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

INFORME EXPERIENCIA 2 ANÁLISIS DE DATOS

AGRUPAMIENTO CON ALGORITMO K-MEDIAS

Integrantes:

Marcela Rivera Castro

Kevin Alvarez

Profesor:

Max Chacón

Ayudante:

Adolfo Guzmán

Santiago - Chile

18 de abril de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

Antiguamente cerca de los años 1880, se tenía completo desconocimiento a cerca de la glándula tiroides y sus funciones. No existía advertencia de lo importante que es para el organismo humano. Entre los conocimientos que se manejaban en aquel momento, se sabía del cretinismo y los casos de mixedema del adulto de Gull, sin embargo no se conocía su origen tiroideo.

En 1883 el cirujano Teodoro Emilio Kocher, realizó una publicación sobre las consecuencias funestas de la tiroidectomía radical. Resolviendo en 1888 que cretinismo, mixedema y "caquexia" postiroidectomía eran síndromes estrechamente relacionados, si no idénticos, y se debían los tres a la pérdida de la función tiroidea.(Aguirre, 2002)

En la actualidad se conocen muchas de las enfermedades relacionadas con la tiroides. En el presente informe se abordará el hipotiroidismo. Hipotiroidismo significa "poca hormona tiroidea". Ocurre cuando la glándula tiroidea esta dañada y no es capaz de producir las hormonas tiroideas suficientes para mantener el metabolismo del cuerpo normal. El exceso de TSH puede causar que la glándula tiroidea aumente de tamaño lo que se llama bocio. Existen otras causas de hipotiroidismo como las tiroiditis autoinmunes o virales que pueden generar el mismo cuadro final pero sin bocio.(de Endocrinología Facultad de Medicina UC, s.f.)

1.2 ORGANIZACIÓN DEL DOCUMENTO

El documento consta de cuatro secciones principales: descripción de problema, donde se aborda la problemática y la base de datos que es utilizada, pre-procesamiento de datos (limpieza de datos), obtención de cluster y su respectivo análisis de resultados y finalmente conclusiones, donde se indica el aprendizaje obtenido a partir del desarrollo de la experiencia.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

- Para el estudio de los datos se utilizará el programa R studio.
- La base de datos a utilizar es: allhypo.data y allhypo.names

CAPÍTULO 2. MARCO TEÓRICO

2.1 CLUSTERING

Clustering es una técnica de minería de datos (data mining) dentro de la disciplina de Inteligencia Artificial que identifica de forma automática agrupaciones o clústeres de elementos de acuerdo a una medida de similitud entre ellos. El objetivo fundamental de las técnicas de clustering consiste en identificar grupos o clústeres de elementos tal que:

- La similitud media entre elementos del mismo clúster sea alta. Similitud intra-clúster alta.
- La similitud media entre elementos de distintos clústeres sea baja. Similitud inter-clúster baja.

La identificación de clústeres o grupos de elementos se basa en una medida de similitud. Diferentes medidas de similitud dan lugar a diferentes clústeres. (Romero-Campero, 2013)

Las aplicaciones del clustering son diversas, es posible utilizarla en áreas tan distintas como lo son la biologıa y el marketing. Existen diversos algoritmos de agrupamiento, entre ellos se encuentra K-means o K-medias, el cual será utilizado en la presente experiencia.

2.2 ALGORITMO K-MEANS

Es un algoritmo de aprendizaje no supervisado, el cual sirve para reconocer formas, pre-procesar datos, entre otros. Este algoritmo aproxima los grupos (o clusters) de manera iterativa. La cantidad de grupos que se formarán es prefijado, por lo que es una variable de entrada para este algoritmo, el cual es definido por la persona que realiza el estudio. El algoritmo de K-means se presenta a continuación:

1. Especificar el número deseado de clusters K.
2. Asignar aleatoriamente cada punto de datos a un grupo.
3. Calcular los centroides del grupo. 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. Reasignar cada punto al centroide más cercano del clúster.

5. Volver a calcular los centroides del clúster.

6. Se repiten los pasos 4 y 5 hasta que no hayan mejoras posibles. Cuando no hayá más conmutación de puntos de datos entre dos clusters para dos repeticiones sucesivas, se marcará la terminación del algoritmo. (KAUSHIK, 2016)

2.3 DISTANCIAS UTILIZADAS

2.3.1. Distancia de Gower

Se utiliza cuando 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). Su fórmula es la siguiente:

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

(Grané, s.f.)

2.4 MÉTODO DE SILUETAS

El objetivo de este índice es identificar el número óptimo de agrupamientos, para esto se calcula la media del coeficiente de silueta de cada objeto de la muestra. Mientras más grande, mejor es la distribución de conglomerados. (Alvarez, 2013)

2.5 T-SNE

Es una técnica para reducir la dimensionalidad de conjuntos de datos, es una alternativa moderna a MDS o PCA. La técnica se puede implementar a través de aproximaciones BarnesHut, lo que le permite ser aplicado en grandes conjuntos de datos del mundo real. (Molina, 2014)

CAPÍTULO 3. PRE-PROCESAMIENTO

Dado a que en el dataset utilizado para el estudio existen variedad de datos tomados de forma aleatoria, es necesario realizar un análisis previo para verificar que los datos sean consistentes y no hayan anomalías que afecten al resultado del estudio.

3.1 ELIMINACIÓN DE REGISTROS NA

Dentro de los datos del dataset puede ocurrir que algunos de éstos datos sean nulos para ciertas variables, es decir, que no fueron tomados o que posteriormente fueron eliminados, éstos datos son representados en el dataset como "?" y es a lo que se llama registro NA o desconocido.

La eliminación de los registros NA se realiza por completo en el set de datos, es decir, si existe un registro NA para una variable en una observación (o fila), ésta se elimina ya que hay un dato faltante, de ésta forma se evitan resultados o conclusiones anómalas en el estudio, o posibles errores que puedan ocurrir por la existencia de estos registros.

3.2 ELIMINACIÓN DE VARIABLES

La eliminación de variables puede darse debido a razones variadas, pero específicas, para este caso el criterio utilizado, es que la variable no contenga ningún dato, que contenga solo registros desconocidos (NA), que el dato sea anómalo (concluido de acuerdo a una métrica), o que para el objetivo del estudio ésta no entregue mucha información.

3.2.1 TBG

Esta variable corresponde a el resultado de una medición de TBG, el cual es un exámen para medir el nivel de una proteína que lleva hormona tiroidea a través de la sangre, dentro del dataset estás mediciones no fueron realizadas, ya que todos los datos son registros desconocidos, dado esto, la variable no entrega información y por ende se puede eliminar.

3.2.2 Variables de medición

En el dataset, existen variables booleanas para señalar si un examen de hormona fue realizado o no, esta variable es usada para filtrar los registros NA, dado a que si es falsa, significa que el dato es desconocido, éstas variables son:

- **TSH measured:** Medición de hormona TSH realizada.

- **T3 measured:** Medición de hormona T3 realizada.
- **TT4 measured:** Medición de hormona TT4 realizada.
- **T4U measured:** Medición T4U realizada.
- **FTI measured:** Medición FTI realizada, ésta indica la cantidad de T4 libre.

Estas variables por si mismas y luego del filtro, no entregan información útil para el objetivo del estudio, por lo que se tornan innecesarias y pueden ser eliminadas.

3.2.3 Fuente de referencia

Esta variable indica la fuente de la cual se obtuvo los datos de un sujeto (u observación) en particular, dentro de los objetivos del estudio, es decir, el agrupamiento de los sujetos, esta variable no entrega información útil que permita realizar tal agrupamiento, por lo que se puede eliminar.

3.2.4 Variable de clasificación

Esta variable indica la clasificación de cada sujeto del dataset, que puede ser el resultado de una agrupación y posterior clasificación mediante algún método, para efectos del estudio actual, no entrega información relevante, dado que no constituye un criterio fuerte para el agrupamiento, por lo que puede ser eliminada.

Cabe destacar que junto a la variable se encontraba el id del sujeto respectivo, este dato no entrega información por lo que es dispensable y no fue tomado en cuenta.

3.3 SELECCIÓN DE DATOS

Para el caso de las variables continuas tales como: TSH, T3, TT4, T4U y FTI, se ha decidido establecer rangos de valores aceptables, esto basándose en los datos que fueron expuestos como normales en la experiencia anterior.

Para ello, se ha decidido que el rango máximo para cada una de las variables, se basa en calcular la distancia entre el valor mínimo y máximo, multiplicarlo por 2 y finalmente sumarle el valor máximo. Este procedimiento entrega la cota superior.

En el caso de la cota inferior, se ha decidido utilizar el valor 0 ya que al aplicar el mismo criterio del máximo, se obtienen valores negativos.

Por lo tanto los nuevos rangos son:

Tabla 3.1: Rangos para las variables continuas.

Hormona	**Rango normal**	**Rango a utilizar**
Edad	0 - 117	0 - 100
TSH	0.4 - 4.0	0 - 11.2
T3	1.07 - 3.37	0 - 7.97
TT4	64 - 164	0 - 364
T4U	0.7 - 1.8	0 - 4
FTI	33.108 - 135.191	205.966

Una vez hecho el pre-procesamiento de datos, se obtiene que el total de registros de sujetos estudiados, es de 1819 observaciones, lo cual equivale al $64,96\,\%$ del total. Además el dataset del estudio queda con un total de 21 variables, las cuales son:

1. age (variable continua)
2. sex (variable categórica - binaria)
3. on thyroxine (variable categórica - binaria)
4. query on thyroxine (variable categórica - binaria)
5. on antithyroid medication (variable categórica - binaria)
6. sick (variable categórica - binaria)
7. pregnant (variable categórica - binaria)
8. thyroid surgery (variable categórica - binaria)
9. I131 treatment (variable categórica - binaria)
10. query hypothyroid (variable categórica - binaria)
11. query hyperthyroid (variable categórica - binaria)

12. lithium (variable categórica - binaria)
13. goitre (variable categórica - binaria)
14. tumor (variable categórica - binaria)
15. hypopituitary (variable categórica - binaria)
16. psych (variable categórica - binaria)
17. TSH (variable continua)
18. T3 (variable continua)
19. TT4 (variable continua)
20. T4U (variable continua)
21. FTI (variable continua)

CAPÍTULO 4. OBTENCIÓN DE CLUSTER

El agrupamiento de datos se realiza con el algoritmo k-medias, donde este recibe como parámetros el k, que es la cantidad de grupos a formar.

4.1 DISTANCIA

Para calcular la dissimilaridad o proximidad de las observaciones, se hace uso de la distancia de Gower, esto es debido a que en el dataset existen variables de dos tipos, binarias y continuas, por lo que no se puede usar distancias que solo puedan ser calculadas con variables continuas, como la euclideana, o solo con variables binarias, como el coeficiente de Jacard.

La distancia de Gower permite realizar el cálculo de distancia tomando en cuenta variables categóricas (en este caso binarias) y continuas dentro del mismo set de datos, es por ello que es la medida indicada para el problema, las distancias son calculadas a través de una matriz de dissimilaridad o distancia, que posteriormente será usado como parámetro de entrada para realizar el clustering con el algoritmo k-medias.

4.2 NÚMERO DE GRUPOS

Antes de realizar el agrupamiento, es necesario definir la cantidad de grupos a formar, que es el parámetro del algoritmo k-medias, una forma es probar diferente número de grupos hasta observar que éstos estén bien definidos, el problema de este método es que la responsabilidad de medir la calidad de los grupos recae en el observador que está analizando los grupos y no tiene una medida numérica de la cual concluir que realmente es un buen número de grupos. Para solucionar este problema se hace uso del método de las siluetas, con el cual se calcula un coeficiente que ayuda a definir cual número de grupo es mejor, mientras mas alto el coeficiente, mejor es la cantidad de grupos.

Para esta instancia se toma como máximo 15 grupos y se calcula el coeficiente para la ejecución del algoritmo desde 2 grupos hasta los 15, al realizar la ejecución y el cálculo de los coeficientes se obtuvieron los resultados mostrados en la figura 4-1:

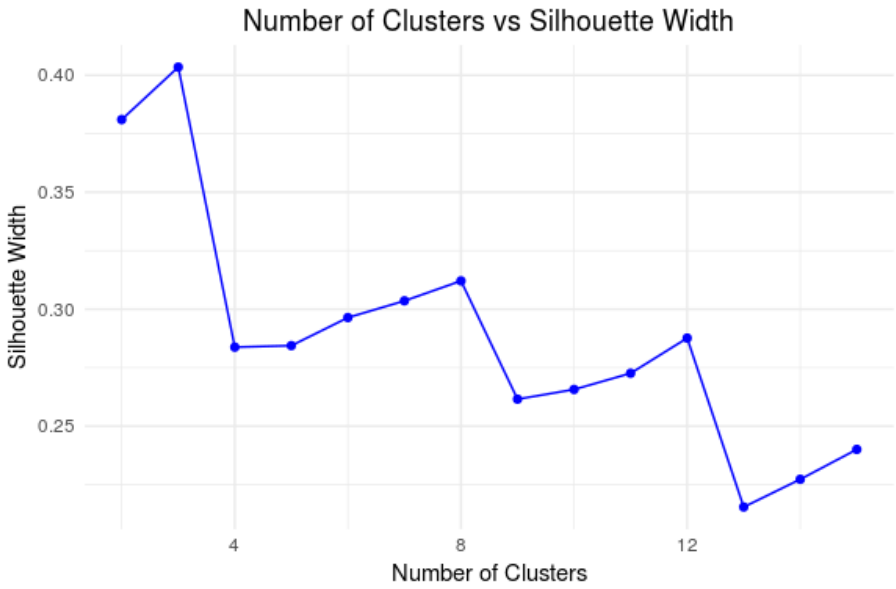

Figura 4-1: Resultados de coeficientes con el método de la silueta

De acuerdo a lo observado en el gráfico mostrado en la figura 4-1, el mejor número de grupos para el problema es 3, entonces el algoritmo se ejecuta con un k=3.

4.3 CLUSTERING

Ya con la matriz de distancias calculada y el número de grupos definido, se puede realizar el agrupamiento. Dado a que en el set de datos se encuentran 21 variables, no es posible para el ojo humano observar gráficamente las observaciones y poder identificar los grupos, es por ello que se hace uso de t-SNE, que es una técnica de reducción de dimensionalidad para llevar los grupos a un plano bidimensional y así poder observar de forma aproximada los grupos formados, en la figura 4-2 se pueden observar los resultados.

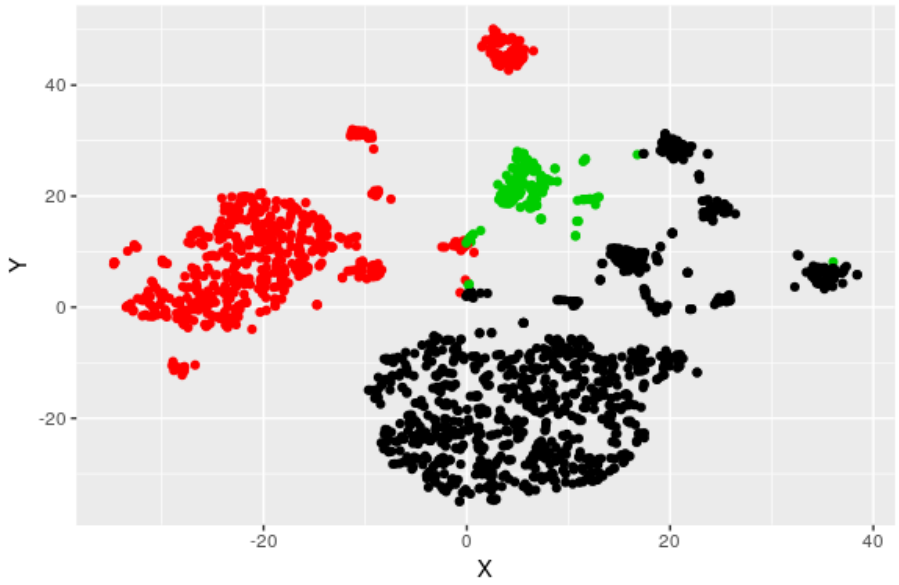

Figura 4-2: Agrupamiento en plano bidimensional

CAPÍTULO 5. ANÁLISIS DE RESULTADOS

Los resultados del agrupamiento por si solo no entregan mucha información, es por ello que se contrastan con el set de datos original que se usó, para ver la pertenencia de cada observación a su cluster determinado, con ello se pueden realizar distintos tipos de análisis.

5.1 ANÁLISIS ESTADÍSTICO

Es interesante conocer en que se diferencia cada cluster de otros, para ello, se puede hacer uso de la estadística descriptiva y calcular medidas de tendencia central o proporciones, con las cuales se puede observar el comportamiento de cada variable en cada cluster.

5.1.1 Cluster 1

De acuerdo a lo que es posible observar en las tablas 5.1 y 5.2, en el cluster 1 se agrupó a solo sujetos con sexo femenino, y dentro del grupo se pueden observar características como que están en espera de tratamiento de hipotiroidismo o hipertiroidismo, etc. Por otro lado las variables continuas presentan valores normales. En definitiva, el agrupamiento está claramente de acuerdo a las variables binarias.

Tabla 5.1: Proporción variables binarias Cluster 1.

Variable	**F/f**	**M/t**
Sex	1051	0
On Thyroxine	1051	0
Query On Thyroxine	1043	8
On Antithyroid Medication	1031	20
Sick	995	56
Pregnant	1025	26
Thyroid Surgery	1036	15
I131 Treatment	1036	15
Query Hypothyroid	999	52
Query Hyperthyroid	971	80
Lithium	1043	8
Goitre	1044	7
Tumor	1018	33
Hypopituitary	1051	0
Psych	995	56

Tabla 5.2: Medidas variables continuas Cluster 1.

Variable	**Min**	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**
Age	2	36	56	53.5	70	93
TSH	0.005	0.500	1.300	1.896	2.400	11.100
T3	0.050	1.600	2.000	2.044	2.400	7.000
TT4	19.00	91.00	106.0	110.8	126.0	301.0
T4U	0.310	0.900	1.000	1.029	1.110	2.030
FTI	17.0	93.0	105.0	108.8	121.0	204.0

5.1.2 Cluster 2

De acuerdo a lo que es posible observar en las tablas 5.3 y 5.4, en el cluster 2 se agrupó solo sujetos con sexo masculino, y se pueden observar una menor cantidad de observaciones con las características que representan las variables. Al igual que el cluster anterior,

las variables continuas no presentan mayor información, ya que se encuentran dentro de los valores normales y similares a los del cluster anterior, es decir, su diferenciación no se encuentra en estos datos.

Tabla 5.3: Proporción variables binarias Cluster 2.

Variable	**F/f**	**M/t**
Sex	0	626
On Thyroxine	604	22
Query On Thyroxine	619	7
On Antithyroid Medication	622	4
Sick	599	27
Pregnant	626	0
Thyroid Surgery	624	2
I131 Treatment	624	2
Query Hypothyroid	610	16
Query Hyperthyroid	607	19
Lithium	625	1
Goitre	619	7
Tumor	623	3
Hypopituitary	625	1
Psych	556	70

Tabla 5.4: Medidas variables continuas Cluster 2.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Age	1	39	55	52.76	67	94
TSH	0.005	0.600	1.300	1.687	2.200	11.000
T3	0.200	1.500	2.000	1.934	2.300	7.100
TT4	38.0	86.0	99.0	101.2	115.0	225.0
T4U	0.4100	0.8300	0.9200	0.9309	1.0200	1.6800
FTI	41.0	96.0	108.0	109.3	121.0	205.0

5.1.3 Cluster 3

De acuerdo a lo que es posible observar en las tablas 5.5 y 5.6, en el cluster 3, a diferencia de los dos anteriores el sexo de los sujetos es variado, la característica principal de este cluster se encuentra en que los sujetos están en tratamiento de tiroxina, que es usado para casos de hipotiroidismo e hipertiroidismo para equilibrar la cantidad de hormona tiroidea. Al igual que los clusters anteriores, los valores de las variables continuas se encuentran en valores normales, y similares entre los tres.

Tabla 5.5: Proporción variables binarias Cluster 3.

Variable	**F/f**	**M/t**
Sex	131	11
On Thyroxine	0	142
Query On Thyroxine	140	2
On Antithyroid Medication	141	1
Sick	138	4
Pregnant	136	6
Thyroid Surgery	138	4
I131 Treatment	138	4
Query Hypothyroid	129	13
Query Hyperthyroid	136	6
Lithium	140	2
Goitre	141	1
Tumor	142	0
Hypopituitary	142	0
Psych	141	1

Tabla 5.6: Medidas variables continuas Cluster 3.

Variable	**Min**	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**
Age	12	39	55.5	52.82	65	84
TSH	0.00500	0.06625	0.30000	1.41690	1.95000	9.70000
T3	0.30	1.70	2.15	2.19	2.50	6.70
TT4	37.0	112.0	132.5	136.6	160.0	289.0
T4U	0.720	0.920	1.015	1.057	1.117	1.800
FTI	51.0	109.2	128.0	129.6	148.8	197.0

5.2 CLUSTERING CON VARIABLES CONTINUAS

Al analizar los cluster obtenidos anteriormente, se puede inferir que no se obtienen resultados que entreguen información relevante y de manera clara, esto porque a simple vista es difícil reconocer el patrón o las variables que son determinantes a la hora de realizar el agrupamiento.

Por otro lado, al investigar acerca de esta enfermedad fue posible detectar que en muchos estudios y diagnósticos, las variables continuas son muy importantes, ya que con éstos índices se reflejan de mejor forma cuando un paciente puede estar en riesgo de sufrir hipotiroidismo. Es por esta razón que se ha decidido realizar un estudio de agrupamiento utilizando únicamente a las variables continuas.

Para lograrlo se deben repetir los pasos previamente explicados en el capítulo anterior, sin embargo se debe realizar un único cambio, el cual corresponde a la distancia a utilizar, en este caso se usa la distancia Euclidiana y no la de Gower. Esto ya que a diferencia del estudio anterior, ahora se trabaja únicamente con variables continuas. A continuación se presenta el agrupamiento final.

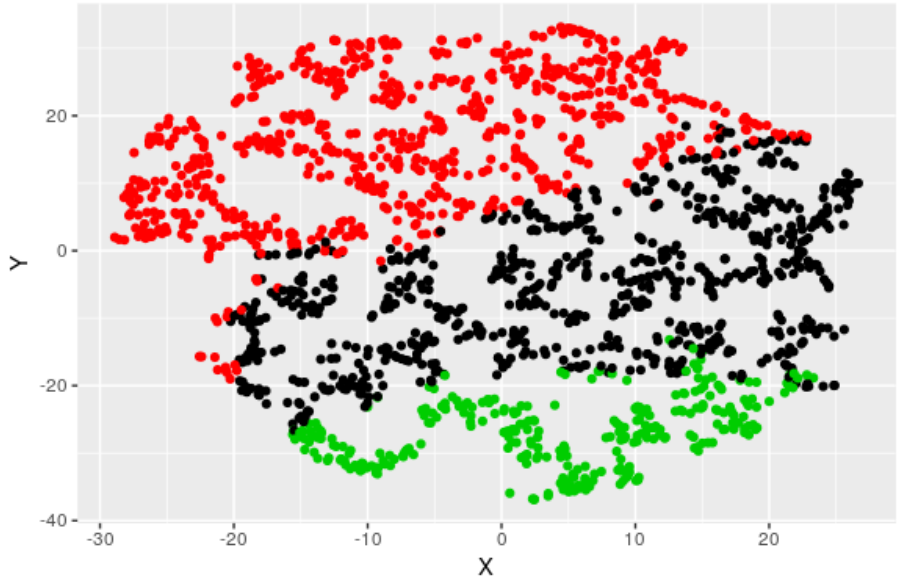

Figura 5-1: Agrupamiento en plano bidimensional utilizando variables continuas

Como se puede ver en la figura 5-1, el agrupamiento es más claro en comparación al caso anterior, esto se puede asegurar gracias a la distribución de las observaciones las

cuales se encuentran ordenadas, con muy pocas observaciones alejadas de sus respectivos grupos. No obstante, se presentan los detalles de cada cluster para así facilitar la comprensión de este nuevo agrupamiento.

5.2.1 Cluster 1

Tabla 5.7: Medidas variables continuas Cluster 1.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Age	1	39	57	54.01	70	94
TSH	0.005	0.600	1.300	1.694	2.200	11.000
T3	0.100	1.600	2000	2.023	2.400	7.000
TT4	50.0	105.0	113.0	114.1	122.0	157.0
T4U	0.310	0.880	0.990	0.9961	1.0800	1.700
FTI	76.0	108.0	115.0	116.3	124.0	165.0

5.2.2 Cluster 2

Tabla 5.8: Medidas variables continuas Cluster 2.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Age	1	38	55.00	53.33	69	93
TSH	0.005	0.7175	1.55	2.2066	2.700	11.100
T3	0.05	1.500	1.900	1.829	2.200	4.100
TT4	19.0	78.0	87.0	85.69	95.0	130.0
T4U	0.410	0.860	0.950	0.9547	1.0500	1.830
FTI	17.0	83.0	93.0	90.76	100.0	137.0

5.2.3 Cluster 3

Tabla 5.9: Medidas variables continuas Cluster 3.

Variable	Min	1st Q	Median	Mean	3rd Q	Max
Age	2	34	51.00	50.81	67	87
TSH	0.005	0.6125	0.3	0.95808	1.525	11.00
T3	0.5	1.900	2.300	2.478	2.900	7.100
TT4	105.0	139.0	152.0	157.7	169.0	301.0
T4U	0.610	0.920	1.030	1.109	1.200	2.030
FTI	92.0	131.0	145.0	146.3	163.0	205.0

Al observar los nuevos cluster, se puede ver que el valor de las medidas para cada variable, ha cambiado. Esto se debe básicamente a que los nuevos grupos están compuestos por un nuevo conjunto de observaciones, lo que conlleva a estas pequeñas variaciones en los grupos finales.

Por último, a partir de la figura 5-1 se puede observar como los grupos se pueden dividir según la tendencia al hipotiroidismo, de acuerdo únicamente a los niveles hormonales.

CAPÍTULO 6. CONCLUSIONES

En esta experiencia fue posible aprender de la aplicación del agrupamiento de datos mediante el uso del algoritmo k-medias, sin embargo, a pesar de la correcta implementación del algoritmo, el análisis de datos ha sido difícil debido a la complejidad para leer y comprender los datos presentados ante un agrupamiento en el plano bidimensional.

Entre las principales dificultades, se tiene el poco conocimiento y especialización del tema, ya que al no tener bastos conocimientos de salud y específicamente de esta enfermedad, se hace difícil el manejo de datos, especialmente para establecer límites y decidir qué es un dato normal y que es un dato anómalo. Esto es un proceso meticuloso que depende exclusivamente de la especialización de quién hace el estudio.

Para el caso del pre-procesamiento de datos, se puede decir que su principal dificultad también se encuentra en la especialización del tema a estudiar, esto porque es necesario tener amplios conocimientos para poder decidir de manera correcta y eficiente, cuáles son los datos que realmente aportan información relevante, cuáles no condicionan el agrupamiento, entre otros. Por otro lado, es importante mencionar que este paso sigue siendo importante para realizar un buen análisis de datos, esto ya que se requiere de un estudio sólido que utilice información de calidad para no obtener conclusiones erróneas, ocasionadas por un mal empleo con los datos.

De acuerdo a los resultados que fueron obtenidos y posteriormente analizados, en primera instancia el estudio realizado trajo como resultados que con los datos, el algoritmo realizó el agrupamiento de acuerdo al sexo de los sujetos, tomando en cuenta características que fueron representadas por las variables binarias, 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 del cuadro de un paciente, en este caso, hipotiroidismo.

Por lo mencionado anteriormente, es que se decide realizar un segundo agrupamiento con las mismas características del anterior, utilizando únicamente las variables continuas, esto con el fin de poder analizar como son agrupadas las observaciones. Con esto, se obtiene como resultado un agrupamiento más limpio, donde en un gráfico bidimensional se

puede claramente identificar los grupos, se tiene que las medidas obtenidas son concordantes con lo conocido, se puede observar una relación entre las hormonas T3, TT4 y TSH, dando como resultado que los grupos tengan diferencias en éstas, donde un grupo tiene menores niveles hormonales y va ascendiendo hasta un máximo, por lo que se podría definir los grupos como tendencia del hipotiroidismo, de acuerdo a los niveles hormonales.

Por último, haciendo una breve comparación entre la experiencia anterior y la actual, se puede deducir que el actual estudio entrega mayor información, ya que a diferencia de la experiencia pasada, ahora fue posible agrupar las observaciones en base a las similitudes entre las medidas de cada variable. Esto puede ser de gran ayuda para estudios que necesitan diferenciar las características que se encuentran dentro de una muestra, donde las observaciones de éstas pueden ser agrupadas y posteriormente clasificadas.

CAPÍTULO 7. BIBLIOGRAFÍA

Aguirre, C. P. (2002). Emil Theodor Kocher (1841-1917). Recuperado desde http://www.historiadelamedicina.org/kocher.html

Alvarez, A. G. (2013). Patrones de multimorbilidad mediante Análisis Clúster con R. Recuperado desde http://masteres.ugr.es/moea/pages/tfm-1213/tfm_garciaalvarezarturo/!

de Endocrinología Facultad de Medicina UC, D. (s.f.). HIPOTIROIDISMO. Recuperado desde http://redsalud.uc.cl/ucchristus/VidaSaludable/Glosario/H/hipotiroidismo.act

del Cáncer de los Institutos Nacionales de la Salud de EE.UU., I. N. (2017). Diccionario de cáncer. Recuperado desde https://www.cancer.gov/espanol/publicaciones/diccionario?cdrid=533434

Grané, A. (s.f.). Distancias estadísticas y Escalado Multidimensional (Análisis de Coordenadas Principales). Recuperado desde http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf

KAUSHIK, S. (2016). Introducción al Clustering y diferentes métodos de agrupación. Recuperado desde https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

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

Romero-Campero, F. J. (2013). Búsqueda de patrones: técnicas de clustering. Recuperado desde https://www.cs.us.es/~fran/curso_unia/clustering.html

CAPÍTULO 8. ANEXO: CÓDIGO EN R

```
library("cluster")
library("ggplot2")
library("Rtsne")

rawdata <- read.csv("allhypo.data", header = FALSE, sep = ",",
    stringsAsFactors = FALSE)
colnames(rawdata) <- 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", "clasification|id")

#delete id from clasification|id column
d <- c()
for(i in rawdata$`clasification|id`){
  d <- c(d, strsplit(i, ".", fixed = TRUE)[[1]][1])
}
colnames(rawdata)[30] <- "clasification"
rawdata$clasification <- d

#Data pre-processing
#delete NA values
data <- rawdata[(rawdata$age!="?" & rawdata$sex!="?" & rawdata$on.
    thyroxine!="?" & rawdata$query.on.thyroxine!="?" & rawdata$on.
    antithyroid.medication!="?" & rawdata$sick!="?" & rawdata$pregnant!=
    "?" & rawdata$thyroid.surgery!="?" & rawdata$I131.treatment!="?" &
    rawdata$query.hypothyroid!="?" & rawdata$query.hyperthyroid!="?" &
    rawdata$lithium!="?" & rawdata$goitre!="?" & rawdata$tumor!="?" &
    rawdata$hypopituitary!="?" & rawdata$psych!="?" & rawdata$TSH.
    measured!="f" & rawdata$T3.measured!="f" & rawdata$TT4.measured!="f"
```

```
    & rawdata$T4U.measured!="f" & rawdata$FTI.measured!="f"),]
#Delete variable TBG
data$TBG.measured <- NULL
data$TBG <- NULL
#Delete measuring variables

data$TSH.measured <- NULL
data$T3.measured <- NULL
data$TT4.measured <- NULL
data$T4U.measured <- NULL
data$FTI.measured <- NULL
#Delete referral.source variable
data$referral.source <- NULL
#Delete clasification variable
data$clasification <- NULL

#Data format transform
#Nominal variables
data$sex <- as.factor(data$sex)
data$on.thyroxine <- as.factor(data$on.thyroxine)
data$query.on.thyroxine <- as.factor(data$query.on.thyroxine)
data$on.antithyroid.medication <- as.factor(data$on.antithyroid.
    medication)
data$sick <- as.factor(data$sick)
data$pregnant <- as.factor(data$pregnant)
data$thyroid.surgery <- as.factor(data$thyroid.surgery)
data$I131.treatment <- as.factor(data$thyroid.surgery)
data$query.hypothyroid <- as.factor(data$query.hypothyroid)
data$query.hyperthyroid <- as.factor(data$query.hyperthyroid)
data$lithium <- as.factor(data$lithium)
data$goitre <- as.factor(data$goitre)
data$tumor <- as.factor(data$tumor)
data$hypopituitary <- as.factor(data$hypopituitary)
data$psych <- as.factor(data$psych)
```

```
#Continuous variables
data$age <- as.numeric(data$age)
data$TSH <- as.numeric(data$TSH)
data$T3 <- as.numeric(data$T3)
data$TT4 <- as.numeric(data$TT4)
data$T4U <- as.numeric(data$T4U)
data$FTI <- as.numeric(data$FTI)

# Data selection - set max value for detection of data anomalies
TSHMAX <- 4.0 + abs(0.4 - 4.0) * 2
T3MAX <-  3.37 + abs(1.07 - 3.37) * 2
TT4MAX <- 164 + abs(64 - 164) * 2
T4UMAX <- 1.8 + abs(0.7 - 1.8) * 2
FTIMAX <- 1.8 + abs(33.108 - 135.191) * 2
data <- subset(data, (age <= 100) & (TSH <= TSHMAX) & (T3 <= T3MAX) & (
    TT4 <= TT4MAX) & (T4U <= T4UMAX) & (FTI <= FTIMAX))

#Calculating Dissimilarity with Gower Distance
data_dist <- daisy(data, metric = "gower")

#Calculating optimal number of groups with Silhouette width
sil <- c()
for (i in 2:15){
  fit <- pam(data_dist, diss = TRUE, k = i)
  sil[i] <- fit$silinfo$avg.width
}

clust_num_plot <- ggplot(data.frame(clust_num = 2:15, sil_width = sil
    [2:15]), aes(x = clust_num, y = sil_width))+labs(x = "Number of
    Clusters", y = "Silhouette Width")+geom_line(color = "blue")+geom_
    point(color = "blue")+ggtitle("Number of Clusters vs Silhouette
    Width")+theme_minimal()+theme(plot.title = element_text(hjust = 0.5)
    )
```

```
#Clustering [k = 3]
data_cluster <- pam(data_dist, diss = TRUE, k = 3)

#t-SNE plot
tsne <- Rtsne(data_dist, is_distance = TRUE)
plot_groups <- ggplot(data.frame(tsne$Y), aes(x = X1, y = X2))+labs(x =
    "X", y = "Y")+geom_point(color = factor(data_cluster$clustering))

#Adding a new column with the cluster number
data["cluster"] <- data_cluster$clustering

#cluster summary
sum_c1 <- summary(data[data$cluster == 1, ])
sum_c2 <- summary(data[data$cluster == 2, ])
sum_c3 <- summary(data[data$cluster == 3, ])

#Clustering using only continuous variables
datac <- data[c(1, 17:21)]
#Calculating Dissimilarity with Euclidean Distance
datac_dist <- daisy(datac, metric = "euclidean")

#Clustering [k = 3]
datac_cluster <- pam(datac_dist, diss = TRUE, k = 3)

#t-SNE plot
tsne_c <- Rtsne(datac_dist, is_distance = TRUE)
plot_groupsc <- ggplot(data.frame(tsne_c$Y), aes(x = X1, y = X2))+labs(
    x = "X", y = "Y")+geom_point(color = factor(datac_cluster$clustering
    ))

#Adding a new column with the cluster number
datac["cluster"] <- datac_cluster$clustering
```

```
#cluster summary
sumc_c1 <- summary(datac[datac$cluster == 1, ])
sumc_c2 <- summary(datac[datac$cluster == 2, ])
sumc_c3 <- summary(datac[datac$cluster == 3, ])
```