestras las otras variables que a priori podrían ser interesantes, que son el azúcar en sangre y el 'Rest_ecg'.

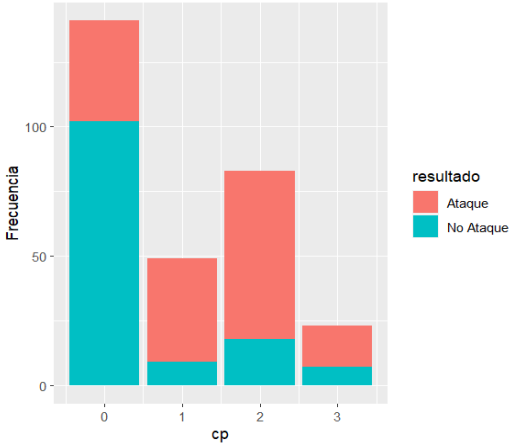


Figura 4: Frecuencia de at.card. en función del Cp

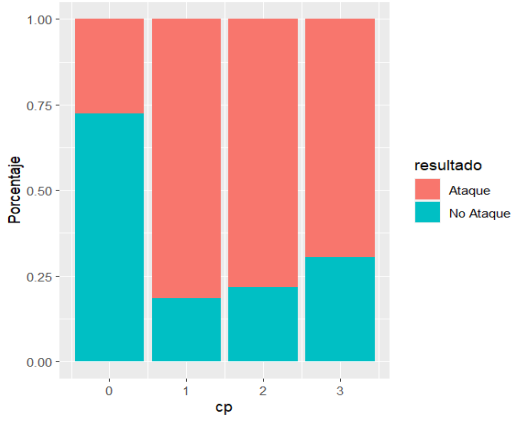


Figura 5: Porcentaje de ataques cardíacos en función de los dolores del pecho

Como se pueden observar en los gráficos de la Figura 6 y Figura 7, no parece que el azúcar en sangre sea relevante para tener mas probabilidad de ataques cardiacos, en cambio si que observamos que los que tienen una 'anomalía de la onda ST-T' en los resultados del cardio en reposo tienen mas probabilidades de tener un ataque cardíaco.

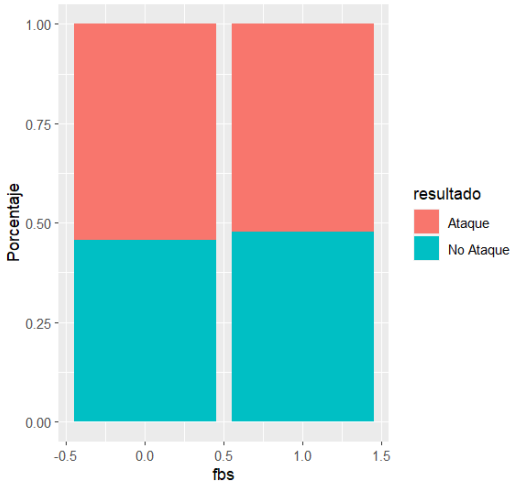


Figura 6: Azúcar en sangre, (1 = mas de 120mg/dl y 0 = menos de ese valor)

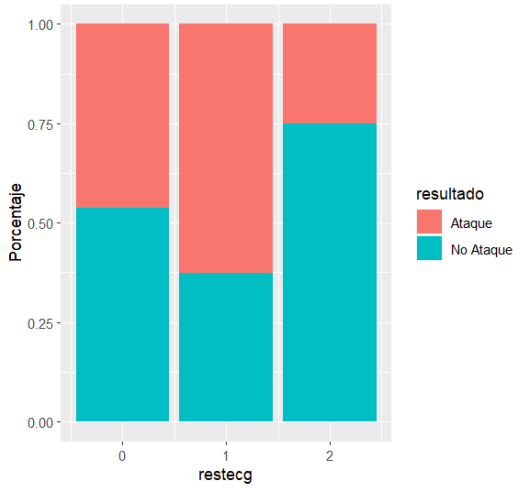


Figura 7: Resultados electro cardíacos en reposo

Otro gráfico interesante para intentar ver correlaciones entre variables se encuentra en el gráfico de la Figura 8, aunque a primera vista cuesta distinguir alguna correlación evidente.

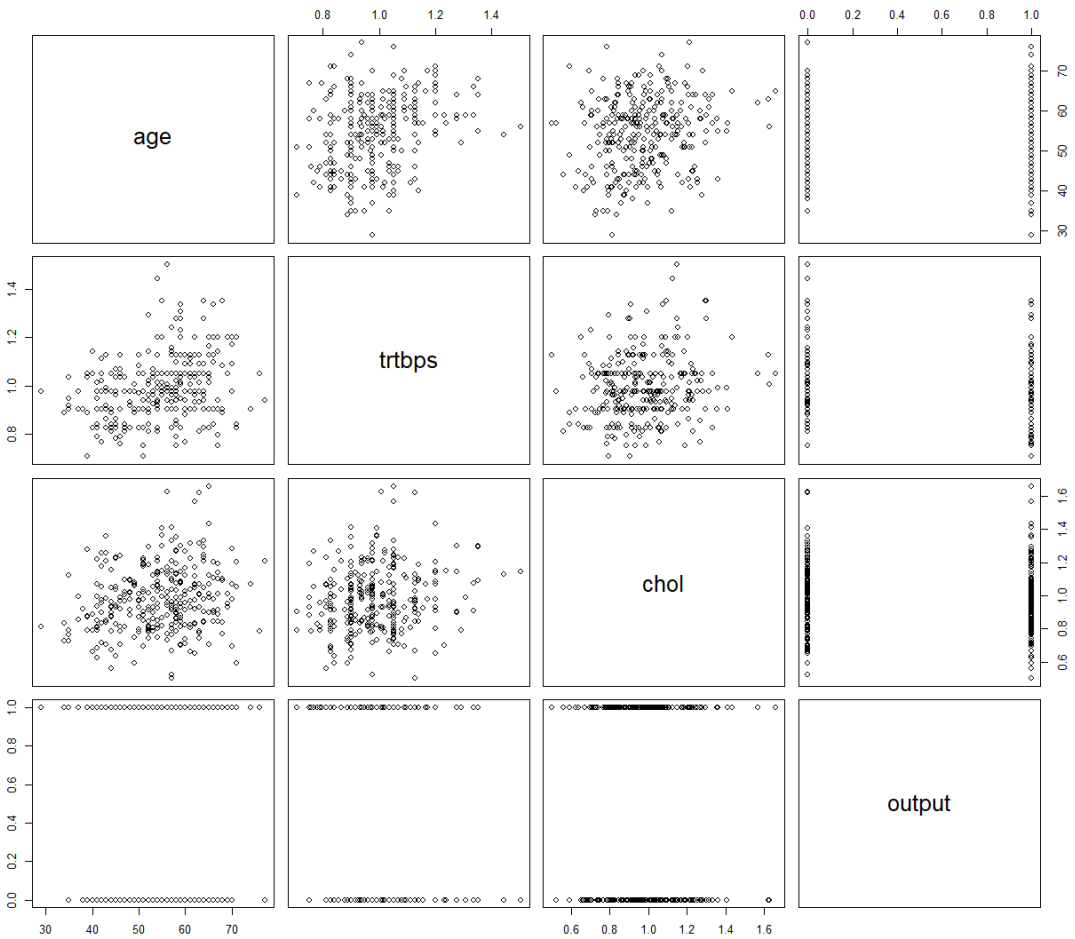


Figura 8: Correlación entre variables

Para discretizar valores habría varios métodos, uno de ellos sería totalmente manual, proponiendo nosotros una discretización de en tramos. Otra manera podría ser usando clustering. Éste es el método que se ha querido hacer para el colesterol.

El cluster se realiza mediante la función ‘discretize’ de ‘R’ usando el método ‘cluster’ y en 3 grupos.

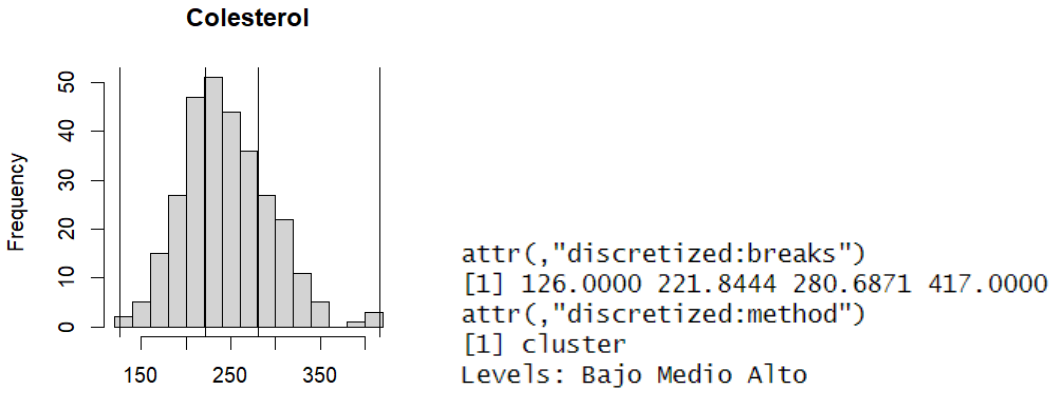


Figura 9: División en 3 grupos en los análisis de colesterol.

De 126 a 221 colesterol bajo, de 221 al 280 medio, y de 280 al 417 colesterol alto son los resultados obtenidos.

Una vez divididos en 3 grupos se comprueba su relación con la probabilidad de ataque cardíaco, Figura 10.

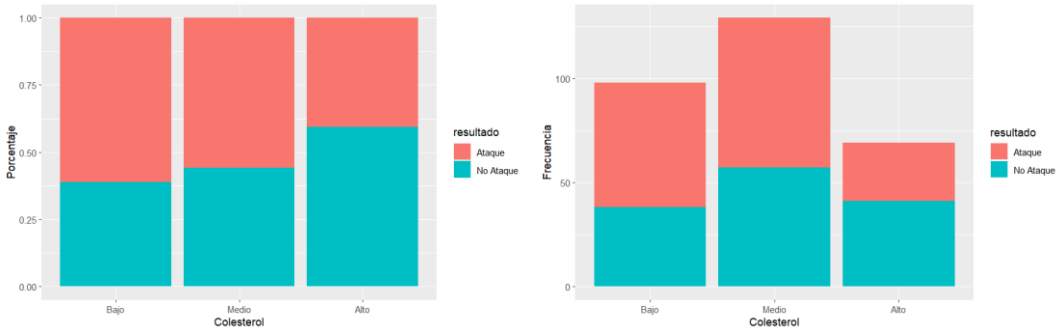
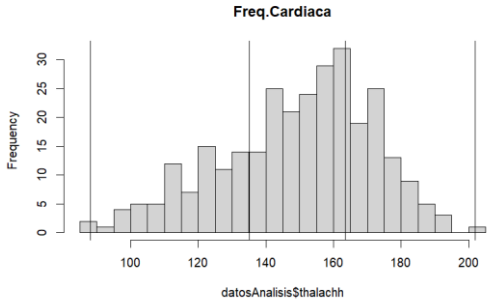


Figura 10: Relación entre el colesterol y los ataques cardíacos.

Contra lógica se obtiene que los que tienen bajos niveles de colesterol tienen mas probabilidad de obtener un ataque cardíaco. Estos resultados hay que cogerlos con cuidado, ya que puede que no sea el factor determinante para el resultado del ataque al corazón.

Realizando exactamente el mismo estudio para el valor de la frecuencia cardíaca se obtiene lo siguiente, Figura 11 y Figura 12.




```
attr(,"discretized:breaks")
[1]  88.0000 129.8192 157.6088 202.0000
attr(,"discretized:method")
[1] cluster
Levels: Bajo Medio Alto
```

Figura 11: División en 3 grupos en los análisis de frecuencia cardiaca.

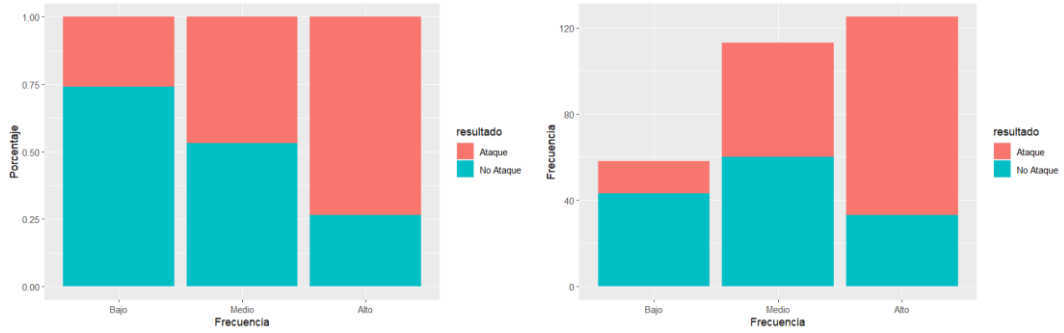


Figura 12: : Relación entre el Frecuencia cardíaca y los ataques cardíacos

En este caso si que se puede observar como hay una mayor probabilidad de tener un ataque cardiaco con valores mayores en la frecuencia cardiaca.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para realizar una comprobación de la normalidad, lo vamos a realizar de los valores de la frecuencia cardiaca.

Se va a realizar el test de Shapiro-Wilk en 'R', los datos obtenidos han sido los siguientes.

```
> shapiro.test(datosAnalisis$thalachh)

        Shapiro-Wilk normality test

data:  datosAnalisis$thalachh
W = 0.97819, p-value = 0.0001746
```

Al no ser el valor de p-valor se concluye que los datos no cuentan con una distribución normal.

4.3. Aplicación de pruebas estadísticas.

Se va a realizar un modelo de aprendizaje supervisado, a partir de una serie de datos de entrenamiento, con esto se persigue predecir posteriormente el resultado de nuevos datos desconocidos.

Se dividen los datos de muestreo en 2/3 para el entrenamiento y 1/3 para el muestreo.

```
> #División para el muestreo
> h <- holdout(datosAnalisis$resultado, ratio = 2/3, mode = "stratified")
> data_train <- datosAnalisis[h$tr,]
> data_test <- datosAnalisis[h$ts,]
> print(table(data_train$resultado))

   Ataque No Ataque
      102        95
> print(table(data_test$resultado))

   Ataque No Ataque
       58        41
```

Una vez realizado se entrena el modelo mediante el método '*random forest'* con 4 folds. Tras eso se realiza una predicción con los modelos de test y después se compara con el resultado original que deberían dar. Con esto se puede obtener la calidad del modelo mediante la función confusionMatrix.

```
data_train_y = data_train$resultado
data_train_x = data_train[,1:10]

data_test_x = data_test[,1:10]
data_test_y = data_test$resultado

train_control <- trainControl(method = "cv", number = 4)
mod <- train(data_train_x, data_train_y, method="rf", trControl = train_control)

pred = predict(mod, newdata = data_test_x)
df1 <- factor(pred)
df2 <- factor(data_test_y)
confusionMatrix(df1, df2)
```

Obteniéndose los siguientes resultados.

```
Confusion Matrix and Statistics

          Reference
Prediction  Ataque No Ataque
  Ataque        46         8
  No Ataque      9        36

               Accuracy : 0.8283
                 95% CI : (0.7394, 0.8967)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 8.862e-09

                  Kappa : 0.6531

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8364
            Specificity : 0.8182
         Pos Pred Value : 0.8519
         Neg Pred Value : 0.8000
             Prevalence : 0.5556
         Detection Rate : 0.4646
   Detection Prevalence : 0.5455
      Balanced Accuracy : 0.8273

       'Positive' Class : Ataque
```

Con este modelo se tiene una posibilidad de acierto del 82%.

5. Resolución

Se ha logrado encontrar un modelo predictivo con un buen porcentaje de acierto, tal y como se ha mostrado a lo largo del documento, las variables por si solas indican poca información sobre la posibilidad de tener un ataque cardíaco. Sin embargo, cuando se recogen todas las variables a la vez se observa que es mas fácil de obtener una predicción mas ajustada a la realidad.

Resulta curioso los resultados del colesterol y lo poco relevante que es por si sola en comparación con las otras variables.

6. Contribución

Contribuciones	Firma
Investigación previa	Ricardo López
Redacción de las respuestas	Ricardo López
Desarrollo del código	Ricardo López