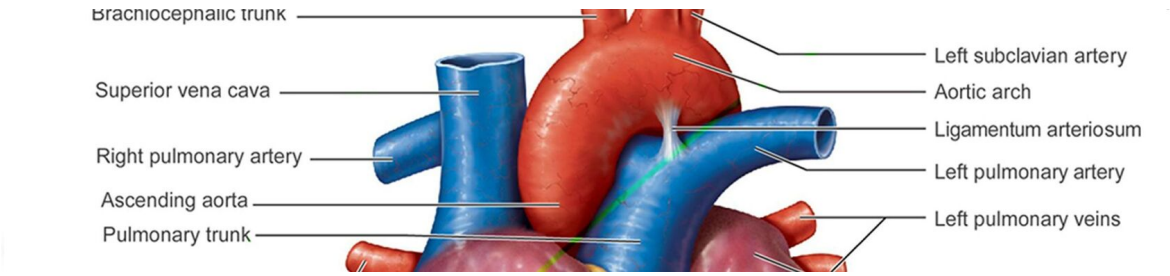
Heart Attack Analysis & Prediction Dataset

A dataset for heart attack classification



Oriol Gonzàlez Dalmau

Índice

1. Descripción del dataset.

Heart Attack Analysis & Prediction Dataset es un conjunto de datos de Kaggle. En este conjunto de dos datasets se puede encontrar dos archivos separados por comas, el primero contiene los atributos relacionados con enfermedades cardiovasculares y el otro un conjunto de datos sobre la saturación de oxígeno en sangre, si más no, el estudio del último csv hace plantear la eficacia de este al no obtener una relación clara entre los dos archivos ya que no coinciden en el número de filas y no existe un identificador para su unificación.

Otro hecho remarcable es la falta de una fuente oficial para la obtención de los datos. Estos fueron subidos a Kaggle por RASHIK RAHMAN pero no se indica donde ni como se recolectaron. La única información remarcable es la licencia que se indica como: "CC0: Public Domain" y se indica como fuente "online" con metodología "crawling". Cabe comentar que para la elaboración de la práctica se ha buscado en internet más información sobre este conjunto de datos, concretamente la web del repositorio oficial que extraordinariamente no aparece citada en Kaggle.

El primer archivo, con el que se trabajará contiene 303 registros con 11 atributos , que según la documentación de Kaggle son:

1. Age : Age of the patient
2. Sex : Sex of the patient sex (1 = male; 0 = female)
3. exang: exercise induced angina (1 = yes; 0 = no)
4. ca: number of major vessels (0-3)
5. cp : Chest Pain type chest pain type
    a. Value 1: typical angina
    b. Value 2: atypical angina
    c. Value 3: non-anginal pain
    d. Value 4: asymptomatic
6. trtbps : resting blood pressure (in mm Hg)
7. chol : cholestoral in mg/dl fetched via BMI sensor
8. fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
9. rest_ecg : resting electrocardiographic results
    a. Value 0: normal
    b. Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    c. Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
10. thalach : maximum heart rate achieved
11. target : 0= less chance of heart attack 1= more chance of heart attack

Se trabaja con supuestos pacientes, cada registro representa una persona con edad, sexo y ciertos atributos concretos, estos atributos describen ,métricas de el estado del cuerpo humano y además otras métricas más sujetivas del tipo de dolor del paciente. El dataset es útil para extraer conclusiones de cuales son las principales causas de los ataques al corazón . Aquí la variable

etiqueta u objetivo se describe como la probabilidad de ataque al corazón aunque realmente se introduce una variable dicotómica, siendo 1 riesgo de ataque y 0 riesgo bajo de ataque.

Conociendo la principal estructura de los datos se puede concluir que la importancia de estos recaerá en la elaboración de un modelo de minería de datos para la predicción de ataques al corazón. Se cree que para este modelo de minería de datos supervisado podría ser interesante el uso de algoritmos de clasificación como el k-means o árboles de decisión.

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

Previamente se ha considerado que el uso del segundo archivo incluido en los datos originales referente al nivel de saturación de oxígeno en sangre no puede ser usado debido a la falta de contexto e información respecto a los otros datos.

Por el momento, hasta un análisis más en profundidad de la variabilidad de los atributos consideramos que los propuestos son todos relevantes para la elaboración de un modelo predictivo.

Primer análisis descriptivo.

Para la práctica usaremos un entorno virtual en el que incluimos las librerias instaladas con las respectivas versiones en el archivo requirements.txt

- Carga de datos:

  Al realizar la carga de datos se observa que las columnas que describen el dataset en la web de Kaggle no coinciden con las columnas que encontramos en el csv ofrecido, como se muestra en la siguiente imagen:

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
# carga de datos

heartAttack = pd.read_csv('../data/heart.csv')
```

In [3]:
```
heartAttack.head()
```

Out[3]:

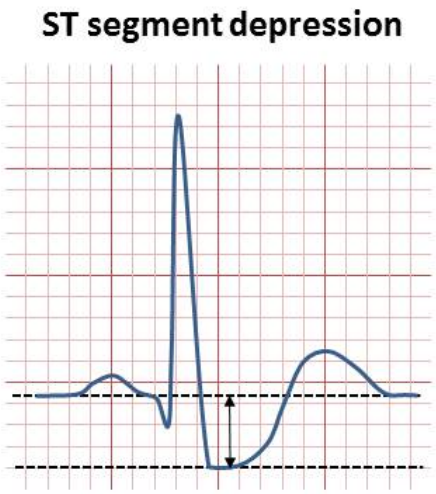
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figura 1

Las columnas que se obtienen son 'age', 'sex', 'cp', 'trtbps', 'chol', 'fbs', 'restecg', 'thalachh', 'exng', 'oldpeak', 'slp', 'caa', 'thall' y 'output', sin embargo las variables que se describen anteriormente no contemplan 'Oldpeak', 'Slp' ni 'output'. Se desconocen las variables Oldpeak y slp, thall y output (que se cree hace referencia a Target).

Una breve búsqueda en internet proporciona la información referente a Oldpeak y slp, ambas directamente relacionadas con exng *Exercise induced angina* y algo de información sobre thal [Deshmukh. Hardik] :

- Oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot, **ST depression** refers to a finding on an electrocardiogram,[1][2] wherein the trace in the ST segment is abnormally low below the baseline.)



- slp: the slope of the peak exercise ST segment , valores posibles -> 0: downsloping; 1: flat; 2: upsloping
- thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
    1. Value 1: fixed defect (no blood flow in some part of the heart)
    2. Value 2: normal blood flow
    3. value 3: reversible defect (a blood flow is observed but it is not normal)

- Primera vista a los datos

```
In [4]: heartAttack.describe()
```

Out[4]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000

exng	oldpeak	slp	caa	thall	output
303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figura 2 y 3

- Análisis de las componentes principales.

  El último punto de selección de datos se basa en el estudio de las componentes principales para así estudiar si se debe realizar algún tipo de reducción de la dimensionalidad. Para ello antes es conveniente realizar la limpieza de datos, de lo contrario el algoritmo PCA podría perder eficacia o incluso dejar de funcionar.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos?

Como se puede observar en las figuras 2 y 3, los valores que contienen los atributos son valores aparentemente correctos para el rango que deberían tener. Sí que existen ceros pero son correctos ya que por ejemplo en la variable sex (1 = male, 0 = female). Si que se encuentran valores vacíos en la variable En cuanto a los valores nulos sin codificar, no encontrcomo podemos ver en la figura 4:

```
No encontramos valores nullos/nas

In [15]: print(heartAttack.isna().sum())

age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

Figura 4

3.2. Identifica y gestiona los valores extremos.

Para estudiar si existen valores extremos se puede mirar de nuevo la distribución de los datos de las figuras 2 y 3 pero es más fácil una visualización con boxplots.

Posibles outliers:

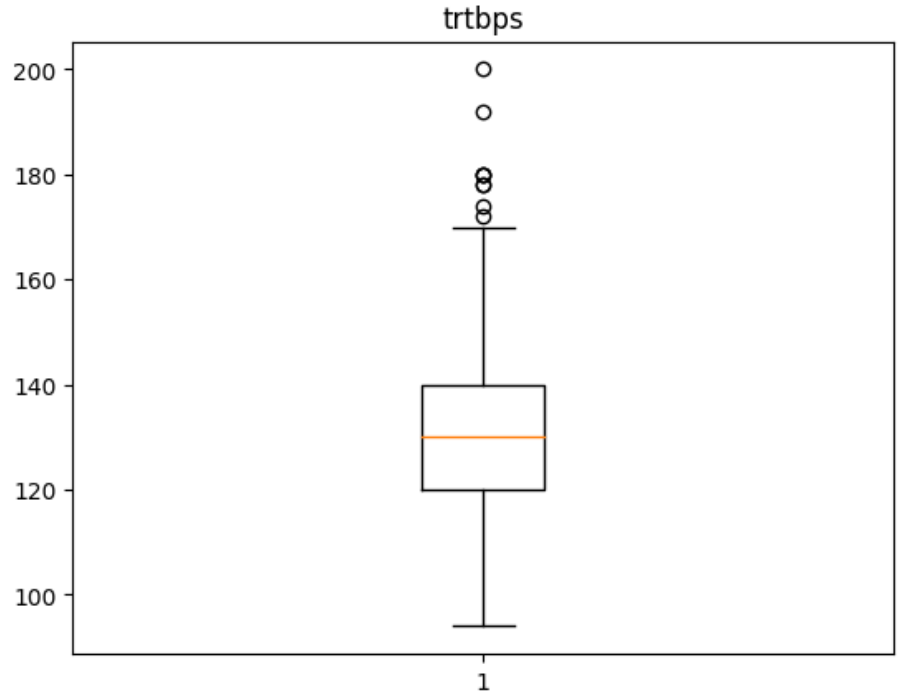


Figura 5

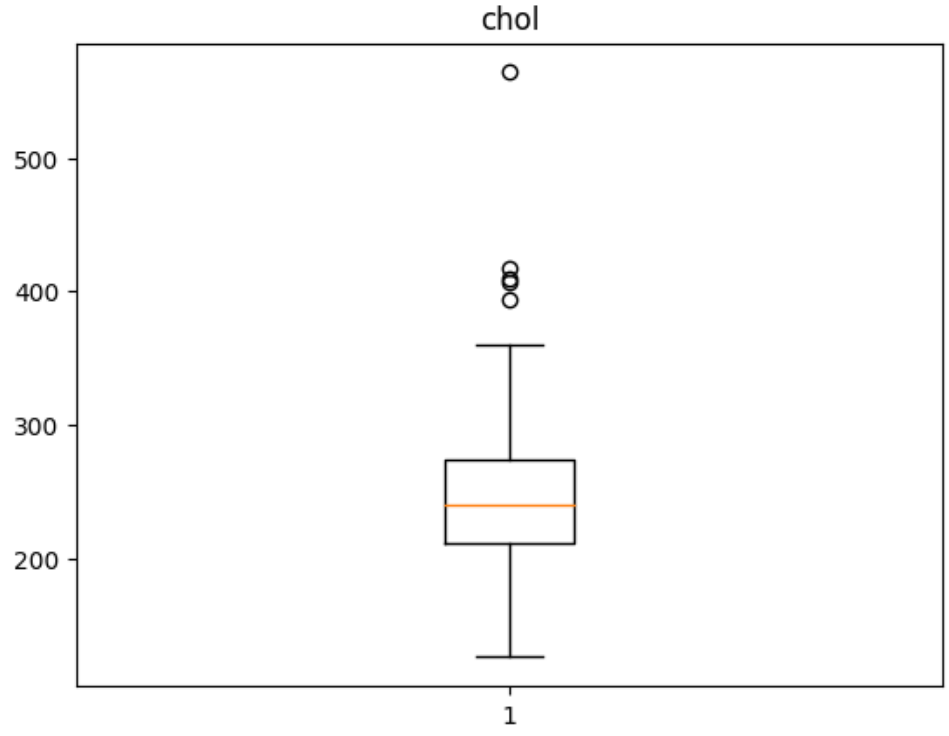


Figura 6

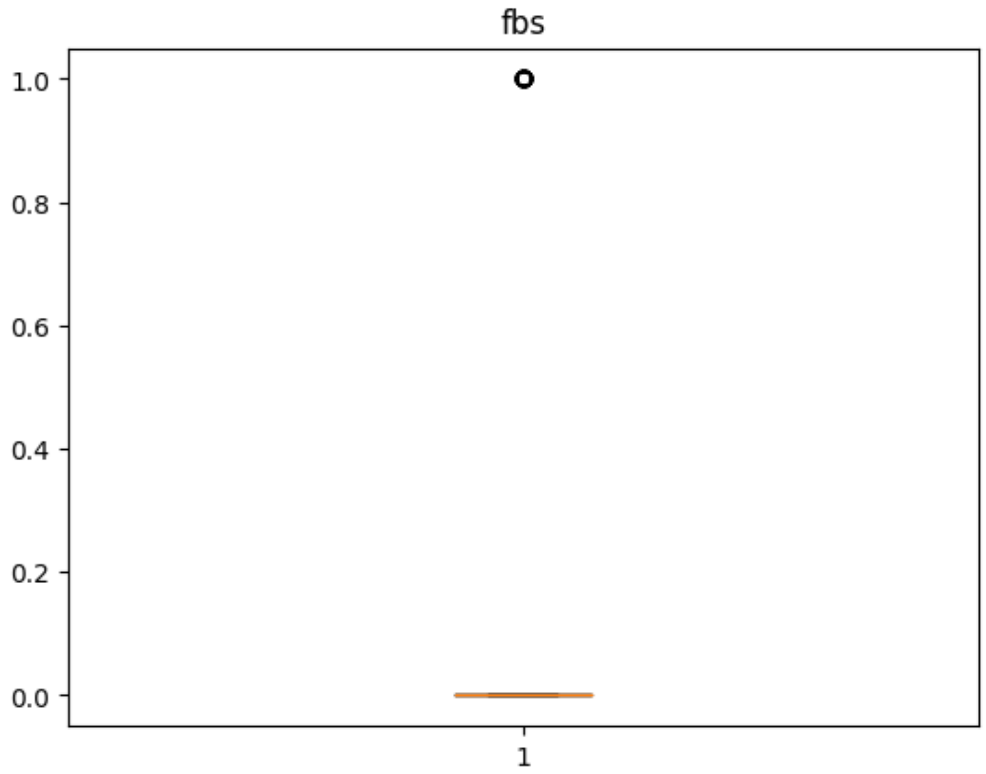


Figura 7

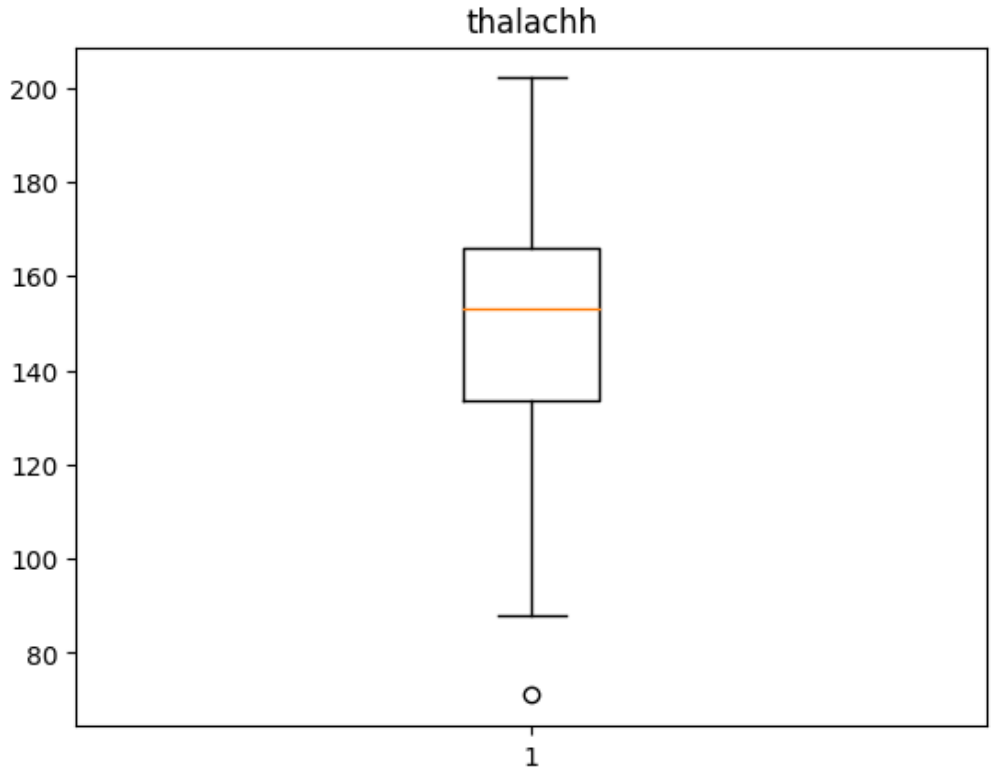


Figura 8

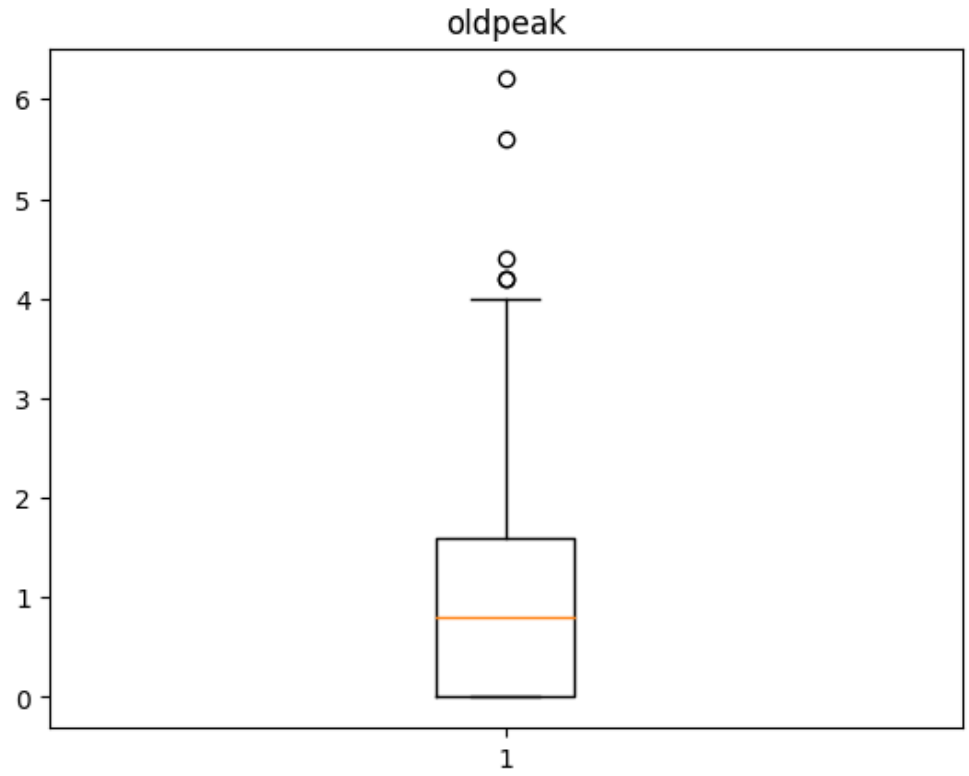


Figura 9

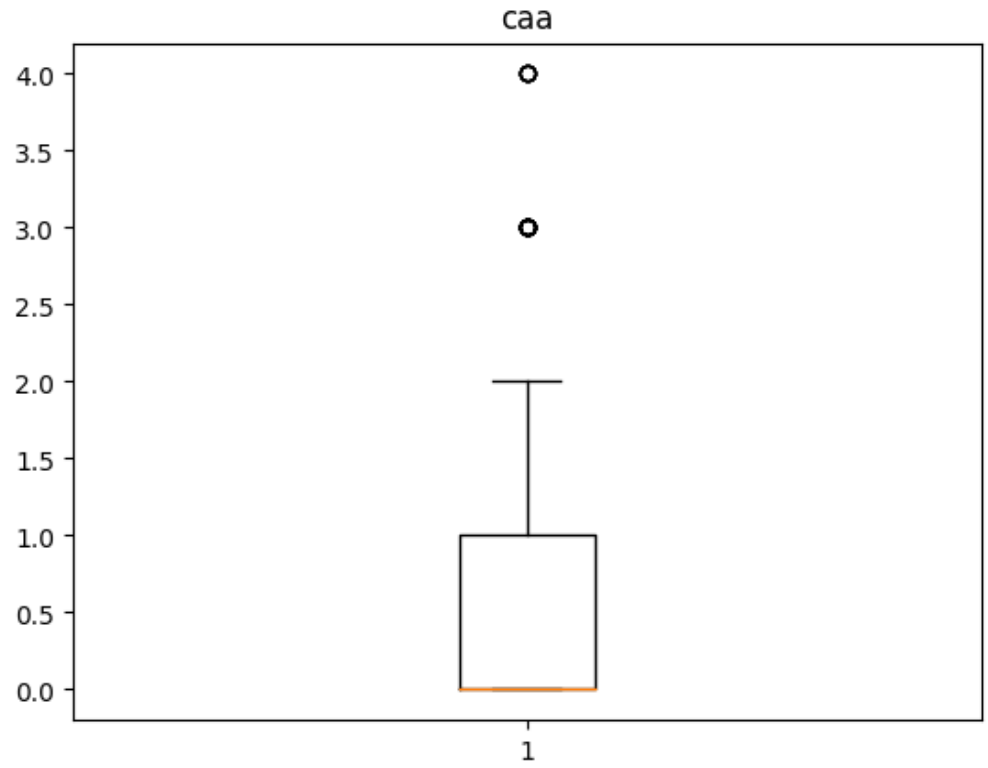


Figura 10

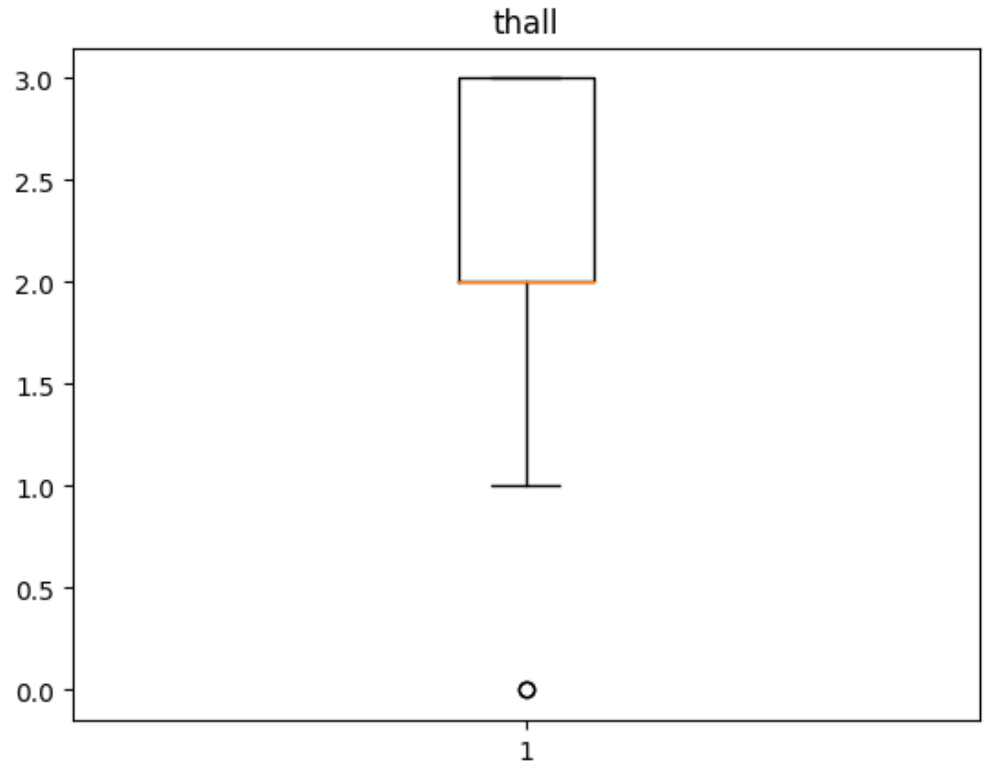


Figura 11

Después de analizar los boxplots que tienen círculos, es decir, los valores que están alejados más de dos desviaciones estándares de la mediana, existen varias figuras que resaltan respecto a las demás.

Los valores de *resting blood pressure*(trtbps) y *serum cholestoral* (chol), contienen valores extremos pero que son frecuentes en pacientes.

Para el caso de (fbs) *fasting blood sugar* > 120 mg/dl, pese encontrar “outliers” lo que representan los círculos del boxplot es una baja frecuencia de valores true (1 = true; 0 = false).

Oldpeak, los valores máximos son muy extremos y hacen que la media de oldpeak suba a más de 1 mientras que la mediana está en 0.8, falta contexto para poder identificar si estos valores son comunes dentro de lo extremo.

En cuanto al *number of major vessels* (0-3) (caa) se contempla lo que se cree un error al existir un paciente con 4 siendo el rango de 0 a 3, podría ser que el paciente tuviese algún tipo de malformación.

En los valores de *maximum heart rate achieved* (thalach) hay un valor que es 71, lo que resulta muy extraño para una frecuencia cardiaca máxima, es posible que sea un error en la prueba,como que en lugar de 171 se escribiera 71. Lo ideal sería estudiar la influencia de este atributo en la variabilidad del modelo y si es importante, eliminar este registro.

Por último Thal contiene un valor outlier que es un 0, como podemos observar en la figura 11. Sin embargo, tras consultar la documentación oficial [David W. Aha] se explica que un 0 es un valor nulo, como identificamos en el punto anterior.

3.3 Análisis de las componentes principales.

Una vez limpiados los datos, procedemos a estudiar la variabilidad de los datos para valorar reducir la dimensionalidad de estos, como se comentó en el punto 2.

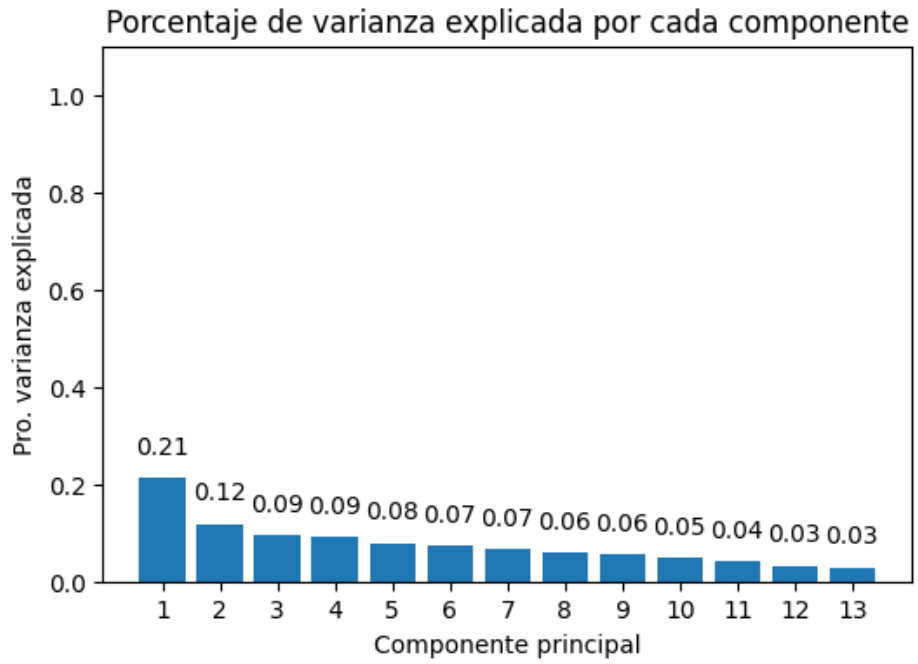


Figura 12

Se observa que las diferentes variables explican de manera similar la variabilidad en los datos, siendo la primera la que replica casi el 21% de la variabilidad, sin embargo las demás componentes van decreciendo muy progresivamente así que no es adecuado eliminar ninguna de ellas ya que estaríamos perdiendo información relevante.

Se sabe a posteriori que el dataset ya es un dataset reducido del original con más de 58 dimensiones, es por esto que casi todas las variables son importantes.

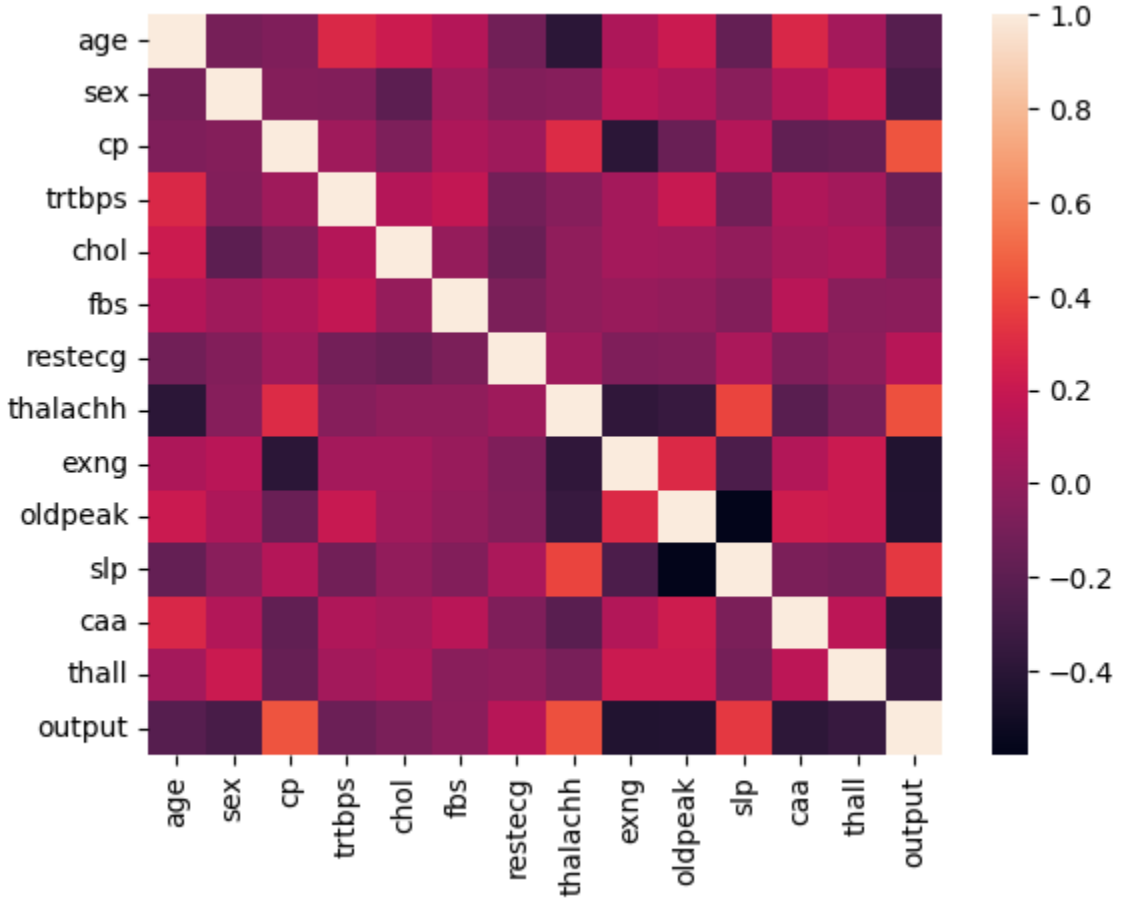
4. Análisis de los datos

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

(p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

- Estudiamos la correlacion

Matriz de covarianza:



La matriz de correlación presenta pocos atributos que estén altamente correlacionados, de los que más podrían interesar respecto a la variable etiqueta Output encontramos Oldpeak, exng, cp y thalachh. Se observa la correlación obtenida:

cp 0.433798
thalachh 0.421741
exng -0.436757
oldpeak -0.430696
slp 0.345877
caa -0.391724
thall -0.344029

Además, pese a no mostrar una correlación notable, se quiere estudiar la influencia del género en el riesgo de contraer enfermedades del corazón. Para decidir que factores son los más influyentes no se debe descartes ninguno hasta que se realice un estudio en profundamente, esto es debido a que posiblemente un solo atributo no tenga suficiente importancia pero la concatenación de múltiples sea clave para descubrir el origen del problema, así que el conjunto de valores previos puede ser extendido. Sin embargo este problema podrá ser tratado en otro instante como propuesta de mejora con el uso, por ejemplo, de árboles de decisión que nos devuelvan reglas con significado profundo.

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

Después de la aplicación de diferentes test para la normalidad como son el test de Shapiro y d'Agostino se llega a la conclusión que los atributos no siguen una distribución normal:

Se muestra un ejemplo, las demás variables están explicadas en el código.

```
AGOSTINO  age : Estadisticos=8.748, p=0.013
   se RECHAZA H0 :  age  proviene de una distribución Normal
SHAPIRO  age : Estadisticos=0.986, p=0.006
   se RECHAZA H0 :  age  proviene de una distribución Normal
AGOSTINO  sex : Estadisticos=947.846, p=0.000
   se RECHAZA H0 :  sex  proviene de una distribución Normal
SHAPIRO  sex : Estadisticos=0.586, p=0.000
   se RECHAZA H0 :  sex  proviene de una distribución Normal
AGOSTINO  cp : Estadisticos=168.439, p=0.000
   se RECHAZA H0 :  cp  proviene de una distribución Normal
SHAPIRO  cp : Estadisticos=0.790, p=0.000
   se RECHAZA H0 :  cp  proviene de una distribución Normal
```

Las variables no siguen una distribución Normal, para la prueba de homocedasticidad de la varianza se usará una prueba no paramétrica como el test de Fligner-killeen.

Antes de empezar con los test de homocedasticidad se plantean las preguntas que se quieren resolver:

1. ¿La media de edad afecta al riesgo de padecer una enfermedad del corazón?¿Y el género?
2. ¿El dolor de pecho ayuda a detectar un posible infarto?
3. ¿Durante la prueba el máximo de pulsaciones influye en el riesgo de un ataque al corazón?
4. ¿ Un electrocardiograma con una caída en oldpeak puede suponer un indicio de un fallo?

Al obtener las preguntas se puede continuar con el estudio de la homocedasticidad para no dar “palos de ciego”. De las diferentes variables dividimos en grupos según media y/o mediana y miramos la homocedasticidad de la variable etiqueta output.

Edad. La edad es relativa, pero en este estudio la media de edad está en los 54,7 años y la mediana en 55, así que consideraremos dos grupos de edad para estudiar la varianza, según el grupo joven y el grupo mayor. El p-valor es mayor de alpha(0.05) no se rechaza la hipótesis nula, y se considera que se cumple el principio de homocedasticidad de la varianza:

```
fligner(heartAttack[heartAttack['age']>=55]['output'],heartAttack[heartAttack['age']<55]['output'])
```

```
FlignerResult(statistic=3.485892926506156, pvalue=0.06189396173947675)
```

En cuanto al género, se estudia los dos grupos, 0 y 1, como el p-valor es menor de alpha(0.05), se rechaza la hipótesis nula, y se considera que no se cumple el principio de homocedasticidad de la varianza:

```
fligner(heartAttack[heartAttack['sex']==1]['output'],heartAttack[heartAttack['sex']==0]['output'])
```

```
FlignerResult(statistic=10.951079464403477, pvalue=0.0009354909924622863)
```

Dolor de pecho

```
# cp mayor de 1 indica dolor en el pecho
fligner(heartAttack[heartAttack['cp']>=1]['output'],heartAttack[heartAttack['cp']<1]['output'])
```

```
FlignerResult(statistic=1.492789809951808, pvalue=0.22178410911219326)
```

Pulsaciones maximas

```
#media de pulsaciones elevadas de 150, mediana 153
fligner(heartAttack[heartAttack['thalachh']>=151]['output'],heartAttack[heartAttack['thalachh']<151]['output'])
```

```
FlignerResult(statistic=1.1113188032107524, pvalue=0.29179544958814935)
```

Oldpeak

```
# oldpeak media 1.039604, mediana 0.8
fligner(heartAttack[heartAttack['oldpeak']>=0.8]['output'],heartAttack[heartAttack['oldpeak']<0.8]['output'])
```

```
FlignerResult(statistic=3.1120709996625617, pvalue=0.07771408581442851)
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

La prueba de Kruskal-Wallis plantea la hipótesis nula de que la mediana de las poblaciones de los grupos estudiados es igual. Esta es una prueba no paramétrica. Si el p-valor es menor a alpha(0.05) se rechaza la hipótesis nula lo que indica que las medianas difieren, pero no indica cómo difieren las medianas.

Así pues, retomando las preguntas anteriores:

1. ¿La media de edad afecta al riesgo de padecer una enfermedad del corazón?¿Y el género?

```
stats.kruskal(heartAttack[heartAttack['sex']==1]['output'],heartAttack[heartAttack['sex']==0]['output'])
```

```
KruskalResult(statistic=23.835458555307394, pvalue=1.049315707336381e-06)
```

Podemos rechazar la hipótesis nula, la mediana de ataques al corazón es diferente entre hombres y mujeres.

```
stats.kruskal(heartAttack[heartAttack['age']>=55]['output'],heartAttack[heartAttack['age']<55]['output'])
```

```
KruskalResult(statistic=24.776490637852383, pvalue=6.437789525854331e-07)
```

Podemos rechazar la hipótesis nula, la mediana de ataques al corazón es diferente entre grupos de edad.

2. ¿El dolor de pecho ayuda a detectar un posible infarto?

```
stats.kruskal(heartAttack[heartAttack['cp']>=1]['output'],heartAttack[heartAttack['cp']<1]['output'])
```

```
KruskalResult(statistic=80.41387531440874, pvalue=3.0365545934645777e-19)
```

Podemos rechazar la hipótesis nula, la mediana de ataques al corazón es diferente entre las personas que tienen dolor en el pecho y las que no.

3. ¿Durante la prueba el máximo de pulsaciones influye en el riesgo de un ataque al corazón?

```
stats.kruskal(heartAttack[heartAttack['thalachh']>=151]['output'],heartAttack[heartAttack['thalachh']<151]['output'])
```

```
KruskalResult(statistic=50.321219935263386, pvalue=1.3053043147691371e-12)
```

Podemos rechazar la hipótesis nula, la mediana de pulsaciones máxima influye en el riesgo de ataque al corazón.

4. ¿ Un electrocardiograma con una caída en oldpeak puede suponer un indicio de un fallo?

```
stats.kruskal(heartAttack[heartAttack['oldpeak']>=0.8]['output'],heartAttack[heartAttack['oldpeak']<0.8]['output'])
```

```
KruskalResult(statistic=48.765659803491, pvalue=2.884462408395387e-12)
```

Podemos rechazar la hipótesis nula, la mediana de ataques al corazón es diferente entre las personas con un electrocardiograma con caída en oldpeak.

5. Resolución del problema

Se obtienen resultados importantes pese a no tener una gran cantidad de registros. Aumentar el número de registros podría ser de especial interés si se quiere profundizar en el tema o crear modelos predictivos eficaces para la aportación de recursos al sistema. Pese a encontrar fallos en la normalidad de las variables, el test de Kruskal ha sido una herramienta clave para

aprender del set de datos. Se considera oportuno como propuesta de mejora realizar un segundo estudio sobre los resultados de los contrastes de hipótesis previos para ver qué grupos predominan sobre los ataques al corazón.

6. Vídeo y código

Vídeo:
https://drive.google.com/file/d/1Y4IbXjJCZ0AVD06edE8WAHxXc3nPhCap/view?usp=sharing

Código: https://github.com/oriolGonzalezDS/heartAttack

7. Webgrafia

[Deshmukh. Hardik]
https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7#70f9

[David W. Aha] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[Amat Rodrigo J] https://www.cienciadedatos.net/documentos/py19-pca-python.html