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

ANÁLISIS DE DATOS

LABORATORIO 2: AGRUPAMIENTO K-MEDIAS

Autores:
Nicolás Mariángel Toledo
Juan Pablo Rojas Rojas
Profesor:
Max Chacón Pacheco
Ayudante:
Ignacio Ibáñez Aliaga

Santiago - Chile
27 de noviembre de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

El cuerpo humano está conformado por distintos sistemas que hacen de sí un ente completo y complejo con un organismo en equilibrio. Uno de estos sistemas es el *Sistema Endocrino* que tiene la función principal de regular la producción de hormonas, las cuales se encargan de influir en la mayoría de las funciones del organismo, por ejemplo: funcionamiento de órganos, control de las distintas funciones del organismo, auto-regulación, comportamientos del individuo, control de homeostasis, entre otras funciones. Así el sistema endocrino posee órganos que ayudan a formar el sistema como tal, uno es estos órganos es la llamada *Tiroides*, la cual es la encargada de producir hormonas tiroideas que ayudan al cuerpo utilizar energía, mantener la temperatura corporal y a que el cerebro, el corazón, los músculos y otros órganos funcionen normalmente [1].

No siempre nuestros sistemas y órganos funcionan como corresponde, en el caso de la Tiroides uno de los problemas más comunes son el *Hipotiroidismo* e *Hipertiroidismo* que tiene que ver principalmente con el desequilibrio de producción de *hormonas tiroideas*, en esta oportunidad el tema de interés se centra en el Hipotiroidismo, que se ocurre cuando la tiroides no es capaz de producir suficientes hormonas tiroideas T3 y T4. Esto último es un tema con gran potencial de estudio, ya que la población actual gradualmente ha incrementado este tipo de problemas, por lo cual se busca tener medicina suficientemente apta para dar soluciones eficientes a estos problemas, para esto se tiene evidencia de distintos datos de pacientes que nos pueden permitir determinar patrones de comportamientos y causas de tal enfermedad, lo que puede concluir en una gran aporte a la medicina para determinar causas de estos fallos en el organismo.

1.2 OBJETIVOS

Extraer el conocimiento del problema asignado, mediante el uso del software R, utilizando el algoritmo de *clustering* K-means (pam) y realizar el análisis respectivo.

Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.

Analizar por grupo e identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

1.3 ORGANIZACIÓN DEL DOCUMENTO

Este documento consta de 4 secciones principales: Marco teórico, donde se entregan definiciones de conceptos necesario para comprender este documento. Pre-procesamiento, obtención de los clústeres, seguido por un capítulo dedicado al análisis de estos. Para finalizar con una conclusión sobre el trabajo realizado.

CAPÍTULO 2. MARCO TEÓRICO

2.1. Clustering

El *Clustering* es el proceso de agrupar datos en clases o clústeres de tal forma que los objetos de un clúster tengan una similaridad alta entre ellos, y baja (sean muy diferentes) con objetos de otros clústeres [2].

El *clustering* es una técnica utilizada en la minería de datos que clasifica los objetos en grupos a partir de los atributos de estos, por ello existen distintos algoritmos para establecer los grupos de modo que la clasificación cumpla la característica más importante: **similitud intra clúster**

Los objetivos más importantes que el *clustering* es tratar de cumplir:

- Capacidad de manejar diferentes tipos de atributos: Numéricos (lo más común), binarios, nominales, ordinales, etc.
- Manejo de ruido: Muchos son sensibles a datos erróneos.
- Que los clústeres sean interpretables y utilizables, como también la capacidad de añadir restricciones [2].

2.2. Algoritmo K-medias

K-means es un algoritmo que agrupa objetos en k grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o clúster. Se suele usar la distancia cuadrática.

El algoritmo consta de tres pasos:

- **Inicialización**: una vez escogido el número de grupos, k, se establecen k centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
- **Asignación objetos a los centroides**: cada objeto de los datos es asignado a su centroide más cercano.
- **Actualización centroides**: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo k-means resuelve un problema de optimización, siendo la función por optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su clúster [3].

2.3. Distancias a utilizar

2.3.1. Distancia de Gower

Se dispone de un conjunto de datos mixto, es decir, un conjunto de individuos sobre los que se han observado tanto variables cuantitativas como cualitativas (o categóricas). Se define la distancia de Gower como $d_{ij}^2 = 1 - s_{ij}$, donde:

$$s_{ij} = \frac{\sum_{p_1}^{h=1}(1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p_1 + (p_2 - d) + p_3} \tag{2.1}$$

es el coeficiente de similaridad de Gower,
p_1 es el número de variables cuantitativas continuas,
p_2 es el número de variables binarias,
p_3 es el número de variables cualitativas(no binarias),
a es el número de coincidencias (1, 1) en las variables binarias,
d es el número de coincidencias (0, 0) en las variables binarias,
α es el número de coincidencias en las variables cualitativas (no binarias) y
G_h es el rango (o recorrido) de la h-ésima variable cuantitativa.[4]

2.3.2. Método de Silueta

El coeficiente de Silueta es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de *clustering*. El objetivo de Silueta es identificar cuál es el número óptimo de agrupamientos [5]. En el método de silueta se obtiene un coeficiente que indicará la calidad de la formación de conglomerados, mientras más grande sea el coeficiente es mejor.

2.3.3. T-SNE

t-Distributed Stochastic Neighbor Embedding es una técnica para la reducción de dimensionalidad que es especialmente adecuado para la visualización de conjuntos de datos de alta dimensión. La técnica se puede implementar a través de aproximaciones BarnesHut, lo que le permite ser aplicado en grandes conjuntos de datos del mundo real[6].

CAPÍTULO 3. PRE-PROCESAMIENTO

El trabajo con el dataset implica manejar una gran cantidad de datos, es por esto que es necesario analizar la conformación de datos de modo que la manipulación en procedimientos futuros sea más consistente. Para esto se debe analizar registros perdidos, valores atípicos o datos que no aporten información relevante para el estudio del problema, de forma que las conclusiones no se vean afectadas por anomalías de la base de datos.

3.1 ELIMINACIÓN DE REGISTROS NULOS

El dataset utilizado de hipotiroidismo posee 2800 sujetos originalmente, existiendo varios sujetos que tienen registros nulos (marcados con '?' en esta base de datos), es decir, hubo varios sujetos los cuales no se realizaron alguno de los exámenes. La cantidad de sujetos con datos incompletos se muestra especificada en la siguiente tabla:

Tabla 3.1: Detalles de variables con datos incompletos.

Variable	Porcentaje '?' [%]	Número de incidencias
sex	3.928571	110
TSH	10.14286	284
T3	20.89286	585
TT4	6.571429	184
T4U	10.60714	297
FTI	10.53571	295
TBG	100	2800

Se decide en contra de la imputación, dado que esta asume que los datos son aleatorios, pero en este caso es más probable que el doctor decida realizar todos los exámenes sólo cuando el sujeto presente síntomas de hipotiroidismo, en concordancia a los resultados obtenidos parcialmente por algunos de los exámenes. Si el doctor decide lo contrario, es decir, no someter al sujeto a todos los exámenes, esto podría significar que el sujeto se encuentra saludable (o que padece de otra enfermedad no relacionada con la tiroides), efecto que sería ocultado por la imputación de datos. Por estas razones se decide eliminar los pacientes que tienen registros incompletos.

3.2 ELIMINACIÓN DE VARIABLES

Se decide eliminar ciertas variables de acuerdo con los siguientes criterios:

- Variables que no contengan ningún dato o que todos sus datos sean registros desconocidos o nulos (NA o '?').
- Que tal variable no entregue información útil para el estudio que se quiere realizar.

3.2.1 TBG

Esta variable corresponde a el resultado de una medición de TBG, el cuál es un examen para medir el nivel de globulina fijadora de tiroxina, glucoproteína que lleva hormona tiroidea a través de la sangre. Ninguno de los pacientes fue sometido a este examen en la base de datos, razón suficiente para eliminar esta variable del análisis, dado que todos los datos de esta variable son '?'. Se eliminan las variables *TBG* y *TBG measured* de la base de datos.

3.2.2 Fuente de referencia

Esta variable indica las referencias que ocupa el autor de la base de datos, es decir, la persona que recopiló toda la información que tiene esta base de datos. Para el agrupamiento de datos, esta variable no entrega información útil, todo lo contrario, podría crear agrupaciones según donde se obtuvieron los datos, lo que sería contraproducente para el análisis de los datos. Se elimina la variable *referral.source* de la base de datos.

3.2.3 Variables de medición

Son variables que definen si un cierto sujeto fue sometido o no a un examen. Como se mencionó anteriormente, en la Sección 3.1, se eliminan todos los sujetos que no fueron sometidos a todos los exámenes que presenta la base de datos, por consecuencia de esto, las variables de medición ahora presentan todas un solo valor que indica que el sujeto se realizó el examen ('t'), por lo que esta variable ya no entrega información útil para cualquier tipo de análisis. Las variables de medición a eliminar son las siguientes:

- **TSH measured:** Indica si el paciente ha sido sometido a pruebas para medir TSH.
- **T3 measured:** Indica si el paciente ha sido sometido a pruebas para medir la hormona triiodotiroxina (T3).
- **TT4 measured:** Indica si el paciente ha sido sometido a mediciones de T4.
- **T4U measured:** Indica si el paciente ha sido sometido a pruebas de T4U.
- **FTI measured:** Indica si el paciente ha sido sometido a mediciones de FTI.

3.2.4 Variable de clasificación

Esta variable trae consigo un identificador del sujeto, el cual fue eliminado para poder tratar la variable simplemente según el resultado del diagnóstico para cada sujeto (Negativo, Hipotiroidismo Compensado, Hipotiroidismo Secundario o Hipotiroidismo Primario). Se omite esta variable del análisis de agrupamiento porque podría acondicionar negativamente los datos.

3.3 SELECCIÓN DE DATOS

En el caso de las variables continuas, hay algunas variables que presentan valores atípicos que causarían un sesgo inesperado en el análisis, por lo que se decide eliminar estos valores haciendo una selección de sujetos de acuerdo con un rango de valores.

Estos rangos de valores aceptables se establecen en base al estudio de las variables continuas realizado en la última experiencia. La cota inferior será 0, para que esta cota no sea negativa al aplicar el mismo criterio que en la cota superior. La cota superior será la media de la variable continua más 3 veces la desviación estándar de tal variable continua $mean(variable) + 3 * sd(variable)$. Tomando este como el criterio de selección, se obtiene el siguiente rango aceptable para cada variable continua:

Tabla 3.2: Rangos aceptables de variables continuas para la selección de sujetos.

Variable	Rango aceptable
age	0 - 120
TSH	0 - 59.67471
T3	0 - 4.489722
TT4	0 - 215.1227
T4U	0 - 1.595783
FTI	0 - 207.1889

Luego de realizar el pre-procesamiento de la base de datos argumentado en esta sección, la base de datos original de 2800 sujetos y 30 variables, se reduce a una base de datos de 1843 sujetos (65.82 % de la información) y 21 variables. Estas 21 variables se encuentran enumeradas a continuación:

1. age (Variable continua)
2. sex (Variable booleana)
3. on thyroxine (Variable booleana)
4. query on thyroxine (Variable booleana)
5. on antithyroid medication (Variable booleana)

6. sick (Variable booleana)
7. pregnant (Variable booleana)
8. thyroid surgery (Variable booleana)
9. I131 treatment (Variable booleana)
10. query hypothyroid (Variable booleana)
11. query hyperthyroid (Variable booleana)
12. lithium (Variable booleana)
13. goitre (Variable booleana)
14. tumor (Variable booleana)
15. hypopituitary (Variable booleana)
16. psych (Variable booleana)
17. TSH (Variable continua)
18. T3 (Variable continua)
19. TT4 (Variable continua)
20. T4U (Variable continua)
21. FTI (Variable continua)

CAPÍTULO 4. OBTENCIÓN DEL CLÚSTER

El agrupamiento de datos se realiza utilizando la función *pam* del paquete *cluster* de R, que es una versión más robusta del algoritmo de las k-medias. Para poder realizar bien el agrupamiento se debe determinar lo siguiente, distancia a utilizar y cuantos grupos tendrá el clúster.

4.1 DISTANCIAS ENTRE SUJETOS

Para poder agrupar la base de datos de hipotiroidismo, es necesario crear una matriz de disimilitud la cual permite determinar qué tan distinto es un sujeto de otro, por lo que se utiliza la distancia de Gower que es la más indicada en este caso, dado que la base de datos posee variables booleanas (categóricas) y continuas (cuantitativas).

4.2 CANTIDAD DE GRUPOS (K)

Para determinar la cantidad de grupos óptima, se realizan agrupamientos desde 2 grupos hasta 10 grupos, evaluando la calidad de cada agrupamiento utilizando el método de Silueta, obteniendo el siguiente gráfico:

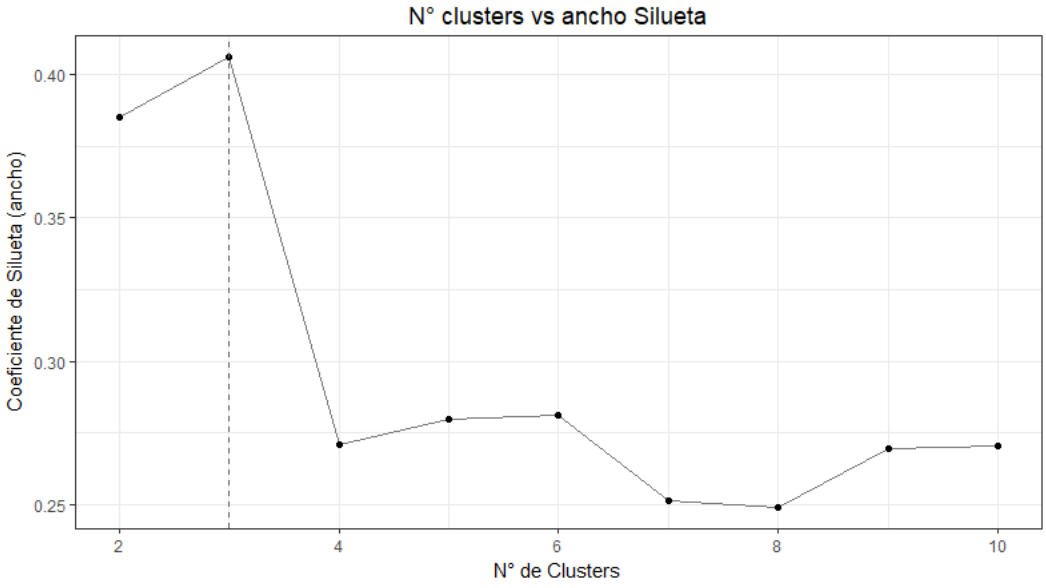

Figura 4.1: Resultados obtenidos por método de silueta para determinar agrupamiento óptimo.

De acuerdo a la Figura 4.1, la agrupación óptima obtenida por medio del método de Silueta es de k=3.

4.3 OBTENCIÓN DE CLÚSTER

Para poder visualizar los datos es necesario reducir la dimensionalidad de la base de datos, ya que esta posee 21 variables, lo que no es posible graficar, por lo que se utiliza la técnica t-SNE para llevar el agrupamiento a un plano bidimensional y así poder observar de forma aproximada los grupos formados.

Utilizando la matriz de disimilitud creada usando distancias de Gower y la agrupación óptima de k=3 grupos obtenida por el método de Silueta se obtiene el siguiente clúster reducido a un plano bidimensional usando la técnica t-SNE:

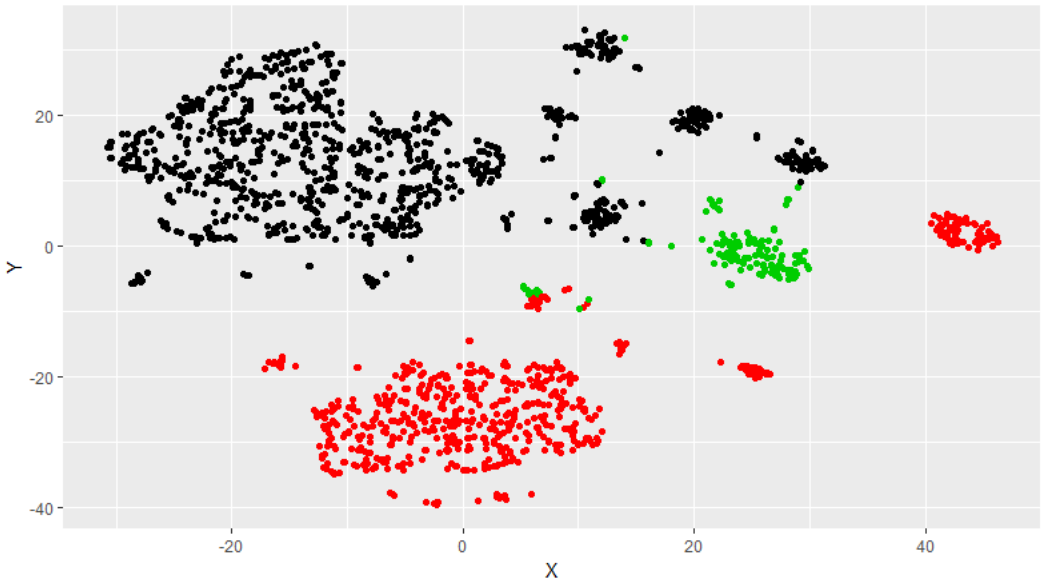

Figura 4.2: Base de datos clústerizada reducida a dos dimensiones por medio de la técnica t-SNE.

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS

Los clústeres, así como se plantean en el capítulo anterior no entregan la totalidad de información que se necesitaría para analizar fenómenos que ocurren dentro de ellos, es por eso que en este capítulo mediante las herramientas que proporciona R, se realizan distintos tipos de análisis a cada uno de los clústeres obtenidos.

5.1 ANÁLISIS ESTADÍSTICO

A continuación, se presentan una serie de datos estadísticos para cada uno de los clústeres de los cuales se desprende el respectivo análisis.

5.1.1 Clúster 1

En el primer clúster como se puede observar en la tabla 5.1, solo fueron agrupados sujetos de sexo femenino. Luego en la tabla 5.2 se presentan las distintas variables binarias de estudio que podrían tener incidencia en el hipotiroidismo, con la respectiva cantidad de sujetos que cumplen o no con la condición de dichas variables.

Por último, en la tabla 5.3 se presentan distintos componentes estadísticos para las variables continuas del clúster en cuestión.

Tabla 5.1: Distribución del género de los sujetos en clúster 1.

Variable	**F**	**M**
Género	1056	0

Tabla 5.2: Distribución de variables binarias en clúster 1.

Variable	f	t
on thyroxine	1056	0
query on thyroxine	1048	8
on antithyroid medication	1039	17
sick	1001	55
pregnant	1044	12
thyroid surgery	1040	16
I131 treatment	1037	19
query hypothyroid	1002	54
query hyperthyroid	986	70
lithium	1048	8
goitre	1049	7
tumor	1021	35
hypopituitary	1056	0
psych	1001	55

Tabla 5.3: Mediciones de la distribución de variables continuas en clúster 1.

Variable	Min.	1st Q.	Median	Mean	3rd Q.	Max.
Edad	2.00	37.00	57.00	53.98	70.25	93.00
TSH	0.005	0.600	1.500	3.063	2.625	55.000
T3	0.050	1.600	2.000	1.941	2.300	4.300
TT4	2.0	89.0	103.5	106.3	122.0	207.0
T4U	0.310	0.890	1.000	1.005	1.100	1.590
FTI	2.0	91.0	104.0	106.8	121.0	207.0

5.1.2 Clúster 2

A diferencia del clúster 1, en el clúster 2 como se puede observar en la tabla 5.4 solo se agrupan sujetos de sexo masculino como también una menor cantidad de observaciones.

Luego en la tabla 5.5 se presentan las distintas variables binarias de estudio que podrían tener incidencia en el hipotiroidismo, con la respectiva cantidad de sujetos que cumplen o no con la condición de dichas variables.

Por último, en la tabla 5.6 se presentan distintos componentes estadísticos para las variables continuas del clúster en cuestión.

Tabla 5.4: Distribución del género de los sujetos en clúster 2.

Variable	**F**	**M**
Género	0	635

Tabla 5.5: Distribución de variables binarias en clúster 2.

Variable	**f**	**t**
on thyroxine	615	20
query on thyroxine	628	7
on antithyroid medication	631	4
sick	606	29
pregnant	635	0
thyroid surgery	634	1
I131 treatment	628	7
query hypothyroid	616	19
query hyperthyroid	615	20
lithium	634	1
goitre	628	7
tumor	632	3
hypopituitary	634	1
psych	564	71

Tabla 5.6: Mediciones de la distribución de variables continuas en clúster 2.

Variable	**Min.**	**1st Q.**	**Median**	**Mean**	**3rd Q.**	**Max.**
Edad	1.00	39.00	55.00	52.82	67.00	94.00
TSH	0.005	0.600	1.300	2.195	2.300	45.000
T3	0.200	1.500	2.000	1.905	2.300	4.100
TT4	23.00	85.00	98.00	99.76	114.00	183.00
T4U	0.4100	0.8300	0.9200	0.9293	1.0200	1.5300
FTI	33	94	107	108	120	204

5.1.3 Clúster 3

A diferencia de los clústeres anteriores, en esta oportunidad como se puede observar en la tabla 5.7 que los sujetos agrupados son de ambos sexos, con mayor cantidad de sujetos en sexo femenino. Además, el número de observaciones disminuyo en comparación a los dos clústeres anteriores.

En la tabla 5.8 se presentan las distintas variables binarias de estudio que podrían tener incidencia en el hipotiroidismo, con la respectiva cantidad de sujetos que cumplen o no con la condición de dichas variables. En esta tabla ocurre un caso particular, el cual se da en sujetos que están en tratamiento con tiroxina, los cuales en su totalidad de encuentran en dicho tratamiento, caso contrario a lo que ocurre en el primer clúster, donde la totalidad de sujetos no se encuentra en este tratamiento.

Por último, en la tabla 5.9 se presentan distintos componentes estadísticos para las variables continuas del clúster en cuestión.

Tabla 5.7: Distribución del género de los sujetos en clúster 3.

Variable	**F**	**M**
Género	136	16

Tabla 5.8: Distribución de variables binarias en clúster 3.

Variable	**f**	**t**
on thyroxine	0	152
query on thyroxine	150	2
on antithyroid medication	150	2
sick	148	4
pregnant	149	3
thyroid surgery	147	5
I131 treatment	147	5
query hypothyroid	135	17
query hyperthyroid	145	7
lithium	150	2
goitre	151	1
tumor	151	1
hypopituitary	152	0
psych	151	1

Tabla 5.9: Mediciones de la distribución de variables continuas en clúster 3.

Variable	**Min.**	**1st Q.**	**Median**	**Mean**	**3rd Q.**	**Max.**
Edad	5.00	39.75	55.50	52.62	65.00	84.00
TSH	0.0050	0.0875	0.5400	2.9541	2.6000	44.0000
T3	0.30	1.70	2.00	2.11	2.40	4.20
TT4	37.0	108.5	129.5	130.8	151.0	213.0
T4U	0.720	0.920	1.010	1.036	1.110	1.550
FTI	51	108	126	127	147	197

CAPÍTULO 6. CONCLUSIONES

Por medio del último documento confeccionado, se llega a tener un conocimiento básico del hipotiroidismo y de las distribuciones de las variables en la base de datos, de esta forma se puede comenzar a realizar un análisis más profundo del tema. En esta experiencia se logra realizar un análisis por grupo e intentar identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

Sobre los resultados obtenidos, en el pre-procesamiento de datos, basándose en los conocimientos obtenidos en la última experiencia, se identifican los sujetos con registros nulos, los cuales son prontamente eliminados de la base de datos, dado que se asume que un doctor no cometería el error de no realizarle los exámenes pertinentes a un sujeto que puede estar padeciendo un problema a la tiroides. También se definen criterios para eliminación de variables y datos, eliminando variables que tienen todos los registros nulos (como es el caso de las variables *TBG* y *TBG measured*), las variables que no aportan nada de información luego de eliminar a los sujetos con registros nulos (como es el caso de las variables TSH measured, T3 measured, TT4 measured, T4U measured y FTI measured), variables que no aportan información y que perjudicarían el agrupamiento si se encontraran en la base de datos (como es el caso de las variables *referral source* y *resultados*). Sobre la eliminación de datos, se crea un rango de datos aceptados para las variables continuas, el cuál es desde 0 hasta la media sumada 3 veces la desviación estándar de tal variable continua, si un sujeto tiene un dato que se encuentre fuera del rango aceptado, tal sujeto no se incluirá en el análisis del agrupamiento.

Gracias al pre-procesamiento del dataset se obtiene una base de datos con datos más consistentes. Para un mejor pre-procesamiento y posteriormente, un mejor análisis, sería necesaria la ayuda de un especialista en el tema, con el cuál se pueda definir mejor que variables aportan información y cuáles no, además de poder definir un mejor rango de selección de datos, de forma de poder ignorar los sujetos con datos técnicamente imposibles y robustecer los clústeres que se obtendrían.

Luego para la obtención de los clústeres, utilizando la base de datos pre-procesada, se obtiene una matriz de disimilaridad utilizando como base la distancia de Gower que sirve para definir una distancia en conjuntos de datos mixtos (variables continuas y booleanas), que es el caso con esta base de datos. Luego se define la cantidad de grupos a utilizar en el *clustering* por medio del método silueta para evaluar el resultado de agrupar el dataset con valores de k desde 2 hasta 10 utilizando el método *pam* en RStudio, el cuál es una versión más robusta del método de las k-medias. Se llega a la conclusión por medio del método silueta que se obtendrían mejores resultados agrupando la base de datos con $k = 3$ (Figura 4.1), por lo que se utiliza la técnica t-SNE para obtener una visualización en un plano bidimensional del *clustering* realizado (Figura 4.2).

Más adelante, en el análisis de los clústeres el estudio realizado trajo como resultado, que el algoritmo realiza el agrupamiento de acuerdo al sexo de los sujetos, tomando en cuenta características que son repre-

sentadas por las variables booleanas, es decir, las variables continuas no tuvieron un gran impacto, contrario a lo conocido de la literatura donde las hormonas (que fueron las variables continuas estudiadas) son muy relevantes al momento de decidir el estado de la tiroides de un paciente. En consecuencia a esto, aunque haya sido útil el agrupamiento realizado, no fue lo suficientemente declarativo, por lo que se sugeriría lo siguiente para mejorar el desarrollo realizado. Hacer el mismo análisis pero sin las variables binarias, de manera de intentar obtener las relaciones que la literatura sugiere que debiera haber entre las variables continuas de esta base de datos.

CAPÍTULO 7. BIBLIOGRAFÍA

[1] A. T. Association, *Pruebas De Función Tiroidea*, 2016. dirección: https://www.thyroid.org/las-pruebas-de-funcion-tiroidea/.

[2] H. J. E. Eduardo Morales, *Clustering*, 2014. dirección: https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/clustering.pdf.

[3] U. de Oviedo, *El algoritmo k-means aplicado a clasificación y procesamiento de imágenes*, 2015. dirección: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html.

[4] A. Grané, *Distancias estadísticas y Escalado Multidimensional(Análisis de Coordenadas Principales)*, 2014. dirección: http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf.

[5] D. d. C. B. Universidad nacional de Lujan, *Calidad del agrupamiento: Coeficiente de Silueta*, 2015. dirección: http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/bdm/coeficiente-silueta.pdf.

[6] M. J. Molina, *Desarrollo de un método de reducción dimensional no lineal y clustering para la visualización e interpretación de single cell RNA-seq data*, 2014. dirección: http://www.masterbioinformatica.com/wp-content/uploads/tfm_2013_2014/TFM_MIGUEL_JULIA.pdf.

CAPÍTULO 8. ANEXO - CÓDIGO R

```r
library(cluster)
library(ggplot2)
library(Rtsne)

# Lectura de los datos + asignarle nombres a las columnas de acuerdo a lo escrito en allhypo.names
allhypo <- read_csv("allhypo/allhypo.data", col_names = FALSE)
allhypoRownames <- c("age", "sex", "on thyroxine", "query on thyroxine", "on antithyroid medication", "sick", "pregnant", "thyroid surgery", "I131 treatment", "query hypothyroid", "query hyperthyroid", "lithium", "goitre", "tumor", "hypopituitary", "psych", "TSH measured", "TSH", "T3 measured", "T3", "TT4 measured", "TT4", "T4U measured", "T4U", "FTI measured", "FTI", "TBG measured", "TBG", "referral source", "results")
colnames(allhypo) <- allhypoRownames

# Mostrar la cantidad de sujetos con datos incompletos por variable
lostsex <- length(allhypo$sex[allhypo$sex == '?']) + length(allhypo$sex[allhypo$sex == 'NA'])
lostTSH <- length(allhypo$TSH[allhypo$TSH == '?']) + length(allhypo$TSH[allhypo$TSH == 'NA'])
lostT3 <- length(allhypo$T3[allhypo$T3 == '?']) + length(allhypo$T3[allhypo$T3 == 'NA'])
lostTT4 <- length(allhypo$TT4[allhypo$TT4 == '?']) + length(allhypo$TT4[allhypo$TT4 == 'NA'])
lostT4U <- length(allhypo$T4U[allhypo$T4U == '?']) + length(allhypo$T4U[allhypo$T4U == 'NA'])
lostFTI <- length(allhypo$FTI[allhypo$FTI == '?']) + length(allhypo$FTI[allhypo$FTI == 'NA'])
lostTBG <- length(allhypo$TBG[allhypo$TBG == '?']) + length(allhypo$TBG[allhypo$TBG == 'NA'])

# Calcular el porcentaje de perdida por cada variable con datos incompletos
percentageLostsex <- lostsex / 2800 * 100
percentageLostTSH <- lostTSH / 2800 * 100
```

```
22 percentageLostT3 <- lostT3 / 2800 * 100
23 percentageLostTT4 <- lostTT4 / 2800 * 100
24 percentageLostT4U <- lostT4U / 2800 * 100
25 percentageLostFTI <- lostFTI / 2800 * 100
26 percentageLostTBG <- lostTBG / 2800 * 100
27
28 # Mostrar en pantalla los datos de perdida
29 cat("El numero de incidencias para la variable sex es", lostsex, "con un % de
       perdida de:", percentageLostsex, "\n")
30 cat("El numero de incidencias para la variable TSH es", lostTSH, "con un % de
       perdida de:", percentageLostTSH, "\n")
31 cat("El numero de incidencias para la variable T3 es", lostT3, "con un % de
       perdida de:", percentageLostT3, "\n")
32 cat("El numero de incidencias para la variable TT4 es", lostTT4, "con un % de
       perdida de:", percentageLostTT4, "\n")
33 cat("El numero de incidencias para la variable T4U es", lostT4U, "con un % de
       perdida de:", percentageLostT4U, "\n")
34 cat("El numero de incidencias para la variable FTI es", lostFTI, "con un % de
       perdida de:", percentageLostFTI, "\n")
35 cat("El numero de incidencias para la variable TBG es", lostTBG, "con un % de
       perdida de:", percentageLostTBG, "\n")
36
37 # Limpieza de los resultados (no se usa el ".|numero" solo interesa la clase)
38 allhypo$results <- vapply(strsplit(allhypo$results,"\\."), `[`, 1, FUN.VALUE=
       character(1))
39
40 # Se eliminan estas columnas porque no fue medida la TBG en la base de datos.
41 allhypo$TBG <- NULL
42 allhypo$`TBG measured` <- NULL
43
44 # Luego se procede a cambiar el caracter '?' por NA para poder ejecutar
       complete.cases()
45 allhypo[allhypo=="?"] <- NA
46 allhypo$age[allhypo$age >= 123] <- NA
47
48 # Se eliminan los datos que tienen NA (originalmente '?') por los motivos
49 # descritos en el informe
```

```
50 allhypo <- allhypo[complete.cases(allhypo), ]
51
52 # Luego como todas las columnas de si es que se mide o no algo son siempre 't'
53 # por haber eliminado los sujetos, no se usaran estas columnas en el analisis
54 allhypo$`T3 measured` <- NULL
55 allhypo$`T4U measured` <- NULL
56 allhypo$`TSH measured` <- NULL
57 allhypo$`TT4 measured` <- NULL
58 allhypo$`FTI measured` <- NULL
59
60 # Eliminar referral.source y results por las razones mencionandas en este
       informe
61 allhypo$`referral source` <- NULL
62 allhypo$results <- NULL
63
64 # Transformar todas las columnas numericas a variables numericas.
65 allhypo <- transform(allhypo, age = as.numeric(age))
66 allhypo <- transform(allhypo, TSH = as.numeric(TSH))
67 allhypo <- transform(allhypo, T3 = as.numeric(T3))
68 allhypo <- transform(allhypo, TT4 = as.numeric(TT4))
69 allhypo <- transform(allhypo, T4U = as.numeric(T4U))
70 allhypo <- transform(allhypo, FTI = as.numeric(FTI))
71
72 # Transformar todas las columnas booleanas a factores
73 allhypo <- transform(allhypo, sex = as.factor(sex))
74 allhypo <- transform(allhypo, on.thyroxine = as.factor(on.thyroxine))
75 allhypo <- transform(allhypo, query.on.thyroxine = as.factor(query.on.thyroxine
       ))
76 allhypo <- transform(allhypo, on.antithyroid.medication = as.factor(
       on.antithyroid.medication))
77 allhypo <- transform(allhypo, sick = as.factor(sick))
78 allhypo <- transform(allhypo, pregnant = as.factor(pregnant))
79 allhypo <- transform(allhypo, thyroid.surgery = as.factor(thyroid.surgery))
80 allhypo <- transform(allhypo, I131.treatment = as.factor(I131.treatment))
81 allhypo <- transform(allhypo, query.hypothyroid = as.factor(query.hypothyroid))
82 allhypo <- transform(allhypo, query.hyperthyroid = as.factor(query.hyperthyroid
       ))
```

```
allhypo <- transform(allhypo, lithium = as.factor(lithium))
allhypo <- transform(allhypo, goitre = as.factor(goitre))
allhypo <- transform(allhypo, tumor = as.factor(tumor))
allhypo <- transform(allhypo, hypopituitary = as.factor(hypopituitary))
allhypo <- transform(allhypo, psych = as.factor(psych))

# Eliminar los valores atipicos
sdRange <- 3
MaxTSH <- mean(allhypo$TSH) + sdRange * sd(allhypo$TSH)
MaxT3 <- mean(allhypo$T3) + sdRange * sd(allhypo$T3)
MaxTT4 <- mean(allhypo$TT4) + sdRange * sd(allhypo$TT4)
MaxT4U <- mean(allhypo$T4U) + sdRange * sd(allhypo$T4U)
MaxFTI <- mean(allhypo$FTI) + sdRange * sd(allhypo$FTI)
allhypo <- subset(allhypo , (age <= 120) & (TSH <= MaxTSH) & (T3 <= MaxT3) & (
    TT4 <= MaxTT4) & (T4U <= MaxT4U) & ( FTI <= MaxFTI) )

# Calcular distancias de gower
allhypoDistances <- daisy(allhypo, metric="gower")

# Usar el metodo de las siluetas para obtener la cantidad optima de clusters
k.max <- 10
k.vector <- 2:k.max
silAvgWidth <- unlist(sapply(k.vector, function(k.cluster){pam(allhypoDistances
    , diss = TRUE, k = k.cluster)$silinfo$avg.width}))

clustersPlot <- ggplot(data.frame(k.vector, silAvgWidth), aes(x = k.vector, y =
     silAvgWidth)) +
  labs(x = "N de Clusters", y = "Coeficiente de Silueta (ancho)") +
  geom_line(color = "#777777") +
  geom_point(color = "black") +
  theme_bw() +
  ggtitle("N clusters vs ancho Silueta") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_vline(xintercept = 3, linetype = 2, color="#666666")
show(clustersPlot)

# Por medio del metodo de siluetas, la agrupacion optima es con k=3 clusters
```

```
117 clusters <- pam(allhypoDistances, diss = TRUE, k = 3)
118
119 # Mostrar informacion/resumen de los datos de cada cluster
120 allhypo["cluster"] <- clusters$clustering
121 cat("------------------------------------------------------------\n")
122 cat("Resumen de datos para el cluster 1 \n")
123 show(summary(allhypo[allhypo$cluster == 1, ]))
124 cat("------------------------------------------------------------\n")
125 cat("Resumen de datos para el cluster 2 \n")
126 show(summary(allhypo[allhypo$cluster == 2, ]))
127 cat("------------------------------------------------------------\n")
128 cat("Resumen de datos para el cluster 3 \n")
129 show(summary(allhypo[allhypo$cluster == 3, ]))
130
131 # Para poder graficar los datos en dos dimensiones se utiliza t-SNE
132 set.seed(6748)
133 tsne <- Rtsne(allhypoDistances, is_distance = TRUE)
134 groupingPlot <- ggplot(data.frame(tsne$Y), aes(x = X1, y = X2)) +
135   labs(x = "X", y = "Y") +
136   geom_point(color = factor(clusters$clustering))
137 show(groupingPlot)
```