Analisis de Cluster para datos de Cancer Mamario

Alfredo Barra, Felipe Altamirano

2023-08-02

Abstract Según la OMS, el cáncer de mama es una enfermedad que afectó a aproximadamente 2.3 millones de mujeres en 2020, siendo una de las principales causas de mortalidad en mujeres en todo el mundo ("Cáncer de Mama" n.d.). Sin embargo, se indica que una de las principales formas de combate a esta enfermedad está en el diagnóstico temprano. Teniendo esto en cuenta, ¿cuales son los parámetros de los signos vitales de una persona más relevantes para poder determinar si está afectada por el cáncer de mama?. A continuación se analizarán los datos de 116 pacientes, de las cuales 64 poseen cáncer de mama.

1 Descripción de los datos

Los datos que usaremos a continuación corresponden a 116 pacientes mujeres, de los cuales 64 han sido diagnosticados con cáncer de mama. Las variables de estos datos corresponden a:

- **Age** (years): *Edad en años*
- **BMI** (kg/m2): *Índice de Masa Corporal* El BMI (IMC) es una medida médica ampliamente utilizada, evalúa la relación peso-altura para determinar si una persona tiene un peso saludable. Es un método conveniente para detectar problemas de peso y clasificarlos como bajo, normal, sobrepeso u obesidad (Gonzalez, Correia, and Heymsfield 2017).
- **Glucose** (mg/dL): *Glucosa en sangre* La concentración de glucosa, un tipo de azúcar, presente en el torrente sanguíneo de una persona se denomina "glucosa en sangre". La glucosa es una fuente de energía esencial para el cuerpo y se deriva principalmente de los alimentos que consumimos, en particular de aquellos con alto contenido de carbohidratos (Galant, Kaufman, and Wilson 2015).
  La falta de insulina o la resistencia a su acción es lo que hace que los niveles de glucosa en sangre sean constantemente altos en la diabetes, lo que puede ser perjudicial para el bienestar del organismo.
- **Insulin** (µU/mL): *Insulina* La insulina es una hormona producida por el páncreas en el cuerpo humano, actúa en la regulación del metabolismo de la glucosa y es esencial para mantener niveles adecuados de glucosa en sangre (Quianzon and Cheikh 2012). Cuando se consumen alimentos el nivel de glucosa (azúcar) en la sangre aumenta. En respuesta, las células beta del páncreas liberan insulina al torrente sanguíneo. La insulina permite que la glucosa entre en las células del cuerpo, donde se utiliza como fuente de energía. Ésta promueve la absorción de glucosa después de la alimentación, inhibe la producción de glucosa ayudando a evitar que los niveles sean demasiado altos y almacena el exceso en forma de glucógeno en hígado y músculos y en forma de grasa en las células adiposas.
- **HOMA**: *Índice de resistencia a la insulina Homeostasis Model Assessment* es un método utilizado para estimar la resistencia a la insulina y la función de las células beta del páncreas. Se basa en la medición de glucosa e insulina en sangre durante el ayuno y sigue la siguiente fórmula descrita por Mattheus (Matthews et al. 1985):

$$HOMA_{IR} = \frac{(Glucosa * Insulina)}{22.5}$$

- **Leptin** (ng/mL): *Leptina* Hormona proteica producida principalmente por el tejido adiposo en el cuerpo humano. Esta hormona juega un papel fundamental en la regulación del peso corporal y el apetito al actuar como una señal de saciedad; cuando los niveles de leptina son altos, se siente menos hambre y el cuerpo tiende a quemar más calorías para mantener el equilibrio energético, por el contrario, cuando los

niveles son bajos el cerebro interpreta esto como una situación de "escasez" de energía y se activan mecanismos para aumentar el apetito y reducir el gasto energético.

- **Adiponectin** (µg/mL): *Adiponectina* Según (Palomer, Pérez, and Blanco-Vaca 2005) la adiponectina es una citoquina secretada por el tejido adiposo, que regula el metabolismo energético, estimula la oxidación de ácidos grasos, reduce los triglicéridos plasmáticos y mejora el metabolismo de la glucosa mediante aumento de la sensibilidad a la insulina.
- **Resistin** (ng/mL): *Resistina* Acorde a lo descrito por (Wellen and Hotamisligil 2005) la resistina es una hormona polipeptídica responsable de ser la conexión en la bien conocida asociación entre la inflamación y la resistencia a la insulina
- **MCP-1**(pg/dL): *Proteína quimiotáctica de monocitos 1* La proteína quimiotáctica de monocitos 1 pertenece a la familia de quimioquinas C-C, caracterizadas por tener dos residuos de cisteína adyacentes. Las quimioquinas son citoquinas con actividad quimioatrayente cuya función, ejercida mediante la unión a receptores con 7 dominios transmembrana acoplados a proteínas G (GPCRs), está relacionada fundamentalmente con el tránsito de células del sistema inmune. Las alteraciones en MCP-1 y su receptor se asocian con distintas enfermedades inflamatorias como Artritis reumatoide, según los estudios de (Ogata et al. 1997).
- **Classification**: como etiquetado si el paciente está enfermo (2) o está sano (1)

2 Preparación de los datos

Dado que los datos presentan distintas escalas, será necesario normalizar los datos. Para ello se utilizará la siguiente función.

```
normalize_min_max <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
```

Y posteriormente crearemos un nuevo *DataFrame* con los datos ya normalizados en una escala de 0 a 1.

```
columnas_a_normalizar = c(
  "Age", "BMI", "Glucose", "Insulin", "HOMA", "Leptin", "Adiponectin", "Resistin", "MCP.1")

data_normalizada = data

data_normalizada[columnas_a_normalizar] =
  lapply(data_normalizada[columnas_a_normalizar], normalize_min_max)
```

Ahora tenemos una data normalizada como la siguiente.

```
head(data_normalizada, n = 3)
```

```
##         Age       BMI    Glucose    Insulin       HOMA     Leptin Adiponectin
## 1 0.3692308 0.2538503 0.07092199 0.00490826 0.00000000 0.05229908   0.2211517
## 2 0.9076923 0.1148262 0.22695035 0.01219033 0.00974207 0.05272598   0.1037068
## 3 0.8923077 0.2352777 0.21985816 0.03687442 0.02205768 0.15852575   0.5710211
##     Resistin     MCP.1 Classification
## 1 0.06066485 0.2246591              1
## 2 0.01082583 0.2559263              1
## 3 0.07690645 0.3079117              1
```

3 Análisis de agrupamiento

3.1 Cantidad de grupos a buscar

Dada la descripción de los datos y la problemática presentada, es que sabemos que los pacientes se dividen entre sanos y enfermos de Cáncer de Mama, por lo tanto, buscaremos dos grupos de datos que describan la

misma clasificación de pacientes.

3.2 Procesamiento de datos

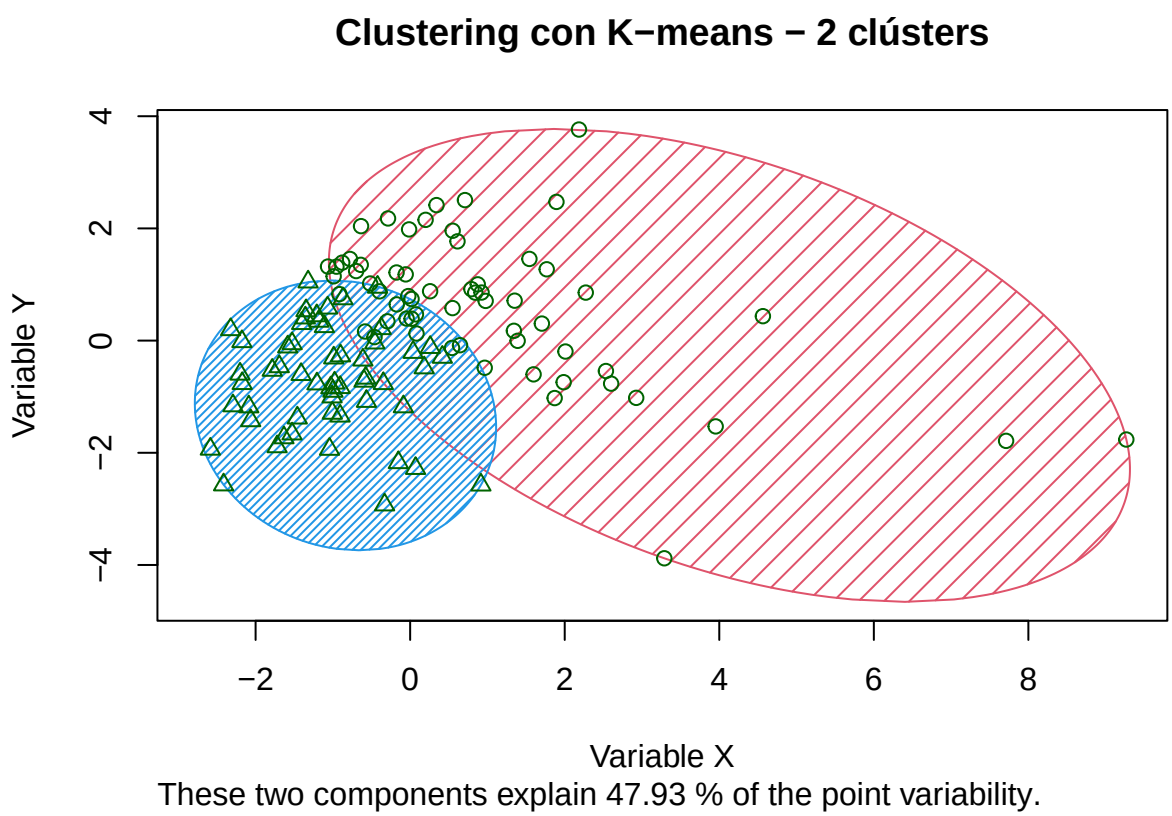
Con la data ya normalizada previamente realizaremos el siguiente análisis.

```
#Primero, crearemos un cluster sólo usando los datos y sin clasificación, y generando sólo dos grupos

clusterPacientes_2 = kmeans(data_normalizada[,1:9], center= 2)
table(clusterPacientes_2$cluster, data_normalizada$Classification)
```

```
##
##      1  2
##   1 27 34
##   2 25 30
```

Y luego graficamos este cluster

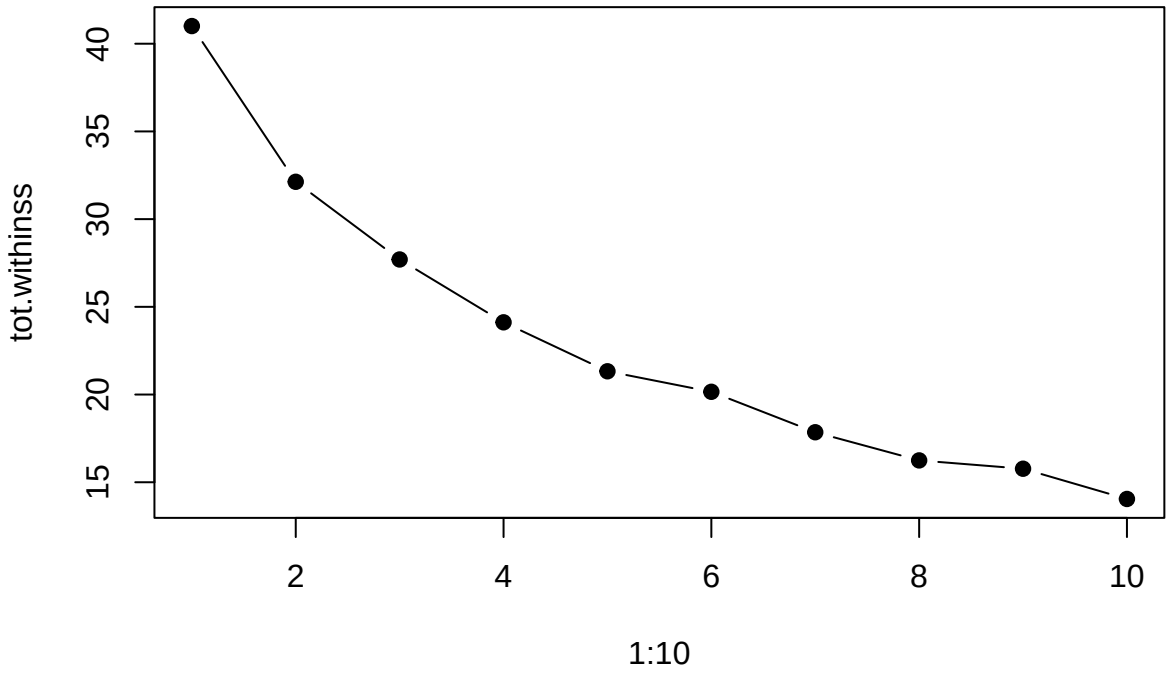
```
clusplot(data_normalizada, clusterPacientes_2$cluster, color=T, shade=T,
         main = "Clustering con K-means - 2 clústers",
         xlab = "Variable X",
         ylab = "Variable Y")
```



De este gráfico, ya se puede desprender que existen dos grupos y que en general coinciden en número respecto a la clasificación incial de pacientes, pero ¿es esta la mejor forma de agrupar los datos?

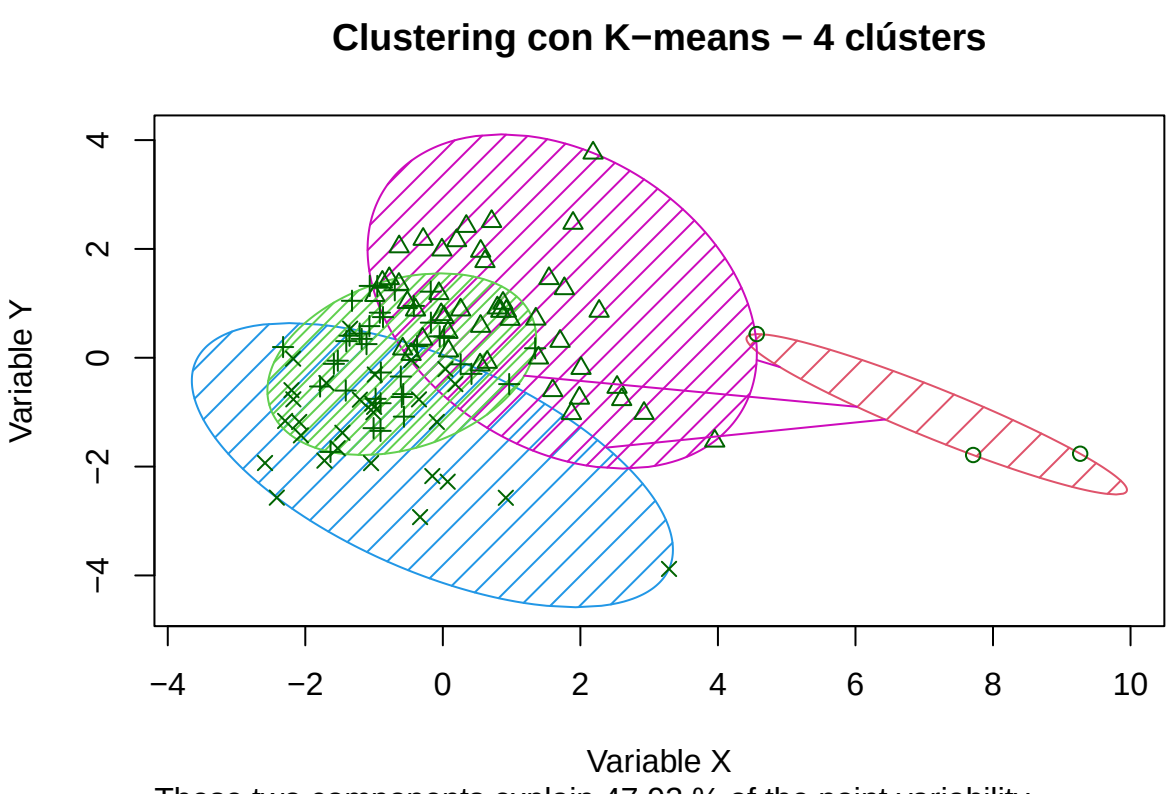
3.3 Verificación de cantidad de grupos

```
tot.withinss = vector(mode="character", length = 10)

for(i in 1:10){
  clusterPacientes = kmeans(data_normalizada[,1:9], center=i)
  tot.withinss[i] = clusterPacientes$tot.withinss
}

plot(1:10, tot.withinss, type="b", pch=19)
```



Como se puede notar en el gráfico anterior, existe un "codo" en el punto dos 2, por lo que ésta podría ser la forma más optima de agrupar. Sin embargo, existe incluso otro codo en el punto 4 y 8; probemos uno de estos grupos para ver si existe un mejor agrupamiento.

```
clusterPacientes = kmeans(data_normalizada[,1:9], center= 4)

clusplot(data_normalizada, clusterPacientes$cluster, color=T, shade=T, main = "Clustering con K-means -
         xlab = "Variable X",
         ylab = "Variable Y")
```



Como se puede observar, usar 4 grupos no se gana mucho, además de no coincidir con la problemática planteada en los datos, por lo que no sería conveniente para este trabajo el usar más de 2 grupos.

Adicionalmente, complementario al análisis del codo, al calcular el índice Davies-Bouldin y el de silueta, para ambos casos (2 y 4 clústers) siempre el mejor escenario resultó ser el de 2 agrupaciones.

Bibliografía

"Cáncer de Mama." n.d. Accessed July 24, 2023. https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer.

Galant, A. L., R. C. Kaufman, and J. D. Wilson. 2015. "Glucose: Detection and Analysis." *Food Chemistry* 188 (December): 149–60. https://doi.org/10.1016/j.foodchem.2015.04.071.

Gonzalez, Maria Cristina, Maria Isabel T. D. Correia, and Steven B. Heymsfield. 2017. "A Requiem for BMI in the Clinical Setting." *Current Opinion in Clinical Nutrition and Metabolic Care* 20 (5): 314–21. https://doi.org/10.1097/MCO.0000000000000395.

Matthews, D. R., J. P. Hosker, A. S. Rudenski, B. A. Naylor, D. F. Treacher, and R. C. Turner. 1985. "Homeostasis Model Assessment: Insulin Resistance and Beta-Cell Function from Fasting Plasma Glucose and Insulin Concentrations in Man." *Diabetologia* 28 (7): 412–19. https://doi.org/10.1007/BF00280883.

Ogata, Hiroomi, Motohiro Takeya, Teizo Yoshimura, Katsumasa Takagi, and Kiyoshi Takahashi. 1997. "The Role of Monocyte Chemoattractant Protein-1 (Mcp-1) in the Pathogenesis of Collagen-Induced Arthritis in Rats." *The Journal of Pathology* 182 (1): 106–14. https://doi.org/10.1002/(SICI)1096-9896(199705)182:1%3C106::AID-PATH816%3E3.0.CO;2-A.

Palomer, Xavier, Antonio Pérez, and Francisco Blanco-Vaca. 2005. "Adiponectina: Un Nuevo Nexo Entre Obesidad, Resistencia a La Insulina y Enfermedad Cardiovascular." *Medicina Clínica* 124 (10): 388–95. https://doi.org/10.1157/13072576.

Quianzon, Celeste C., and Issam Cheikh. 2012. "History of Insulin." *Journal of Community Hospital Internal Medicine Perspectives* 2 (2): 18701. https://doi.org/10.3402/jchimp.v2i2.18701.

Wellen, Kathryn E., and Gökhan S. Hotamisligil. 2005. "Inflammation, Stress, and Diabetes." *Journal of*

Clinical Investigation 115 (5): 1111–19. https://doi.org/10.1172/JCI200525102.