UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

Análisis de Datos

Laboratorio 2: Clustering

Gustavo Hurtado

Patricia Melo

Profesor: Max Chacón

Ayudante: Javier Arredondo

Santiago – Chile

2020

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

En esta segunda experiencia se sigue trabajando con la base de datos $allhyper$ que contiene datos correspondientes al año 1987, y fueron obtenidos desde la página UCI Machine Learning Repository. Si bien se realizó un estudio preliminar de esta base de datos, conociendo un poco más sus datos y la relación entre ellos - que por cierto ayudó bastante para esta ocasión -, aún queda mucho de dónde se podría obtener información útil con respecto a estos datos y el hipertiroidismo.

Los objetivos de este laboratorio son en primer lugar, extraer conocimiento del problema asignado, que en este caso es hipertiroidismo haciendo uso de la base de datos $allhyper$, mediante el uso del software R, utilizando un algoritmo de Clustering y realizando su análisis respectivo. Por otro lado el comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido, para finalmente analizar por grupo e identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

El informe en un inicio consta de un marco teórico en el cual se darán las algunas definiciones a utilizar a lo largo del documento, mientras que luego se muestra un proceso de pre-procesamiento (limpieza) a los registros de la base de datos. Se continúan mostrando los fundamentos para generar un Clúster representativo para los datos, para después realizar el análisis ante el Clúster presetenado. Finalmente se tiene las conclusiones y referencias del laboratorio.

CAPÍTULO 2. MARCO TEÓRICO

2.1 CLUSTERING

Como menciona Moya (2016), Clustering se enmarca dentro del aprendizaje no supervisado; dado que mediante una entrada de la cual no se tienen las respuestas esperadas, se busca obtener información o realizar una acción dependiendo de esta entrada.

Teniendo en cuenta lo anterior, Clustering se puede definir como una tarea de agrupamiento de datos, mediante la información proporcionada por la misma entrada, y siendo separados en diferentes grupos llamados "$Clusters$" que permiten obtener nueva información sobre los datos.

2.2 K-MEDIAS Y PAM

El algoritmo de k-medias, es un método para generar agrupamientos de manera no supervisada (Clustering), que como indica Unioviedo (S.F.) es capaz de agrupar objetos en una cantidad de grupos establecida por el investigador según las propias características de los datos. Este agrupamiento es realizado minimizando la suma de las distancias entre cada uno de los objetos y el centroide de cada Cluster. Este centroide, es el promedio de todo el grupo.

Si bien PAM cumple el mismo objetivo que K-medias, este, realiza el procedimiento de agrupamiento de manera diferente, ya que en vez de tener un centroide, usa un el concepto de medioide, que como indica Joaquín Amat (2017), es un elemento dentro de un grupo en el cual la distancia promedio entre él y todos los demás elementos del grupo, es la menor posible. Esta diferencia es la que hace de PAM una mejor herramienta de Clustering para datos con datos atípicos.

En estos algoritmos, a diferencia de otros de Clustering, necesitan que se ingrese la cantidad de Clústers que se buscan generar con los datos. Aún así, existe métodos para encontrar el número óptimo de Clústers para un conjunto determinado, dentro de los cuales se encuentran

Elbow y Silhouette .

2.3 DISTANCIAS

Para establecer a qué grupo debe pertenecer un sujeto, es necesario establecer una medida de similaridad, o lo que es su contraparte, una de distancia. Esto permite ver cuan iguales o diferentes son un conjunto de datos con respecto a otro y permite ir clasificándolos. Existe una gran variedad de medidas de distancia, desde la más simple como puede ser la Euclideana, a unas más complejas y que sirven para diferentes tipos de casos.

2.3.1 Distancia Manhattan

Esta distancia que también es llamada geometría del taxista, según Joaquín Amat (2017) define la distancia entre dos puntos p y q como la sumatoria de las diferencias absolutas entre cada dimensión o coordenada. Tiene la característica de que se ve menos afectada por datos atípicos, por lo que se considera más robusta que la distancia euclídea dado que no eleva al cuadrado las diferencias: $d_{man}(p,q) = \sum_{i=1}^{n} |(p_i - q_i)|$

2.3.2 Distancia Gower

La distancia Gower fue ideada especialmente para datos de tipo mixto, que como menciona Prieto, R (2006) fue propuesto por Gower un coeficiente general de semejanza que busca combinar diferentes descriptores y los procesa de acuerdo al propio tipo de este. En la fórmula S_{ijk} denota la constribución por la k-ésima variable y w_{ijk} es usualmente 1 o 0 dependiendo si la comparación es válida para la k-ésima variable.

$$s_{ij} = \frac{\sum_{i=k}^{n} w_{ijk} s_{ijk}}{\sum_{i=k}^{n} w_{ijk}}$$

CAPÍTULO 3. PRE-PROCESAMIENTO

En el laboratorio anterior se realizó una especie de pre-procesamiento para la base de datos, generando la limpieza de algunos datos que no tenían sentido, o eliminando columnas que no tenían información que pareciera relevante. Por otro lado, también se eliminaron todas las filas que tuviesen elementos del tipo NULL, esto con el fin de facilitar el manejo de la base de datos. Si bien esto fue de utilidad para esa experiencia, en este caso se buscará realizar un proceso más elaborado, que no sólo incluya lo antes mencionado, si no, que también permita normalizar los datos y rescatar todos los registros posibles.

En este capítulo se utilizará como guía el artículo "$6\ Pasos\ para\ realizar\ un\ pre-procesamiento\ de\ datos\ óptimo$" en el cual P. Aguilera (S.F.) menciona y muestra estos 6 pasos, entre los cuales están: Integración de las Bases de Datos, Reconocimiento de Datos, Detección y tratamiento de datos faltantes y atípicos (outliers), Reducción de dimensionalidad por transformación de variables, Discretización y numerización de las variables, y finalmente, Normalización del rango de los datos. Dado que algunos de estos pasos, o partes de ellos ya fueron implementados anteriormente, acá se hará referencia al informe anterior y se mostrarán los nuevos procedimientos.

3.1 INTEGRACIÓN DE LAS BASES DE DATOS

Esta etapa tiene más sentido cuando se tienen 2 o más bases de datos que pueden tener datos similares, o incluso, a veces repetidos, en la cual los registros podrían tener problemas de duplicación o pérdida de datos.

En esta experiencia se hace uso de sólo una base de datos, por lo que la integración consiste en traer los datos y manejarlos dentro de un dataframe. Lo único que se le realiza a este dataframe dentro de la etapa de integración es darle nombres apropiados a las columnas.

3.2 RECONOCIMIENTO DE DATOS

En esta sección se debe incluir todo lo que es el análisis descriptivo de las variables, es decir, cantidad total de datos, medidas de tendencia central, mínimo, máximo, desviación estándar, etc. Esto permite conocer a grandes rasgos los datos, además se puede hacer uso de gráficos tales como histogramas, de barras, entre otros.

Gran parte de este proceso ya fue realizado en el laboratorio anterior, pero acá se mostrarán algunos de ellos nuevamente para tener una idea de lo que se hará en los siguientes pasos.

Existen variables discretas que sólo indican si se midió o no una variable continua de la base de datos, por lo que no tienen relevancia real en los procedimientos que se realizarán posteriormente en este informe. Debido a esto, no son consideradas en esta etapa ninguna variable que contenga ”measured”.

En la Tabla 3.1 se puede apreciar las variables discretas, exceptuando sexo, con las cantidades de sus respectivas opciones binarias V o F. Mientras que en la Tabla 3.2 se aprecia las cantidades de hombres y mujeres.

Tabla 3.1: Variables discretas y cantidad de datos (sin considerar sexo)

Variable	Verdaderos	Falsos
On thyroxine	330	2470
Query on thyroxine	40	2760
On Antithyroid medication	34	2766
Sick	110	2690
Pregnant	41	2759
Thyroid surgery	39	2761
I131 treatment	48	2752
Query on thyroxine	163	2637
Query hypothyroid	173	2627
Query hyperthyroid	14	2786
Lithium	14	2786
Goitre	25	2775
Tumor	71	2729
Hypopituitary	1	2799
Psych	135	2665

Tabla 3.2: Variables sexo y cantidad de datos

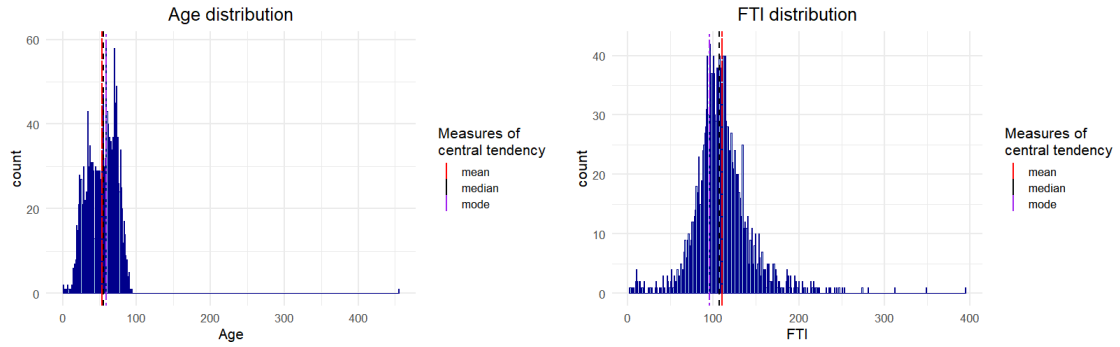
Variable	Mujeres	Hombres
Sex	1830	860

En la Tabla 3.3 se pueden apreciar la media, mediana, desviación estándar, como también los mínimos y máximos de cada una de las variables continuas. La variable TBG no tiene estos datos, ya que todos sus valores son NULL.

Tabla 3.3: Variables continuas y sus datos de tendencia central, mínimo y máximo

Variable	Media	Mediana	min	max	sd
Edad	51.84	54.00	1.00	455.00	20.46
TSH	4.67	1.40	0.005	478.00	21.45
T3	2.03	2.00	0.05	10.60	0.83
TT4	109.07	104.00	2.00	430.00	35.39
T4U	0.99	0.98	0.310	2.12	0.19
FTI	110.79	107.00	2.00	395.00	32.88
TBG	?	?	?	?	?

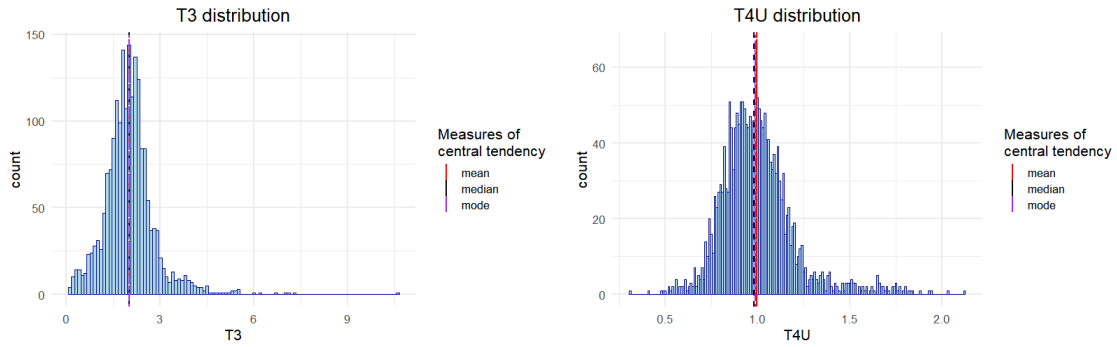
Los gráficos de la Figura 3.1 a la 3.3 muestran la distribución que tienen cada una de las variables continuas. Estos serán relevantes en la siguiente etapa para la detección y corrección de datos atípicos.



(a) Gráfico distribución Edad.

(b) Gráfico distribución FTI.

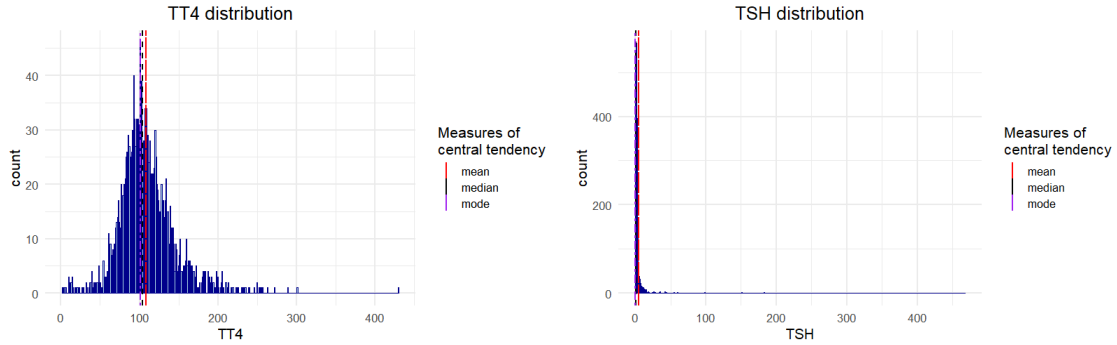
Figura 3.1: Gráficos variables continuas Pt.1



(a) Gráfico distribución T3.

(b) Gráfico distribución T4U.

Figura 3.2: Gráficos variables continuas Pt.2



(a) Gráfico distribución TT4.

(b) Gráfico distribución TSH.

Figura 3.3: Gráficos variables continuas Pt.3

Cabe mencionar que dentro de la base de datos existe una gran cantidad de datos faltantes que se representan como "?", es decir, como NULL. Esto debido a que mucho de los sujetos no tenían las mediciones de algunos datos como TSH, T3, TT4, entre otros.

3.3 DETECCIÓN Y TRATAMIENTO DE DATOS FALTANTES Y ATÍPICOS

Dado que en grandes volúmenes de datos se suelen tener algunos de ellos que se encuentran fuera de los rangos normales establecidos por el investigador, es necesario establecer cuáles son estos rangos, para luego detectar y tratar estos datos atípicos. Lo mismo ocurre con los datos faltantes, ya que puede ser relevante tener la mayor cantidad de datos posibles.

3.3.1 Definición de rangos normales

En el laboratorio anterior se definieron los rangos normales que se tenían para las variables continuas según la literatura, los cuales se pueden apreciar en la tabla 3.4.

Tabla 3.4: Variables continuas y sus valores normales

Variable	Rango valores normales	Unidad
Edad	0 a 110	Años
TSH	0,4 a 4,5	mU/L
T3	0.92 a 2.76	nmol/L
TT4	54 a 115	nmol/L
T4U	0.71 a 1.85	ng/dL
FTI	45 a 117	ng/dL

De los datos mostrados, el rango normal de FTI fue cambiado con respecto a la experiencia anterior, ya que tanto los valores, como las unidades que antes fueron entregadas carecían de sentido con respecto a los datos. De endocrinology (S.F) indican que el rango normal de FTI para personas $>=$ a 20 años es de 4.8-12.7 mcg/dl, que en ng/ml queda entre 48 y 127.

Por otro lado, en la tabla 3.5 se puede apreciar los valores esperados para las variables discretas. No se consideraron las variables ”measured”.

Tabla 3.5: Variables discretas y sus valores normales

Variable	Valores normales
Sex	M o F
On thyroxine	V o F
Query on thyroxine	V o F
On Antithyroid medication	V o F
Sick	V o F
Pregnant	V o F
Thyroid surgery	V o F
I131 treatment	V o F
Query on thyroxine	V o F
Query hypothyroid	V o F
Query hyperthyroid	V o F
Lithium	V o F
Goitre	V o F
Tumor	V o F
Hypopituitary	V o F
Psych	V o F

3.3.2 Detección y tratamiento de datos atípicos

Teniendo en cuenta los rangos normales mostrados en las tablas anteriores y en el análisis realizado en la fase de reconocimiento de datos, es posible comenzar a detectar si entre los datos existen datos atípicos o que no tengan sentido. El procedimiento a realizar con estos datos es reemplazarlos por una medida que no genere un gran impacto en el promedio o la desviación estándar, por lo que la mediana cumple bien esta característica.

3.3.2.1 Variables continuas

- **Edad**: En esta variable sólo hay un dato que supera los rangos normales, la cual es la edad máxima de 455 años.
- **TSH**: Los datos mayores a 10 mU/L fueron considerados como atípicos para esta variable.
- **T3**: Los datos mayores a 5 mU/L fueron considerados como atípicos para esta variable.
- **TT4**: Los datos mayores a 220 mU/L fueron considerados como atípicos para esta variable.
- **T4U**: No se apreciaron datos fuera de lo común.
- **FTI**: Los datos mayores a 200 mU/L fueron considerados como atípicos para esta variable.

3.3.2.2 Variables discretas

En la base de datos no existen datos atípicos para las variables discretas, por lo que no se realizará ningún procedimiento para éstas.

3.3.3 Detección y tratamiento de datos faltantes

Para realizar el procedimiento de detección es simplemente buscar todos los datos que contengan un '?', lo que equivale a un NULL.

Por otro lado, el proceso para el tratamiento es un tanto más complejo, esto debido a que lo que se busca es tener la mínima pérdida y sesgo en los datos.

Existen variadas formas de tratar los datos faltantes, donde la más simple siempre será no hacer nada con ellos, pero también hay otras posibilidades que involucran el eliminar al sujeto que tiene estos datos faltantes, siendo esta última opción viable cuando la cantidad de estos datos no es muy grande. Además de los métodos mencionados, también existe el concepto de imputación o reemplazo, en el cual se suelen utilizar valores como la media o mediana de los datos como sustituto al valor faltante, lo cual tiene sus pro's y contras, dentro de los cuales indica Will Badr (2019) en los pro's está su facilidad y velocidad de implementación , mientras que en los contras está el hecho de que sólo funciona a nivel de columnas y más importante, es que no es demasiado preciso. Cabe mencionar que también existen otras formas de realizar este proceso, pero son más complejas, por lo que en este informe sólo se reducirá a la eliminación de filas y la imputación.

3.3.3.1 Procedimiento realizado

Lo que se hizo para el tratamiento de los datos faltantes, como se mencionó anteriormente, fue la eliminación de filas y a la vez, la imputación de datos mediante su mediana. Esto, siempre buscando mantener la mayor cantidad de datos.

Se buscó dentro de todas las variables de la base de datos cuál era la que tenía una mayor cantidad de datos faltantes, que para este caso fue T3 con 585 datos nulos, y se eliminaron todas esas muestras, para así disminuir en la mayor medida posibles estos datos faltantes, pero no eliminando la de todas las columnas. De esta manera, en esta variable existe un 0 % de datos nulos, mientras que en las otras columnas, estos datos están alrededor del 5 % del total de aquella

variable. Este procedimiento permite poder imputar todos los datos nulos de cada columna por su mediana.

Así, de tener 2800 registros, sólo se redujeron a 2215, mientras que si se hubiesen eliminado todos los registros con datos nulos, estos se hubiesen reducido a 1946 casos.

3.4 REDUCCIÓN DE DIMENSIONALIDAD

En esta fase se busca disminuir la cantidad de variables a estudiar. Razones para esto hay muchas, pero dentro de las principales se encuentra el tiempo de procesamiento y análisis de los datos. Por otro lado, existe el concepto de la ”Navaja de Ockham”, el cual indica en simples palabras que la respuesta más simple suele ser la más probable”, es decir, que si se logran reducir las variables que no aportar información relevante al problema, se podría estar más cerca de encontrar información útil.

3.4.1 Variables continuas

Dado que existe una variable llamada TBG que sólo posee valores nulos, esta se quitará dado que no aporta ninguna información.

3.4.2 Variables discretas

Dentro de las variables discretas hay una gran cantidad de ellas que no aportar información relevante, ya sea porque todos sus valores son verdaderos o falsos, o porque sólo indican la existencia de una variable continua.

Por lo anterior, todas las variables discretas que indican una medición (measured) fueron quitadas. Dentro de estas se encuentran TSH measured, T3 measured, entre otras.

Además de las variables anteriormente quitadas, también se encuentra Hypopituatary, dado que en la tabla 3.1 se puede apreciar que sólo existe 1 caso verdadero y todos los demás son falsos. Esto no aporta ninguna información relevante al análisis.

Por último, las columnas llamadas referral source y class también fueron removidas, ya que la primera sólo indica de dónde se obtuvieron los datos y la segunda, indica el diagnóstico del paciente.

3.4.3 Método de Componente principales

Una de las soluciones más comunes para la reducción de dimensionalidad es el uso del método de Componentes principales, el cual busca concentrar las variables del problema en diferentes componentes que se vean caracterizadas por estas variables. Esto permite seleccionar una cantidad menor de Componentes a la cantidad de variables que se tienen y así reducir la dimensionalidad.

Se intentó realizar este procedimiento, pero dado que la cantidad de CP necesarias para explicar las variables que se tienen era tan alto, que resultaba ser inútil. En la Figura 3.4 se puede apreciar que la primera componente principal sólo cubre el 13.6 %, mientras que todas las demás están bajo del 10 %. Por otro lado, para entender qué significa cada una de estas componentes se debían caracterizar para entender qué variables explican, proceso que hace aún más complejo el procesamiento de los datos. Considerando lo anterior, se decidió no utilizar este método.

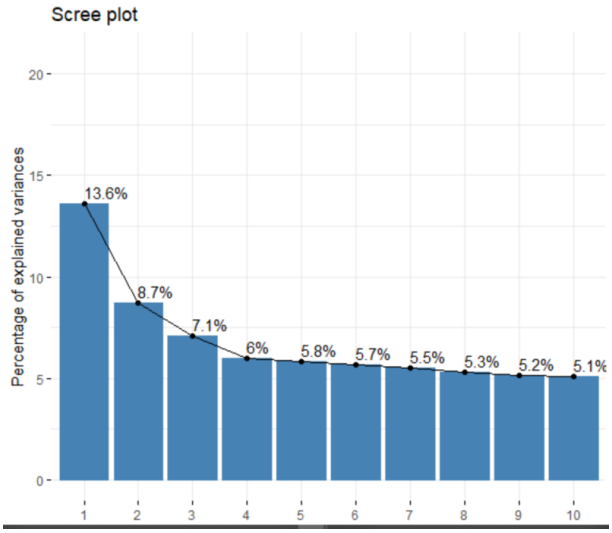


Figura 3.4: Gráfico de cuánta varianza explican las primeras 10 CP

3.5 DISCRETIZACIÓN Y NUMERIZACIÓN DE LAS VARIABLES

Ya habiendo realizado todas las fases anteriores, identificando cada tipo de variable con las cuales se trabajará, entonces ahora se debe determinar si es necesario o conveniente transformar las variables a numéricas o nominales.

Dado la naturaleza de nuestros datos en que una parte son continuos y en otra nominales binarios, lo más simple es tomar estas variables nominales binarias que pueden ser F o V como también F o M para el caso de sexo, y llevarla a los valores numéricos 0 o 1 respectivamente. De esta manera se tienen todas las variables con valores numéricos, hecho que simplifica mucho la utilización de algunos algoritmos o métodos para Clustering.

3.6 NORMALIZACIÓN DEL RANGO DE LOS DATOS

Cuando se habla de normalización en este tipo de contexto, es básicamente el pasar toda variable continua al rango de 0 a 1, lo que si bien no es necesario para muchos de los métodos existentes para agrupaciones o machine learning en general, como menciona P. Aguilera

(S.F.) en técnicas basadas en distancias como lo podría ser Componentes principales o K-medias, es necesario realizar este procedimiento, ya que permite mantener la relación entre los valores. Aunque luego de finalizar los algoritmos, se recomienda desnormalizar estas variables para así facilitar la compresión y análisis de los resultados.

En este caso se utilizó la forma de normalización que se aprecia en la Figura 3.5, en la cual a un elemento X de cierta columna se le resta el mínimo de aquella columma, para luego esto dividirlo por el máximo, menos el mínimo de la misma columna.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Figura 3.5: Normalización de una variable X

CAPÍTULO 4. OBTENCIÓN DE CLÚSTER

Para la realización de este capítulo se utilizó como guía un artículo publicado en Rpubs por Joaquín Amat (2017), en el que se explica cómo hacer uso del algoritmo de PAM, pero también se hace uso de métodos como el de Elbow y Silhouette para obtener un número de Clústers óptimo a tener en consideración al momento de utilizar el parámetro K (cantidad de Clusters) en PAM.

Dada la naturaleza de los datos mostrados en el capítulo anterior y a todos los procedimientos de preprocesamiento realizados, hacer uso del algoritmo de K-means no debería ser ningún inconveniente, considerando que se hizo una limpieza de datos, eliminando e imputando datos tanto inexistentes como atípicos, aún así, se consideró como una mejor opción utilizar PAM (Partitioning Around Medoids), dado que según Kassambara (S.F.) este es un algoritmo robusto contra datos atípicos, pero más importante aún, que los medioides son sujetos representativos de los grupos y no sólo un promedio. Esto permite realizar un análisis preciso de cada uno de estos grupos considerando a la clase a la cual pertenecen.

Otro punto importante que se mostrará en este capítulo son las distancias utilizadas en el algoritmo de Clustering.

4.1 DISTANCIAS UTILIZADAS

Para la selección de distancias a utilizar se deben tener en cuenta el tipo de datos que se tienen en la base de datos. En este caso, no existe un único tipo de dato, si no, una mezcla de ellos. Una gran cantidad es de tipo binario, ya sea True o False para la mayoría de las variables discretas, así como M o F en el caso de Sexo, mientras que todas las demás son de tipo numéricas continuas. Por lo tanto, teniendo en cuenta esta mezcla, es necesario hacer uso de una medida de distancia que soporte datos mixtos, o transformar los datos binarios a numéricos y así trabajar con ellos.

En este laboratorio se decidió realizar ambas aproximaciones, haciendo uso de

distancia Manhattan para los datos numéricos y transformando los datos binarios a 0's y 1's, mientras que en la distancia Gower se utilizaron los datos mixtos.

Es importante tener en cuenta que la distancia Gower es la más adecuada en este caso, ya que sirve exactamente para el tipo de datos que se tiene, pero a modo de ilustración se compararán sus resultados con la distancia Manhattan.

4.2 NÚMERO ÓPTIMO DE CLUSTERS

Para determinar el valor K que se debe ingresar como parámetro a PAM existen variadas formas de hacerlo. Si bien una opción es probar un gran rango de valores, esto es bastante ineficiente, ya que no sólo basta con correr el algoritmo de Clustering, si no, también se debe analizar cada uno de los clusters, considerando si la información presentada allí tiene algún valor.

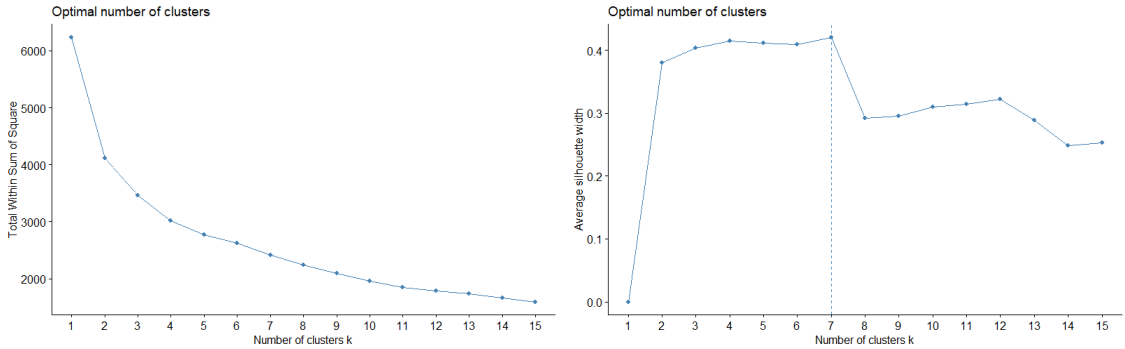
Considerando lo anterior, se emplearon 2 métodos con el objetivo de encontrar un número óptimo de Clusters para el conjunto de datos que se desea agrupar, que si bien no necesariamente dan el número exacto de Clusters, entregan una muy buena aproximación de lo buscado.

4.2.1 K Óptimo para distancia Manhattan

En la Figura 4.1 (a) se puede apreciar el K óptimo entregado por el método Elbow. La idea de este método es buscar el número de Clústers óptimo en la parte del gráfico donde se vea una regularizacion en la curva, o dicho de otra manera, según Amat (2017), un punto de la curva donde la mejora deja de ser sustancial, considerando como medida la varianza total intra-clusters en función de la cantidad de Clústers que se tengan.

En la Figura 4.1 (b) por otro lado, muestra el K óptimo mediante el método Silhouette, donde el valor más alto en el eje Y es el K óptimo. Amat (2017) indica que a diferencia de

Elbow que busca minimizar el WSS (total inter-cluster sum of squares), en Silhouette se busca maximizar el índice de silueta, que en simples palabras considera qué tan buena fue la asignación realizada a una observación comparando su similitud con las demás observaciones de su Clúster frente a la de los otros grupos. Este índice se encuentra entre -1 y 1, donde un valor mayor indica que la observación ha sido clasificada correctamente en el Clúster.



(a) K óptimo mediante Elbow

(b) K óptimo mediante Silhouette

Figura 4.1: Comparación Método Elbow y Silhouette para PAM con distancia Manhattan

4.2.2 K Óptimo para distancia Gower

Al igual que en el K óptimo para PAM con distancia Manhattan, acá se realiza mediante el método Silhouette, el cual se puede ver en la Figura 4.2. Se puede apreciar que en este caso no se aplicó Elbow para obtener el K, pero esto se debe a que Silhouette da una muy buena opción, que es 4 grupos. Si se mira la Figura 4.1 (b),si bien el método Silhouette muestra un K óptimo, este no difiere mucho de los que se encuentran en el rango de 4 a 7, mientras que en la Figura 4.2 es claro que K = 4 es una buena opción a considerar.

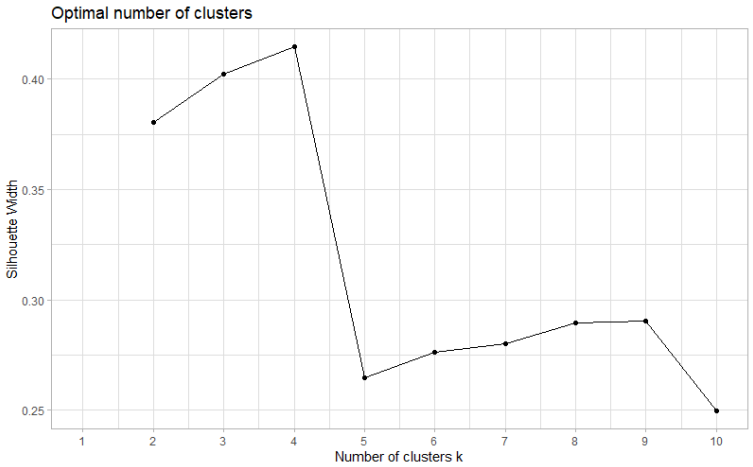


Figura 4.2: Gráfico K óptimo método Silhouette con distancia Gower

4.3 CLÚSTERS OBTENIDOS

Como se mencionó anteriormente, para esta base de datos se realizaron 4 Clústers diferentes, donde se varió tanto el K (cantidad de grupos), como el tipo de distancia a utilizar.

En la Figura 4.3 se pueden apreciar los Clústers obtenidos usando un K = 4 y haciendo uso de distancia Gower y Manhattan respectivamente. A simple vista no pareciera haber mucha diferencia en ellos, ya que incluso sus medioides que se indican en "$id.med$" en donde se determina la fila del dataframe en el que se encuentra el sujeto, son exactamente los mismos. Es decir, los sujetos más representativo en cada grupo, para ambos casos, son iguales. Ahora, si se comienza a observar en mayor detalle la variable "$clustering$", en la cual se indica a qué grupo pertenece cada registro de la base de datos, se pueden ver algunas diferencias en cuanto a la asignación que se les hizo a estos registros.

pam4_gower	list [9] (S3: pam, partition)	List of length 9
medoids	character [4]	'2373' '1475' '1240' '744'
id.med	integer [4]	1877 1167 981 592
clustering	integer [2215]	1 1 2 1 1 1 ...
objective	double [2]	0.045 0.045
isolation	factor	Factor with 3 levels: "no", "L", "L*"
clusinfo	double [4 x 5]	1.21e+03 1.80e+02 6.93e+02 1.28e+02 1.75e-01 2.35e-01 2.14e-01 1.79e-01 4.39e-02 ...
silinfo	list [3]	List of length 3
diss	NULL	Pairlist of length 0
call	language	pam(x = gower_dist, k = 4, diss = TRUE)

(a) Clúster PAM con k= 4 y distancia Gower

pam4_manhattan	list [10] (S3: pam, partition)	List of length 10
medoids	double [4 x 20]	0.5591 0.5699 0.5591 0.4301 1.0000 1.0000 0.0000 1.0000 0.0000 1.0000 0.0000 0.0 ...
id.med	integer [4]	1877 1167 981 592
clustering	integer [2215]	1 1 2 1 1 1 ...
objective	double [2]	0.9 0.9
isolation	factor	Factor with 3 levels: "no", "L", "L*"
clusinfo	double [4 x 5]	1214.000 180.000 693.000 128.000 3.497 4.709 4.273 3.574 0.878 ...
silinfo	list [3]	List of length 3
diss	NULL	Pairlist of length 0
call	language	pam(x = sep_data_normalized[1:20], k = 4, metric = "manhattan")
data	double [2215 x 20]	0.4301 0.2366 0.7419 0.7419 0.8495 0.6989 1.0000 1.0000 1.0000 1.0000 1.0000 1.0 ...
(attributes)	list [2]	List of length 2

(b) Clúster PAM con k= 4 y distancia Manhattan

Figura 4.3: Comparación Clústers obtenidos con K = 4 y diferentes distancias

pam7_gower	list [9] (S3: pam, partition)	List of length 9
medoids	character [7]	'2623' '1475' '2449' '1240' '744' '2742' ...
id.med	integer [7]	2081 1167 1937 981 592 2174 ...
clustering	integer [2215]	1 1 2 3 3 3 ...
objective	double [2]	0.0399 0.0394
isolation	factor	Factor with 3 levels: "no", "L", "L*"
clusinfo	double [7 x 5]	5.49e+02 1.78e+02 5.90e+02 6.00e+02 1.28e+02 9.20e+01 1.60e-01 2.35e-01 1.54e-01 ...
silinfo	list [3]	List of length 3
diss	NULL	Pairlist of length 0
call	language	pam(x = gower_dist, k = 7, diss = TRUE)

(a) Clúster PAM con k= 7 y distancia Gower

pam7_manhattan	list [10] (S3: pam, partition)	List of length 10
medoids	double [7 x 20]	0.3871 0.5699 0.7312 0.5591 0.4301 0.3871 1.0000 1.0000 1.0000 0.0000 1.0000 0.0 ...
id.med	integer [7]	2081 1167 1937 981 592 2174 ...
clustering	integer [2215]	1 1 2 3 3 3 ...
objective	double [2]	0.798 0.787
isolation	factor	Factor with 3 levels: "no", "L", "L*"
clusinfo	double [7 x 5]	549.0000 178.0000 590.0000 600.0000 128.0000 92.0000 3.1929 4.7093 3.0807 ...
silinfo	list [3]	List of length 3
diss	NULL	Pairlist of length 0
call	language	pam(x = sep_data_normalized[1:20], k = 7, metric = "manhattan")
data	double [2215 x 20]	0.4301 0.2366 0.7419 0.7419 0.8495 0.6989 1.0000 1.0000 1.0000 1.0000 1.0000 1.0 ...

(b) Clúster PAM con k= 7 y distancia Manhattan

Figura 4.4: Comparación Clústers obtenidos con igual K = 7 y diferentes distancias

Por otro lado, en la Figura 4.4 se tiene la comparación entre los Clústers con un número de grupos K = 7 y diferentes distancias. Similar a como ocurre con los Clústers de K = 4, sus medioides son idénticos, pero al mirar a fondo çlustering", se pueden notar algunas diferencias en la asignación de grupo para los sujetos.

Es interesante notar que entre los Clústers de la Figura 4.3 y 4.4 existen medioides que son iguales, lo que podría llegar a indicar que estos son fundamentales en las agrupaciones, ya que representan de mejor manera los grupos a los que pertenecen, que cualquier otro sujeto dentro de toda la base de datos.

CAPÍTULO 5. ANÁLISIS DE LOS RESULTADOS

En este capítulo se analizará un caso de los clúster obtenidos, el cual corresponde al de la distancia Gower con K = 4, comparando las clases a las que pertenecen las observaciones y el grupo al cual son asignadas.

Como se mencionó en los capítulos anteriores, la distancia Gower es la más adecuada para trabajar con datos mixtos, y PAM entrega los medioides que corresponden a observaciones de la base de datos. Esto último permite analizar a los medioides como representantes de cada grupo, donde se toma dicha observación y se analiza cada valor que toman las variables, identificando así características que personifican al grupo. Además, se decidió analizar solo el Clúster con K=4, ya que como se presenció en la figura 4.2, éste es el óptimo de los grupos para la base de datos con que se trabaja.

Las frecuencias de las clases (diagnósticos) que se encuentran en cada grupo aparecen en la Tabla 5.1 como porcentajes y en la Tabla 5.2 con números.

Tabla 5.1: Porcentajes de las clases en cada grupo

Grupo	Negativos	T3 toxic	Goitre	Hipertiroide
1	96,79 %	0,25 %	0,49 %	2,47 %
2	99,44 %	0,56 %	0 %	0 %
3	98,99 %	0,15 %	0 %	0,87 %
4	83,59 %	2,34 %	0,78 %	13,28 %

Tabla 5.2: Frecuencias de las clases en cada grupo

Variables	Negativos	T3 toxic	Goitre	Hipertiroide
1	1175	3	6	30
2	179	1	0	0
3	686	1	0	6
4	107	3	1	17

La siguiente Tabla contiene los medioides de cada grupo, indicando los valores que toman algunas de las variables.

Tabla 5.3: Datos de medioides

Variables	Medioide 1	Medioide 2	Medioide 3	Medioide 4
Edad	53	54	53	41
Sexo	Femenino	Femenino	Masculino	Femenino
Con tiroxina	No	Si	No	No
Consulta de hipertiroidismo	No	No	No	Si
TSH	1,50	0,93	1,40	0,15
T3	1,8	1,8	2,0	2,0
TT4	105	115	103	103
T4U	0,98	0,94	0,98	1,09
FTI	107	123	107	107
Clase	Negativo	Negativo	Negativo	Hipertiroide

Solo se exponen 10 variables de las 21 que se tienen, esto debido a que solo las que se muestran son las que presentan diferencias entre los medioides, el resto comparten los mismos valores, los cuales son $falsos$.

5.1 ANÁLISIS DE GRUPOS

Utilizando los datos presentados en las Tablas es que se procede a analizar cada grupo del clúster.

5.1.1 Grupo 1

A partir de los datos expuestos en la Tabla 5.1 y 5.2 se puede observar que el grupo 1 es el que más contiene individuos con diagnóstico de hipertiroidismo, y a su vez el que más contienen a personas con diagnóstico negativo. Esto se puede deber a que en su mayoría las observaciones pertenecen a la clase negativo (un 96,93 %), es decir, no clasifican en ninguna de las otras enfermedades. Por otro lado, pese a contener varios datos con hipertiroidismo y algunos con goitre y T3 toxic, el porcentaje de ellos en el grupo es bajo, comparado con la clase negativo. Esto da a suponer que los individuos que se encuentran en este grupo en su mayoría no presentan enfermedades como hipertiroidismo, T3 toxic o goitre. Para confirmar lo anterior es que se analiza los valores que toma el medioide del grupo 1. Estos datos se encuentra en la Tabla 5.3.

El medioide del grupo 1 es una mujer de 53 años, que no esta en tratamiento con tiroxina ni tampoco ha hecho alguna consulta por hipertiroidismo, y el resto de las variables se encuentran en un rango normal, es por esto que finalmente se cree que el grupo 1 representa a las mujeres sanas y con datos similares al medioide.

5.1.2 Grupo 2

Para el segundo grupo se observa que la mayoría de sus datos pertenecen a individuos de la clase negativo, dando a entender que aquellos individuos pueden estar sanos, o por lo menos no presentar ninguna de las enfermedades como hipertiroidismo, goitre o T3 toxic. Analizando el medioide 2 de la Tabla 5.3 es que obtiene los datos de una mujer de 54 años, la cual si esta en tratamiento con tiroxina, pero no ha realizado alguna consulta por hipertiroidismo. El resto de los datos se encuentran en un rango normal, exceptuando el FTI que esta más elevado, esto se puede explicar por el tratamiento que lleva el individuo. La clase a la que pertenece finalmente es negativa, coincidiendo con el 99,44 % de los datos. Además, cabe mencionar que este grupo tiene solo 180 observaciones, siendo uno de los más pequeños si se compara con el primer grupo.

5.1.3 Grupo 3

Este grupo es similar al anterior, presentando un 98,99 % de casos negativos, un 0,15 % a T3 toxic, un 0 % con goitre y un 0,87 % tiene hipertiroidismo. Se puede suponer a partir de estos valores que el grupo contiene a individuos en general sanos, o que están lejos de tener enfermedades relacionadas al hipertiroidismo. Para ver las diferencias con respecto a los grupos anteriores es que se analizan los datos de la Tabla 5.3.

Los valores que toma el medioide pertenecen a un hombre de 53 años, que no esta en tratamiento con tiroxina ni tampoco ha realizado alguna consulta por hipertiroidismo, además, los valores de TSH, T3, TT4, T4U y FTI se encuentran en los rangos normales, junto con esto

también se observa que pertenece a la clase negativo. Todo lo anterior permite suponer que este grupo representa a individuos varones sanos.

5.1.4 Grupo 4

El último grupo del clúster contiene 128 datos, es el más pequeño de todos, pese a eso tiene ciertas características que marcan diferencia con el resto de los grupos. Observando la Tabla 5.2 se distinguen 17 casos con hipertiroidismo, y al igual que el grupo 1, la mayoría de los individuos están en la clase negativa, pese a eso la gran diferencia entre estos grupos es el porcentaje que tienen dichas clases, donde el hipertiroidismo equivale al 13,28 %, este es el valor más alto entre todos los grupos sobre los que contienen alguna enfermedad.

A priori se podría decir que este grupo representa a los individuos con hipertiroidismo, esto se confirma analizando su medioide de la Tabla 5.3, el cual indica que es una mujer de 41 años, sin tratamiento con tiroxina, pero que si ha consultado previamente por hipertiroidismo, junto a lo anterior se observa que el TSH es el único valor fuera del rango normal, siendo más bajo. La clase a la cual pertenece esta observación es hipertiroide.

Lo anterior permite suponer que una de las variables más relacionadas con la enfermedad del hipertiroidismo es el TSH, ya que para este medioide el resto de los valores se encuentran en un rango normal, y comparando con los otro grupos, pese a tener algunas de las otra variables alteradas, siguen perteneciendo a la clase negativo.

CAPÍTULO 6. CONCLUSIONES

A lo largo de esta experiencia se realizaron una gran cantidad de procedimientos de los que muchos de ellos no llegaron a figurar en este informe, ni tampoco el código que fue programado. Dentro de ellos se encuentra reducción por componentes principales, algoritmo de K-medias, distancias euclidianas, entre otros. Esto da a entender la complejidad que se tiene al analizar datos, ya que no sólo basta con ejecutar método tras método, si no, que se debe de alguna manera tomar los datos y estrujarlos de forma tal que puedan entregar toda la información posible, pero no lo hacen fácilmente. No cualquier método sirve para cualquier dato. Se debe saber a priori la naturaleza de los datos y de qué manera se puede obtener lo que se busca de ellos.

Considerando lo anterior, es clave elegir las herramientas correctas. Si bien seguramente existen métodos mucho más potentes que los utilizados dentro de esta experiencia, PAM, Gower y Silhouette, fueron piezas claves para poder analizar los datos. PAM y Silhouette son algoritmos ampliamente documentados, por lo que su uso se facilita bastante, pero la distancia Gower no es tan manejada, por lo que implementarla junto a los algoritmos anteriores se complicó de una u otra manera.

6.1 SOBRE LOS RESULTADOS

Sería el ideal decir que se pudo obtener información contundente y sólida del Clúster obtenido y sus grupos, pero eso no es la realidad. Si bien se buscó el mejor método de Clustering dentro de los conocidos en cátedra para los datos, también se buscó la mejor medida de distancia para datos mixtos y se aplicaron los mejores algoritmos para obtener un número de grupos óptimo, los datos no fueron fáciles de analizar.

Se obtuvieron 4 grupos diferentes, de los cuales sólo 1 se podría decir que representa a personas enfermas por Hipertiroidismo, ya que al mirar su medioide, se aprecia que tiene un índice hormonal fuera de lugar, donde el más notorio es el TSH, que lo tiene por debajo

de 0.15, cuando lo normal es entre 0.4 y 4.5. Mientras en otros grupos el medioide también tenía algunos índices hormonales anormales pero eran diagnosticados como negativo, en el grupo 4 sólo teniendo el TSH bajo determinó la enfermedad.

Por otro lado, en la Tabla 5.1 queda clara la predominancia de diagnósticos de clase negativa (sanos) en todos los grupos, pero en el Grupo 4 es donde este valor es menor, e incluso, la cantidad de casos como T3 Toxic, Goitre y sobretodo, Hipertiroidismo es notoria.

Si se comparan los grupos 1, 2 y 3 las diferencias no son muy notorias. Lo más destacable tal vez es que el representante del Grupo 3 es una persona de sexo masculino, pero si se comparan sus datos con el medioide del grupo 1, más allá de eso, no hay diferencias. Por otro lado, es destacable el hecho de que el grupo 2 es representado por una persona que se encuentra en un tratamiento con tiroxina, lo que explica algunos de sus altos índices hormonales.

Así, se podría concluir lo siguiente con respecto a los datos:

- Grupo 1: Grupo sano, representado por sexo femenino.
- Grupo 2: Grupo en tratamiento, representado por sexo femenino.
- Grupo 3: Grupo sano, representado por sexo masculino.
- Grupo 4: Grupo enfermo, representado por sexo femenino.

Si bien los objetivos del laboratorio fueron cumplidos a cabalidad, haber conocido más sobre los métodos existentes y para qué tipo de datos se aplican, hubiesen ahorrado mucho del trabajo experimental, que a prueba y error, se fue complicando. Como se mencionó anteriormente, muchos métodos utilizados no fueron de utilidad para los datos, por lo que se tuvieron que desechar completamente. Aún así, la experiencia de ver cómo el trabajo va tomando forma y todo comienza a tener sentido, es altamente gratificante.

BIBLIOGRAFÍA

Aguilera, P. (S.F.). 6 pasos para realizar un preprocesamiento de datos óptimo). [Online] https://aml.stradata.co/6-pasos-para-realizar-un-preprocesamiento-de-datos-optimo/.

Amat, J. (2017). Clustering y heatmaps: aprendizaje no supervisado. [Online] https://rpubs.com/Joaquin_AR/310338.

Badr, W. (2019). 6 different ways to compensate for missing values in a dataset (data imputation with examples). [Online] https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779.

Delgado, R. (2018). Introducción a los modelos de agrupamiento (clustering) en r). [Online] https://rpubs.com/rdelgado/399475.

endocrinology (S.F.). Free thyroxine index (fti), serum. [Online] https://endocrinology.testcatalog.org/show/FRTUP.

Kassambara, A. (S.F.). K-medoids in r: Algorithm and practical examples. [Online] https://www.datanovia.com/en/lessons/k-medoids-in-r-algorithm-and-practical-examples/.

Moya, R. (2016). ¿que es el clustering? [Online] https://jarroba.com/que-es-el-clustering/.

Prieto, R. (2006). Técnicas estadísticas de clasificación. [Online] https://www.uaeh.edu.mx/docencia/Tesis/icbi/licenciatura/documentos/Tecnicas%20estadisticas%20de%20clasificacion.pdf.

Unioviedo (S.F.). El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. [Online] https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html.