
Richard Douglas y Marcelino Madriz
Grupo# 2.

# Dataset SAheart

# Introducción

- El aprendizaje estadístico desempeña un papel clave en muchas áreas de la ciencia.
- Identificar los factores de riesgo para problemas cardiacos, en función de la clínica y variables demográficas. La ciencia del aprendizaje juega un papel clave en los campos de la estadística, los datos minería e inteligencia artificial, cruzando con áreas de ingeniería y otras disciplinas.

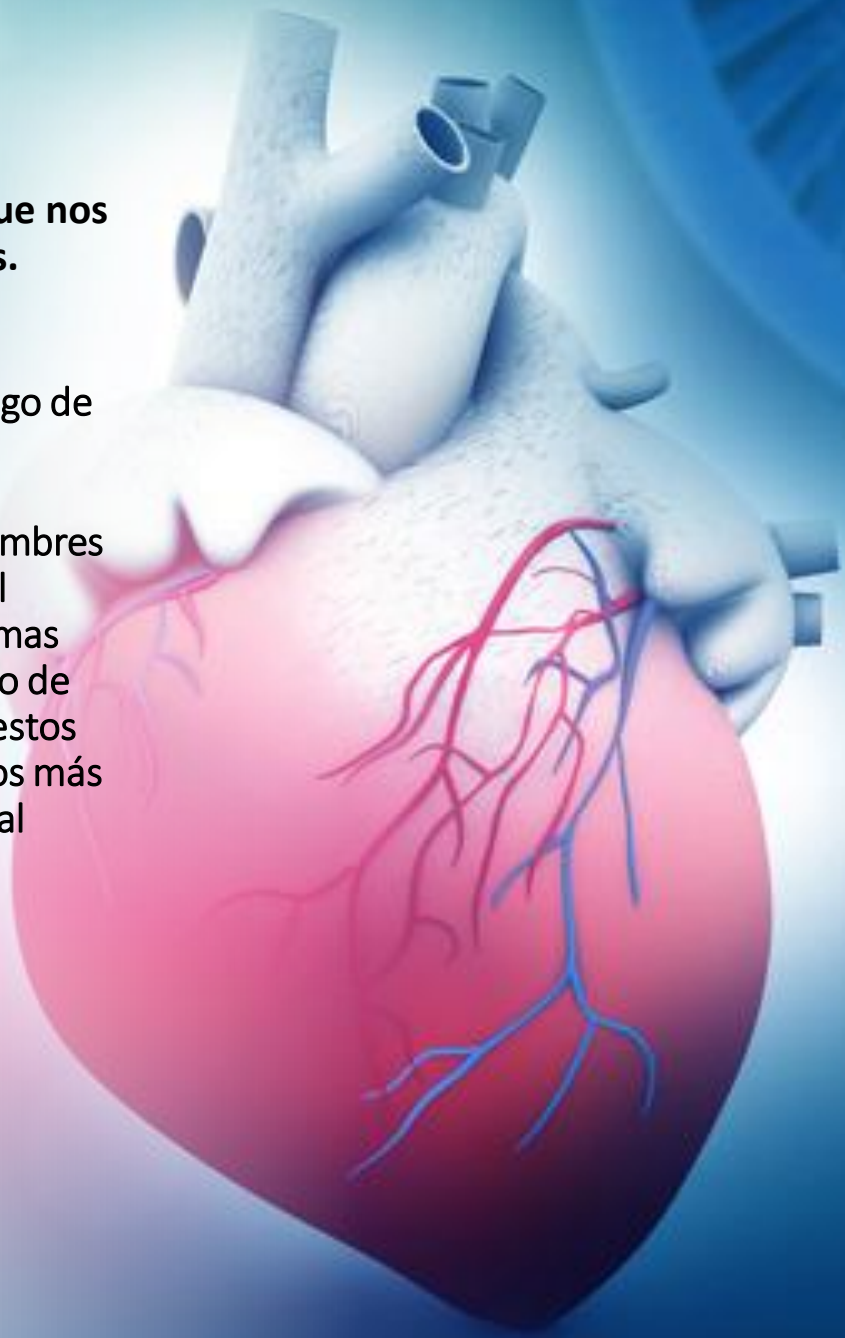
- En un escenario típico, tenemos una medición de resultados, generalmente cuantitativa o categórico, que deseamos predecir basado en un conjunto de características.

# Introducción

- **Usando estos datos construimos un modelo de predicción, que nos permitirá predecir el resultado para nuevos objetos invisibles.**

- Una muestra retrospectiva de varones en una región de alto riesgo de enfermedades cardíacas del Cabo Occidental, Sudáfrica.

- Hay áspero dos controles por el caso de CHD. Muchos de los hombres positivos de CHD(Coronary Heart Disease) han experimentado el tratamiento de la reducción de la presión arterial y otros programas para reducir sus factores de riesgo después de su acontecimiento de CHD. En algunos casos las mediciones se realizaron después de estos tratamientos. Estos datos están tomados de un conjunto de datos más grande, descrito en Rousseauw et al, 1983, South African Medical Journal.



# Carga del Dataset SAHeart

## Procesamiento


- Fiabilidad de los Datos.
- Limpieza de los datos.
- EDA.
- Anova.
- Machine Learning

| | Unnamed: 0 | tobacco | ldl | adiposity | famhist | typea | obesity | alcohol | age |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 160 | 12.00 | 5.73 | 23.11 | Present | 49 | 25.30 | 97.20 | 52 |
| 1 | 144 | 0.01 | 4.41 | 28.61 | Absent | 55 | 28.87 | 2.06 | 63 |
| 2 | 118 | 0.08 | 3.48 | 32.28 | Present | 52 | 29.14 | 3.81 | 46 |
| 3 | 170 | 7.50 | 6.41 | 38.03 | Present | 51 | 31.99 | 24.26 | 58 |
| 4 | 134 | 13.60 | 3.50 | 27.78 | Present | 60 | 25.99 | 57.34 | 49 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 457 | 214 | 0.40 | 5.98 | 31.72 | Absent | 64 | 28.45 | 0.00 | 58 |
| 458 | 182 | 4.20 | 4.41 | 32.10 | Absent | 52 | 28.61 | 18.72 | 52 |
| 459 | 108 | 3.00 | 1.59 | 15.23 | Absent | 40 | 20.09 | 26.64 | 55 |
| 460 | 118 | 5.40 | 11.61 | 30.79 | Absent | 64 | 27.35 | 23.97 | 40 |
| 461 | 132 | 0.00 | 4.82 | 33.41 | Present | 62 | 14.70 | 0.00 | 46 |

462 rows × 9 columns

# DataSet SAheart

- En el que cada columna de la tabla corresponde a una variable las filas representan los diferentes registros que almacena cada una de las columnas o variables de la tabla.

- Estas filas y columnas, junto con los valores, conforman el dataset o conjunto de datos en cuestión, uno de los ejemplo podrían ser la tabla se una base de datos conformados por:

- Umnamed 0 : systolic blood pressure(presion arterial)
- Tabacco: cumulative tobacco (kg) (consumo tabaco o fumado)
- LDL: : low densiity lipoprotein cholesterol (tambien conocido como colesterol dañino o malo)
- Adiposity: adiposidad o grasa localizada

| | Unnamed: 0 | tobacco | ldl | adiposity | famhist | typea | obesity | alcohol | age |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 160 | 12.00 | 5.73 | 23.11 | Present | 49 | 25.30 | 97.20 | 52 |
| 1 | 144 | 0.01 | 4.41 | 28.61 | Absent | 55 | 28.87 | 2.06 | 63 |
| 2 | 118 | 0.08 | 3.48 | 32.28 | Present | 52 | 29.14 | 3.81 | 46 |
| 3 | 170 | 7.50 | 6.41 | 38.03 | Present | 51 | 31.99 | 24.26 | 58 |
| 4 | 134 | 13.60 | 3.50 | 27.78 | Present | 60 | 25.99 | 57.34 | 49 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 457 | 214 | 0.40 | 5.98 | 31.72 | Absent | 64 | 28.45 | 0.00 | 58 |
| 458 | 182 | 4.20 | 4.41 | 32.10 | Absent | 52 | 28.61 | 18.72 | 52 |
| 459 | 108 | 3.00 | 1.59 | 15.23 | Absent | 40 | 20.09 | 26.64 | 55 |
| 460 | 118 | 5.40 | 11.61 | 30.79 | Absent | 64 | 27.35 | 23.97 | 40 |
| 461 | 132 | 0.00 | 4.82 | 33.41 | Present | 62 | 14.70 | 0.00 | 46 |

462 rows × 9 columns

# DataSet SAheart

- En el que cada columna de la tabla corresponde a una variable las filas representan los diferentes registros que almacena cada una de las columnas o variables de la tabla.

- Estas filas y columnas, junto con los valores, conforman el dataset o conjunto de datos en cuestión, uno de los ejemplo podrían ser la tabla se una base de datos conformados por:

- Famhist: family history of heart disease(Present=1, Absent=0) (presenta o no historial de problemas cardiacos).
- Typea: type-A behavior type of personality concerns how people respond to stress (Comportamiento de Tipo A, muetra como las personas reaccionan al estres=)
- Obesity: obesidad
- Alcohol: current alcohol consumption (consumo de alcochol)
- Age: age at onset (edad ).

# Fiabilidad y limpieza del Dataset

- Carga de Data

- Importación Dataset

- Exploración Inicial

- Revisión de datos nulos o espacios

- Limpieza de los datos

# Fiabilidad y limpieza del Dataset

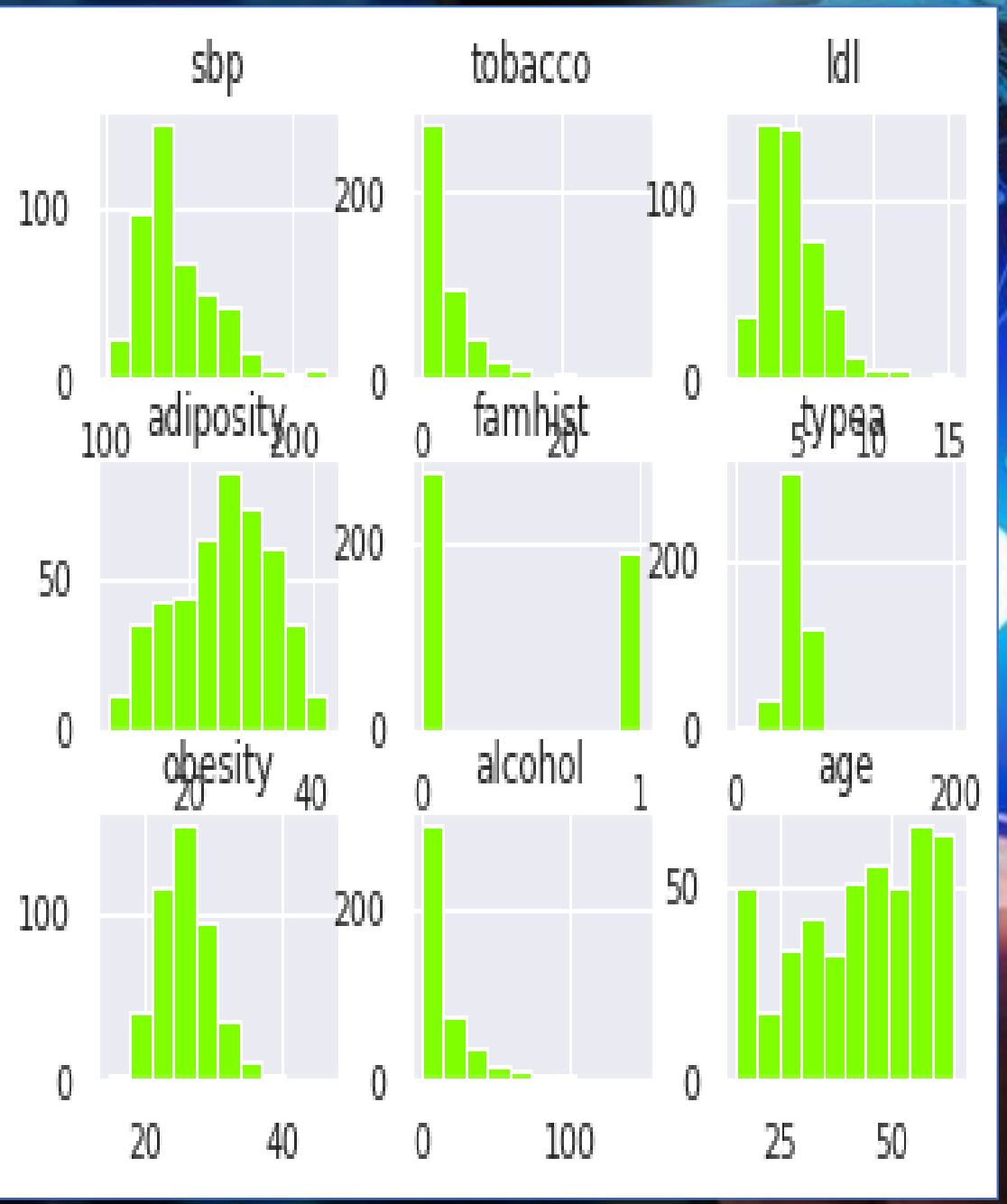
- Carga de Archivo: Cargarmos el Archivo Saheart,
- Importación del Dataset: Importamos el Archivo a Google Colad,
- Exploracion Inicial: Revisamos como esta compuesto nuestros Dataset.
- Revision de Datos: Revisamos los datos si hay datos nulos y espacios.
- Limpieza de los datos: Corregimos los espacios, datos nulos y nombres incorrectos de las columnas del Dataset.

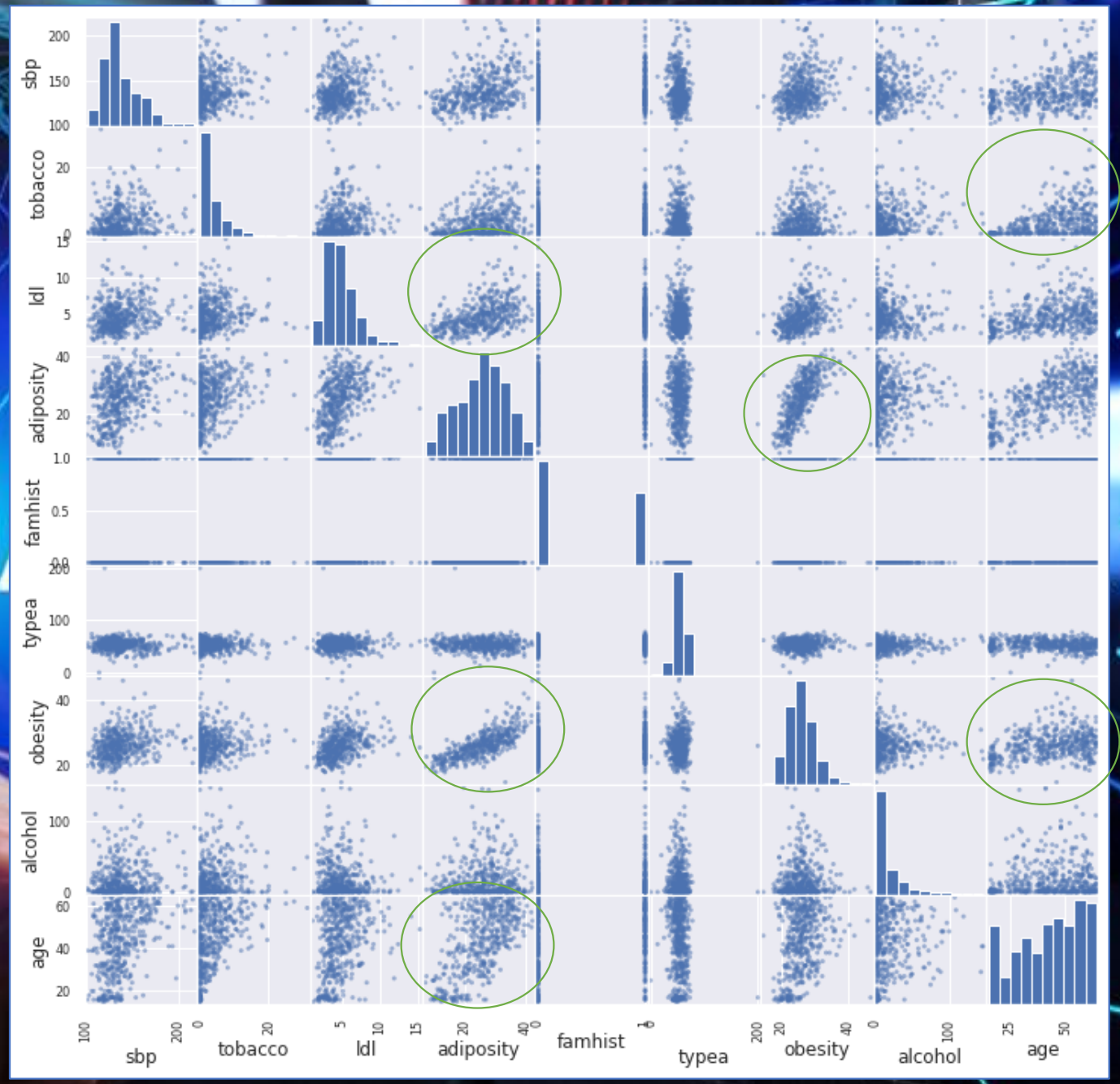

# EDA del Dataset

- Histograma
- Análisis
- Correlaciones
- Tablas Pivote

# EDA del Dataset

- Histogramas: Realizamos un histograma para cada variable.
- Analisis: Analizamos la media, mediana, moda, promedios, maximos y Boxplot del Dataset.
- Correlaciones: Explicamos las correlaciones de los gráficos utilizados.
- Tablas Pivote: Realizamos tablas Pivote para el trabajo de los datos según se solicito.


sbp
tobacco
ldl
adiposity
famhist
typea
obesity
alcohol
age


sbp
tobacco
ldl
adiposity
famhist
typea
obesity
alcohol
age

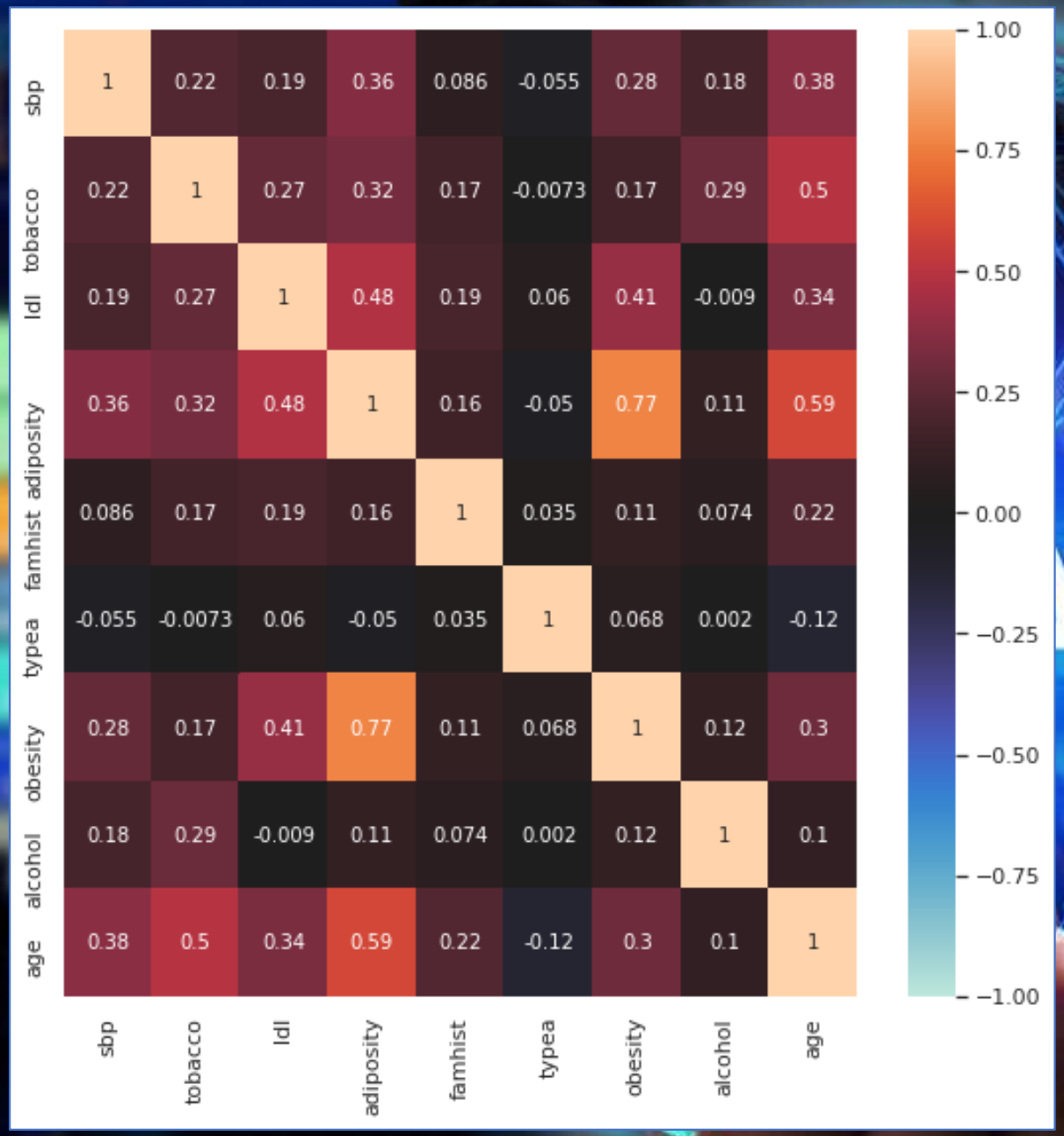



Tabla Pivote que muestra el nivel promedio de **adiposity** = adiposidad o grasa localizada (colesterol malo) entre aquellos que presentan o no historial familiar de enfermedad cardiaca y por rangos de edades

```
[234] round(data2.pivot_table('adiposity', index= 'age_range', columns= 'famhist', aggfunc= 'mean',fill_value=0),2)
```

| famhist | 0 | 1 |
|---|---|---|
| age_range | | |
| 15 | 15.44 | 15.19 |
| 20 | 17.61 | 14.77 |
| 25 | 19.51 | 18.98 |
| 30 | 23.50 | 21.82 |
| 35 | 22.30 | 22.79 |
| 40 | 26.66 | 29.18 |
| 45 | 27.83 | 28.65 |
| 50 | 31.72 | 30.30 |
| 55 | 29.09 | 28.73 |
| 60 | 29.37 | 30.10 |

# Ejemplo Tablas Pivote

Tabla pivote en relación del nivel de obesidad entre aquellos que tienen histirial familiar de enfermedad cardiaca (fanhist) por rangos de edad

```
[378] round(data2.pivot_table('obesity', index= 'age_range', columns= 'famhist', aggfunc= 'mean',fill_value=0),2)
```

| famhist | 0 | 1 |
|---|---|---|
| age_range | | |
| 15 | 22.69 | 21.97 |
| 20 | 24.12 | 21.95 |
| 25 | 25.08 | 24.85 |
| 30 | 26.44 | 26.51 |
| 35 | 25.51 | 26.49 |
| 40 | 26.04 | 27.10 |
| 45 | 26.91 | 27.51 |
| 50 | 29.03 | 27.67 |
| 55 | 26.27 | 26.82 |
| 60 | 26.36 | 26.41 |

# Ejemplo Tablas Pivote

```
[375] round(data2.pivot_table('sbp', index= 'age_range', columns= 'famhist', aggfunc= 'mean',fill_value=0),2)

      # De la siguiente tabla pivote se desprende la siguiente información, según el rango de edad la presion arterial sbp
      # es mayor segun la edad.
```

| famhist | 0 | 1 |
|---|---|---|
| age_range | | |
| 15 | 124.33 | 132.00 |
| 20 | 126.83 | 129.60 |
| 25 | 130.92 | 138.67 |
| 30 | 129.67 | 126.50 |
| 35 | 133.76 | 133.00 |
| 40 | 138.32 | 139.80 |
| 45 | 137.96 | 135.78 |
| 50 | 153.00 | 143.19 |
| 55 | 142.51 | 146.29 |
| 60 | 152.44 | 149.37 |


# Anova

- Detalle DataSet
- Pvalues
- Interpretación

# Anova

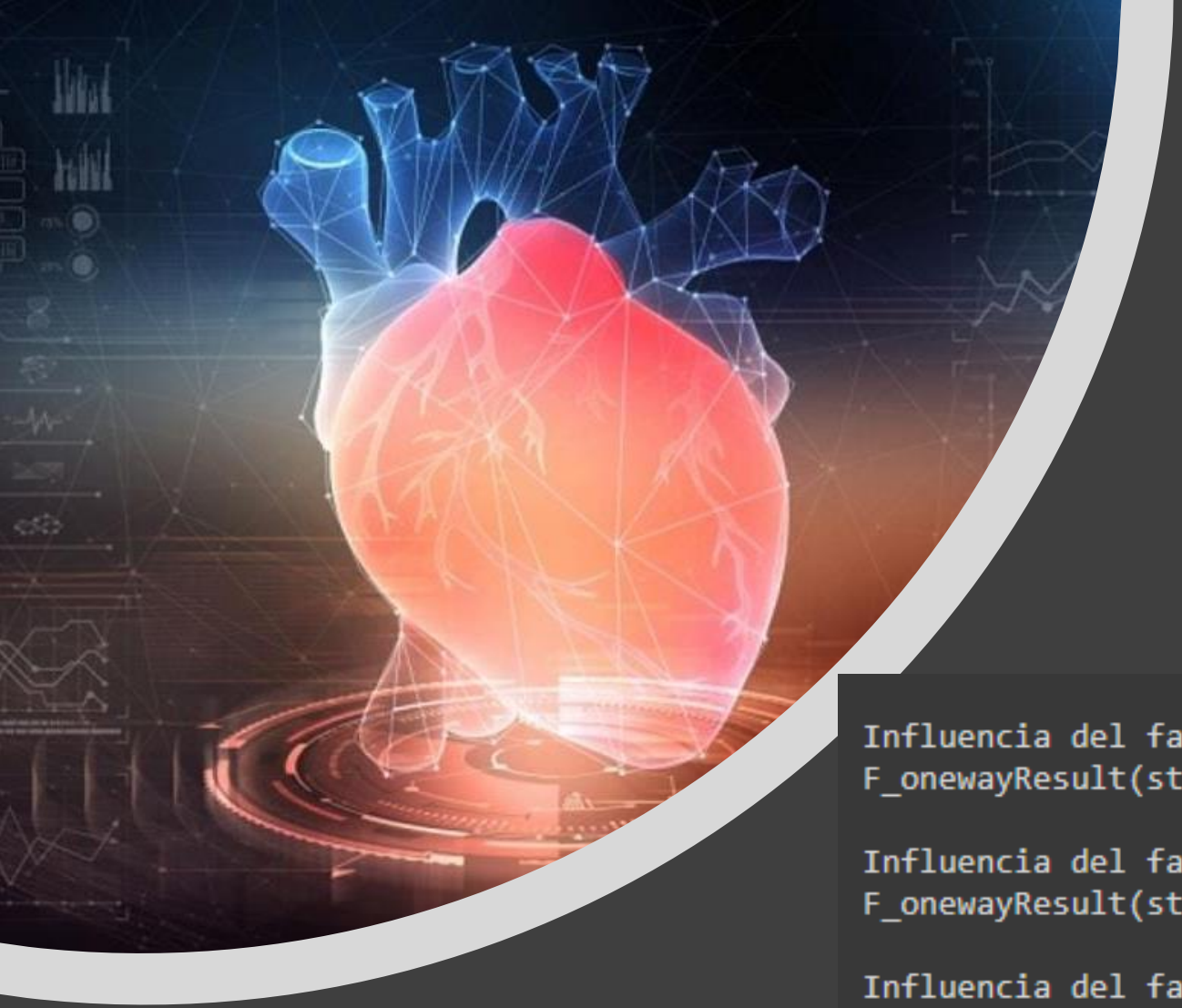
- **Detalle Dataset:** Realizamos el estudio está enfocado en población masculina de una edad entre 15 a 64 años en la región de Sudafrica

| | sbp | tobacco | ldl | adiposity | famhist | typea | obesity | alcohol | age |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 160 | 12.00 | 5.73 | 23.11 | 1 | 49.0 | 25.30 | 97.20 | 52 |
| 1 | 144 | 0.01 | 4.41 | 28.61 | 0 | 55.0 | 28.87 | 2.06 | 63 |
| 2 | 118 | 0.08 | 3.48 | 32.28 | 1 | 52.0 | 29.14 | 3.81 | 46 |
| 3 | 170 | 7.50 | 6.41 | 38.03 | 1 | 51.0 | 31.99 | 24.26 | 58 |
| 4 | 134 | 13.60 | 3.50 | 27.78 | 1 | 60.0 | 25.99 | 57.34 | 49 |

# Anova

- Pvalues: Analizamos cual de ellas se cumple y detallamos.

```
Influencia del famhist/ historial familiar sobre SBP presion arterial
F_onewayResult(statistic=3.1283087125281015, pvalue=0.07760618353588801)

Influencia del famhist/ historial familiar sobre fumado/tobaco
F_onewayResult(statistic=4.228471650801951, pvalue=0.040314814171165635)

Influencia del famhist/ historial familiar sobre LDL colesterol
F_onewayResult(statistic=12.35259223927411, pvalue=0.0004839705777565132)

Influencia del famhist/ historial familiar sobre adiposity/ grasa corporal
F_onewayResult(statistic=13.267259990639152, pvalue=0.0003008075023510091)

Influencia del famhist/ historial familiar sobre obesity/ obesidad
F_onewayResult(statistic=4.828885255572376, pvalue=0.028484503297451197)

Influencia del famhist/ historial familiar sobre alcohol
F_onewayResult(statistic=1.8913696629395682, pvalue=0.16971690028699915)

Influencia del famhist/ historial familiar sobre age/edad
F_onewayResult(statistic=26.872948746894696, pvalue=3.2602776527500005e-07)
```



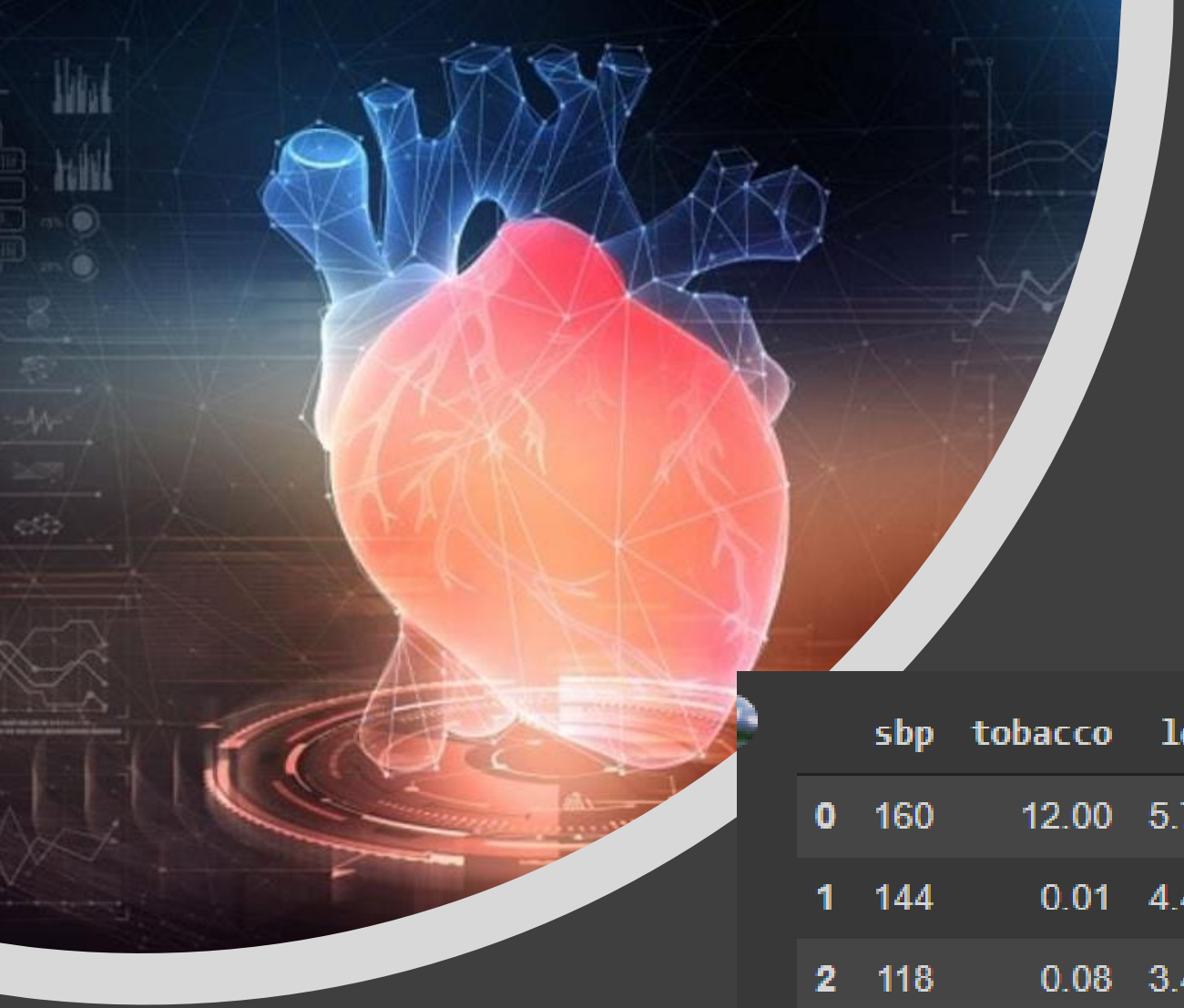
# Anova

- Interpretacion:  Realizamos el analisis y justificacion de los Pvalues que cumplen con lo solicitado.

```
Influencia del famhist/ historial familiar sobre fumado/tobaco
F_onewayResult(statistic=4.228471650801951, pvalue=0.040314814171165635)

Influencia del famhist/ historial familiar sobre LDL colesterol
F_onewayResult(statistic=12.35259223927411, pvalue=0.0004839705777565132)

Influencia del famhist/ historial familiar sobre adiposity/ grasa corporal
F_onewayResult(statistic=13.267259990639152, pvalue=0.0003008075023510091)

Influencia del famhist/ historial familiar sobre obesity/ obesidad
F_onewayResult(statistic=4.828885255572376, pvalue=0.028484503297451197)
```


# Machine learning

- Detalle DataSet Copia
- Extracción de Datos
- Interpretación y diagramas,



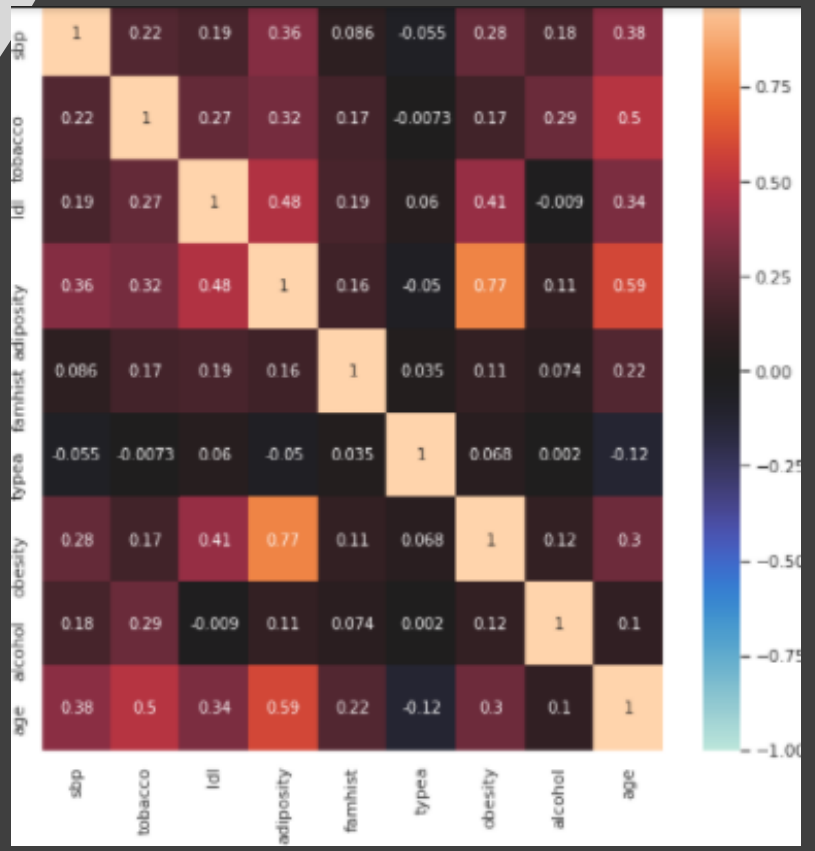
# Machine learning

- Detalle DataSet Copia:  Realizamos Realizamos Una copia de los datos.

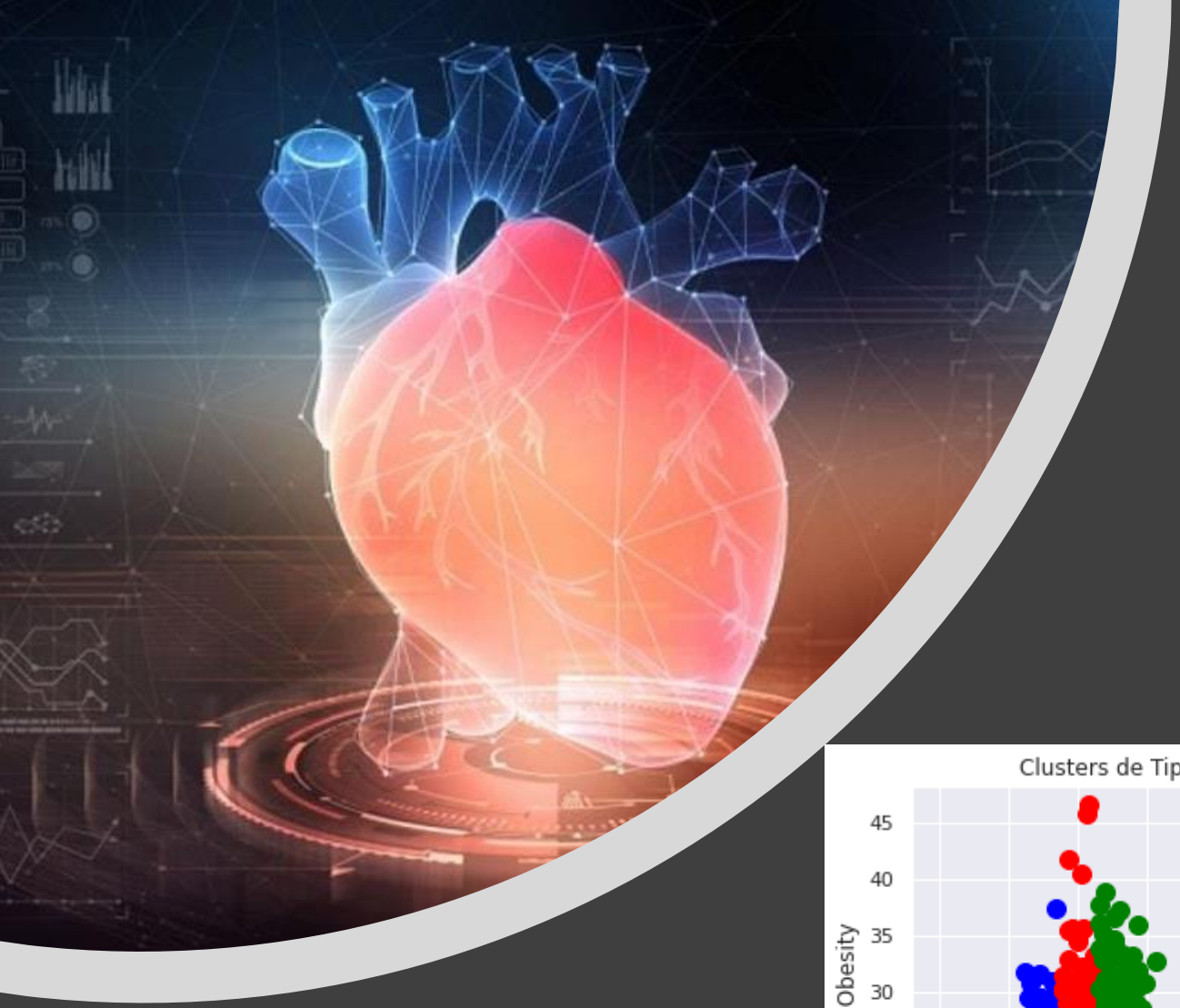
| | sbp | tobacco | ldl | adiposity | famhist | typea | obesity | alcohol | age |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 160 | 12.00 | 5.73 | 23.11 | 1 | 49.0 | 25.30 | 97.20 | 52 |
| **1** | 144 | 0.01 | 4.41 | 28.61 | 0 | 55.0 | 28.87 | 2.06 | 63 |
| **2** | 118 | 0.08 | 3.48 | 32.28 | 1 | 52.0 | 29.14 | 3.81 | 46 |
| **3** | 170 | 7.50 | 6.41 | 38.03 | 1 | 51.0 | 31.99 | 24.26 | 58 |
| **4** | 134 | 13.60 | 3.50 | 27.78 | 1 | 60.0 | 25.99 | 57.34 | 49 |

# Machine learning

- Extracción De Losa Datos:  Extraemos las variables que utilizaremos

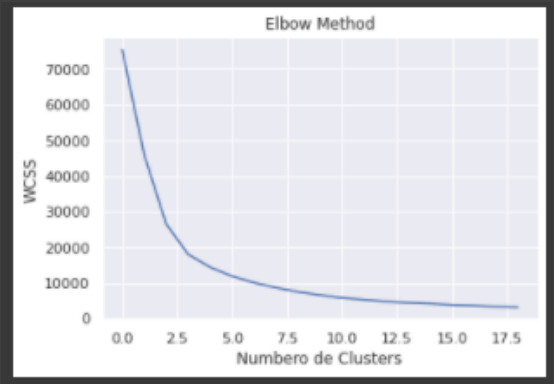


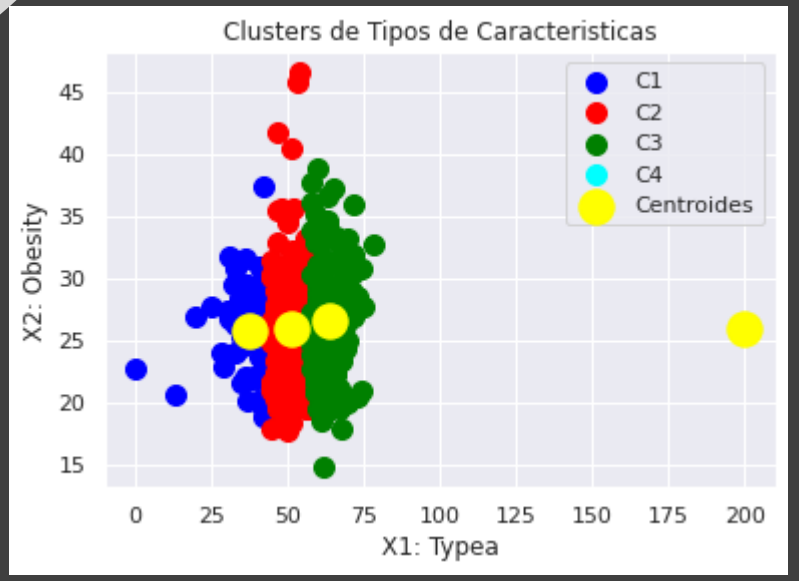

```
array([[ 49.  ,  25.3 ],
       [ 55.  ,  28.87],
       [ 52.  ,  29.14],
       [ 51.  ,  31.99],
       [ 60.  ,  25.99],
       [ 62.  ,  30.77],
       [ 59.  ,  20.81],
       [ 62.  ,  23.11],
       [ 49.  ,  24.86],
       [ 69.  ,  30.11],
       [ 72.  ,  26.81],
       [ 65.  ,  23.09],
       [ 59.  ,  21.57],
       [ 49.  ,  23.63],
       [ 54.  ,  23.53],
```

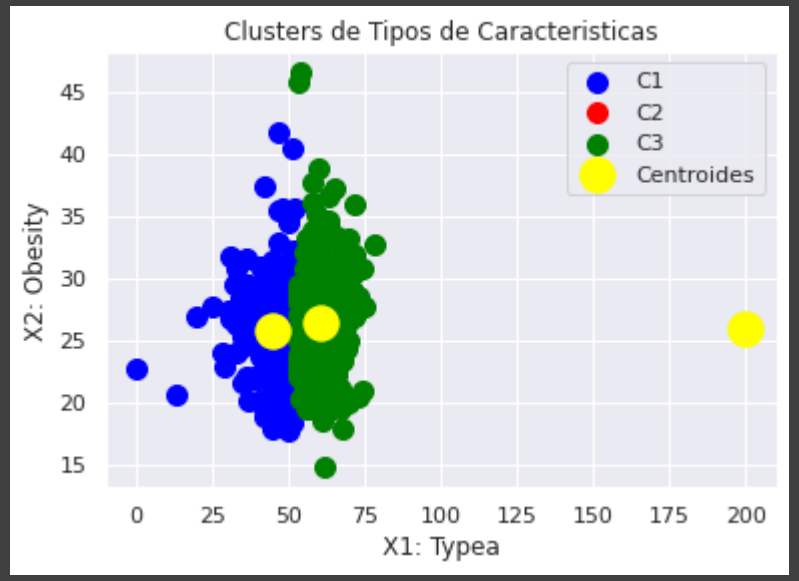


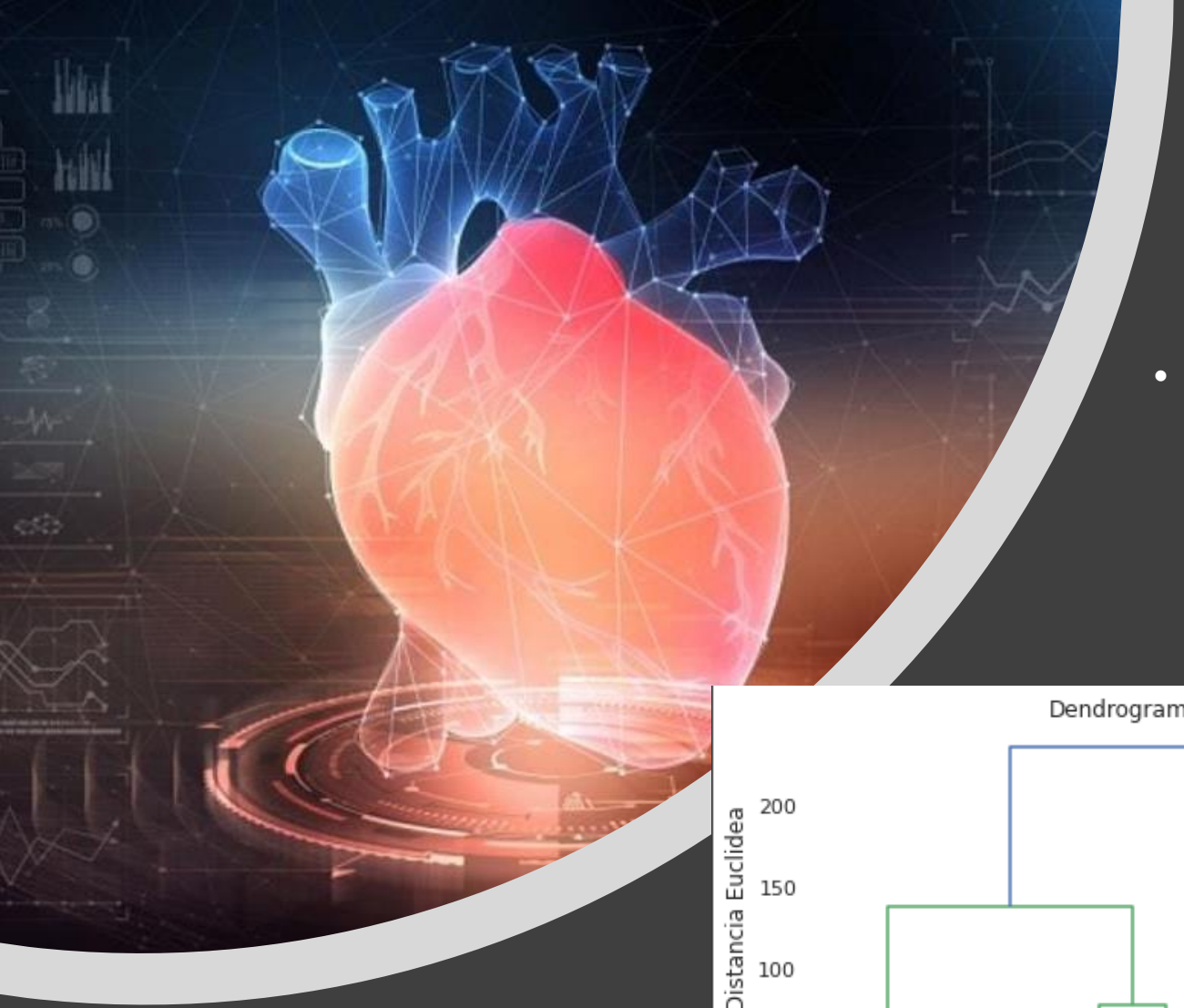
# Machine learning

- Interpretación y diagramas :  Revisamos las graficas y Kmeas e identificamos la posición de los clusters



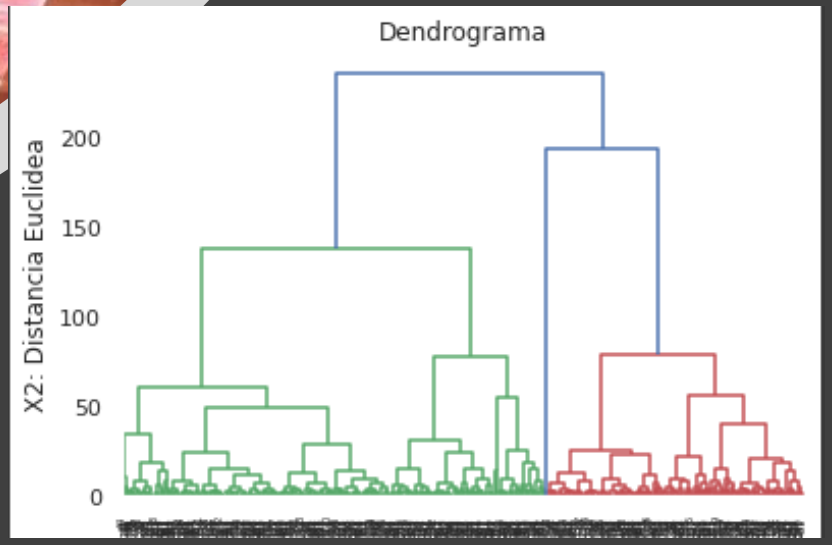



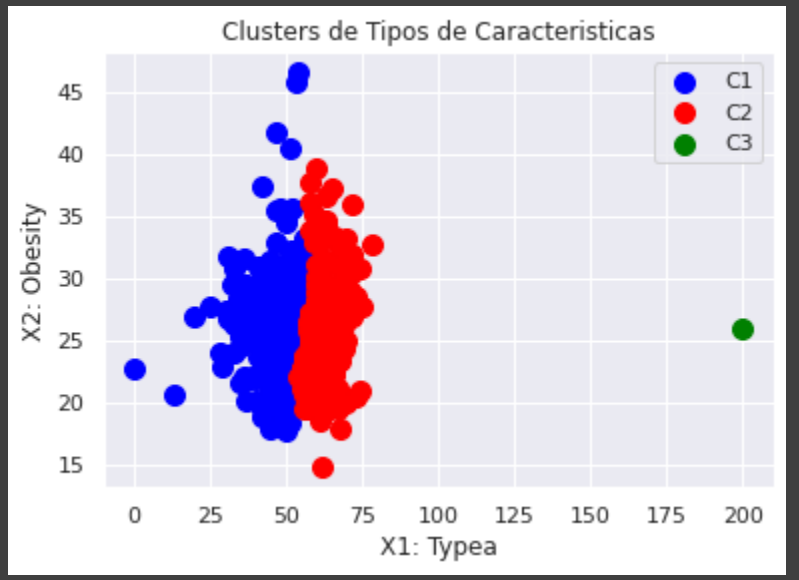





# Machine learning

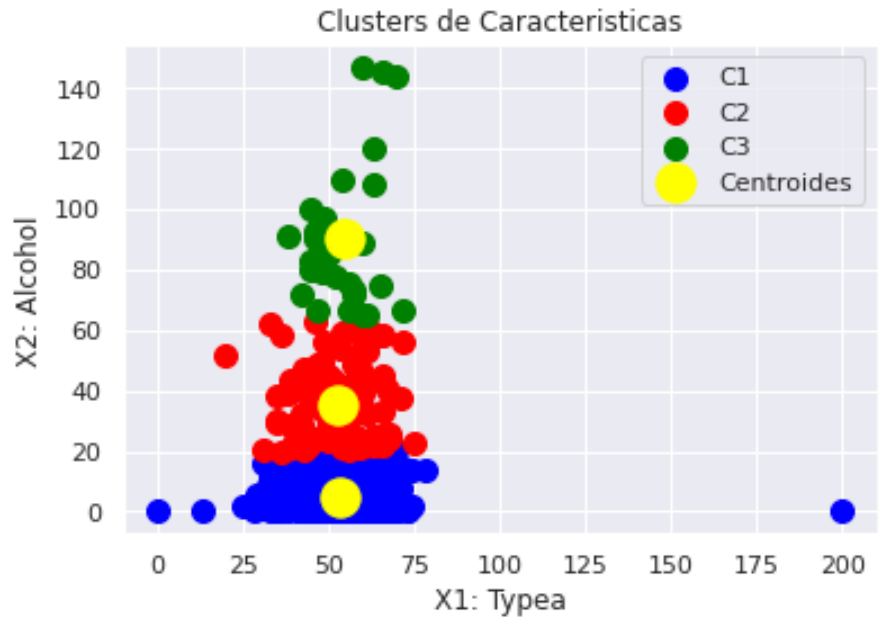
- Interpretación y diagramas :  Revisamos las graficas y Kmeas e identificamos la posición de los clusters

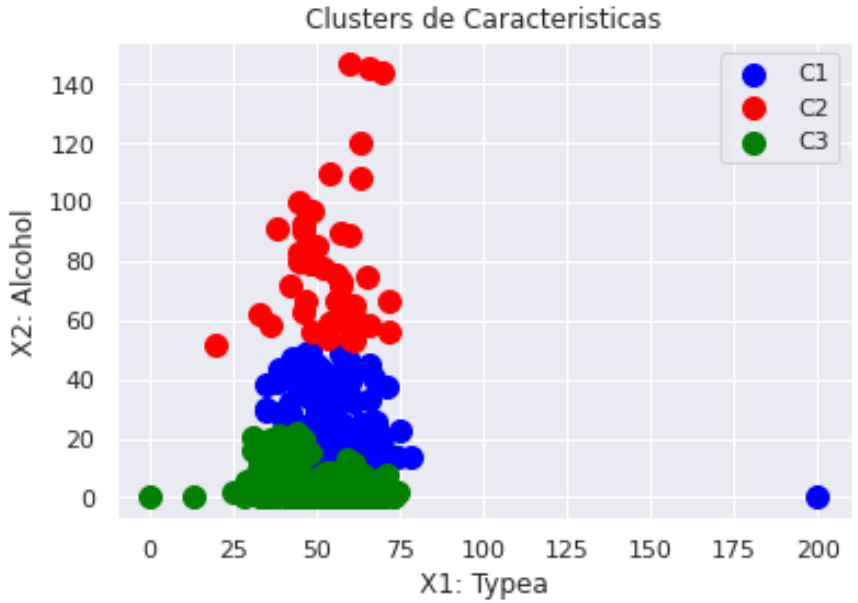


Machine learning

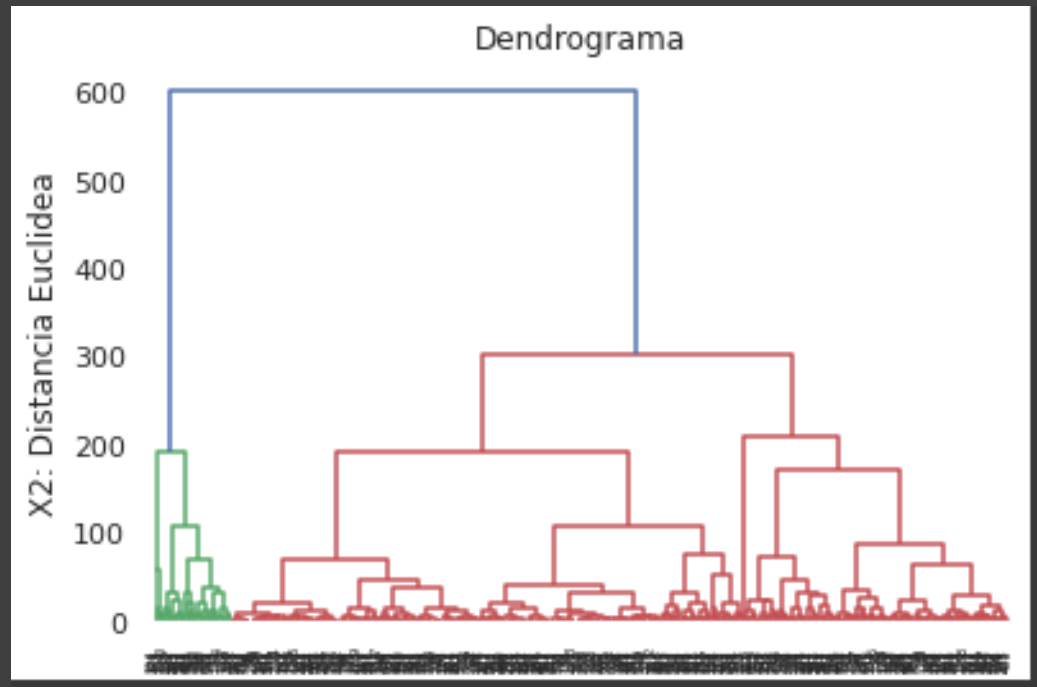
- Interpretación y diagramas : Revisamos las graficas y Kmeas e identificamos la posición de los clusters



# Machine learning

- Interpretación y diagramas :  Revisamos las graficas y Kmeas e identificamos la posición de los clusters

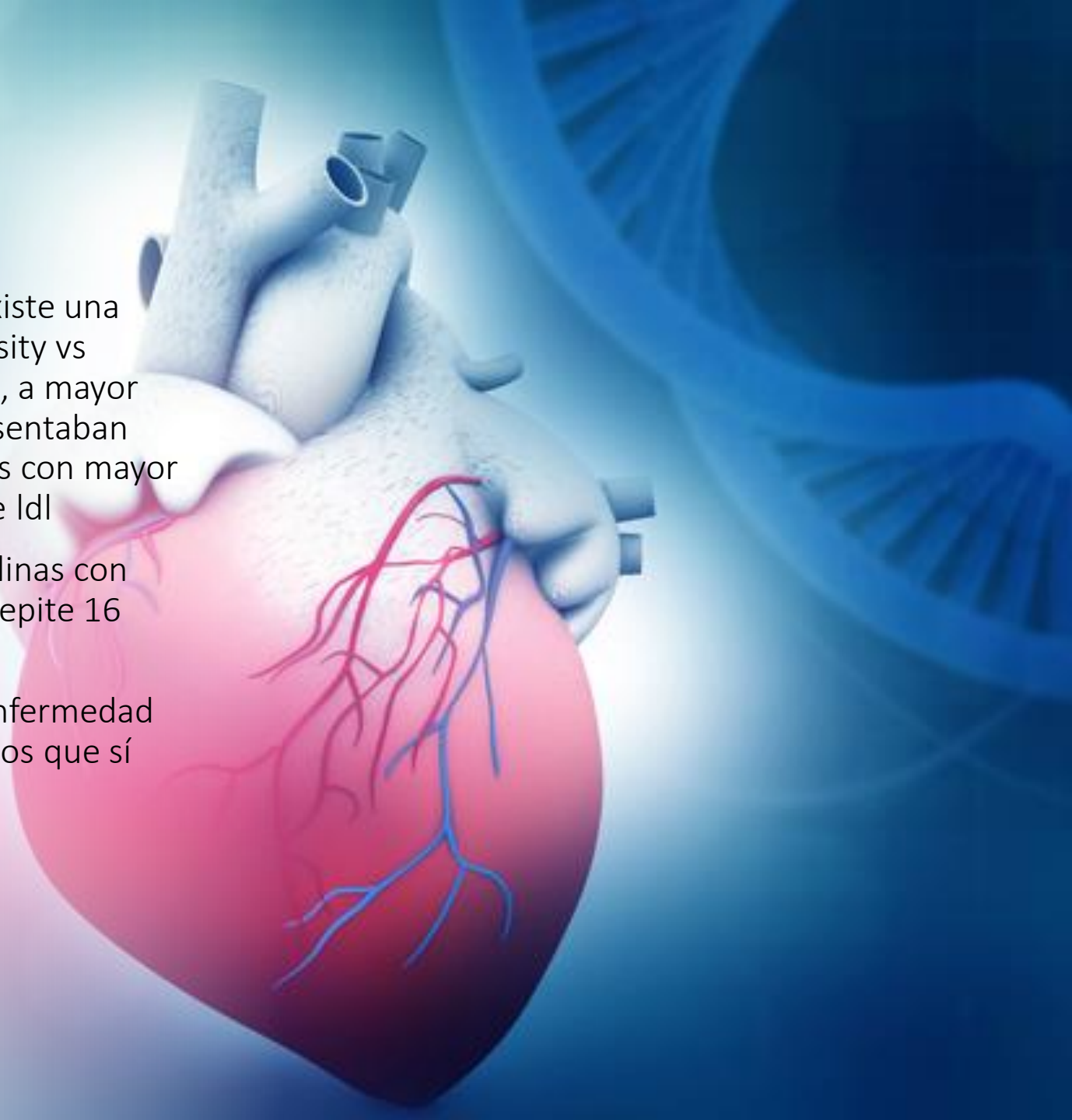


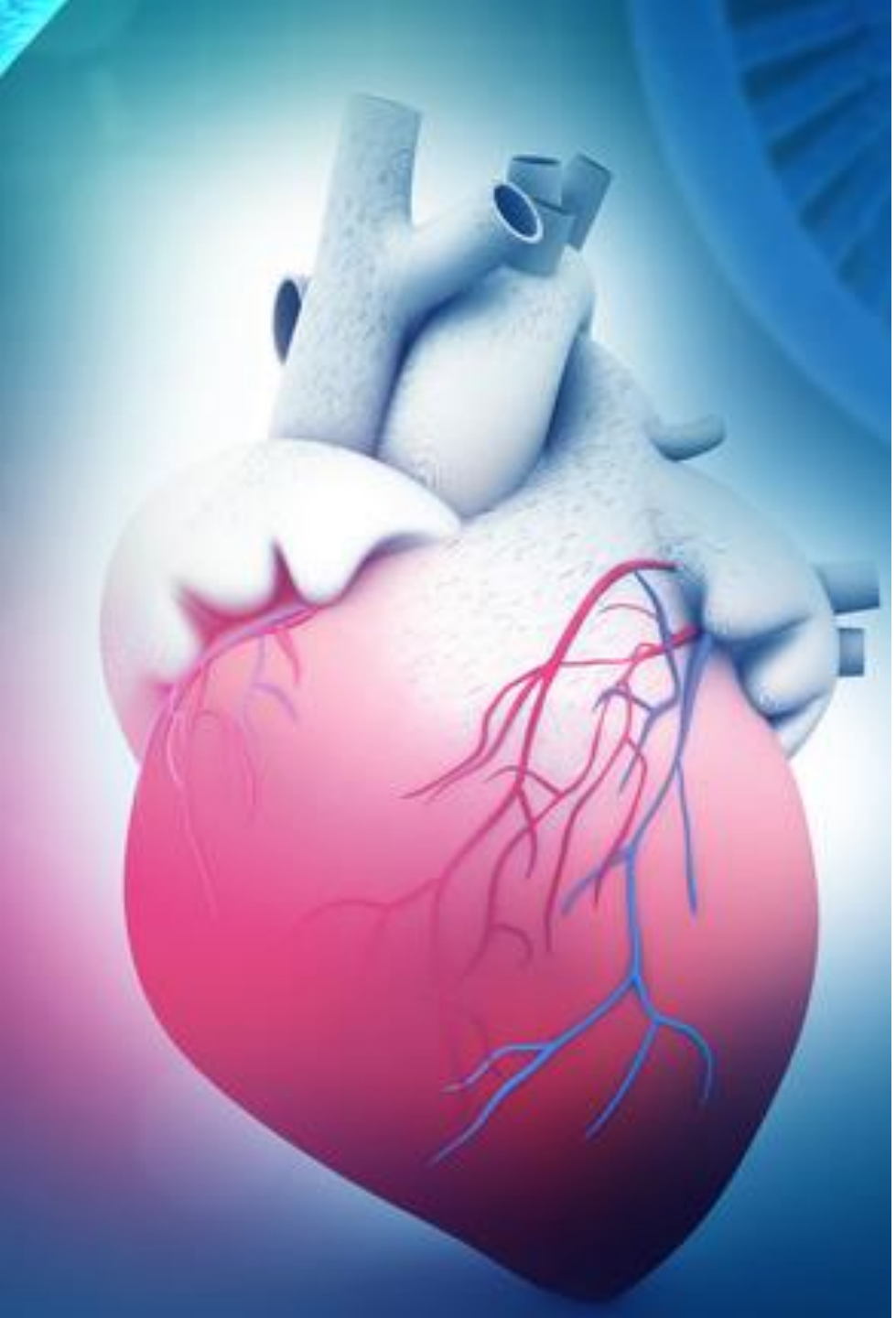
# Resumen

Según los datos procesados se pudo observar que existe existe una correlación positiva fuerte entre Adiposity vs obesity - Obesity vs Adiposity y Age vs Adiposity - Adiposity vs Age. por lo tanto, a mayor edad se presenta un mayor nivel de obesidad, quienes presentaban mayor nivel de consumo de tabaco acumulado son aquellos con mayor edad, así también el nivel de obesidad influye en el nivel de ldl

La muestra de las personas corresponde a personas masculinas con edades entre los 15 a 64 años, siendo la edad que más se repite 16 años.

Además, que aquellos que no tienen historial familiar de enfermedad cardiaca (famhist) tienden a fumar un poco más que aquellos que sí presentan un historial familiar de enfermedad cardiaca

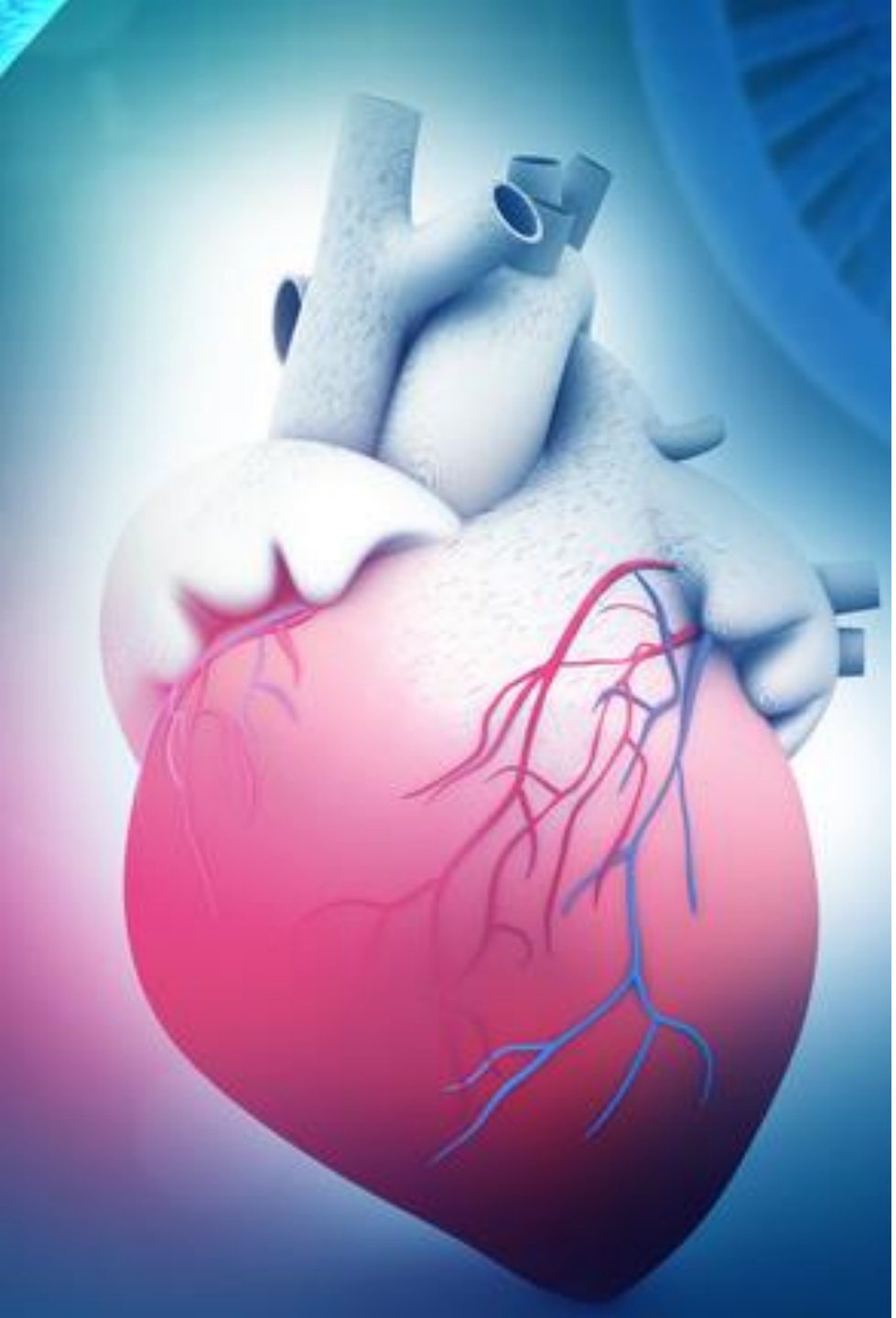




# Conclusiones y Recomendaciones

Es importante el cuido de la salud, y este dataset pone en evidencia una problemática actual, los problemas cardiovasculares, sin embargo un infarto puede ser producto de alguna otra afectación «fantasma» que puede presentar o no algunos síntomas, hasta terminar en un evento fulminante, y aquellos factores se pueden agravar si se suman conductas o hábitos que incidan en los factores de riesgo. En este dataset SAHeart, a falta de una característica final en el la que indique que los participantes sufrieron o no un problema o afectación cardiaca, se tomaron las variables que se consideran de mayor peso en incidir en un daño coronario, y se procede a indicar algunos datos de investigaciones.

2009 «La PAS, como factor de riesgo independiente y modificable, sino que se le señala como el principal determinante del riesgo cardiovascular en personas mayores de 50 años de edad, y especialmente en ancianos». «Los valores de PAS por encima de 140 mmHg pueden elevar considerablemente el riesgo cardiovascular, especialmente en ancianos. Por ese motivo, en caso de objetivarse valores superiores a los citados en personas con más de 50 años de edad, deberá intensificar su actividad educativa para mejorar el control»



# Conclusiones y Recomendaciones

Habito de Fumado y el Riesgo cardiovascular o coronario El Tabaquismo es uno de los factores de riesgo más importantes para el desarrollo de la enfermedad cardiovascular, Este riesgo es importante para cualquier grupo etario y es proporcional al número de cigarrillos diarios consumidos y al tiempo de duración de la adicción o exposición . Tema central: Riesgo cardiovascular páginas 699-705 Colesterol-LDL elevado y el Riesgo cardiovascular o coronario Colesterol-LDL elevado se asocia con un riesgo aumentado de enfermedad cardiovascular, tal y como lo indica Briceño y Acquatella El colesterol elevado es uno de los factores de riesgo cardiovascular modificable más importante junto con las modificaciones del estilo de vida.


BIG DATA