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

INFORME EXPERIENCIA 3 ANÁLISIS DE DATOS

REGLAS DE ASOCIACIÓN

Integrantes:

Marcela Rivera Castro

Kevin Alvarez

Profesor:

Max Chacón

Ayudante:

Adolfo Guzmán

Santiago - Chile

18 de abril de 2018

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS ..... **v**

ÍNDICE DE CUADROS ..... **vi**

CAPÍTULO 1. INTRODUCCIÓN **7**

1.1 MOTIVACIÓN 7

1.2 ORGANIZACIÓN DEL DOCUMENTO 7

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS 7

CAPÍTULO 2. MARCO TEÓRICO **9**

2.1 REGLAS DE ASOCIACIÓN 9

2.1.1 Soporte 9

2.1.2 Confianza 9

2.1.3 Lift 10

2.1.4 Monotonicidad 10

CAPÍTULO 3. OBTENCIÓN DE REGLAS **11**

3.1 PRE-PROCESAMIENTO 11

3.1.1 Eliminación de registros NA 11

3.1.2 Eliminación de variables 11

3.1.2.1 TBG 11

3.1.2.2 Variables de medición 11

3.1.2.3 Fuente de referencia 11

3.1.3 Trasnformación de datos 12

3.1.3.4 Age 12
3.1.3.5 Hormonas 12
3.2 FUNCIÓN UTILIZADA 13
3.3 REGLAS OBTENIDAS 13

CAPÍTULO 4. ANÁLISIS..... 17
4.1 REGLAS OBTENIDAS 17
4.2 MEDIDAS DE CALIDAD 19
4.3 COMPARACIÓN CON K-MEDIAS 20

CAPÍTULO 5. CONCLUSIONES 23

CAPÍTULO 6. BIBLIOGRAFíA..... 25

CAPÍTULO 7. ANEXO: CÓDIGO EN R..... 27

ÍNDICE DE FIGURAS

Figura 3-1: Scatter Plot de reglas obtenidas. 15

ÍNDICE DE CUADROS

Tabla 3.1: Rangos de edad. 12
Tabla 3.2: Rangos de valores para hormonas. 13
Tabla 3.3: Medidas de calidad de reglas. 15

CAPÍTULO 1. INTRODUCCIÓN

1.1 MOTIVACIÓN

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

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

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

1.2 ORGANIZACIÓN DEL DOCUMENTO

El documento consta de cuatro secciones principales: marco teórico para entender el dominio del problema, obtención de reglas seguido de su respectivo análisis de resultados y finalmente conclusiones, donde se indica el aprendizaje obtenido a partir del desarrollo de la experiencia.

1.3 METODOLOGÍAS Y HERRAMIENTAS UTILIZADAS

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

CAPÍTULO 2. MARCO TEÓRICO

2.1 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. (Center, s.f.)

La manera general de obtener las conclusiones es a través de reglas, donde : $X \Rightarrow$Y

Donde X e Y son conjuntos de ítems. X es nominado antecedente (condición) de la regla e Y su consecuente (conclusión). Es necesario medir de alguna forma la calidad de las reglas de asociación, para esto es necesario atender las reglas en base a diferentes métricas y diferentes medidas de calidad, las cuales son:

2.1.1 Soporte

Probabilidad de encontrar elementos o conjuntos de elementos X e Y en una transacción. Se estima por el número de veces que ambos elementos o conjuntos de elementos se encuentran en todas las transacciones disponibles. Este valor se encuentra entre 0 y 1 (de soporte XLSTAT, s.f.). El soporte normalizado se expresa de la forma:

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

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 (de soporte XLSTAT, s.f.). La confianza se expresa de la forma:

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

2.1.3 Lift

Indica cuándo una regla es mejor prediciendo el resultado que asumiendo el resultado de forma aleatoria. Si el resultado es mayor que uno, la regla es buena, pero si es menor que uno, es peor que elegir un resultado aleatorio. Se muestra en la ecuación 2.107 (López, 2006). Su ecuación es:

$$lift(A \Rightarrow B) = \frac{confianza(A \Rightarrow B)}{P(B)} \tag{2.3}$$

Cabe destacar que en el actual proyecto, se hará uso de lift.

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 (Chacón, 2015).

CAPÍTULO 3. OBTENCIÓN DE REGLAS

3.1 PRE-PROCESAMIENTO

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

3.1.1 Eliminación de registros NA

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

3.1.2 Eliminación de variables

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

3.1.2.1 TBG

Esta variable dentro del dataset tiene mediciones que no fueron realizadas, ya que todos los datos son registros desconocidos, dado esto, la variable no entrega información y por ende se puede eliminar.

3.1.2.2 Variables de medición

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

TSH measured, **T3 measured**, **TT4 measured**, **T4U measured**, **FTI measured**.

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

3.1.2.3 Fuente de referencia

Esta variable indica la fuente de la cual se obtuvo los datos de un sujeto (u observación) en particular, por lo que esta variable no entrega información útil, por lo que se puede

eliminar.

3.1.3 Trasnformación de datos

Dado que para obtener las reglas de asociación es necesario que las variables de entrada sean binarias, es necesario hacer una transformación de las que son continuas a binarias, en este caso, las variables a trasformar son: age, TSH, T3, TT4, T4U y FTI.

3.1.3.4 Age

Para el caso de la variable de edad, dentro de ésta se pueden encontrar tres grupos: niños, adultos y ancianos, los cuales se dividen de acuerdo a los rangos de edad, los cuales son:

Tabla 3.1: Rangos de edad.

Grupo	**Mínimo**	**Máximo**
Niño	0	17
Adulto	18	59
Anciano	60	-

Para representar estos datos, se agregan tres variables binarias nuevas a la base de datos, las cuales son: age.child, age.adult, age.oldman, las cuales representan si el sujeto es niño, adulto o anciano. Cabe destacar que solo una de estas puede ser verdadera por observación.

3.1.3.5 Hormonas

Para el caso de las hormonas se analiza si ésta esta en el rango aceptable, para esto se puede verificar si la medición está por debajo del rango normal, o por encima, para ello se crean dos variables binarias nuevas en la base de datos para cada hormona con el sufijo ünder"para indicar si el valor está por debajo del normal, o .ºver"para indicar si esta por encima de lo normal. Cabe destacar que si un sujeto tiene niveles hormonales normales ambas variables van a ser falsas.

Los rango para la transformación son los mismo utilizados en la experiencia anterior, los cuales son:

Tabla 3.2: Rangos de valores para hormonas.

Hormona	Rango Mínimo (under)	Rango Máximo (over)
TSH	0.4	4.0
T3	1.07	3.37
TT4	64	164
T4U	0.7	1.8
FTI	33.108	135.191

3.2 FUNCIÓN UTILIZADA

La función utilizada es apriori la cual pertenece a la librería .ªrules", el algoritmo de Apriori emplea la búsqueda del nivel-sabio para los itemsets frecuentes. Esta recibe como entrada: los datos, parámetro objeto de clase para definir soporte y confianza, se recibe un parámetro apariencia con la cual se puede restringir (implementa plantillas de reglas), finalmente se tiene un parámetro controlar el cual controla el rendimiento algorítmico del algoritmo de minería (rdocumentation, s.f.).

Para el consecuente de las reglas se hará uso de la clasificación del dataset, dado a que es necesario que la variable sea binaria, se realiza una transformación de ésta para que indique si el sujeto fue clasificado con hipotiroidismo, por lo que se ignora el tipo del cuadro, y solo se verifica si lo padece.

3.3 REGLAS OBTENIDAS

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 mas robusta que la confianza.

Para obtener reglas que sean mejores y entreguen mas información para el estudio, se aplicaron restricciones en los parámetros del algoritmo, éstas son que el mínimo de antecedentes y consecuentes que debe tener una regla son 2, y así mismo se estableció un máximo de 5, cabe destacar que se quiere encontrar como único consecuente la variable que clasificación, por lo que finalmente cada regla puede tener un máximo de 4 antecedentes. Con esto se obtuvo un total de 8367, a continuación se muestran las 10 reglas con

más lift obtenidas:

1. $\{I131.treatment=f, TSH.over=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
2. $\{thyroid.surgery=f, TSH.over=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
3. $\{I131.treatment=f, TSH.under=0, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
4. $\{thyroid.surgery=f, TSH.under=0, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
5. $\{I131.treatment=f, TSH.over=1, TT4.under=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
6. $\{thyroid.surgery=f, TSH.over=1, TT4.under=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
7. $\{I131.treatment=f, TSH.under=0, TT4.under=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
8. $\{thyroid.surgery=f, TSH.under=0, TT4.under=1, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
9. $\{I131.treatment=f, TSH.over=1, TSH.under=0, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
10. $\{thyroid.surgery=f, TSH.over=1, TSH.under=0, FTI.under=1\} \Rightarrow \{hypothyroid=1\}$

Las medidas de cada regla son las siguientes:

Tabla 3.3: Medidas de calidad de reglas.

Regla	**Soporte**	**Confianza**	**Lift**
1	0.01129944	0.9565217	11.86209
2	0.01129944	0.9565217	11.86209
3	0.01129944	0.9565217	11.86209
4	0.01129944	0.9565217	11.86209
5	0.01129944	0.9565217	11.86209
6	0.01129944	0.9565217	11.86209
7	0.01129944	0.9565217	11.86209
8	0.01129944	0.9565217	11.86209
9	0.01129944	0.9565217	11.86209
10	0.01129944	0.9565217	11.86209

En la figura 3-1 se puede observar como se distribuyen las reglas de acuerdo a los niveles de soporte y confianza, en la esquina superior izquierda del gráfico se ubican las reglas mostradas anteriormente, que tienen mayor *lift*

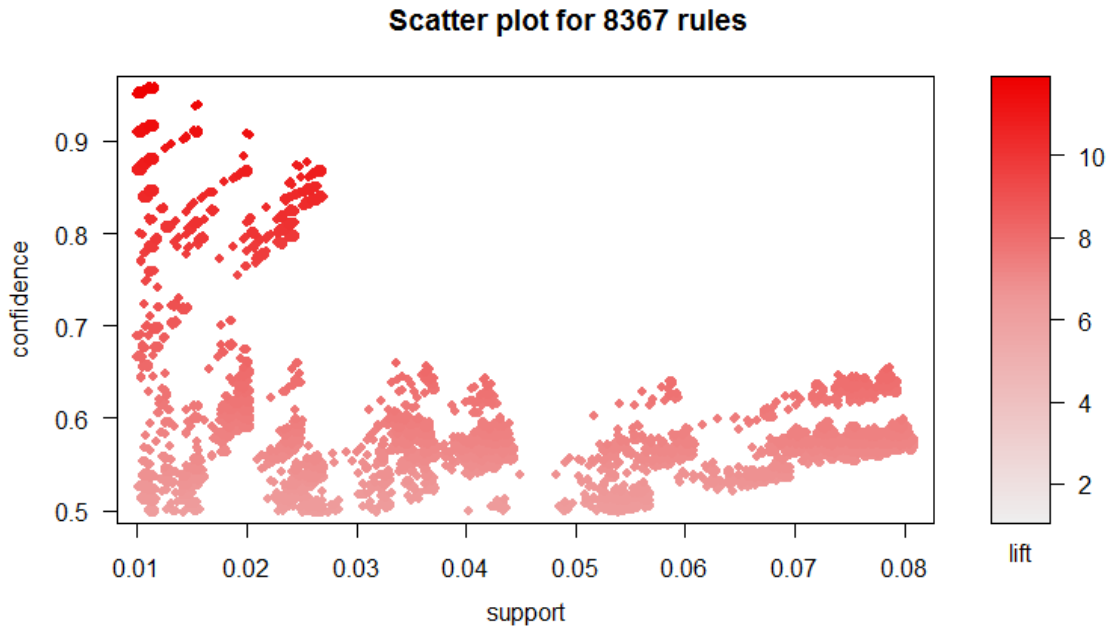

Figura 3-1: Scatter Plot de reglas obtenidas.

CAPÍTULO 4. ANÁLISIS

4.1 REGLAS OBTENIDAS

A partir de los resultados expuestos en el capítulo anterior, se puede observar que las reglas obtenidas son 8367, todas éstas se encuentran ordenadas según el valor lift obtenido para cada una de las reglas, donde las primeras 10 son las que mayor valor tienen.

A partir del estudio de las primeras 10 reglas, es posible observar que se tienen como variables involucradas a: **TSH**, **thyroid.surgery**, **TT4 measured**, **I131.treatment**, **FTI**. Donde a partir de la combinación de los valores de cada uno, se obtiene como consecuente hipotiroidismo. A continuación se realizará un análisis detallado para cada una de las reglas.

1. Cuando {**I131.treatment=f, TSH.over=1, FTI.under=1**}: se obtiene que el consecuente es hipotiroidismo, lo cual concuerda con la teoría ya que se sabe que cuando los pacientes padecen de este cuadro, tienden a sufrir una disminución de secreción de la hormona T3, la cual produce una hipersecreción de TSH (valores altos). Lo anterior, sumado a un bajo nivel de FTI, genera como consecuencia hipotiroidismo.

2. Cuando {**thyroid.surgery=f, TSH.over=1, FTI.under=1**}: ocurre un fenómeno similar al anterior ya que nuevamente TSH es alta y FTI baja, por lo que al no tener cirugías en su tiroides, se tiene como consecuente la condición de hipotiroidismo.

3. Cuando {**I131.treatment=f, TSH.under=0, FTI.under=1**}: al observar que TSH.under es igual a 0, se tiene que la hormona TSH es alta. Por lo tanto, esta regla es totalmente redundante ya que se está diciendo lo mismo que en la regla 1.

4. Cuando {**thyroid.surgery=f, TSH.under=0, FTI.under=1**}: en esta regla ocurre lo mismo que en el caso 3, se está diciendo que TSH es alta, lo que genera una regla idéntica a la dos, lo que convierte a esta regla en redundante.

5. Cuando {**I131.treatment=f, TSH.over=1, TT4.under = 1, FTI.under=1**}: en este caso, se tiene la condición de hipotiroidismo cuando TSH es alta y TT4 con FTI baja.

Esto se encuentra acorde a la teoría, ya que la FTI indica la cantidad de T4 libre que está presente en comparación con T4 atado, además puede ayudar a determinar si las cantidades anormales de T4 están presentes debido a cantidades anormales de globulina fijadora de tiroxina. (Templates, 2013).

Por lo tanto, si la FTI es baja, la TT4 también y esto sumado a un TSH alto, se obtiene como consecuencia hipotiroidismo.

6. Cuando {**thyroid.surgery=f, TSH.over=1, TT4.under = 1, FTI.under=1**}: en este caso, ocurre un suceso similar al 5 ya que se tiene la FTI, TT4 baja y la TSH alta, por lo tanto, esto sumado a que no se tiene una operación de tiroides, provoca que la persona padezca de hipotiroidismo.

7. Cuando {**I131.treatment=f, TSH.under=0, TT4.under = 1, FTI.under=1**}: como en el caso de la regla 3, nuevamente se tiene una regla redundante, esto porque se construye la regla con TSH.under = 0, dejando expresado lo dicho por la regla 5.

8. Cuando {**thyroid.surgery=f, TSH.under=0, TT4.under = 1, FTI.under=1**}: igual al caso anterior, solo que esto ocurre cuando la cirugía a tiroides no ha sido hecha.

9. Cuando {**I131.treatment=f, TSH.over=1, TSH.under = 0, FTI.under=1**}: en este caso, se tiene una regla compuesta por dos variables que dicen lo mismo, es decir, TSH.over = 1 indica que tiene alto el nivel de TSH, por otro lado TSH.under = 0 indica que no es bajo el nivel de TSH. Finalmente, esto es trivial ya que decir que TSH es alto y no es bajo, es lo mismo.

   Basándose en este último punto, se puede observar que la regla es igual a la regla 1, ya que ambas consideran un nivel alto de TSH. Esto implica que la actual regla es redundante.

10. Cuando {**thyroid.surgery=f, TSH.over=1, TSH.under = 0, FTI.under=1**}: aquí se repite el caso anterior, por lo que es una regla redundante al decir lo mismo que la regla dos.

Finalmente, las reglas que aportan información relevante son:

1. $\{I131.treatment=f,\ TSH.over=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
2. $\{thyroid.surgery=f,\ TSH.over=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
3. $\{I131.treatment=f,\ TSH.over=1,\ TT4.under=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$
4. $\{thyroid.surgery=f,\ TSH.over=1,\ TT4.under=1,\ FTI.under=1\} \Rightarrow \{hypothyroid=1\}$

Al analizar estas reglas, es posible ver como las hormonas son las capaces de determinar si un paciente padece o no de hipotiroidismo. Esta conclusión está acorde con la teoría ya que en la actualidad, una de las formas de detectar hipotiroidismo, es a partir de las mediciones de estas hormonas.

4.2 MEDIDAS DE CALIDAD

Las medidas de calidad nos permiten ver que reglas son mas interesantes dentro del conjunto, en esta instancia se utilizan 3: soporte, confianza y lift. El soporte es la proporción de observaciones que tienen las variables de la regla, la confianza permite definir que tan fiable es la regla y el *lift*, que dado el supuesto de independencia de las observaciones, muestra la proporción del soporte observado con respecto al soporte teórico.

De acuerdo a lo visto en la tabla 3.3, las reglas presentan un nivel alto de confianza y *lift*, pero tienen un nivel bajo de soporte, esto quiere decir que las reglas no son tan frecuentes, pero que cada una tiene una gran fiabilidad, con un 95 % de que si se se tienen esas variables como antecedentes, el consecuente va a ser el cuadro de hipotiroidismo. El *lift* permite observar las variables desde un supuesto de independencia entre éstas, para este estudio, se obtuvo un *lift* de 11.86209, un número mucho mayor a 1, por lo que se puede deducir que existe una fuerte relación entre las variables utilizadas en el estudio, que hace que la cantidad de observaciones sea mayor a lo esperado, de acuerdo a la teoría esto se ve apoyado, dado a que los niveles hormonales no son independientes y están fuertemente relacionados entre sí.

En la figura 3-1 se muestra la distribución de todas las reglas obtenidas de acuerdo a soporte, confianza y *lift*, donde es posible observar que las reglas con más lift están

ubicadas en las reglas con menor soporte y mayor confianza, también se ve que existen reglas con menor confianza pero con un soporte grande, estas reglas son las mas simples, con 1 o 2 antecedentes, por lo que las variables involucradas son bastante frecuente, es por ello que tienen un gran soporte, pero en si no entregan tanta información. Un detalle importante que se puede observar, es que el *lift* de las reglas en todas es mayor a 1 lo que denota una relación entre las variables, es decir, posiblemente ninguna de estas son independientes entre si.

4.3 COMPARACIÓN CON K-MEDIAS

Se puede realizar una comparación entre el método de reglas de asociación usado en esta instancia, y el algoritmo de clustering k-medias utilizado en la experiencia anterior. La diferencia principal de estos métodos recae en que reglas de asociación es un método supervisado y k-medias, al ser clustering, es no supervisado, es decir, al usar reglas de asociación se tiene una variable objetivo(en este caso la variable ”hypothyroid”) y en k-medias no se puede asegurar que éste agrupe de acuerdo a lo que se quiere hallar.

En relación a los resultados obtenidos de ambos laboratorios, en primera instancia k-medias agrupó por sexo, dado a que daba mayor significancia a las primeras variables, para obtener mas información sobre el cuadro del paciente se realizó un nuevo agrupamiento solo con las variables de edad y hormonas para así observar el comportamiento de los sujetos en relación a éstas, lo cual resultó en 3 grupos donde se podía observar la tendencia al hipotiroidismo, para reglas de asociación las reglas obtenidas muestran una relación entre TSH alto, FTI bajo y TT4 bajo, donde se tiene por consecuencia el cuadro de hipotiroidismo.

Si se comparan los resultados obtenidos, no se puede observar claramente una relación entre ellos, pero si se complementan entre sí, se puede observar que en k-medias, el agrupamiento realizado se hace en relación a los niveles hormonales, donde los sujetos que tienen mayor tendencia al hipotiroidismo son los que tienen niveles altos de TSH y bajos de las otras hormonas, llevando esto a lo obtenido con las reglas de asociación, se tienen reglas que indican que un alto nivel de TSH en conjunto con un nivel bajo de T4, dado lo mostrado en los niveles bajos de FTI y TT4, tiene como consecuente un cuadro

de hipotiroidismo, con esto se puede decir que con el algoritmo k-medias se agruparon los sujetos de acuerdo a las variables observadas, y con reglas de asociación se clasificó, o identificó, uno de los grupos como los que padecen el cuadro de hipotiroidismo dado sus niveles hormonales.

CAPÍTULO 5. CONCLUSIONES

En la experiencia se realizó el estudio utilizando el método de reglas de asociación para determinar si un sujeto tiene o no un cuadro de hipotiroidismo, dado que el método es supervisado, se enfoca el estudio en la variable de clase como consecuente.

Para lograr el objetivo y obtener las reglas, primero se realizó un procesamiento de los datos, dado a que es necesario que éstos sean binarios, se transformaron las variables continuas de tal forma que los valores binarios sean representativos, también se transformó la variable de clase de tal forma que indicara si el sujeto tiene un cuadro de hipotiroidismo o no.

Los resultados obtenidos fueron concluyentes en cuanto a lo esperado, dado a que se observa una clara relación entre los niveles hormonales y padecer el cuadro de hipotiroidismo, ya que se puede observar que las variables mas involucradas en las reglas con mejores medidas de calidad son las que hacen referencia a los niveles de hormonas TSH y T4, aún así las reglas no entregan información sobre los niveles hormonales de hormona T3, por lo que se puede decir que no tiene una relación determinista con el cuadro.

Si se compara con los resultados obtenidos de la experiencia anterior, el agrupamiento de los sujetos en relación a las hormonas realizado con el algoritmo k-medias, mostró una distribución donde se tienen niveles altos de TSH y niveles bajos de T4 y T3, además en la distribución se pudieron identificar 3 grupos, donde se asignaron a la tendencia al padecimiento de la enfermedad. Al analizar en conjunto con el estudio actual, se pudieron observar las mismas relaciones de las hormonas con el cuadro, donde se pueden clasificar a los sujetos de acuerdo a las reglas obtenidas.

Como aprendizaje, esta fue un primer acercamiento a un método de aprendizaje supervisado, donde se tiene una variable objetivo la cual se obtiene de acuerdo al análisis de las variables de entrada. Para este caso, se tiene como objetivo la clasificación de los sujetos, y con las reglas fue posible analizar cuales variables son las mas involucradas en el procedimiento.

Las dificultades halladas en la experiencia fueron darle un significado a las reglas y analizar como interactúan las variables (antecedentes) para llegar a un consecuente,

además de no saber con exactitud si el modelo representa la realidad, dado al escaso conocimiento del problema y sus factores.

Para una próxima instancia se busca estudiar nuevos métodos de este tipo, para poder compararlos, y averiguar cual es mejor para el problema, para así realizar una mejor clasificación y tener mejores conclusiones del estudio.

CAPÍTULO 6. BIBLIOGRAFÍA

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

Center, I. K. (s.f.). Reglas de asociación. Recuperado desde https://www.ibm.com/support/knowledgecenter/es/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/nodes_associationrules.htm

Chacón, M. (2015). "Reglas de Asociación". Recuperado desde http://www.udesantiagovirtual.cl/moodle2/pluginfile.php?file=%5C%2F215450%5C%2Fmod_resource%5C%2Fcontent%5C%2F1%5C%2FCapitulo%5C%20IV%5C%20Reglas%5C%20de%5C%20Asociaci%5C%C3%5C%B3n.pdf

de soporte XLSTAT, C. (s.f.). Reglas de asociación para análisis cesta de compra. Recuperado desde https://help.xlstat.com/customer/es/portal/articles/2062425-reglas-de-asociaci%5C%C3%5C%B3n-para-an%5C%C3%5C%A1lisis-cesta-de-compra?b_id=9283

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

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

López, J. M. M. (2006). TÉCNICAS DE ANÁLISIS DE DATOS. Recuperado desde http://www.academia.edu/6078707/T%5C%C3%5C%89CNICAS_DE_AN%5C%C3%5C%81LISIS_DE_DATOS_APLICACIONES_PR%5C%C3%5C%81CTICAS_UTILIZANDO_MICROSOFT_EXCEL_Y_WEKA

rdocumentation. (s.f.). a priori. Recuperado desde https://www.rdocumentation.org/packages/arules/versions/1.5-3/topics/apriori

Templates, M. B. (2013). Las pruebas de la hormona tiroidea. Recuperado desde http://listaexamenesmedicos.blogspot.cl/2013/04/las-pruebas-de-la-hormona-tiroidea.html

CAPÍTULO 7. ANEXO: CÓDIGO EN R

```
library("arules")
library("arulesViz")



rawdata <- read.csv("allhypo.data", header = FALSE, sep = ",",
    stringsAsFactors = FALSE)
colnames(rawdata) <- c("age", "sex", "on.thyroxine", "query.on.
    thyroxine", "on.antithyroid.medication", "sick", "pregnant", "
    thyroid.surgery", "I131.treatment", "query.hypothyroid", "query.
    hyperthyroid", "lithium", "goitre", "tumor", "hypopituitary", "psych
    ", "TSH.measured", "TSH", "T3.measured", "T3", "TT4.measured", "TT4"
    , "T4U.measured", "T4U", "FTI.measured", "FTI", "TBG.measured", "TBG
    ", "referral.source", "clasification|id")

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

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

```
    & rawdata$T4U.measured!="f" & rawdata$FTI.measured!="f"),]
#delete variable TBG
data$TBG.measured <- NULL
data$TBG <- NULL
#delete measuring variables
data$TSH.measured <- NULL
data$T3.measured <- NULL
data$TT4.measured <- NULL
data$T4U.measured <- NULL
data$FTI.measured <- NULL
#delete referral.source variable
data$referral.source <- NULL

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

#continuous variables
data$age <-as.numeric(data$age)
```

```
data$TSH <- as.numeric(data$TSH)
data$T3 <- as.numeric(data$T3)
data$TT4 <- as.numeric(data$TT4)
data$T4U <- as.numeric(data$T4U)
data$FTI <- as.numeric(data$FTI)

#continuous variables and clasification to binary
#age values
child.adult_border <- 18
adult.oldman_border <- 60
#min values
TSH.min <- 0.4
T3.min <- 1.07
TT4.min <- 64.0
T4U.min <- 0.7
FTI.min <- 33.108
#max values
TSH.max <- 4.0
T3.max <- 3.37
TT4.max <- 154.0
T4U.max <- 1.8
FTI.max <- 135.191

#vectors(zeros)
child <- integer(length(data[[1]]))
adult <- integer(length(data[[1]]))
oldman <- integer(length(data[[1]]))
TSH.under <- integer(length(data[[1]]))
T3.under <- integer(length(data[[1]]))
TT4.under <- integer(length(data[[1]]))
T4U.under <- integer(length(data[[1]]))
FTI.under <- integer(length(data[[1]]))
TSH.over <- integer(length(data[[1]]))
T3.over <- integer(length(data[[1]]))
```

```
TT4.over <- integer(length(data[[1]]))
T4U.over <- integer(length(data[[1]]))
FTI.over <- integer(length(data[[1]]))

#change data to binary
for(i in 1:length(data[[1]])){
  #age
  if(data$age[i] < child.adult_border){
    child[i] <- 1
  }else if(data$age[i] >= child.adult_border & data$age[i] < adult.
      oldman_border){
    adult[i] <- 1
  }else if(data$age[i] >= adult.oldman_border){
    oldman[i] <- 1
  }
  #hormones
  if(data$TSH[i] >= TSH.max){
    TSH.over[i] <- 1
  }else if(data$TSH[i] <= TSH.min){
    TSH.under[i] <- 1
  }
  if(data$T3[i] >= T3.max){
    T3.over[i] <- 1
  }else if(data$T3[i] <= T3.min){
    T3.under[i] <- 1
  }
  if(data$TT4[i] >= TT4.max){
    TT4.over[i] <- 1
  }else if(data$TT4[i] <= TT4.min){
    TT4.under[i] <- 1
  }
  if(data$T4U[i] >= T4U.max){
    T4U.over[i] <- 1
  }else if(data$T4U[i] <= T4U.min){
```

```
    T4U.under[i] <- 1
  }
  if(data$FTI[i] >= FTI.max){
    FTI.over[i] <- 1
  }else if(data$FTI[i] <= FTI.min){
    FTI.under[i] <- 1
  }
}

data$clasification <- ifelse(data$clasification %in% c("primary
    hypothyroid", "secondary hypothyroid", "compensated hypothyroid"),
    1, 0)

#replace vectors on data frame
data$age <- NULL
data$TSH <- NULL
data$T3 <- NULL
data$TT4 <- NULL
data$T4U <- NULL
data$FTI <- NULL
data$age.child <- as.factor(child)
data$age.adult <- as.factor(adult)
data$age.oldman <- as.factor(oldman)
data$TSH.over <- as.factor(TSH.over)
data$T3.over <- as.factor(T3.over)
data$TT4.over <- as.factor(TT4.over)
data$T4U.over <- as.factor(T4U.over)
data$FTI.over <- as.factor(FTI.over)
data$TSH.under <- as.factor(TSH.under)
data$T3.under <- as.factor(T3.under)
data$TT4.under <- as.factor(TT4.under)
data$T4U.under <- as.factor(T4U.under)
data$FTI.under <- as.factor(FTI.under)
data$clasification <- as.factor(data$clasification)
```

```
names(data)[names(data) == "clasification"] <- "hypothyroid"

#rules without restrictions, with min 1 antecedents and max 4
    antecedents [support=0.01, Confidence=0.5]
rules <- apriori(data, parameter = list(minlen=2, support=0.01,
    confidence=0.5, maxlen=5), appearance = list(rhs=c("hypothyroid=1"),
     default="lhs"))

#sort rules
rules.sorted <- sort(rules, by="lift")
inspect(head(rules.sorted, 100))

#scatter plot for rules
scatter.rules <- plot(rules.sorted)
```