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

Laboratorio 2

Integrantes: Nicole Reyes
Diego Águila

Curso: Análisis de Datos

Profesor: Max Chacón

Ayudante: Francisco Muñoz

5 de Mayo de 2019

Tabla de contenidos

1. Introducción

El descubrimiento de conocimiento en una base de datos (del inglés knowledge discovery in databases, KDD) es un proceso dedicado a identificar patrones de comportamiento en datos potencialmente útiles. Compuesto por los siguientes pasos: 1) comprensión del dominio de estudio, 2) obtención de un conjunto de datos objetivo, 3) limpieza y procesamiento de datos, 4) minería de datos, 5) interpretación de patrones y 6) obtención de conocimiento. El KDD es usado en áreas de negocios con múltiples fines, tales como, detectar fraudes, aumentar ventas dado el comportamiento de clientes, etc. Es usado también en áreas de ciencia como el estudio de la genética humana y en la ingeniería eléctrica para monitorizar las condiciones de instalaciones de alta tensión.

Esta experiencia contempla la ejecución de este proceso con respecto a la base de datos "Allhypo" estudiada de la entrega anterior correspondiente al Hipotiroidismo, enfermedad que afecta a la glándula tiroides en donde la baja producción de la hormona Tiroxina (T4) es un indicio de esta enfermedad. Cabe destacar que esa entrega correspondía a la primera etapa del proceso de KDD enfocada en la compresión del dominio de los datos.

En el presente documento se describe un marco teórico con conceptos claves para comprender el desarrollo de la experiencia, se presenta el pre-procesamiento de los datos, donde se define como se realiza la limpieza de la base de datos y así pasar a la etapa de obtención de los clústers, mostrándose distintos criterios de obtención de estos con el fin de obtener el adecuado para los datos, se analizan los resultados a modo de identificar características relevantes que por sí solas no son observables en los datos pudiendo enriquecer estos, para finalmente obtener conclusiones respecto a la base de datos y el desarrollo de la experiencia.

1.1. Objetivos

1. Identificar y extraer elementos de la base de datos que sean inadecuado.
2. Aplicar el algoritmo K-Means a la base de datos para obtener agrupaciones entre estos.
3. Analizar las agrupaciones para identificar características valiosas que aporten conocimiento a la base de datos.

2. Marco Teórico

En esta sección se describen los conceptos que explican las técnicas utilizadas durante el desarrollo de esta experiencia.

2.1. Clustering

Un algoritmo de agrupamiento o clustering, es un proceso usado muy recurrentemente en la minería de datos o en Machine learning, el cual dado un conjunto de datos y un criterio, se realizan agrupaciones de datos de los cuales cada uno comparte características o propiedades similares con el fin de obtener similitudes (entre elementos del mismo grupo) o diferencias (entre elementos de distintos grupos). Dentro de los algoritmos que existen para realizar esta técnica están el algoritmo de Mean-shift (cambio de medias) o el algoritmo de K-means (K-medias).

2.2. K-means

Es uno de los algoritmos de agrupamiento más conocido y más usado, su objetivo es particionar los datos en un conjunto de k grupos generados en base a la distancia de los datos con respecto a un centroide. Este algoritmo es iterativo y cuenta con 5 pasos de ejecución que se describen a continuación:

1. Formar k agrupaciones con los datos de forma aleatoria.
2. Obtener el centroide de cada grupo mediante el cálculo de la media.
3. Calcular la distancia de cada elemento con respecto a todos los centroides y asignarlo al grupo cuyo centroide sea el más cercano.
4. Recalcular los centroides de cada grupo.
5. Repetir los pasos 3 y 4 hasta que los centroides dejen de variar y los datos dejen de cambiar de grupo.

Una vez realizado este procedimiento se obtiene un conjunto de k grupos de datos los cuales poseen una relación entre sí y de la cual se puede obtener información.

2.3. Distancias

Los métodos y algoritmos de Clustering requieren encontrar patrones de similitud para identificar los grupos, la más conocida es la medida de las distancias. Dentro de las mas importantes está la *distancia de Gower*, la cual tiene como característica generar una medida de distancia con datos de distintos tipos como por ejemplo en donde se tienen valores continuos y binarios. Su fórmula es la siguiente:

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

- p_1, p_2, p_3 corresponden al número de variables cuantitativas continuas, variables binarias y cualitativas (que no sean binarias) respectivamente.
- a y d corresponden al número de coincidencias de las variables binarias (1,1) en el caso de a y (0,0) en el caso de d.
- α es el número de coincidencia en las variables cualitativas.
- G_h corresponde al recorrido o rango de h-ésima variable.

3. Pre-procesamiento

En la experiencia anterior, se pudo visualizar la existencia de valores anómalos en la base de datos usada "Allhypo", por lo que para esta experiencia se presentan los siguientes procedimientos para la limpieza de los datos y así poder trabajar sobre la base de datos con ningún problema y realizar el posterior análisis.

3.1. Eliminación de valores vacíos

Dentro de la base de datos, existen celdas con valores nulos (representados con el símbolo "?"), por lo que se opta por eliminar todos los sujetos de la muestra que contengan este símbolo para alguna variable ya que no aportan nada al análisis, para que así no entorpezca el estudio de esta base de datos, además de que no hayan dificultades en la plataforma R. Además de que no se puede seguir con la experiencia si se tienen estos datos ya que al aplicar el agrupamiento posteriormente esto no tendrá sentido alguno.

3.2. Eliminación de variables

Dentro del conjunto de datos usados, hay columnas que no formarán parte del estudio, debido a que no aportan información importante para este pudiendo estropear el agrupamiento de los datos a utilizar como también lo hacen los valores vacíos. Entonces, las columnas correspondientes a la variables que son eliminadas son:

- Variable TBG: Corresponde a la proteína fijador de la Tiroxina (T4). Como toda su columna representa valores vacíos, no entrega información relevante debido a que se puede inferir a que no se midió la presencia de esta proteína en ningún paciente de la muestra, por lo tanto se opta por eliminarla de la base de datos.

- Variables de verificación de medición (measured): Indican si el paciente se ha realizado alguna medición, siendo esta información no relevante para el estudio. Lo que si importa es el valor numérico de estas mediciones de las hormonas por litro de sangre en los pacientes.

Entonces, se decide dejar fuera las siguientes variables de medición: *TSH measured, T3 measuted, TT4 measured, T4U measured, FTI measured y TBG measured.*

- Variable de Estado y de fuente: Como también sucede con las variables anteriores, la variable *referral.source* no entrega información que sea pertinente para el agrupamiento de datos, al igual que la variable de estado del paciente (que indica la condición de la tiroides) tampoco es un criterio importante para después utilizar el método de las k-medias. Entonces así, se eliminan ambas columnas de la base de datos.

3.3. Con respecto a valores numéricos

Al realizar un análisis en profundidad de la base de datos, se pudo apreciar algo que no se vio en la entrega pasada, que tiene que ver con ciertos datos anómalos, como por ejemplo, la edad de un paciente de mas de 400 años. Por lo tanto, lo que se hace es tener un tope máximo del valor aceptable para cada variable numérica.

- Edad: Valor aceptable hasta los 100 años.
- TSH: Valor aceptable hasta 15.0
- T3: Valor aceptable hasta 10.0
- TT4: Valor aceptable hasta 460
- T4U: Valor aceptable hasta 3.45
- FTI: Valor aceptable hasta 405.0

3.4. Muestra final a estudiar

Al final, la base de datos a utilizar tiene una cantidad de 21 variables, tanto binarias como continuas, además de un total de 1874 sujetos, de un total original de 2800, lo que corresponde al 66.92 % del total original de pacientes en la base de datos. A pesar de haber perdido una cantidad considerable de sujetos en la limpieza, esta es necesaria para el estudio a realizar sobre esta misma y no tener algún inconveniente con alguna variable y/o valor al momento de calcular las distancias Euclidiana y de Gower en el agrupamiento.

4. Obtención del Clúster

Se realizan dos tipos de agrupamiento, uno teniendo en cuenta variables binarias (o dicotómicas) y continuas, que suman en total 21 variables, y otro de solo variables continuas que son 6. A continuación se presenta el uso del algoritmo de agrupamiento de K-medias.

4.1. Distancias a utilizar

Para agrupar los datos, se necesita un paso previo, el cual es la distancia de estos datos entre sí, lo que se conoce como matriz de disimilaridades que sirve como entrada para el agrupamiento de las K-medias, en donde habrán dos matrices de disimilaridades una para cada Clustering.

Para el primer Clustering, se utiliza la *Distancia de Gower* para el cálculo de esta matriz, en donde dentro de las 21 variables hay valores que son binarios (o dicotómicos) y continuos por lo tanto es la más adecuada para este caso. Para el segundo Clustering, como son solo valores numéricos se usa la *Distancia Euclidiana* ya que esta es la distancia común que se puede medir con una regla entre este tipo de datos.

4.2. Cantidad de grupos a formar

Otro parámetro que se necesita para el agrupamiento de las K-medias es el número de grupos a formar. El problema que se tiene para determinar el número de grupos es que la calidad de estos depende de la observación realizada por el investigador, teniendo en cuenta que los integrantes de esta experiencia no son expertos en temáticas de salud hormonal que es lo relacionado al tema específico que se está analizando, el Hipotiroidismo.

Como solución se utiliza el método de las siluetas conocido como *Anchura de siluetas*. La finalidad de este método es encontrar la mejor consistencia entre los datos y la cantidad de grupos a utilizar, en otras palabras, presenta una medida de que tan correctamente se ajusta cada instancia en su grupo. El método está expuesto en la documentación del paquete *Cluster* de R, en donde se probaron distintos números para los grupos, obteniendo la siguiente Figura.

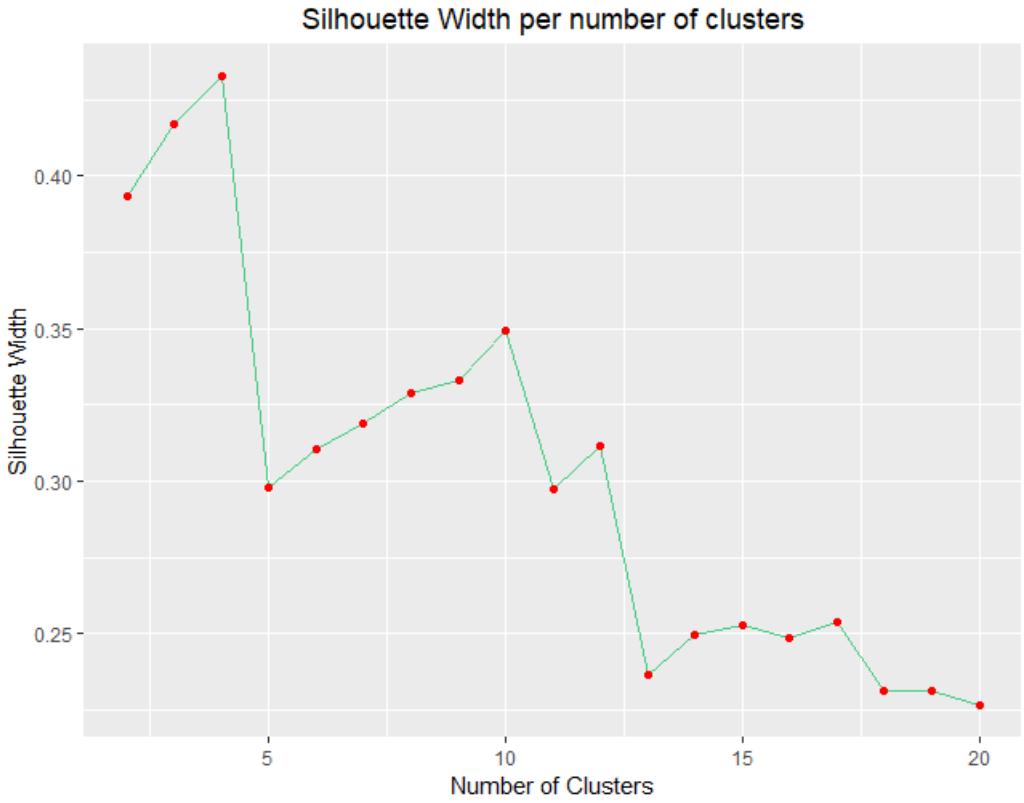


Figura 1: Ancho de las siluetas por número de grupos.

De acuerdo al método, el grupo que tenga un ancho de las siluetas mayor será el número de grupos a usar en K-medias, por lo que en este caso se utiliza un **K=3**.

4.3. Agrupamiento con K-medias

Los siguientes resultados de agrupamiento son graficados con la librería *Rtsne*, ya que esta permite de alguna manera "aplanar" todas las variables, a modo de poder visibilizar los grupos en un plano de dos dimensiones.

Para el primer clustering aplicado al conjunto de datos binarios y numéricos se obtiene el gráfico de la Figura 2, del cual, como aproximación inicial, inmediatamente se observa que el grupo amarillo es pequeño en comparación a los grupos verde y rojo, también se observa que en estos dos grupos existen pequeños conjuntos de datos aislados.

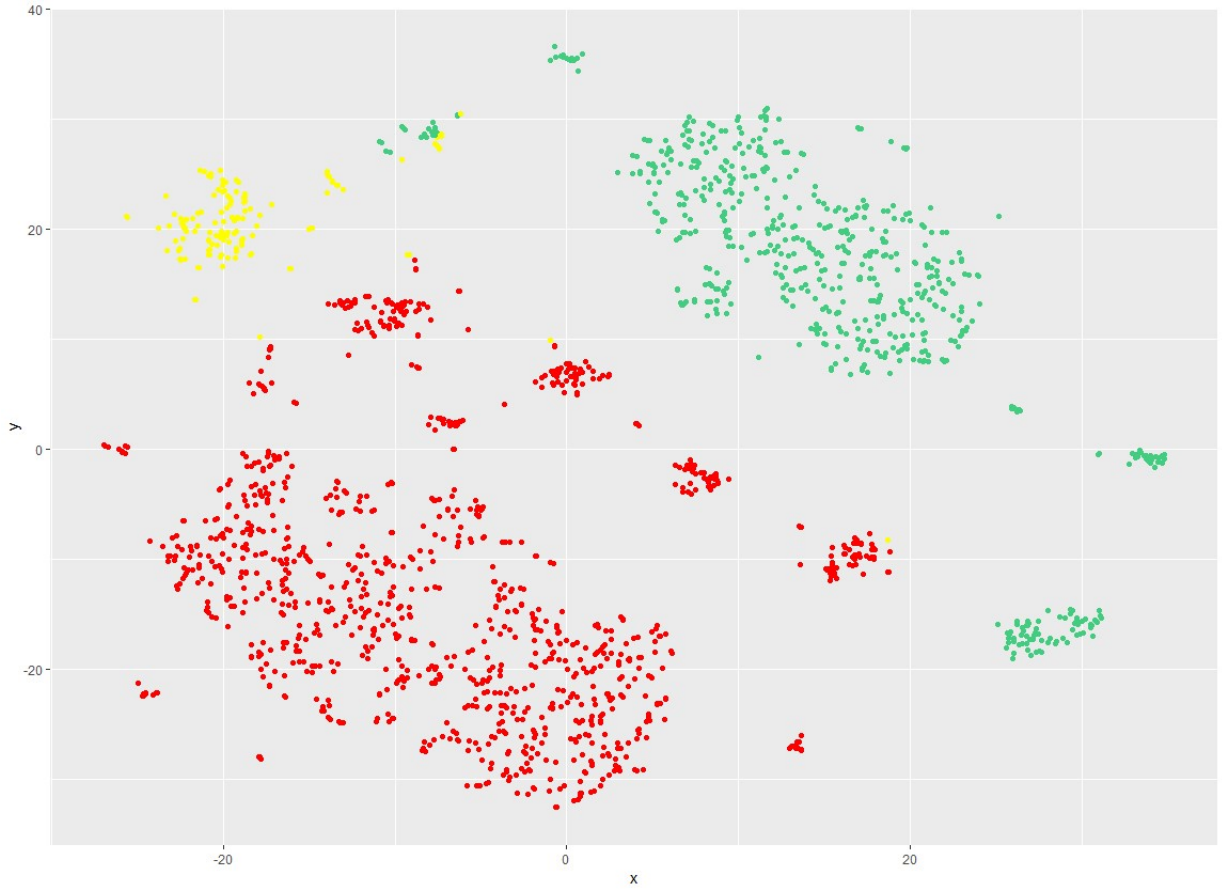


Figura 2: Gráfico de los 3 grupos generados para los datos binarios y numéricos.

Para el segundo gráfico, correspondiente a los grupos del conjunto de datos numéricos se obtiene la distribución que se observa en la Figura 3, nuevamente el grupo amarillo resulta ser el más pequeño, sin embargo, en esta ocasión no se presentan datos aislados y las distribuciones de los grupos resultan ser similares.

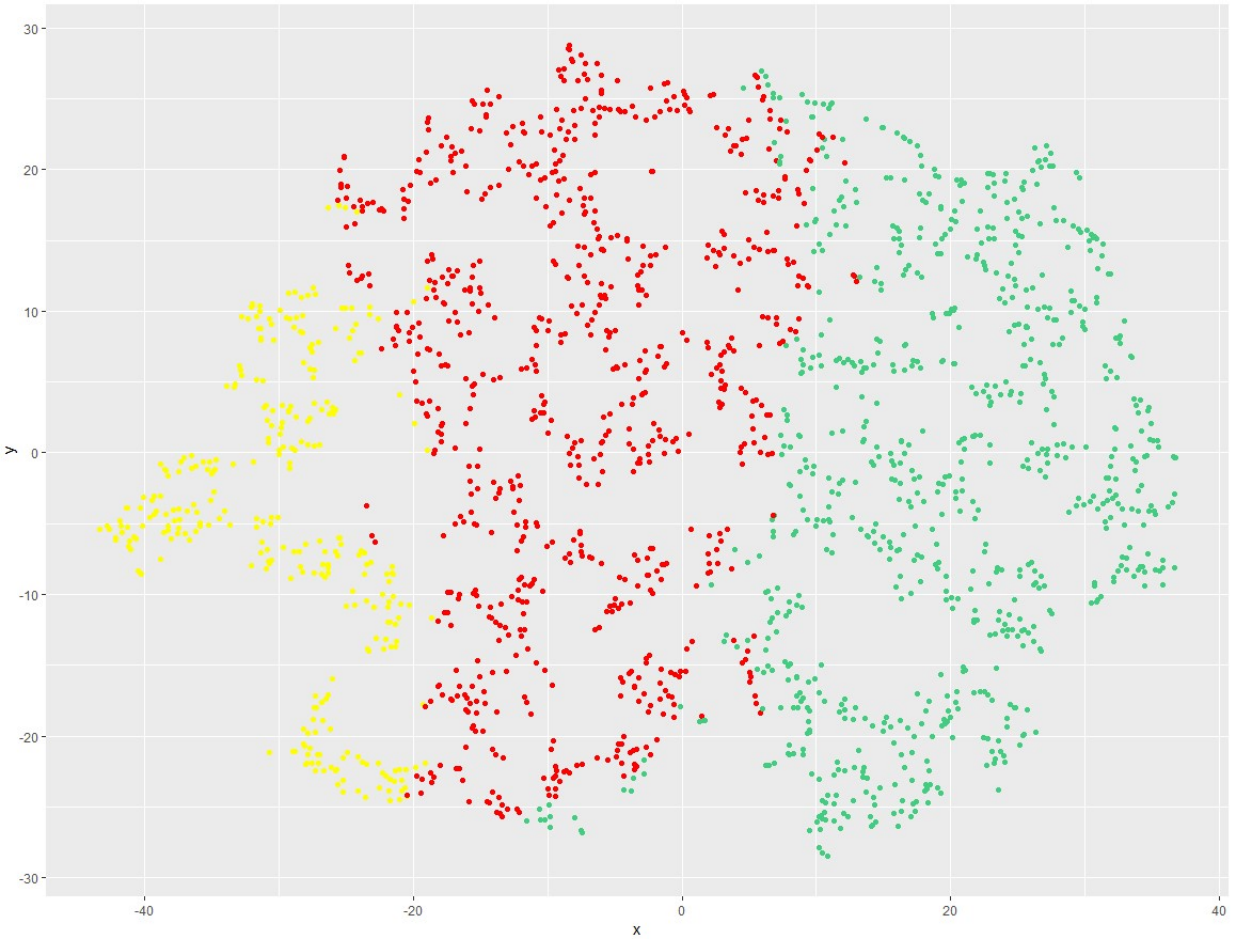


Figura 3: Gráfico de los 3 grupos generados para los datos numéricos.

5. Análisis de los resultados

En esta sección se analizan los resultados de la ejecución del algoritmo k-means para los conjuntos de datos binarios y numéricos, se obtiene información acerca de las representaciones visuales para explicar las similitudes de los datos dentro de los grupos y las diferencias entre los grupos.

5.1. Primer Clustering

El primer clustering corresponde a los 3 conjuntos de datos binarios y numéricos de la base de datos obtenidos del algoritmo k-means como se observa en las Figuras 4, 5 y 6:

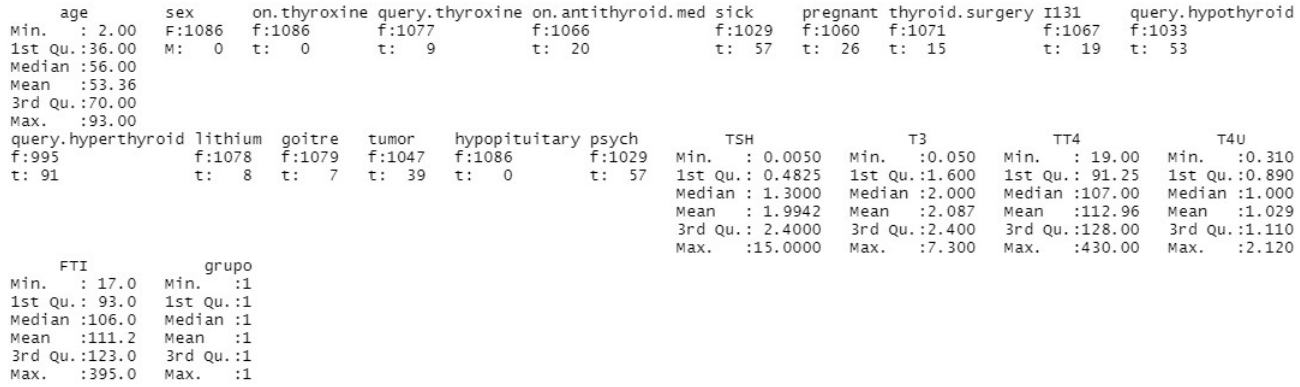


Figura 4: Resumen de los resultados del primer grupo para los datos binarios y numéricos.

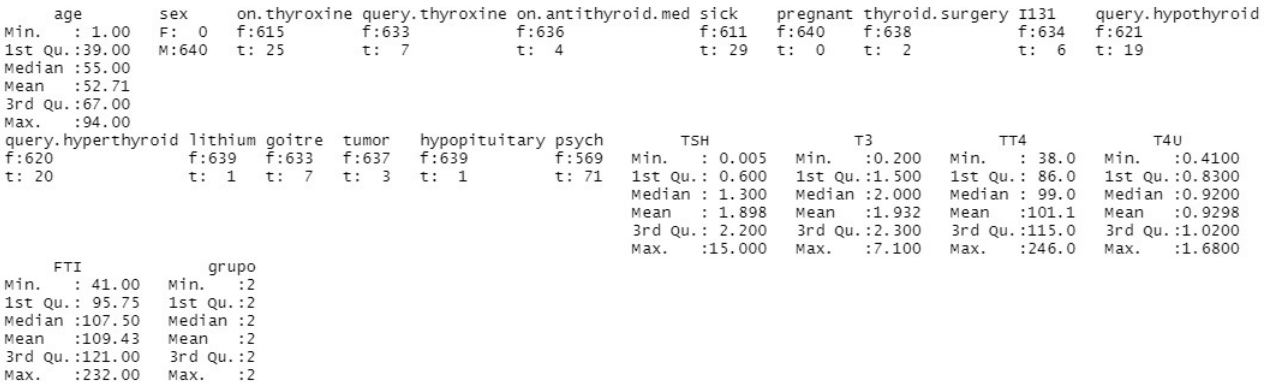


Figura 5: Resumen de los resultados del segundo grupo para los datos binarios y numéricos.

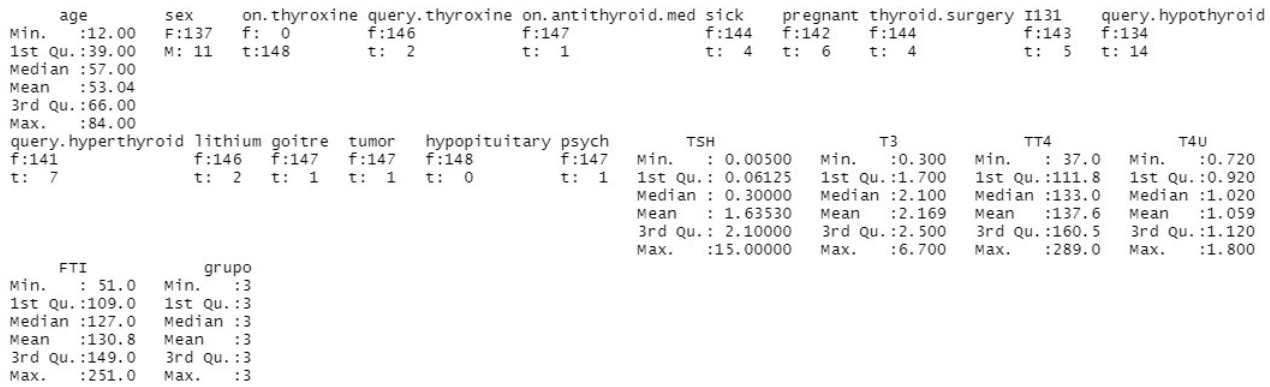


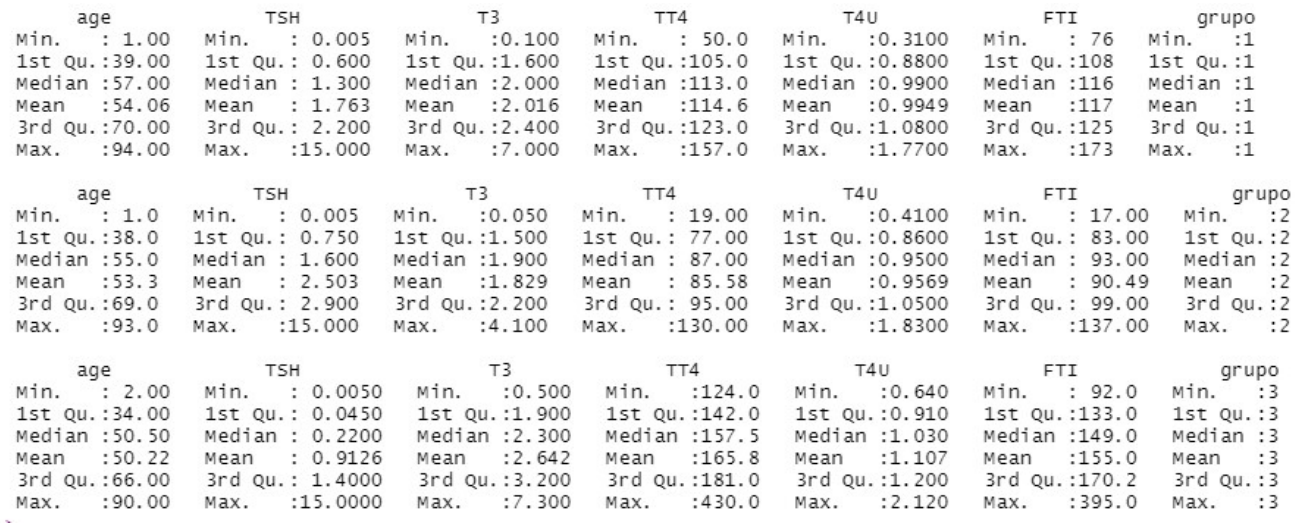
Figura 6: Resumen de los resultados del tercer grupo para los datos binarios y numéricos.

Para este clustering, una de las características más notorias, es que el grupo 1 esta compuesto solo por mujeres y el grupo 2 solo por hombres, sin embargo, en el grupo 3 hay hombres y mujeres, esto indica que uno de los factores predominantes para la generación de los grupos fue el sexo del paciente. Luego, del grupo 3 se observa que el total de los pacientes se encuentra en tratamiento de reemplazo de tiroxina, lo opuesto del grupo 1 en el cual los 1086 pacientes no se encuentran realizando este tratamiento, claramente el grupo 3 se ha formado con respecto a esta característica.

Por otra parte, se observa que en el grupo 2, de hombres, de un total de 640 pacientes, 71 de ellos presentan problemas psiquiátricos, lo opuesto al caso del grupo 1, mujeres, donde del total de 1086 pacientes, solo 57 de ellas presentan problemas psiquiátricos. También se observa que en el grupo 1 de mujeres tienden a tener más cirugías a la tiroides que el grupo 2 de hombres. Finalmente, se observa también que en el grupo 1 un total de 39 pacientes ha padecido de algún tumor, no así en el grupo 2, donde solamente 3 pacientes presentan esta característica.

5.2. Segundo Clustering

El segundo clustering corresponde a los 3 conjunto de datos numéricos obtenidos del algoritmo k-means como se observa en las Figura 7:



age	TSH	T3	TT4	T4U	FTI	grupo
Min. : 1.00	Min. : 0.005	Min. :0.100	Min. : 50.0	Min. :0.3100	Min. : 76	Min. :1
1st Qu.:39.00	1st Qu.: 0.600	1st Qu.:1.600	1st Qu.:105.0	1st Qu.:0.8800	1st Qu.:108	1st Qu.:1
Median :57.00	Median : 1.300	Median :2.000	Median :113.0	Median :0.9900	Median :116	Median :1
Mean :54.06	Mean : 1.763	Mean :2.016	Mean :114.6	Mean :0.9949	Mean :117	Mean :1
3rd Qu.:70.00	3rd Qu.: 2.200	3rd Qu.:2.400	3rd Qu.:123.0	3rd Qu.:1.0800	3rd Qu.:125	3rd Qu.:1
Max. :94.00	Max. :15.000	Max. :7.000	Max. :157.0	Max. :1.7700	Max. :173	Max. :1

age	TSH	T3	TT4	T4U	FTI	grupo
Min. : 1.0	Min. : 0.005	Min. :0.050	Min. : 19.00	Min. :0.4100	Min. : 17.00	Min. :2
1st Qu.:38.0	1st Qu.: 0.750	1st Qu.:1.500	1st Qu.: 77.00	1st Qu.:0.8600	1st Qu.: 83.00	1st Qu.:2
Median :55.0	Median : 1.600	Median :1.900	Median : 87.00	Median :0.9500	Median : 93.00	Median :2
Mean :53.3	Mean : 2.503	Mean :1.829	Mean : 85.58	Mean :0.9569	Mean : 90.49	Mean :2
3rd Qu.:69.0	3rd Qu.: 2.900	3rd Qu.:2.200	3rd Qu.: 95.00	3rd Qu.:1.0500	3rd Qu.: 99.00	3rd Qu.:2
Max. :93.0	Max. :15.000	Max. :4.100	Max. :130.00	Max. :1.8300	Max. :137.00	Max. :2

age	TSH	T3	TT4	T4U	FTI	grupo
Min. : 2.00	Min. : 0.0050	Min. :0.500	Min. :124.0	Min. :0.640	Min. : 92.0	Min. :3
1st Qu.:34.00	1st Qu.: 0.0450	1st Qu.:1.900	1st Qu.:142.0	1st Qu.:0.910	1st Qu.:133.0	1st Qu.:3
Median :50.50	Median : 0.2200	Median :2.300	Median :157.5	Median :1.030	Median :149.0	Median :3
Mean :50.22	Mean : 0.9126	Mean :2.642	Mean :165.8	Mean :1.107	Mean :155.0	Mean :3
3rd Qu.:66.00	3rd Qu.: 1.4000	3rd Qu.:3.200	3rd Qu.:181.0	3rd Qu.:1.200	3rd Qu.:170.2	3rd Qu.:3
Max. :90.00	Max. :15.0000	Max. :7.300	Max. :430.0	Max. :2.120	Max. :395.0	Max. :3

Figura 7: Resumen de los resultados de los tres grupos para los datos numéricos.

De los datos que se presentan en la Figura 7 existe una serie de columnas que se diferencian considerablemente entre los grupos, se puede ver que entre los tres grupos existe una diferencia muy notoria, por ejemplo, en la columna del FTI, el grupo 1 posee valores entre 76 y 173, no así en el caso del grupo 2, que los valores van entre 17 y 137, y en el caso del grupo 3, los valores van de 92 a 395, algo similar ocurre en las columnas del TT4, y en las columnas del TSH y T3 también ocurre pero en cantidades numéricas menores. Esto demuestra que la generación de las agrupaciones provocó una división con respecto a las cantidades de hormonas TSH, T3, TT4 y el indice FTI mayormente. Las agrupaciones obtenidas permiten suponer que existe una correlación lineal entre las cantidades de ciertas hormonas presentes, se observa que en los 3 grupos dada la cantidad de la hormona TT4 existe una cantidad similar del índice FTI.

6. Conclusiones

Gracias a esta experiencia de laboratorio, se puede desarrollar efectivamente el proceso de KDD, ya que se realiza una limpieza a la base de datos, al eliminar los valores vacíos y algunas variables irrelevantes que pueden entorpecer el proceso, luego se realiza la obtención de los clústers adecuados mediante el uso del algoritmo de k-medias y las distancias de Gower y Euclidiana, finalizando así la tercera y cuarta etapa del KDD con 2 clústers de 3 grupos cada uno. Luego se realiza la quinta etapa de interpretación de patrones, donde se analizan los clústers obtenidos para detectar las diferencias existentes entre los grupos y las similitudes de los datos de cada grupo que motivaron a la formación de estos.

Una de las dificultades que surgieron al momento de realizar la experiencia, fueron las dudas sobre ciertos contenidos sobre la teoría como los ejemplo las distancias, en donde en los apuntes del curso (Chacón, 2018) no quedaba tan claro (a modo personal) el como usar estas distancias al momento de obtener luego los clústers, ya que se muestran variadas distancias, pero no indican en que momento se utilizan o bajo que variables. Lo mismo pasa con la teoría de los agrupamientos, que de la manera de como lo tiene el profesor expuesto no fue suficiente para entenderlo de mejor manera.

Finalmente, pese a haber obtenido las agrupaciones y haber identificado dichas similitudes y diferencias, los patrones obtenidos no entregan información concreta, pero si permite establecer hipótesis para futuros estudios en relación con el hipotiroidismo y sus posibles consecuencias, como por ejemplo, que los hombres tienden a sufrir más problemas psiquiátricos, o que las mujeres tienden a requerir más cirugías a la tiroides.

7. Anexos

A continuación se presenta el script en R utilizado para el desarrollo de esta experiencia.

```
require(ggplot2)
require(cluster)
require(Rtsne)

#Se obtiene los datos.
data <- read.csv("C:\\Users\\diego\\Dropbox\\Usach\\ANLISIS DE DATOS\\Lab\\allhypo.data",
                 col.names = c("age","sex", "on.thyroxine", "query.thyroxine", "on.antithyroid.med", "sick","pregnant", "thyroid.surgery", "I131", "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","state"),
                 header=FALSE, sep=",", stringsAsFactors=FALSE)

#### ETAPA DE LIMPIEZA ####

#Se limpia la variable de estado del paciente eliminando numeros
state <- data$state
state <- strsplit(state, ".", fixed = TRUE)
state <- lapply(state, `[[`, 1)
state <- unlist(state)
data$state <- state

#Se elimina la variable TBG y TBG.measured
data$TBG <- NULL
data$TBG.measured <- NULL

#dentro del clustering no vamos a usar ni el state ni el source
data$state <- NULL
data$referral.source <- NULL

```

```r
#Se eliminan las variables que indican si se realizo la medicion de la hormona
data$TSH.measured <- NULL
data$T3.measured <- NULL
data$TT4.measured <- NULL
data$T4U.measured <- NULL
data$FTI.measured <- NULL

# Se limpian los datos, eliminando los sujetos que contengas algun NA
#sex
error = data$sex == '?'
data = data[!error,]

#age
error = data$age == '?'
data = data[!error,]

#on thyroxine
error = data$on.thyroxine == '?'
data = data[!error,]

#query on antithyroid thyroxine
error = data$query.thyroxine == '?'
data = data[!error,]

#on antithyroid medication
error = data$on.antithyroid.med == '?'
data = data[!error,]

#sick
error = data$sick == '?'
data = data[!error,]

#pregnant
error = data$pregnant == '?'
data = data[!error,]

```

```
#thyroid surgery
error = data$thyroid.surgery == '?'
data = data[!error,]

#I131 treatment
error = data$I131 == '?'
data = data[!error,]

#query hypothyroid
error = data$query.hypothyroid == '?'
data = data[!error,]

#query hyperthyroid
error = data$query.hyperthyroid == '?'
data = data[!error,]

#lithium
error = data$lithium == '?'
data = data[!error,]

#goitre
error = data$goitre == '?'
data = data[!error,]

#tumor
error = data$tumor == '?'
data = data[!error,]

#hypopituitary
error = data$hypopituitary == '?'
data = data[!error,]

#psych
error = data$psych == '?'
data = data[!error,]

```

```
#tsh
error = data$TSH == '?'
data = data[!error,]

#t3
error = data$T3 == '?'
data = data[!error,]

#tt4
error = data$TT4 == '?'
data = data[!error,]

#t4u
error = data$T4U == '?'
data = data[!error,]

#fti
error = data$FTI == '?'
data = data[!error,]



#Se transforman los datos correspondientes

#variables no numericas
data$sex <- as.factor(data$sex)
data$on.thyroxine <- as.factor(data$on.thyroxine)
data$query.thyroxine <- as.factor(data$query.thyroxine)
data$on.antithyroid.med <- as.factor(data$on.antithyroid.med)
data$sick <- as.factor(data$sick)
data$pregnant <- as.factor(data$pregnant)
data$thyroid.surgery <- as.factor(data$thyroid.surgery)
data$I131 <- as.factor(data$I131)
data$query.hypothyroid <- as.factor(data$query.hypothyroid)
data$query.hyperthyroid <- as.factor(data$query.hyperthyroid)
data$lithium <- as.factor(data$lithium)
```

```r
data$goitre <- as.factor(data$goitre)
data$tumor <- as.factor(data$tumor)
data$hypopituitary <- as.factor(data$hypopituitary)
data$psych <- as.factor(data$psych)

#variables numericas y continuas
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)

#Se eliminan datos numericos que puedan ser anomalos
#age
error = data$age > 100
data = data[!error,]

#tsh
error = data$TSH > 15
data = data[!error,]

#t3
error = data$T3 > 10
data = data[!error,]

#tt4
error = data$TT4 > 460
data = data[!error,]

#t4u
error = data$T4U > 3.45
data = data[!error,]

#fti
error = data$FTI > 405.0
```

```r
data = data[!error,]

#Ahora quedan 1874 sujetos de un total de 2800
#corresponde al %66.92 aprox del total de datos

#### ETAPA DE CLUSTERING ####

#Se obtiene la matriz de disimilaridades con distancia de Gower ya que la muestra tiene datos continuos y binarias
disimilaridades <- daisy(data, metric = "gower")

#Se obtiene el numero de grupos para el clustering usando el metodo de las siluetas
numero.cluster = 2:20
siluetas <- c()
for (i in numero.cluster){
  kmedia <- pam(disimilaridades, diss = TRUE, k = i)
  siluetas[i] <- kmedia$silinfo$avg.width
}

plot <- ggplot(data.frame(numero = numero.cluster, ancho = siluetas[2:20]), aes(x = numero, y = ancho))+
  labs(x = "Number of Clusters", y = "Silhouette Width")+
  geom_line(color = "seagreen3")+geom_point(color = "red")+
  ggtitle("Silhouette Width per number of clusters")+
  theme(plot.title = element_text(hjust = 0.5))

#Se forman los 3 grupos [k = 3] obtenido con el metodo anterior
clusters <- pam(disimilaridades, diss = TRUE, k = 3)

#Se agrega la columna del cluster al cual pertenece el sujeto
data["grupo"] <- clusters$clustering

#t-SNE plot para graficar los grupos en el plano de dos dimensiones
tsne <- Rtsne(disimilaridades, is_distance = TRUE)
plot.clusters <- ggplot(data.frame(tsne$Y), aes(x = X1, y = X2))+
```

```
204   labs(x = "x", y = "y")+
205   geom_point(aes(color = factor(clusters$clustering)))+
206   scale_colour_manual(values = c("red", "seagreen3","yellow"))
207
208
209 #cluster summary
210 resumen.c1 <- summary(data[data$grupo == 1, ])
211 resumen.c2 <- summary(data[data$grupo == 2, ])
212 resumen.c3 <- summary(data[data$grupo == 3, ])
213
214
215 #Clustering usando solo variables continuas
216 data.continuo <- data[c(1, 17:21)]
217
218 #Se obtiene la matriz de disimilaridades con distancia euclideana solo para
      los valores continuos
219 disimilaridades.continuo <- daisy(data.continuo, metric = "euclidean")
220
221 #Clustering [k = 3]
222 clusters.continuo <- pam(disimilaridades.continuo, diss = TRUE, k = 3)
223
224 #Se Repiten todos los pasos
225 tsne.c <- Rtsne(disimilaridades.continuo, is_distance = TRUE)
226 plot.clusters.continuo <- ggplot(data.frame(tsne.c$Y), aes(x = X1, y = X2))+
227   labs(x = "x", y = "y")+
228   geom_point(aes(color = factor(clusters.continuo$clustering)))+
229   scale_colour_manual(values = c("red", "seagreen3","yellow"))
230
231 data.continuo["grupo"] <- clusters.continuo$clustering
232
233 resumen.continuo.c1 <- summary(data.continuo[data.continuo$grupo == 1, ])
234 resumen.continuo.c2 <- summary(data.continuo[data.continuo$grupo == 2, ])
235 resumen.continuo.c3 <- summary(data.continuo[data.continuo$grupo == 3, ])
```

Bibliografía

Chacón, M. (2018). Análisis de agrupamientos. [Online] https://udesantiagovirtual.cl/moodle2/pluginfile.php?file=%2F115775%2Fmod_resource%2Fcontent%2F1%2FCapitulo%20III%20An%C3%A1lisis%20de%20Datos_AA%202016.pdf.

de la Fuente, N. (2019). K-means clustering: Agrupamiento con minería de datos. [Online] https://estrategiastrading.com/k-means/.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. in proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics, (pp. 281–297). [Online] https://projecteuclid.org/euclid.bsmsp/1200512992.

Seif, G. (2018). The 5 clustering algorithms data scientists need to know. [Online] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.