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

Laboratorio 2 - Agrupamiento K-means

Integrantes:	Matías Escudero
	Joaquín Macías
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco
Ayudante:	Daniel Calderón

12 de Junio de 2023

Tabla de contenidos

1. Introducción

El algoritmo de K-means es un método de agrupamiento o clustering (Preeti Arora a, 2016) ampliamente utilizado en el campo del aprendizaje automático y la minería de datos. Su objetivo principal es agrupar un conjunto de datos en K clusters, donde K es un número predefinido.

En K-means, el centroide es un punto que representa el centro de un cluster. Es la media de todas las muestras que pertenecen a ese cluster en términos de sus características. El centroide se actualiza iterativamente durante el proceso de agrupamiento para minimizar la distancia entre los puntos de datos y el centroide asignado a su cluster.

A diferencia del centroide, el medoide es una muestra real del conjunto de datos que pertenece a un cluster y que mejor representa ese cluster. En lugar de ser la media de las características, el medoide es una muestra existente en el conjunto de datos.

A continuación, se abordarán los objetivos de este segundo laboratorio, que serán desde la aplicación de agrupamiento no jerárquico (clustering) al análisis de resultados.

1.1. Objetivos

1.1.1. Objetivo General

- Interpretar y aplicar el método de agrupamiento no jerárquico a un conjunto de datos relacionado a la enfermedad de hipertiroidismo.

1.1.2. Objetivos Específicos

1. Examinar las variables de nuestro conjunto de datos con el fin de verificar la necesidad de la estandarización o normalización.
2. Justificar la designación de clusters (cantidad) a utilizar, según un k mínimo de prueba.
3. Examinar los centroides y medoides obtenidos tras la aplicación del algoritmo K-means.
4. Designar las medidas de calidad para evaluar el método de agrupamiento.

2. Marco Teórico

2.1. Agrupamiento

Se refiere al proceso de asignar puntos de datos a diferentes clusters en función de su proximidad a los centroides. El objetivo es agrupar los puntos de datos de manera que los puntos dentro de cada cluster sean similares entre sí y que los clusters estén bien separados unos de otros. En esta etapa, se calcula la distancia entre cada punto de datos y los centroides de los clusters. El punto de datos se asigna al cluster cuyo centroide está más cercano en términos de distancia euclidiana.

2.2. K-means

Es una técnica de aprendizaje no supervisado que agrupa un conjunto de datos en k grupos o clusters basados en sus características y similitudes. El algoritmo busca minimizar la varianza intra-cluster y maximizar la varianza inter-cluster para obtener una separación óptima entre los grupos. Es importante tener en cuenta que K-means es un algoritmo heurístico y no garantiza encontrar la solución óptima global (Han, 2011).

2.2.1. Aplicaciones de agrupamiento

K-means tiene varias aplicaciones en diferentes campos, como la ciencia de datos, la inteligencia artificial, la minería de datos, la biología, la medicina y la investigación de mercado, entre otros. Algunas de las aplicaciones más comunes del algoritmo de K-means son: Segmentación de mercado, Análisis de imagen, Minería de datos, Detección de anomalías, Biología, Medicina y Sistemas de recomendación.

2.2.2. Intuición de agrupamiento de K-means

En el algoritmo de agrupamiento K-means, el centroide es un punto que representa el centro de cada cluster. Durante el proceso de agrupamiento, los datos se agrupan en clusters, y cada cluster se define por su centroide. El centroide es calculado como la media de todos los puntos que pertenecen a ese cluster.

El objetivo es minimizar la distancia entre cada punto y el centroide de su cluster asignado. Por lo tanto, la elección de los centroides iniciales puede afectar significativamente la calidad de los clusters finales y la eficiencia del algoritmo.

2.3. Medoide

El medoide es un concepto utilizado en el contexto del agrupamiento de datos para representar de manera efectiva un cluster. Es una muestra real del conjunto de datos que pertenece a un cluster y se considera como el punto más representativo de ese cluster (Leticia Laura Ochoa et al., 2017).

A diferencia del centroide, que se calcula como la media de las características de todas las muestras en un cluster, el medoide es una de las muestras existentes en el conjunto de datos. Representa la muestra que tiene la menor distancia promedio a todas las demás muestras del mismo cluster.

El cálculo del medoide implica encontrar la muestra que minimiza la suma de las distancias entre esa muestra y todas las demás muestras dentro del mismo cluster. En otras palabras, se busca la muestra que es más similar o cercana al resto de las muestras en términos de sus características.

2.4. Distancia de Manhattan

La medida de distancia Manhattan (Rodríguez, 2015), también conocida como distancia de la ciudad o distancia L1, es una métrica utilizada para medir la distancia entre dos puntos en un espacio euclidiano. A diferencia de la distancia euclidiana, que se calcula como la longitud de la línea recta entre dos puntos, la distancia Manhattan se calcula como la suma de las diferencias absolutas entre las coordenadas de los puntos a lo largo de cada dimensión.

3. Pre-procesamiento de datos

Para poder utilizar de la mejor manera el algoritmo de K-means, es recomendado poder utilizar nuestras variables en el tipo de codificación numérica continua. Las variables categóricas pueden ser problemáticas para el algoritmo, ya que no se pueden calcular fácilmente las distancias entre ellas. Por lo tanto, en general se recomienda que las variables categóricas sean convertidas a variables numéricas antes de aplicar el algoritmo de K-means (Amir Ahmad a, 2007). Según esto, debemos tener presente los siguientes aspectos:

- **Tipo de codificación**: existen diferentes técnicas de codificación de variables categóricas, como la codificación one-hot, la codificación ordinal y la codificación de frecuencia. Es importante seleccionar la técnica adecuada para el tipo de variable categórica y para el objetivo del análisis.
- **Tratamiento de valores faltantes**: es común que los datos categóricos tengan valores faltantes. En este caso, se deben decidir si se imputarán los valores faltantes o si se eliminarán las filas correspondientes.
- **Escala de las variables**: es importante escalar las variables numéricas para que tengan la misma importancia en el proceso de agrupamiento que las variables categóricas codificadas.
- **Interpretación de los resultados**: después de aplicar el algoritmo de K-means a los datos codificados, es importante interpretar los resultados de manera adecuada, teniendo en cuenta la naturaleza de las variables originales y el contexto del análisis.

3.1. Imputación de datos

En primera instancia, se observan los valores faltantes en aquellas variables en las que efectivamente existen NA's y así, entender qué decisión tomar sobre estas columnas. Este porcentaje se muestra a continuación en la Tabla 1:

Variable	% Valores Faltantes
age	0.04
sex	3.93
TT4	6.57
TSH	10.14
FTI	10.54
T4U	10.61
T3	20.89
TBG	100.00

Cuadro 1: Porcentaje de valores faltantes en variables de estudio.

A partir de estos resultados, se toman las siguientes decisiones sobre el set de datos:

- Se elimina el único registro con edad faltante.
- Se elimina la variable *TBG* ya que no posee información y, por lo tanto, también *TBG Measured*.

Ahora, queda saber qué hacer con las variables *sex*, *TT4*, *TSH*, *FTI*, *T4U* y *T3*. Para evitar utilizar imputación de datos realizada únicamente en base a promedios, medianas o modas. Primero se busca entender si existe un patrón en la forma en que los valores faltantes se distribuye a lo largo del set de datos. Para esto, se utiliza un gráfico *UpSet* para entender si es que los valores faltantes a lo largo de las variables de interés se relacionan entre sí. El gráfico se muestra a continuación en la Figura 1:

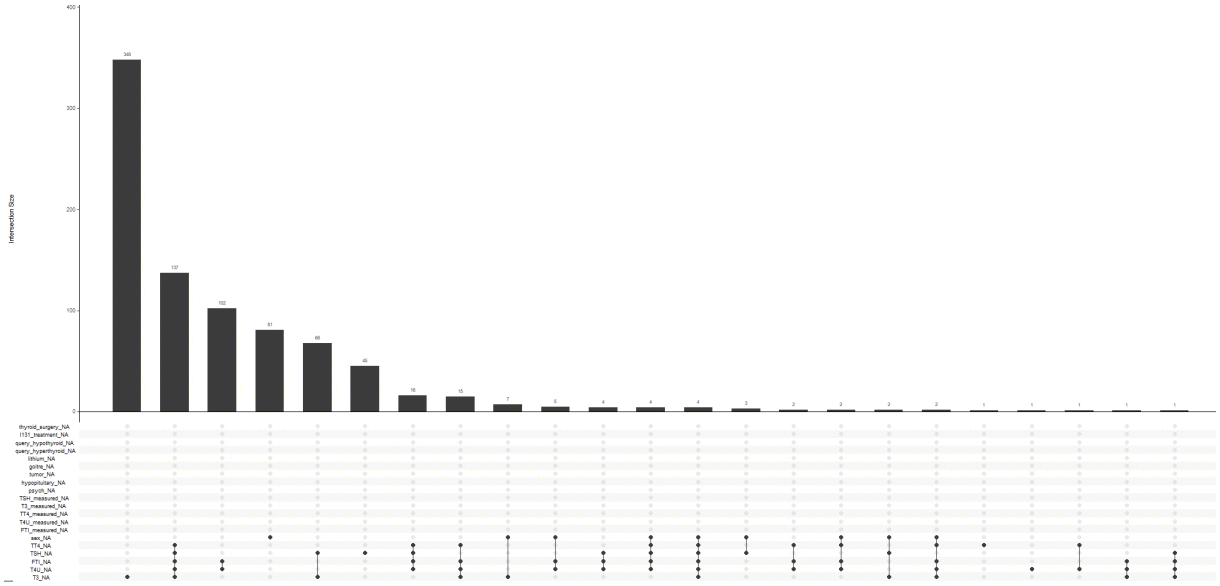

Figura 1: Gráfico UpSet para valores faltantes.

Esta figura muestra el número de intersecciones de valores faltantes en las distintas variables. Por ejemplo, la segunda barra muestra que en 137 registros existen **simultáneamente** valores faltantes en las variables *TT4*, *TSH*, *FTI*, *T4U* y *T3*. A partir de este gráfico, es posible notar que los datos faltantes de la variable *sex* no siguen un patrón claro respecto a los valores faltantes de otras variables. Además, el porcentaje de valores faltantes es relativamente bajo (3.9 %), **por lo cual los valores faltantes de la variable *sex* pasarán a ser el valor de la moda de esta**, que corresponde a 'F' (65 % de los registros).

Ahora, con el resto de variables numéricas (con valores faltantes), se busca imputar en los casos que sea posible, utilizando algún tipo de regresión. Para esto, en primera instancia se estudia el grado de correlación que existe entre estas variables y así, ver si es posible estimar valores faltantes en base a otra variable predictora. A continuación, se muestra la matriz de correlación en la Figura 2:

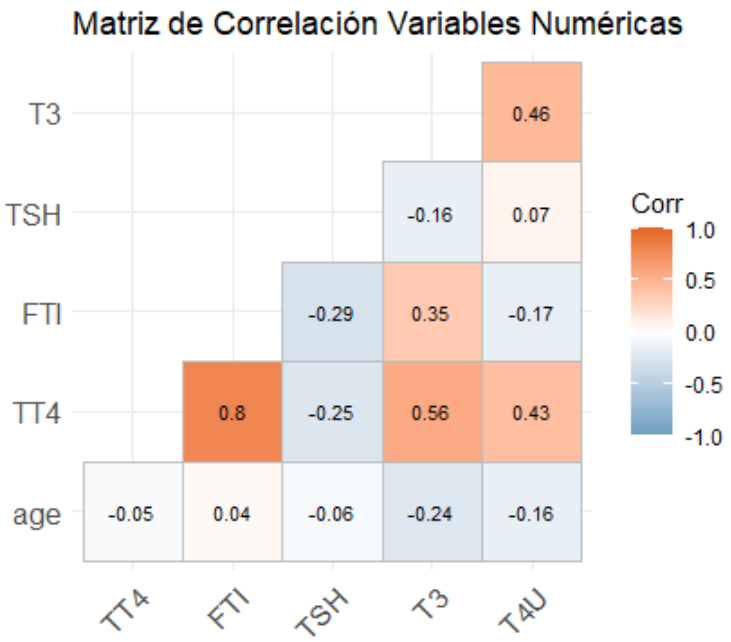

Figura 2: Matriz de correlación entre variables numéricas.

A partir de la matriz de correlaciones es posible notar que **(i)** *TT4* y *FTI* guardan una correlación bastante alta y **(ii)** *TT4* y *T3* una correlación razonable. Respecto a las demás variables no se observa ninguna otra correlación importante. De acuerdo a las correlaciones observadas, se utiliza el método de imputación *Predictive Mean Matching* (Tim P Morris, 2014) mediante la librería `mice` de R. Para cada una de las variables a imputar se utilizan cómo variables predictoras aquellas que guardan más relación con ellas. De esta forma, ahora el set de datos con el que se trabaja ya no tiene valores faltantes.

3.2. Codificación de variables categóricas

Ya que la función que se utilizará para realizar el agrupamiento mediante k-means recibe cómo entrada un set de datos únicamente con valores numéricos, se recodifican los valores de las variables categóricas y booleanas. De esta forma la variable *sex* cuyos valores eran 'M' y 'F' se transforman ahora a 1 y 0, respectivamente.

Por otra parte, la principal variable a reclasificar es la variable de clase *class* que indica la condición del paciente, la cual puede tomar los valores *negative*, *hyperthiroid*, *goitre* o

T3 toxic. Cómo se busca tranformar esta variable categórica a variable numérica y se está estudiando el fenómeno de hipertiroidismo hay que decir que hacer con los valores *goitre* y *T3 Toxic*. Por una parte, de acuerdo a revisión de literatura, *T3 Toxic* hace referencia a la condición llamada toxicosis por T3 que es un sub-tipo de hipertiroidismo (Burch and Cooper (2015)), por lo cual **se reclasifican estos casos *T3 toxic* a *hyperthiroid***. Por otra parte, el valor *goitre* hace referencia a la condición del bocio, que se refiere a una hinchazón de la glándula tiroides, la cual se puede deber a muchas condiciones diferentes (Vanderpump (2011)), no particularmente hipertiroidismo, por lo cual **se decide eliminar los 7 registros que contienen este valor en la variable de clase**.

Finalmente, de acuerdo a lo visto en el laboratorio anterior, la variable *referral source* contiene información irrelevante para el caso de estudio, por lo cual se decide eliminarla, mientras que el resto de variables booleanas, se transforman los valores 'True' en 1 y 'False' en 0.

3.3. Normalización/Estandarización

Consiste en ajustar las características para que tengan una escala comparable o una distribución estándar, lo que puede ser útil en varios algoritmos de aprendizaje automático (Chaganti et al., 2022).

Para nuestro caso, implica aplicar alguna técnica de normalización a las características numéricas del conjunto de datos. Esto se hace para asegurarse de que todas las características tengan una escala similar y no dominen el análisis o el modelado debido a sus diferentes rangos o magnitudes. Podrías aplicar las siguientes técnicas:

Normalización Min-Max: Ajusta los valores de las características dentro de un rango específico, generalmente entre 0 y 1.

Estandarización (Z-score): Transforma los valores de las características para que tengan una media de 0 y una desviación estándar de 1.

La elección entre normalización Min-Max y estandarización (Z-score) depende de la distribución y la escala de tus características. Si tus características tienen valores atípicos o una

distribución no normal, la estandarización puede ser más apropiada. Si deseas mantener los rangos originales de las características, puedes optar por la normalización Min-Max.

Para este caso, utilizaremos la normalización Min-Max, con el fin de mantener los rangos originales de las características (Kirubha et al., 2019). El proceso a aplicar será el siguiente:

- Reconocer las características que son numéricas y necesitan ser normalizadas. Puedes hacer esto revisando la descripción de las características en el conjunto de datos o explorando su tipo de datos.
- Calcular el valor mínimo $Xmin$ y el valor máximo $Xmax$ para cada característica numérica que deseas normalizar.
- Aplicar la Fórmula de normalización Min-Max.
- Repetir el proceso para todas las características numéricas. Puedes utilizar bucles o funciones de mapeo para aplicar la fórmula a cada valor en cada característica.
- Reemplazar los valores originales en el conjunto de datos con los valores normalizados para cada característica numérica. Esto asegurará que el conjunto de datos esté completamente normalizado en el rango específico.

Ya habiendo normalizado todas las variables numéricas, se procederá a ejecutar los algoritmos de agrupamiento. Cabe destacar que una vez que ya se realice el agrupamiento por k-means se vuelve a utilizar las variables sin normalizar para facilitar la interpretación de los grupos.

4. Obtención de clústers

4.1. Justificación de distancia utilizada

Entendiendo que algoritmo k-means puede utilizar distintos tipos de distancia cómo valor de entrada, es importante definir qué distancia será utilizada de acuerdo al conjunto de datos con el que se trabaja.

La elección de la **distancia de Manhattan** como medida de distancia, según nuestro conjunto de datos, depende de las características y la naturaleza de los datos, por ello, se establecen las siguientes razones por lo que podría ser apropiado utilizarla (Aggarwal et al., 2001):

- El conjunto de datos contiene variables categóricas que representan características como el *sex*, la fuente de derivación y la clase binaria. La distancia de Manhattan es adecuada para medir la similitud o la diferencia entre variables categóricas, ya que se basa en la suma de las diferencias absolutas entre las categorías. Esto permite tratar las variables categóricas de manera adecuada en el cálculo de la distancia entre los puntos de datos.
- Es probable que las variables numéricas tengan diferentes escalas y rangos. La distancia de Manhattan es útil en este caso porque considera las diferencias absolutas entre las coordenadas de los puntos en cada dimensión, sin verse afectada por las diferencias de escala. Esto asegura que las variables con rangos muy diferentes no dominen la medida de distancia.
- La distancia de Manhattan mide la distancia a lo largo de las calles de una cuadrícula ortogonal. Esta métrica puede ser relevante en el contexto del conjunto de datos si se desea dar más importancia a los movimientos en líneas rectas ortogonales y se desea penalizar las diferencias en cada dimensión de manera igualmente importante.
- La distancia de Manhattan es más interpretable que otras medidas de distancia, como la distancia euclidiana, ya que se basa en las diferencias absolutas. Esto facilita la interpretación y la comparación de las distancias entre los puntos de datos. Además, la distancia de Manhattan también puede ofrecer una mejor interpretación visual cuando se representan los datos en un espacio bidimensional o tridimensional.

4.2. Valor de K

El algoritmo K-means depende de encontrar la cantidad de grupos y etiquetas de datos para un valor predefinido de K. Para encontrar la cantidad de grupos en los datos, necesitamos ejecutar el algoritmo de agrupación en clústeres K-means para diferentes valores de K y comparar los resultados. Entonces, el rendimiento del algoritmo K-means depende del valor de K. Debemos elegir el valor óptimo de K que nos brinde el mejor rendimiento. Hay diferentes técnicas disponibles para encontrar el valor óptimo de K. La técnica más común es el método del codo que se describe a continuación.

El método del codo (en inglés *elbow method*) es una técnica utilizada para determinar el número óptimo de clusters en el algoritmo de agrupamiento K-means. Este método se basa en el análisis de la variabilidad explicada por el modelo en función del número de clusters (Edy Umargono, 2000).

La idea es trazar un gráfico de la variación explicada por el modelo (por ejemplo, la suma de los cuadrados de las distancias intra-cluster) en función del número de clusters, y encontrar el 'codo' en la curva, es decir, el punto donde la adición de un cluster adicional no mejora significativamente la variación explicada por el modelo.

El método del codo es una técnica gráfica que proporciona una guía útil para la selección del número óptimo de clusters en el algoritmo de K-means. Sin embargo, es importante tener en cuenta que esta técnica es subjetiva y que la elección final del número de clusters debe basarse en una combinación de técnicas objetivas y subjetivas, y en el conocimiento del dominio del problema.

A continuación se muestra el gráfico del método del codo generado para el conjunto de datos ya pre-procesados de acuerdo a la sección anterior, utilizando distancia de Manhattan:

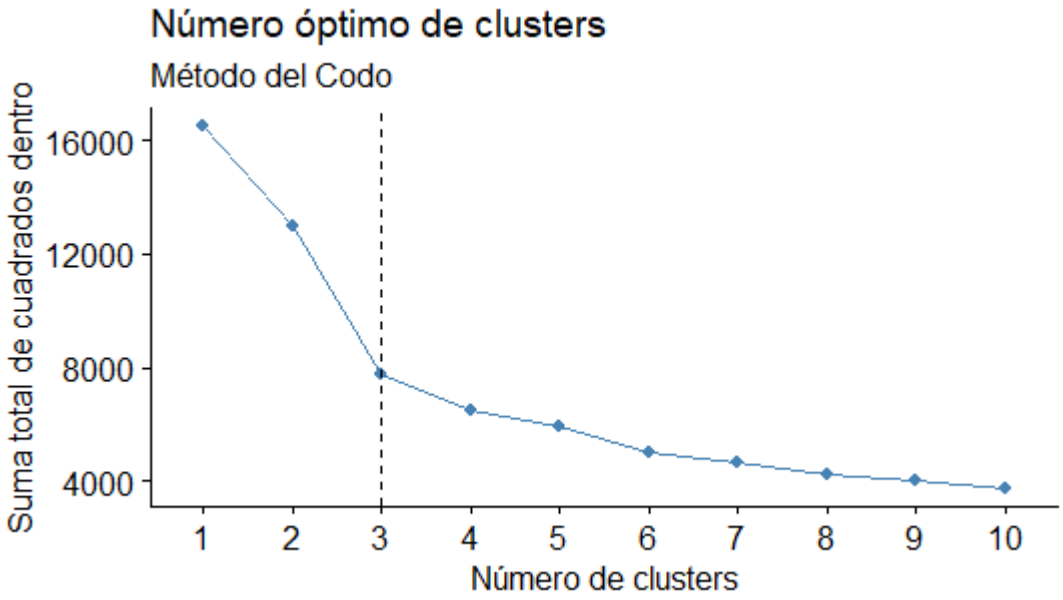

Figura 3: Número óptimo de clústers obtenido mediante método del codo.

A partir de la Figura 3 anterior es posible visualizar que el 'codo' se muestra al realizar 3 clústers. Esto significa que al generar 3 clústers se minimiza la suma de las distancias cuadradas dentro de cada clúster y maximiza la distancia entre diferentes clústers, lo que indica que **3 es el número óptimo de grupos a generar para el presente caso de estudio**. Esto indica que el generar un cuarto clúster no reduciría la variabilidad dentro del clúster de manera significativa, en comparación con la reducción obtenida al pasar de dos a tres clústers.

5. Análisis de Resultados

Ya entendiendo el tipo de distancia a utilizar y el número de k óptimo, se procede a ejecutar el algoritmo de agrupación utilizando $k = 3$ y posterior a eso con $k = 7$ para luego contrastar los resultados y demostrar por qué al utilizar $k = 3$ se estaría generando un mejor agrupamiento.

5.1. Agrupamiento con k = 3

En **primera instancia se genera agrupamiento con k=3**, para lo cual se busca entender cómo está compuesto cada uno de los clústers generados. A continuación se muestra el tamaño de cada uno de ellos:

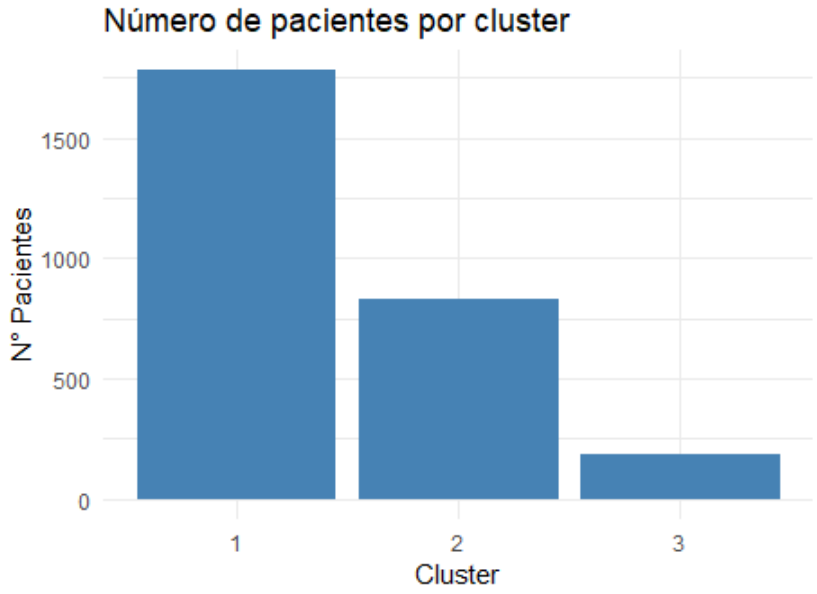

Figura 4: Número de pacientes por clúster.

Es posible notar que el clúster con más pacientes asignados corresponde al clúster 1 con un 63,7 % de los pacientes, seguido del clúster 2 con un 29,6 % de los pacientes y finalmente el clúster 3 con un 6,7 % de los pacientes. Ya conociendo el tamaño de cada uno de los grupos generados, para cada uno de estos se estudia el promedio y desviación estándar de las variables numéricas del conjunto de datos, además de las estadísticas asociadas a las variables categóricas:

Variable	Cluster 1	Cluster 2	Cluster 3
Age	52,22 (19.16)	51,80 (18.23)	46,66 (19.71)
TSH	5,15 (24.08)	3,40 (12.39)	2,78 (5.53)
T3	2,07 (0.88)	1,93 (0.76)	2,06 (0.79)
TT4	113,00 (37.79)	100,57 (27.94)	108,81 (33.92)
T4U	1,03 (0.20)	0,93 (0.15)	1,00 (0.19)
FTI	111,42 (34.85)	108,74 (27.94)	111,26 (29.57)

Cuadro 2: Estadísticas asociadas a cada clúster para variables numéricas.

Variable	**Cluster 1**		**Cluster 2**		**Cluster 3**	
	0	**1**	**0**	**1**	**0**	**1**
sex	1778	0	0	828	154	32
on_thyroxine	1527	251	768	60	168	18
query_on_thyroxine	1757	21	809	19	186	0
on_antithyroid_medication	1754	24	823	5	181	5
sick	1709	69	794	34	179	7
pregnant	1740	38	828	0	186	0
thyroid_surgery	1745	33	822	6	186	0
I131_treatment	1741	37	818	10	186	0
query_hypothyroid	1661	117	790	38	178	8
query_hyperthyroid	1654	124	799	29	167	19
lithium	1766	12	826	2	186	0
goitre	1765	13	817	11	185	1
tumor	1728	50	821	7	179	7
hypopituitary	1778	0	827	1	186	0
psych	1717	61	754	74	186	0
TSH_measured	73	1705	46	782	165	21
T3_measured	320	1458	105	723	160	26
TT4_measured	1	1777	2	826	181	5
T4U_measured	71	1707	40	788	186	0
FTI_measured	70	1708	39	789	186	0
hyperthiroid	1716	62	820	8	186	0

Cuadro 3: Estadísticas asociadas a cada clúster para variables booleanas.

A partir de la tablas anteriores es posible inferir que:

- Los pacientes del **clúster 1** está compuesto únicamente por mujeres y tienen, en promedio, una edad más alta en comparación con los otros clusters. El nivel promedio de TSH (hormona estimulante de la tiroides) es significativamente más alto en este grupo, además de los valores de FTI, T4U, TT4 y T3. De acuerdo a la variable de clase, este

es el grupo con más pacientes con hipertiroidismo de todos.

- Los pacientes del **clúster 2** está compuesto únicamente por hombres, los cuales tienen el nivel promedio de T3 más bajo en comparación con los otros clusters. Además, la media de la hormona T4U (tiroxina no unida) también es más baja en este grupo.
- Los pacientes del **clúster 3** son los más jóvenes en promedio, en comparación con los otros clusters. Además, los niveles promedio de TSH son los más bajos en este grupo. No hay gente con hipertiroidismo en este grupo.

En complemento a las estadísticas revisadas, se visualizan las matrices de correlación y gráficos de dispersión para variables numéricas en los distintos clústers generados:

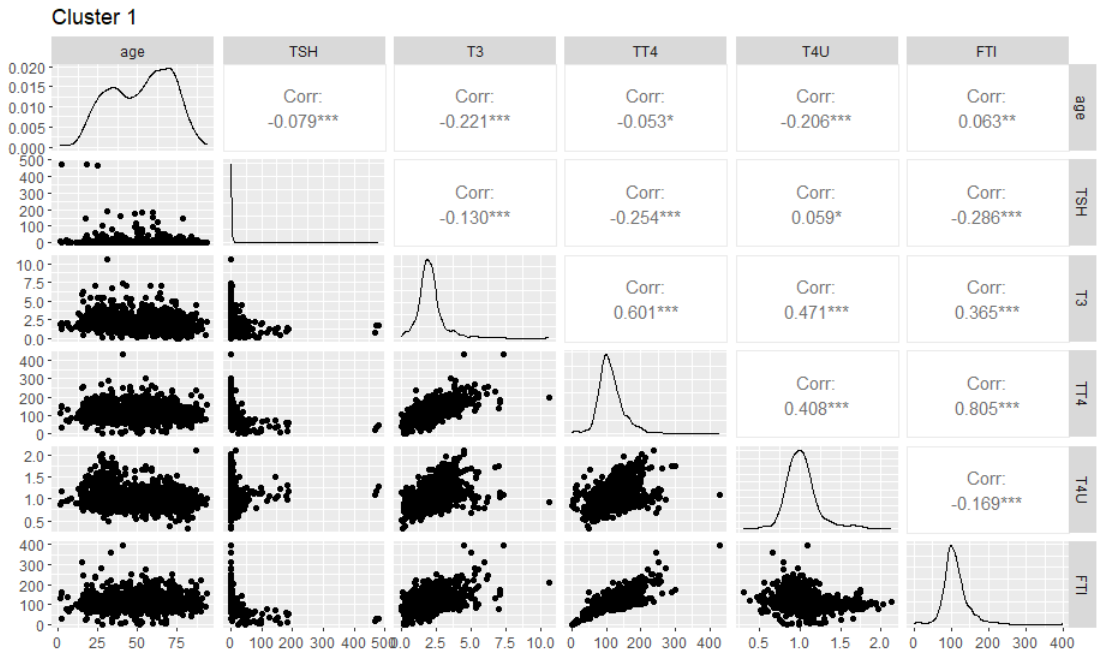

Figura 5: Matriz de correlación para variables numéricas Cluster 1

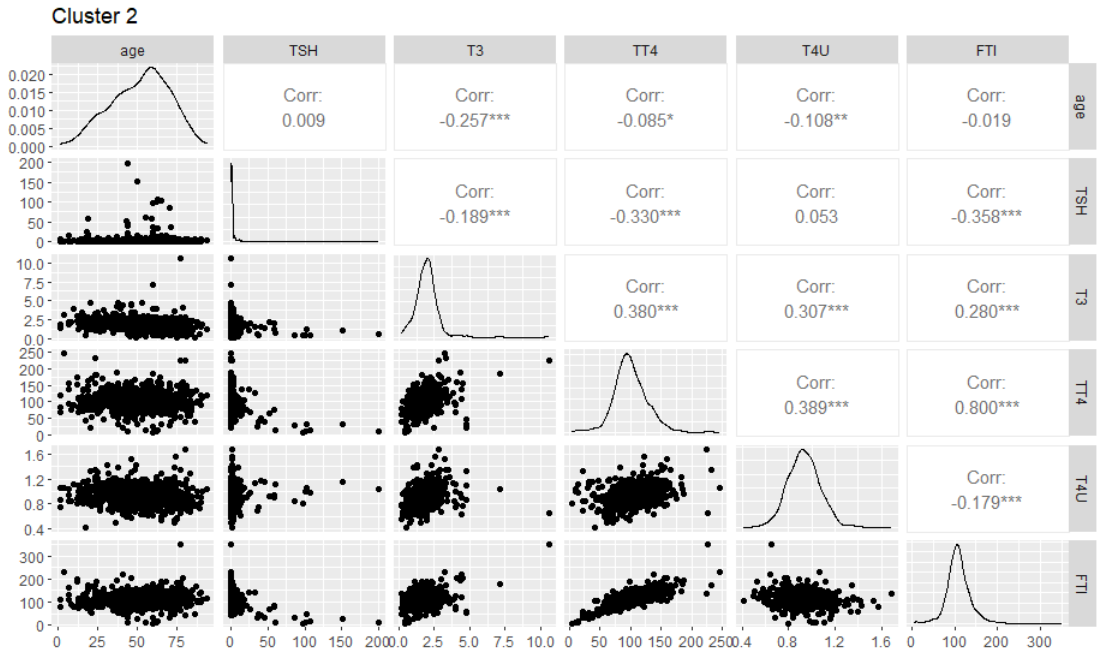

Figura 6: Matriz de correlación para variables numéricas Cluster 2

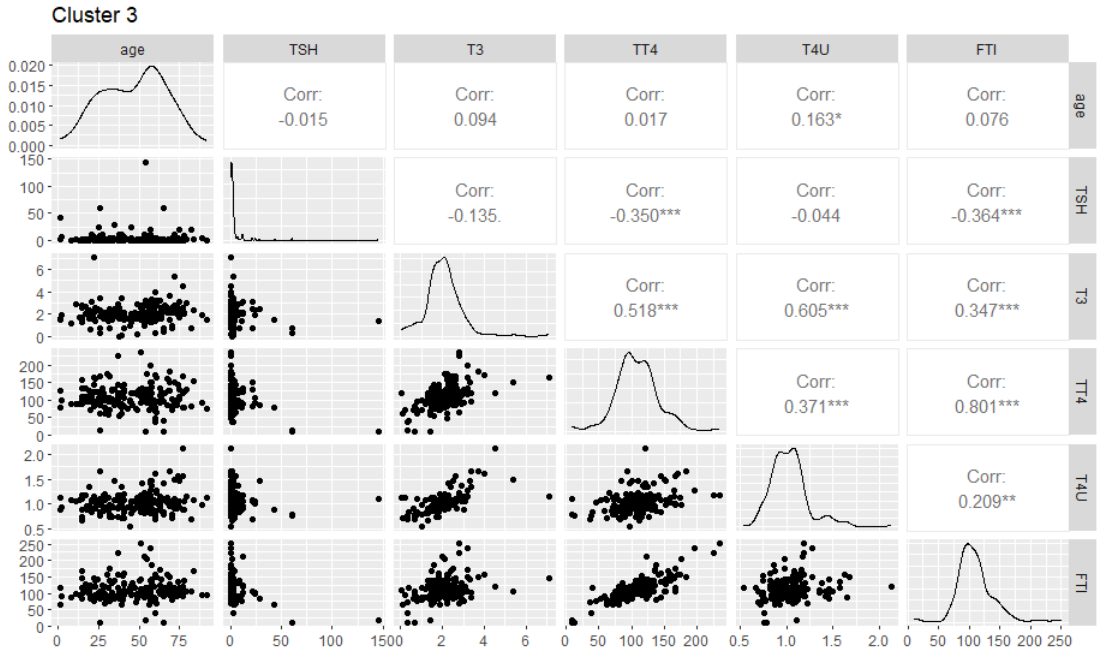

Figura 7: Matriz de correlación para variables numéricas Cluster 3

A partir de las distintas matrices de correlación no se observan grandes diferencias entre las correlaciones en los distintos cluústers. Sin embargo, hay un par de variaciones relevantes:

- La correlación entre T4U y T3 es notablemente más alta para el clúster 3 (0,605). Esto indica que estos niveles están más relacionados entre sí para este clúster.
- En el clúster 3 también existe una tendencia distinta entre la edad y los niveles de T3, en comparación a los otros dos grupos.

A modo de resumen los clústers obtenidos bajo los parámetros mencionados anteriormente se caracterizan por:

- **Clúster 1**: Grupo de mujeres que presentan una edad media mayor que la de los otros grupos. Este grupo presenta los valores más altos de las hormonas TSH, FTI, T4U, TT4 y T3. Los niveles altos de estas 3 últimas hormonas están asociadas a hipertiroidismo, **lo que indica que este cluster podría representar a mujeres mayores con hipertiroidismo más avanzado o más severo**.
- **Clúster 2**: Grupo de hombres con niveles de T3 y T4U más bajo. La correlación entre TSH y FTI es fuerte en este grupo (-0.36), y esta tendencia inversa podría indicar una respuesta compensatoria del cuerpo al bajo nivel de hormonas tiroideas (Hoermann et al. (2015)). Dado que los niveles de T3 y T4U son bajos en este grupo, podríamos inferir que estos **hombres pueden estar en las primeras etapas del hipertiroidismo o tener una forma más leve de la enfermedad**.
- **Clúster 3**: Este grupo es el más joven en promedio y tiene los niveles promedio más bajos de TSH. No hay pacientes con hipertiroidismo en este grupo. El hecho de que la correlación entre T4U y T3 sea notablemente alta en este grupo podría indicar que su función tiroidea está regulada de manera diferente o más eficaz en comparación con los otros dos grupos Biondi and Wartofsky (2014). **Sería el grupo más joven y más sano**.

Se logra identificar que los clústers se dividieron a lo largo de las líneas de género. Esto hace mucho sentido de acuerdo a la regresión logística hecha en la experiencia pasa y en base a bibliografía, ya que está ampliamente documentado que las mujeres tienden a presentar mayores tasas de hipertiroidismo (Dunn and Turner (2016)). Sin embargo, llama la atención

que los niveles de TSH sean los más altos dentro del grupo 1 de mujeres con hipertensión, ya que está documentado que los niveles bajos de TSH están asociados a hipertiroidismo (Magner (1993)). Esto demuestra que el para el agrupamiento otras variables tuvieron una mayor incidencia sobre la forma en que se seleccionaron los clústers.

5.2. Agrupamiento con k = 7

En segunda instancia **se utiliza k=7** y se estudian las estadísticas descriptivas asociadas a cada clúster generado y de esta forma lograr entender la composición de cada uno de estos:

Figura 8: Número de pacientes por clúster para k = 7.

Es posible notar que clúster con más pacientes termina siendo el clúster 1 (40,54 %), seguido del clúster 3 (29,05 %) y finalmente, el clúster 5 (9,42 %). Ya conociendo el tamaño de cada uno de los grupos generados (7), se estudiará el promedio y desviación estándar de las variables numéricas del conjunto de datos, además las estadísticas asociadas a las variables categóricas:

Variable	C1	C2	C3	C4	C5	C6	C7
Age	53,26 (19.55)	43,72 (19.23)	52,30 (18.00)	52,50 (16.71)	50,44 (19.54)	48,17 (17.91)	46,69 (19.67)
TSH	5,38 (25.07)	4,41 (10.35)	3,50 (12.73)	3,84 (15.70)	6,36 (30.91)	2,66 (10.31)	4,31 (13.07)
T3	1,98 (0.78)	2,05 (0.81)	1,93 (0.75)	2,33 (1.03)	2,07 (0.91)	2,78 (1.28)	2,03 (0.85)
TT4	107,44 (33.68)	101,21 (32.64)	100,66 (27.51)	140,12 (38.02)	107,34 (28.18)	140,00 (61.45)	108,60 (33.64)
T4U	1,02 (0.20)	1,00 (0.16)	0,93 (0.15)	1,07 (0.20)	1,00 (0.17)	1,11 (0.25)	1,03 (0.21)
FTI	106,74 (30.47)	104,48 (29.72)	109,06 (27.91)	131,89 (32.47)	108,22 (27.77)	131,72 (64.47)	110,39 (32.72)

Cuadro 4: Estadísticas asociadas a cada clúster para variables numéricas con k = 7.

Variable	**C1**		**C2**		**C3**		**C4**		**C5**		**C6**		**C7**	
	0	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
sex	1132	0	70	16	0	811	201	0	263	0	112	0	154	33
on_thyroxine	1132	0	77	9	751	60	0	201	227	36	107	5	169	18
query_on_thyroxine	1122	10	85	1	793	18	198	3	255	8	112	0	187	0
on_antithyroid_medication	1117	15	84	2	806	5	199	2	263	0	107	5	182	5
sick	1076	56	86	0	777	34	197	4	256	7	110	2	180	7
pregnant	1114	18	86	0	811	0	197	4	260	3	99	13	187	0
thyroid_surgery	1115	17	82	4	805	6	197	4	258	5	109	3	187	0
I131_treatment	1114	18	84	2	801	10	196	5	255	8	108	4	187	0
query_hypothyroid	1077	55	81	5	776	35	181	20	228	35	107	5	179	8
query_hyperthyroid	1132	0	82	4	783	28	197	4	258	5	0	112	168	19
lithium	1126	6	86	0	809	2	199	2	261	2	110	2	187	0
goitre	1124	8	86	0	800	11	200	1	259	4	112	0	186	1
tumor	1100	32	85	1	804	7	199	2	254	9	106	6	180	7
hypopituitary	1132	0	86	0	810	1	201	0	263	0	112	0	187	0
psych	1074	58	84	2	738	73	200	1	262	1	112	0	187	0
TSH_measured	24	1108	0	86	46	765	3	198	40	223	6	106	165	22
T3_measured	0	1132	0	86	105	706	52	149	263	0	5	107	160	27
TT4_measured	1	1131	86	0	23	778	0	201	0	263	0	112	187	0
T4U_measured	1	1131	86	0	23	788	0	201	0	263	0	112	187	0
FTI_measured	0	1132	86	0	22	789	0	201	0	263	0	112	187	0
hyperthyroid	1099	33	85	1	804	7	199	2	255	8	93	19	187	0

Cuadro 5: Estadísticas asociadas a cada clúster para variables booleanas con k = 7.

A partir de las tablas anteriores es posible inferir que:

- **Cluster 1**: Este grupo está compuesto únicamente por pacientes mujeres ya mayores con una edad media de 53.26 años. Este grupo también tiene uno de los valores medios más altos de TSH (5.38).
- **Cluster 2**: Este grupo es notablemente más joven que el resto, con una edad media de 43.72 años. Aunque este grupo tiene un valor de TSH ligeramente menor que el Cluster 1, los niveles de TSH siguen siendo relativamente altos.
- **Cluster 3**: Este grupo son sólo hombres que tienen una edad media similar a los Clusters 1 y 4, y sus niveles de TSH son los más bajos, lo que podría indicar una tiroides más saludable o una tendencia hacia el hipertiroidismo. Además, los niveles medios de TT4, T4U y FTI son ligeramente más bajos que en otros grupos.
- **Cluster 4**: Este grupo tiene sólo mujeres que presentan la mayor media de TT4, T4U y FTI, lo que podría indicar que la tiroides está produciendo demasiada hormona tiroidea. La edad media de este grupo es similar a la del Cluster 1.
- **Cluster 5**: Este grupo tiene una media de TSH mayor que la mayoría de los otros grupos, lo que podría sugerir una tendencia hacia el hipotiroidismo. En cuanto a las otras variables numéricas, este grupo se parece mucho a los Clusters 1 y 2.
- **Cluster 6**: Este grupo es notable por tener el valor medio más alto de T3 y los valores más altos de TT4 y FTI, lo que podría indicar una fuerte tendencia hacia el hipertiroidismo. La edad media de este grupo es menor que la del Cluster 1 pero similar a la del Cluster 7.
- **Cluster 7**: Este grupo tiene una edad media de 46.69 años, la más baja después del Cluster 2. Sus niveles medios de TT4, T4U y FTI son similares a los de los Clusters 1, 2 y 5, mientras que su nivel medio de TSH es similar al del Cluster 2.

A partir de lo anterior, es posible notar que los Clusters 1, 5 y 7 parecen ser bastante similares entre sí en términos de sus características numéricas, así como el Cluster 2 y 7 en términos

de la edad media. Los Clusters 4 y 6 son notablemente diferentes de los demás debido a sus altos niveles de TT4, T4U y FTI. Si bien los grupos generados logran agrupar algunas características que no se lograban con k = 3, **pareciera que se forman grupos demasiado similares entre sí, y que quizá no es necesario desagregar hasta este punto**.

6. Conclusiones

Recordando el laboratorio 1, el modelo parece ser adecuado para predecir la presencia de hipertiroidismo utilizando las variables seleccionadas por el método *Stepwise*. La desviación residual **132.50** del modelo es menor que su desviación nula **419.03**, lo cual indica que el modelo con las variables incluidas explica una mayor variabilidad en los datos que el modelo sin ellas.

A partir del modelo final se demostró que las variables T3, T4U, TSH, TT4 y sexo de un paciente son significativas sobre la predicción del hipertiroidismo.

Sobre la evaluación de la cantidad óptima de clusters, se recomienda el uso del método del codo para determinar la cantidad óptima de clusters en función de la variabilidad dentro de los datos y la separación entre los clusters. Se comprobó empíricamente por qué realizar 3 clústers (número óptimo) puede resultar mejor que otro valor de k.

Antes de aplicar K-means, es importante realizar un preprocesamiento adecuado de los datos. Esto puede incluir la normalización de las características numéricas y la codificación adecuada de las características categóricas. El preprocesamiento garantiza que todas las características tengan una escala comparable y una representación numérica adecuada, por lo que el uso de la normalización MinMax.

De acuerdo a la interpretación de los resultados, una vez que se haya aplicado K-means y obtenido los clusters, es importante representar y comprender los resultados. Examinar las características de los clusters y realizar un análisis descriptivos para comprender las características comunes y las diferencias entre los clusters, ayudará a identificar patrones y obtener información sobre los subgrupos presentes en el Thyroid Disease Dataset, tras identificar grupos distintos con características tiroideas diferentes. El análisis permite identificar patrones y características específicas en cada grupo, lo que puede ser útil para comprender la prevalencia de diferentes condiciones tiroideas y ayudar en la toma de decisiones clínicas y tratamientos adecuados.

Por otra parte, según la iteración y ajuste, es posible que se deba iterar y ajustar los parámetros del algoritmo de K-means, como el número de clusters, el número de iteraciones o los

criterios de convergencia, para obtener resultados óptimos. Sería interesante probar distintos parámetros de entrada para el algoritmo y evaluar cómo varían los resultados obtenidos.

Además, se observó que la agrupación funciona mejor para el número de clústers sugerida por el método del codo. Sin embargo, esto hasta cierto punto puede ser subjetivo, ya que dependiendo del contexto uno puede buscar un número de grupos distinto al óptimo.

Finalmente, contrastando los resultados obtenidos en esta experiencia versus los obtenidos por la regresión logística se obtienen conclusiones relevantes. Si bien en líneas generales los resultados hacen sentido, los objetivos de estas distintas técnicas son distintas: mientras que una regresión logística entrega un modelo que puede predecir la probabilidad de hipertiroidismo basado en un conjunto de características, el algoritmo de agrupamiento utilizado en esta experiencia busca agrupar a los pacientes en base a sus características clínicas y demográficas. En este sentido, si bien ya se tenía una idea sobre qué variables inciden más sobre el hipertiroidismo el utilizar técnicas de agrupamiento nos permitió entender mejor el conjunto de datos sobre el cual se trabaja y se revelaron patrones que la regresión logística pasa por alto. El elegir una técnica sobre otra dependerá del contexto en el que se está trabajando o la pregunta de investigación que se busca responder.

Bibliografía

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space.

Amir Ahmad a, L. D. b. (2007). A k-mean clustering algorithm for mixed numeric and categorical data.

Biondi, B. and Wartofsky, L. (2014). Treatment with thyroid hormone. *Endocrine reviews*, 35(3):433–512.

Burch, H. B. and Cooper, D. S. (2015). Management of graves disease: a review. *Jama*, 314(23):2544–2554.

Chaganti, R., Rustam, F., De La Torre Díez, I., and Vidal Mazón, J. L. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques.

Dunn, D. and Turner, C. (2016). Hypothyroidism in women. *Nursing for women's health*, 20(1):93–98.

Edy Umargono, Jatmiko Endro Suseno, S. V. G. (2000). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula.

Han, X. J. . J. (2011). K-Means Clustering.

Hoermann, R., Midgley, J. E., Larisch, R., and Dietrich, J. W. (2015). Homeostatic control of the thyroid–pituitary axis: perspectives for diagnosis and treatment. *Frontiers in endocrinology*, 6:177.

Kirubha, M., Prinitha, R., Preethika, P., and Samyuktha, A. (2019). Analysis of Thyroid Disease Using K Means and Fuzzy C Means Algorithm.

Leticia Laura Ochoa, M., Karina Rosas Paredes, M., and José Esquicha Tejada, M. (2017). Estudio Comparativo de Técnicas no Supervisadas de Minería de Datos para Segmentación de Alumnos.

Magner, J. A. (1993). Tsh-mediated hyperthyroidism. *The Endocrinologist*, 3(4):289–296.

Preeti Arora a, Deepali Dr. b, S. V. (2016). Analysis of K-Means and K-Medoids Algorithm For Big Data.

Rodríguez, M. A. (2015). Comparación de métricas de distancia en el algoritmo K-Vecinos Más Cercanos para el problema de Reconocimiento Automático de Dígitos Manuscritos, INFORME DEL PROYECTO DE TESIS.

Tim P Morris, I. R. W. . P. R. (2014). Tuning multiple imputation by predictive mean matching and local residual draws.

Vanderpump, M. P. (2011). The epidemiology of thyroid disease. *British medical bulletin*, 99(1).