Práctica 2

Sebastian Maya Hernández

12/27/2021

Contents

1. Descripción del Dataset

El conjunto de datos seleccionado se llama "Heart Failure Prediction Dataset" y consiste en 11 características clínicas para predecir posibles eventos de enfermedad cardíaca. Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial, representando un 31% del total de muertes de todo el mundo[1]. En este estudio, queremos saber si la presencia de uno o más factores de riesgo como la hipertensión, diabetes, hiperlipidemia, etc, son la principal causa de las ECV y la insuficiencia cardíaca. También investigaremos si la edad y el género pueden ser factores de riesgo para las enfermedades cardíacas.

En este dataset se representan 918 observaciones o registros de los cuales las columnas o atributos que lo forman son las siguientes:

- **Age:** Edad del paciente en años.
- **Sex:** Sexo del paciente [M: Masculino, F: Femenino].
- **ChestPainType:** Tipo de dolor de pecho [TA: angina típica, ATA: angina atípica, NAP: dolor no anginal, ASY: asintomático].
- **RestingBP:** Presión arterial en reposo [mm Hg].
- **Cholesterol:** Colesterol sérico [mm/dl].
- **FastingBS:** Azúcar en sangre en ayunas [1: si BS en ayunas> 120 mg/dl, 0: en caso contrario].
- **RestingECG:** Resultados del electrocardiograma en reposo [Normal: Normal, ST: con anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST> 0,05 mV), HVI: que muestra una hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes].
- **MaxHR:** Frecuencia cardíaca máxima alcanzada [Valor numérico entre 60 y 202].
- **ExerciseAngina:** Angina inducida por el ejercicio [Y: Sí, N: No].
- **Oldpeak:** Valor numérico medido en depresión que representa ST (ST es un segmento en un electrocardiograma normalmente representa un área eléctricamente neutra entre la despolarización ventricular y la repolarización).
- **ST_Slope:** La pendiente del segmento ST del pico de ejercicio [Up: uploping, Flat: flat, Down: downsloping].
- **HeartDisease:** Clase de salida [1: Enfermedad cardíaca, 0: Normal].

Comenzamos leyendo el fichero de entrada con la información de los pacientes y comprobaremos que tipo de dato ha sido asignado automáticamente.

```
heart_df <- read.csv("heart_input.csv", sep=",")
str(heart_df)
```

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : chr  "M" "F" "M" "F" ...
##  $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

Como podemos observar todos los atributos han sido cargados y asignados a su clase correspondiente correctamente por lo tanto no aplicaremos ningún método de cambio de clase.

2. Integración y selección de los datos de interés a analizar

Como hemos mencionado anteriormente, a partir del conjunto de datos se plantea el problema de determinar qué variables tienen mayor influencia sobre la presencia de enfermedades cardíacas. Además, las personas con enfermedades cardiovasculares o que se encuentran el alto riesgo cardiovascular necesitan una detección y control temprano donde un modelo de regresión podría intervenir y estudios de hipótesis puedan ayudar a identificar propiedades con mayor significancia.

En aras de simplificación y para que este estudio pueda llegar a un publico más amplio y menos especializado en medicina, el análisis se realizará sobre un número reducido de atributos que son seleccionados considerando la probabilidad de mayor impacto que pueden tener en la existencia o no de ECV. Intuitivamente, las variables a relacionar y analizar serán las categorizadoras del paciente como son la edad y el sexo. También trabajaremos con el nivel de colesterol, la presión arterial, los niveles de azúcar en sangre, frecuencia cardíaca y la depresión ST (signo más temprano de infarto agudo del miocardio y generalmente está relacionado con la oclusión de una arteria coronaria)[2].

```
# Eliminamos los atributos que no son de interés
col_to_exclude <- c("ChestPainType", "RestingECG", "ExerciseAngina", "ST_Slope")
heart_df <- heart_df[, !(colnames(heart_df) %in% col_to_exclude)]
head(heart_df)
```

```
##   Age Sex RestingBP Cholesterol FastingBS MaxHR Oldpeak HeartDisease
## 1  40   M       140         289         0   172     0.0            0
## 2  49   F       160         180         0   156     1.0            1
## 3  37   M       130         283         0    98     0.0            0
## 4  48   F       138         214         0   108     1.5            1
## 5  54   M       150         195         0   122     0.0            0
## 6  39   M       120         339         0   170     0.0            0
```

3. Limpieza de los datos

En este apartado revisaremos la consistencia de los valores y normalizaremos los datos en caso necesario. También, buscaremos valores atípicos y los imputaremos aplicando diferentes criterios según el tipo de dato.

3.1 Ceros y elementos vacíos

Primero vamos a comprobar para que atributos no deberían existir elementos vacíos y si es pertinente substituirlos con el valor más adecuado o eliminarlos. Comenzamos con la normalización de los datos cualitativos.

```
# Reemplazamos nulls, NA y valores vacíos por U (Unknown).
heart_df$Sex[is.null(heart_df$Sex)] <- 'U'
heart_df$Sex[is.na(heart_df$Sex)] <- 'U'
# Pasamos los estados a mayúsculas para unificar el formato
heart_df$Sex <- toupper(heart_df$Sex)
unique(heart_df$Sex)
```

```
## [1] "M" "F"
```

```
# Marcamos como 'U' cualquier valor fuera de los admitidos
possible_gender_values <- c('M', 'F', 'U')
heart_df$Sex[!is.element(heart_df$Sex, possible_gender_values)] <- 'U'
unique(heart_df$Sex)
```

```
## [1] "M" "F"
```

Aplicamos un proceso similar para la normalización de los datos cuantitativos, pero en este caso procedemos a eliminar los registras ya que no podemos determinar si los valores faltante provienen de errores humanos,

fallo en el sistema, etc.

```
# Eliminamos los registros donde existe un valor vació
heart_df <- heart_df[complete.cases(heart_df), ]
```

Algunos de nuestros datos pueden tener valores iguales a 0 por lo tanto vamos a determinar en cuales atributos no tendría sentido encontrarnos con este valor para posteriormente imputarlos utilizando la media aritmética.

- Tenemos en cuenta que para la variable edad no deberían haber valores igual a cero ni menores a 18 ya que nos centraremos en pacientes mayores de edad. En este caso, eliminaremos el registro.
- La depresión que representa ST (Oldpeak) es un intervalo que puede tomar valores negativos, positivos e iguales a cero[3]. Por lo tanto no imputaremos estos valores.
- El nivel de azúcar en sangre (FastingBS) es una variable binaria por lo tanto no imputaremos los valores iguales a 0.

```
# Excluimos todos los pacientes con edad menor a 18 y mayores a 99 años.
heart_df <- heart_df[heart_df$Age > 18 | heart_df$Age <= 99, ]
# Imputamos los registros en donde no es posible tener un valor 0.
col_names <- c("RestingBP", "Cholesterol", "MaxHR")
for (column in col_names) {
  # Asignamos Na a los valores 0
  heart_df[column][heart_df[column] == 0] <- NA
  missing_values <- is.na(heart_df[column])
  # Imputación de la columna con media aritmética
  heart_df[column][is.na(heart_df[column])] <- as.integer(mean(heart_df[[column]], na.rm = TRUE))
  # Compruebo que no hayan NAs
  unique(is.na(heart_df[column]))
  print(paste("Número de valores imputados", sum(missing_values), "para", column))
}
```

```
## [1] "Número de valores imputados 1 para RestingBP"
## [1] "Número de valores imputados 172 para Cholesterol"
## [1] "Número de valores imputados 0 para MaxHR"
```

3.2 Identificación y tratamiento de valores extremos

El objetivo es evaluar los valores atípicos y aplicar una imputación por vecinos más cercanos, considerando en el cómputo de los vecinos más cercanos el resto de variables cuantitativas en el conjunto de datos. Por otra parte, la imputación se realizará con registros del mismo género.

Voy a realizar un análisis para cada una de los atributos ya que no quiero aplicar una imputación sobre valores que quizá tiene sentido en nuestro contexto.

```
head(boxplot.stats(heart_df$Age)$out)
```

```
## integer(0)
```

```
head(boxplot.stats(heart_df$RestingBP)$out)
```

```
## [1] 190 180 180 180 200 180
```

```
head(boxplot.stats(heart_df$Cholesterol)$out)
```

```
## [1] 468 518 365 412 529 100
```

```
head(boxplot.stats(heart_df$FastingBS)$out)
```

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

```
head(boxplot.stats(heart_df$MaxHR)$out)
```

```
## [1] 63 60
```

```
head(boxplot.stats(heart_df$Oldpeak)$out)
```

```
## [1]  4.0  5.0 -2.6  4.0  4.0  4.0
```

En los resultados anteriores podemos ver que los valores detectados como outliers para la presión arterial y el colesterol son los más altos dentro del conjunto de datos, no obstante estos pueden ser una perfecta representación de lo que esta ocurriendo con el paciente y una fuente de información valiosa para el análisis posterior. No considero que se deba realizar una imputación sobre estos valores pero crearemos un conjunto de datos alternativo en donde estos valores se han imputado con la técnica anteriormente mencionada y posteriormente veremos su influencia en le análisis. Por lo tanto, procedemos a imputar algunas de las columnas.

```
# Creo una copia del conjunto de datos para su posterior análisis con/sin imputación
heart_original_df <- cbind(heart_df)
# Listamos los valores posibles por tipo de género
gender_type_list <- unique(heart_df$Sex)
col_names <- c("RestingBP", "Cholesterol", "MaxHR", "Oldpeak")

# Primero asignamos outliers
for (column in col_names) {
  # Buscamos los valores atípicos usando las estadísticas del diagrama de cajas
  outliers = boxplot.stats(heart_df[[column]])$out
  # Con el which obtenemos el índice o posición del valor
  outlier_idx <- which(heart_df[[column]] %in% c(outliers))
  heart_df[[column]][outlier_idx] <- NA
  # Mostramos los valores atípicos encontrado
  print(paste("outliers", column, ":", toString(head(outliers)), "-> total", length(outliers)))
}
```

```
## [1] "outliers RestingBP : 190, 180, 180, 180, 200, 180 -> total 27"
## [1] "outliers Cholesterol : 468, 518, 365, 412, 529, 100 -> total 41"
## [1] "outliers MaxHR : 63, 60 -> total 2"
## [1] "outliers Oldpeak : 4, 5, -2.6, 4, 4, 4 -> total 16"
```

```
# Imputamos los outliers
for (gender_type in gender_type_list) {
  # Obtenemos los valores por género
  set_by_gender <- heart_df[heart_df$Sex == gender_type, ]
  # KNN imputation sobre las columnas deseadas
  set_imputed <- kNN(set_by_gender, variable=col_names, imp_var=FALSE)
  # Visualizamos los valores antes imputar
  set_by_gender_simplified <- set_by_gender[col_names]

  if (!identical(set_by_gender_simplified, set_imputed)){
    caption_msg <- paste("Valores para el género", gender_type, "antes de imputar:")
    print(kable(summary(set_by_gender_simplified[complete.cases(set_by_gender_simplified), ]), caption=
    # Visualizamos los valores después de imputar
    caption_msg <- paste("Valores para el género", gender_type, "después de imputar:")
    print(kable(summary(set_imputed[col_names]), caption=caption_msg, format="markdown"))
    # Actualizo mi conjunto con los nuevos valores
    heart_df[heart_df$Sex == gender_type, ] <- set_imputed
  } else {
```

```
    print("El dataset antes y después de ser imputado es exactamente igual")
  }
}
```

```
##
##
## Table: Valores para el género M antes de imputar:
##
## |   |  RestingBP   | Cholesterol  |    MaxHR     |   Oldpeak      |
## |:--|:-------------|:-------------|:-------------|:---------------|
## |   |Min.   : 94.0 |Min.   :139.0 |Min.   : 67.0 |Min.   :-2.0000 |
## |   |1st Qu.:120.0 |1st Qu.:215.0 |1st Qu.:117.0 |1st Qu.: 0.0000 |
## |   |Median :130.0 |Median :244.0 |Median :135.0 |Median : 0.7000 |
## |   |Mean   :131.2 |Mean   :238.8 |Mean   :134.9 |Mean   : 0.8956 |
## |   |3rd Qu.:140.0 |3rd Qu.:257.2 |3rd Qu.:154.0 |3rd Qu.: 1.6000 |
## |   |Max.   :170.0 |Max.   :342.0 |Max.   :202.0 |Max.   : 3.7000 |
##
##
## Table: Valores para el género M después de imputar:
##
## |   |  RestingBP   | Cholesterol  |    MaxHR     |   Oldpeak      |
## |:--|:-------------|:-------------|:-------------|:---------------|
## |   |Min.   : 92.0 |Min.   :139.0 |Min.   : 67.0 |Min.   :-2.0000 |
## |   |1st Qu.:120.0 |1st Qu.:216.0 |1st Qu.:117.0 |1st Qu.: 0.0000 |
## |   |Median :130.0 |Median :244.0 |Median :134.0 |Median : 0.8000 |
## |   |Mean   :131.5 |Mean   :239.3 |Mean   :134.5 |Mean   : 0.9026 |
## |   |3rd Qu.:140.0 |3rd Qu.:258.0 |3rd Qu.:152.0 |3rd Qu.: 1.6000 |
## |   |Max.   :170.0 |Max.   :342.0 |Max.   :202.0 |Max.   : 3.7000 |
##
##
## Table: Valores para el género F antes de imputar:
##
## |   |  RestingBP   | Cholesterol |    MaxHR     |   Oldpeak     |
## |:--|:-------------|:------------|:-------------|:--------------|
## |   |Min.   : 94.0 |Min.   :141  |Min.   : 90.0 |Min.   :0.0000 |
## |   |1st Qu.:120.0 |1st Qu.:210  |1st Qu.:130.0 |1st Qu.:0.0000 |
## |   |Median :130.0 |Median :244  |Median :150.0 |Median :0.0000 |
## |   |Mean   :129.2 |Mean   :244  |Mean   :146.2 |Mean   :0.5711 |
## |   |3rd Qu.:140.0 |3rd Qu.:274  |3rd Qu.:163.0 |3rd Qu.:1.0000 |
## |   |Max.   :170.0 |Max.   :344  |Max.   :192.0 |Max.   :3.6000 |
##
##
## Table: Valores para el género F después de imputar:
##
## |   |  RestingBP | Cholesterol  |    MaxHR     |   Oldpeak     |
## |:--|:-----------|:-------------|:-------------|:--------------|
## |   |Min.   : 94 |Min.   :141.0 |Min.   : 90.0 |Min.   :0.0000 |
## |   |1st Qu.:120 |1st Qu.:211.0 |1st Qu.:130.0 |1st Qu.:0.0000 |
## |   |Median :130 |Median :244.0 |Median :150.0 |Median :0.0000 |
## |   |Mean   :130 |Mean   :245.8 |Mean   :146.1 |Mean   :0.6078 |
## |   |3rd Qu.:140 |3rd Qu.:276.0 |3rd Qu.:163.0 |3rd Qu.:1.0000 |
## |   |Max.   :170 |Max.   :344.0 |Max.   :192.0 |Max.   :3.6000 |
```

4. Análisis de los datos

4.1 Selección de los grupos de datos a analizar

Para analizar los datos definimos dos agrupaciones para nuestro dataset. La primera agrupación es el sexo que nos permitirá identificar que genero tiene una incidencia más elevada de afecciones cardiovasculares. La segunda agrupación está determinado por tramos de edades en donde se busca resaltar que el riesgo de padecer una ECV aumenta progresivamente con la edad, siendo más frecuente entre las personas mayores de 65 años[4].

```
# Agrupación por genero
heart_df.male <- heart_df[heart_df$Sex == "M",]
heart_df.female <- heart_df[heart_df$Sex == "F",]

# Agrupación por rango de edad
heart_df.age_group1 <- heart_df[heart_df$Age <= 35, ]
heart_df.age_group2 <- heart_df[heart_df$Age > 35 & heart_df$Age <= 45, ]
heart_df.age_group3 <- heart_df[heart_df$Age > 45 & heart_df$Age <= 55, ]
heart_df.age_group4 <- heart_df[heart_df$Age > 55 & heart_df$Age <= 65, ]
heart_df.age_group5 <- heart_df[heart_df$Age > 65, ]
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Para justificar que las variables tienen una distribución normal, aplicamos el contraste de normalidad de Lilliefors de la librería nortest.

```
alpha = 0.05
col_to_exclude <- c("FastingBS", "HeartDisease")

for (column in colnames(heart_df)) {
  if ((is.integer(heart_df[[column]]) | is.double(heart_df[[column]])) & !(column %in% col_to_exclude))
    p_value <- lillie.test(heart_df[[column]])$p.value
    if (p_value > alpha){
      print(paste(column, "sigue una distribución normal."))
    } else {
      print(paste(column, "No sigue una distribución normal."))
    }
  }
}
```

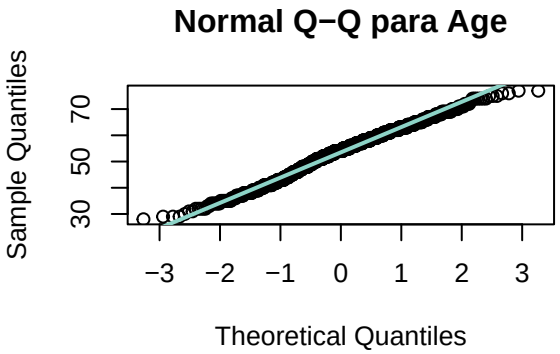
```
## [1] "Age No sigue una distribución normal."
## [1] "RestingBP No sigue una distribución normal."
## [1] "Cholesterol No sigue una distribución normal."
## [1] "MaxHR No sigue una distribución normal."
## [1] "Oldpeak No sigue una distribución normal."
```

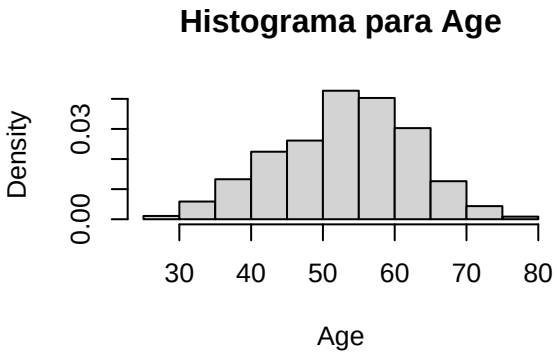
Las pruebas anteriores nos indican que ninguna de las variables tiene un p-value menor al nivel de significancia alfa y por lo tanto no podemos rechazar la hipótesis nula, en otras palabras, las variables no siguen una distribución normal.

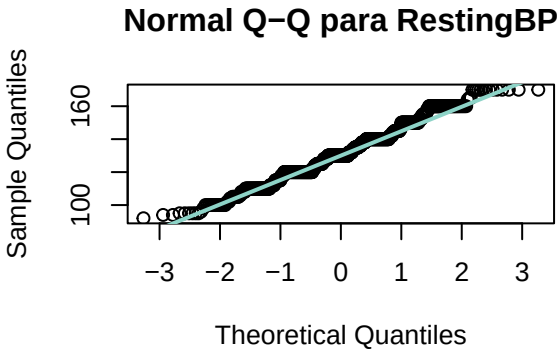
Para contrastar si las variables siguen una distribución normal gráficamente, visualizamos el diagrama Q-Q y el histograma.

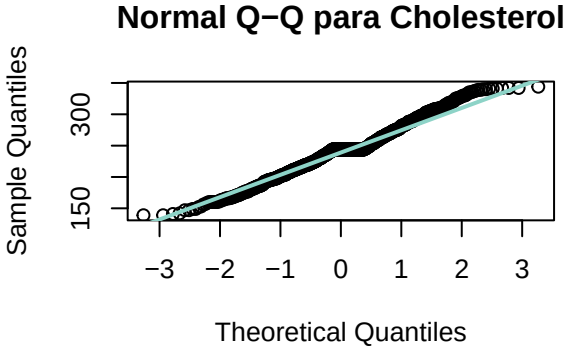
```
par(mfrow=c(2,2))
for (column in colnames(heart_df)) {
  if ((is.integer(heart_df[[column]]) | is.double(heart_df[[column]])) & !(column %in% col_to_exclude))
    qqnorm(heart_df[[column]], main=paste("Normal Q-Q para", column))
    qqline(heart_df[[column]], col=palette, lwd=2)
```

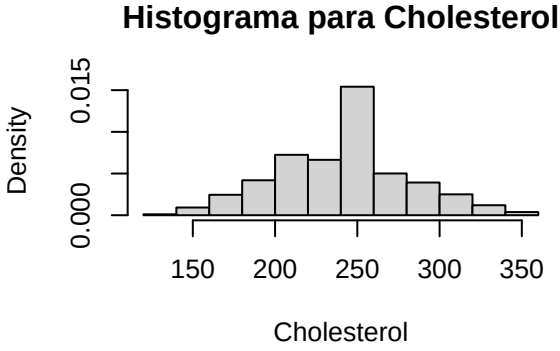
```
    hist(heart_df[[column]], main=paste("Histograma para", column), xlab=column, freq=FALSE)
  }
}
```

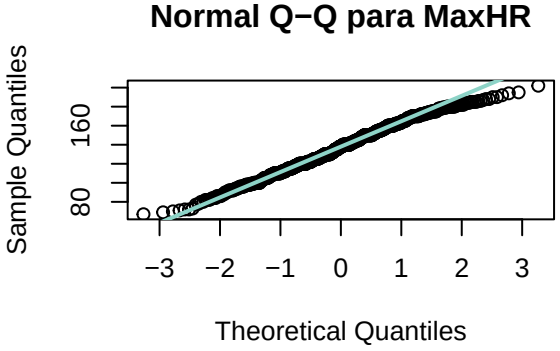


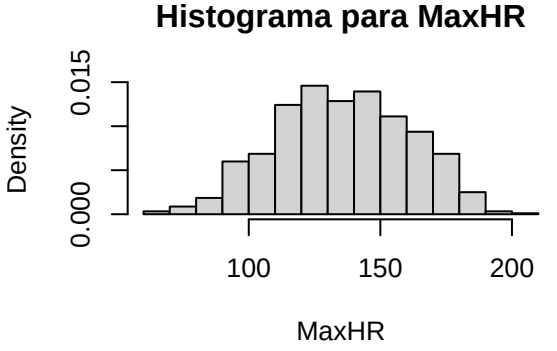


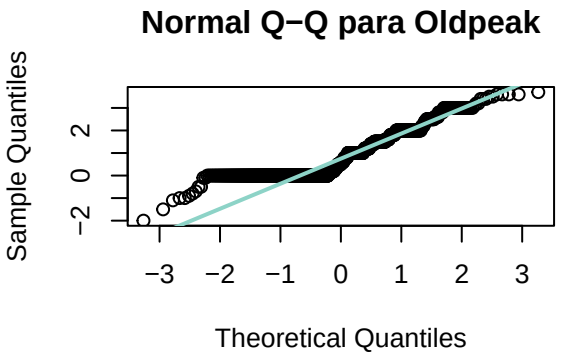


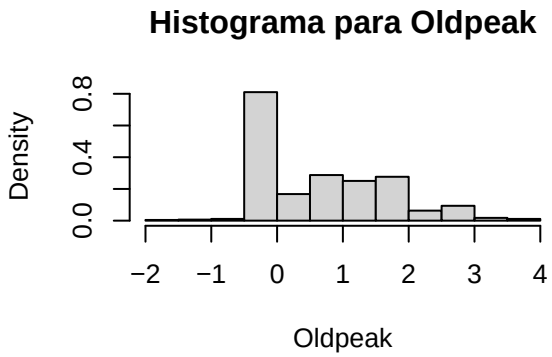












Aunque las variables no pase el test de normalidad de Lilliefors, podemos suponer la normalidad de los datos basándonos en el teorema del limite centrar al tener más de 30 observaciones. También podríamos aplicar una transformación a los datos (por ejemplo una transformación logarítmica) y comprobaríamos tanto con el test de normalidad como con el gráfico Q-Q que se sigue una distribución normal.

A continuación, realizamos el test de homocedasticidad para saber si las varianzas son iguales o diferentes. También se podría asumir normalidad porque las dos muestras tienen un tamaño suficientemente grande y, por lo tanto, se aplicaría el teorema del límite central. En este caso, analizamos la homogeneidad en cuanto a la segmentación por género.

```
for (column in colnames(heart_df)) {
  if ((is.integer(heart_df[[column]]) | is.double(heart_df[[column]])) & !(column %in% col_to_exclude))
    p_value <- var.test(heart_df.female[[column]], heart_df.male[[column]])$p.value
    if (p_value < alpha){
      print(paste("P-value de", round(p_value, 4), "Las varianzas son iguales para ", column))
    } else {
      print(paste("P-value de", round(p_value, 4), "Las varianzas No son iguales para ", column))
    }
  }
}
```

```
## [1] "P-value de 0.8079 Las varianzas No son iguales para  Age"
## [1] "P-value de 0.9363 Las varianzas No son iguales para  RestingBP"
## [1] "P-value de 0.0015 Las varianzas son iguales para  Cholesterol"
## [1] "P-value de 0.0202 Las varianzas son iguales para  MaxHR"
## [1] "P-value de 4e-04 Las varianzas son iguales para  Oldpeak"
```

En algunos casos el p-value es menor al valor de significancia por lo tanto, podemos asumir que las varianzas son iguales, entonces procedemos a aplicar algún test de contraste sobre las variables y ambos géneros, por ejemplo un test sobre la media de dos poblaciones independientes con varianza desconocida iguales.

4.3 Aplicación de pruebas estadísticas

4.3.1 Análisis de correlaciones: Primero comenzamos analizando la correlación entre las distintas características y la posibilidad de tener un ECV, queremos identificar cuáles de ellas tiene un mayor peso a la hora existir la patología.

```
generate_cor_table <- function(table, caption) {
  kable(round(table, 4), caption=caption, format="markdown")
}
# Creamos las columnas de interes
filter_disease_col <- c("HeartDisease")
filter_charact_col <- c("Age", "RestingBP", "Cholesterol", "MaxHR", "Oldpeak")

generate_cor_table(cor(heart_df[filter_disease_col], heart_df[filter_charact_col]), "Matriz de correlac
```

Table 1: Matriz de correlación características y Existencia ECV

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
HeartDisease	0.282	0.1163	0.1249	-0.3994	0.4112

A partir de los resultados anteriores obtenemos que existe una mayor relación de padecer una ECV a partir de la frecuencia cardiaca (MaxHR), la edad (Age) y la depresión ST (Oldpeak). No obstante, como hemos visto que algunos de nuestros datos no siguen una distribución normal, realizaremos una segunda comprobación de la correlación de Spearman para validar dichos resultados.

```
# Creamos una matriz para guardar los datos
corr_matrix <- matrix(nc=2, nr=0)
colnames(corr_matrix) <- c("Estimate", "P-value")
# Calculamos los coeficientes de correlación con respecto a HeartDisease
for (column in filter_charact_col){
  spearman_test <- cor.test(heart_df[[column]], heart_df[[filter_disease_col[1]]], method="spearman")
  corr_coef <- spearman_test$estimate
  p_value <- spearman_test$p.value
  # Añadimos valores a la matriz
  pair <- matrix(ncol=2, nrow=1)
  pair[1][1] <- corr_coef
  pair[2][1] <- p_value
  corr_matrix <- rbind(corr_matrix, pair)
  rownames(corr_matrix)[nrow(corr_matrix)] <- column
}

corr_matrix
```

```
##                Estimate      P-value
## Age           0.2895757 3.419034e-19
## RestingBP     0.1186893 3.134114e-04
## Cholesterol   0.1424165 1.482828e-05
## MaxHR        -0.4038065 2.549063e-37
## Oldpeak       0.4152573 1.421685e-39
```

Podemos contrastar que las características con mayor peso son **MaxHR, Oldpeak y Age** con buen nivel de significancia p-value.

4.3.2 Contraste de hipótesis: La segunda prueba que realizaremos es un contraste de hipótesis sobre dos muestras para determinar si la opinión generaliza de que las mujeres suelen padecer más ECV que los hombres es cierto o no. La pregunta que realizaremos sobre los datos es:

- ¿La proporción de personas que padecen una ECV es mayor para hombres que para mujeres?

Así pues planteamos el siguiente contraste de hipótesis de dos muestras sobre la diferencia de dos proporciones (z-test for independent proportions). Por lo tanto, se trata de un contraste de dos variables sobre la proporción de la muestra y unilateral

$$H_0 : \rho_H - \rho_M = 0$$
$$H_1 : \rho_H - \rho_M > 0$$

Donde ρ es la proporción de personas que sufren una ECV.

```r
# Calculamos el tamaño de la muestra
n1 <- nrow(heart_df.male)
n2 <- nrow(heart_df.female)
# Calculamos las probabilidad de sufrir ECV
p1 <- sum(heart_df.male$HeartDisease==1) / n1
p2 <- sum(heart_df.female$HeartDisease==1) / n2
# Contrate de hipótesis usando prop.test
success <- c(p1*n1, p2*n2)
n <- c(n1, n2)
prop.test(success, n, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of n
## X-squared = 85.646, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.3129991 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.6317241 0.2590674
```

Concluimos que el p-value es menor que el nivel de significancia alfa (0.05), estamos en la zona de no aceptación de la hipótesis nula. Por tanto, podemos afirmar que los hombres sufren más ECV que las mujeres con un nivel de confianza del 95%. Concluimos que el porcentaje de personas que sufren ECV es mayor entre hombres que mujeres.

4.3.3 Modelo de regresión lineal: El objetivo de este apartado es establecer el modelo de regresión lineal que mejor se adapta a nuestras variables para poder predecir sobre la existencia de enfermedades cardíacas dadas ciertas características. Crearemos el modelos utilizando regresores tanto cuantitativos como cualitativos, pero para obtener el modelo más eficiente primero crearemos varios modelos utilizando las características más correlacionadas con la variable "HeartDisease" que hemos obtenido anteriormente, para después elegir el mejor modelo empleando como criterio el coeficiente de determinación (R^2).

```r
# Generación de modelos
modelo1 <- lm(HeartDisease ~ Age, data=heart_df)
modelo2 <- lm(HeartDisease ~ Age + MaxHR, data=heart_df)
modelo3 <- lm(HeartDisease ~ Age + Oldpeak, data=heart_df)
modelo4 <- lm(HeartDisease ~ MaxHR, data=heart_df)
modelo5 <- lm(HeartDisease ~ MaxHR + Oldpeak, data=heart_df)
modelo6 <- lm(HeartDisease ~ Oldpeak, data=heart_df)
modelo7 <- lm(HeartDisease ~ Age + MaxHR + Oldpeak, data=heart_df)

coef_table <- matrix(c(1, summary(modelo1)$r.squared,
                       2, summary(modelo2)$r.squared,
                       3, summary(modelo3)$r.squared,
                       4, summary(modelo4)$r.squared,
                       5, summary(modelo5)$r.squared,
                       6, summary(modelo6)$r.squared,
                       7, summary(modelo7)$r.squared), ncol=2, byrow=TRUE)
colnames(coef_table) <- c("Num Modelo", "R2")
coef_table
```

```
##      Num Modelo         R2
## [1,]          1 0.07954572
## [2,]          2 0.17884510
## [3,]          3 0.20175226
## [4,]          4 0.15951064
## [5,]          5 0.27979983
## [6,]          6 0.16906682
## [7,]          7 0.28420795
```

Podemos apreciar que el séptimo modelo tiene un mayor coeficiente de determinación y por lo tanto sería el más adecuado.No obstante, el quinto modelo también tiene un coeficiente de determinación similar y por lo tanto nos hace pensar que el peso que tiene la variable Age sobre el modelo es ligeramente significativo.

Utilizamos le modelo para predecir la existencia de enfermedades cardíacas.

```r
# Ejemplo de como emplear el modelo
to_predict <- data.frame(Age=45, MaxHR=130, Oldpeak=1)
round(predict(modelo7, to_predict))
```

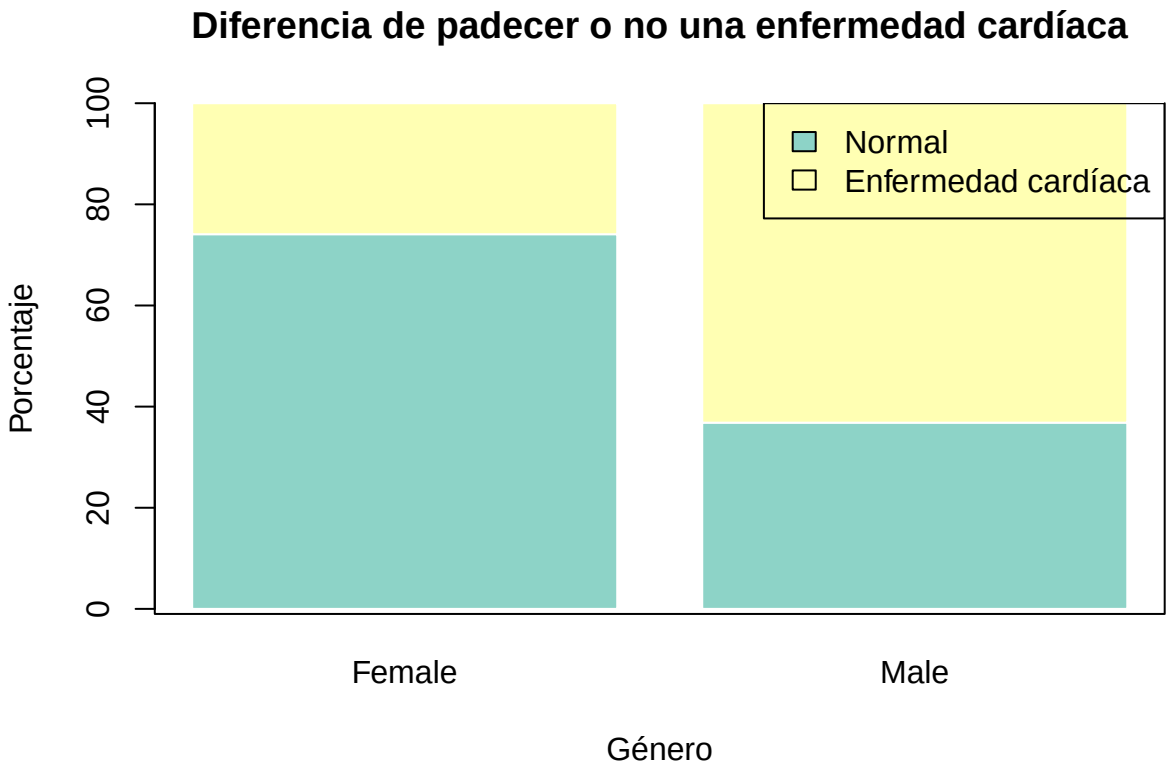
```
## 1
## 1
```

Con 45 años, frecuencia cardíaca máxima de 130 y una depresión ST de 1 existe la posibilidad de sufrir una enfermedad cardíaca.

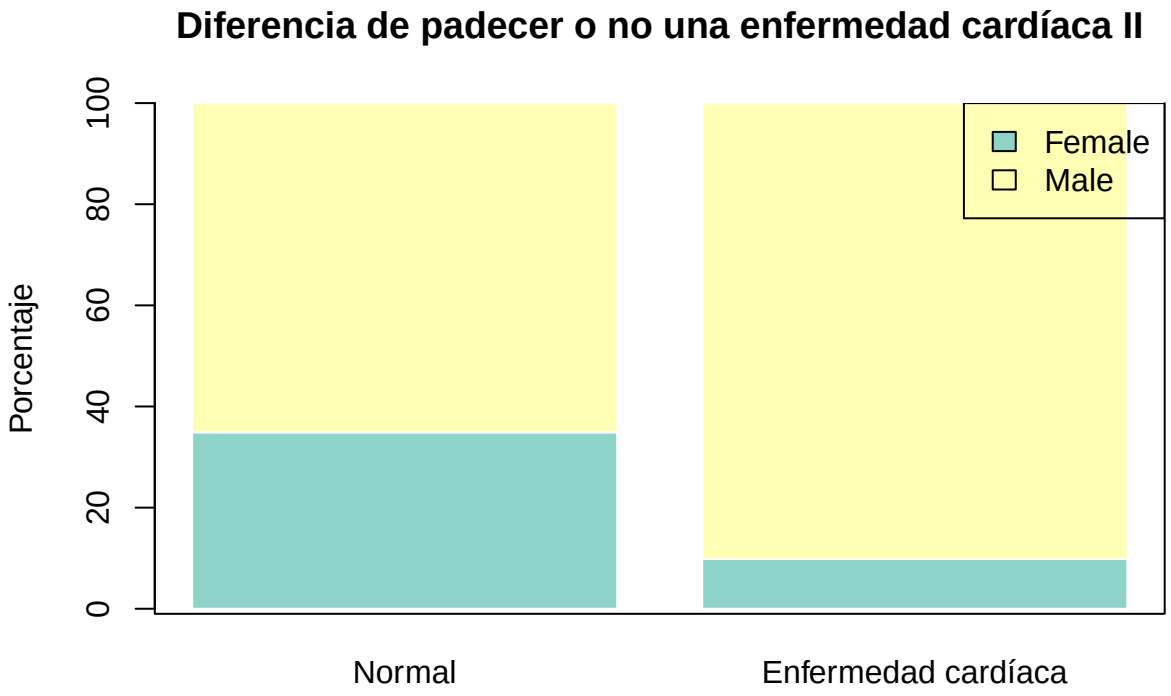
5. Representación de los resultados a partir de tablas y gráficas.

Una de las cuestiones que se han presentado en este trabajo esta relacionada con la comparativa del género en cuanto a padecer una enfermedad cardíaca. En los siguientes gráficos queremos visualizar dicha comparativa.

```
cols <- c("HeartDisease", "Sex")
# Usamos un dataframe temporal para extraer los datos de las columnas deseadas y lo convertimos en tabl
table_df <- table(heart_df[, cols])
colnames(table_df) <- c("Female","Male")
rownames(table_df) <- c("Normal","Enfermedad cardíaca")
# Transformamos los datos a porcentaje
table_percentage <- apply(table_df, 2, function(x){ x*100/sum(x, na.rm=T) })
# Visualizamos un stacked barplot por género
barplot(table_percentage, col=palette , border="white", main='Diferencia de padecer o no una enfermedad
box(bty="U")
legend('topright', fill=palette, legend=rownames(table_percentage))
```




```
# Visualizamos invirtiendo los datos
table_t_df <- t(table_df)
table_percentage <- apply(table_t_df, 2, function(x){ x*100/sum(x, na.rm=T) })
# Visualizamos un stacked barplot por existencia de enfermedad o no
barplot(table_percentage, col=palette , border="white", main='Diferencia de padecer o no una enfermedad
box(bty="U")
legend('topright', fill=palette, legend=rownames(table_percentage))
```

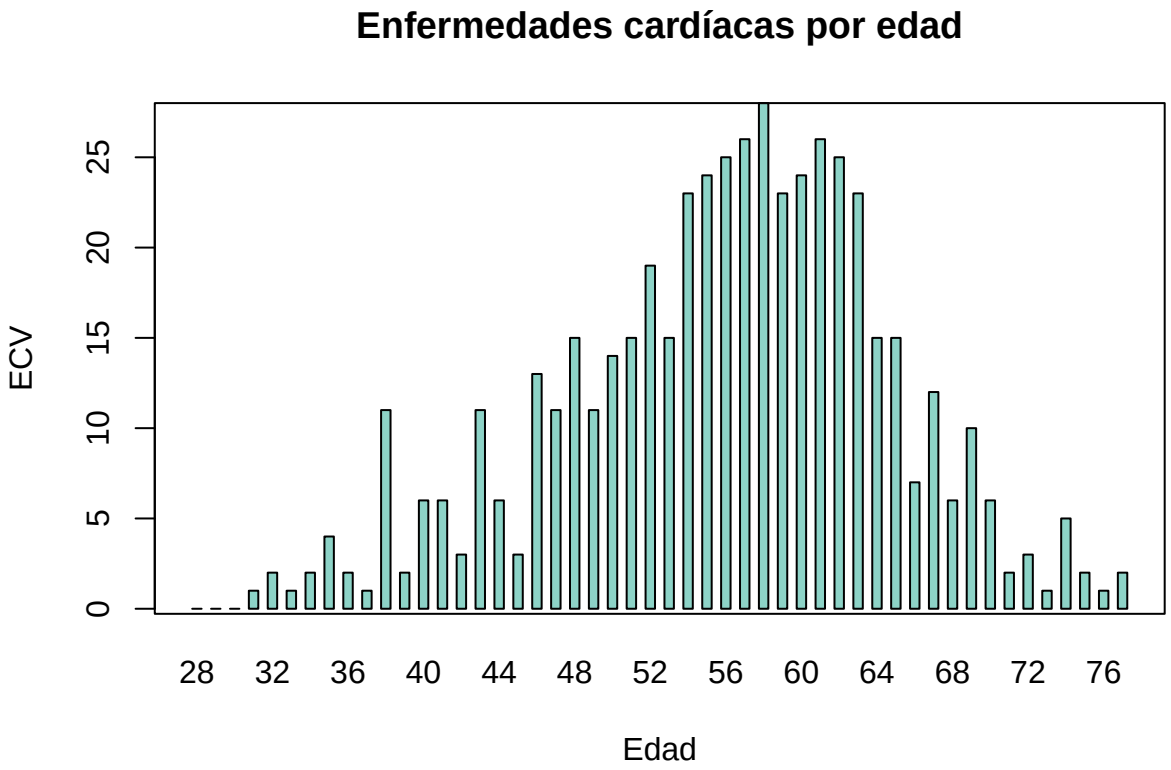


Como podemos ver en las gráficas y alineado con las conclusiones de nuestro contraste de hipótesis, podemos apreciar que en nuestro conjunto de datos es más habitual que los hombres sufran una enfermedad cardíaca que las mujeres. También podemos pensar que existe un desequilibrio en los datos de entrada pero dicha comprobación queda fuera de este trabajo.

Por lo tanto, esta gráfica nos da una idea significativa de la proporción de padecer una enfermedad cardíaca entre hombres y mujeres.

Otra de las cuestiones interesantes de estudio era conocer si la probabilidad de padecer una enfermedad cardíaca aumenta con el paso de los años. Para ello hemos creado la siguiente gráfica.

```
col_list <- c("HeartDisease", "Age")
# Usamos un dataframe temporal para extraer los datos de las columnas deseadas
test_disease <- heart_df[, col_list]
# Compartimos los mismos colores entre visualizaciones
colours = c(palette[1], palette[2], palette[3])
# Agrego los datos por edad y los transformo para el barplot
# Note que transpongo los datos para no añadir la edad como una fila extra
temporal_aggreagation <- aggregate(.~Age, test_disease, sum)
tmp_matrix <- t(as.matrix(temporal_aggreagation[-1]))
colnames(tmp_matrix) <- temporal_aggreagation$Age
# Visualizamos las ECV por edad
barplot(tmp_matrix, main='Enfermedades cardíacas por edad', ylab='ECV', xlab='Edad', beside=TRUE, col=pa
box()
```



Como podemos apreciar existe un incremento de las enfermedades cardíacas con el paso de los años, teniendo los picos máximos entre las edades de 55 y 65 años. a partir de los 65 vemos un decremento de los casos ya bien porque existe menos probabilidad de sufrir la enfermedad si no la has sufrido en los años anteriores o bien porque no hay suficiente representación de los casos de esas edades en nuestro conjunto de datos y/o el volumen de la población es mucho menor (de esta manera esta representado en nuestros datos). No obstante y teniendo en cuenta los anteriores análisis estadísticos, podemos asegurar que la edad es un factor de riesgo para sufrir enfermedades cardiovasculares.

6. Resolución del problema

En este proyecto hemos trabajado sobre un conjunto de datos para predecir posibles eventos de enfermedades cardíacas. Como hemos comentado es una de las principales causas de muerte a nivel mundial, cuatro de cada cinco muertes por ECV se deben a ataques cardíacos y accidentes cardiovasculares, y un tercio de estas muertes ocurren prematuramente en personas menores de 70 años. Por esto queríamos saber si la edad y el género eran factores de riesgo y/o determinantes en la aparición de problemas cardíacos.

Hemos podido observar que los problemas cardiovasculares ocurren en mayor proporción sobre la población masculina, y la incidencia se ve incrementada con el paso de los años, obteniendo unos picos máximos en edades comprendidas entre 55 y 65 años. También hemos visto mayor correlación para la existencia de enfermedades cardiovasculares con variables como la frecuencia cardíaca máxima y el segmento del electrocardiograma ST, siendo estos los factores más determinantes.

Cabe destacar, que hay otros factores que no son tan determinantes, como son el colesterol y la presión arterial en reposo, pero pueden tener mayor influencia poniéndolos en un contexto de un profesional medico y uniéndolos al resto de datos de este estudio.

Podemos concluir que este trabajo ha sido un ejercicio a nivel estadístico en donde es mas difícil tener un resultado final sobre los factores que influyen en la aparición de enfermedades cardíacas, pero ademas de que hemos podido responder a las preguntas planteadas, los gráficos, datos extraídos y las conclusiones de los análisis son relevantes para saber cuales son los marcadores más peligrosos y tener una conciencia social sobre lo que hay que evitar en cuanto a las enfermedades cardiovasculares.

7. Código y archivo final

El código de este proyecto se puede encontrar en el siguiente enlace conjuntamente con el conjunto de datos utilizado y el fichero final con los datos preprocesados en un archivo llamado “heart_output.csv”.

- https://github.com/damaro05/Practica2_Tipologia_y_ciclo_de_vida_de_los_datos

Referencias

1. https://www.kaggle.com/fedesoriano/heart-failure-prediction
2. https://www.my-ekg.com/como-leer-ekg/segmento-st.html
3. https://www.ncbi.nlm.nih.gov/books/NBK459364/
4. https://es.statista.com/estadisticas/576912/prevalencia-de-las-enfermedades-cardiovasculares-por-grupo-de-edad-espana/
5. Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.
6. Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
7. Squire, Megan (2015). Clean Data. Packt Publishing Ltd.

Contribuciones

Contribuciones	**Firmas**
Investigación previa	Sebastian Maya
Redacción de las respuestas	Sebastian Maya
Desarrollo del código	Sebastian Maya