Métodos de agrupamiento usando Machine Learning grupo 7

Integrantes:

-GALVAN CAMARENA, GABRIEL ALEXANDER
-GABRIEL ENRIQUE JIMENEZ CARHUAS
- PATILLA FIERRO, MIGUEL ANGEL
- ERAZO MEJIA, HAROLD JEAN PIERRE
- GUTIERREZ MENDIZABAL, JOSEPH ALDAIR

Nuestro grupo entendió que estandarizar las variables antes de aplicar el análisis de agrupamiento jerárquico es un paso fundamental, especialmente cuando trabajamos con datos clínicos. En medicina, cada variable se mide en unidades diferentes y tiene rangos muy distintos. Por ejemplo, el IMC se mide en kg/m² y suele ir de 18 a 35, mientras que la creatinina sérica se mide en mg/dL y puede estar entre 0.5 y 15. Si no estandarizamos estas variables, las que tienen números más grandes terminan influyendo más en el análisis, aunque clínicamente no sean las más importantes.

#2.3.2 La importancia de estandarizar

```{r}
head(hemo_data_1)
```

Description: df [6 × 52]

	Edad <dbl>	Peso <dbl>	Talla <dbl>	IMC <dbl>	Presion_Sistolica <dbl>	Presion_Diastolica <dbl>	Volumen_Urina <dbl>	Hemoglobina <dbl>	Leucoci
1	52.48	75.66894	176.1509	24.38642	133.9204	73.84135	1387.602	14.68396	5.627
2	49.31	68.35500	161.9624	26.05804	128.7547	80.94792	1254.397	13.85548	7.369
3	53.24	68.57009	159.8712	26.82838	127.9980	86.99798	1312.429	14.66810	8.577
4	57.62	50.65430	160.9521	19.55344	150.2472	77.76286	1417.917	13.81079	6.628
5	48.83	71.50659	164.0500	26.57013	121.4674	81.70308	2089.276	14.57930	6.901
6	48.83	69.84285	159.4455	27.47246	129.8572	79.93111	1440.833	14.49729	8.211

6 rows | 1-10 of 52 columns

un vistazo después del escalamiento:

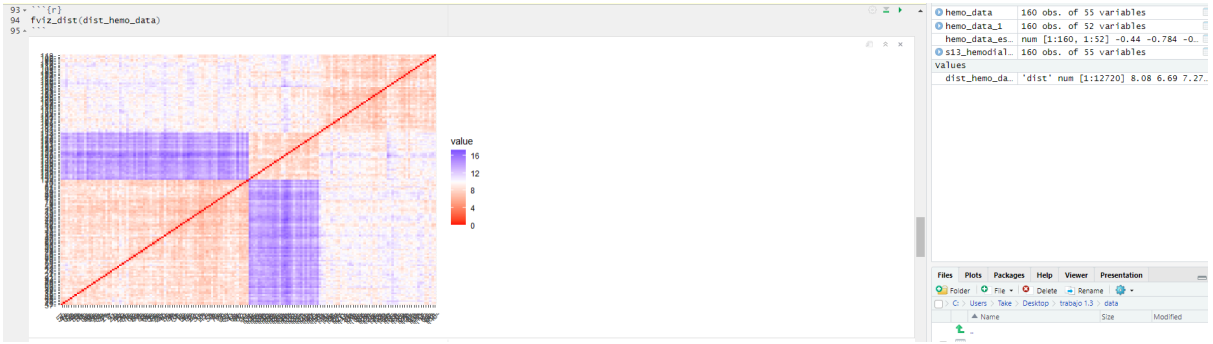
```{r}
head(hemo_data_escalado)
```

```
       Edad       Peso      Talla         IMC Presion_Sistolica Presion_Diastolica Volumen_Urina Hemoglobina Leucocitos    Plaquetas Proteinas_Totales
1 -0.4395599  0.8608999  1.4917950 -0.007099125        0.1625814         -1.6853653    0.08327311   1.0179089 -1.2562843 -0.246522554         0.8315151
2 -0.7843341  0.2006188 -0.4324701  0.340138307       -0.3035065         -0.3626873   -0.27530736   0.5217133 -0.2026073  1.576983448         0.5100876
3 -0.3569011  0.2200369 -0.7160898  0.500156887       -0.3717907          0.7633511   -0.11908905   1.0084096  0.5283781  0.006131148         0.8062482
4  0.1194746 -1.3973473 -0.5694952 -1.011028210        1.6357143         -0.9554923    0.16488087   0.4949514 -0.6503291  0.820727681         0.4175002
5 -0.8365396  0.4851354 -0.1493502  0.446511783       -0.9610289         -0.2221383    1.97214565   0.9552271 -0.4853017  0.523776978        -0.4467250
6 -0.8365396  0.3349379 -0.7738226  0.633948406       -0.2040350         -0.5519370    0.22657075   0.9061088  0.3066671 -0.584901135         0.5728494
    Albunina        AST        ALT        GGT Fosfatasa_Alcalina Acido_Urico        BUN Creatinina      Sodio     Potasio     Cloruro      Calcio     Fosforo
1  0.5603534 -0.3116269 -0.8626442 -2.2563842         -1.0103592 -0.95631332  0.4390603 -1.2865625 -0.5395536 -1.1742213 -0.6679867  0.91523260 -0.2714892
```

Para poder agrupar pacientes que se parezcan, necesitamos una forma de medir qué tan similares son. En este análisis usamos la distancia euclidiana, que calcula qué tan distintos son dos pacientes según sus valores clínicos. Mientras más parecidos sean, menor será la distancia entre ellos.

2.4.1 (opcional) Visualizando las distancias euclidianas con un mapa de calor

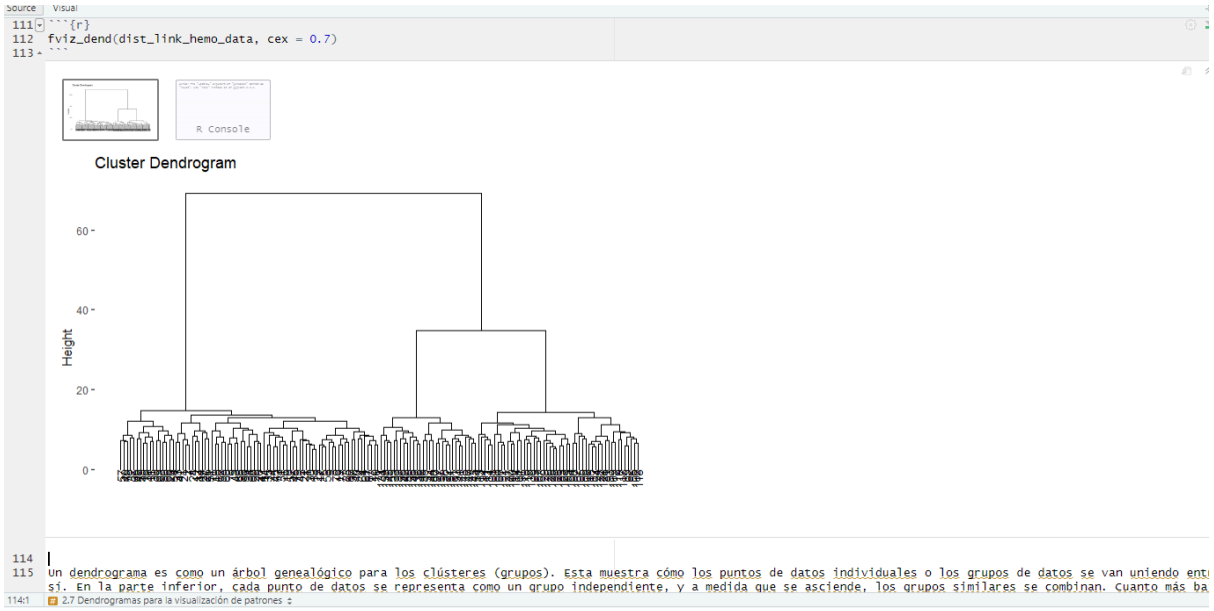
En R usamos la función dist() para calcular estas distancias entre todos los pacientes del dataset. El resultado es una matriz de distancias, donde cada valor representa qué tan distintos son dos pacientes.



El agrupamiento jerárquico comienza uniendo las observaciones más parecidas, pero una vez que se forman grupos, hay que decidir cómo calcular la distancia entre esos grupos y los demás. A esto se le llama función de enlace o linkage. Existen distintos métodos, como el enlace completo, el mínimo, el promedio o el método de Ward.

En este análisis usamos el método de Ward, que busca formar grupos que tengan la menor variación interna posible. Es decir, intenta que los pacientes dentro de cada grupo sean lo más parecidos posible.

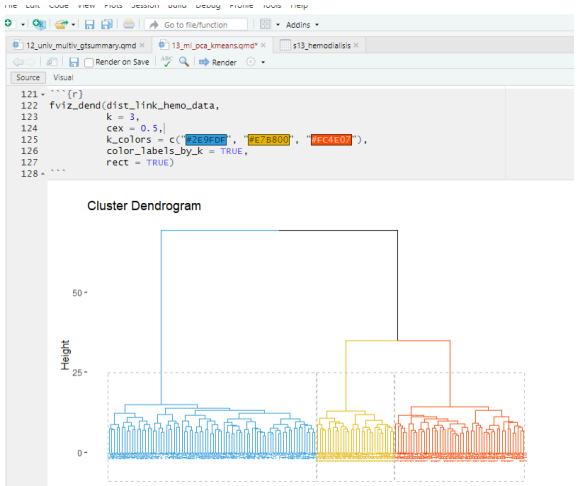
2.7 Dendrogramas para la visualización de patrones



El dendrograma es un gráfico que muestra cómo se van formando los grupos paso a paso. Cada paciente empieza como su propio grupo y, a medida que subimos en el gráfico, los que se parecen más se van uniendo. Si dos pacientes o grupos se unen en la parte baja del dendrograma, eso significa que son muy parecidos. Es como ver un árbol de cómo se organizan los pacientes según sus características clínicas.

Una limitación del agrupamiento jerárquico es que no indica directamente cuántos grupos hay, así que el corte del dendrograma depende del criterio del investigador. En nuestro caso, al observar el dendrograma, se ve claramente que se forman tres grupos bien definidos.

2.8 ¿Cúantos grupos se formaron en el dendrograma?

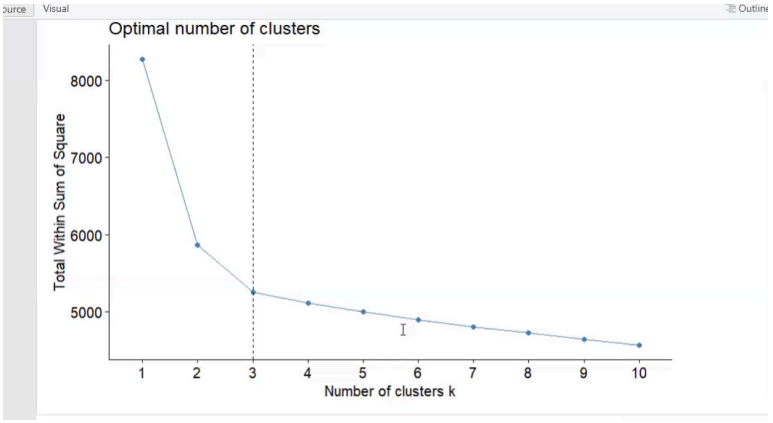


Este gráfico muestra cada grupo con un color distinto y los separa con rectángulos. Así podemos visualizar mejor qué pacientes están en cada clúster.

3 Agrupamiento con el algoritmo K-Means

En este caso usamos el algoritmo de K-Means, que es otra forma de agrupar pacientes. A diferencia del método jerárquico, en K-Means sí tenemos que decidir cuántos grupos queremos formar desde el comienzo. Por eso, uno de los pasos más importantes es elegir bien ese número.

La idea del algoritmo es formar grupos donde los pacientes se parezcan entre sí lo más posible, y que sean distintos a los pacientes de los otros grupos. Cada grupo tiene un "centroide", que es como el paciente promedio del grupo. El algoritmo empieza asignando centros al azar, luego va agrupando pacientes al centro más cercano, y recalcula esos centros hasta que los grupos se estabilizan.

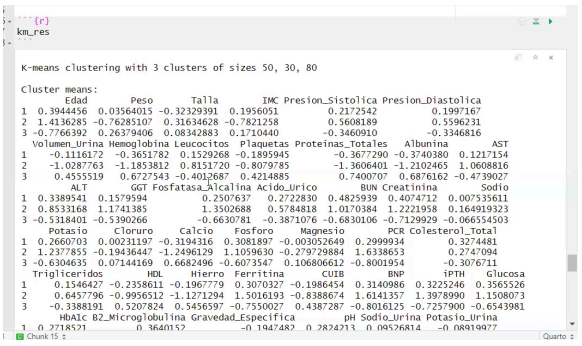


Como dijimos, K-Means necesita que uno diga cuántos grupos quiere K, pero eso no siempre es obvio. Para decidirlo, usamos el método del (codo) o elbow method, que consiste en probar varios valores de K.

3.3 Cálculo del agrupamiento k-means

Después de ver que el mejor número de grupos es 3, aplicamos el algoritmo de K-Means con ese valor.

Algo importante que aprendimos es que los resultados de K-Means pueden cambiar porque el algoritmo empieza con puntos aleatorios. Para que el resultado sea más confiable, usamos el argumento nstart = 25, lo que significa que R va a probar 25 inicios diferentes y se quedará con el que tenga la menor variación interna entre los grupos.



```{r}
km_res
```

```
K-means clustering with 3 clusters of sizes 50, 30, 80

Cluster means:
       Edad       Peso       Talla        IMC Presion_Sistolica Presion_Diastolica
1  0.3944456  0.03564015 -0.32329391  0.1956051         0.2172542          0.1997167
2  1.4136285 -0.76285107  0.31634628 -0.7821258         0.5608189          0.5596231
3 -0.7766392  0.26379406  0.08342883  0.1710440        -0.3460910         -0.3346816
  Volumen_Urina Hemoglobina Leucocitos  Plaquetas Proteinas_Totales   Albunina        AST
1    -0.1116172  -0.3651782  0.1529268 -0.1895945        -0.3677290 -0.3740380  0.1217154
2    -1.0287763  -1.1853812  0.8151720 -0.8079785        -1.3606401 -1.2102465  1.0608816
3     0.4555519   0.6727543 -0.4012687  0.4214885         0.7400707  0.6876162 -0.4739027
         ALT        GGT Fosfatasa_Alcalina Acido_Urico        BUN Creatinina       Sodio
1  0.3389541  0.1579594          0.2507637   0.2722830  0.4825939   0.4074712  0.007535611
2  0.8533168  1.1741385          1.3502688   0.5784818  1.0170384   1.2221958  0.164919323
3 -0.5318401 -0.5390266         -0.6630781  -0.3871076 -0.6830106  -0.7129929 -0.066554503
     Potasio     Cloruro     Calcio     Fosforo     Magnesio         PCR Colesterol_Total
1  0.2660703  0.00231197 -0.3194316  0.3081897 -0.003052649  0.2999934         0.3274481
2  1.2377855 -0.19436447 -1.2496129  1.1059630 -0.279729884  1.6338653         0.2747094
3 -0.6304635  0.07144169  0.6682496 -0.6073547  0.106806612 -0.8001954        -0.3076711
  Triglyceridos        HDL      Hierro  Ferritina        CUIB         BNP        iPTH    Glucosa
1     0.1546427 -0.2358611 -0.1967779  0.3070327 -0.1986454  0.3140986  0.3225246  0.3565526
2     0.6457796 -0.9956512 -1.1271294  1.5016193 -0.8388674  1.6141357  1.3978990  1.1508073
3    -0.3388191  0.5207824  0.5456597 -0.7550027  0.4387287 -0.8016125 -0.7257900 -0.6543981
     HbAlc B2_Microglobulina Gravedad_Especifica         pH Sodio_Urina Potasio_Urina
1  0.2718521        0.3640152          -0.1947482  0.2824213  0.09526814   -0.08919977
```

-Los centros (o medias) de los clústeres:
Esto es una tabla donde cada fila representa uno de los tres grupos, y cada columna es una variable del dataset. Es como el "paciente promedio" de cada grupo, y nos ayuda a entender las características clínicas generales de cada clúster.

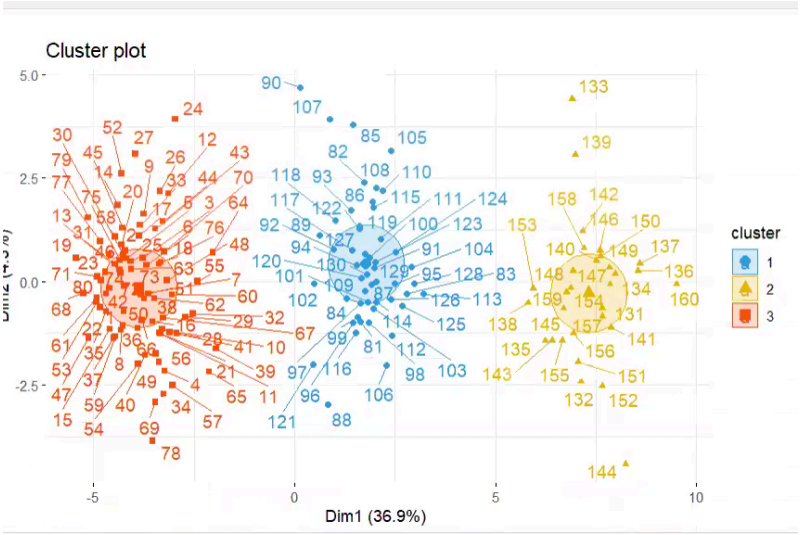
-El vector de asignación de clúster:
Es una lista que indica a qué grupo fue asignado cada paciente. Por ejemplo, si el paciente número 10 tiene un "2", significa que pertenece al clúster 2. Así podemos saber cómo se distribuyeron los pacientes entre los tres grupos.

3.4 Visualización de los clústeres k-means

Después de aplicar el algoritmo K-Means, quisimos visualizar los grupos en un gráfico. Pero como tenemos más de 50 variables clínicas, no es posible graficarlas todas juntas. Entonces surge la pregunta: ¿qué variables usamos en el eje X y en el eje Y?

La solución que usamos fue aplicar PCA, que es una técnica para reducir la cantidad de variables. El PCA convierte las 52 variables originales en dos nuevas variables que resumen la mayor parte de la información. Estas dos se pueden usar como ejes en el gráfico.

Para graficar usamos la función fviz_cluster() del paquete factoextra. Esta función permite mostrar los grupos formados por el K-Means, usando colores diferentes para cada clúster. El resultado es un gráfico donde cada punto es un paciente, y los grupos aparecen separados visualmente.



3.4.1 ¿Cómo interpretar?

En el gráfico de clústeres, cada punto representa a un paciente. Como teníamos muchas variables clínicas, usamos PCA para reducirlas a dos dimensiones nuevas que resumen la mayor parte de la información. Así, pudimos ver cómo se agrupan los pacientes en 3 grupos, según sus características clínicas.

Es importante aclarar que el gráfico nos muestra la forma de los grupos, pero no nos dice qué tan útiles son esos grupos para la práctica médica. Para eso, tendríamos que hacer más análisis, como ver si los grupos tienen diferencias en la supervivencia, o si la función renal cambia entre grupos.

Conclusión grupal

Conclusión 1

Aplicar métodos de agrupamiento como K-Means o clustering jerárquico nos permitió identificar grupos de pacientes con perfiles clínicos similares, lo cual podría ser muy útil para personalizar tratamientos o anticipar riesgos en la práctica médica. Esto es especialmente importante en pacientes con enfermedades crónicas como los que inician hemodiálisis.

Conclusión 2

Como grupo, creemos que aprender técnicas de machine learning aplicadas a salud nos abre una nueva forma de entender los datos. Aunque al principio parezcan complejas, estas herramientas pueden complementar nuestra formación médica y ayudarnos a tomar decisiones más informadas y basadas en evidencia.

Conclusión 3

El uso de técnicas de agrupamiento nos ayudó a encontrar patrones dentro de un grupo de pacientes sin necesidad de tener una variable de referencia. Creemos que este tipo de análisis puede ser una herramienta útil para estratificar riesgos, entender mejor a nuestros pacientes y apoyar decisiones clínicas de forma más objetiva.