Trabajo final de estadística con R:

Análisis de diversas técnicas aplicadas a datos de fallos cardiacos

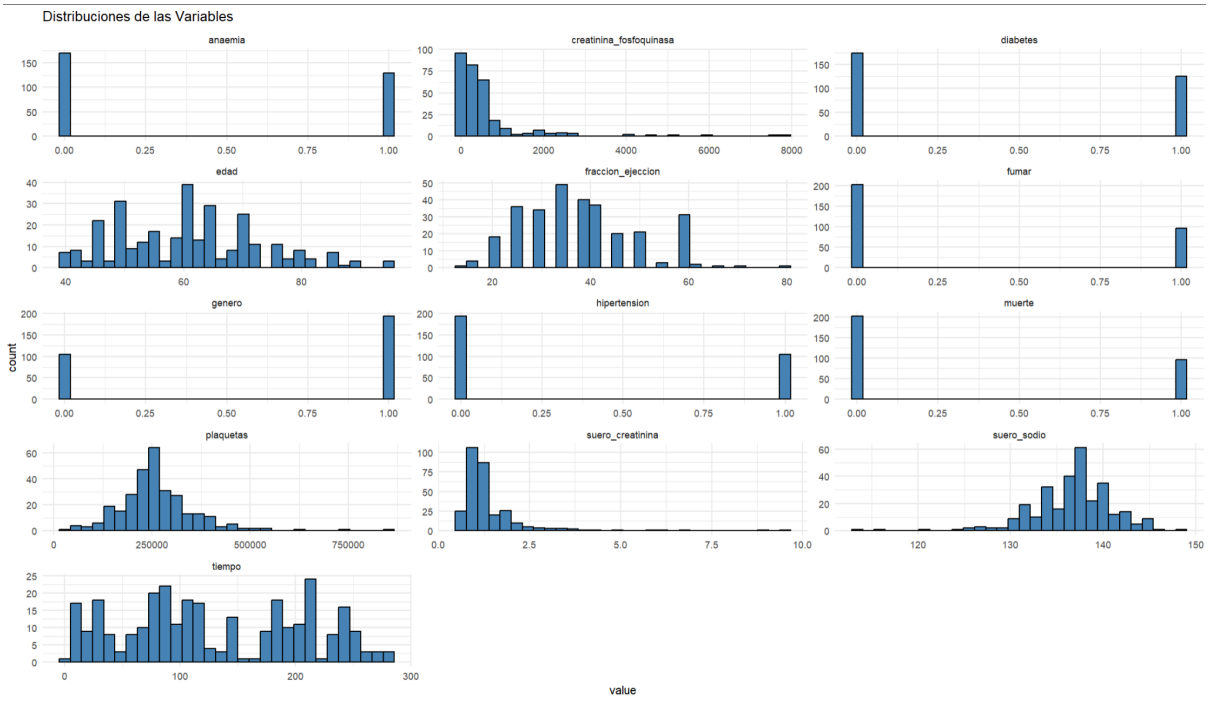
Adam Maltoni
Ibón de Mingo Arroyo
Antonio Peña

UAM
Escuela Politécnica Superior

Índice: tabla de contenidos del proyecto

Introducción

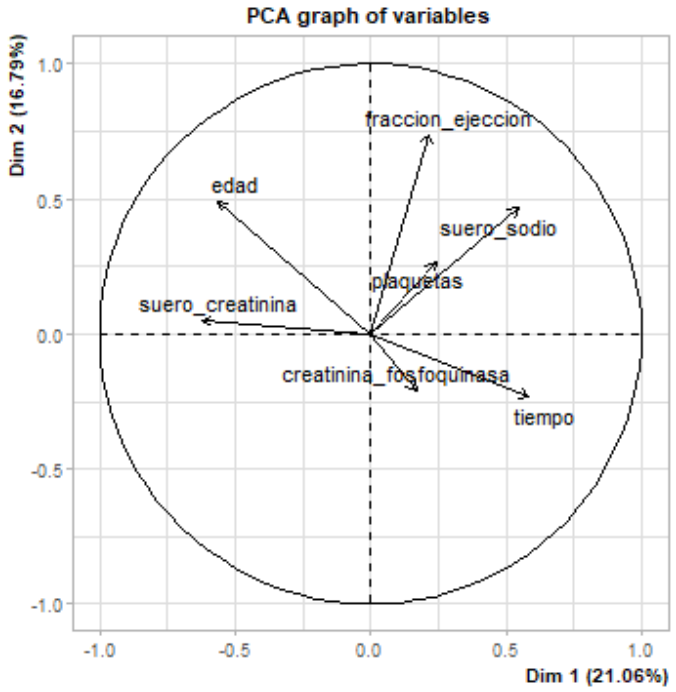
En este trabajo, se explora un dataset de fallos cardíacos utilizando análisis y clasificación de datos. Se aplican métodos de reducción de dimensionalidad como PCA y PLS, junto con clustering para agrupar pacientes con características similares. Finalmente, se emplea LDA para evaluar la capacidad de clasificación de las variables, con el objetivo de aportar conocimiento clínico sobre perfiles de riesgo y posibles intervenciones. A modo de breve análisis exploratorio inicial, mostramos las distribuciones de los datos (299 obs., 15 variables).



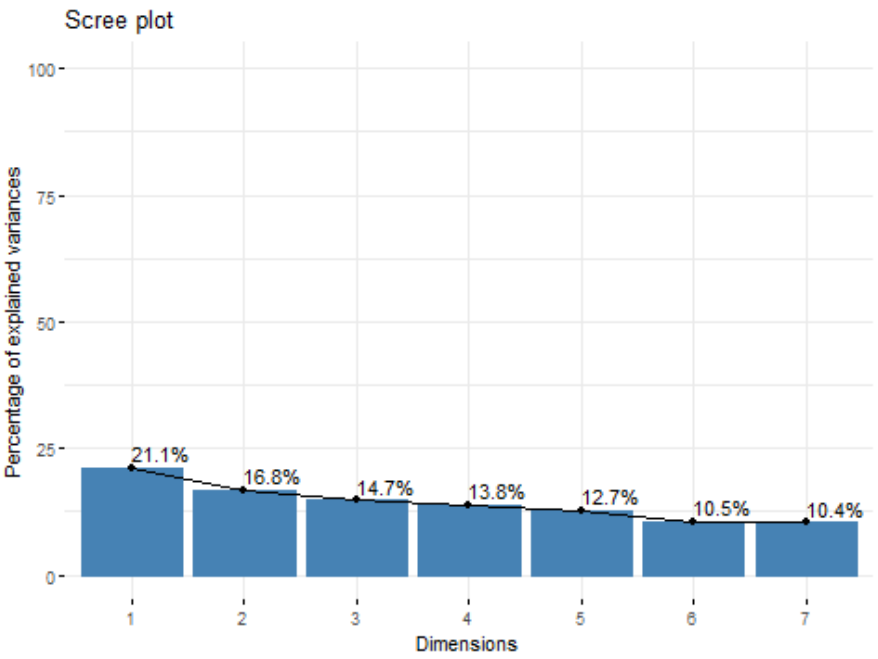
Análisis de componentes principales (ACP)

El objetivo de un ACP siempre es el mismo: (1) identificar patrones subyacentes en las variables continuas del conjunto de datos; (2) reducir la dimensionalidad a la par que se conserva la mayor parte de la varianza; y (3) visualizar los datos en un espacio de una dimensión manejable (por ejemplo, visualizar en 2 dimensiones un dataset con 27).

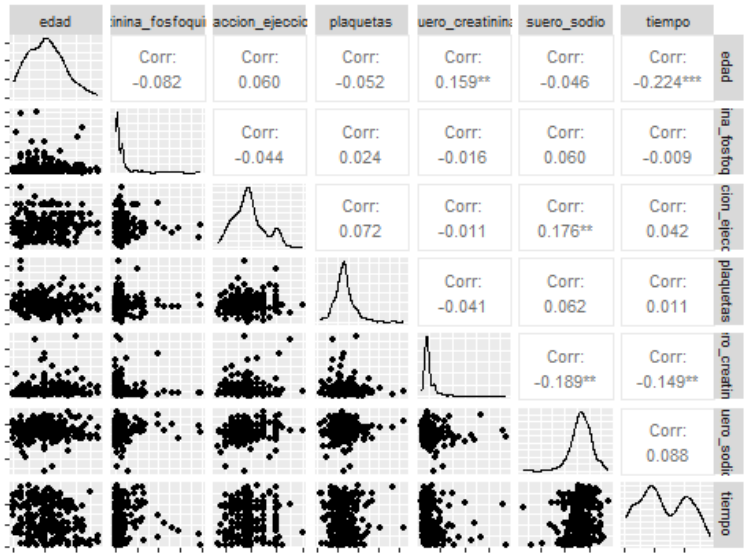
Para ello, como ACP solo funciona con variables numéricas vamos a escoger única y exclusivamente las variables continuas que encontramos en el conjunto. A saber: `edad`, `creatinina_fosfoquinasa`, `fraccion_ejeccion`, `plaquetas`, `suero_creatinina`, `suero_sodio` y `tiempo`. Tras calcular las variables correspondientes con el método `PCA` de R nos encontramos con lo siguiente



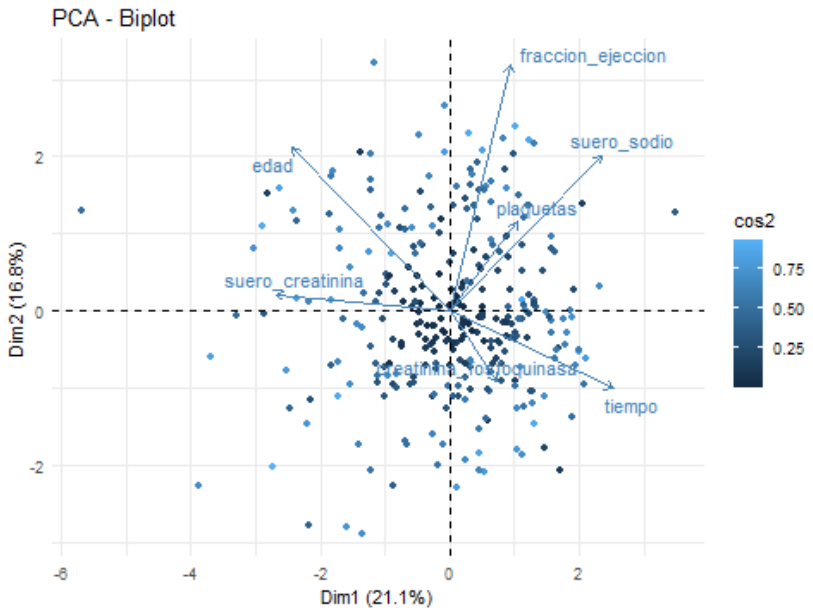
Lo cual viene a decir que las variables no están especialmente correlacionadas ni tampoco muy bien representadas en el espacio de dos dimensiones. De hecho, si nos fijamos podemos ver que la varianza explicada por las dos primeras componentes es de tan solo un 37,85%. Si nos fijamos en el screeplot:



Con lo que necesitamos hasta 5 variables para llegar a acumular casi un 80% de la variabilidad (79,1% con las primeras cinco variables). Al estar considerando 7 para el análisis, no resulta muy útil como herramienta de reducción de variables. La mayor parte de la información (varianza) se pierde en el proceso a no ser que mantengamos la gran mayoría de las variables. La baja correlación de las variables entre sí está muy clara al contemplar el siguiente gráfico



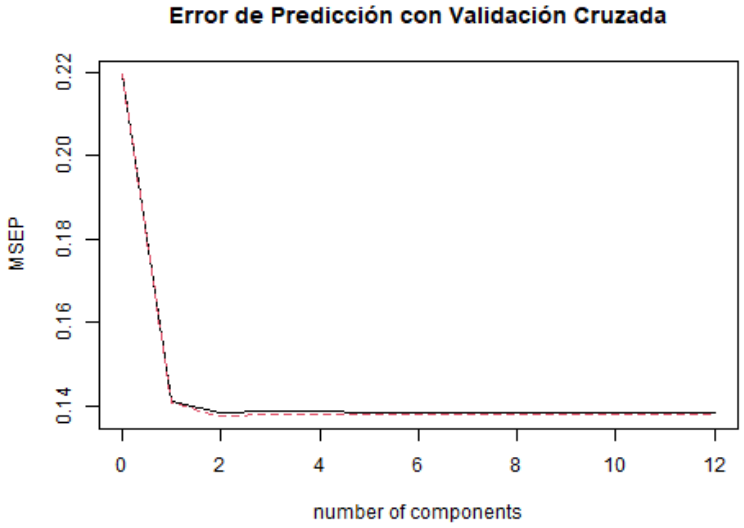
La correlación más fuerte se presenta entre `edad` y `tiempo`, con tan solo un -0,224. Por último, la representación de los datos puede visualizarse mediante un biplot con cierto `alpha` (para no sobrecargar ciertas zonas con muchos puntos). Añadiendo colores claros, el resultado del código es



Como puede verse fácilmente los puntos se distribuyen de forma más o menos aleatoria y uniforme a lo largo del gráfico, con una notoria concentración de puntos alrededor del centro.

Partial Least Squares (PLS)

La utilidad fundamental de PLS es alcanzar una elevada capacidad predictiva descontando posibles correlaciones entre las variables predictoras. En este caso queremos predecir si una persona morirá o sobrevivirá de acuerdo con los datos de los que disponemos, que son los proporcionados en el dataset. Tras entrenar el modelo de la forma vista en clase vemos que con unos pocos componentes podemos reducir el error drásticamente hasta casi cero.



El error medio comienza siendo bajo pero se reduce hasta niveles irrisorios de un salto al incluir más de uno o dos componentes. Como resultado final habiendo escogido tan solo 2 componentes obtenemos una exactitud muy alta, de 0,84280; con una matriz de confusión caracterizada por ser bastante fiel a los datos reales.

Matriz de confusión

Real \ Predicho	Falso	Verdadero
Falso	185	18
Verdadero	29	67

El análisis revela que ciertos tipos de pacientes pueden tener un mayor riesgo de muerte que otros según ciertas variables clave. Por ejemplo, las personas de edad avanzada con diabetes o que fumen. Las personas jóvenes con relativa salud tienen un riesgo ínfimo de padecer un paro cardíaco.

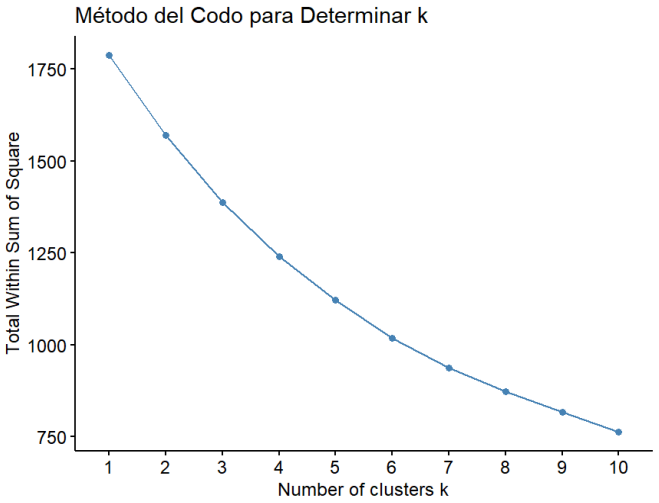
Análisis de clúster

El análisis de clusters es una técnica de aprendizaje no supervisado que busca identificar patrones en conjuntos de datos al agrupar observaciones similares en grupos o «clusters». Cada cluster contiene elementos que son más parecidos entre sí que a los elementos de otros clusters, según alguna métrica de similitud o distancia.

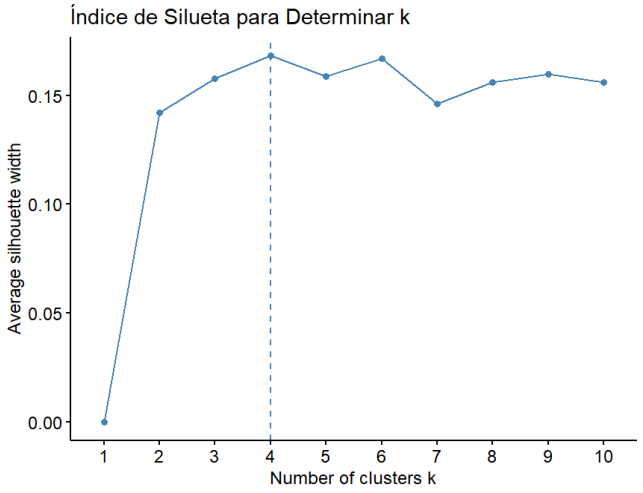
En el contexto de datos médicos, como en nuestro caso, esta técnica es útil para identificar grupos de pacientes con características clínicas similares. Esto puede ser clave para diseñar tratamientos personalizados, detectar factores de riesgo comunes y mejorar la precisión de diagnósticos. Por ende, podemos intuir que a priori se trata de una técnica adecuada para nuestro dataset.

Para el análisis se han seleccionado las variables continuas y enteras que representan indicadores de la salud, es decir, `edad`, `creatinina fosfoquinasa`, `fracción de eyección`, `suero creatinina`, `suero sodio` y `plaquetas`. No se han seleccionado las variables categóricas debido a que, por lo general, no están incluidas en las técnicas comunes de clustering que vamos a utilizar (sobretodo K-means); usualmente, se suelen mezclar los clusters obtenidos con otras variables categóricas para su interpretación al final del análisis, o se pueden hacer uso de otros algoritmos como las K-modes, las K-means con distancia de Gower o el algoritmo KAMILA.

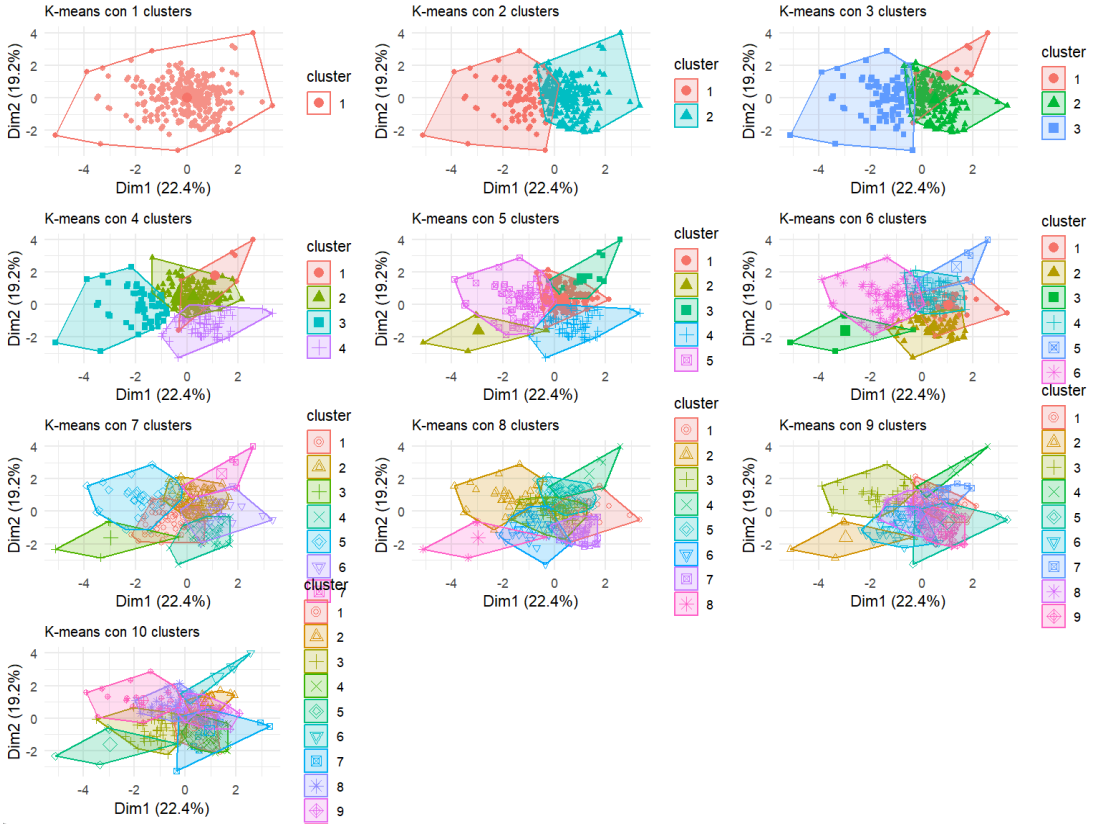
Comenzando con el análisis en sí, primero escalaremos los datos para asegurarnos de que pesamos todas las variables de igual manera en el cálculo de distancias. A continuación, intentaremos determinar el número óptimo de clusters a utilizar mediante dos métodos: el método del codo (WSS) y el índice de la silueta.



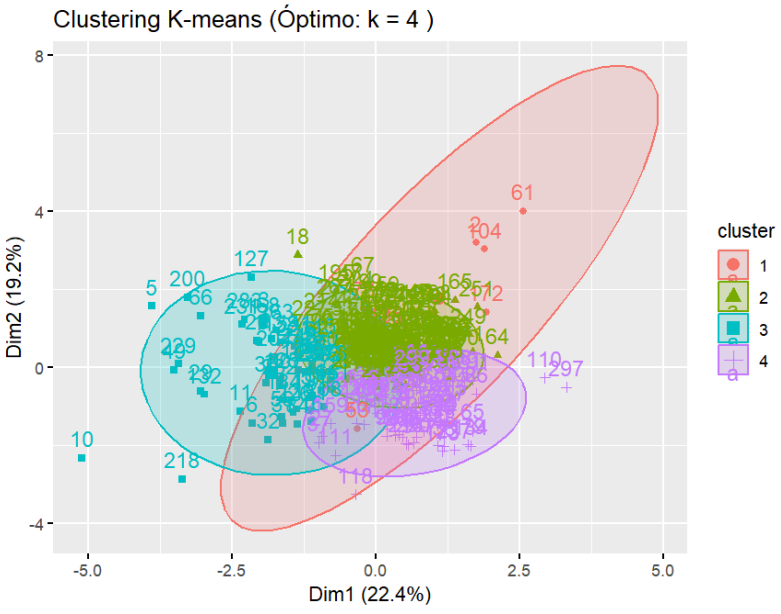
En este caso, ciertamente elegiríamos considerar unos 6 clusters como el número óptimo de ellos, aunque no parece haber un punto de inflexión claro en la curva definida por la WSS. Por otro lado, según el método de silueta obtenemos como clustering óptimo 4 grupos, como podemos ver a continuación buscando el punto con mayor índice de silueta en la gráfica:



Además, y para terminar de confirmar nuestras hipótesis, se elabora un plot donde probamos cómo quedarían agrupados los datos para cada número de clusters entre 1 y 10 según un algoritmo de K-means.



Se puede ver que, a partir de los datos, añadir más clusters a partir de 4 o 5 no parece hacer mucho efecto a nivel de separación, aunque esto puede ser debido a la estructura intrínseca de la representación (en 2D, se están representando solamente en función de los dos atributos que más variabilidad explican, y, como se menciona en el apartado de PCA, la variabilidad del dataset está muy bien repartida entre variables, como podemos ver por el 22.4% y 19.2% de los ejes). Los datos dan la impresión de estar muy centralizados en este plot.



Analizamos los centroides obtenidos para comprender cómo se diferencia cada grupo, clínicamente:

Cluster 1:
- Edad media: 60
- Creatinina media (suero): 1.854286
- Fracción de eyección media: 37.85714
- Plaquetas medias: 277388
- Suero Sodio medio: 138.5714
- Tamaño del cluster: 7

Pacientes mayores con creatinina elevada, función cardíaca moderada, y niveles normales de sodio, posiblemente con daño renal.

Cluster 2:
- Edad media: 55.77019
- Creatinina media (suero): 1.119938
- Fracción de eyección media: 32.98137
- Plaquetas medias: 252995.8
- Suero Sodio medio: 137.4534
- Tamaño del cluster: 161

Adultos con función cardíaca reducida y creatinina normal, probablemente en estadios iniciales de insuficiencia cardíaca.

Cluster 3:
- Edad media: 69.78182
- Creatinina media (suero): 2.516
- Fracción de eyección media: 34.18182
- Plaquetas medias: 226791
- Suero Sodio medio: 131.5455
- Tamaño del cluster: 55

Ancianos con severa disfunción renal, niveles bajos de sodio y moderada insuficiencia cardíaca avanzada.

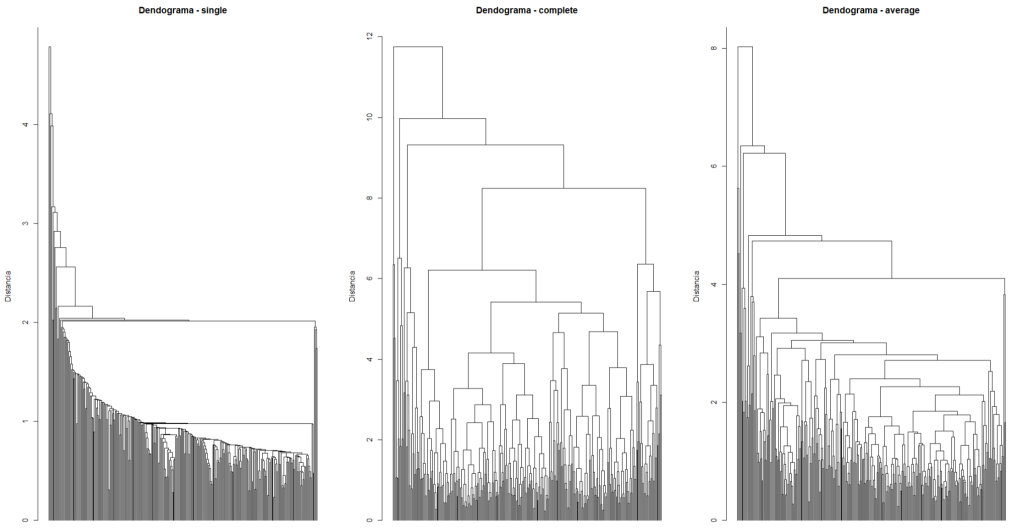
Cluster 4:
- Edad media: 65.17105
- Creatinina media (suero): 1.119737
- Fracción de eyección media: 51.73684
- Plaquetas medias: 310480.3
- Suero Sodio medio: 138.3684
- Tamaño del cluster: 76

Pacientes mayores con función cardíaca preservada, creatinina normal y buen estado general, posiblemente en recuperación o con control adecuado.

Vemos que si, por ejemplo, pusiéramos 10 clusters, los centroides no variarían mucho entre sí en atributos, y, además, terminaríamos con varios clusters con muy pocos individuos.

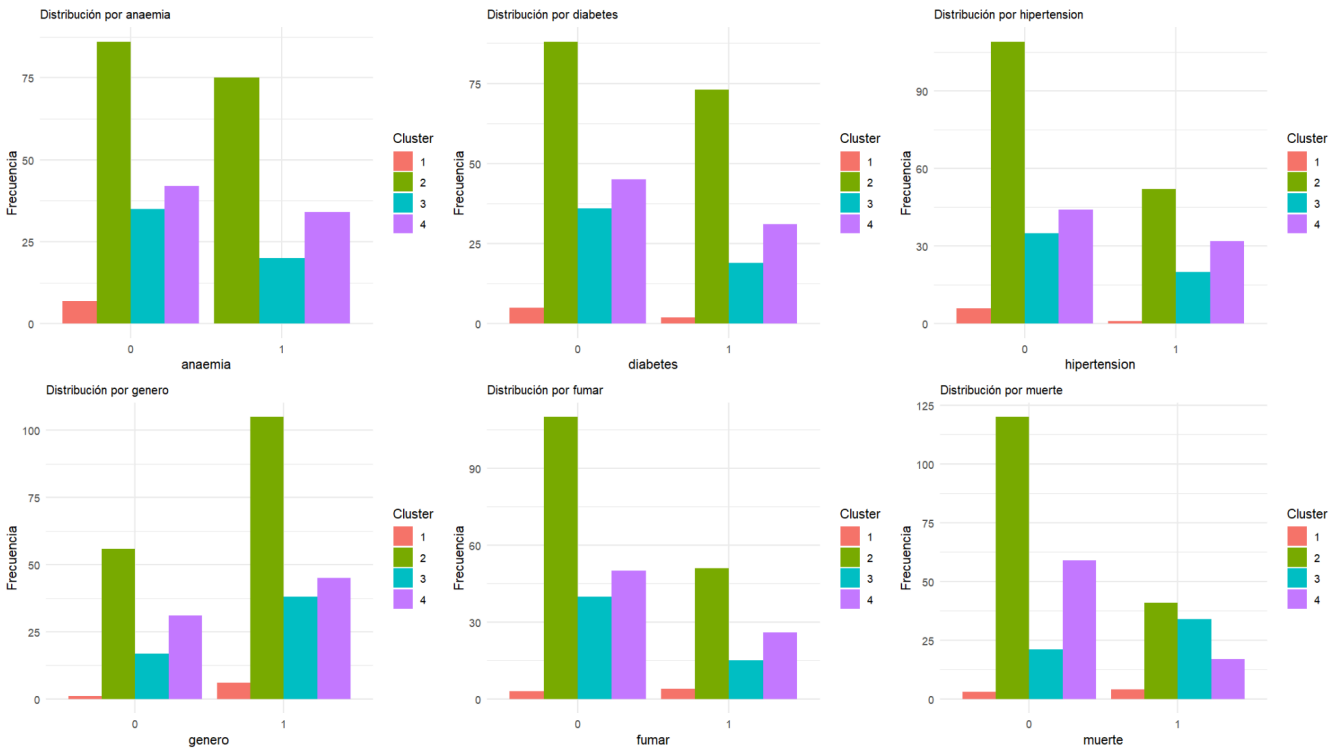
(Nótese que los "diagnósticos" han sido emitidos con ChatGPT a falta de conocimientos médicos por parte de los estudiantes).

Ahora, pasaremos al clustering jerárquico, con tres estrategías de *linkage*, considerando la distancia mínima, máxima y media entre puntos/clusters durante la construcción de estos (*single*, *complete*, *average*). Obtenemos tres dendogramas que podemos comparar:



Como vemos en los dendogramas, las tres métricas de distancia no se corresponden particularmente a la hora de crear clusters, lo que sugiere que no se están encontrando relaciones del todo naturales entre los datos.

A posteriori, los resultados muestran que el análisis de clusters ha revelado agrupaciones que tienen sentido clínico y estadístico, dado que los grupos se diferencian en términos de edad, creatinina sérica, fracción de eyección, etc. Sin embargo, vemos baja concordancia entre los tres métodos jerárquicos (cambio de la distancia) utilizados, lo que puede sugerir que los datos podrían no estar perfectamente separados en grupos naturales. Además, sería ideal validar estos resultados con conocimiento clínico o correlaciones con desenlaces médicos, por ejemplo con las variables categóricas dejadas atrás antes como `fumar` o `muerte`.



Como conclusión extraída, el análisis de clusters ha proporcionado una segmentación útil del dataset, destacando diferencias clínicas significativas entre los 4 grupos considerados, pero hay espacio de mejora en la técnica mediante la consideración de las variables binarias y a causa de la baja concordancia entre los métodos jerárquicos, entre sí, y de estos con el método de K-medias no jerárquico considerado, ya que al final del estudio obtenemos una concordancia de tan solo el 2.01% entre los métodos jerárquicos y el no jerárquico.

Análisis Discriminante Lineal

El análisis discriminante lineal (LDA) es un análisis equivalente al de regresión lineal en el que se clasifican variables categóricas. El objetivo es identificar relaciones lineales entre las variables continuas para que podamos diferenciar entre los grupos prefijados. Además, se intenta definir una regla de decisión que permita clasificar correctamente un nuevo objeto, del que no sabemos a qué grupo pertenece, asignándole al grupo más adecuado. Para llevarlo a cabo es importante que se cumplan una serie de suposiciones:

Las variable objetivo debe ser categórica, lo cual en nuestro conjunto de datos cumplimos ya que es binaria. Y las variables que se eligen para hacer el análisis deben ser continuas, por ellos se seleccionan las siguiente variables: `edad`, `creatinina fosfoquinasa`, `fracción de eyección`, `suero creatinina`, `suero`

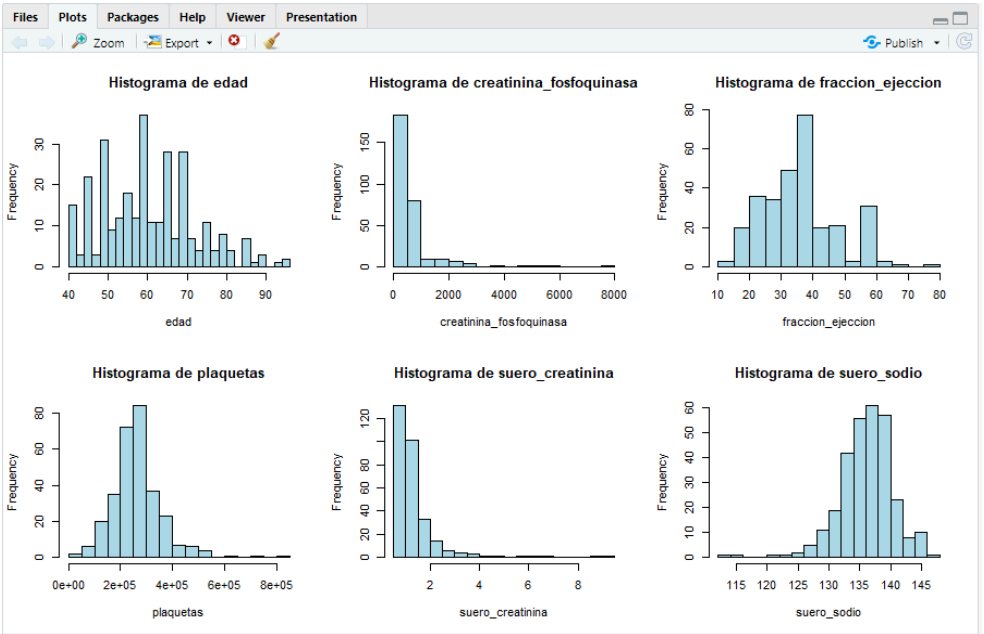
`sodio` y `plaquetas`. También es necesario que para cada grupo de la variable binaria, se disponga de dos o más observaciones lo cual se cumple en nuestro conjunto de datos. También a la hora de realizar el análisis el número de variables discriminantes debe ser menor que el número de observaciones menos dos. Es decir si tenemos p variables discriminantes y n individuos, entonces p < (n − 2).

Antes de realizar LDA se debe comprobar que ninguna variable discriminante es combinación lineal de las otras variables discriminantes. Para ello mostramos la matriz de covarianzas.

```
> print("Matriz de correlación entre variables continuas:")
[1] "Matriz de correlación entre variables continuas:"
> print(correlation_matrix)
                          edad creatinina_fosfoquinasa fraccion_ejeccion   plaquetas
edad                1.00000000             -0.08167188        0.06004941 -0.05229358
creatinina_fosfoquinasa -0.08167188              1.00000000       -0.04407955  0.02446339
fraccion_ejeccion        0.06004941             -0.04407955        1.00000000  0.07217747
plaquetas               -0.05229358              0.02446339        0.07217747  1.00000000
suero_creatinina         0.15916101             -0.01640848       -0.01130247 -0.04119808
suero_sodio             -0.04599247              0.05955016        0.17590228  0.06212462
                    suero_creatinina suero_sodio
edad                      0.15916101 -0.04599247
creatinina_fosfoquinasa  -0.01640848  0.05955016
fraccion_ejeccion        -0.01130247  0.17590228
plaquetas                -0.04119808  0.06212462
suero_creatinina          1.00000000 -0.18909521
suero_sodio              -0.18909521  1.00000000
>
```

En la matriz no se observa ninguna correlación alta (la más alta es cercana a 0.2 lo cual es adecuado) por lo que seguiremos comprobando el resto de suposiciones.

El LDA asume que las variables continuas deben seguir una distribución normal. Para ello lo visualizamos con histogramas y haremos una comprobación con el código mediante la prueba Kolmogorov-Smirnov estudiada en cursos anteriores.



En los histogramas parece que en general las variables no siguen una distribución normal y tras realizar la prueba observamos que la mayoría de p-valores son menores a 0.05 por lo que no se sigue una distribución normal. La única con un p-valor mayor a 0.05 es la edad que tiene un p-valor de 0.1. Seguiremos comprobando las suposiciones aunque realizar

Finalmente, la última suposición es que las matrices de covarianzas dentro de cada grupo deben ser aproximadamente iguales. Para ello emplearemos boxM() que medirá si son iguales o no.

```
> boxM_result <- boxM(data[, variables_continuas], data$muerte)
> print("Prueba de Box M para igualdad de matrices de covarianza:")
[1] "Prueba de Box M para igualdad de matrices de covarianza:"
> print(boxM_result)

        Box's M-test for Homogeneity of Covariance Matrices

data:  data[, variables_continuas]
Chi-Sq (approx.) = 167.16, df = 21, p-value < 2.2e-16
```

La hipótesis nula de esta prueba es que las matrices de covarianza son iguales entre los diferentes grupos. Al realizar la prueba vemos que el p-valor es mucho menor que 0.05 por lo que rechazamos la hipótesis nula de que las matrices de covarianza son iguales entre los grupos.

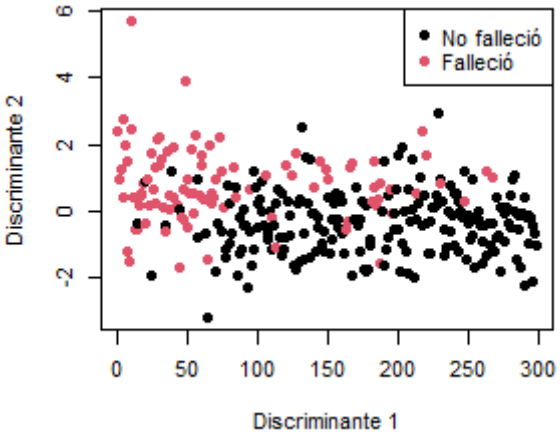
A pesar de que no se cumplen las suposiciones y que por ello el modelo pueda estar sesgado o sea ineficiente realizaremos el análisis para estudiar los resultados. Antes de realizar el modelo estandarizamos todas las variables.

```
[1] "Resumen del modelo LDA:"
> print(lda_model)
Call:
lda(muerte ~ edad + creatinina_fosfoquinasa + diabetes + fraccion_ejeccion +
    hipertension + plaquetas + suero_creatinina + suero_sodio,
    data = data)

Prior probabilities of groups:
        0         1
0.6789298 0.3210702

Group means:
        edad creatinina_fosfoquinasa  diabetes fraccion_ejeccion hipertension   plaquetas suero_creatinina suero_sodio
0 -0.1742419             -0.04306482 0.4187192         0.1844045    0.3251232  0.03373535       -0.2020307   0.1340133
1  0.3684489              0.09106416 0.4166667        -0.3899387    0.4062500 -0.07133622        0.4272107  -0.2833823

Coefficients of linear discriminants:
                                LD1
edad                     0.54316728
creatinina_fosfoquinasa  0.20344168
diabetes                 0.12010440
fraccion_ejeccion       -0.61620006
hipertension             0.37800443
plaquetas               -0.02535247
suero_creatinina         0.56709951
suero_sodio             -0.23497701
```

Observando los discriminantes vemos que la variable fracción de eyección tiene un coeficiente de -0.62 lo que indica que los valores más bajos en esta variable tienen más probabilidades de pertenecer al grupo que falleció. Por otro lado, la

variable suero creatinina con un coeficiente de 0.57 indica que cuanto mayor sea el valor de esta variable más probabilidad tiene el paciente de fallecer. Una mayor edad también contribuye significativamente en la probabilidad de muerte. Estas tres variables son las más importantes para diferenciar el grupo al que pertenece cada paciente. Sin embargo, variables como plaquetas (0.025) y diabetes (0.12) parece que tienen menos influencia en el modelo.

Tras realizar el modelo hacemos las predicciones con las que obtenemos los siguientes resultados:



	Actual	
Predicted	0	1
0	184	54
1	19	42

Visualmente parece que el modelo ha sido capaz de diferenciar con cierta precisión ambas clases, pero para asegurarnos, evaluaremos el accuracy del modelo que nos da: 0.7558528

Aunque el modelo logra un accuracy del 75%, el hecho de que no se cumplan los supuestos hace que los resultados no sean completamente confiables. Es posible que el modelo esté sesgado o que la capacidad de generalización con otros datos no sea buena.

Conclusión

En este estudio, se han combinado técnicas como PCA, PLS, clustering y LDA para analizar datos de fallos cardíacos. PCA ayudó a reducir la dimensionalidad, PLS destacó variables clave para predicciones, el clustering identificó perfiles de pacientes, y LDA evaluó su clasificación. Juntas, estas técnicas ofrecen una visión más completa del dataset, orientándose a la mejora del análisis clínico y predicción y prevención de posibles muertes.

Bibliografía

1. Herramientas y software

[1] R Core Team. "The R Project for Statistical Computing." https://www.r-project.org/.

[2] RStudio. "RStudio IDE Documentation." https://posit.co/products/open-source/rstudio/.

2. Información sobre las técnicas utilizadas

[3] Brownlee, J. "A Gentle Introduction to Principal Component Analysis (PCA)." Machine Learning Mastery. https://machinelearningmastery.com/introduction-to-principal-component-analysis/.

[4] Statology. "Partial Least Squares (PLS) Regression in R." https://www.statology.org/pls-regression-in-r/.

[5] Wikipedia. "Cluster Analysis." https://en.wikipedia.org/wiki/Cluster_analysis.

[6] Tutorials Point. "Linear Discriminant Analysis." https://www.tutorialspoint.com/linear_discriminant_analysis.htm.

3. Recursos de aprendizaje y tutoriales

[7] Towards Data Science. "A Guide to PCA in R with Examples." https://towardsdatascience.com/a-guide-to-principal-component-analysis-in-r-with-examples-654d856633be.

[8] GeeksforGeeks. "Clustering in R." https://www.geeksforgeeks.org/clustering-in-r/.