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

Laboratorio 2 - Agrupamiento K-medias

Integrantes:	Christian Méndez Acosta
	Israel Arias Panez
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco
Ayudante:	Gustavo Hurtado A.

5 de Octubre de 2022

Tabla de contenidos

1. Introducción

El hipotiroidismo es una de las patologías endocrinas más frecuentes. En Chile según la ENS 2009-2010, la prevalencia total de hipotiroidismo detectada fue de un 19,4 %. En hombres fue de 17.3 % y en mujeres de 21.5 %. En mujeres, la cifra aumenta con la edad, llegando a un 31,3 % en las mayores de 65 años (Martín, sf). A nivel mundial afecta hasta al 5 % de la población, y se calcula que otro 5 % no está diagnosticado. (Chiovato et al., 2019).

Es debido a lo común de esta patología, sumado a las consecuencias que trae a la salud de quienes padecen de ella que estudiar este problema se hace importante e interesante. El estudio de esta enfermedad se llevará a cabo al igual que en la experiencia de laboratorio anterior con la base de datos *Allhypo* la cual fue donada dentro de un grupo de distintas bases de datos con registros sobre enfermedades a la tiroides aportados por el instituto de Garavan y la Universidad de California durante la visita de Ross Quinlan en el taller de Machine Learning en 1987 (Dua and Graff, 2017). Esta base de datos contiene específicamente datos sobre el hipotiroidismo.

El presente informe presenta seis secciones, correspondientes a: introducción, marco teórico, pre-procesamiento, obtención de clúster, análisis de resultados y conclusiones. En la siguiente sección se presentará una breve descripción de cada algoritmo o concepto relevante ocupado en esta experiencia que permitirá comprender lo realizado de mejor forma, en las secciones posteriores se presentará la forma de obtener los clusters, los resultados conseguidos, el análisis de estos para finalmente realizar una conclusión en base a todo lo realizado, recalcando los puntos más relevantes. Los objetivos de esta experiencia de laboratorio son:

- Extraer conocimiento o información relevante respecto al hipotiroidismo, mediante el uso del software R, utilizando algoritmos de clustering como K-means y posteriormente realizar el análisis respectivo.
- Comparar los resultados obtenidos con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- Analizar por grupo e identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

2. Marco Teórico

2.1. Clustering

El Clustering consiste en la agrupación de objetos de manera que los objetos del mismo clúster o grupo sean más similares entre sí que con los de otro clúster. La clasificación en clusters se realiza utilizando criterios como las distancias más pequeñas, la densidad de los puntos de datos, los gráficos o diversas distribuciones estadísticas. El análisis de clústeres tiene una amplia aplicabilidad, incluso en el aprendizaje automático no supervisado, la minería de datos, la estadística, el análisis de gráficos, el procesamiento de imágenes y numerosas aplicaciones de las ciencias físicas y sociales (Nvidia, sf).

2.2. Algoritmo K-means

El clustering de K-means es un algoritmo de aprendizaje no supervisado para clasificar datos no etiquetados agrupándolos por características, en lugar de por categorías o clases predefinidas. La variable K representa el número de grupos o categorías creadas (DeepAi, sf). Para saber si los datos son parecidos o diferentes el algoritmo K-means utiliza la distancia entre los datos. Las observaciones que se parecen tendrán una menor distancia entre ellas. En general, como medida se utiliza la distancia euclideana aunque también se pueden utilizar otras funciones. Ademas, K-means necesita como dato de entrada el número de grupos en los que vamos a segmentar la población. A partir de este número k de clusters, el algoritmo coloca primero k puntos aleatorios (centroides). Luego asigna a cualquiera de esos puntos todas las muestras con las distancias más pequeñas. A continuación, el punto se desplaza a la media de las muestras más cercanas. Esto generará una nueva asignación de muestras, ya que algunas muestras están ahora más cerca de otro centroide. Este proceso se repite de forma iterativa y los grupos se van ajustando hasta que la asignación no cambia más moviendo los puntos. Este resultado final representa el ajuste que maximiza la distancia entre los distintos grupos y minimiza la distancia intragrupo (Duk2, 2019).

2.3. Algoritmo K-mediodes

El algoritmo k-mediodes funciona bajo el mismo concepto que el algoritmo de k-means, pero incorporando una diferencia, que es el usar medioides en vez de centroides.

Para entender esto de mejor forma (Soleymani, sf) menciona que la relación entre centroides y medoides es similar a la relación entre medias y medianas. Los medoides y las medianas siempre serán una de las observaciones de los datos, mientras que ese no es necesariamente el caso de los centroides y las medias. La principal diferencia entre K-means y K-medoides es que K-means formará clusters basados en la distancia de las observaciones a cada centroide, mientras que K-medoides forma clusters basados en la distancia a los medoides.

K-medioides es una alternativa robusta al algoritmo de k-means. Esto significa que, el algoritmo es menos sensible al ruido y a los valores atípicos, en comparación con k-means, porque utiliza medioides como centros de cluster en lugar de medias (utilizadas en k-means) (Kassambara, sf).

2.4. Distancia euclidiana

Es la distancia en línea recta entre dos puntos. Por ejemplo en un plano con un punto $p1$ en $(x1, y1)$ y un punto $p2$ en $(x2, y2)$, la distancia euclidiana es: $\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$ (Black, 2004).

2.5. Distancia de Gower

La distancia de Gower permite medir la disimilaridad entre los registros de un conjunto de datos mixto. Un conjunto de datos mixto se refiere a un conjunto de datos en el que existen conjuntamente características numéricas (variables) y características categóricas (variables). La idea básica de la distancia de Gower es aplicar una métrica de distancia diferente a cada variable en función del tipo de datos: Para las variables numéricas y los factores ordenados (variable categórica ordenada), la distancia se calcula como el valor absoluto de la diferencia entre dos registros (distancia Manhattan). Para las variables categóricas, la distancia es 1 si las categorías entre dos registros son diferentes y la distancia es 0 si las categorías son iguales (distancia de Dice). La fórmula matemática de la distancia de Gower es:

$$S_{ij} = \frac{\sum_k \omega_{ijk} S_{ijk}}{\sum_k \omega_{ijk}} \tag{1}$$

donde S_{ij} es la matriz de disimilitud con los registros, S_{ijk} es la matriz de disimilitud de una variable k-ésima con los registros, ω_{ijk} es el peso de cada matriz de disimilitud S_{ijk}. (Lee, 2021).

3. Pre-procesamiento

Para poder trabajar satisfactoriamente en la creación de clusters y en el estudio del problema, es necesario estudiar la base de datos.

La base de datos presenta 2800 observaciones y 30 variables. Al realizar una exploración inicial es posible el observar que existen variables que no aportan nada al estudio del problema como lo son las variables TSH measured, T3 measured, TT4 measured, T4U measured, FTI measured, TBG measured, todas estas variables tienen relación con indicar si fue medido el nivel de la hormona que indican. Sumado a esto, la variable TBG no presenta registrado ningún dato, solo NA's (no disponible) por lo que tampoco aporta nada al estudio. Por otro lado, la variable referral source hace referencia al lugar de donde provienen las distintas muestras, por lo que tampoco es una variable que aporte al estudio del problema. Una vez identificadas estas ocho variables que no aportan información, se procedió a su eliminación de la base de datos, sumándose también una novena variable correspondiente a la clase, que identifica si los pacientes pertenecen a algún grupo con hipotiroidismo o de personas sanas, esta decisión se realizó debido a que clustering es un método de aprendizaje no supervisado y como sostiene (Mishra, 2017) el objetivo en los problemas de aprendizaje no supervisado puede ser el descubrir grupos de ejemplos similares dentro de los datos. En términos sencillos en las muestras no se proporcionan las verdaderas etiquetas de clase para cada muestra, es de ahí que se conozca al aprendizaje no supervisado como aprendizaje sin profesor. Para reforzar está idea, la (Universidad de Oviedo, sf) menciona que en el aprendizaje no supervisado el dataset viene sin variables (o no se deben ocupar) y la data es clasificada de acuerdo a su propia estructura interna.

A continuación, se estudió respecto a los valores faltantes que presenta la base de datos, en la Figura 1 es posible visualizar el resumen de todos los datos faltantes de la base de datos. Es posible apreciar que se presentan datos faltantes en las variables T3, T4U, FTI, TSH, TT4, sex y age, llegando a presentar una importante cantidad de datos faltantes como un 20 % en la variable T3 y un 10 % en T4U. Respecto a datos faltantes (Lodder et al., 2013) sostienen que la regla general es que si falta menos del 5-10 % de los datos, la eliminación de muestras ya no supone una amenaza importante para la potencia estadística. Por otro lado, (Alice, 2015) recomienda que si los datos que faltan para una determinada característica o muestra son superiores al 5 %, probablemente debería dejar

fuera esa característica o muestra. Tomando en consideración estos factores, se decidió el eliminar las muestras que presentan menos del 5 % de datos faltantes, como ocurre en las variables sex y age. Como una perdida mayor al 5 % afecta al estudio se tomó la decisión de imputar los datos faltantes para las variables correspondientes a la medición de las hormonas tiroideas, que son las únicas variables que presentan perdida de datos en este punto.

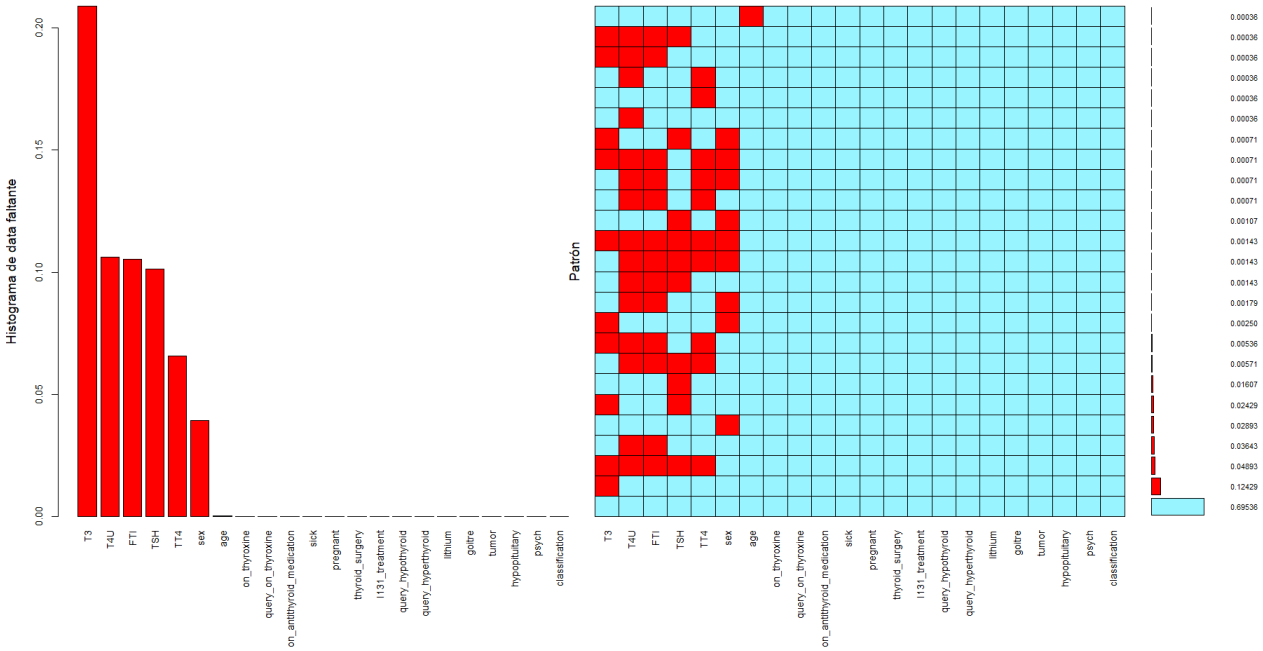

Figura 1: Histograma y gráfica de patrón de datos faltantes

Para la imputación de los datos se decidió usar el método de predictive mean matching, como menciona (Allison, 2015) usar predictive mean matching es una forma atractiva de realizar la imputación de múltiples datos faltantes. En comparación con los métodos estándar basados en la regresión lineal y la distribución normal, predictive mean matching produce valores imputados que se parecen mucho más a los valores reales. El método de predictive mean matching funciona haciendo coincidir la distancia media predictiva de las observaciones incompletas con las de las observaciones completas (Akmam et al., 2019).

Finalmente, la base de datos quedó con 21 variables y 2688 observaciones útiles para el estudio del problema.

Otro punto importante que se debe tomar en consideración es el estudiar la existencia de valores atípicos dentro de la base de datos, considerando que las variables categóricas de la base de datos son dicotómicas, solo hay que estudiar las variables numéricas, en las Fi-

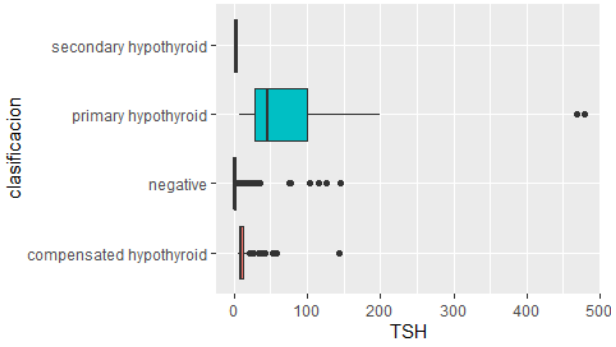

Figura 2: Diagrama de caja: Variable TSH

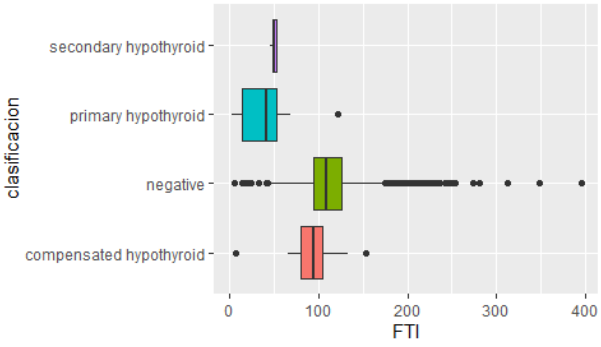

Figura 3: Diagrama de caja: Variable T3

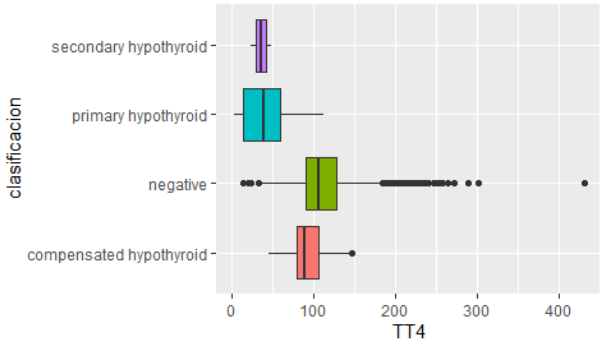

Figura 4: Diagrama de caja: Variable TT4

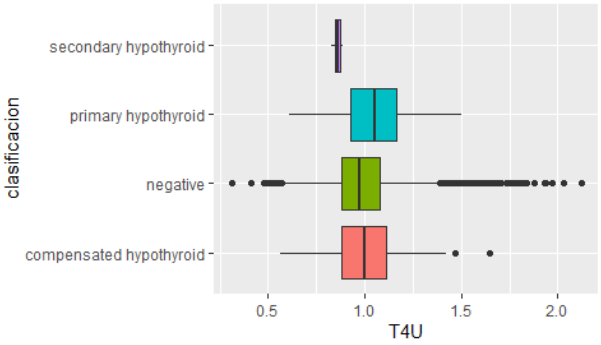

Figura 5: Diagrama de caja: Variable T4U

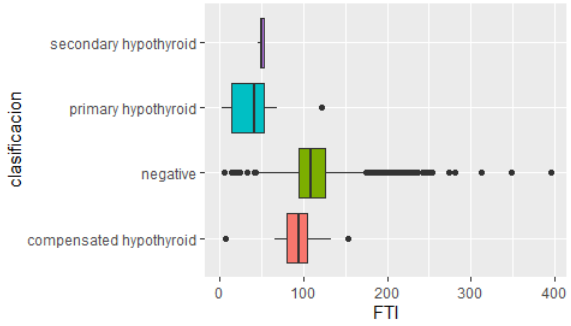

Figura 6: Diagrama de caja: Variable T4U

guras 2, 3, 4, 5 y 6 se presentan diagramas de caja que permiten observar a simple vista los datos atípicos, cabe destacar que no se gráfico la edad debido a que todas las observaciones se encuentran dentro de rangos normales (1 - 94 años).

De los diagramas de caja de las Figuras 2, 3, 4, 5 y 6 se puede observar claramente la presencia de muchos datos atípicos para la medición de las distintas hormonas en las distintas clases. A pesar de esto y tomando en consideración la Tabla 1 elaborada en base a

Hormona	Rango normal
TSH	0,4–5,5 $[\mu U/mL]$
T3	94–170 $[ng/dL]$
TT4	$60 - 150$ $[nmol/L]$
T4U	$0{,}6 - 1{,}3$ $[ug/dL]$
FTI	70–142 $[nmol/L]$

Tabla 1: Hormonas tiroideas y sus rangos normales

lo expuesto por (Actualidad sanitaria, 2020) y (Pantalone et al., 2015) que presenta los rangos normales para cada hormona tiroidea es que se puede apreciar que quienes escapan de los rangos normales corresponden justamente a personas que presentan algún tipo de afección a la tiroides, estos datos son importantes en el estudio y además los datos atípicos de una clase pueden ser valores que caen dentro de los rangos de los cuartiles 1-3 de otra clase, entonces considerando que uno de los objetivos es estudiar la afección del hipotiroidismo y además que se trata de un estudio de clusters, el cual al ser un tipo de aprendizaje no supervisado no considerará las clases, se ha decidido no eliminar ningún dato que se pueda considerar atípico, ya que esto implicaría perder información relevante.

Con el fin de poder generar comparaciones en el estudio, es que se optó por usar la técnica de One Hot Encoding para crear un segundo set de datos, esta técnica permite binarizar variables categóricas, convirtiéndolas a numéricas, se logra creando una columna para cada valor distinto que exista en la variable categórica que estamos codificando y, para cada muestra, marcar con un 1 la columna a la que pertenezca dicha muestra y dejar las demás con 0 (InteractiveChaos, sf). Este nuevo set de datos al estar compuesto solo con variables numéricas puede ser ingresado al algoritmo de K-means. Complementario a este set de datos generados, también se creó otro set que solo contiene las variables numéricas de la base de datos.

Finalmente se normalizan todos los datos numéricos de los sets de datos descritos anteriormente, ya que como (Sharma, 2019) explica: K-means es un algoritmo basado en distancia, mientras menos la distancia entre puntos más es la similaridad y viceversa. Todos los algoritmos basados en distancia son afectados por la escala de las variables, por ejemplo si

se están estudiando datos que tengan la variable edad en años y otra variable que registra el ingreso de una persona en rupias, las edades de las personas varían en rangos de 25 a 40 por ejemplo, mientras que el ingreso varía en rangos de 50.000 a 110.000, al ocurrir esta diferencia de magnitud se puede notar que la alta magnitud del ingreso afecta a la distancia entre los dos puntos y esto impacta a su vez a los resultados, ya que dará un peso mucho mayor a las variables con mayor magnitud (peso en este caso). Entonces para que el algoritmo no se vea afectado por sesgo a variables con mayor magnitud es necesario el trasladar todas las variables a una misma escala, la forma más común de hacer esto es normalizando.

4. Obtención del cluster

Como se estableció en la sección anterior, se crearon tres distintos sets de datos, a continuación, se indica una breve descripción y el método de clustering que se realizará sobre cada uno de estos:

- **Set de datos pre-procesados**: compuesto por las **variables categóricas y numéricas** de la base de datos luego de ser procesada. Dado que se tiene un conjunto de datos mixto, el algoritmo de **K-mediodes con distancia de Gower** se presenta como una opción adecuada para realizar el cluster en este set de datos ya que según (Lee, 2021) la distancia de Gower permite medir la disimilaridad entre los registros de un conjunto de datos mixto. Un conjunto de datos mixto se refiere a un conjunto de datos en el que existen conjuntamente características numéricas (variables) y características categóricas (variables).

- **Set de datos binarizado** (One hot encoding): Contiene todas las variables del set de datos pre-procesado, pero las **variables categóricas fueron cambiadas a binarias**. La técnica de one hot enconding es utilizada cuando se tienen variables con valores categóricos nominales, por cada posible valor se incorpora una nueva variable donde sus valores son mapeados a 1 o 0, siendo 1 si presente la característica y 0 si no la presenta. (Madaan, 2022). Al aplicar el cambio sobre las variables categóricas se tiene únicamente números en la base de datos, por lo cual el método de clustering adecuado a realizar sobre esta base de datos corresponde a **K-means con distancia euclideana**.

- **Set de datos de solo numéricos:** Contiene únicamente aquellas variables que **desde**

un principio contenían valores numéricos (edad, TSH, T3, TT4, T4U y FTI). El método de clustering ideal sobre esta base de datos corresponde a **K-means con distancia euclideana**.

Una vez mencionados los métodos de agrupamiento que serán aplicados sobre los sets de datos, resulta interesante calcular un K adecuado para realizar los clusters. En particular, se realizan dos métodos para la estimación de un K adecuado los cuales son: **método de codo** y **método de la silueta**.

En primer lugar se tiene el método del codo, según (Kumar, 2020) consiste en encontrar el número óptimo de clusters para un set de datos, en este método se designa un rango de posibles valores de K los cuales son aplicados iterativamente sobre el algoritmo K-Means con cada valor de K, el algoritmo en cada iteración encuentra la distancia promedio entre cada punto de un cluster a su centroide y lo representa en un gráfico, finalmente el valor K seleccionado es aquel en que la distancia promedio disminuya repentinamente. Además, Kumar describe el método de la silueta como otro método que también permite encontrar el número K óptimo de clusters. El método de la silueta, computa coeficientes de silueta para cada punto el cual mide que tan parecido es el punto a su propio cluster en comparación a los otros clusters. Entrega una representación gráfica precisa del mejor valor K encontrado.

En la Tabla 2 el resultado de los K mediante los métodos del codo y la silueta para cada base de datos.

Set de datos	**Método del codo**	**Método de la silueta**
Set de datos pre-procesados	4	6
Set de datos binarizado	5	2
Set de datos de solo numéricos	5	2

Tabla 2: Resultado de los K mediante los métodos de codo y silueta.

Para cada set de datos se realizan dos clusters, uno con el valor de K obtenido mediante el método del codo y otro con el valor de K obtenido mediante el método de la silueta.

Para el **Set de datos pre-procesados**, en la Figura 7 se puede apreciar el cluster obtenido con K = 4 (método del codo) y en la Figura 8 el cluster con K = 6 (método de la silueta).

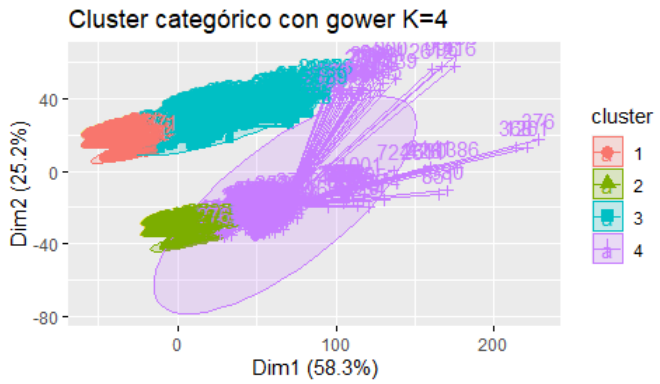

Figura 7: Set de datos pre-procesados, K-medioides con distancia de Gower, K = 4

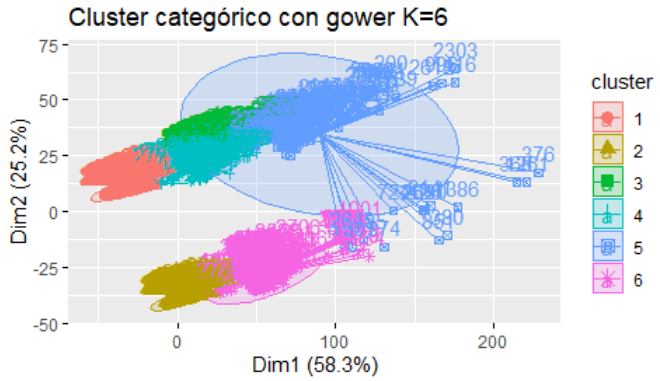

Figura 8: Set de datos pre-procesados, K-medioides con distancia de Gower, K = 6

Para el **Set de datos binarizado**, en la Figura 9 se puede apreciar el cluster obtenido con K = 5 (método del codo) y en la Figura 10 el cluster con K = 2 (método de la silueta).

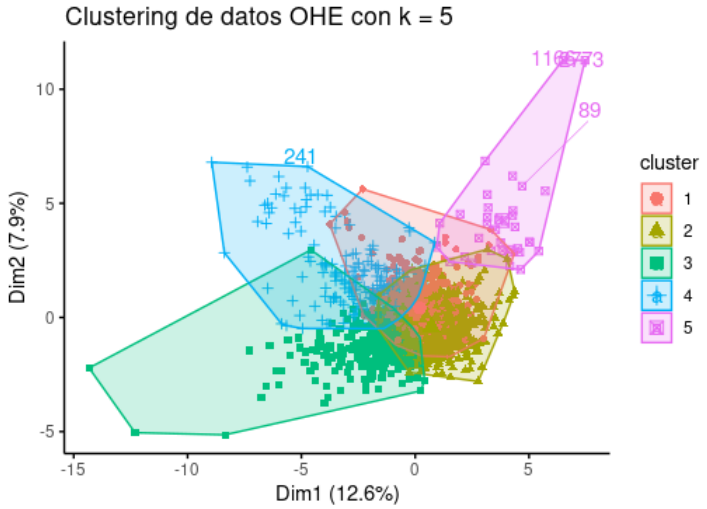

Figura 9: Set de datos binarizado, K-means con K = 5

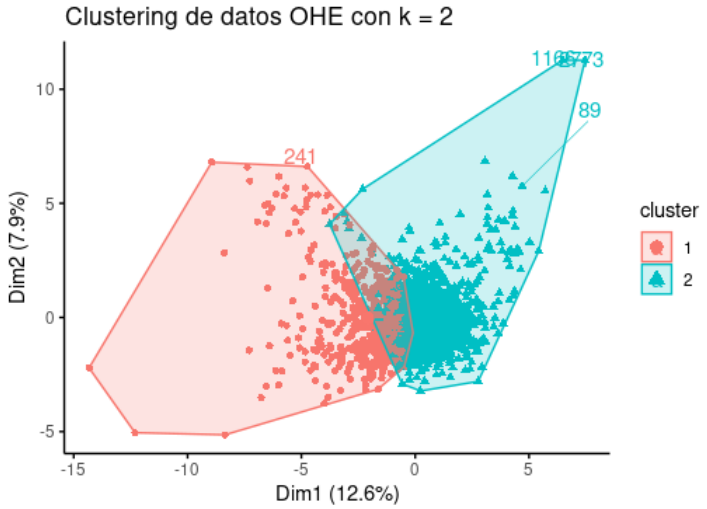

Figura 10: Set de datos binarizado, K-means con K = 2

Para el **Set de datos solo numéricos**, en la Figura 11 se puede apreciar el cluster obtenido con K = 5 (método del codo) y en la Figura 12 el cluster con K = 2 (método de la silueta).

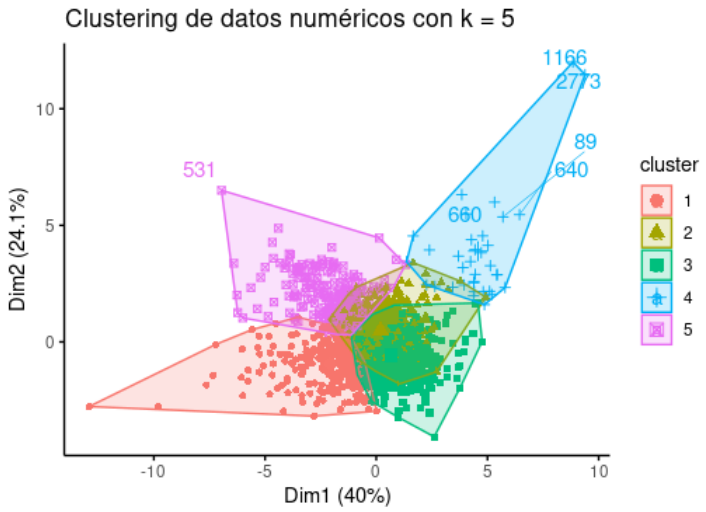

Figura 11: Set de datos solo numéricos, K-means con K = 5

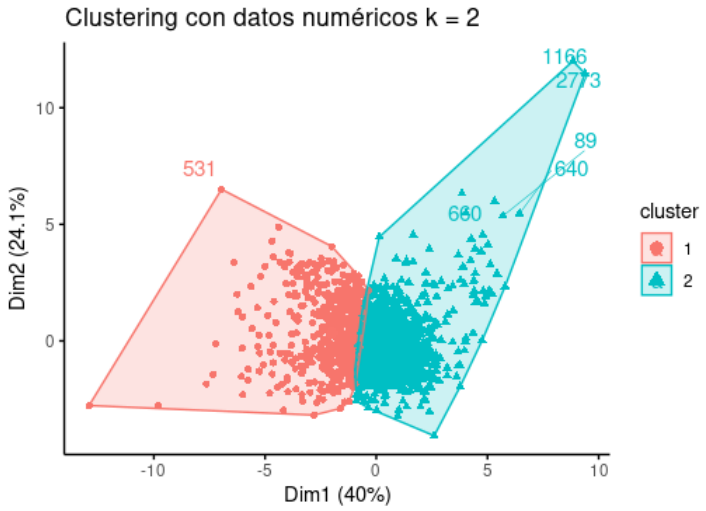

Figura 12: Set de datos solo numéricos, K-means con K = 2

5. Análisis de los resultados

Como fase inicial del análisis de los resultados, se calcula la varianza explicada para cada cluster, la varianza explicada representa el porcentaje de información retenida al haber reducido su dimensionalidad, en este caso a dos componentes. En la tabla 3 se presenta el resumen del cálculo de las varianzas para cada cluster.

Cluster	**Figura**	**Total (%)**
Cluster categórico con K = 4	7	58.3+25.2 = 83.5
Cluster categórico con K = 6	8	58.3+25.2 = 83.5
Clustering con datos One Hot Encoding con K = 5	9	12.6+7.9 = 20.5
Clustering con datos One Hot Encoding con K = 2	10	12.6+7.9 = 20.5
Clustering con datos numéricos con K = 5	11	40+24.1 = 64.1
Clustering con datos numéricos con K = 2	12	40+24.1 = 64.1

Tabla 3: Varianza explicada para cada cluster

Es posible observar en la Figura 3 que los clusters generados con el método One hot encoding tiene una varianza explicada bastante baja, siendo de un 20.5 %. A su vez, los clusters que solo incluyen variables numéricas tienen una varianza del 64 %. Finalmente se toma la decisión de estudiar los cluster categóricos que fueron conseguidos mediante el método de K-mediodes usando la distancia de Gower, al tener una varianza explicada del 83.5 %, (IBM, 2022) define como recomendación que la varianza explicada debe ser de al menos 80 %.

5.1. Análisis de los clusters

A continuación se realizará el análisis de los clusters seleccionados.

5.1.1. Cluster Categórico K = 4

Con el fin de entender este cluster, se realizará una descripción y análisis de todos los grupos formados, en la Figura 13 se puede observar el resumen de este primer grupo o cluster, cabe destacar que se incluye la clasificación de hipotiroidismo en este y en los demás

clusters a presentar, pero esta solo fue agregada al final a modo de poder observar si existe alguna relación, no estuvo contenida en ningún proceso del desarrollo del cluster.

Es posible observar que en este cluster se agrupan mujeres no embarazadas, además cabe destacar que ninguna presenta algún tratamiento enfocado a tratar el hipotiroidismo, detallado por las variables categóricas que se encuentran en false. Respecto a las hormonas se observan datos máximos fuera de los valores de referencia (detallados en la Figura 1), sin embargo, presentan una media normal dentro de estos valores. En conclusión, este cluster se caracteriza por agrupar solo a mujeres no embarazadas que no se encuentran en ningún tratamiento relacionado al hipotiroidismo, además tampoco presentan alguna característica relacionada a la patología como lo es tumores, bocio, etc.

```
[[1]]
      age        sex      on_thyroxine query_on_thyroxine on_antithyroid_medication sick     pregnant thyroid_surgery
 Min.   : 2.00   F:1049   f:1049       f:1049             f:1049                    f:1049   f:1049   f:1049
 1st Qu.:36.00   M:   0   t:   0       t:   0             t:   0                    t:   0   t:   0   t:   0
 Median :55.00
 Mean   :52.72
 3rd Qu.:70.00
 Max.   :94.00
 I131_treatment query_hypothyroid query_hyperthyroid lithium  goitre   tumor    hypopituitary psych         TSH
 f:1049         f:1049            f:1049             f:1049   f:1049   f:1049   f:1049        f:1049   Min.   :  0.005
 t:   0         t:   0            t:   0             t:   0   t:   0   t:   0   t:   0        t:   0   1st Qu.:  0.650
                                                                                                      Median :  1.500
                                                                                                      Mean   :  3.680
                                                                                                      3rd Qu.:  2.800
                                                                                                      Max.   :151.000
       T3              TT4             T4U             FTI           cluster                  clasificacion
 Min.   :0.050   Min.   :  3.0   Min.   :0.480   Min.   :  3.0   Min.   :1   compensated hypothyroid: 74
 1st Qu.:1.600   1st Qu.: 88.0   1st Qu.:0.880   1st Qu.: 92.0   1st Qu.:1   negative               :955
 Median :2.000   Median :103.0   Median :0.980   Median :104.0   Median :1   primary hypothyroid    : 19
 Mean   :1.944   Mean   :105.9   Mean   :0.995   Mean   :107.4   Mean   :1   secondary hypothyroid  :  1
 3rd Qu.:2.300   3rd Qu.:121.0   3rd Qu.:1.080   3rd Qu.:120.0   3rd Qu.:1
 Max.   :5.100   Max.   :230.0   Max.   :1.650   Max.   :219.0   Max.   :1
```

Figura 13: Tabla resumen: Cluster 1 con K=4

Respecto al segundo cluster cuyo resumen se puede observar en la Figura 14, se caracteriza por agrupar solo a hombres que no presentan ninguna condición relacionada al hipotiroidismo, como tampoco se encuentran en algún tratamiento relacionado. Al igual que el cluster anterior se puede observar la existencia de máximos atípicos para las hormonas, pero medias y medianas dentro de los rangos normales.

Respecto al tercer cluster, cuyo resumen se puede observar en la Figura 15, se caracteriza por agrupar solo a mujeres que además pueden presentar algunas características relacionadas al hipotiroidismo como la presencia de bocio, tumores, tratamiento de yodo radiactivo, etc. Además de estar en tratamientos relacionados al hipotiroidismo, como lo son el estar con medicación antitiroidea o estar en tratamiento de tiroxina.

```
[[2]]
      age        sex     on_thyroxine query_on_thyroxine on_antithyroid_medication sick    pregnant thyroid_surgery
 Min.   : 1.00   F:  0   f:577        f:577              f:577                     f:577   f:577    f:577
 1st Qu.:40.00   M:577   t:  0        t:  0              t:  0                     t:  0   t:  0    t:  0
 Median :55.00
 Mean   :52.55
 3rd Qu.:67.00
 Max.   :94.00
 I131_treatment query_hypothyroid query_hyperthyroid lithium goitre  tumor   hypopituitary psych        TSH
 f:577          f:577             f:577              f:577   f:577   f:577   f:577         f:577   Min.   :  0.005
 t:  0          t:  0             t:  0              t:  0   t:  0   t:  0   t:  0         t:  0   1st Qu.:  0.610
                                                                                                 Median :  1.400
                                                                                                 Mean   :  2.761
                                                                                                 3rd Qu.:  2.300
                                                                                                 Max.   :103.000
       T3              TT4              T4U              FTI           cluster                 clasificacion
 Min.   :0.200   Min.   : 14.00   Min.   :0.5000   Min.   : 13.0   Min.   :2   compensated hypothyroid: 26
 1st Qu.:1.500   1st Qu.: 84.00   1st Qu.:0.8300   1st Qu.: 92.0   1st Qu.:2   negative               :543
 Median :1.900   Median : 98.00   Median :0.9300   Median :106.0   Median :2   primary hypothyroid    :  8
 Mean   :1.894   Mean   : 99.64   Mean   :0.9358   Mean   :107.1   Mean   :2   secondary hypothyroid  :  0
 3rd Qu.:2.300   3rd Qu.:114.00   3rd Qu.:1.0200   3rd Qu.:121.0   3rd Qu.:2
 Max.   :4.800   Max.   :187.00   Max.   :1.6500   Max.   :204.0   Max.   :2
```

Figura 14: Tabla resumen: Cluster 2 con K=4

```
[[3]]
      age        sex     on_thyroxine query_on_thyroxine on_antithyroid_medication sick    pregnant thyroid_surgery
 Min.   : 2.00   F:745   f:503        f:727              f:721                     f:673   f:725    f:713
 1st Qu.:35.00   M:  0   t:242        t: 18              t: 24                     t: 72   t: 20    t: 32
 Median :55.00
 Mean   :51.16
 3rd Qu.:66.00
 Max.   :93.00
 I131_treatment query_hypothyroid query_hyperthyroid lithium goitre  tumor   hypopituitary psych        TSH
 f:716          f:632             f:627              f:733   f:732   f:689   f:745         f:687   Min.   :  0.005
 t: 29          t:113             t:118              t: 12   t: 13   t: 56   t:  0         t: 58   1st Qu.:  0.200
                                                                                                 Median :  1.100
                                                                                                 Mean   :  6.199
                                                                                                 3rd Qu.:  2.800
                                                                                                 Max.   :468.000
       T3              TT4              T4U              FTI           cluster                 clasificacion
 Min.   :0.100   Min.   :  2.0   Min.   :0.310   Min.   :  2   Min.   :3   compensated hypothyroid: 37
 1st Qu.:1.700   1st Qu.: 93.0   1st Qu.:0.910   1st Qu.: 93   1st Qu.:3   negative               :682
 Median :2.100   Median :113.0   Median :1.020   Median :111   Median :3   primary hypothyroid    : 26
 Mean   :2.218   Mean   :119.6   Mean   :1.057   Mean   :115   Mean   :3   secondary hypothyroid  :  0
 3rd Qu.:2.500   3rd Qu.:141.0   3rd Qu.:1.130   3rd Qu.:132   3rd Qu.:3
 Max.   :7.100   Max.   :301.0   Max.   :2.120   Max.   :312   Max.   :3
```

Figura 15: Tabla resumen: Cluster 3 con K=4

Respecto al cuarto cluster, cuyo resumen se puede observar en la Figura 16 se caracteriza por tener una mayor cantidad de hombres (89.2 %) que mujeres (10.8 %) y al igual que en el cluster anterior algunas de estas personas presentan condición relacionada al hipotiroidismo y/o el estar en algún tratamiento relacionado al hipotiroidismo.

Tomando en consideración las conclusiones recopiladas de cada cluster y con el fin de caracterizar de forma global este agrupamiento, es que en la Figura 17 se presenta de forma gráfica la caracterización general de cada cluster, sumado a la explicación de las dimensiones, de las cuales se concluyó que la dimensión 1 caracteriza la presencia de condiciones o tratamientos ligados al hipotiroidismo, mientras que la dimensión 2 caracteriza el

```
[[4]]
      age          sex      on_thyroxine query_on_thyroxine on_antithyroid_medication sick    pregnant thyroid_surgery
 Min.   : 1.00   F: 34   f:238        f:295              f:307                      f:281   f:297    f:310
 1st Qu.:34.00   M:283   t: 79        t: 22              t: 10                      t: 36   t: 20    t:  7
 Median :50.00
 Mean   :48.52
 3rd Qu.:62.00
 Max.   :84.00
 I131_treatment query_hypothyroid query_hyperthyroid lithium goitre  tumor   hypopituitary psych        TSH
 f:302          f:270             f:267              f:315   f:305   f:303   f:316         f:241   Min.   :  0.005
 t: 15          t: 47             t: 50              t:  2   t: 12   t: 14   t:  1         t: 76   1st Qu.:  0.250
                                                                                                  Median :  1.100
                                                                                                  Mean   :  5.623
                                                                                                  3rd Qu.:  2.200
                                                                                                  Max.   :478.000
       T3              TT4              T4U              FTI          cluster                clasificacion
 Min.   : 0.20   Min.   :  5.8   Min.   :0.4100   Min.   :  7.0   Min.   :4   compensated hypothyroid: 12
 1st Qu.: 1.60   1st Qu.: 86.0   1st Qu.:0.8600   1st Qu.: 95.0   1st Qu.:4   negative               :297
 Median : 2.00   Median :102.0   Median :0.9400   Median :108.0   Median :4   primary hypothyroid    :  7
 Mean   : 2.22   Mean   :110.8   Mean   :0.9926   Mean   :113.1   Mean   :4   secondary hypothyroid  :  1
 3rd Qu.: 2.50   3rd Qu.:127.0   3rd Qu.:1.0700   3rd Qu.:125.0   3rd Qu.:4
 Max.   :10.60   Max.   :430.0   Max.   :1.9700   Max.   :395.0   Max.   :4
```

Figura 16: Tabla resumen: Cluster 4 con K=4

sexo de la persona.

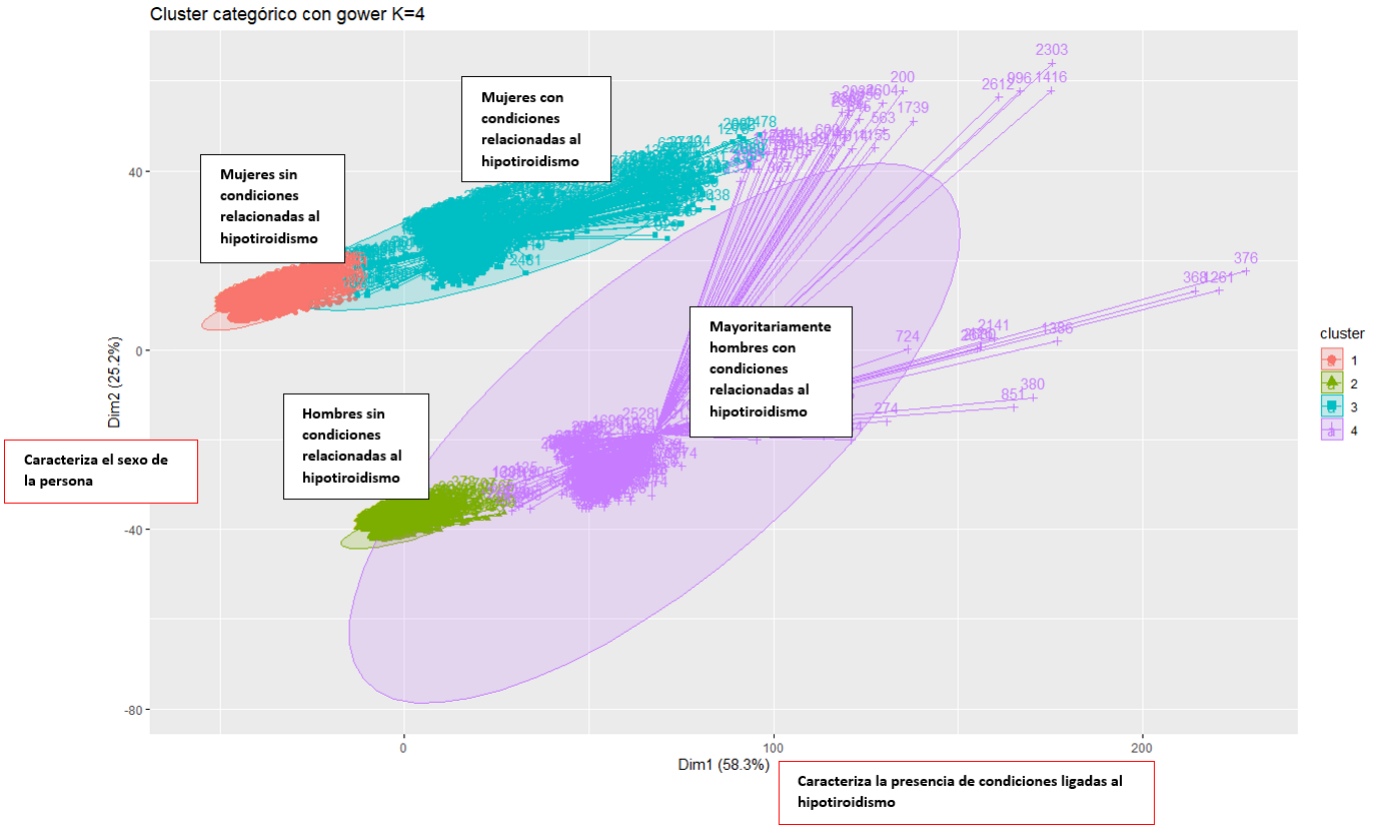

Figura 17: Explicación gráfica de las dimensiones y de los clusters K = 4

5.1.2. Cluster categórico K = 6

Este cluster al ser estudiado dio a conocer que repite el mismo comportamiento que el cluster presentado en la sección anterior (Sección 5.1.1.), es por esto que se ha decidido presentar su comportamiento resumido en la Figura 18. Se puede observar que solo agrega dos nuevos grupos, pero no aporta nueva información.

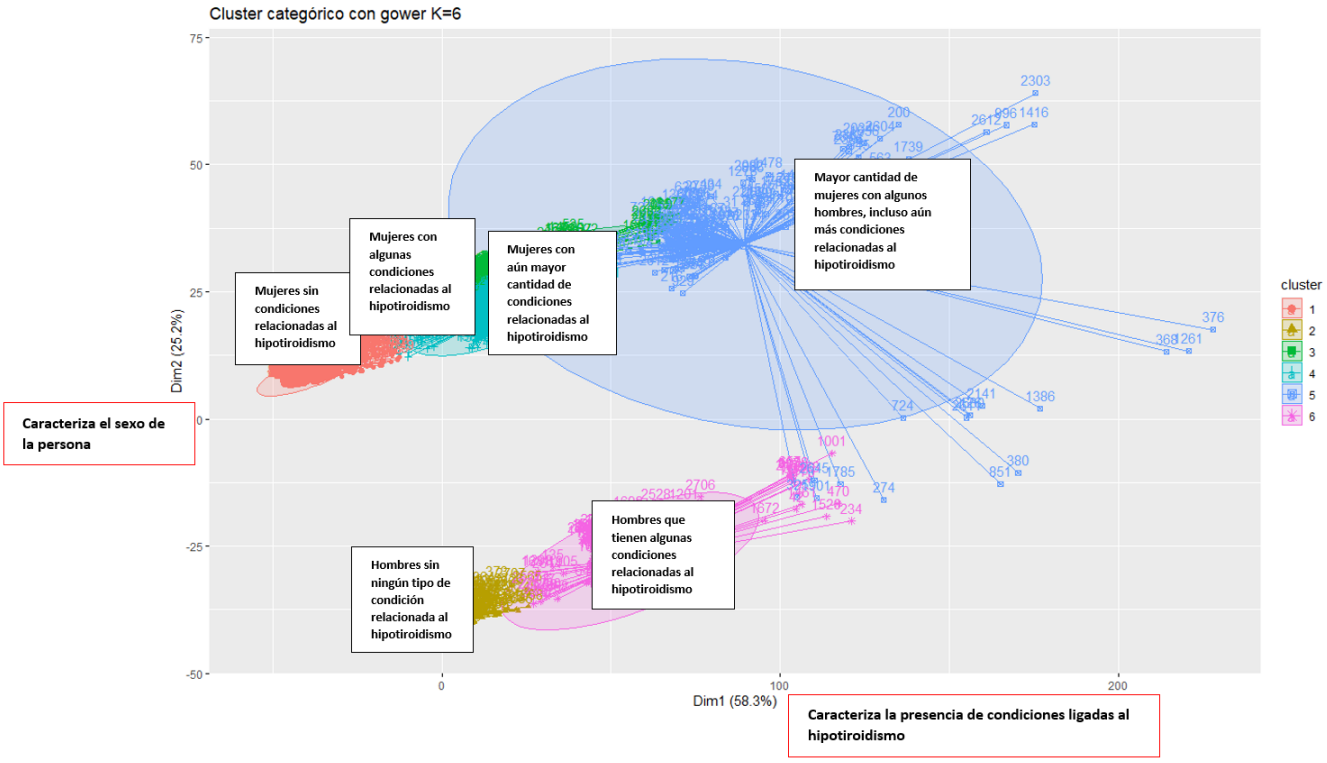

Figura 18: Explicación gráfica de las dimensiones y de los clusters K = 6

Como se pudo observar en ambos clusters, el agrupamiento se realizó tomando en consideración el sexo de las personas y la presencia de algunas condiciones ligadas al hipotiroidismo, denotadas en las variables categóricas. Sin embargo no hay mucha información del problema que se pueda extraer con estos resultados, incluso es posible visualizar el cómo no había tendencia en las clases (negativo, hipotiroidismo primario, hipotiroidismo compensado, hipotiroidismo secundario) en ningún grupo y las variables numéricas tampoco parecían entregar ninguna información, esto causó intriga, debido a que como se menciona en (McMaster Textbook of Internal Medicine, 2021) el diagnóstico del hipotiroidismo se hace a través de exámenes de hormonas, en el cual se registran las concentraciones de las hormonas TSH y FTI principalmente, siendo a veces requerida la concentración de hormona T3. En-

tonces tomando de guía la literatura se decidió el estudiar el algoritmo de clustering k-Means ocupando el set de datos que solo contiene las variables continuas, que corresponde en su mayoría a las medidas hormonales, esto debiese entregar información del problema que los clusters presentados no entregaron.

5.2. Cluster K-means númerico K = 3

Para la creación de este cluster se decidió el valor de la variable K = 3 debido a que se tienen cuatro clases, sin embargo, se tienen muy pocas observaciones de la clase hipotiroidismo secundario, por lo que se optó por técnicamente igualar la cantidad de grupos a la cantidad de clases. El cluster generado se puede observar en la Figura 24. Tiene una varianza explicada de 64.1 %.

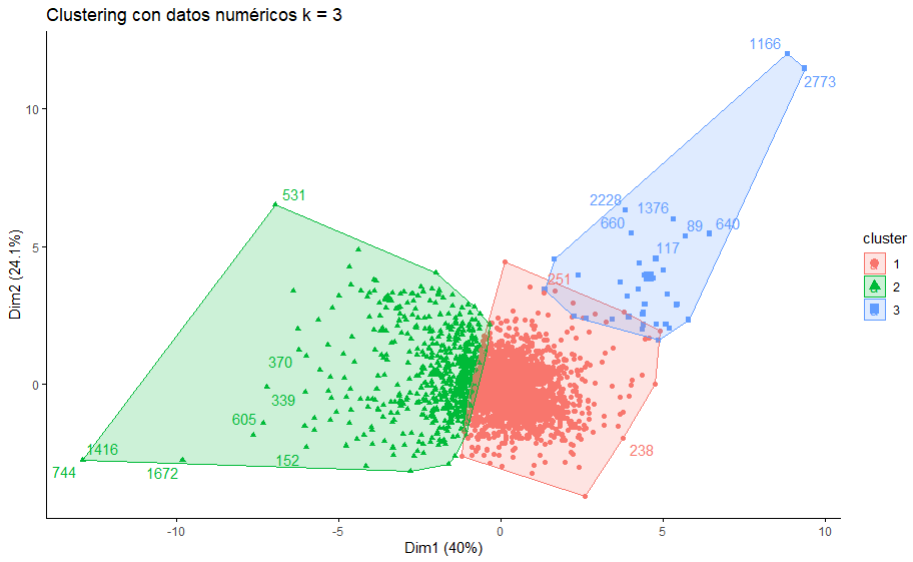

Figura 19: Cluster K-means númerico con K = 3

Tomando en consideración la Tabla 4 que muestra el porcentaje de distribución de clases en los tres cluster generados, es posible el observar como en el Cluster 3 se presenta un porcentaje mayoritario de 46.66 % de pacientes con hipotiroidismo primario, en la Figura 22 se detalla que en este cluster hay 28 pacientes con hipotiroidismo primario, 1 con hipotiroidismo compensado, 6 con hipotiroidismo negativo y 0 secundarios. En resumen, el cluster 3 agrupa a los pacientes con hipotiroidismo primario.

Por otro lado en el Cluster 2, como menciona la Tabla 4 se observa un 22.6 % de la

totalidad de pacientes negativos, esto corresponde a 560 pacientes como detalla la Figura 21, a su vez hay 12 pacientes con hipotiroidismo compensado o subclínico, no hay pacientes de las otras dos clases restantes. En resumen, el Cluster 2 agrupa a los pacientes que no tienen hipotiroidismo, o sea que son de la clase negativa.

Finalmente en el Cluster 1, como se observa en la Tabla 4 presenta un porcentaje alto de observaciones de las distintas clases, al ver las cifras exactas en la Figura 20 se puede apreciar que corresponden a 136 pacientes con hipotiroidismo compensado, 1911 pacientes con hipotiroidismo negativo, 32 pacientes con hipotiroidismo primario y 2 con hipotiroidismo secundario. En resumen, hay una gran cantidad de pacientes que deberían de tener niveles normales en sus hormonas (Clase negativa), pero al estar el 91 % del total de pacientes con hipotiroidismo compensado y el 53.33 % de pacientes del total de pacientes con hipotiroidismo primario, se esperaría una tendencia en la medidas estadísticas de las hormonas del cluster en comparación a las presentes del cluster 2 que agrupa solo a negativos. Este cluster no agrupa a una clase en especifico aparentemente.

```
[[1]]
      age              TSH               T3              TT4              T4U              FTI           cluster
 Min.   : 1.00   Min.   : 0.005   Min.   :0.050   Min.   :  3.00   Min.   :0.3100   Min.   :  3.0   Min.   :1
 1st Qu.:39.00   1st Qu.: 0.640   1st Qu.:1.500   1st Qu.: 84.00   1st Qu.:0.8600   1st Qu.: 90.0   1st Qu.:1
 Median :57.00   Median : 1.500   Median :1.900   Median : 98.00   Median :0.9400   Median :103.0   Median :1
 Mean   :53.82   Mean   : 3.108   Mean   :1.825   Mean   : 97.08   Mean   :0.9476   Mean   :103.6   Mean   :1
 3rd Qu.:69.00   3rd Qu.: 2.800   3rd Qu.:2.200   3rd Qu.:111.00   3rd Qu.:1.0400   3rd Qu.:117.0   3rd Qu.:1
 Max.   :94.00   Max.   :58.000   Max.   :4.800   Max.   :165.00   Max.   :1.6800   Max.   :189.0   Max.   :1
                 clasificacion
 compensated hypothyroid: 136
 negative               :1911
 primary hypothyroid    :  32
 secondary hypothyroid  :   2
```

Figura 20: Tabla resumen: Cluster N°1 K-means numérico con K = 3

```
[[2]]
      age              TSH               T3              TT4              T4U              FTI           cluster
 Min.   : 2.00   Min.   : 0.005   Min.   : 1.100   Min.   : 90.0   Min.   :0.640   Min.   : 58.0   Min.   :2
 1st Qu.:29.00   1st Qu.: 0.065   1st Qu.: 2.300   1st Qu.:134.0   1st Qu.:0.990   1st Qu.:109.0   1st Qu.:2
 Median :41.50   Median : 0.540   Median : 2.600   Median :149.0   Median :1.130   Median :131.0   Median :2
 Mean   :44.44   Mean   : 1.338   Mean   : 2.899   Mean   :156.7   Mean   :1.181   Mean   :138.5   Mean   :2
 3rd Qu.:60.00   3rd Qu.: 1.700   3rd Qu.: 3.400   3rd Qu.:170.0   3rd Qu.:1.310   3rd Qu.:159.0   3rd Qu.:2
 Max.   :94.00   Max.   :26.000   Max.   :10.600   Max.   :430.0   Max.   :2.120   Max.   :395.0   Max.   :2
                 clasificacion
 compensated hypothyroid: 12
 negative               :560
 primary hypothyroid    :  0
 secondary hypothyroid  :  0
```

Figura 21: Tabla resumen: Cluster N°2 K-means numérico con K = 3

A continuación se realizará una comparación y análisis más exhaustivo, tomando en consideración lo establecido en la descripción de cada cluster.

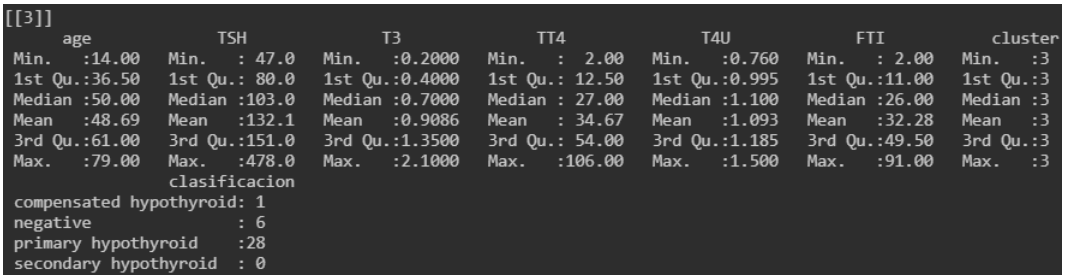

Figura 22: Tabla resumen: Cluster N°3 K-means numérico con K = 3

Cluster	**H. Compensado**	**H. primario**	**H. secundario**	**Negativos**
1	91.27 %	53.34 %	100 %	77.14 %
2	8.06 %	0 %	0 %	22.61 %
3	0.67 %	46.66 %	0 %	0.25 %

Tabla 4: Tabla porcentaje distribución de clases en los clusters

Como se mencionó, el cluster 3 agrupa a los pacientes con hipotiroidismo primario, pero **¿qué caracteriza a un paciente con hipotiroidismo primario?** En (McMaster Textbook of Internal Medicine, 2021) se define un algoritmo para el diagnostico del hipotiroidismo, el cual se presenta en la Figura 23. Según el algoritmo una persona con hipotiroidismo primario se caracteriza por tener **niveles hormonales de TSH alta y FTI (Tiroxina libre) baja**. Los niveles normales de la TSH se sitúan entre 0,37 y 4,7 mUI/L. Los niveles de T4 en plasma (FTI) se sitúan entre 60 y 150 nmol/L y los de T3 se sitúan entre 1,2 y 2,7 nmol/L (Gonzalez, 2021).

Es posible observar que la media de TSH se escapa con creces en el cluster 3, siendo de 132.1 mUI/L en comparación al rango normal de [0.37-4.7] mUI/L, además la media de FTI es de 32.28 nmol/L siendo mucho más baja que los rangos normales de [60-150] nmol/L, se puede observar entonces que esta agrupación de personas con hipotiroidismo primario sigue a cabalidad lo que se expresa en la literatura, cumpliendo con ambos requisitos expuestos en el algoritmo de la Tabla 23. En consecuencia, es posible el afirmar que estas personas al poseer las características particulares del hipotiroidismo primario han sido agrupadas en un mismo cluster. Cabe recordar que el cluster no conoce las clases de antemano, ya que fueron removidas para su creación, al tratarse de un algoritmo de aprendizaje no supervisado.

Respecto al cluster 2 que agrupa principalmente a las personas de la clase negativa,

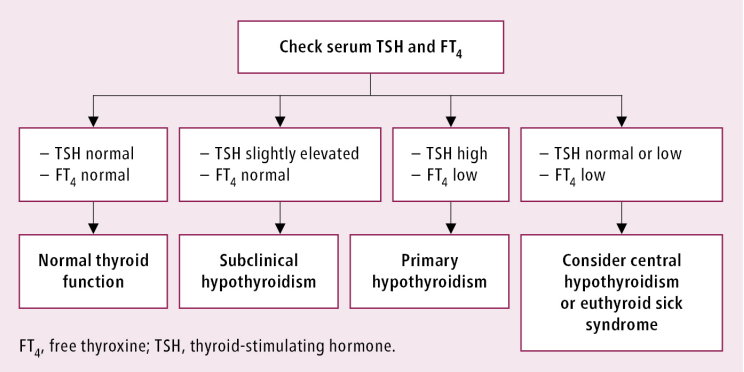

Figura 23: Algoritmo de diagnostico de hipotiroidismo basado en los niveles de TSH y T45 libre. *Adaptado de: McMaster Textbook of Internal Medicine (2021). https://empendium.com/mcmtextbook/chapter/B31.II.9.1*

es posible observar que la media de TSH es de 1.338 mUI/L, lo cual se encuentra dentro del rango normal expuesto de [0.37-4.7] mUI/L, además la hormona FTI tiene una media de 138.5 nmol/L, también cumpliendo los rangos normales expuestos [60-150] nmol/L. Para este cluster también se cumplió lo expuesto en la literatura (algoritmo de la Tabla 23), lo que significa que el algoritmo de K-means fue capaz de agrupar a quienes no tienen problemas de hipotiroidismo en un solo grupo. Incluso es posible el comparar a este grupo con el del cluster 3, se puede observar como la media de TSH es menor para este grupo que el del cluster 3 (1.338 [mUI/L] v/s 132.1 [mUI/L]) y a su vez la media de FTI es mayor a la del cluster 3 (138.5 [nmol/L] v/s 32.28[nmol/L]). Lo que concuerda con (Romero and Almazán, 2015) quienes sostienen que un patrón de hipotiroidismo primario se observa cuando se tiene una TSH elevada con una T4(FTI) baja respecto a los niveles hormonales normales.

Respecto al cluster 1, se mencionó que se esperaba una posible tendencia en las hormonas en comparación al cluster 2. Tomando en consideración al (McMaster Textbook of Internal Medicine, 2021) con su algoritmo de detección de patologías de hipotiroidismo de la Figura 23, una persona con hipotiroidismo subclínico o compensado tendrá una TSH ligeramente elevada y un FTI normal, también que quienes tienen hipotiroidismo primario tendrán un TSH alto y una FTI baja. Por lo que considerando que en el cluster 1 hay 136 pacientes con hipotiroidismo compensado, 1911 pacientes con hipotiroidismo negativo, 32 pacientes con hipotiroidismo primario y 2 con hipotiroidismo secundario, se esperaría que la media de las hormonas tienda a parecerse a los del cluster 2 (grupo de pacientes sin

hipotiroidismo), pero al haber pacientes de dos clases (un total de 8 % del cluster) que se caracterizan por tener una TSH más alta a la normal (h.primario y h.compensado) y una FTI más baja (h.primario) se espera el que estos valores tiendan a alterar la media de los rangos normales de las hormonas, aumentando el TSH y disminuyendo el FTI. Esto efectivamente se cumple, la media de TSH del grupo de los negativos (cluster 2) es de 1.338 [mUI/L], mientras que la media de TSH del cluster 1 es de 3.108 [mUI/L], en el caso de la FTI, en el cluster 2 la media es de 138.5 nmol/L, mientras que la media del cluster 1 es de 103.6 nmol/L.

Para finalizar el estudio de este cluster, la dimensión 1 del cluster se encuentra caracterizada por los niveles de TSH, aumentando a medida que se desplaza hacia la derecha, en cuanto a la dimensión 2 se encuentra caracterizada por los niveles de FTI, disminuyendo a medida que se desplaza hacia arriba. Este comportamiento se puede verificar por medio de las Figuras 20, 21 y 22.

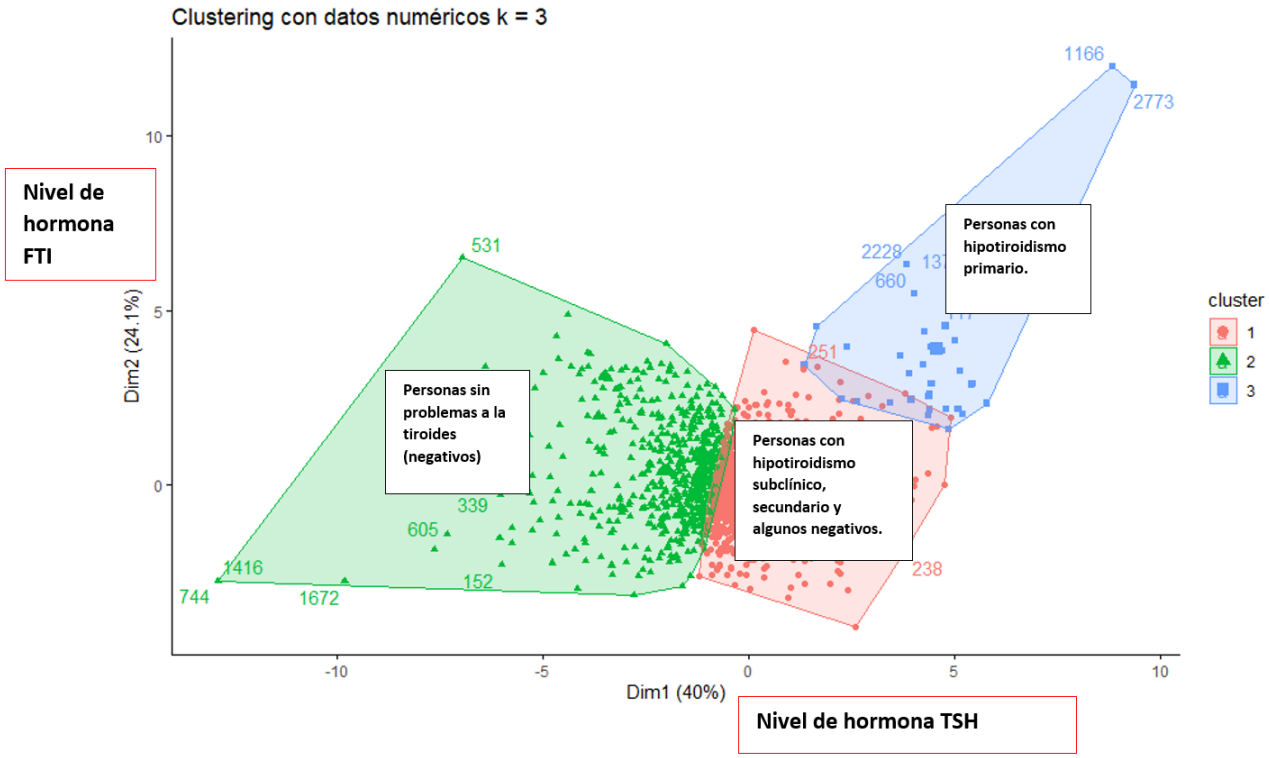

Figura 24: Explicación gráfica cluster K-means numérico K=3

6. Conclusión

En la presente experiencia de laboratorio se realizó el estudio de la enfermedad hipotiroidismo mediante algoritmos de clustering como lo son K-means y K-medioides, con uno de los objetivos de extraer información relevante de la estructura de la base de datos. El clustering de datos ayuda a discernir la estructura y simplifica la complejidad de cantidades masivas de datos (Villagra et al., 2009).

El estudio comenzó estableciendo un marco teórico para la comprensión de los distintos algoritmos o distancias utilizadas, posteriormente se realizó y detalló la fase de pre-procesamiento de datos, en la cual se utilizaron métodos basados en la literatura para eliminar datos atípicos e imputar datos faltantes, además de crear distintos sets de datos para la obtención de clusters. Los clusters fueron obtenidos mediante K-medias y K-mediodies utilizando el método del codo y la silueta para encontrar un valor K optimo, además de las distancias euclidiana y de Gower. Posteriormente en la obtención de resultados se realizarón comparaciones con la literatura y con otros clusters, con el fin de analizar si realmente se estaba consiguiendo nueva información.

Respecto a los resultados relevantes que se encontraron durante el estudio, se puede mencionar que fue posible el encontrar información oculta en la estructura de los datos, logrando identificar distintas clases gracias a las características de sus hormonas, esto no hubiese sido posible sin la investigación y la comparación con la literatura, debido a que en un inicio se intentó el realizar clustering con todas las variables categóricas y continuas, lo cual no arrojó ningún comportamiento o información adicional respecto al problema, fue gracias a la investigación que se descubrió que el diagnóstico del hipotiroidismo se hace a través de exámenes de hormonas, en el cual se registran las concentraciones de las hormonas TSH y FTI principalmente (McMaster Textbook of Internal Medicine, 2021). En base a esto se realizó un nuevo cluster, utilizando solo las variables continuas que corresponden a los niveles hormonales y edad del paciente. Este nuevo resultado fue satisfactorio.

Respecto a los métodos utilizados, podemos concluir por la experiencia desarrollada que los algoritmos de clustering no siempre logran aportar información al estudio de un fenómeno o problema, sin embargo, si se complementan con una investigación exhaustiva respecto al contexto de la problemática abordada se podrá realizar una correcta interpretación y hacer frente a los posibles problemas que pueden aparecer al usar estos algoritmos,

como los documentados en esta experiencia de laboratorio.

Respecto a los objetivos específicos detallados en la sección de introducción fueron cumplidos en su totalidad se extrajo conocimiento relevante del hipotiroidismo implementando algoritmos de Clustering en el software R. Además, se le dio sentido a los resultados obtenidos, mediante la comparación con la literatura, lo cual arrojó que efectivamente sustentaba el conocimiento que se había obtenido. Finalmente se realizó un análisis por grupos de los distintos clusters, en donde se identificaron las características más relevantes de cada grupo, como también se caracterizaron las dimensionalidades de cada cluster y se caracterizó cada grupo individualmente. Lo que permitió adquirir y comprobar nuevo conocimiento, como lo fue el obtener las clasificaciones de hipotiroidismo solo a través de los niveles hormonales presentes en los datos.

Para concluir, respecto a las posibles mejoras en el desarrollo de esta experiencia se podría haber investigado sobre distintos tipos de algoritmos de clustering que no fueron abordados en el estudio, como por ejemplo algoritmos de clustering jerárquico, con el fin de realizar más comparaciones que podrían haber mejorado los resultados.

Bibliografía

Actualidad sanitaria (2020). ¿cuáles son los niveles normales de tiroides? https://actualidadsanitaria.com/vida-saludable/cuales-son-los-niveles-normales-de-tiroides/.

Akmam, E. F., Siswantining, T., Soemartojo, S. M., and Sarwinda, D. (2019). Multiple imputation with predictive mean matching method for numerical missing data. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–6. IEEE.

Alice, M. (2015). Imputing missing data with r; mice package. *Data Science Plus*.

Allison, P. (2015). Imputation by predictive mean matching: Promise & peril. *Statistical Horizons*.

Black, P. E. (2004). Euclidean distance. https://xlinux.nist.gov/dads/HTML/euclidndstnc.html.

Chiovato, L., Magri, F., and Carlé, A. (2019). Hypothyroidism in context: where we've been and where we're going. *Advances in therapy*, 36(2):47–58.

DeepAi (s.f.). K-means. https://deepai.org/machine-learning-glossary-and-terms/k-means.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Duk2 (2019). K-means clustering: Agrupamiento con minería de datos. https://estrategiastrading.com/k-means/.

Gonzalez, A. (2021). Pruebas de función tiroidea. https://www.cun.es/enfermedades-tratamientos/pruebas-diagnosticas/funcion-tiroidea.

IBM (2022). Proporción de varianza explicada. https://www.ibm.com/docs/es/cloud-paks/cp-data/4.5.x?topic=overview-proportion-explained-variance.

InteractiveChaos (s.f.). One hot encoding. https://interactivechaos.com/es/manual/tutorial-de-machine-learning/one-hot-encoding.

Kassambara, A. (s.f.). K-medoids in r: Algorithm and practical examples. https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/.

Kumar, S. (2020). Why is scaling required in knn and k-means? https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891.

Lee, D. (2021). Gower's distance. https://daesoolee.tistory.com/112.

Lodder, P. et al. (2013). To impute or not impute: That's the question. *Advising on research methods: Selected topics*, pages 1–7.

Madaan, M. (2022). Handling categorical variables with one-hot encoding. https://www.naukri.com/learning/articles/handling-categorical-variables-with-one-hot-encoding/.

Martín, M. S. (s.f.). Hipotiroidismo en adultos: ¿lo sabemos todo? *Meta*, 70(80):4–6.

McMaster Textbook of Internal Medicine (2021). Hypothyroidism. https://empendium.com/mcmtextbook/chapter/B31.II.9.1.

Mishra, S. (2017). Unsupervised learning and data clustering. https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a.

Nvidia (s.f.). Cluster analysis. https://www.nvidia.com/en-us/glossary/data-science/clustering/.

Pantalone, K. M., Hatipoglu, B., Gupta, M. K., Kennedy, L., and Hamrahian, A. H. (2015). Measurement of serum free thyroxine index may provide additional case detection compared to free thyroxine in the diagnosis of central hypothyroidism. *Case reports in endocrinology*, 2015.

Romero, C. M. and Almazán, E. M. (2015). Hipo e hipertiroidismo. *Tratado de Geriatría para residentes*, pages 605–614.

Sharma, P. (2019). Why is scaling required in knn and k-means? https://medium.com/analytics-vidhya/why-is-scaling-required-in-knn-and-k-means-8129e4d88ed7.

Soleymani, A. (s.f.). What is k-medoids clustering and when should you use it instead of k-means. https://medium.com/@ali.soleymani.co/beyond-scikit-learn-is-it-time-to-retire-k-means-and-use-this-method-instead-b8eb9ca9079a.

Universidad de Oviedo (s.f.). k-means algorithm applied to image classification and processing. https://www.unioviedo.es/compnum/labs/new/kmeans.html.

Villagra, A., Guzmán, A., Pandolfi, D., and Leguizamón, G. (2009). An´ alisis de medidas no-supervisadas de calidad en clusters obtenidos por k-means y particle swarm optimization. *Ciencia y Tecnología*.