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

ANÁLISIS DE DATOS

LABORATORIO 3: REGLAS DE ASOCIACIÓN

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

Santiago - Chile
11 de diciembre de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

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

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

1.2 OBJETIVOS

- Extraer conocimiento del problema asignado, por medio de las reglas de asociación a través del software R.
- Utilizar el package: *arulesViz*.

1.3 ORGANIZACIÓN DEL DOCUMENTO

Este documento consta de 4 partes: Primero se presenta un marco teórico de conceptos de importancia, luego la identificación de reglas de asociación con el respectivo análisis de resultados y por último conclusiones respecto a la experiencia.

CAPÍTULO 2. MARCO TEÓRICO

2.1 REGLAS DE ASOCIACIÓN

Las reglas de asociación relacionan una determinada conclusión (por ejemplo, la compra de un producto dado) con un conjunto de condiciones (por ejemplo, la compra de otros productos). Los algoritmos de reglas de asociación buscan automáticamente las asociaciones que se podrían encontrar manualmente usando técnicas de visualización, como en el nodo Malla. La ventaja de los algoritmos de reglas de asociación sobre los algoritmos más estándar de árboles de decisión es que las asociaciones pueden existir entre cualquiera de los atributos. Un algoritmo de árbol de decisión generará reglas con una única conclusión, mientras que los algoritmos de asociación tratan de buscar muchas reglas, cada una de las cuales puede tener una conclusión diferente [2].

Luego de obtenidas las reglas de asociación que corresponde a un antecedente(condición) y consecuente(conclusión) es necesario evaluar la calidad de esta regla, por lo que se presentan las siguientes métricas:

2.1.1. Soporte de una regla

Probabilidad de encontrar un elemento o un conjunto de elementos X en una transacción. Se estima por el número de veces que un elemento o conjunto de elementos se encuentra en todas las transacciones disponibles. Este valor se encuentra entre 0 y 1 [3].

La expresión de soporte es la siguiente:

$$Sop(A \Rightarrow B) = \frac{P(A \cap B)}{n} \tag{2.1}$$

2.1.2. Confianza

Probabilidad de encontrar un elemento o conjunto de elementos Y en una transacción, sabiendo que el elemento o conjunto de elementos X está en la transacción. Se estima por la frecuencia correspondiente observada (número de veces que X e Y se encuentran en todas las transacciones, dividido por el número de veces que se encuentra X). Este valor se encuentra entre 0 y 1 [3].

La confianza se expresa de la forma:

$$Conf(A \Rightarrow B) = \frac{Sop(A \cap B)}{Sop(A)} = \frac{P(A \cap B)}{P(A)} = P(B|A) \tag{2.2}$$

2.1.3. *Lift*

Lift indica la proporción entre el soporte observado de un conjunto de ítems respecto del soporte teórico de ese conjunto dado el supuesto de independencia [4].

Un valor de *lift* $= 1$ indica que ese conjunto aparece una cantidad de veces acorde a lo esperado bajo condiciones de independencia.

Un valor de *lift* > 1 indica que ese conjunto aparece una cantidad de veces superior a lo esperado bajo condiciones de independencia (por lo que se puede intuir que existe una relación que hace que los ítems se encuentren en el conjunto más veces de lo normal).

Un valor de *lift* < 1 indica que ese conjunto aparece una cantidad de veces inferior a lo esperado bajo condiciones de independencia (por lo que se puede intuir que existe una relación que hace que los ítems no estén formando parte del mismo conjunto más veces de lo normal)

El *lift* se expresa de la siguiente forma:

2.1.4. Monotonicidad

El problema del cumplimiento de las restricciones está asociado con la monotonicidad de la restricción, en función de la especialización. Si se tienen dos especializaciones del antecedente, se generan dos reglas tales que $|A1| < |A2|$ y dos restricciones o medidas $med(Ai)$; $i = 1, 2$, asociadas a cada una de las reglas.

- Se dice que la medida es monótona si: $med(A1) \leq med(A2)$.
- La medida es anti-monótona si: $med(A1) \geq med(A2)$.

Para realizar una pre-poda eficiente se requiere usar restricciones monótonas o anti-monótonas. Con lo cual se descartan ramas completas en el proceso de especialización [5].

CAPÍTULO 3. OBTENCIÓN DE REGLAS

Para poder obtener reglas de asociación de la base de datos de hipotiroidismo, es necesario primero realizar una limpieza de los datos (eliminando variables que no entreguen información y datos incompletos o erróneos), además de transformar las variables continuas en variables booleanas o categóricas para poder utilizar la función *apriori* que posee el paquete *arulesViz* de R para obtener las reglas de asociación.

3.1 PRE-PROCESAMIENTO

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

3.1.1 Eliminación de registros nulos

El dataset utilizado de hipotiroidismo posee 2800 sujetos originalmente, existiendo varios sujetos que tienen registros nulos (marcados con '?' en esta base de datos), es decir, hubo varios sujetos los cuales no se realizaron alguno de los exámenes. Se decide eliminar estos sujetos que tienen datos incompletos para este estudio.

3.1.2 Eliminación de variables

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

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

3.1.2.1 TBG

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

3.1.2.2 Fuente de referencia

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

lo que sería contraproducente para el análisis de los datos. Se elimina la variable *referral.source* de la base de datos.

3.1.2.3 Variables de medición

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

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

3.1.3 Transformación de variables y datos

Como se menciona anteriormente, para poder obtener las reglas de asociación, la base de datos debe estar en formato booleano o categórico, por lo que se deben transformar las siguientes variables continuas: age, TSH, T3, TT4, T4U y FTI.

Para transformar estas variables se deciden en rangos numéricos de acuerdo a lo estudiado en la literatura, definiendo los siguientes rangos que reemplazarán a las variables continuas:

Tabla 3.1: Rangos utilizados para transformar las variables continuas en varias variables booleanas.

Variable	Low range	Normal range	High range
age	**child:** 0 - 17	**adult:** 18 - 64	**old:** 65+
TSH	0 - 0.4	0.4 - 4.0	4.0+
T3	0 - 1.07	1.07 - 3.37	3.37+
TT4	0 - 64	64 - 164	164+
T4U	0 - 0.7	0.7 - 1.8	1.8+
FTI	0 - 33.108	33.108 - 135.191	135.191+

3.2 OBTENCIÓN DE REGLAS

Luego de pre-procesar la base de datos, se puede utilizar la función *apriori* para obtener las reglas de asociación. Se utiliza como consecuente de las reglas si es que el paciente padece de hipotiroidismo, lo que corresponde a la variable *results* que indica los resultados obtenidos de los exámenes que se realizó cada sujeto, por lo que se transforma este atributo para que sea una variable binaria, ignorando el tipo de hipotiroidismo y solo tomando en cuenta si es que el paciente padece de algún tipo de hipotiroidismo o no.

Para la obtención de las reglas con los datos procesados y la función mencionada se especifica un soporte mínimo y una confianza mínima de $0{,}01$ y $0{,}5$ respectivamente, además como medida de calidad se utiliza la medida *lift* ya que permite comparar la proporción del soporte observado con el teórico, por lo que es más robusta que la confianza.

Para obtener reglas que sean más fáciles de estudiar, se limita el algoritmo a encontrar reglas que tengan como máximo 6 elementos y como mínimo 2 elementos, esto significa dado que hay un solo consecuente (si el sujeto padece de hipotiroidismo), cada regla puede tener un máximo de 5 antecedentes. Con estas limitaciones se obtuvieron más de 100 reglas con el mismo *lift*, por lo que se decide limitar el máximo de elementos a 5, mientras que el resultado anterior se utiliza para crear la Figura 3.1, las 10 reglas con más *lift* obtenidas con la nueva restricción se muestra a continuación:

1. $\{thyroid.surgery = f, FTI.low = 1, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
2. $\{thyroid.surgery = f, TSH.low = 0, FTI.low = 1\} \Rightarrow \{hypothyroid = 1\}$
3. $\{thyroid.surgery = f, TT4.low = 1, FTI.low = 1, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
4. $\{thyroid.surgery = f, TSH.low = 0, TT4.low = 1, FTI.low = 1\} \Rightarrow \{hypothyroid = 1\}$
5. $\{thyroid.surgery = f, FTI.low = 1, TT4.normal = 0, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
6. $\{thyroid.surgery = f, FTI.low = 1, FTI.normal = 0, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
7. $\{thyroid.surgery = f, FTI.low = 1, TSH.normal = 0, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
8. $\{thyroid.surgery = f, TSH.low = 0, FTI.low = 1, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$
9. $\{thyroid.surgery = f, FTI.low = 1, TSH.high = 1, FTI.high = 0\} \Rightarrow \{hypothyroid = 1\}$
10. $\{on.thyroxine = f, thyroid.surgery = f, FTI.low = 1, TSH.high = 1\} \Rightarrow \{hypothyroid = 1\}$

Obteniendo las siguientes medidas para cada una de las reglas:

Tabla 3.2: Medidas de calidad obtenidas para las 10 reglas con más lift.

Reglas	**Soporte**	**Confianza**	***Lift***
1	0.01130524	0.9565217	11.85600
2	0.01130524	0.9565217	11.85600
3	0.01130524	0.9565217	11.85600
4	0.01130524	0.9565217	11.85600
5	0.01130524	0.9565217	11.85600
6	0.01130524	0.9565217	11.85600
7	0.01130524	0.9565217	11.85600
8	0.01130524	0.9565217	11.85600
9	0.01130524	0.9565217	11.85600
10	0.01130524	0.9565217	11.85600

En las siguientes figuras se puede observar cómo se distribuyen las reglas de acuerdo con los niveles de soporte y confianza, en la esquina superior izquierda del gráfico se ubican las reglas que tienen mayor *lift*.

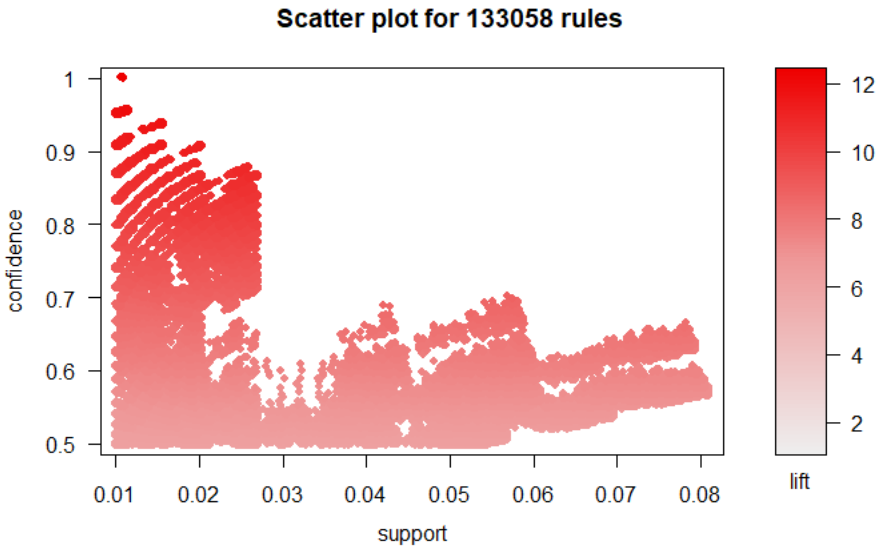

Figura 3.1: Distribución de las reglas de acuerdo al soporte y confianza, mostrando el *lift* de cada regla (con un máximo de 5 antecedentes).

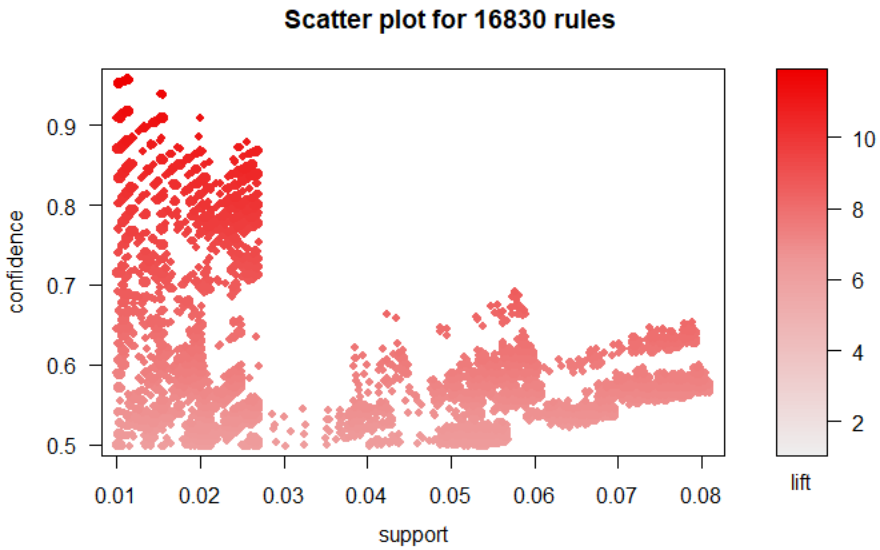

Figura 3.2: Distribución de las reglas de acuerdo al soporte y confianza, mostrando el *lift* de cada regla (con un máximo de 4 antecedentes).

CAPÍTULO 4. ANÁLISIS DE RESULTADOS Y COMPARACIÓN

4.1 ANÁLISIS DE LAS REGLAS OBTENIDAS

Se obtuvieron 16830 reglas con máximo 4 antecedentes (Figura 3.2 en el capítulo anterior, usando el programa anexo a este informe se pueden ver las 40 reglas con mayor *lift*, de estas reglas se escogen las 10 que poseen el mayor *lift* para analizarlas en esta sección.

Teniendo en cuenta que el consecuente de todas las reglas es el hipotiroidismo, se encuentran las siguientes variables involucradas: **thyroid.surgery, FTI.low, TSH.high, TSH.low, TT4.low, TT4.normal, FTI.normal, TSH.normal, FTI.high, on.thyroxine**. Algo interesante de notar antes de analizar individualmente cada una de las 10 reglas con mayor *lift*, es que todas estas reglas *lift* poseen una variable común, esta es {**thyroid.surgery=f**}, lo que significa que hay más casos de gente con algún caso de hipotiroidismo cuando estos pacientes no se han realizado una cirugía que implique a la tiroides. A continuación, se muestra un análisis detallado de las 10 reglas con mayor *lift*:

1. {**thyroid.surgery=f,FTI.low=1,TSH.high=1**}: Revisando la teoría sobre los síntomas que presente un paciente con hipotiroidismo [6], se puede apreciar que en estos sujetos los síntomas comunes son una elevación en el nivel de TSH (hormona reguladora de la producción de hormonas tiroideas) y un descenso del nivel de T4 en el paciente, teniendo en cuenta que el FTI es directamente proporcional al nivel de tiroxina en el sujeto, esta regla concuerda con lo analizado en la teoría.

2. {**thyroid.surgery=f,TSH.low=0,FTI.low=1**}: Ocurre algo similar al caso anterior, solo que decir que el TSH no es bajo puede significar que se encuentre dentro del rango normal, por lo que es más vago y entrega menos información que la regla 1.

3. {**thyroid.surgery=f,TT4.low=1,FTI.low=1,TSH.high=1**}: Al igual que en la regla 1, el FTI es directamente proporcional con el nivel de tiroxina en la sange, por lo que es algo redundante decir que el nivel de tiroxina en el paciente es bajo y que el FTI también es bajo.

4. {**thyroid.surgery=f,TSH.low=0,TT4.low=1,FTI.low=1**}: El mismo caso que sucedió en la regla 2 se presenta en esta con la regla 3, solo que decir que el TSH no es bajo puede significar que se encuentre dentro del rango normal, por lo que es más vago y entrega menos información que la regla 3.

5. {**thyroid.surgery=f,FTI.low=1,TT4.normal=0,TSH.high=1**}: Esta regla no entrega más información que la regla 1, dado que decir que el nivel de tiroxina en el sujeto no está dentro del rango de normalidad es redundante, ya que el FTI del paciente es bajo por lo que el nivel de tiroxina también debería ser bajo (TT4 no es normal). Eliminado esta trivialidad, esta regla se convierte en la regla 1.

6. {**thyroid.surgery=f,FTI.low=1,FTI.normal=0,TSH.high=1**}: Esta regla no entrega más información que la regla 1, dado que decir que el FTI está bajo los rangos normales y que el FTI no está en los rangos normales es obvio. Eliminado esta trivialidad, esta regla se convierte en la regla 1.

7. {**thyroid.surgery=f,FTI.low=1,TSH.normal=0,TSH.high=1**}: Al igual que en la regla anterior los antecedentes TSH.normal=0 y TSH.high=1 realizan una observación inobjetable. Eliminado esta trivialidad, esta regla se convierte en la regla 1.

8. {**thyroid.surgery=f,TSH.low=0,FTI.low=1,TSH.high=1**}: Al igual que en la regla anterior los antecedentes TSH.low=0 y TSH.high=1 realizan una aclaración evidente. Eliminado esta trivialidad, esta regla se convierte en la regla 1.

9. {**thyroid.surgery=f,FTI.low=1,TSH.high=1,FTI.high=0**}: Al igual que en la regla anterior los antecedentes FTI.low=1 y FTI.high=0 realizan un apunte elemental. Eliminado esta trivialidad, esta regla se convierte en la regla 1.

10. {**on.thyroxine=f,thyroid.surgery=f,FTI.low=1,TSH.high=1**}: Esta regla incluye un nuevo antecedente, el cual indica que hay más casos de hipotiroidismo en los pacientes que no están en algún tratamiento con medicación de tiroxina.

Otro antecedente interesante encontrado en el análisis de las reglas, pero que no forma parte de las 10 reglas con mayor *lift* es $\{child = 0\}$, lo que indica que hay más casos de gente adulta o de tercera edad que presenta hipotiroidismo, lo que es lo mismo que decir que hay menos casos de hipotiroidismo en sujetos menores de edad (menores a 18 años).

4.1.1 Simplificación de las reglas

Finalmente, sintetizando todas las reglas obtenidas y eliminando las reglas redundantes, quedan las siguientes reglas:

1. {**thyroid.surgery=f,FTI.low=1,TSH.high=1**}

2. {**thyroid.surgery=f,TT4.low=1,FTI.low=1,TSH.high=1**}

3. {**on.thyroxine=f,thyroid.surgery=f,FTI.low=1,TSH.high=1**}

Donde se repiten 3 antecedentes (que también se repitieron en las 10 reglas con mayor *lift*), estos son: **thyroid.surgery=f, FTI.low=1** y **TSH.high=1**, que son los antecedentes presentes en la regla 1.

4.2 COMPARACIÓN CON LABORATORIOS ANTERIORES

Hay ciertos aspectos obtenidos en este laboratorio que deben ser destacados en comparación a resultados obtenidos en laboratorios anteriores:

- En el laboratorio 1, mientras se analizaban las variables booleanas, se encontró que cuando un sujeto fue operado de la tiroides, la probabilidad era muy baja de padecer de algún tipo de hipotiroidismo, lo que luego se ve reflejado en las reglas obtenidas, dado que todas las reglas con mayor *lift* tienen como antecedente que los sujetos no hayan sido operados de la tiroides anteriormente.
- En cuanto al laboratorio 2, en el clúster 1 obtenido, se tiene la mayor cantidad de pacientes que no fueron operados de la tiroides, exactamente 1040 sujetos de los 1921 que no han sido operados de la tiroides, mientras que este clúster también es el que presenta valores de la FTI en promedio más bajos que los otros clústeres, igualmente con los valores de la TSH que son en promedio más altos que los otros clústeres. Con los conocimientos obtenidos gracias a las reglas obtenidas, ahora se puede inferir que este clúster es el que probablemente tiene una mayor proporción de sujetos que padecen de algún tipo de hipotiroidismo.

CAPÍTULO 5. CONCLUSIONES

En esta experiencia se logran obtener reglas de asociación de la base de datos de hipotiroidismo usando el paquete *arulesViz*, específicamente la función *aprioiri* que trae este paquete y que pide una base de datos transaccional para obtener las reglas de asociación de acuerdo a los parámetros entregados (soporte, confianza, mínimo largo y máximo largo de elementos en la regla). Se tuvo que transformar las variables continuas que trae la base de datos de hipotiroidismo en variables binarias, por lo que se definieron varios rangos donde estas variables iban a ser verdaderas o falsas (*low*, *normal* y *high*).

Realizando el pre-procesamiento de la base de datos y considerando que el consecuente de todas las reglas es el que el sujeto padece de hipotiroidismo, se obtienen 133058 reglas de asociación calculando con 5 antecedentes como máximo (Figura 3.1) y 16830 reglas usando como restricción un máximo de 4 antecedentes (Figura 3.2), considerando que en el primer caso son demasiadas reglas, es muy probable que haya redundancia, por lo que se decidió utilizar el segundo caso, donde se mostraron las mejores reglas de acuerdo a su *lift*.

Aunque se haya decidido utilizar el caso que poseía menos reglas, igual se llega a tener varias reglas redundantes donde la mayoría tienen 3 antecedentes en común, estas son: **thyroid.surgery=f, FTI.low=1** y **TSH.high=1**, mientras que sus otros antecedentes no entregan nada de información, por ejemplo tener **TSH.normal=0** junto a los 3 antecedentes mencionados, no entrega más información. Hay varios otros casos como este que podrían haber sido eliminados, es decir, eliminar las reglas donde aparece la misma variable continua varias veces, pero en distintos estados (*low*, *normal* y *high*), sabiendo que siempre solo 1 de estos será verdadero.

Otros problemas que llevan a que las reglas obtenidas sean redundantes es la base de datos en sí, dado que hay variables que pueden ser calculadas desde otras, como es el caso con el FTI y la TT4, lo que lleva a un crecimiento exponencial en la cantidad de reglas que se pueden encontrar (por las combinaciones que se pueden hacer, sería lo mismo en una regla poner una, la otra o ambas juntas). Solucionar este problema llevaría a tener reglas menos redundantes y que entreguen más información.

CAPÍTULO 6. BIBLIOGRAFÍA

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

[2] I. K. Center, *Reglas de asociación*, 2014. dirección: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/nodes_associationrules.html.

[3] XLSATS, *Reglas de asociación para análisis cesta de compra*, 2017. dirección: https://help.xlstat.com/customer/es/portal/articles/2062425.

[4] A. Monteserin, *Reglas de asociación*, 2018. dirección: http://www.exa.unicen.edu.ar/catedras/optia/public_html/2018%20Reglas%20de%20asociaci%C3%B3n.pdf.

[5] M. Chacón, *Reglas de Asociación*, 2015. dirección: http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=/215450/mod_resource/content/1/Capitulo%20IV%20Reglas%20de%20Asociaci%C3%B3n.pdf.

[6] D. J. C. G. Ferrater, *Hipotiroidismo*, 2018. dirección: https://www.cun.es/enfermedades-tratamientos/enfermedades/hipotiroidismo.

CAPÍTULO 7. ANEXO 1 - CÓDIGO R

```r
#_____________Universidad de Santiago de Chile_____________________
#________Departamento de Ingenieria en Informatica_________________
#____________________Analisis de Datos_________________________________
#
# Laboratorio 3: Reglas de asociacion
# Integrantes: Nicolas Mariangel | Juan Pablo Rojas
# Profesor: Max Chacon
# Ayudante: Ignacio Ibanez Aliaga
#
# http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

library(readr)
library(ggplot2)
library("arulesViz")

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

# Limpieza de los resultados (no se usa el ".|numero" solo interesa la clase)
allhypo$results <- vapply(strsplit(allhypo$results,"\\."), `[`, 1, FUN.VALUE=
    character(1))

# Se eliminan las siguientes columnas porque no fue medida la TBG en la base de
     datos.
allhypo$TBG <- NULL
allhypo$`TBG measured` <- NULL
```

```
27
28 # Luego se procede a cambiar el caracter '?' por NA para poder ejecutar
       complete.cases()
29 allhypo[allhypo=="?"] <- NA
30 allhypo$age[allhypo$age >= 123] <- NA
31
32 # Se eliminan los datos que tienen NA (originalmente '?') por los motivos
33 # descritos en el informe
34 allhypo <- allhypo[complete.cases(allhypo), ]
35
36 # Luego como todas las columnas de si es que se mide o no algo son siempre 't'
37 # por haber eliminado los sujetos, no se usaran estas columnas en el analisis
38 allhypo$`T3 measured` <- NULL
39 allhypo$`T4U measured` <- NULL
40 allhypo$`TSH measured` <- NULL
41 allhypo$`TT4 measured` <- NULL
42 allhypo$`FTI measured` <- NULL
43
44 # Eliminar referral.source porque no entrega informacion relevante.
45 allhypo$`referral source` <- NULL
46
47 # Transformar todas las columnas numericas a variables numericas.
48 allhypo <- transform(allhypo, age = as.numeric(age))
49 allhypo <- transform(allhypo, TSH = as.numeric(TSH))
50 allhypo <- transform(allhypo, T3 = as.numeric(T3))
51 allhypo <- transform(allhypo, TT4 = as.numeric(TT4))
52 allhypo <- transform(allhypo, T4U = as.numeric(T4U))
53 allhypo <- transform(allhypo, FTI = as.numeric(FTI))
54
55 # Transformar todas las columnas booleanas a factores
56 allhypo <- transform(allhypo, sex = as.factor(sex))
57 allhypo <- transform(allhypo, on.thyroxine = as.factor(on.thyroxine))
58 allhypo <- transform(allhypo, query.on.thyroxine = as.factor(query.on.thyroxine
       ))
59 allhypo <- transform(allhypo, on.antithyroid.medication = as.factor(
       on.antithyroid.medication))
60 allhypo <- transform(allhypo, sick = as.factor(sick))
```

```
61 allhypo <- transform(allhypo, pregnant = as.factor(pregnant))
62 allhypo <- transform(allhypo, thyroid.surgery = as.factor(thyroid.surgery))
63 allhypo <- transform(allhypo, I131.treatment = as.factor(I131.treatment))
64 allhypo <- transform(allhypo, query.hypothyroid = as.factor(query.hypothyroid))
65 allhypo <- transform(allhypo, query.hyperthyroid = as.factor(query.hyperthyroid
       ))
66 allhypo <- transform(allhypo, lithium = as.factor(lithium))
67 allhypo <- transform(allhypo, goitre = as.factor(goitre))
68 allhypo <- transform(allhypo, tumor = as.factor(tumor))
69 allhypo <- transform(allhypo, hypopituitary = as.factor(hypopituitary))
70 allhypo <- transform(allhypo, psych = as.factor(psych))
71
72 # Para poder utilizar el paquete arulesViz y utilizar la funcion apriori para
       obtener
73 # las reglas de asociacion, primero se deben transformar todos los datos a
       variables
74 # booleanas, por lo que se crean rangos para transformar las variables
       continuas.
75
76 # Definir limites para las edades
77 age.adult <- 18
78 age.old <- 65
79
80 # Definir valores minimos y maximos para los valores de las hormonas
81 TSH.min <- 0.4
82 T3.min <- 1.07
83 TT4.min <- 64.0
84 T4U.min <- 0.7
85 FTI.min <- 33.108
86
87 # Definicion de valores maximos
88 TSH.max <- 4.0
89 T3.max <- 3.37
90 TT4.max <- 154.0
91 T4U.max <- 1.8
92 FTI.max <- 135.191
93
```

```
94  # Crear vectores con valores 0 para iniciar con la transformacion a valores
        booleanos (0 o 1)
95  allhypo$child <- integer(length(allhypo[[1]]))
96  allhypo$adult <- integer(length(allhypo[[1]]))
97  allhypo$old <- integer(length(allhypo[[1]]))
98  allhypo$TSH.low <- integer(length(allhypo[[1]]))
99  allhypo$T3.low <- integer(length(allhypo[[1]]))
100 allhypo$TT4.low <- integer(length(allhypo[[1]]))
101 allhypo$T4U.low <- integer(length(allhypo[[1]]))
102 allhypo$FTI.low <- integer(length(allhypo[[1]]))
103 allhypo$TSH.normal <- integer(length(allhypo[[1]]))
104 allhypo$T3.normal <- integer(length(allhypo[[1]]))
105 allhypo$TT4.normal <- integer(length(allhypo[[1]]))
106 allhypo$T4U.normal <- integer(length(allhypo[[1]]))
107 allhypo$FTI.normal <- integer(length(allhypo[[1]]))
108 allhypo$TSH.high <- integer(length(allhypo[[1]]))
109 allhypo$T3.high <- integer(length(allhypo[[1]]))
110 allhypo$TT4.high <- integer(length(allhypo[[1]]))
111 allhypo$T4U.high <- integer(length(allhypo[[1]]))
112 allhypo$FTI.high <- integer(length(allhypo[[1]]))
113
114 # Usar los rangos para transformar las variables continuas en "booleanas"
115 # Entregando valores 1 a los vectores recien creados cuando corresponda
116 for(i in 1:length(allhypo[[1]])){
117
118   if(allhypo$age[i] < age.adult){
119     allhypo$child[i] <- 1
120   }else if(allhypo$age[i] >= age.adult & allhypo$age[i] < age.old){
121     allhypo$adult[i] <- 1
122   }else if(allhypo$age[i] >= age.old){
123     allhypo$old[i] <- 1
124   }
125
126   if(allhypo$TSH[i] >= TSH.max){
127     allhypo$TSH.high[i] <- 1
128   }else if(allhypo$TSH[i] <= TSH.min){
129     allhypo$TSH.low[i] <- 1
```

```
130     } else {
131       allhypo$TSH.normal[i] <- 1
132     }
133     if(allhypo$T3[i] >= T3.max){
134       allhypo$T3.high[i] <- 1
135     }else if(allhypo$T3[i] <= T3.min){
136       allhypo$T3.low[i] <- 1
137     } else {
138       allhypo$T3.normal[i] <- 1
139     }
140     if(allhypo$TT4[i] >= TT4.max){
141       allhypo$TT4.high[i] <- 1
142     }else if(allhypo$TT4[i] <= TT4.min){
143       allhypo$TT4.low[i] <- 1
144     } else {
145       allhypo$TT4.normal[i] <- 1
146     }
147     if(allhypo$T4U[i] >= T4U.max){
148       allhypo$T4U.high[i] <- 1
149     }else if(allhypo$T4U[i] <= T4U.min){
150       allhypo$T4U.low[i] <- 1
151     } else {
152       allhypo$T4U.normal[i] <- 1
153     }
154     if(allhypo$FTI[i] >= FTI.max){
155       allhypo$FTI.high[i] <- 1
156     }else if(allhypo$FTI[i] <= FTI.min){
157       allhypo$FTI.low[i] <- 1
158     } else {
159       allhypo$FTI.normal[i] <- 1
160     }
161 }
162
163 allhypo$results <- ifelse(allhypo$results %in% c("primary hypothyroid", "
        secondary hypothyroid", "compensated hypothyroid"), 1, 0)
164
165 # Transformar a factores y eliminar las variables continuas
```

```
166 allhypo$age <- NULL
167 allhypo$TSH <- NULL
168 allhypo$T3 <- NULL
169 allhypo$TT4 <- NULL
170 allhypo$T4U <- NULL
171 allhypo$FTI <- NULL
172 allhypo <- transform(allhypo, child = as.factor(child))
173 allhypo <- transform(allhypo, adult = as.factor(adult))
174 allhypo <- transform(allhypo, old = as.factor(old))
175 allhypo <- transform(allhypo, TSH.low = as.factor(TSH.low))
176 allhypo <- transform(allhypo, T3.low = as.factor(T3.low))
177 allhypo <- transform(allhypo, TT4.low = as.factor(TT4.low))
178 allhypo <- transform(allhypo, T4U.low = as.factor(T4U.low))
179 allhypo <- transform(allhypo, FTI.low = as.factor(FTI.low))
180 allhypo <- transform(allhypo, TSH.normal = as.factor(TSH.normal))
181 allhypo <- transform(allhypo, T3.normal = as.factor(T3.normal))
182 allhypo <- transform(allhypo, TT4.normal = as.factor(TT4.normal))
183 allhypo <- transform(allhypo, T4U.normal = as.factor(T4U.normal))
184 allhypo <- transform(allhypo, FTI.normal = as.factor(FTI.normal))
185 allhypo <- transform(allhypo, TSH.high = as.factor(TSH.high))
186 allhypo <- transform(allhypo, T3.high = as.factor(T3.high))
187 allhypo <- transform(allhypo, TT4.high = as.factor(TT4.high))
188 allhypo <- transform(allhypo, T4U.high = as.factor(T4U.high))
189 allhypo <- transform(allhypo, FTI.high = as.factor(FTI.high))
190 allhypo$results <- as.factor(allhypo$results)
191 names(allhypo)[names(allhypo) == "results"] <- "hypothyroid"
192
193 # Obtener las reglas que sean de largo 2 minimo y largo 6 maximo, teniendo
      minimo soporte de 0.01
194 # y confianza minima de 0.5, se busca encontrar reglas que indiquen que
      atributos llevan a padecer
195 # de hipotiroides
196 #rules <- apriori(allhypo, parameter = list(minlen=2, support=0.01, confidence
      =0.5, maxlen=6), appearance = list(rhs=c("hypothyroid=1"), default="lhs"))
197
198 # Graficar las reglas
199 #plot(rules)
```

```
200
201 # Se analizan solo las reglas que tengan un largo maximo de 5 elementos.
202 rules <- apriori(allhypo, parameter = list(minlen=2, support=0.01, confidence=0
        .5, maxlen=5), appearance = list(rhs=c("hypothyroid=1"), default="lhs"))
203
204 # Revisar las mejores reglas segun lift
205 inspect(head(rules, n = 40, by ="lift"))
206
207 # Graficar las reglas
208 plot(rules)
```