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

ANÁLISIS DE DATOS

HIPERTIROIDISMO

FELIPE JARA

FRANCISCO MUÑOZ

Profesor: Max Chacón

Ayudantes: Adolfo Guzmán

Santiago - Chile

April 29, 2017

TABLA DE CONTENIDOS

CAPÍTULO 1. INTRODUCCIÓN 5

CAPÍTULO 2. MARCO TEÓRICO 7

2.1 CLUSTERING 7

2.2 ALGORITMO K-MEANS 7

2.3 DISTANCIAS UTILIZADAS 8

2.3.1 Distancia de Gower 8

2.4 MÉTODO DE SILUETAS 8

2.5 T-SNE 9

CAPÍTULO 3. PRE-PROCESAMIENTO 11

CAPÍTULO 4. OBTENCIÓN DEL CLÚSTER 15

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS 25

CAPÍTULO 6. CONCLUSIÓN 27

CAPÍTULO 7. REFERENCIAS 29

CAPÍTULO 1. INTRODUCCIÓN

La glándula tiroidea es uno de los órganos más relevantes del sistema endocrino. Es la responsable de estimular la producción de proteínas en casi todos los tejidos del organismo e incrementar la cantidad de oxígeno que utilizan las células. De alguna forma se encarga de controlar el metabolismo y la sensibilidad del cuerpo. Si dicha glándula secreta una cantidad excesiva de hormonas tiroideas (T3 y T4), se acelera el metabolismo, y en consecuencia la persona afectada padece de síntomas como taquicardias, perdida de peso, nerviosismo, temblores, entre otros síntomas y enfermedades. Cómo fue posible apreciar en la primera experiencia del laboratorio de Análisis de Datos, el comportamiento anormal del cuerpo descrito anteriormente se conoce como hipertiroidismo y corresponde al objeto de estudio de las distintas experiencias que se realizan en dicho ramo por los autores de este documento.

En la primera experiencia se vivió una fase de interiorización con la base de datos, en donde se realizaron definiciones de las variables existentes, análisis estadístico de aquellas variables y conclusiones en base a los análisis. Con lo anterior se dio el primer paso y se realizó al primer acercamiento a la seguidilla de experiencias del análisis de una base de datos sobre el hipertiroidismo.

El documento actual corresponde a la segunda experiencia a realizar, aquí se plantea entender las características que tienen en común los distintos síntomas del hipertiroidismo y en consecuencia los distintos casos del hipertiroidismo. Lo anterior se pretende abordar mediante el agrupamiento de datos, para los cuales se deben usar técnicas definidas, y así, analizar y caracterizar los grupos generados.

El informe consta de la introducción, donde se presenta el contexto y los objetivos del documento. Un marco teórico, en el cual se definen los conceptos principales de la experiencia. Luego se exponen las decisiones de pre-procesamiento con sus respectivas justificaciones. Posteriormente se presentan los agrupamiento realizados y sus análisis. Finalmente se realizan las conclusiones de la experiencia y se presentan las referencias usadas. De manera anexa se agrega el código en R usado para llevar a cabo el trabajo.

CAPÍTULO 2. MARCO TEÓRICO

2.1 CLUSTERING

El clustering es el proceso de agrupar datos en clases o grupos. Consiste en el agrupamiento de vectores bajo un criterio dado, este criterio en general corresponde a la distancia de los vectores o la similitud de ellos. La cercanía entre datos se determina mediante una función de distancia, la cual puede ser una función euclídea, o cualquier otra función que pueda medir distancias. La similitud de los datos puede ser determinada con una matriz de correlaciones, aunque también existen algoritmos que maximizan una propiedad estadística llamada "verosimilitud".

Por lo general, los grupos generados utilizando clustering, cuentan con características similares. De esta forma es posible obtener conocimiento mediante el análisis o minería de datos. Las aplicaciones del clustering son diversas, es posible utilizarla en áreas tan distintas como lo son la biología y el marketing, incluso es posible mencionar que las áreas donde puede ser de utilidad el clustering, se encuentran limitadas simplemente por la imaginación de las mentes humanas.

Existen diversos algoritmos de agrupamiento, entre ellos se encuentra uno de los que se usa en la presente experiencia, este corresponde al algoritmo de las K-means o K-medias. Cabe destacar que el clustering busca generar grupos de tal forma que en su interior exista una alta semejanza y de igual forma en el exterior exista una gran diferencia.

2.2 ALGORITMO K-MEANS

Es el algoritmo más simple de aprendizaje no supervisado y que realiza clustering. Los algoritmos de aprendizaje no supervisado sirven para reconocer formas, pre-procesar datos, reconocer imágenes, reconocer voz, entre otros. Este algoritmo aproxima los grupos (o clusters) de manera iterativa. Cabe destacar que la cantidad de grupos que formará este algoritmo es prefijado, esto quiere decir que la cantidad de grupos a formar es una decisión humana y por lo tanto es una variable de entrada del algoritmo.

El algoritmo de K-mean se presenta a continuación:

1. Se sitúan los K centroides en el espacio donde se "encuentran" los datos que se quieren agrupar.

2. Se asigna cada dato al grupo con el centroide más cercano.

3. Una vez que todos los datos se asignaron a un centroide, se recalculan los centroides, generalmente esto es realizado calculando el punto medio de los datos de cada grupo, de tal forma que el centroide se mueve a las coordenadas calculadas como punto medio según corresponda.

4. Se repiten los pasos 2 y 3 hasta que los grupos permanezcan sin variación.

2.3 DISTANCIAS UTILIZADAS

La distancia más común y utilizada es la distancia Euclidiana, esta corresponde a la distancia más corta entre dos puntos. En el caso de nuestra experiencia no es posible utilizar dicha forma de calcular la distancia, debido a que la distancia Euclidiana sólo es válida para valores continuos. Por lo anterior se debe buscar otro tipo de distancia que se acomode al problema. En base a la investigación, fue posible determinar que una distancia muy usada cuando se tienen datos cualitativos y cuantitativos es la distancia de Gower.

2.3.1 Distancia de Gower

Esta distancia es usada para determinar distancias para valores mixtos. Dicha distancia sirve para calcular distancia en conjuntos de datos con variables cuantitativas y cualitativas. A continuación se presenta la fórmula de esta distancia.

$$d_{ij}^2 = 1 - \frac{\sum_{h=1}^{p1}(1 - |x_{ih} - x_{jh}|/G_h) + a + \alpha}{p1 + (p2 - d) + p3} \tag{2.1}$$

2.4 MÉTODO DE SILUETAS

Este método busca determinar la mejores medidas de similitud intra-cluster e inter-cluster. En otras palabras busca maximizar el ancho de las siluetas y que de esta forma los elementos se clasifiquen de forma correcta. El método funciona creando grupos de distintos tamaños y calculando el ancho promedio de sus siluetas. El mejor agrupamiento corresponde al que tiene la mayor silueta.

2.5 T-SNE

Esto corresponde a una forma o técnica para reducir la dimensionalidad de un conglomerado de datos. El computador está preparado para calcular grandes cantidades de datos a la vez, y sin problemas trabajar con múltiples dimensiones. En cambio, para el humano es muy difícil visualizar múltiples dimensiones, por lo cual este algoritmo es usado para reducir la cantidad de dimensiones a 2 o 3. A grandes rasgos, el concepto de este algoritmo es que los datos cercanos se atraen y los datos lejanos se repelen.

CAPÍTULO 3. PRE-PROCESAMIENTO

Uno de los problemas con la base de datos proporcionada consiste en la ausencia de información en ciertas variables o en el conjunto de ellas. En la mayoría de los casos, son datos que no pueden ser inferidos o completados de ninguna manera, lo cual obliga a eliminar los registros cuyas pérdidas obstaculizan el desarrollo de un análisis con respecto a la enfermedad del hipertiroidismo. Es decir, aquellos registros que presentan pérdida de información de alguna de las variables relevantes para el estudio serán ignorados.

A continuación, se presenta una tabla con el porcentaje de pérdida o ausencia de información solo para las variables registradas en la base de datos que presenten datos sin información:

Tabla 3.1: Porcentaje de pérdida de variables.

Variable	**Porcentaje pérdida**	**Número de incidencias**
age	0.036	1
sex	3.929	110
TSH	10.143	284
T3	20.892	585
TT4	6.5714	184
T4U	10.607	297
FTI	10.536	295
TBG	100	2800

De la tabla anterior, se puede apreciar como la variable TBG (Niveles de globulina fijadora de tiroxina en el torrente sanguíneo) presenta un porcentaje de pérdida del 100%, razón suficiente como para descartar la variable en el estudio sin dar mayores explicaciones.

Adicionalmente, se considera que no es necesario incluir la variable FTI en el estudio si se tiene en cuenta que esta puede ser obtenida a partir del cuociente entre el TT4 (nivel de tiroxina en el torrente sanguíneo) y el T4U (Tasa de utilización de tiroxina), y que con-

siderarla en el estudio sería prácticamente una redundancia de información. Sin embargo, conociendo el FTI es posible recuperar el TT4 y T4U de los registros que presentan pérdida o ausencia de alguna de las dos variables mencionadas. Dicho esto, fue posible realizar esta recuperación de información en dos de los registros del dataset proporcionado.

Así como el FTI no aporta información adicional para el estudio, hay otras variables que también lo hacen o que derechamente son irrelevantes para el estudio a realizar. Esto sería el caso de:

- query on thyroxine, query hypothyroid, query hyperthyroid: Ya que solo indican si el individuo en estudio se sometió a un examen médico o no, pero no los resultados.

- TSH measured, T3 measured, TT4 measured, T4U measured, FTI measured, TBG measured: Ya que solo indican si el individuo se realizó un análisis de la hormona correspondiente o no, pero no los niveles de concentración hormonal en el torrente sanguíneo (esa información están registradas en otras variables).

- Referral source: Ya que solo indica la fuente desde donde se obtuvieron los datos.

- Diagnostico (hyperthyroid, T3 toxic, goitre, secondary toxic, negative): Ya que se tiene la suposición de la existencia de diagnósticos incorrectos en el dataset, por lo que saber de antemano si un individuo presenta hipertiroidismo o no, de acuerdo a su diagnóstico, podría condicionar los resultados de los análisis por realizar. Sin embargo, se considerará para el posterior algoritmo de agrupamiento k-means. la obtención de 4 grupos diferentes que deberían estar relacionados con los 4 tipo de diagnósticos proporcionados (Hyperthyroid, T3toxic, goitre, negativo).

- Sick: Debido a que el concepto de enfermedad es muy amplio para considerarlo en el estudio, ya que la enfermedad podría llegar a ser desde un simple resfriado común a una condición de anemia. Al no conocer la enfermedad específica, nos deja un rango muy amplio de posibilidades que podrían ofuscar el desarrollo de conclusiones adecuadas con respecto al análisis de la información proporcionada.

Asimismo, conociendo los rangos de niveles normales de la hormona TSH, TT4, T4U y T3 en la sangre, los cuales se encuentran registrados en el informe de la anterior experiencia de laboratorio, se realiza una limpieza de los datos cuyos valores se escapan de manera exagerada al rango que se considera normal. Para ello, y debido al desconocimiento de los valores máximos que podría llegar la concentración de una hormona en casos de enfermedad, como medida arbitraria se decide calcular la distancia entre los valores mínimos y máximos del rango aceptable, multiplicarla por 3 y establecer una cota superior e inferior nueva en base al valor obtenido.

Los nuevos rangos vendrían siendo los siguientes:

Tabla 3.3: Rangos de variables.

Hormona	**Rango normal**	**Rango utilizado**
TSH	0.5 - 4.7 (mlU/L)	0 - 17.3 (mlU/L)
T3	58 - 161 (nmol/L)	0 - 470 (nmol/L)
TT4	0.8 - 1.3 (TBI)	0 - 2.8 (TBI)
T4U	0.9- 2.8 (nmol/L)	0 - 8.5 (nmol/L)

Por último, se determina un rango de valores posibles para la edad de 0 a 117 años. Siendo 117 años el récord de la cantidad más alta de años que ha vivido un ser humano. Esto permite eliminar del dataset un registro con una edad imposible de 455 años.

Finalmente, tras haber realizado un pre-procesamiento de los datos, se cuenta con un dataset de 16 variables y 1880 registros en total (de los cuales 46 corresponden a pacientes a los que se les diagnosticó hipertiroidismo). Entonces, las variables que serán consideradas en el estudio son:

1. age (variable discreta)
2. sex (variable categórica)
3. on thyroxine (variable categórica)
4. on antithyroid medication (variable categórica)

5. pregnant (variable categórica)
6. thyroid surgery (variable categórica)
7. I131 treatment (variable categórica)
8. lithium (variable categórica)
9. goitre (variable categórica)
10. tumor (variable categórica)
11. hypopituitary (bivariable categóricanario)
12. psych (variable categórica)
13. TSH (variable continua)
14. T3 (variable continua)
15. TT4 (variable continua)
16. T4U (variable continua)

CAPÍTULO 4. OBTENCIÓN DEL CLÚSTER

Como ya se mencionó anteriormente en el marco teórico, para la obtención del cluster se ha decidido utilizar como medida de similaridad la distancia de Gower debido a que está métrica permite trabajar con la combinación de diferentes tipos de variables (En este caso, variables numéricas y categóricas). Lo cual es una ventaja frente a otro tipo de métricas como la distancia Euclidiana, la que es válida sólo para variables de tipo continua. En consecuencia, se estima conveniente el uso del algoritmo de clustering K-medoids para conformar los diferentes grupos que se esperan encontrar en el dataset proporcionado.

Teniendo esto en cuenta, el primer paso vendría siendo el determinar el número de grupos que se estime necesario conformar (valor de K), de tal manera que aporte la mejor información para los posteriores análisis por realizar. Debido a la falta de experticia en temas de medicina, no nos es posible dar a priori un valor estimado de K. Sin embargo, aplicando el método de las siluetas es posible calcular un número óptimo de cluster por conformar. De ésta manera, se obtienen los siguientes resultados:

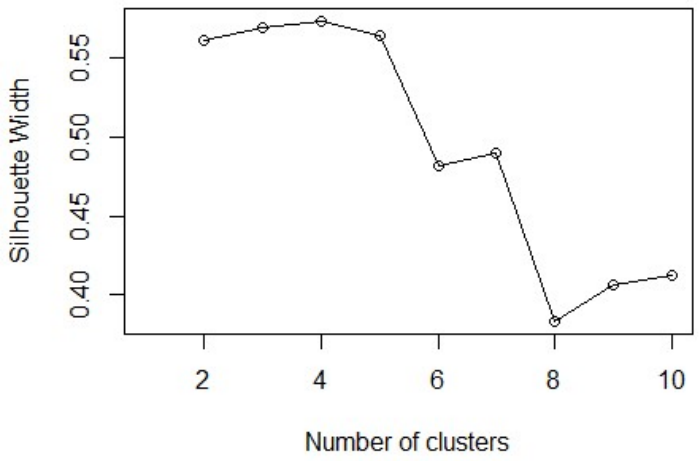

Figura 4-1: Gráfico siluetas.

A priori, se podría decir que el número óptimo de clusters por considerar sería de K = 4, lo cual coincide con los cuatro tipos de diagnósticos registrados en la base de

datos proporcionada (sin considerar el diagnóstico de secondary toxic, ya que ninguno de los registros presentaba dicho diagnóstico) . Sin embargo, ya que el método de la silueta corresponde a una heurística y no a una solución absoluta, se considera además los valores vecinos K=3 y K=5 dado a la posibilidad de que alguno de estos valores entreguen mejores resultados que el otro.

De ésta manera, y graficando los clusters obtenidos para K = 3, 4 y 5 en un gráfico de dispersión basado en el algoritmo t-SNE para reducción de dimensionalidad, se obtienen los siguientes resultados:

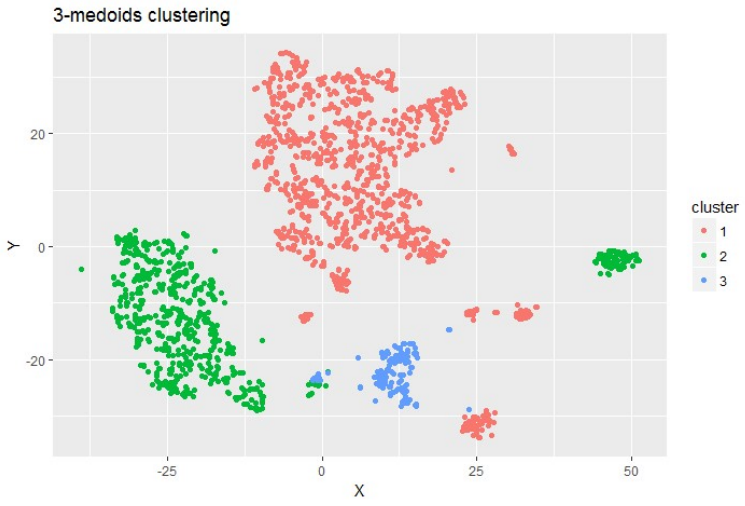

Figura 4-2: Clúster con k = 3.

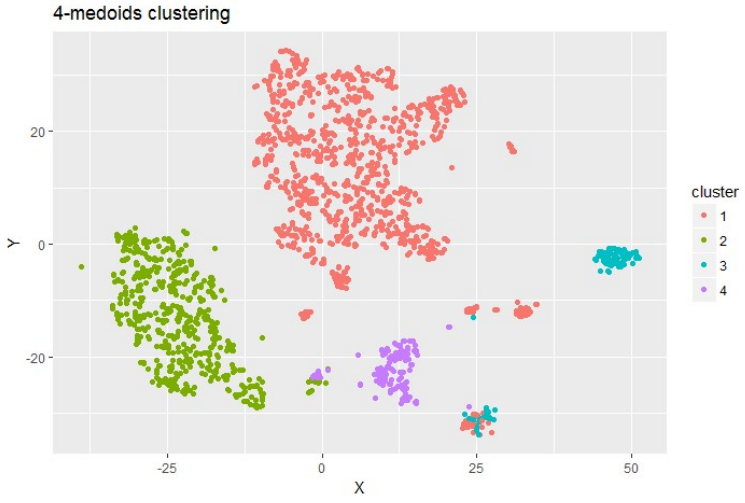

Figura 4-3: Clúster con k = 4.

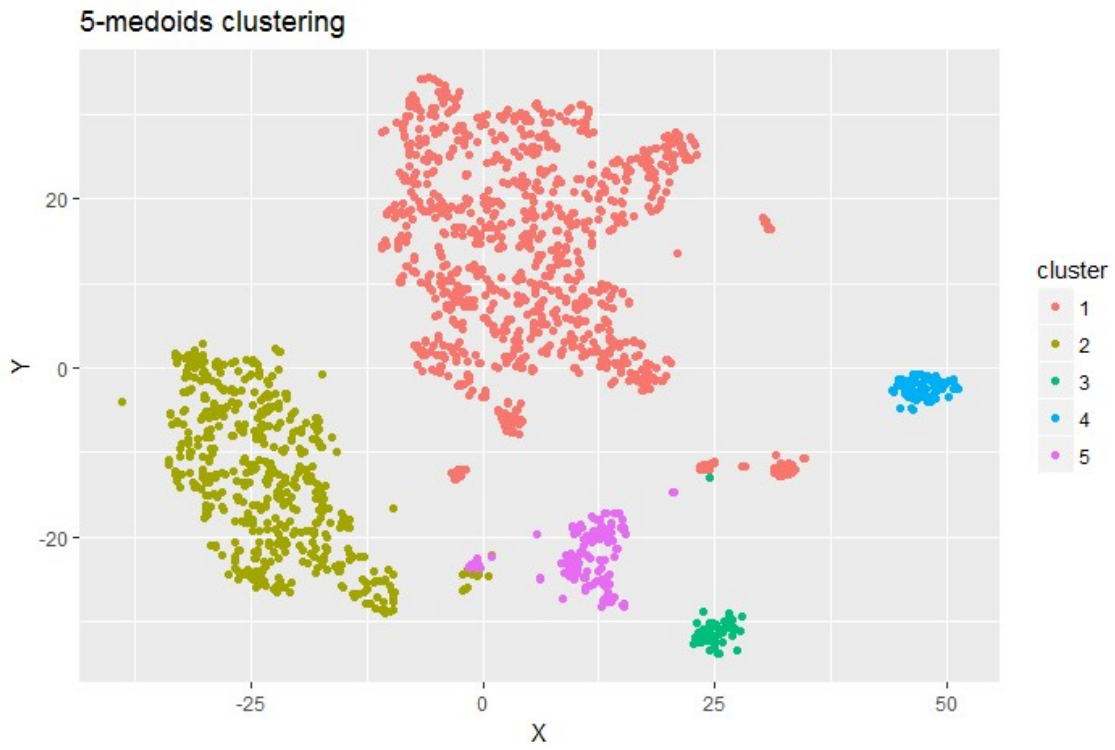

Figura 4-4: Clúster con k = 5.

A partir de los gráficos anteriores, se puede apreciar visualmente como los grupos quedan definidos de mejor manera al utilizar el algoritmo de k-medoids con un k = 5, ya que no hay lugar en el plano en donde las agrupaciones se crucen entre ellas, a diferencia de los resultados obtenidos con un k=3 y k=4.

Las siguientes tablas proveen un resumen del conjunto de datos considerado para cada uno de los cinco clusters conformados:

CLUSTER 1 (DATOS = 1031)						
Variable numérica	**Mínimo**	**1er Cuart.**	**Mediana**	**Media**	**3er Cuart.**	**Máximo**
Age (años)	2.00	36.00	57.00	53.72	70.50	93.00
TSH (mlU/L)	0.005	0.420	1.300	2.019	2.400	16.000
T3 (nmol/L)	0.050	1.600	2.000	2.085	2.400	7.300
TT4 (nmol/L)	19.0	91.0	106.0	112.7	126.0	430.0
T4U (TBI)	0.310	0.890	1.000	1.029	1.110	2.120
Variable categórica	**Verdadero/Female**			**Falso/Male**		
Sex	1031			0		
On thyroxine	0			1031		
On antithyroid med	20			1011		
Pregnant	25			1006		
Thyroid surgery	16			1015		
I131.treatment	19			1012		
Lithium	6			1025		
Goitre	7			1024		
Tumor	38			993		
Hypopituitary	0			1031		
Psych	0			1031		

Figura 4-5: Tabla de clúster 1.

CLUSTER 2 (DATOS = 567)						
Variable numérica	**Mínimo**	**1er Cuart.**	**Mediana**	**Media**	**3er Cuart.**	**Máximo**
Age (años)	1.00	41.00	56.00	53.78	68.00	94.00
TSH (mlU/L)	0.005	0.550	1.300	1.901	2.300	15.000
T3 (nmol/L)	0.2	1.5	1.9	1.9	2.3	7.1
TT4 (nmol/L)	38.0	85.0	98.0	100.4	113.5	246.0
T4U (TBI)	0.41	0.82	0.92	0.93	1.02	1.68
Variable categórica	**Verdadero/Female**			**Falso/Male**		
Sex	0			567		
On thyroxine	21			546		
On antithyroid med	4			563		
Pregnant	0			567		
Thyroid surgery	2			565		
I131.treatment	6			561		
Lithium	1			566		
Goitre	7			560		
Tumor	4			563		
Hypopituitary	1			566		
Psych	0			567		

Figura 4-6: Tabla de clúster 2.

CLUSTER 3 (DATOS = 58)						
Variable numérica	**Mínimo**	**1er Cuart.**	**Mediana**	**Media**	**3er Cuart.**	**Máximo**
Age (años)	18.0	32.5	41.0	46.0	62.0	84.0
TSH (mlU/L)	0.200	1.125	1.550	2.023	2.225	12.000
T3 (nmol/L)	0.400	1.800	2.000	2.088	2.400	3.500
TT4 (nmol/L)	73.0	95.0	111.5	116.7	134.8	187.0
T4U (TBI)	0.600	0.922	1.010	1.030	1.120	1.820
Variable categórica	**Verdadero/Female**			**Falso/Male**		
Sex	58			0		
On thyroxine	1			57		
On antithyroid med	0			58		
Pregnant	1			57		
Thyroid surgery	0			58		
I131.treatment	0			58		
Lithium	2			56		
Goitre	1			57		
Tumor	1			57		
Hypopituitary	0			58		
Psych	0			58		

Figura 4-7: Tabla de clúster 3.

CLUSTER 4 (DATOS = 71)						
Variable numérica	**Mínimo**	**1er Cuart.**	**Mediana**	**Media**	**3er Cuart.**	**Máximo**
Age (años)	15.00	29.00	39.00	43.85	59.00	84.00
TSH (mIU/L)	0.050	0.990	1.400	1.928	2.050	13.00
T3 (nmol/L)	0.600	1.800	2.100	2.139	2.500	3.800
TT4 (nmol/L)	58.0	92.5	102.0	105.3	120.0	141.0
T4U (TBI)	0.5900	0.8700	0.9400	0.9452	1.0200	1.3100
Variable categórica	**Verdadero/Female**			**Falso/Male**		
Sex	0			71		
On thyroxine	0			71		
On antithyroid med	0			71		
Pregnant	0			71		
Thyroid surgery	0			71		
I131.treatment	0			71		
Lithium	0			71		
Goitre	0			71		
Tumor	0			71		
Hypopituitary	0			71		
Psych	0			71		

Figura 4-8: Tabla de clúster 4.

CLUSTER 5 (153)						
Variable numérica	**Mínimo**	**1er Cuart.**	**Mediana**	**Media**	**3er Cuart.**	**Máximo**
Age (años)	5.00	39.00	57.00	53.09	65.00	84.00
TSH (mlU/L)	0.005	0.065	0.300	1.810	2.100	17.00
T3 (nmol/L)	0.30	1.70	2.20	2.18	2.50	6.70
TT4 (nmol/L)	37.0	111.0	131.0	136.7	160.0	289.0
T4U (TBI)	0.720	0.920	1.010	1.057	1.120	1.800
Variable categórica	**Verdadero/Female**			**Falso/Male**		
Sex	138			15		
On thyroxine	0			153		
On antithyroid med	1			152		
Pregnant	6			147		
Thyroid surgery	4			149		
I131.treatment	5			148		
Lithium	2			151		
Goitre	0			153		
Tumor	1			152		
Hypopituitary	0			153		
Psych	0			153		

Figura 4-9: Tabla de clúster 5.

Asimismo, se proporcionan otros gráficos adicionales para obtener una segunda perspectiva de los clusters generados. Esta vez utilizando el algoritmo de reducción de dimen-

sionalidad basada en componentes principales. De está manera, y descartando el clustering de 3 agrupaciones, se obtienen los siguientes gráficos:

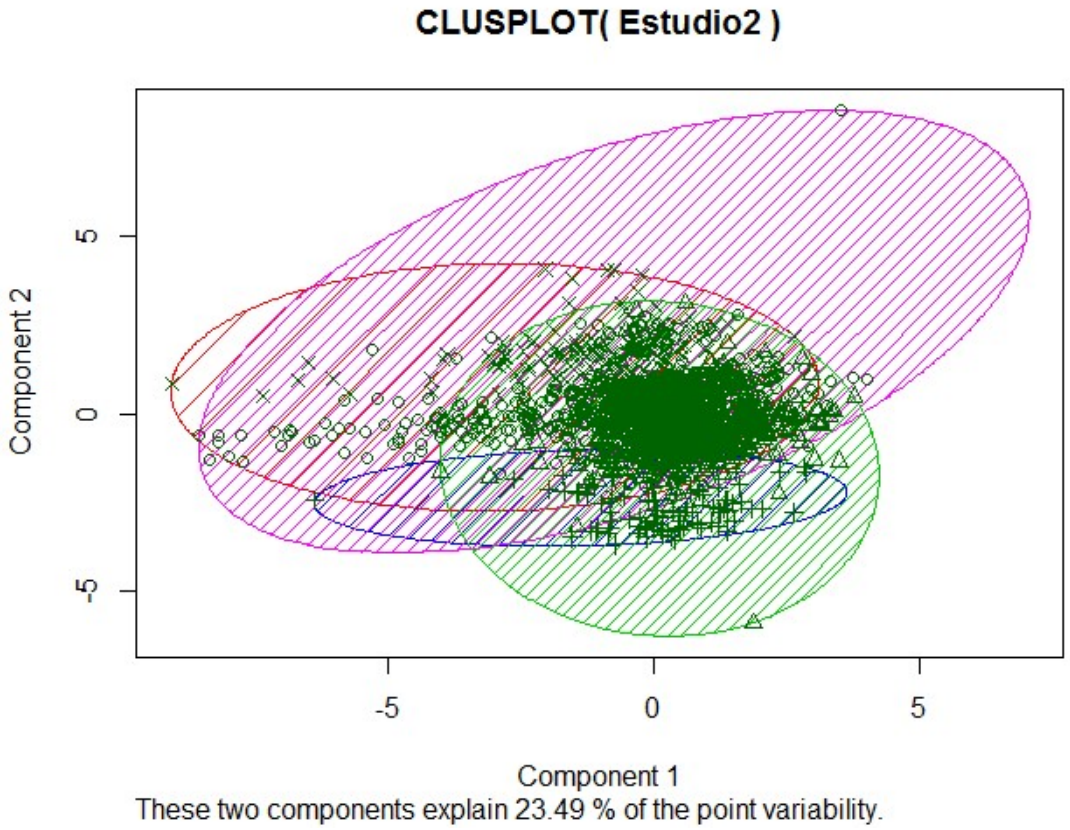

Figura 4-10: Gráfico de componentes principales de 4-medoids clustering.

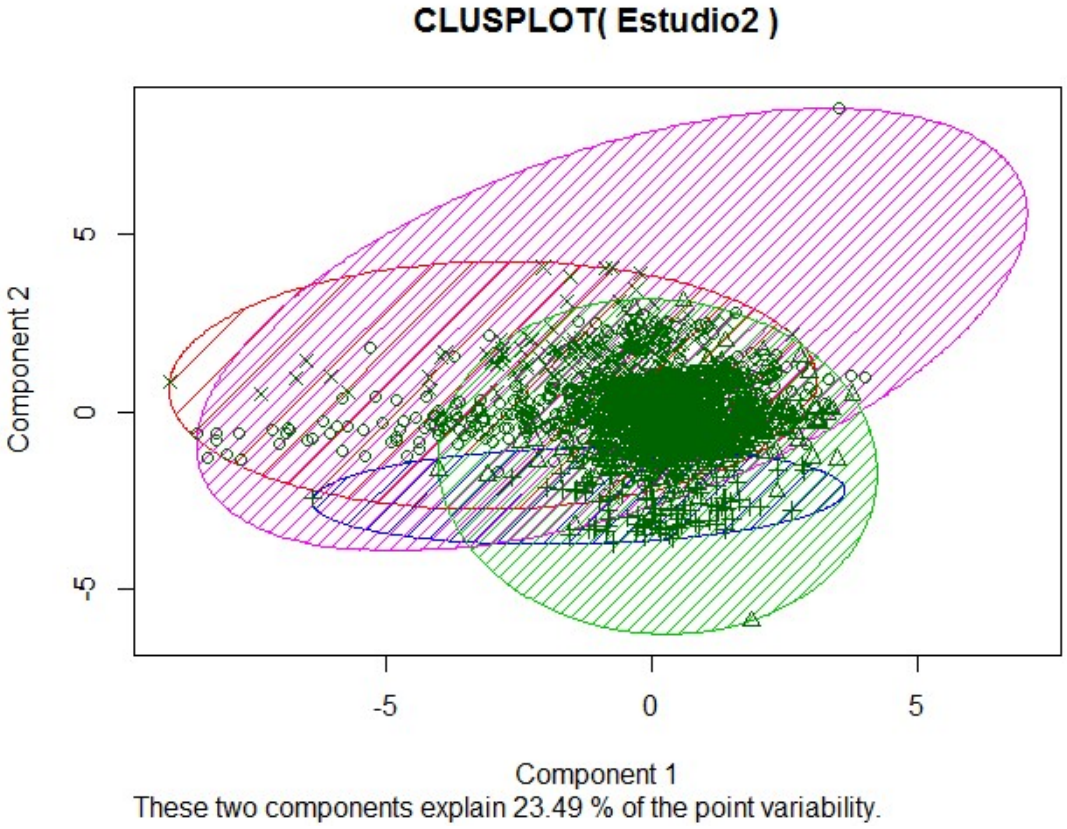

Figura 4-11: Gráfico de componentes principales de 5-medoids clustering.

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS

Lamentablemente, la información contenida en los clusters no entregan información interesante alguna con respecto a la relación existente entre las variables de la base de datos proporcionada. Esto se puede apreciar en la tabla de resumen de cada uno de los clusters generados, en donde a simple vista los criterios para separar un cluster de otro se encuentra fuertemente influenciado por el sexo del individuo y el resto de las variables binarias, dejando completamente de lado los niveles de concentración hormonal en la sangre, las cuales corresponden a las variables claves que permitirían diferenciar un paciente que presenta hipertiroidismo del que no. Por esta misma razón, y debido a que se cree que el problema se debe a la combinación de diferentes tipos de variable en el estudio, se propone realizar otro proceso de clustering. Esta vez dejando de lado las variables de tipo categórica para facilitar el análisis y establecer relaciones únicamente entre las variables numéricas.

Sin dar mayores detalles, y utilizando un k = 5 como el número de conjuntos óptimos, se obtienen los siguientes resultados:

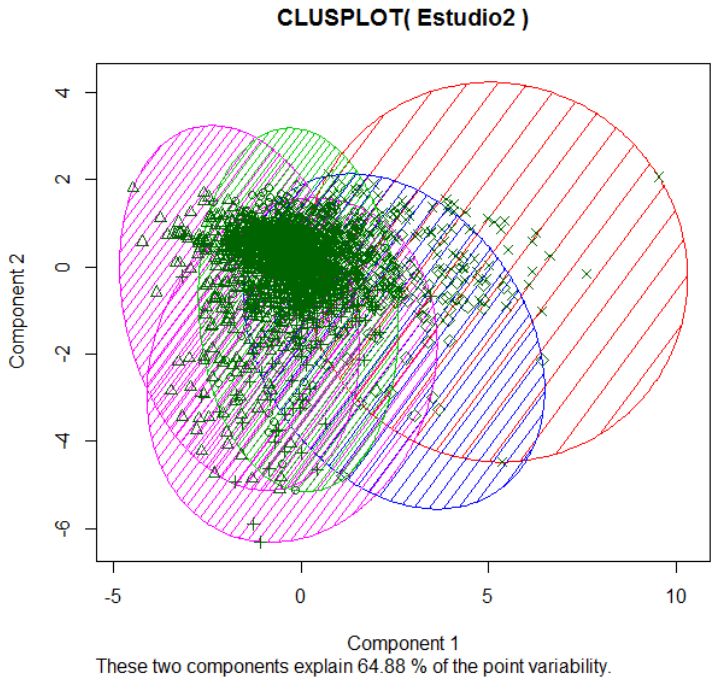

Figura 5-1: Gráfico basado en componentes principales solo con las variables númericas

Tabla 5.1: Resumen de la información contenida por los centroides de cada grupo

Grupo	age	TSH	T3	TT4	T4U
1	67.98876	1.7391386	1.902903	114.82240	0.9936891
2	63.97708	2.6997604	1.623854	80.71042	0.8915917
3	33.11850	2.0520478	2.054678	97.79834	0.9717048
4	46.59302	0.5095349	3.562791	209.24419	1.2733721
5	42.90970	1.4550334	2.485284	144.31773	1.1348161

CAPÍTULO 6. CONCLUSIÓN

Debido a la inexperiencia con respecto a la manipulación de datos de diferentes tipos, en especial de la combinación de datos de tipo numérico y categórico, se considera que a partir del desarrollo de la experiencia no se obtuvieron resultados satisfactorios teniendo en cuenta lo obtenido tras aplicar el algoritmo de clustering k-medoids basado en la distancia de Gower. Tras esta situación, se planteó la posibilidad de transformar las variables de tipo numérico a categórico bajo un criterio definido. Sin embargo, no existe mucha documentación con respecto a los algoritmos de clustering que permite trabajar con este tipo de variables (por ejemplo: k-modes), por lo que no se contaba con los conocimientos necesarios para llevarlo a cabo. También se descartó la posibilidad de transformar las variables binarias a numericas para poder utilizar el algoritmo de clustering k-means, el cual se entiende como el más sencillo. Sin embargo, de la literatura se sabe que esto no es conveniente. Finalmente, se decidió en utilizar únicamente las variables de tipo numérica para llevar a cabo un análisis simple mediante los clusters obtenidos a partir del algoritmo de las k-means. Sin embargo, este análisis no aportó mayor información de lo que se conocía de antemano.

Se piensa que las complicaciones para la conformación de clusters se debe a la poca cantidad de registros en la base de datos a los cuales se les ha diagnosticado la enfermedad del hipertiroidismo (46 de 1880 registros lo tienen).

Por otro lado, y con respecto a la parte del preprocesamiento, se considera que no es buena idea filtrar datos de acuerdo a un rango de valores arbitrarios, aún cuando estos fueron calculados en base a los rangos normales que se espera tener de cada una de las variables. Sin embargo, al desconocer metodologías para la limpieza de datos y al no tener conocimientos avanzados con respecto a la medicina, era la única opción que se tenía en el momento para poder continuar con la experiencia de laboratorio y presentar resultados.

En general la experiencia ayudó a afinar el conocimiento obtenido en la cátedra sobre pre-procesamiento y agrupamiento. En específico, se pudo comprobar que es muy difícil trabajar con datos en los cuales no se tiene experiencia ni conocimiento, y mucho más difícil es cuando no se tiene ninguna ayuda profesional. Además, uno de los grandes

problemas que se tuvo en esta experiencia fue al momento de definir un algoritmo para agrupar, presentaba una complicación tener una combinación de datos, y al tener tantas dimensiones no era posible visualizar de manera concreta el problema.

Por último, se espera que en futuras experiencias se puedan entregar resultados adecuados con respecto al análisis de datos de diferentes tipos. Ya que la idea es encontrar alguna relación interesante entre las variables presentes en el dataset, además de las ya conocidas relación entre los niveles hormonales relacionadas con la glándula tiroidea.

CAPÍTULO 7. REFERENCIAS

[1] Caparrini, F. and Work, W. (2017). Clustering por K-medias - Fernando Sancho Caparrini. [online] Cs.us.es. Desde: http://www.cs.us.es/ fsancho/?e=43 [Recuperado 25 abril de 2017].

[2] clustering?, H. (2017). How to use both binary and continuous variables together in clustering?. [online] Stats.stackexchange.com. Desde: https://stats.stackexchange.com/questions/130974/ [Recuperado 25 abril de 2017].

[3] Cs.us.es. (2017). Búsqueda de patrones: técnicas de clustering — Un Caso Práctico en Biología Molecular de Sistemas: Análisis de Redes de Coexpresión Génica 0 documentation. [online] Desde: https://www.cs.us.es/ fran/curso_unia/clustering.html [Recuperado 26 abril de 2017].

[4] Dabbling with Data. (2017). Clustering categorical data with R. [online] Desde: https://dabblingwithdata.wordpress.com/2016/10/10/clustering-categorical-data-with-r/ [Recuperado 28 abril de 2017].

[5] Datanalytics.com. (2017). Reducción de la dimensionalidad con t-SNE – datanalytics. [online] Desde: https://www.datanalytics.com/2017/03/08/reduccion-de- la-dimensionalidad-con-t-sne/ [Recuperado 29 abril de 2017].

[6] R-bloggers. (2017). Clustering Mixed Data Types in R. [online] Desde: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/ [Recuperado 29 abril de 2017].