Universitat Oberta de Catalunya

uoc.edu

Tipología y ciclo de vida de los datos

¿Cómo realizar la limpieza y análisis de datos?

Autores:

1) William Gabriel Granda Betancourt y 2) Oscar Augusto Díaz Triana.

Mayo 2022

1) Descripción del dataset

Vamos a trabajar con el conjunto de datos Heart Attack Analysis & Prediction dataset, el cual se encuentra disponible en Kaggle.

Problema a resolver:

El conjunto de datos sobre pacientes y las posibilidades de sufrir un infarto es de gran importancia debido a su relevancia para la salud cardiovascular. Al analizar este conjunto de datos, el objetivo es identificar las características o factores que influyen en la propensión de una persona a sufrir un ataque al corazón. (¿Cuáles son las características que influyen en la propensión de una persona a sufrir un ataque cardíaco?)

Mediante la implementación de un modelo de clasificación, es posible predecir con antelación si una persona es propensa a sufrir un ataque cardíaco. Esto permitiría a los especialistas médicos intervenir de manera oportuna y tomar medidas preventivas para reducir el riesgo. Al identificar las características que están fuertemente asociadas con los ataques cardíacos, se podrían desarrollar estrategias y programas de prevención más efectivos para mejorar la salud cardiovascular de la población en riesgo.

Empezaremos haciendo un análisis sobre el número de registros y número de variables que contiene el dataset.

El conjunto de datos contiene 303 registros u observaciones y 14 variables. Contamos con las siguientes variables:

- **age**: Edad del paciente. Es una variable numérica.
- **sex**: Sexo del paciente.
  - 1 = hombre,
  - 0 = mujer.
- **exng**: Angina producida por ejercicio. Es un variable categórica y toma los siguientes valores:
  - 1, si se posee la molestia
  - 0, si no.
- **caa**: Número de vasos principales coloreados por fluroscopia. Es una variable numérica discreta y toma los siguientes valores: (0-3).
- **cp**: Tipo de dolor toráxico. Es una variable categórica que toma los siguientes valores:
  - 1: angina típica,
  - 2: angina atípica,
  - 3: dolor no anginoso,
  - 0: asintomático.
- **trtbps**: Presión arterial en reposo (en $mmHg$).
- **chol**: Colesterol en mg/dl obtenido a través del sensor BMI.
- **fbs**: Azúcar en sangre en ayunas $> 120mg/dl$. Es una variable categórica, que toma los siguientes valores:
  - 1 = verdadero,
  - 0 = falso.
- **restecg**: Resultados electrocardiográficos en reposo. Es un variable categórica que toma los siguientes valores:

  - 1: normal,
  - 2: Indica si el paciente tiene anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del $ST > 0.05mV$)
  - 0: Indica si el paciente muestra hipertrofia ventricular izquierda probable o definitiva, según los criterios de Estes.
- **thalachh**: Frecuencia cardíaca máxima alcanzada. Es una variable numérica continua.
- **oldpeak**: Depresión del ST inducida por el ejercicio en relación con el reposo. Es una variable numérica continua.
- **slp**: Pendiente de un segmento de electrocardiograma. Es una variable categórica que toma los siguientes valores:
  - 0: pendiente descendente,
  - 1: plano,
  - 2: pendiente ascendente.
- **thall**: Indica los resultados de prueba de esfuerzo con talio, la muestra qué tan bien fluye la sangre hacia el músculo cardíaco, tanto en reposo como en actividad. Es una variable categórica y toma los siguientes valores:
  - 1: defecto fijo,
  - 2: normal,
  - 3: defecto reversible.
- **output**: Es nuestra variable binaria a predecir, toma los siguientes valores:
  - 1 = más posibilidades de ataque al corazón,
  - 0 = menos posibilidades de ataque al corazón.

2) Integración y selección

La integración y selección de datos permite utilizar información relevante y de calidad para un análisis preciso y significativo, mejorando la toma de decisiones y obteniendo resultados confiables.

Tras analizar el diccionario de datos del dataset *Heart Attack Analysis & Prediction*, podemos notar que contamos con las variables necesarias para cumplir con el objetivo de la práctica, por lo que, en este caso no vamos a emplear datasets adicionales.

Con respecto a la selección de variables, a lo largo del desarrollo de la PRA vamos a presentar pruebas estadísticas que nos permitan identificar las variables que influyen en que una persona sea más propensa a sufrir un ataque al corazón. Algunas variables de interés podrían ser la *edad* y el *sexo* del paciente, los niveles de colesterol y la presión arterial.

3 Limpieza de datos

La limpieza de datos es crucial para eliminar errores, valores faltantes y ruido, lo que mejora la calidad de los datos y evita sesgos en el análisis, proporcionando resultados más precisos y confiables.

Vamos a emplear todas las variables numéricas y categóricas con la intención de determinar que variables influyen en que una persona sea más propensa a sufrir un ataque al corazón.

Al aplicar la función str()notamos que ciertas variables categóricas fueron identificadas como numéricas, por ello, vamos a transformarlas al tipo factor.

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

En primer lugar, vamos a determinar si existen valores perdidos para cada variable:

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

De este modo podemos observar que no existen valores atípicos. También, identificamos si existen valores duplicados:

```
data[duplicated(data),]
```

```
##     age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
## 165  38   1  2    138  175   0       1      173    0       0   2   4     2
##     output
## 165      1
```

Con lo cual, existen dos filas duplicadas, la función unique() nos permite resolver este problema. Así, nuestro juego de datos posee 302 observaciones únicas.

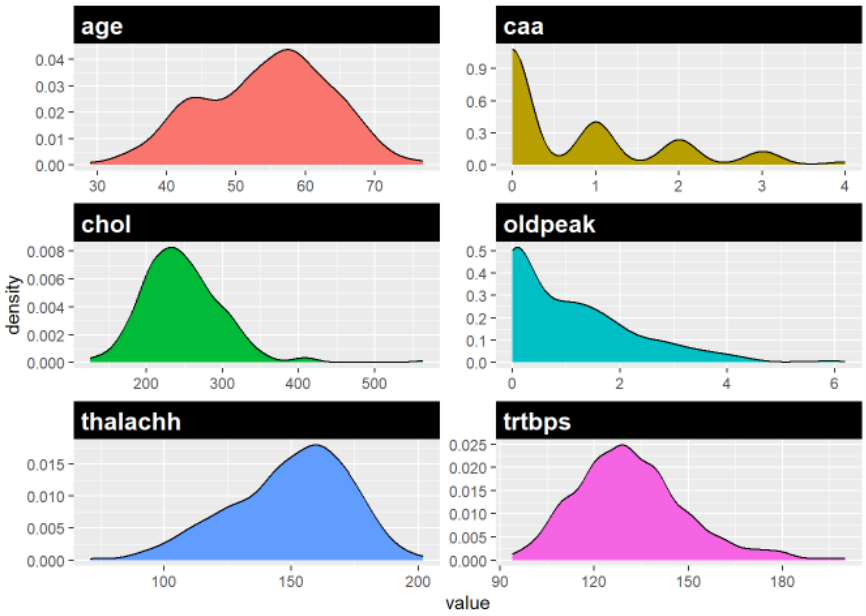
3.2 Identifica y gestiona los valores extremos.

Ahora, vamos a analizar cada una variable, para ello, presentamos sus estadísticos:

```
      age        sex     cp          trtbps           chol       fbs
 Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:257
 1st Qu.:48.00   1:206   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
 Median :55.50           2: 86   Median :130.0   Median :240.5
 Mean   :54.42           3: 23   Mean   :131.6   Mean   :246.5
 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.8
 Max.   :77.00                   Max.   :200.0   Max.   :564.0
 restecg    thalachh     exng       oldpeak      slp          caa
 0:147   Min.   : 71.0   0:203   Min.   :0.000   0: 21   Min.   :0.0000
 1:151   1st Qu.:133.2   1: 99   1st Qu.:0.000   1:140   1st Qu.:0.0000
 2:  4   Median :152.5           Median :0.800   2:141   Median :0.0000
         Mean   :149.6           Mean   :1.043           Mean   :0.7185
         3rd Qu.:166.0           3rd Qu.:1.600           3rd Qu.:1.0000
         Max.   :202.0           Max.   :6.200           Max.   :4.0000
 thall   output
 0:  2   0:138
 1: 18   1:164
 2:165
 3:117
```

3.2.1 Variables numéricas

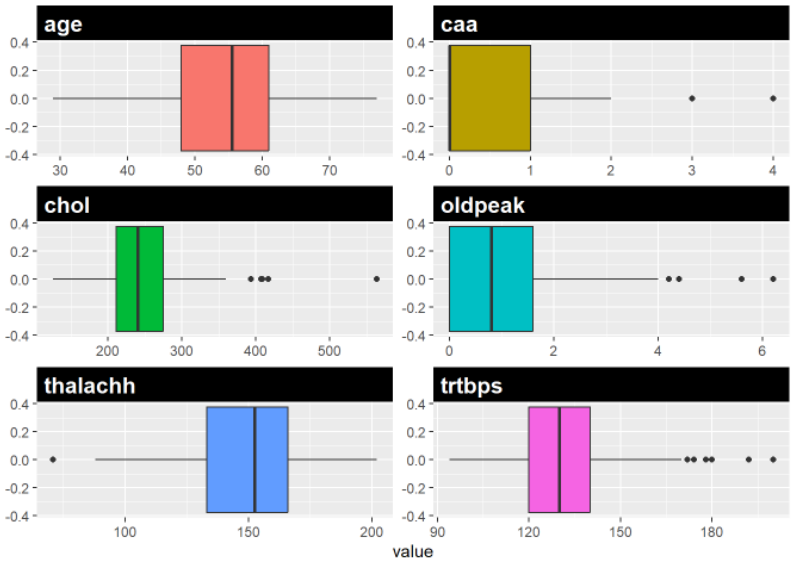
Para complementar esta información vamos a presentar la distribución de las variables:



Se tienen las siguientes observaciones:

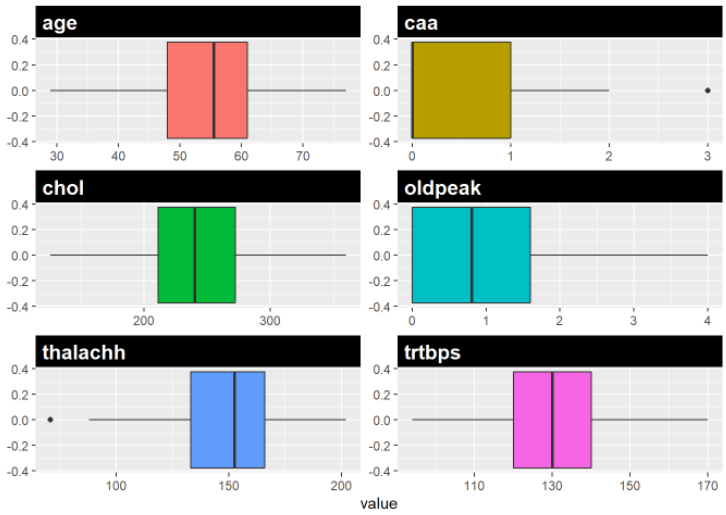
- **age**: La edad promedio de los pacientes es 54 años, la edad mínima es 29 años y la edad máxima es de 77 años. Además, el 75% de los pacientes es menor a 61 años. Con respecto a la distribución de la variable podemos notar que es bimodal, es decir, presenta dos modas.
- **caa**: El valor mínimo de esta variable es 0 y el valor máximo es 4. Como podemos notar la distribución de las variables es sesgada a la izquierda. El 75% de los pacientes ha registrado entre 0 y 1 vasos colorados por fluroscopia.
- **chol**: El valor mínimo del colesterol es 94 y el valor máximo es 200, además, el 75% de los pacientes tiene un colesterol inferior a 274.8. Podemos notar que existen pocos valores en la cola derecha de la distribución los cuales pueden corresponder a valores atípicos.
- **oldpeak**: El valor mínimo de la depresión del segmento ST es 0 y el valor máximo es de 6.2 El 75% de los clientes posee un valor de depresión del segmento ST menor a 1.6. La distribución de esta variable está acumulada a la izquierda.
- **thalachh**: El valor mínimo de a frecuencia cardíaca máxima alcanzada es de 71 y el valor máximo es de 202. Además, el 75% de los pacientes tiene una frecuencia máxima alcanza menor a 166. La distribución de esta variable es sesgada a la derecha.
- **trtbps**: El valor mínimo de la presión arterial es de 94 y el valor máximo es de 200. También podemos afirmar que el 75% de los pacientes tiene una presión arterial menor o igual a 130. Podemos notar que existen pocos valores en la cola derecha de la distribución, los cuales podrían corresponder a valores atípicos.

Ahora, analicemos los valores atípicos, para lo cual emplearemos el diagrama de cajas o boxplot:



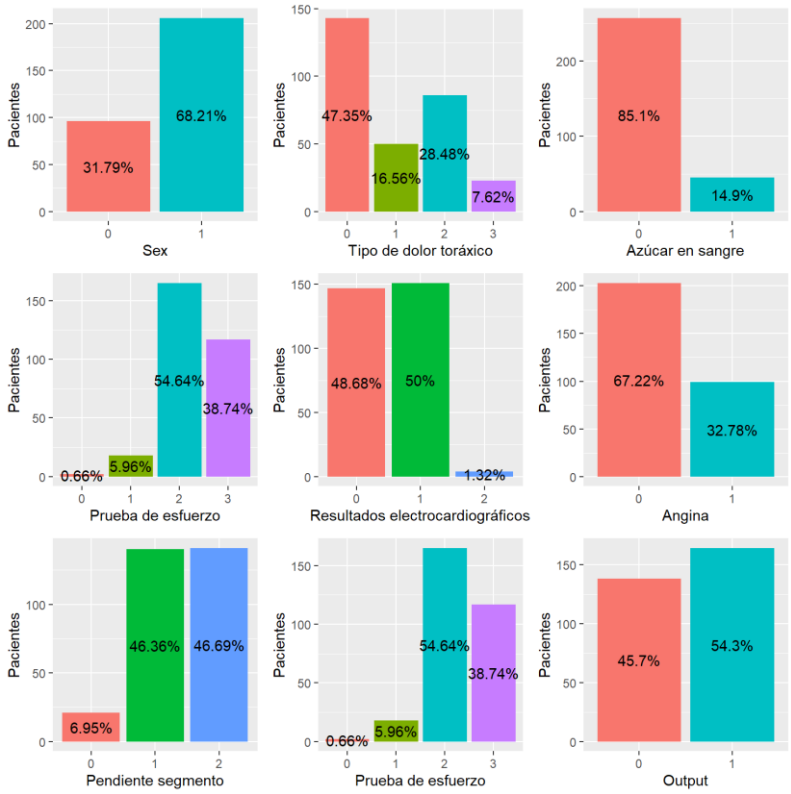
- **caa**: La variable caa posee dos valores atípicos 3 y +, no obstante en la definición de esta variable se especifica que su dominio es (0−3)(0−3), por lo que el único valor atípico es 44. Vamos a reemplazar estos valores atípicos por el valor máximo de esta variable que es 33.
- Para las variables **chol**, **oldpeak** y **trtbps** vamos a reemplazar los valores atípicos por la mediana de cada una de las variables, debido la media está sesgada por los valores atípicos.

De este modo, podemos observar que no existen valores atípicos:



3.2.2 Variables categóricas

Para las variables categóricas vamos a presentar su diagrama de frecuencias:



Se tienen las siguientes observaciones:

- **sex**: El 68.21% de los pacientes son hombres, mientras que el 31.79% son mujeres.
- **cp**: El 43.25% de los pacientes es asintomático, el 16.56% posee angina típica, el 28.48% posee angina atípica y el 7.62% tiene dolor no anginoso.
- **fbs**: El 85.15% de los pacientes no posee azúcar en sangre en ayunas superior a $120mg/dl$ mientras que el 14.85% de los pacientes si posee niveles de azúcar superiores.
- **thall**: El 85.1% de los pacientes no registra azúcar en la sangre mayor a $> 120mg/dl$, mientras que el 14.9% sí.
- **restecg**: El 1.32% de los pacientes posee resultados electrocardiográficos normales, el 48.68% de los pacientes posee anomalías en la onda ST-T y finalmente, el 50% de los pacientes muestra hipertrofia ventricular izquierda probable o definitiva.
- **exng**: El 67.22% de los pacientes no posee el dolor de angina, mientras que el 32.78% sí lo posee.
- **slp**: El 6.95% posee una pendiente descendiente de segmento de electrocardiograma, el 46.36% posee una pendiente plana y el 46.69% posee una pendiente ascendente.
- **thall**: Esta variable solamente debería tomar tres valores, por lo que el valor de cero corresponde a un valor perdido, reemplazaremos este valor por 2, que es el valor que más se repite. El 54.94% de los pacientes posee un defecto fijo, el 5.96% de los pacientes dio un resultado normal y el 38.74% tiene un defecto reversible.
- **output**: El 45.7% de los pacientes tiene menos probabilidades de sufrir un ataque al corazón, mientras que el 54.3% de los pacientes tiene más probabilidades de sufrir un ataque al corazón.

4 Análisis de datos

Aplicando el análisis de datos podemos obtener información relevante sobre los factores que pueden influir en los ataques cardíacos. Al explorar y examinar los datos, podemos identificar patrones, relaciones y características relevantes que ayudarán a comprender mejor los riesgos y tomar medidas preventivas adecuadas para la salud cardiovascular de los pacientes, lo que potencialmente puede salvar vidas y mejorar la calidad de vida de las personas.

En este paso del análisis, se seleccionan los grupos de datos que se desean analizar y comparar, en este caso, en función de la variable “output”. Se aplicará un análisis de variables numéricas, específicamente se examinará el comportamiento de las variables age, trtbps, thalachh y oldpeak en relación con la variable output. Se utilizará la técnica de diagramas de cajas para visualizar las diferencias entre los grupos. Se observa que hay diferencias significativas en estas variables con respecto a output. Para confirmar estas hipótesis, se realizarán pruebas estadísticas, incluyendo la transformación de Box-Cox en las variables no normalmente distribuidas.

Además, se realizará un análisis de las variables categóricas sex, cp, thall, restecg, exng, slp y thall para determinar si existen diferencias significativas entre los grupos definidos por output. Se presentarán diagramas de frecuencias para visualizar el comportamiento de estas variables en relación con output y se realizarán pruebas estadísticas, como el test de chi-cuadrado y el análisis de frecuencias relativas, para identificar posibles asociaciones o diferencias entre los grupos.

Previo a aplicar los análisis mencionados, es importar validar las hipótesis de los mismos, tales como normalidad y homogeneidad de las varianzas.

4.1 Correlaciones

En primer lugar, para determinar el tipo de prueba para analizar las correlaciones de las variables numéricas vamos a verificar si nuestras variables tienen o no distribuciones normales. Para ello, emplearemos la prueba de Shapiro-Wilk, la cual contrasta las siguientes hipótesis:

$$\begin{cases} H_0 : \text{Los datos siguen distribución normal.} \\ H_1 : \text{Los datos no siguen una ditribución normal.} \end{cases}$$

Prueba para **age**:

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.98664, p-value = 0.006745
```

Considerando un nivel de confianza $\alpha = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Prueba para **trtbps**:

```
##
##  Shapiro-Wilk normality test
##
## data:  data$trtbps
## W = 0.98394, p-value = 0.001851
```

Considerando un nivel de confianza $\alpha = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Prueba para **chol**:

```
##
##  Shapiro-Wilk normality test
##
## data:  data$chol
## W = 0.99333, p-value = 0.2007
```

Considerando un nivel de confianza $\alpha = 0.05$ no se rechaza la hipótesis nula, y podemos concluir que los datos siguen una distribución normal.

Prueba para **thalachh**:

```
##
##  Shapiro-Wilk normality test
##
## data:  data$thalachh
## W = 0.97679, p-value = 8.268e-05
```

Para un nivel de confianza $\alpha = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Prueba para **oldpeak**:

De igual forma si se considera un nivel de confianza $\alpha = 0.05$ rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Dado que tres de las cuatro variables numéricas no siguen una distribución normal, vamos a emplear la prueba de correlación de Spearman, la cual es una alternativa no paramétrica cuando las variables no son normales.

No observamos fuertes correlaciones positivas o negativas entre las variables, no obstante, vamos a aplicar pruebas de hipótesis para determinar si la correlación es significativamente distinta de cero. Para lo cual, emplearemos la prueba de **Spearman**, la cual contrasta las hipótesis:

$$\begin{cases} H_0 : \text{Las variables } X \text{ y } Z \text{ son independientes} \\ H_1 : \text{Las variables } X \text{ y } Z \text{ no son independientes} \end{cases}$$

Test para **thalachh vs oldpeak:**

```
##
##  Spearman's rank correlation rho
##
## data:  data$thalachh and data$oldpeak
## S = 6506261, p-value = 3.723e-14
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.417316
```

Para un nivel de confianza $\boldsymbol{\alpha} = 0.05$, se rechaza la hipótesis nula y podemos concluir que la correlación entre las variables **thalachh** y **oldpeak** es de -0.42, no obstante, este valor no es elevado.

Test para **thalachh vs age:**

```
##
##  Spearman's rank correlation rho
##
## data:  data$thalachh and data$age
## S = 6396719, p-value = 1.271e-12
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.3934534
```

De igual forma, considerando un nivel de confianza $\boldsymbol{\alpha} = 0.05$, se rechaza la hipótesis nula y podemos concluir que la correlación entre las variables **thalachh** y **age** es de -0.393.

Test para **thalachh vs trtbps:**

```
##
##  Spearman's rank correlation rho
##
## data:  data$thalachh and data$trtbps
## S = 4784894, p-value = 0.4636
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.04233545
```

Con un nivel de confianza $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor > \alpha}$, no se rechaza la hipótesis nula y podemos concluir que las variables **thalachh** y **trtbps** son independientes.

Test para **thalachh vs chol:**

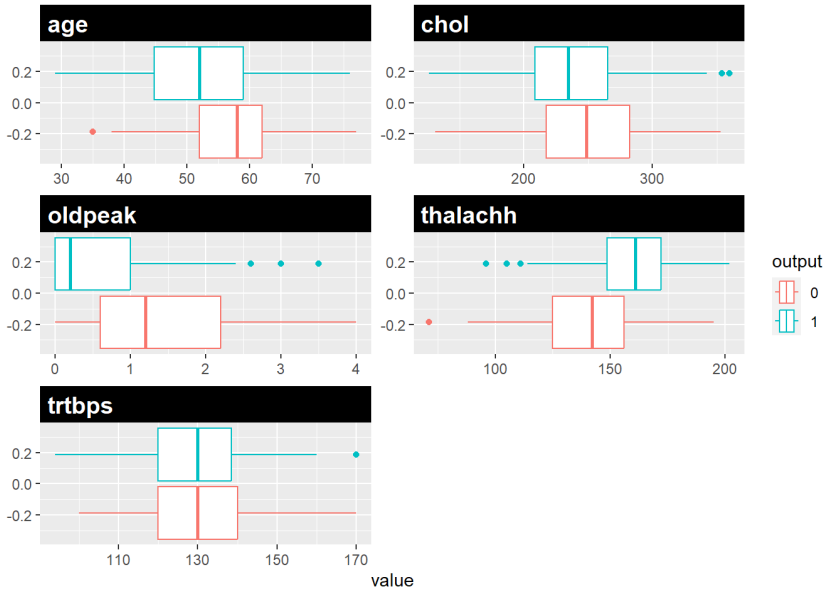
```
##
##  Spearman's rank correlation rho
##
## data:  data$thalachh and data$chol
## S = 4808107, p-value = 0.4119
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.04739214
```

De igual forma, con un nivel de confianza $\boldsymbol{\alpha} = 0.05$ dado que $\boldsymbol{p - valor > \alpha}$, no se rechaza la hipótesis nula y podemos concluir que las variables **thalachh** y **trtbps** son independientes.

4.2 Comparación entre grupos

4.2.1 Variables numéricas

En primer lugar, vamos a estudiar gráficamente el comportamiento de las variables **age**, **trtbps**, **thalachh** y **oldpeak** con respecto a la variable **output**, para lo cual, vamos a diagramas de cajas:



Podemos notar que existen diferencias significativas de las variables **age**, **chol**, **oldpeak** y **thalachh** con respecto a la variable **ouput**. Vamos a comprobar estas hipótesis aplicando pruebas estadísticas.

Para ello, en primer lugar, como pudimos observar anteriormente solamente la variable **chol** sigue una distribución normal por lo que vamos a aplicar la transformación de Box - Cox para el resto de variables y aplicamos la prueba de Shapiro-Wilk a las variables transformadas para comprobar si siguen una distribución normal:

Prueba para **age** con la transformación de Box – Cox**:**

```
##
##  Shapiro-Wilk normality test
##
## data:  dataT$ageT
## W = 0.98848, p-value = 0.01691
```

Considerando un nivel de confianza $\boldsymbol{\alpha} = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Prueba para **trtbps** con la transformación de Box – Cox**:**

```
##
##  Shapiro-Wilk normality test
##
## data:  dataT$trtbpsT
## W = 0.98419, p-value = 0.002079
```

Considerando un nivel de confianza $\boldsymbol{\alpha} = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Prueba para **thalachh** con la transformación de Box – Cox**:**

```
##
##  Shapiro-Wilk normality test
##
## data:  dataT$thalachhT
## W = 0.99174, p-value = 0.0901
```

Para un nivel de confianza $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor > \alpha}$, no se rechaza la hipótesis nula, y podemos concluir que los datos siguen una distribución normal.

Prueba para **oldpeak** con la transformación de Box – Cox**:**

```
##
##  Shapiro-Wilk normality test
##
## data:  dataT$oldpeakT
## W = 0.78649, p-value < 2.2e-16
```

Considerando un nivel de confianza $\boldsymbol{\alpha} = 0.05$ se rechaza la hipótesis nula, y podemos concluir que los datos no siguen una distribución normal.

Por lo cual, solamente vamos a aplicar la prueba de Levene a las variables **chol** y a la variable transformada **thalachh** para verificar la homocedasticidad, debido a que estas variables tienen distribución normal.

Pruebas para **chol**:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.0697 0.3018
##       300
```

Al nivel de significancia $\boldsymbol{\alpha} = 0.05$ no rechazamos la hipótesis nula, es decir, la varianza de la variable **chol** con respecto a la variable **output** es homogénea. Así, podemos aplicar la prueba t de Student:

```
##
##  Welch Two Sample t-test
##
## data:  chol by output
## t = 1.9412, df = 285.7, p-value = 0.05321
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1394118 20.1472768
## sample estimates:
## mean in group 0 mean in group 1
##        248.6594        238.6555
```

Pruebas para **thalachh:**

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.9923   0.32
##       300
```

Al nivel de significancia $\boldsymbol{\alpha} = 0.05$ no rechazamos la hipótesis nula, es decir, las varianzas son homogéneas, por lo cual podemos aplicar la prueba t de Student.

```
##  Welch Two Sample t-test
##
## data:  thalachhT by output
## t = -8.034, df = 285.33, p-value = 2.507e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -3480.782 -2110.851
## sample estimates:
## mean in group 0 mean in group 1
##        9924.245       12720.062
```

En este caso, el $p - valor$ es menor al nivel de significancia $\alpha = 0.05$. Por lo cual, se rechaza la hipótesis nula, es decir, se observan diferencias significativas de la frecuencia máxima alcanzada entre entre las personas menos propensas a sufrir un ataque cardíaco y las personas más propensas a sufrir un ataque cardíaco. De hecho, podemos ver que el promedio estimado del grupo de personas menos propensas es menor al grupo de las personas más propensas.

Para el resto de variables numéricas vamos a trabajar con las variables originales, dado que las trasformaciones de Box y Cox de estas variables no cumplieron el supuesto de normalidad, entonces aplicaremos la prueba de Wilcoxon:

Prueba para **age:**

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by output
## W = 14394, p-value = 4.626e-05
## alternative hypothesis: true location shift is not equal to 0
```

En este caso el $\boldsymbol{p - valor}$ es menor al nivel de significancia $\boldsymbol{\alpha} = 0.05$, por lo cual, sí se observan diferencias estadísticamente significativas en la edad de los pacientes con respecto a la variable **output**.

Prueba para **trtbps:**

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  trtbps by output
## W = 12694, p-value = 0.06744
## alternative hypothesis: true location shift is not equal to 0
```

En este caso el $\boldsymbol{p-valor}$ es mayor al nivel de significancia $\boldsymbol{\alpha} = 0.05$, de este modo, no se observan diferencias estadísticamente significativas en la presión arterial entre los pacientes más propensos y menos propensos a sufrir ataques cardíacos.

Prueba para **oldpeak:**

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  oldpeak by output
## W = 16620, p-value = 9.191e-13
## alternative hypothesis: true location shift is not equal to 0
```

En este caso el $\boldsymbol{p-valor}$ es menor al nivel de significancia $\boldsymbol{\alpha} = 0.05$, por lo cual, sí se observan diferencias estadísticamente significativas en la depresión del segmento ST de los pacientes más probables y menos probables a sufrir ataques cardíacos.

4.2.2 Variables categóricas

En esta sección vamos a determinar si existen diferencias significativas de las variables categóricas **sex**, **cp**, **thall**, **restecg**, **exng**, **slp**, **thall** entre los grupos definidos por la variable categórica **output**. En primer lugar, vamos a presentar diagramas de frecuencias de cada una de estas variables con respecto a la variable **output**, lo cual nos permitirá formar una idea sobre el comportamiento de estas variables.

Para ello, vamos a emplear el test χ^2, el cual plantea la siguiente prueba de hipótesis:

$$\begin{cases} H_0 : \text{La variable } X \text{ es independiente a la variable } Y. \\ H_1 : \text{Las variables } X \text{ y } Y \text{ están asocidadas.} \end{cases}$$

Prueba para sex y output:

```
          0         1
0 0.2500000 0.7500000
1 0.5533981 0.4466019
```

Podemos notar que el $\mathbf{75}\%$ de las mujeres son más propensas a sufrir un ataque cardíaco, mientras que en los hombres solamente el $\mathbf{49.83}\%$ lo son.

Con la prueba χ^2 se obtienen los siguientes resultados:

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  outputsex
## X-squared = 23.084, df = 1, p-value = 1.551e-06
```

Considerando un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p-valor} < \boldsymbol{\alpha}$, podemos rechazar la hipótesis nula y concluir que existe una asociación significativa entre las variables **sex** y **output**. Es decir, podemos afirmar que la distribución de género difiere significativamente entre los grupos de **output**.

Prueba para cp y output:

```
          0         1
0 0.7272727 0.2727273
1 0.1800000 0.8200000
2 0.2093023 0.7906977
3 0.3043478 0.6956522
```

- Solamente para los pacientes asintomáticos la tasa de personas sanas es mayor a la tasa de personas propensas a sufrir una enfermedad. En este caso, la tasa de personas sanas es de 72.72%.
- En las categorías angina típica, angina atípica y dolor no anginoso, la tasa de personas sanas es menor, siendo estas de 18%, 20.69% y 30.43% respectivamente.

Aplicado la prueba χ^2, tenemos que:

```
##  Pearson's Chi-squared test
##
## data:  outputcp
## X-squared = 80.979, df = 3, p-value < 2.2e-16
```

Para un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor < \alpha}$, podemos rechazar la hipótesis nula y concluir que existe una asociación significativa entre las variables **cp** y **output**. Es decir, podemos afirmar que la distribución de la variable **cp** (tipo de dolor torácico) difiere significativamente entre los grupos de **output**.

Prueba para thall y output:

```
          0         1
1 0.6666667 0.3333333
2 0.2215569 0.7784431
3 0.7606838 0.2393162
```

- La tasa de salud es menor a la tasa de enfermedad cuando los pacientes dan resultados normales a la prueba de talio, la cual es de 22.02%.
- Por otro lado, la tasa de salud es alta cuando los pacientes poseen una prueba de esfuerzo con talio que indica un defecto fijo o defectos reversibles, en este caso, la tasa de salud es 66.67% y 76.07%, respectivamente.

Con la prueba χ^2 se obtienen los siguientes resultados:

```
##  Pearson's Chi-squared test
##
## data:  outputthall
## X-squared = 83.978, df = 2, p-value < 2.2e-16
```

Para un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor < \alpha}$, podemos rechazar la hipótesis nula y concluir que existe una asociación significativa entre las variables **thall** y **output**. Es decir, podemos afirmar que la distribución de la variable **thall** (resultados de prueba de esfuerzo con talio) difiere significativamente entre los grupos de **output**.

Prueba para restecg y output:

```
          0         1
0 0.5374150 0.4625850
1 0.3708609 0.6291391
2 0.7500000 0.2500000
```

- Cuando el paciente posee anomalías en la onda STT, la tasa de salud ($\mathbf{36.84}\%$) es menor en comparación a la tasa de enfermedad ($\mathbf{63.16}\%$).

La prueba χ^2 nos indica los siguientes resultados:

```
##  Pearson's Chi-squared test
##
## data:  outputrestecg
## X-squared = 9.7297, df = 2, p-value = 0.007713
```

Para un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor < \alpha}$, podemos rechazar la hipótesis nula y concluir que existe una asociación significativa entre las variables **restecg** y **output**. Es decir, podemos afirmar que la distribución de la variable **restecg** (resultados electrocardiográficos en reposo) difiere significativamente entre los grupos de **output**.

Prueba para exng y output:

```
          0         1
0 0.3054187 0.6945813
1 0.7676768 0.2323232
```

- Si el paciente posee angina la tasa de salud ($\mathbf{76.77}\%$) es mayor a la tasa de sufrir un ataque cardíaco ($\mathbf{23.23}\%$).

Con la prueba χ^2, obtenemos los siguientes resultados:

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  outputexng
## X-squared = 55.456, df = 1, p-value = 9.556e-14
```

Para un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor < \alpha}$, podemos rechazar la hipótesis nula y concluir que existe una asociación significativa entre las variables **exng** y **output**. Es decir, podemos afirmar que la distribución de la variable **exng** (angina producida por ejercicio) difiere significativamente entre los grupos de **output**.

Prueba para slp y output:

```
          0         1
0 0.5714286 0.4285714
1 0.6500000 0.3500000
2 0.2482270 0.7517730
```

- Únicamente cuando la pendiente de un segmento de electrocardiograma del paciente es ascendente, la tasa de salud (24.65%) es menor a la tasa de enfermedad (75.35%).
- Cuando los pacientes tienen pendiente descendente o pendiente normal, entonces sus tasas de salud son mayores, siendo estás de 57.14% y 65%, respectivamente.

La prueba χ^2, nos permite obtener los siguientes resultados:

```
##  Pearson's Chi-squared test
##
## data:  outputslp
## X-squared = 46.889, df = 2, p-value = 6.578e-11
```

Prueba para fbs y output:

```
          0         1
0 0.4513619 0.5486381
1 0.4888889 0.5111111
```

- En ambas categorías la tasa de salud es menor a la tasa de enfermedad.

Aplicando la prueba χ^2, se obtiene que:

```
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  outputfbs
## X-squared = 0.092408, df = 1, p-value = 0.7611
```

Si consideramos un nivel de significancia del $\boldsymbol{\alpha} = 0.05$, dado que $\boldsymbol{p - valor > \alpha}$, no podemos rechazar la hipótesis nula, por lo cual, no existe una asociación significativa entre las variable **fbs** y **output**.

4.3 Regresión logística

En el apartado anterior realizamos un análisis bivariado entre las variables numéricas y las variables categóricas con respecto a la variable **ouput**, por lo cual, en esta sección vamos a presentar un modelo de clasificación con la intención de predecir la probabilidad de que una persona sea más propensa a sufrir un ataque cardíaco, para ello, vamos a utilizar una regresión logística.

En primer lugar, vamos a dividir el conjunto de datos de modo que el 80% corresponda a la muestra de entrenamiento y el 20% a la muestra de validación, para ello, emplearemos la *librería* `caret` y la *función* `createDataPartition()`.
Verificamos que se mantenga la misma proporción de pacientes propensos y menos propensos a sufrir un ataque cardíaco en ambas muestras:

output	n	Percent
0	111	45.68
1	132	54.32

Mientras que para la muestra de validación se tiene que:

output	n	Percent
0	27	45.76
1	32	54.24

Por lo cual, podemos notar que la variable **output** se encuentra distribuida proporcionalmente en ambas muestras.

4.3.1 Selección de variables

En los modelos logit, la prueba para contrastar las hipótesis si los coeficientes son diferentes de cero ($\beta i \neq 0$) es la prueba de Wald.Consideremos:

$$H_0 : \hat{\beta}_j = \beta_{j0},$$
$$H_a : \hat{\beta}_j \neq \beta_{j0}.$$

Bajo la hipótesis nula, el estadístico de Wald sigue una distribución χ^2, y se define por:

$$T = \frac{(\hat{\beta}_j - \beta_{j0})^2}{V[\hat{\beta}_j]}$$

Para seleccionar los predictores que deben formar parte del modelo, emplearemos el método backward, este método primero calcula el modelo con todas las variables disponibles, luego se excluyen las variables una a una buscando una mejoría para finalmente eliminar la peor variable. Este método permite evaluar cada variable en presencia de las otras.

Para seleccionar los parámetros del modelo logit emplearemos el criterio de Akaike, que propone estudiar el problema de la identificación desde la perspectiva de la teoría de decisión estadística, que consiste en elegir como función de pérdida (o criterio de especificación) el AIC (Akaike Information Criterion) mínimo. Este valor se calcula por:

$$AIC = -2ln(\text{ máxima verosimilitud}) + 2(\text{ no. de parámetros independientemente ajustados}).$$

Previo a aplicar el modelo, vamos a transformar las variables categóricas a dummy's, para ello, vamos a emplear la *librería* `fastDummies`:

Aplicamos el modelo con todas las variables y luego consideramos el método backward para seleccionar las variables que formarán parte del modelo final.

```
##
## Call:
## glm(formula = output ~ ., family = "binomial", data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7784  -0.3559   0.1571   0.5363   2.3950
##
## Coefficients: (7 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.808359   4.092423   0.198 0.843416
## age          0.009303   0.025993   0.358 0.720419
## trtbps      -0.016871   0.014976  -1.126 0.259963
## chol        -0.007389   0.005502  -1.343 0.179311
## thalachh     0.021068   0.011451   1.840 0.065795 .
## oldpeak     -0.656341   0.260801  -2.517 0.011848 *
```

```
## caa          -0.910652   0.250324  -3.638 0.000275 ***
## sex_0         1.416172   0.562641   2.517 0.011836 *
## sex_1               NA         NA      NA       NA
## cp_0         -1.687161   0.749041  -2.252 0.024295 *
## cp_1         -0.756450   0.855150  -0.885 0.376382
## cp_2          0.109777   0.749317   0.147 0.883525
## cp_3               NA         NA      NA       NA
## fbs_0        -0.078895   0.601051  -0.131 0.895569
## fbs_1               NA         NA      NA       NA
## restecg_0     0.226174   2.307804   0.098 0.921929
## restecg_1     1.122205   2.303237   0.487 0.626096
## restecg_2           NA         NA      NA       NA
## exng_0        0.663503   0.474974   1.397 0.162436
## exng_1              NA         NA      NA       NA
## slp_0        -0.496047   1.124016  -0.441 0.658983
## slp_1        -0.636832   0.514912  -1.237 0.216169
## slp_2               NA         NA      NA       NA
## thall_1       1.099223   0.827909   1.328 0.184274
## thall_2       1.191371   0.456319   2.611 0.009032 **
## thall_3             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 335.05  on 242  degrees of freedom
## Residual deviance: 163.25  on 224  degrees of freedom
## AIC: 201.25
##
## Number of Fisher Scoring iterations: 6
```

Se puede encontrar variables no significativas cuyo p-valor es mayor a 0.050.05, por lo tanto, hay que estimar un nuevo modelo.

```
step(logit, direction = "backward")
```

De este modo, consideramos el siguiente modelo:

```
##
## Call:
## glm(formula = output ~ chol + thalachh + oldpeak + caa + sex_0 +
##     cp_0 + restecg_1 + thall_2, family = "binomial", data = trainin
g)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4670  -0.3981   0.1800   0.5297   2.3794
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.658266   1.818014  -0.362 0.717293
## chol        -0.009110   0.005012  -1.818 0.069133 .
## thalachh     0.024201   0.009605   2.520 0.011745 *
## oldpeak     -0.750708   0.206308  -3.639 0.000274 ***
## caa         -0.775425   0.220773  -3.512 0.000444 ***
## sex_0        1.154603   0.492382   2.345 0.019030 *
## cp_0        -1.647800   0.408360  -4.035 5.46e-05 ***
## restecg_1    0.975578   0.400211   2.438 0.014783 *
## thall_2      1.212180   0.420171   2.885 0.003914 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 335.05  on 242  degrees of freedom
## Residual deviance: 170.50  on 234  degrees of freedom
## AIC: 188.5
##
## Number of Fisher Scoring iterations: 6
```

El valor de AIC para este modelo es de 187.3. Para interpretar los coeficientes obtenidos vamos a emplear los odd ratios:

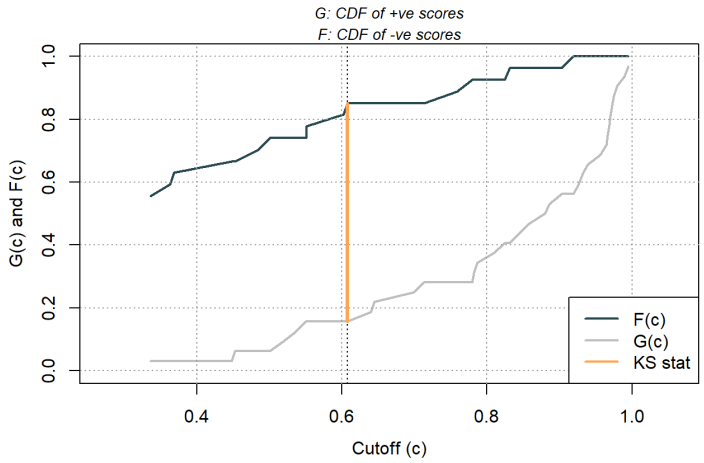
```
## (Intercept)        chol    thalachh     oldpeak         caa       s
ex_0
##   0.5177486   0.9909313   1.0244964   0.4720322   0.4605078   3.172
7632
```

```
##        cp_0   restecg_1     thall_2
##   0.1924728   2.6527008   3.3608046
```

De este modo, vamos a interpretar el coeficiente de cada una de las variables, mientras el resto se mantienen constantes:

- **Intercepto**: Por el signo del coeficiente podemos asegurar que hay menos personas propensas a sufrir un ataque cardíaco. De hecho, existen 1.93 pacientes propensos a sufrir un ataque cardíaco por cada persona que no es propensa.
- **chol**: La probabilidad de que un paciente sea propenso a sufrir un ataque cardíaco está negativamente relacionada con la cantidad de colesterol del paciente. Por cada unidad que se reste de esta variable la probabilidad de tener un ataque cardíaco aumenta en promedio 1.009 veces.
- **thalachh**: La probabilidad de que un paciente sea propenso a sufrir un ataque cardíaco está positivamente relacionada con la frecuencia cardíaca máxima alcanzada. Por cada unidad que se aumente de esta variable la probabilidad de tener un ataque cardíaco aumenta en promedio 1.02.
- **oldpeak**: La probabilidad de que un paciente sea propenso a sufrir un ataque cardíaco está negativamente relacionada con la depresión del ST inducida por el ejercicio en relación con el reposo. Por cada unidad que se reste de esta variable la probabilidad de tener un ataque cardíaco aumenta en promedio 2.12 veces.
- **caa**: La probabilidad de que un paciente sea propenso a sufrir un ataque cardíaco está negativamente relacionada con el número de vasos principales coloreados por fluoroscopio. Por cada unidad que se reste de esta variable la probabilidad de tener un ataque cardíaco aumenta en promedio 2.2 veces.
- **sex**: Cuando el paciente es mujer ($sex = 0$) la probabilidad de sufrir un ataque cardíaco es de 3.17 veces mayor que cuando el paciente es hombre.
- **cp**: Cuando el paciente es asintomático $cp = 0$, la probabilidad de sufrir un ataque cardíaco es 5.2 veces menor que aquellos pacientes que poseen angina típica, angina atípica o dolor no anginoso.
- **restecg**: Cuando el paciente presenta anomalías en la onda ST-T $restecg = 1$, la probabilidad de sufrir un ataque cardíaco es 2.65 veces mayor que aquellos pacientes cuyos resultados fueron normales o que presentan hipertrofia ventricular izquierda.
- **thall**: Cuando la prueba de esfuerzo con talio indica que el paciente posee una prueba normal $thall = 2$, la probabilidad de sufrir un ataque cardíaco es 3.36 veces mayor que aquellos clientes que obtuvieron defectos reversibles o fijos.

Para determinar el punto de corte y clasificar a los pacientes vamos a determinar la medida de Kolmogorov - Smirnov, de este, modo presentamos el gráfico de las distribuciones acumuladas de la razón de verdaderos positivos y la distribución de falsos positivos:

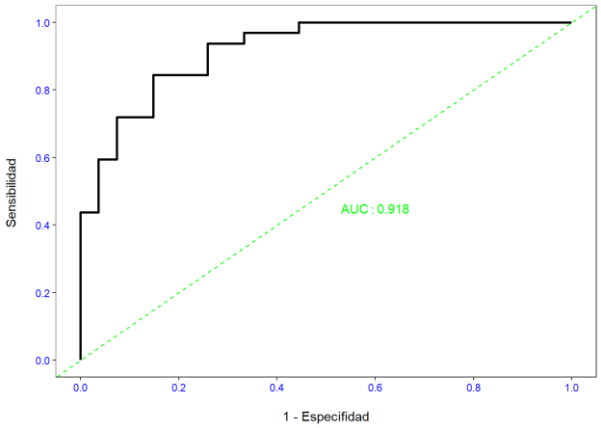


De este modo, el valor del KS es 0.6956019, el cual se alcanza en 0.6077056. Por lo cual, vamos a clasificar como pacientes propensos a sufrir un ataque cardíaco aquellos pacientes cuya probabilidad sea mayor a 0.6077056. De este modo, presentamos la matriz de confusión:

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  0  1
##          0 23  6
##          1  4 26
##
##                Accuracy : 0.8305
##                  95% CI : (0.7103, 0.9156)
##     No Information Rate : 0.5424
##     P-Value [Acc > NIR] : 3.14e-06
##
```

De este modo, el modelo tiene una exactitud del 83.05%, una precisión de 86.67% y un $F1 - score$ del 83.87%.

Ahora, vamos a calcular el valor del área bajo la curva AUC, para ello, consideremos la curva ROC:



Con lo cual, obtenemos un valor de $AUC = 0.92$.

Finalmente, para detectar la multicolinealidad emplearemos el **Factor de Inflación de la Varianza Generalizado (GVIF)** de los parámetros estimados, el cual proporciona un índice que mide hasta que punto la varianza de un coeficiente de regresión estimado se incrementa a causa de la co-linealidad; Los valores de los GVIF calculados de cada variable no deben ser mayores a 4:

chol	thalachh	oldpeak	caa	sex_0	cp_0	restecg_1	thall_2
1.146483	1.147110	1.053059	1.024546	1.198328	1.102594	1.060397	1.151725

Podemos notar que para ninguna variable el valor del $GVIF$ es superior a 4, por lo cual, podemos concluir que no existen problemas de multicolinealidad.

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

A lo largo de la práctica, se utilizaron varias representaciones gráficas y tablas para mostrar los resultados. A continuación, se muestran algunos ejemplos de las representaciones utilizadas:

- Representación gráfica de la distribución de variables numéricas: Se utilizaron gráficos de densidad y diagramas de caja para representar la distribución de las variables numéricas (age, trtbps, chol, thalachh, oldpeak) en relación a la variable “output”. Estos gráficos permiten visualizar la forma de la distribución, los valores atípicos y las diferencias entre los grupos.
- Representación gráfica de variables categóricas: Se utilizaron gráficos de barras para representar la frecuencia de las variables categóricas (sex, cp, fbs, restecg, exng, slp, thall, output). Estos gráficos permiten comparar las proporciones de diferentes categorías y observar posibles patrones o tendencias.
- Tabla de correlaciones: Se calculó la matriz de correlaciones (método de Spearman) entre las variables numéricas (age, trtbps, chol, thalachh, oldpeak) y se mostró en una tabla. Esta tabla permite identificar las relaciones de correlación entre las variables.
- Gráfico de cajas para variables numéricas y output: Se utilizó un gráfico de cajas para representar la distribución de las variables numéricas (age, trtbps, chol, thalachh, oldpeak) en relación a la variable “output”. Este gráfico muestra las diferencias en las distribuciones entre los grupos de “output”.

- Gráfico ROC (Receiver Operating Characteristic): Se utilizó un gráfico ROC para evaluar el rendimiento del modelo predictivo. Este gráfico muestra la sensibilidad frente a la especificidad del modelo a diferentes puntos de corte y calcula el área bajo la curva (AUC), que indica la capacidad de discriminación del modelo.

6) Resolución del problema

A partir de los resultados obtenidos en el análisis estadístico y la evaluación del modelo de regresión logística, podemos llegar a las siguientes conclusiones:

- Las variables que influyen en que una persona sea más o menos propensa a sufrir un ataque cardíaco son: **chol**, **thalachh**, **oldpeak**, **caa**, **sex**, **cp**, **restecg** y **thall**.
- El valor del estadístico de Kolmogorov-Smirnov obtenido, con un valor de 0.6956019, indica que el modelo tiene una buena capacidad para distinguir entre los pacientes propensos y no propensos a sufrir un ataque cardíaco.
- El punto de corte identificado en 0.6077056 nos permite clasificar eficientemente a los pacientes propensos a sufrir un ataque cardíaco. Aquellos con una probabilidad estimada superior a este punto de corte se consideran propensos.
- El modelo de regresión logística ha mostrado un desempeño satisfactorio, con una exactitud de 83.05% una precisión del 86.67% y un $F1 - score$ del 83.87%. Estos indicadores demuestran la capacidad del modelo para predecir correctamente los casos positivos (pacientes propensos).
- El valor del área bajo la curva (AUC) obtenido, con un valor de 0.920.92, indica que el modelo tiene una buena capacidad de discriminación entre los casos positivos y negativos.
- El análisis del Factor de Inflación de la Varianza Generalizado (GVIF) ha revelado que no existen problemas significativos de multicolinealidad entre las variables utilizadas en el modelo. Esto refuerza la confiabilidad de las estimaciones de los coeficientes de regresión.

7 Código

Durante el desarrollo del ejercicio, se empleó el lenguaje de programación R, el cual se encuentra alojado en un repositorio de GitHub. Se utilizaron diversas herramientas y técnicas para abordar los siguientes elementos:

- Descripción del dataset: Se realizó una descripción detallada de los datos utilizados en el análisis, incluyendo la naturaleza de las variables, la estructura de los datos y la distribución de los valores.
- Limpieza de datos: Se realizaron tareas de limpieza y pre procesamiento de los datos para garantizar su calidad y adecuación para el análisis. Esto incluyó el manejo de valores faltantes, la corrección de errores y la transformación de variables si fuera necesario.
- Variables numéricas: Se llevaron a cabo técnicas como el imputado de valores faltantes, la detección y tratamiento de outliers, y la normalización o estandarización de variables si fuera requerido.
- Variables categóricas: Se realizó el manejo de variables categóricas, incluyendo la codificación de variables categóricas en variables numéricas utilizando técnicas como la codificación one-hot o la codificación ordinal.

- Correlaciones: Se exploraron las correlaciones entre las variables para identificar posibles relaciones o dependencias entre ellas.
- Comparación entre grupos: Se realizaron comparaciones entre grupos de datos utilizando técnicas estadísticas adecuadas. Esto incluyó la comparación de variables numéricas entre diferentes grupos utilizando pruebas de hipótesis o análisis de varianza, y la comparación de variables categóricas utilizando tablas de contingencia y pruebas de chi-cuadrado.
- Variables numéricas: Se compararon las distribuciones de variables numéricas entre diferentes grupos para evaluar posibles diferencias estadísticamente significativas.
- Variables categóricas: Se compararon las frecuencias de las categorías de variables categóricas entre diferentes grupos para determinar si existían diferencias significativas.
- Regresión logística: Se aplicó el modelo de regresión logística para predecir la probabilidad de que una persona sea propensa a sufrir un ataque cardíaco. Se realizaron técnicas de selección de variables para identificar las variables más relevantes en el modelo.
- Conclusiones: A partir de los resultados obtenidos, se elaboraron conclusiones sobre el análisis realizado. Estas conclusiones pueden incluir hallazgos importantes, relaciones identificadas entre variables, el rendimiento del modelo de regresión logística y su capacidad para predecir la probabilidad de ataques cardíacos, entre otros aspectos relevantes.

8 Vídeo

Enlace

Referencias

Calvo, M., Subirats, L., & Pérez, D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

Squire, M. (2015). Clean Data. Packt Publishing Ltd.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. Morgan Kaufmann.

Osborne, J. W. (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews, 10(1), 15-27.

Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.

McKinney, W. (2012). Python for Data Analysis. O'Reilly Media, Inc.

GitHub. (n.d.). Tutorial de Github. Recuperado de https://guides.github.com/activities/hello-world

Data to Viz. (n.d.). Herramienta para realización de gráficas. Recuperado de https://www.data-to-viz.com/

Contribuciones	Firma Integrantes
Investigación previa	WGGB, OADT
Redacción de las respuestas	WGGB, OADT
Desarrollo del código	WGGB, OADT
Participación en el vídeo	WGGB, OADT