**UNIVERSIDAD DE SANTIAGO DE CHILE**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA INFORMÁTICA**


# Laboratorio 2 - Agrupamiento K-medias

| | |
|---|---|
| Integrantes: | Chun-zen Yu |
| | Matias Pizarro |
| Curso: | Análisis de Datos |
| Sección | A-1 |
| Profesor: | Max Chacón Pacheco |

12 de Diciembre de 2020

# Tabla de contenidos

# 1. Introducción

Las enfermedades al corazón siempre han sido una gran preocupación para muchas personas, hace años que este tipo de enfermedad ha estado aumentando significativamente, de acuerdo a la World Health Organization (WHO), el numero total de personas que han muerto por enfermedades cardiovasculares llega a los 17,3 millones al año. Sin embargo, un diagnostico durante las primeras etapas de esta enfermedad seguido con el tratamiento adecuado puede salvar una gran cantidad de vidas. Desafortunadamente, el diagnostico correcto de una enfermedad cardiovascular en sus primeras etapas es algo complejo que requiere de un diagnostico médico realizado por expertos en el área, esto es un problema debido a que no todos los centros médicos cuentan con la cantidad adecuada de expertos para todos los pacientes. Es por eso que existe la necesidad de desarrollar un sistema de diagnostico médico de tal forma de poder asistir con este proceso.

Este proyecto se concentrara principalmente en seguir conociendo los datos (como en el primer laboratorio ), ocupando esta vez clustering el cual ayuda a conocer los datos, a través de un proceso de agrupación, que permite formar grupos con datos de similares características.

## 1.1. Objetivos

- Extraer el conocimiento del problema asignado, mediante el uso del software R, utilizando el algoritmo de clustering K-means y realizar el análisis respectivo.
- Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- Analizar por grupo e identificar aquellas caracterısticas mas relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

# 2. Marco Teórico

## 2.1. Clustering

El agrupamiento o también conocido como *clustering* es una técnica utilizada como primer paso para conocer un conjunto de datos, esta técnica consiste en agrupar casos que tengan características similares entre si para así formar grupos dentro del conjunto de datos. Luego estos grupos son utilizaos para simplificar los datos en estructuras mas fáciles de interpretar y manipular.

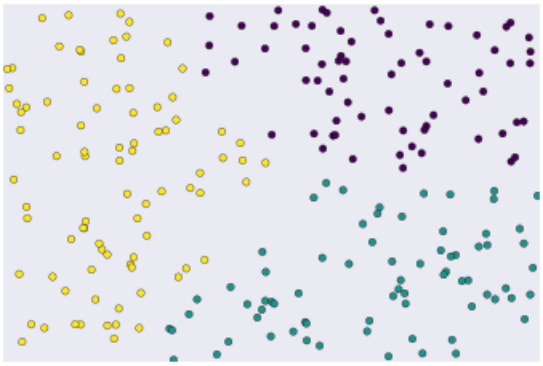

Figura 1: Ejemplo agrupamiento dentro de un conjunto de datos.

## 2.2. Distancia

Antes de realizar un cluster primero se necesita encontrar una forma de medir la similitud entre dos casos dentro del conjunto de datos. Es por eso que es necesario definir una medida de distancia entre dos datos.

Existen distintos tipos de distancias definidas para medir la similitud entre dos casos distintos, algunas de las mas utilizadas son la *Distancia Euclidiana* o también la *Distancia Manhattan* las cuales se basan en calcular la distancia entre dos vectores.

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Figura 2: Formula distancia euclidiana.

El problema de estas medidas que solo consideran datos que contengan variables numéricas. En el caso de que se tenga una mezcla de variables categóricas, numéricas o binarias es necesario definir realizar una matriz de disimilitud donde se comparen los datos entre si, y se tiene que definir una distancia que tome en cuenta todos los datos ya sean numéricos o categóricos.

Para esto existen otros tipos de distancia mas robustas como por ejemplo la distancia *Gower* la cual es utilizada cuando el conjunto de datos tiene una mezcla de datos numéricos, categóricos o lógicos. Esta distancia siempre mantiene el valor entre 0 a 1 para mantener ponderaciones justas entre las variables.

## 2.3. K-mean

K-medias es un algoritmo de agrupamiento cuyo objetivo es encontrar $K$ grupos dentro de un conjunto con $N$ datos. Este algoritmo funciona de forma iterativa, asignando un centroide a cada K grupo y luego asignando los casos a distintos grupos según la similitud que tenga con el centroide del grupo.

Este algoritmo de clustering requiere de tener definido previamente el numero de grupos que se quiere generar dentro del conjunto de datos. Para esto se necesita utilizar distintos algoritmos para encontrar el mejor numero de grupos para no sesgar los datos.

# 3. Pre-procesamiento

Antes de comenzar a trabajar con los datos es necesario adecuarlos para el estudio que se pretende realizar:

## 3.1. Datos faltantes

La base de datos consistía con 303 casos distintos, 6 de ellos contaban con algún tipo de atributo no asignado, para evitar cualquier tipo de alteración de datos se decidió eliminar esos casos, debido a que no conformaban una gran parte de los datos y la resultados no se verán afectados.

## 3.2. Binarizacion de los datos

En este conjunto de datos existen variables numéricas y categóricas, dado esto es necesario encontrar una forma para que las variables tengan una ponderación justa. Para esto es necesario binarizar los las variables categóricas y normalizar las variables numéricas entre 0 a 1.

Las variables sex, cp, fbs, restecg, exang, thal y num se separan en cada uno de sus factores de tal forma que solo tengan valores binarios. En la figura 4 se puede observar los datos después de este proceso.

| | age | sexfemale | sexmale | cpasymptomatic | cpatypical angina | cpnon-anginal pain | cptypical angina | trestps | chol | fbsfbs < 120 mg/dl |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 63 | 0 | 1 | 0 | 0 | 0 | 1 | 145 | 233 | 0 |
| 2 | 67 | 0 | 1 | 1 | 0 | 0 | 0 | 160 | 286 | 1 |
| 3 | 67 | 0 | 1 | 1 | 0 | 0 | 0 | 120 | 229 | 1 |
| 4 | 37 | 0 | 1 | 0 | 0 | 1 | 0 | 130 | 250 | 1 |
| 5 | 41 | 1 | 0 | 0 | 1 | 0 | 0 | 130 | 204 | 1 |
| 6 | 56 | 0 | 1 | 0 | 1 | 0 | 0 | 120 | 236 | 1 |
| 7 | 62 | 1 | 0 | 1 | 0 | 0 | 0 | 140 | 268 | 1 |
| 8 | 57 | 1 | 0 | 1 | 0 | 0 | 0 | 120 | 354 | 1 |
| 9 | 63 | 0 | 1 | 1 | 0 | 0 | 0 | 130 | 254 | 1 |
| 10 | 53 | 0 | 1 | 1 | 0 | 0 | 0 | 140 | 203 | 0 |

Figura 3: Primeros 10 datos binarizacion.

## 3.3. Normalización de los datos

Una vez binarizada las variables categóricas es posible utilizar algún algoritmo de clusterizacion para agrupar los casos, sin embargo existe un problema debido a que las variables numéricas tienen números de escalas mas grandes por ejemplo el colesterol va desde los 200-300 lo que ocasionara que el colesterol tenga una ponderación mas grande con respecto las variables de menor escalas, y en especial las variables binarias. Es por esto que es necesario normalizar las variables numéricas para que vayan por el rango de 0-1.

Para normalizar las variables numericas se utilizara una ecuacion basada en el maximo y el minimo de cada variables, la ecuacion utilizada sera la siguiente:

$$z = \frac{x \text{ - } \min(x)}{\max(x) \text{ - } \min(x)}$$

Figura 4: Ecuación normalizar.

# 4. Obtención del cluster

## 4.1. Matriz de disimilitud

Antes de realizar el algoritmo para obtener el cluster, es necesario definir una distancia para determinar que tan similares son las variables, una vez definido esto es posible construir una matriz de similitud o de disimilitud. Esta matriz relacionara cada uno de los individuos entre si.

Para esta ocasión la distancia que se utiliza corresponde a la *distancia Gower*, esta distancia se caracteriza por ser una métrica la cual mide la disimilitud de dos casos sin importar si las variables son binarias o numéricas. Esta distancia esta basada según la siguiente formula.

$$D_{Gower}(x_1, x_2) = 1 - \left( \frac{1}{p} \sum_{j=1}^{p} s_j(x_1, x_2) \right)$$

Figura 5: Distancia Gower.

Luego, se define una matriz de disimilitud utilizando la distancia *Gower* con la función *daisy* incluida en MATALB.

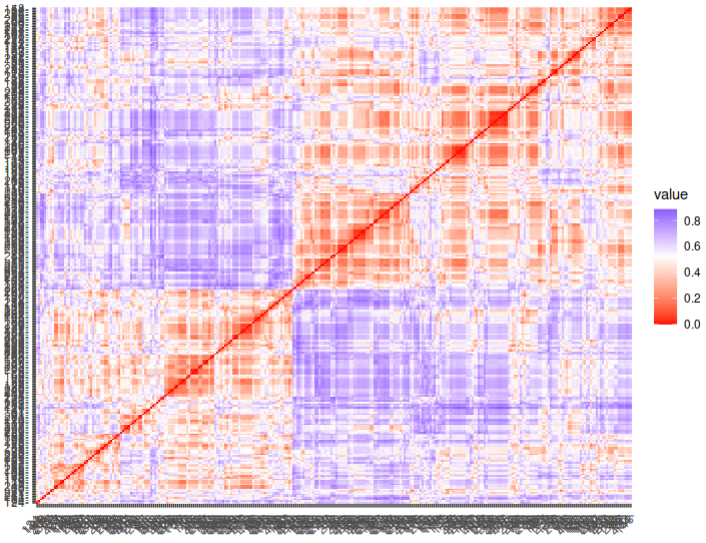


Figura 6: Matriz de disimilitud.

## 4.2. Obtención de K

Además de definir la distancia, también es necesario definir el numero de grupos el cual se desea, esto se debe a que el algoritmo de K-medias requiere saber de antemano cuantos grupos se van a construir. Desafortunadamente no existe una respuesta definitiva para el numero de grupos, este valor dependerá de método utilizado para obtener la distancia y los parámetros utilizados para la partición los grupos.

Para esta experimento se opto por el método Silhouette para encontrar el numero de grupos óptimo, este método consiste en calcular el coeficiente Silhouette para cada punto del plano, el cual indicara que tan similar es un punto con el cluster el cual esta ubicado. Una vez calculado el coeficiente para cada uno de los puntos se obtiene un promedio de todos los valores y se obtiene una puntuación general. Este proceso se realiza para varios K de forma incremental y luego se genera una gráfica donde se puede observar la puntuación Silhouette de cada uno de los K.

Para realizar este proceso en R se utilizo la siguiente función *fviz_nbclust*, donde se especifica el algoritmo de agrupamiento *pam* y el método Silhouette.

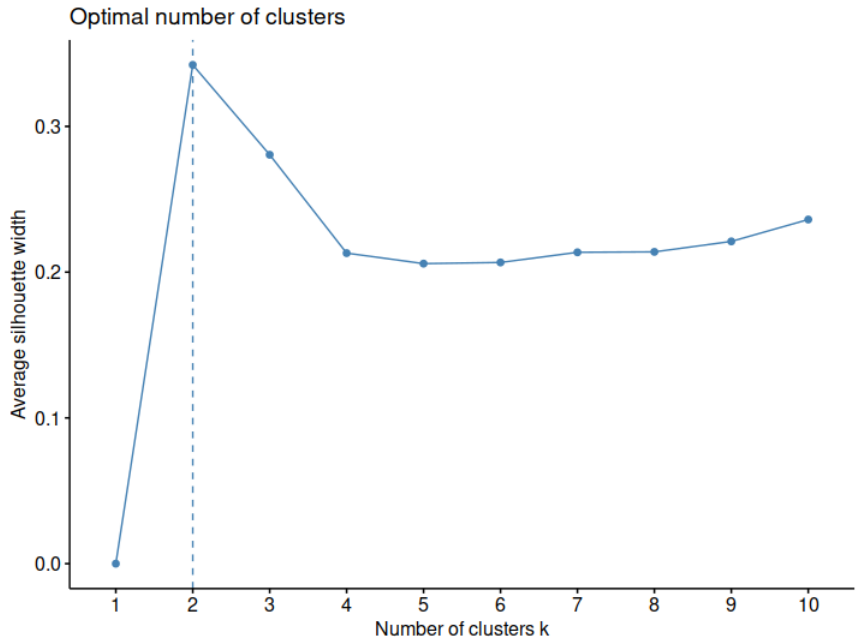


Figura 7: Numero de cluster determinado por Silhouette.

Como se puede observar en el gráfico el numero óptimo de grupos para el cluster es 2 con un coeficiente de Silhouette promedio de 0.33, luego al incrementar mas el numero de grupos se puede notar que el promedio del coeficiente Silhouette se reduce hasta un 0.2. A pesar de que el promedio de k=2 no es muy alto aun siendo el mejor K para este conjunto de datos.

## 4.3. Algoritmo de agrupamiento

Finalmente, una vez definido el K y la distancia utilizada para definir la matriz de disimilitud se procede a utilizar el algoritmo K-medias.

En R se utiliza la función *pam* (Partitioning Around Medoids), Esta función recibe una matriz de disimilitud, el numero de grupos y con esto utiliza el algoritmo de K-medias para generar un cluster. En la siguiente figura se puede observar el cluster obtenido en R.

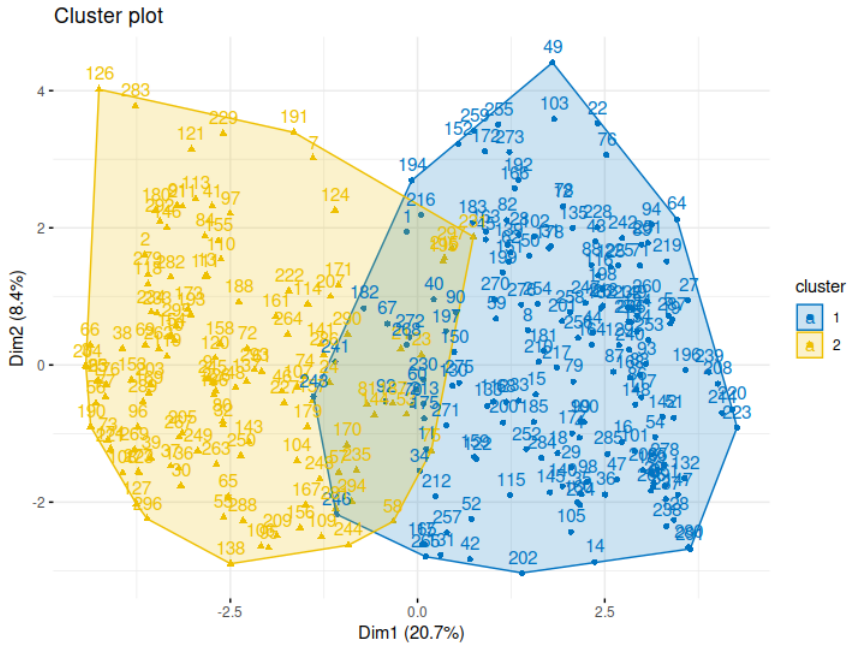


Figura 8: Cluster obtenido mediante K-mean.

# 5. Análisis de resultado

Al observar el cluster se puede notar los dos grupos mediante el algoritmo de k-medias, los grupos formados comparten una forma similar, ambos corresponden a círculos del mismo tamaño con los datos centrados en sector en particular, como se puede apreciar el cluster amarillo concentra sus datos principalmente a la izquierda mientras que el cluster azul los centra mas a la derecha. Los datos los cuales están mas alejados pueden corresponder a datos atipicos dentro de la base de datos, es por esto que los grupos puede aparentar se mas grande de los esperado.

Observando la posición de cada uno de los puntos se puede notar que existen dos áreas donde los puntos tienden a estar mas cerca, estos corresponden al área inferior izquierda del cluster amarillo y el área superior derecha del cluster azul. Dado esto, es posible realizar una categorización de los datos tal como muestra el cluster, para determinar los valores de las variables que predominan dentro de cada cluster es necesario graficar los vectores de cada variable.

Otra observación importantes es que los grupos se encuentran superpuestos, según el algoritmo de K-medias este comportamiento no debería ser posible debido a que los puntos siempre se asignaran al centroide mas cercano. Sin embargo, este gráfico esta representado en solo 2 variables, por lo tanto es posible que ambos grupos no estén realmente superpuestos al graficar al eje Z.

# 6. Conclusiones

A partir del trabajo desarrollado a lo largo de este informe, se logro la aplicación y el entendimiento de lo aprendido en clases, ya que se tuvo que hacer una investigación mas profunda con respecto a que métodos existían para la realización de los cluster tanto teóricamente, como en el lenguaje solicitado (R), al realizar este laboratorio se noto que existe una gran cantidad de literatura acerca de los algoritmos de clusterizacion.

Con respecto a los objetivos de este laboratorio no se cumplieron en su totalidad, ya que falto el apartado de "Analizar por grupo e identificar aquellas caracterısticas mas relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello". Lo anterior es un aspecto a mejorar para los siguientes laboratorios, para lograr lo anterior se debe como grupo mejorar en la organización del tiempo. Con respecto al los aspectos positivos de este laboratorio, es que hubo investigación suficiente, para llegar a un real entendimiento de lo obtenido. Con respecto a los objetivos cumplidos, se ven reflejado a lo largo de este informe en especial en el Análisis de resultado, ya que en el se llega a un resultado favorable, en el cual se cumple con el avance en el entendimiento de las variables que están dentro de este dataset. Cabe mencionar que según el algoritmo de k-medias, los grupos formados que en este caso son 2 que corresponden a grupos del mismo tamaño y los datos se distribuyen de manera similar. Por ultimo se recuerda que en la gráfica de los grupos se ven superpuestos, pero esto puede ser debido a que esta en dos dimensiones, por lo que para ver lo anterior se debería agregar una tercera variable para interpretar de forma mas acertada en ese sector

# Bibliografía

Anónimo (2016). Clustering algorithms. [Online] https://developers.google.com/machine-learning/clustering/clustering-algorithms.

datanovia (2015). Cluster validation essentials. [Online] https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.

Team, E. D. S. (2020). Clustering — when you should use it and avoid it. [Online] https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/#:~:text=Clustering%20is%20an%20unsupervised%20machine,more%20easily%20understood%20and%20manipulated.