UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Informática

Análisis de Datos

Laboratorio 3 Análisis Estadístico

Gustavo Hurtado

Patricia Melo

Profesor: Max Chacón

Ayudante: Javier Arredondo

Santiago – Chile

2020

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

CAPÍTULO 1. INTRODUCCIÓN

Esta experiencia corresponde a la tercera etapa del conjunto de laboratorios que se desarrollan, todos ellos trabajan con la base de datos $allhyper$, la cual contiene datos correspondientes al año 1987. Esta base de datos se obtuvo desde la página UCI Machine Learning Repository.

El objetivo principal de esta experiencia es extraer conocimiento o información relevante de la base de datos utilizando reglas de asociación.

El informe en un inicio consta de un marco teórico en el cual se darán las algunas definiciones a utilizar a lo largo del documento, luego se muestra el proceso de la obtención de reglas sobre los registros de la base de datos. Se continúa con el análisis de los resultados, realizando una comparación entre los datos obtenidos del laboratorio 2 y el presente. Finalmente se tiene las conclusiones y referencias del laboratorio.

CAPÍTULO 2. MARCO TEÓRICO

A continuación se presenta una serie de definiciones que ayudarán para el entendimiento del presente informe.

2.1 REGLAS DE ASOCIACIÓN

Según Cristina Gil (2020), la minería de reglas de asociación es utilizada para descubrir patrones de objetos o atributos que ocurren de manera simultanea. Si se tienen dos conjuntos de ítems, A y B, una regla entre ellos sería la expresión $A => B$, como aparece en la FIgura 2.1. Además, las reglas de asociación permiten establecer relaciones entre variables cualitativas.

Cabe mencionar que el antecedente (parte izquierda de la regla) indica que relacionas provocan algo y el consecuente (parte derecha de la regla) es la conclusión a la que llega la regla.

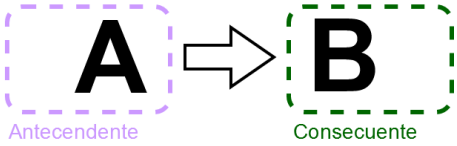

Figura 2.1: Ejemplo regla de asociación

2.2 MEDIDAS DE CALIDAD Y CONFIANZA

Las medidas de calidad utilizadas comúnmente en las reglas de asociación son el soporte y la confianza, pero además en esta experiencia se utilizará lift, por lo que también se procede a explicar.

Para las siguientes definiciones considere los conjuntos de ítems A y B.

- **Soporte**

  Puede aplicarse al conjunto A, $Sop(A)$, o a la regla $A => B$ escrito como $Sop(A => B)$.

  En el primer caso se define como el número transacciones de los ítems pertenecientes a un conjunto A que toman valor verdadero, es decir, que entre todas las transacciones, ve cuantas veces se encuentran todos los ítems que cumplen con el conjunto A.

  En el caso del soporte de una regla $A => B$, entrega el número de todas las transacciones en que A y B se cumplen simultáneamente.

- **Confianza**

  La confianza de una regla $A => B$ se define como la proporción entre el número de casos en que A y B aparecen simultáneamente, sobre el número de casos de A. Lo anterior se expresa en la Ecuación 2.1 .

$$Conf(A => B) = \frac{Sop(A => B)}{Sop(A)} \tag{2.1}$$

  Si a la Ecuación 2.1 se le divide tanto en el numerador como denominador por n (total de las transacciones), da como resultado la probabilidad condicional de B sobre A: $p(B/A)$. Por lo tanto la confianza de la regla $A => B$ representa la probabilidad de que se encuentren los ítems de B en la transacción dado que también contenga los ítems de A.

- **Lift**

  Representa la independencia entre A y B, esto se calcula como la confianza de la regla $A => B$ divido por el sopoorte de B, tal como se aprecia en la Ecuación 2.2

$$lift = \frac{Conf(A => B)}{Sop(B)} \tag{2.2}$$

  Según Ariel Monteserin (2018) se analizan 3 posibles casos de resultado.

  - lift = 1, existe una independencia completa entre los conjuntos A y B.

- lift $>$ 1, el conjunto de los ítems $A\ y\ B$ aparece una cantidad de veces mayor a lo esperado, pudiendo intuir que existe alguna relación que haga que dichos ítems se encuentren más veces de lo normal en conjunto.
- lift $<$ 1, los ítems $A\ y\ B$ aparecen una cantidad de veces menor a lo esperado, intuyendo así que existe alguna relación que haga que los ítems no formen parte del mismo conjunto más veces de lo normal.

2.3 PROPIEDADES DE LA MEDIDA

Existen conjuntos que cumplen con ciertas cualidades, estas se detallan a continuación.

- **Reglas confiables:** Es el conjunto de todas las reglas que cumplen con una confianza mínima ($minconf$).
- **Reglas frecuentes:** Es el conjunto de todas reglas que cumplen con un soporte mínimo ($minsop$).
- **Reglas interesantes:** Son aquellas reglas que cumplen con un soporte mínimo y una confianza mínima.
- **Conjunción:** Se habla de conjunción cuando un conjunto para ser verdadero debe tener todos sus ítems como verdaderos, por ejemplo $A = V_1 \bigwedge V_2 \bigwedge V_3$ debe cumplir con que en una transacción esos tres elementos deben estar presentes para que A sea verdadero.
- **Disyunción:** Cuando un conjunto se hace verdadero si uno o más ítems son verdaderos, por ejemplo $A = V_1 \bigvee V_2 \bigvee V_3$, en este caso para que A sea cierto entonces al menos un ítem perteneciente a A debe ser verdadero.
- **Especialización y generalización:** Para explicar estos temas se utilizará el siguiente ejemplo:

Se tienen los ítems V_i donde $i = 1, 2, 3$, y se define el conjunto $A_1 = V_1, V_2, V_3$ y $A_2 = V_1 \bigwedge V_2, V_1 \bigwedge V_3, V_2 \bigwedge V_3$. En este caso se puede decir que el conjunto A_1 es más generalizado que A_2, y que a su vez el conjunto A_2 esta más especializado que A_1.

2.4 MONOTONICIDAD

La monotonicidad se aplica a las medidas de calidad (soporte, confianza, etc.) en donde indica si es que la medida empleada es monótona o no. Esto quiere decir, que la medida aplicada en todos los conjuntos donde están los mismos ítems, generalizados como especializados, debe siempre estar bajando o siempre estar subiendo.

Continuando con el ejemplo dado en la definición de especialización y generalización, y utilizando el soporte como medida, se sabe que $Sop(A_1) > Sop(A_2)$, esto porque es más probable obtener los ítems de manera separada que varios simultáneamente, y si se tuviera un conjunto A_3 más especializada que A_2, entonces se tendría que $Sop(A_1) > Sop(A_2) > Sop(A_3)$, demostrando así que la medida soporte es monótona, ya que a medida que se especializa éste siempre baja.

Para el caso de la confianza resulta más complicado, esto porque tiene a $Sop(A => B)$ en el numerador y $Sop(A)$ en el denomidador, como se sabe que el soporte es monótono entonces $Sop(A_1) > Sop(A_2)$, y también $Sop(A_1 => B) > Sop(A_2 => B)$. Si tanto numerador como denominador bajan en la misma proporción se tendría como resultado $Conf(A_1 => B) > Conf(A_2 => B)$, pero puede darse el caso en que $Sop(A_2)$ baje mucho más que $Sop(A_2 => B)$, lo que provocaría que $Conf(A_1 => B) < Conf(A_2 => B)$, demostrando así que la confianza no es monótona, ya que puede bajar o subir dependiendo de la situación.

Con respecto al $Lift$, éste es monótono en confianza.

CAPÍTULO 3. OBTENCIÓN DE REGLA

En este capítulo se verá el desarrollo de la obtención de las reglas más interesantes para la base de datos $allhyper$, utilizando las medidas de calidad soporte, confianza y lift.

3.1 PRE-PROCESAMIENTO

Previo a la obtención de las reglas se debe realizar un pre-procesamiento de los datos, el cual ya fue efectuado en las experiencias anteriores, sin embargo, para hacer uso de reglas de asociación aún quedan tratamientos por hacer a los datos, los cuales se presentan a continuación.

3.1.1 Discretización

Se tienen 6 variables continuas, a las que se procede definir los límites y de esta forma discretizarlas.

Las variables TSH, T3, TT4, T4U y FTI se separaron según sus rangos bajos, normales y altos. Estos rangos fueron descritos en la experiencia anterior, por lo que en la Tabla 3.1 se presentan los valores que toman cada variable de las recién mencionadas.

Tabla 3.1: Variables continuas y sus valores normales

Variable	Rangos		
	Bajo	Normal	Alto
TSH	$-\infty - 0{,}39$	$0{,}4 - 4{,}5$	$4{,}6 - \infty$
T3	$-\infty - 0{,}91$	$0{,}92 - 2{,}76$	$2{,}77 - \infty$
TT4	$-\infty - 53$	$54 - 115$	$116 - \infty$
T4U	$-\infty - 0{,}70$	$0{,}71 - 1{,}85$	$1{,}86 - \infty$
FTI	$-\infty - 44$	$45 - 117$	$118 - \infty$

Con respecto a la edad, esta se procede a dividir en 4 etapas, las cuales se presentan en la Tabla 3.2.

Tabla 3.2: Ragos de edad

Rangos	
Infacia	0 a 11 años
Juventud	12 a 26 años
Adultez	27 a 59 años
Vejez	60 a 110 años

Cabe destacar que todas las variables fueron transformadas a factores, sin realizar ningún procedimiento de normalización.

3.1.2 Unión de clases

La base de datos $allhyper$ cuenta con 2215 observaciones a analizar luego de haber realizado el pre-procesamiento en ella. Estos datos en su mayoría corresponden a la clase negativa y tan solo el 3,07 % pertenece a clases con enfermedades. Es por este motivo que se decidió unir todas las clases no negativas (hipertiroidismo, bocio y T3 toxic) nombrándolas como positivas.

3.2 REGLAS INTERESANTES

Como se mencionó anteriormente, estas reglas son aquellas que cumplen con un soporte mínimo y una confianza mínima previamente establecida, considerando que estas últimas varían según reglas y relaciones que se buscan obtener.

En primer lugar se verán reglas para las clases (positivas y negativas) de la base de datos, para luego buscar reglas relacionadas a los niveles de hormonales, tal como TSH, T3, TT4, entre otros. Finalmente se mostrarán reglas relacionadas con la variable edad.

3.2.1 Clase positiva

Para el caso de la clase positiva su soporte es muy bajo, ya que solo representa al 3,07 % del total de los datos, es por este motivo que el soporte mínimo toma el valor de 0.01, mientras que la confianza mínima es de sólo 0.2, por exactamente el mismo motivo.

Con la función $apriori$ de R se obtuvieron una serie de reglas de asociación, las cuales fueron ordenadas y filtradas por aquellas que contaban con un mayor lift, soporte o confianza, dejando sólo aquella que cuenta con el máximo valor de estas medidas, pero dado que la regla de mayor lift y de mayor confianza resultaron ser las mismas, se decidió dejar sólo una de estas, quedando en la Tabla 3.3 las reglas con máximo lift y soporte.

Tabla 3.3: Reglas para la clase positiva

Antecedente	Consecuente	Soporte	Confianza	Lift
sex = F, on thyroxine = f, on antithyroid medication = f, pregnant = f, psych = f, TSH = Low, T3 = High	class = positive	0.013	0.492	16.011
sex = F, on thyroxine = f, on antithyroid medication = f, pregnant = f, psych = f, TSH = Low	class = positive	0.023	0.202	6.592

3.2.2 Clase negativa

Las reglas en este caso no necesitan un soporte bajo, ya que esta clase se encuentra en la mayoría de las observaciones. El soporte mínimo es de 0.8 y la confianza mínima de 0.8. La Tabla 3.4 muestra las primeras 5 reglas con mayor lift y las primeras 5 reglas con menor lift respectivamente.

Tabla 3.4: Reglas para la clase negativa

Orden	Antecedente	Consecuente	Soporte	Confianza	Lift
Mayor lift	pregnant = f, T3 = Normal	class = negative	0.801	0.986	1.017
	on antithyroid medication = f, I131 treatment = f, query hyperthyroid = f, tumor = f	class = negative	0.876	0.982	1.013
	on antithyroid medication = f, I131 treatment = f, query hyperthyroid = f, lithium = f, tumor = f	class = negative	0.872	0.982	1.013
	on antithyroid medication = f, I131 treatment = f, query hyperthyroid = f, tumor = f, T4U = Normal	class = negative	0.844	0.982	1.013
	on antithyroid medication = f, I131 treatment = f, query hyperthyroid = f, goitre = f, tumor = f	class = negative	0.868	0.982	1.013
Menor lift	on thyroxine = f, thyroid surgery = f, goitre = f, psych = f	class = negative	0.801	0.964	0.995
	on thyroxine = f, query on thyroxine = f, goitre = f, psych = f	class = negative	0.802	0.964	0.995
	on thyroxine = f, thyroid surgery = f, lithium = f, psych = f	class = negative	0.805	0.964	0.995
	on thyroxine = f, query on thyroxine = f, lithium = f, psych = f	class = negative	0.805	0.964	0.995
	on thyroxine = f, thyroid surgery = f, psych = f	class = negative	0.808	0.964	0.995

3.2.3 TSH

Se decide buscar las reglas relacionadas con la variable TSH, ya que esta es una hormona producida en el cerebro, más específicamente en la glándula pituitaria, la cual tiene la función de estimular la tiroides para su producción de hormonas tiroidales, por lo que puede proporcionar información relevante para el análisis.

El soporte mínimo con el que se trabajará es de 0.001 y la confianza mínima de 0.8. A continuación se presentan las reglas relacionadas a TSH bajo y TSH alto.

3.2.3.1 TSH bajo

La Tabla 3.5 presenta las reglas ordenadas según mayor lift y mayor soporte respectivamente, cabe destacar que la mayor confianza ya se encuentra en las reglas del mayor lift con un valor de 1.

Tabla 3.5: Reglas para TSH bajo

Orden	Antecedente	Consecuente	Soporte	Confianza	Lift
Mayor lift	on antithyroid medication = t, pregnant = t	TSH = Low	0.014	1.000	4.293
	I131 treatment = t,T3 = High	TSH = Low	0.014	1.000	4.293
	query on thyroxine = t, sick = t, FTI = High	TSH = Low	0.014	1.000	4.293
Mayor soporte	sex=F, thyroid surgery = f, query hypothyroid = f, query hyperthyroid = f, tumor = f, psych = f, T3 = High, TT4 = High, FTI = High	TSH = Low	0.015	0.811	3.481

3.2.3.2 TSH alto

Al igual que en TSH bajo, aquí se presentan las reglas con el mayor lift y mayor soporte respectivamente, dado a que la mayor confianza ya se encuentra incluida en la regla de mayor lift.

Tabla 3.6: Reglas para TSH alto

Antecedente	Consecuente	Soporte	Confianza	Lift
age = Adulthood, sick = t, T3 = Low	TSH = High	0.001	1.000	14.477
age = Adulthood, query hypothyroid = t, T3 = Low	TSH = High	0.002	0.833	12.064

3.2.4 Hormonas en la tiroides

Las variables de $allhyper$ que tienen relación con las hormonas producidas en la tiroides son T4U, T3, TT4 y FTI, es por este motivo que se proceden a obtener las reglas más

interesantes de dichas variables, de esta forma en el consecuente se tendrán a algunas de esas variables con los valores de alto y bajo.

El soporte mínimo establecido es de 0.01 y la confianza mínima de 0.8. En la Tabla 3.7 se puede apreciar las reglas ordenadas por mayor lift, mayor soporte y mayor confianza respectivamente.

Tabla 3.7: Reglas para hormonas de la tiroides

Antecedente	Consecuente	Soporte	Confianza	Lift
age = Adulthood, sex = F, pregnant = t	T3 = High	0.011	0.828	7.245
on thyroxine = t, sick = f, thyroid surgery = f, query hyperthyroid = f, TSH = Low	FTI = High	0.029	0.802	2.830
age = Adulthood, sex = F, pregnant = t, query hypothyroid = f	TT4 = High	0.011	0.893	2.720

3.2.5 Edad

A continuación se presentan las reglas más interesantes que en su consecuente tenga la variable edad. Estas son calculadas con un soporte mínimo de 0.01 y confianza mínima de 0.8, además, en la Tabla 3.8 se muestra las reglas ordenadas por las dos primeras con mayor lift, y las dos primeras con mayor soporte respectivamente.

Tabla 3.8: Reglas para la edad

Orden	Antecedente	Consecuente	Soporte	Confianza	Lift
Mayor lift	sex = M, TSH = Normal, T3 = Low, TT4 = Normal	age = Eld	0.011	0.893	2.180
	query on thyroxine = f, pregnant = t, tumor = f	age = Adulthood	0.012	0.929	1.935
Mayor soporte	query hypothyroid = f, goitre = f, T3 = Low, TT4 = Normal	age = Eld	0.034	0.800	1.954
	sick = f, pregnant = t	age = Adulthood	0.013	0.806	1.679

CAPÍTULO 4. ANÁLISIS DE RESULTADOS Y COMPARACIÓN

En este capítulo se realizará un análisis de las reglas obtenidas y mostradas en el Capítulo 3, considerando las respectivas medidas de soporte y confianza mínima que fueron utilizadas para su obtención, así como el lift para verificar su calidad.

Por otro lado, se realizará una comparación entre los resultados obtenidos en esta experiencia y en la anterior, la cual consistía en hacer uso de un algoritmo de Clustering para obtener y analizar sus diferentes grupos.

Antes de comenzar el análisis como tal, es interesante notar los diferentes valores mínimos para el soporte y la confianza, que si bien podrían parecer arbitrarios, estos tienen directa relación con la gran cantidad de cierto tipo de datos dentro de los registros, que como se mencionó anteriormente, se tiene de ejemplo la clase negativa, pero también, es evidente como en variables continuas los niveles normales predominan de sobremanera en contraste de los bajos o altos. Esto lo que produce es que al intentar buscar reglas con altos niveles de soporte y/o confianza, se obtengan siempre de los mismos niveles, que en el caso de esta base de datos, tienen una gran tendencia a ser datos que se encuentran dentro de los rangos hormonales normales y de clase negativa, quitando la posibilidad de encontrar otras reglas que si bien podrían tener menos representatividad dentro de estos registros, podrían indicar información relevante para el estudio. Considerando lo anterior, en algunos casos se llegó a bajar el soporte mínimo a un 1 %, lo que podría parecer un % irrelevante, pero finalmente, termina siendo una proporción aceptable dadas las condiciones que indicaron anteriormente.

4.1 ANÁLISIS DE REGLAS INTERESANTES

4.1.1 Clase positiva

Dada la baja cantidad de registros con una clase diferente a negativo, se decidió unir Hipertiroidismo, Goitre y T3 toxic en una sola clase, para así lograr soportes más altos para las

reglas que involucrasen a estas clases. De esta manera se creó la clase positiva, la que indica la existencia de una enfermedad relacionada a hormonas tiroideas (sin considerar hipotiroidismo).

A pesar de este arreglo de las clases, si se aprecia la Tabla 3.3, el soporte para ambas reglas seleccionadas es de 0.013 y 0.023 respectivamente, es decir, una representación no mayor a un 1.3 % y 2.3 % para estos itemsets dentro de la totalidad de registros. Ahora, si se aprecia la confianza en la primera regla, esta es de 0.492, un número que deja prácticamente al azar de una moneda la ocurrencia de esta regla, sin embargo, al mirar el lift se puede ver una fuerte relación entre el antecedente y el consecuente, lo que permite definir esta regla como interesante. Básicamente lo mismo ocurre para la segunda regla, donde se aprecia una confianza de 0.202, pero su lift es de 6.592.

Ya teniendo claro que estas reglas son de interés, se procede a realizar el análisis de su relación.

La primera regla indica a una persona de sexo femenino, que no se encuentra en tratamiento de tiroxina ni medicación antitiroidea, que tampoco está embarazada ni tiene un tratamiento psicológico, pero lo más relevante de esta regla se encuentra en las últimas dos variables, que indican un nivel de TSH bajo y T3 alto. La importancia de estos dos últimos items radica en que un bajo nivel de hormona TSH se encuentra directamente relacionado a un T3 alto, ya que es el TSH el encargado de regular la producción de T3 y otras hormonas en la tiroides, pero no sólo eso, sino que se tiene por otro lado que un alto nivel de T3 se se considera un factor crucial para determinar si una persona se encuentra afectada por alguna enfermedad tiroidea. Si bien como se mencionó antes, su suporte es de prácticamente un 1 %, el lift permite determinar que estas variables sí afectan o influyen de manera directa a que un paciente sea diagnosticado como clase positiva.

La segunda regla sólo difiere en que no incluye T3 alto, pero su soporte es ligeramente mayor, es por esto que también se consideró como regla interesante, reforzando aún más la relación entre el TSH bajo y la positividad de alguna enfermedad relacionada a la tiroides.

De esta manera, aunque se tenga un soporte pequeño y una confianza no muy alta,

estas reglas permiten dar algunos indicios de cómo se podría detectar a un paciente enfermo.

4.1.2 Clase negativa

Dado a que la mayoría de las observaciones pertenecen a la clase negativa, es por este motivo que el soporte a utilizar puede ser estricto, otorgándole así un valor de 0.8, además, la confianza por defecto de la función $apriori$ es de 0.8, la cual también es un valor alto.

Observando las reglas de la Tabla 3.4, las primeras 5 corresponden a las de mayor lift, coincidiendo con las de mayor confianza (con un valor muy cercano a 1). De estas reglas con un lift superior, la primera podría ser la de mayor interés, y no sólo por su gran soporte o confianza, si no, porque se indica que si el paciente no se encuentra en período gestacional y tiene un T3 normal, entonces pertenece a la clase negativa, es decir, que no tiene alguna enfermedad tiroidal. Lo del embarazo tiene sentido, ya que en experiencias anteriores se indicó que esta variable afecta directamente los niveles hormonales tales como TSH,T3,TT4 u otros, además, la base de datos cuenta con una gran mayoría de muestras de pacientes con sexo femenino. Por otro lado, el tener un T3 normal indica generalmente un correcto funcionamiento de la tiroides. Ahora, el problema de esta y todas las otras reglas relacionadas a la clase negativa, es que su lift es prácticamente 1, lo que indica que el antecedente es independiente del consecuente, pero no sólo eso, ya que como indica Anisha Garg (2018), no es relevante lo que se tenga en el antecedente para un consecuente que muy frecuente, ya que la confianza para una regla de asociación que tiene un muy frecuente consecuente, siempre será alto. Dado que la clase negativa representa a más del 95 % de los registros, es claro que el consecuente es muy frecuente, por lo que no se podría considerar esta como una regla realmente confiable. Si miramos las otras 4 reglas restantes, se puede apreciar que son bastante similares entre ellas, sólo destacando la 4ta, ya que contiene un T4U normal.

Por otro lado, las reglas con un menor lift no difieren demasiado con las anteriores, por lo que tampoco se podrían considerar reglas realmente confiables.

Considerando lo ya mencionado, se puede decir que no fue posible encontrar reglas

de interés teniendo un consecuente con clase negativa, no porque estas no alcanzaran un soporte o confianza mínimos suficientes, sino, por la gran predominancia de esta clase dentro de la base de datos, lo que entrega un lift muy cercano a 1, indicando así, la independencia del antecedente y consecuente.

4.1.3 Hormona TSH

Como se menciona en el capítulo de obtención de reglas, el TSH es una hormona crucial para el corrector funcionamiento de la tiroides, por lo que encontrar reglas que la involucren siempre será de interés. Ahora, considerando que los casos con TSH normal sobrepasan de sobremanera a los de TSH bajo y alto, no tiene caso buscar reglas para ese nivel, ya que tendrán un alto soporte y confianza, pero un lift muy cercano a 1.

4.1.3.1 TSH bajo

En experiencias anteriores se indicó que el tener un bajo nivel de TSH es señal de que en la tiroides se está secretando en exceso hormonas tiroidales (T3 y T4), lo que podría conllevar directamente a alguna enfermedad como hipertiroidismo.

En la Tabla 3.5 se pueden apreciar las reglas que tienen como consecuente un TSH bajo, siendo 3 las que tienen el mayor lift y 1 la que tiene el mayor soporte (aunque no por mucha diferencia). Dentro de las reglas con mayor lift, se puede ver que en la primera de ellas se indica que la persona se encuentra en consumo de medicamentos antitiroideos y que a la vez se encuentra embarazada, teniendo como consecuente un TSH bajo. Esto tiene sentido, ya que durante el embarazo son comunes los desórdenes hormonales, por lo que estar en un tratamiento para controlar estos cambios en los niveles tiroideos afectan directamente la secreción de TSH. En la segunda regla se aprecia un tratamiento de yodo radiactivo y un alto nivel de T3 que son responsable de niveles de TSH bajos. El indicio de estar en un tratamiento de I131 se condice

directamente con un alto T3, dado que estos tratamientos buscan disminuir los niveles de T3 y T4 en la sangre. Por otro lado, en la tercera regla se tiene a una persona que ya había realizado consultas sobre sus niveles tiroideos, que además se encuentra enferma y tiene un alto FTI, el cual es un índice que indica el nivel de tiroxina libre en la sangre, habiendo una relación directa entre un alto nivel en FTI y un bajo nivel de TSH. Si bien el soporte de estas reglas es muy pequeño, siendo de no más de 0.014, su confianza es de 1, es decir, que si bien no es frecuente que que ocurra esta combinación de items, cuando ocurren, es 100 % probable que estarán juntos, siendo estas reglas apoyadas también por un lift bastante alto, lo que permite determinar que sí existe una correlación entre antecedente y consecuente. Por lo tanto, se consideran reglas de interés.

Dentro de la misma Tabla 3.5 se encuentra la regla con mayor soporte para un consecuente con TSH bajo, esta indica que es un paciente de sexo femenino sin mayores problemas, ya que no tuvo cirugía de tiroides, tampoco consultas previas por hipertiroidismo o hipotiroidismo, no cuenta con un tumor o problemas psicológicos, sin embargo, sus niveles de T3, TT4 y FTI son altos, los que son una clara señal de TSH bajo. El soporte de esta regla sigue siendo bajo (0.015), con una confianza de 0.811, pero con un lift que apoya el hecho de que existe una relación en la regla. Esta regla también es considerada de interés.

4.1.3.2 TSH alto

A diferencia de un nivel de TSH bajo, cuando este se encuentra con valores mayores a los normales, la persona es propensa a contraer hipotiroidismo.

En la Tabla 3.6 se tienen 2 reglas, siendo la primera la de una persona adulta que presenta algún tipo de enfermedad, pero que además posee bajos niveles de T3. Nuevamente esto último tiene sentido, ya que existe una relación inversa entre TSH y los niveles tiroidales, en la que al tener un alto TSH, se tiene un bajo T3 y T4 o viceversa. La segunda regla de esta tabla no es muy diferente, ya que es un adulto con niveles de T3 bajos, pero que pueden entenderse debido a que la persona ya había hecho consultas por hipotiroidismo, lo que conlleva también

a que sus niveles de TSH figuren como bajos. El soporte de estas reglas no supera el valor de 0.002, pero mantienen un alta confianza apoyado directamente por un lift. Ambas reglas son de interés.

4.1.4 Hormonas en la tiroides

En esta subsección hará el análisis de las reglas para hormonas producidas en la tiroides como el T4 o T3 y sus índices o variaciones como el FTI o TT4.

En la Tabla 3.7 se aprecian 3 reglas diferentes, donde en la primera se indica una persona adulta, de sexo femenino y además se encuentra embarazada, siendo el consecuente de esta regla un T3 alto. Como ya se ha mencionado anteriormente, el embarazo es una variable a considerar cuando se habla de problemas tiroideos, así que no es extraño ver un T3 alto para una persona en período gestacional. La segunda regla por otro lado, indica el caso de usa persona que se encuentra consumiendo tiroxina, no está enferma, no ha tenido una cirugía de tiroides y tampoco había realizado consultas con respecto a un posible hipertiroidismo, sin embargo, se puede ver que cuenta con un bajo TSH. Tal como indica la American Thyroid Association (S.F), un bajo nivel de TSH en la sangre y un alto nivel de FTI son claros indicios de una persona que podría llegar a tener hipertiroidismos, aunque esto altos niveles de FTI estén dados por el consumo de tiroxina. Finalmente la tercera regla muestra a nuevamente a una persona adulta, de sexo femenino y embarazada, agregando que no tenía consultas previas por hipotiroidismo, aunque ahora el consecuente es que su TT4 es alto. Dado que el TT4 es la hormona T4 total, está directamente ,relacionado a la secreción hormonal producida en la tiroides, por ende, con el embarazo.

De estas 3 reglas, la primera con un lift de 7,245 es la de mayor valor, indicando una clara relación entre antecedente y consecuente, al igual que la segunda y tercer regla, aunque en un menor grado. Aún así, como ya se ha mencionado, el soporte de estas reglas es generalmente muy pequeño debido a la baja representatividad que se tiene de personas con parámetros altos o bajos. Por otro lado, la confianza que mantiene estas reglas se puede considera como alta, ya

que es por sobre el 80 % para las 3. Teniendo en cuenta lo anterior, es posible indicar que estas reglas son de interés y pueden ser de utilidad.

4.1.5 Edad

Generalmente la edad es un factor importante cuando se habla sobre salud, por lo que en este caso sería interesante analizar si existen reglas que relacionen una serie de atributos con la edad de una persona. En esta base de datos la gran mayoría de los registros pertenecen a personas adultas o ancianas, quedando prácticamente sin representación la población infante y joven, incluso con un soporte de 0.01.

En la Tabla 3.8 se pueden apreciar una serie de reglas, donde la 1 y 2 representan las de mayor lift. La primera de ellas muestra a un hombre con un TSH y TT4 normal, pero con un bajo T3, indicando así que se trata de una persona anciana. Mientras que en la segunda lo más interesante podría ser ver la relación entre un embarazo y estar en la adultez. Luego en la tabla se encuentran las reglas con mayor soporte, donde la primera muestra a una persona que no ha hecho consultas previas por hipotiroidismo, no tiene goitre, con un T3 bajo y TT4 normal, indicando así que pertenece a la edad anciana. Si comparamos esta regla con la primera de la tabla, se ve que hay una clara relación entre un T3 bajo y ser anciano. Para la última regla de la tabla se indica que no se está enfermo y que está embarazada, teniendo como consecuente que es una persona adulta. Nuevamente en las reglas con edad adulta en el consecuente se ve la aparición del embarazo, porque claro, es en el rango de 26 a 59 donde más embarazos se presentan dentro de la base de datos.

El soporte en general de estas reglas supera el 10 %, siendo la más representativa la tercera regla con un 34 %, además la confianza de todas ellas es alta, llegando a 93 % aproximadamente, mientras que el menor lift es de 1.679 y el mayor de 2.180, lo que muestra la clara relación positiva entre el lado izquierdo y derecho de las reglas. Teniendo estas medidas de calidad, es claro que estas reglas se pueden considerar de interés, más aún aquellas que contienen items relacionados a hormonas como TT4, T3 o TSH.

4.2 COMPARACIÓN DE EXPERIENCIAS

En la experiencia anterior también se trabajaron los datos como factores, y se buscó determinar las relaciones que tenían los sujetos entre sí. A diferencia de este laboratorio, el cual busca identificar las relaciones existentes entre las variables y no entre los sujetos. Aún así, se ven similitudes entre las reglas obtenidas en este laboratorio y los grupos encontrados en el anterior.

En la Tabla 4.1 se pueden apreciar los medioides de los grupos formados en el laboratorio anterior, con las que se realizarán las comparaciones con las reglas obtenidas.

Tabla 4.1: Datos de medioides laboratorio anterior

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

Luego se tiene un resumen de esta tabla y sus grupos.

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

En la Tabla 3.3 se pueden apreciar 2 reglas que indican a una persona de sexo femenino y con TSH bajo, perteneciente a la clase positiva. Ahora, si se mira la Tabla 4.1 el medioide 4 representa a un grupo enfermo, y se puede apreciar que su TSH también es bajo, confirmando así, la relación que existe entre esta variable y la clase positiva.

Con respecto a la clase negativa, en la Tabla 3.4 se encuentran reglas con un lift muy cercano 1, mostrando que no hay relación entre el antecedente y el consecuente, ahora, si

se observa el medioide 1 y 2 de la Tabla 4.1, se da también que en algunos casos las variables tampoco tienen relación con su clase, ya que en el medioide 1 el sujeto no se encuentra con un consumo de tiroxina, mientras que en el medioide 2 sí lo está, y en ambos casos su clase es negativa. Lo mismo sucede con el FTI entre estos medioides, ya que el primer caso se encuentra con niveles normales y el segundo con niveles altos.

Por otro lado, en la Tabla 3.5 se tienen las reglas para un TSH bajo, que si bien se puede ver que en el Medioide 4 se cumple esta condición, los otros elementos que hacen cumplir esta regla no se aprecian en este medioide. Considerando incluso que altos niveles de T3, TT4 y un alto índice FTI son claras señales de hipertiroidismo. Aún así, esto se puede explicar porque el grupo representa a una gran cantidad de sujetos, mientras que si se miran las reglas para TSH bajo, sus soportes no superan el 1.5 %.

En el caso de de las reglas para TSH alto, es difícil realizar una comparación con los grupos de la experiencia anterior, ya que altos niveles de TSH se relacionan directamente con hipotiroidismo, enfermedad que si bien tiene relación con las hormonas tiroidales, no es el objetivo de estudio para esta base de datos.

Mirando la Tabla 3.7 se pueden encontrar reglas que tienen un soporte mayor al 10 %, y aún así, estas reglas no se ven representadas por los medioides de los grupos, ya que como se ha mencionado en varias ocasiones, dentro de esta base de datos los niveles normales son los que predominan por lejos, así que encontrar alguna de estas hormonas con niveles fuera de lugar es complejo. Sin embargo, la regla que tiene como consecuente un alto nivel de FTI está directamente relacionado con el medioide del grupo 2, ya que ambas cuenta con el hecho de estar en un tratamiento de consumo de tiroxina.

Para las reglas relacionadas a la edad, no existe un grupo que represente a la población anciana, así como tampoco existen personas adultas embarazadas dentro de estos grupos, por lo que las reglas obtenidas no se pueden aplicar allí.

CAPÍTULO 5. CONCLUSIONES

Esta experiencia permitió observar cuan relacionadas están las variables entre sí, logrando obtener información relevante de ellas, como por ejemplo los niveles de hormonas que se producen tanto en la tiroides como en el cerebro, éstas están fuertemente relacionadas ya que en general se tiene un lift alto, determinando así la clase a la que debe pertenecer el sujeto, o el nivel que otra hormona debe tener. Cuando ocurren esas condiciones, en su mayoría suceden pocas veces, pero no porque no sea interesante, sino porque la base de datos $allhyper$ cuenta con pocas observaciones con variables distintas a negativa, normal, juventud, infancia, entre otros. Pero cabe destacar que cuando ocurren dichas reglas, éstas se cumplen la mayoría de las veces, esto porque en general la confianza obtenida de ellas es alta.

Lo recién mencionado confirma que se cumplió el objetivo principal de la experiencia, ya que surgieron varias reglas interesantes que permiten enriquecer el conocimiento del problema que se estudia. Además, con la información obtenida del laboratorio se puede realizar una comparación con la entrega anterior.

Las principales similitudes que se observan entre ambos laboratorios es que en los dos se busca encontrar relaciones entre los datos, ya sea mediante las observaciones o a través de las variables que los componen. Igualmente esto permite observar si los representantes de los grupos obtenidos en la experiencia anterior cumplen con las reglas que se consiguieron en este. Estas comparaciones permiten enriquecer aún más el análisis de los datos.

Por otro lado, se puede decir que el método utilizado en esta experiencia resulta más exacto que en el laboratorio anterior, ya que los datos que se obtienen en este tienen mayor relación y entregan mayor información que los grupos vistos anteriormente.

Para seguir con un buen análisis es necesario conocer bien las variables de la base de datos, ya que de esta forma se pueden entender mejor las relaciones que se tiene entre ellas, y la única forma de conocer dichas variables es mediante la investigación.

BIBLIOGRAFÍA

ASSOCIATION, A. T. (S.F.). Thyroid function tests. [Online] https://www.thyroid.org/thyroid-function-tests/#:~:text=The%20finding%20of%20an%20elevated,in%20individuals%20who%20have%20hyperthyroidism.

Chacón, M. and Arredondo, J. (2020). Laboratorio 3 - reglas de asociación. [Online] https://uvirtual.usach.cl/moodle/pluginfile.php/264844/mod_resource/content/1/An_lisis_de_Datos___Laboratorio_3.pdf.

Garg, A. (2018). Complete guide to association rules. [Online] https://towardsdatascience.com/association-rules-2-aa9a77241654.

Martínez, C. G. (S.F). Reglas de asociaciÓn. [Online] https://rpubs.com/Cristina_Gil/Reglas_Asociacion.

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