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

Laboratorio 3 - Reglas de Asociación

Integrantes:	Ariel Aaron Argomedo Madrid
	Marcelo Israel Guzmán Gutiérrez
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco
Ayudante:	Gustavo Hurtado

29 de Octubre de 2022

Tabla de contenidos

1. Introducción **1**

1.1. Objetivos 1

2. Marco teórico **2**

2.1. Reglas de asociación 2

2.2. Algoritmos más utilizados para la obtención de reglas de asociación 3

2.3. Medidas de calidad y confianza 3

2.3.1. Medidas monótonas en confianza 4

2.3.2. Medidas monótonas en soporte y confianza 4

2.4. Monotonicidad y propiedades de las medidas 5

2.4.1. Medida monótona 5

2.4.2. Medida anti-monótona y principio *Apriori* 5

3. Obtención de reglas **6**

3.1. Pre-procesamiento 6

3.2. Reglas interesantes 7

3.2.1. Clase muerto 8

3.2.2. Clase vivo 8

3.2.3. Sgot 9

3.2.4. Albumin 10

4. Análisis de resultados y comparación **11**

4.1. Análisis de reglas con consecuente “Clase = Muerto” 11

4.2. Análisis de reglas con consecuente “Clase = Vivo” 13

4.3. Análisis de otras reglas interesantes 14

4.4. Comparación experiencias anteriores 15

4.4.1. Grupo 1: Vivos 16

4.4.2. Grupo 2: Muertos 16

5. Conclusiones **18**

Bibliografía **20**

1. Introducción

Cada uno de los elementos que componen el cuerpo humano forman parte de un todo que, al trabajar uno en conjunto con el otro, permiten que el cuerpo funcione adecuadamente. Trabajando juntos estos elementos o sistemas mantienen la estabilidad interna, concepto conocido como homeostasis (RevereHealth, 2016).

En este laboratorio se estudian las asociaciones existentes entre los elementos de la base de datos de Hepatitis, que contiene los niveles de *Bilirrubina*, *Alk phosphate*, *Sgot*, etc. Los algoritmos de reglas de asociación que serán utilizados en el estudio, tienen como objetivo encontrar relaciones dentro de un conjunto de transacciones (ítems o atributos), que ocurren de forma conjunta (Rodrigo, 2018).

Se espera que los resultados de las reglas de asociación obtenidas en el presente estudio se correspondan con los resultados de los anteriores laboratorios, es decir, se espera que los pacientes con hepatitis con mayor riesgo de muerte sean aquellos con los niveles más altos de *Bilirubin*, *Alk phosphate*, *Sgot*.

A lo largo de la experiencia, se explicarán los conceptos más importantes a tener en cuenta en el marco teórico, influyentes en la parte técnica o algorítmica de las reglas de asociaciones. Luego, se explica el proceso de obtención de las reglas de asociación en forma práctica (a partir del código en R), donde se señalan aquellas reglas más relevantes. Más adelante, se realiza un análisis de resultados y comparación con las experiencias de laboratorio anteriores. Por último, se concluye el trabajo a través de una breve retroalimentación y problemas asociados al desarrollo de la experiencia.

1.1. Objetivos

1. Utilizar el paquete `arulesViz`, propio del lenguaje R, para visualizar las reglas de asociación obtenidas, en forma práctica.
2. Extraer conocimiento del problema asignado, por medio de las reglas de asociación.
3. Entender el problema y contrastar los resultados con laboratorios anteriores.
4. Entender la importancia de la aplicación de las reglas de asociación en minería de datos.

2. Marco teórico

2.1. Reglas de asociación

Una regla de asociacion es definida por Rodrigo (2018) como una implicación del tipo "si X entonces Y" ($X \rightarrow Y$), donde X e Y son *itemsets* o *items* individuales. A la izquierda de la regla se encuentra el antecedente o *left-hand-side* (LHS) y al lado derecho esta el consecuente o *right-hand-side* (RHS).

Para entender los diferentes algoritmos de reglas de asociacion, es necesario definir los siguientes conceptos:

- Soporte: El soporte del *item* o *itemset* X es el número de transacciones que contienen X dividido entre el total de transacciones (n).

$$soporte(X) = \frac{freq(X)}{n} \tag{1}$$

  Especialmente, se evalúa el soporte de una regla:

$$soporte(X \Rightarrow Y) = \frac{freq(X \cup Y)}{n} \tag{2}$$

  Tomando *freq* como la frecuencia de X en la lista de *itemsets*.

- Confianza: Se define a la confianza de una regla "si X entonces Y" como:

$$confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)} \tag{3}$$

  Se interpreta a la confianza como la probabilidad $P(Y|X)$, es decir, la probabilidad de que una transacción que contiene los *items* de X, también contenga los *items* de Y (Rodrigo, 2018).

2.2. Algoritmos más utilizados para la obtención de reglas de asociación

Los algoritmos más utilizados para encontrar *itemsets* frecuentes o reglas de asociación son:

- Apriori: Este algoritmo está compuesto de dos etapas. En primer lugar, se deben identificar aquellos *itemsets* más frecuentes, dentro de un límite establecido. A continuación, se convierten estos *itemsets* frecuentes en reglas de asociación (Rodrigo, 2018).

  El proceso explicado a grandes rasgos por (Rodrigo, 2018), se presenta a continuación:

  1. Por cada *itemset* frecuente I, obtener todos los posibles subsets de I. Luego, para cada subset s de I, crear la regla "s $\Rightarrow (I - s)$".
  2. Descartar todas las reglas que no superen un mínimo de confianza.

- FP-Growth: El algoritmo emplea una estructura de árbol que contiene toda la información de las transacciones. Una vez la base de datos se comprimió en una estructura *FP-Tree*, se divide en varias bases de datos condicionales, cada una asociada con un patrón frecuente. Se analiza cada partición de forma separada y se concatenan los resultados obtenidos (Rodrigo, 2018).

- Eclat: Este algoritmo analiza las transacciones en formato vertical, donde cada línea contiene un ítem y las transacciones en las que aparece dicho ítem. Sin embargo, no genera reglas de asociación, tan solo permite identificar los *itemsets* más frecuentes (Rodrigo, 2018).

2.3. Medidas de calidad y confianza

Para evaluar de mejor manera la confianza, se han generado diferentes medidas o métricas de calidad para seleccionar el conjunto de las mejores reglas. Generalmente, se les exige a estas medidas que sean monótonas en confianza, o soporte y confianza, pero manteniendo uno de ellas constante.

2.3.1. Medidas monótonas en confianza

- *Lift*: Representa una medida de independencia entre dos *items* A y B. *lift* toma su valor más bajo (1) cuando A y B son completamente independientes (Martinez, 2020).

$$lift(A \Rightarrow B) = \frac{P(B \cap A)}{(P(A)P(B))}$$

  La interpretación del valor del *lift* es la siguiente:

  - $lift = 1$ indica que A y B son independientes, es decir, que la regla no representa un patrón real.
  - $lift < 1$ indica que A y B están correlacionadas negativamente.
  - $lift > 1$ indica que A y B están correlacionadas positivamente.

- *Convicción*: Al igual que la medida anterior, representa la independencia entre A y B y es monótona en confianza.

$$convicción = \frac{P(A)P(\bar{B})}{P(A \cap B)}$$

2.3.2. Medidas monótonas en soporte y confianza

- *Laplace*: k es un numero entero mayor que 1:

$$Laplace = \frac{Sop(A \Rightarrow B) + 1}{Sop(A \Rightarrow B)/c + k}$$

  Manteniendo la confianza constante, la medida es monótona en soporte. Manteniendo el soporte constante, la medida es monótona en confianza.

- *Ganancia*:

$$Ganancia(A \Rightarrow B) = Sop(A \Rightarrow B)(1 - \theta/c)$$

- *Metrica de Piatetsky-Shapiro (P-S)*:

$$P - S(A \Rightarrow B) = Sop(A \Rightarrow B) - \frac{Sop(A)Sop(B)}{n}$$

2.4. Monotonicidad y propiedades de las medidas

Los algoritmos para la generación de reglas de asociación buscan evitar lo más posible la generación del conjunto de todas las combinaciones de *items* que serán parte de las reglas, o bien, minimizar las comparaciones entre las medidas de calidad. Para ello, se necesita el concepto de medidas monótonas o anti-monótonas, que ayudan a la *poda* de soluciones para los algoritmos nombrados anteriormente.

2.4.1. Medida monótona

Tomando I como un conjunto de *items*, $J = 2^I$ su conjunto potencia, se define f como medida monótona si:

$$\forall X, Y \in J : (X \subseteq Y) \longrightarrow f(X) \leq f(Y)\,, \tag{4}$$

es decir, cualquier subconjunto que se tome de Y, tendrá menor o igual medida que el conjunto completo Y (Tan et al., 2006, p. 334).

2.4.2. Medida anti-monótona y principio *Apriori*

Análogamente, f es antimonótona si:

$$\forall X, Y \in J : (X \subseteq Y) \longrightarrow f(Y) \leq f(X)\,, \tag{5}$$

lo que quiere decir que, desde un conjunto Y de *ítems*, si quitamos elementos, la medida de f sobre el nuevo subconjunto, sea este X, será mayor que la medida de Y. El soporte, entonces, se define como una medida de naturaleza **antimonótona**. Esta definición lleva a la creación del teorema de principio *Apriori*: si un *itemset* es frecuente (sobre o igual al umbral del soporte), todos los subconjuntos deben ser frecuentes. Esto también ocurre al revés, considerando la infrecuencia (menor al umbral o soporte mínimo elegido): de un conjunto infrecuente, todas las combinaciones de subconjuntos son infrecuentes (Tan et al., 2006, pp. 333-335), ayudando a la *poda* del conjunto potencia en los algoritmos, optimizándolos.

3. Obtención de reglas

En el presente capítulo, se presentan las reglas obtenidas de la base de datos *Hepatitis* con la que se ha trabajado a lo largo del curso, a través del proceso de minería de reglas de asociación con el algoritmo *A priori*. Para la obtención de las reglas, se tomaron en cuenta tanto valores mínimos de soporte como confianza, además del *lift*.

Para cada uno de los casos presentados a continuación, se eliminaron aquellas reglas consideradas redundantes y que no aportan mayor información al problema.

3.1. Pre-procesamiento

Antes de realizar el proceso de obtención de reglas de asociación se requirió de un preprocesamiento inicial para tratar la información perdida y convertir las variables a los formatos adecuados, proceso que en el laboratorio anterior fue explicado con mayor detalle.

La mayoría de los métodos de reglas de asociación utilizan un enfoque de manipulación de conjuntos de elementos, en el que el tipo de dato debe ser de naturaleza categórica. Cuando el conjunto de datos tiene atributos numéricos, es necesario discretizarlos antes de la minería de reglas, de forma que los *itemsets* estarán conformados por los atributos binarios del dataset y los atributos discretizados (Tan, 2018).

	Baja	Normal	Alta
Bilirubin	- ∞ - 0.1	0.1 - 1.2	1.2 - ∞
Alk Phosphate	- ∞ - 30	30 - 120	120 - ∞
Sgot	- ∞ - 8	8 - 45	45 - ∞
Albumin	- ∞ - 3.4	3.4 - 5.4	5.4 - ∞

Cuadro 1: Discretización de las variables numéricas

Las Tablas 1 y 2 muestran como se llevó a cabo la discretización, de las medidas de metabolitos y edad, respectivamente. Cabe destacar, que los rangos definidos para el proceso están fundamentados en investigaciones realizadas en la entrega anterior.

Las reglas de asociación fueron obtenidas con la función *apriori* del paquete *arulesViz* de R. Se presentan las reglas con mayor confianza de las obtenidas en el proceso de

	Age
Niño	5 - 13
Adolescente	14 - 17
Adulto joven	18 - 35
Adulto	36 - 64
Tercera edad	65 - 100

Cuadro 2: Discretización de la variable Age

minería.

3.2. Reglas interesantes

Las reglas interesantes son aquellas que cumplen con un soporte mínimo y una confianza mínima. Los valores de estas medidas dependen completamente del problema.

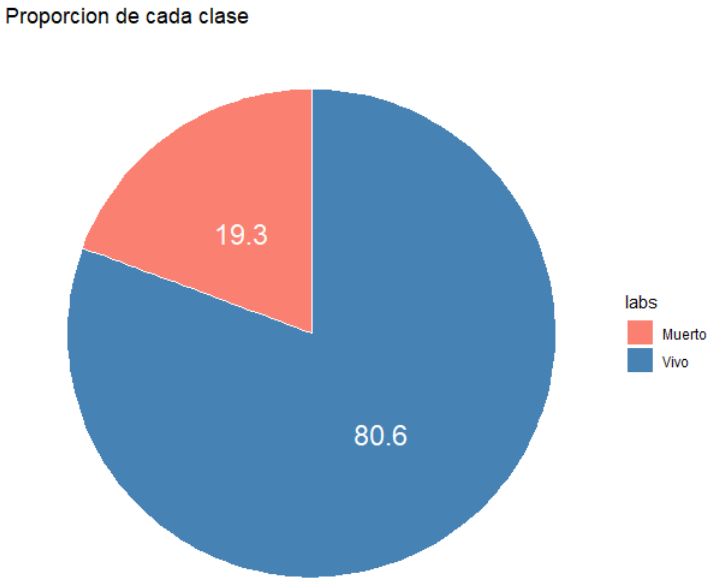

Figura 1: Proporción de cada clase.

Considerando que la clase “Muerto” se encuentra presente, en la base de datos de Hepatitis, menos de un 20 %, no se será tan exigente con el valor mínimo de soporte.

3.2.1. Clase muerto

Teniendo en cuenta que la clase *vivo* se encuentra presente en un 19.3 % en la base de datos, se escogió un soporte mínimo igual 0.08 y una confianza de 0.8. Obteniéndose las reglas de asociación presentes en la Tabla 3, la cual muestra las mejores 4 reglas (las 4 reglas con mayor confianza).

Antecedente	Consecuente	Soporte	Confianza	Lift
Ascites=Yes Ascites, Albumin=Albumin baja, Histology=No Histology	Class=Muerto	0.082758	0.8	4.1428
Sex=Hombre, Fatigue=YesFatigue, Bilirubin=Bilirubin alta, Albumin=Albumin baja	Class=Muerto	0.082758	0.8	4.1428
Malaise=Yes Malaise, Liver_Big=No Liver Big, Alk_Phosphate=Alk Phosphate normal, Histology=No Histology	Class=Muerto	0.082758	0.8	4.1428
Age=Adulto, Sex=Hombre, Fatigue=YesFatigue, Liver_Big=No Liver Big, Spiders=Yes Spiders, Alk_Phosphate=Alk Phosphate normal	Class=Muerto	0.082758	0.8	4.1428

Cuadro 3: Reglas de asociación con consecuente clase Muerto

Es posible observar que las reglas de asociación obtenidas, contemplan excelentes valores de *lift*, es decir, los resultados obtenidos no son casualidad. Además, la confianza es de un 80 % lo cual indica que la frecuencia con la que aparece el antecedente y el consecuente juntos es elevada.

3.2.2. Clase vivo

En cuanto a la clase *vivo*, se considera un soporte mínimo de 0.6, ya que presenta una mayor frecuencia en la base de datos, y una confianza de 0.8. La Tabla 4 muestra las 4 mejores reglas ordenadas por confianza.

En este caso, se puede observar que la confianza es elevada, en todos los casos superior a 90 %, indicando que el antecedente y consecuente se encuentran juntos casi el 100 % de las veces. Sin embargo, se debe ser cuidadoso, ya que el valor de *lift* a pesar de ser superior a 1, se encuentra muy cerca de no ser relevante.

Antecedente	Consecuente	Soporte	Confianza	Lift
Spiders=No Spiders, Varices=No Varices	Class = Vivo	0.60689	0.946	1.1726
Varices=No Varices, Bilirubin=Bilirubin normal	Class = Vivo	0.60000	0.945	1.1719
Spleen_Palpable=No Spleen Palpable, Ascites=No Ascites, Varices=No Varices	Class = Vivo	0.63448	0.938	1.1634
Spleen_Palpable=No Spleen Palpable, Albumin=Albumin normal	Class = Vivo	0.60689	0.926	1.1479

Cuadro 4: Reglas de asociación con consecuente clase Vivo

3.2.3. Sgot

También es interesante estudiar que otras variables están implicadas en una alta concentración de Sgot. Dado que, según comenta KidsHealth (2022), la Sgot es una de las enzimas que ayudan al hígado a transformar los alimentos en energía. Una alta concentración de esta enzima puede indicar que existe daño en el hígado.

Para las reglas de asociación obtenidas con consecuente "Sgot Alta" se consideró un soporte mínimo igual a 0.15 (valor escogido luego de explorar resultados con diferentes soportes) y una confianza de 0.8. La Tabla 5 muestra las 4 primeras reglas con mayor *lift*.

Antecedente	Consecuente	Soporte	Confianza	Lift
Fatigue=Yes Fatigue, Alk_Phosphate=Alk Phosphate alta	Sgot = Sgot Alta	0.17931	0.92857	1.5128
Malaise=Yes Malaise, Bilirubin=Bilirubin alta	Sgot = Sgot Alta	0.17931	0.92857	1.5128
Sex= Hombre, Fatigue= Yes Fatigue, Liver_Big= No Liver Big, Bilirubin=Bilirubin alta	Sgot = Sgot Alta	0.16551	0.92307	1.5038
Fatigue= Yes Fatigue, Spiders= Yes Spiders, Bilirubin=Bilirubin alta	Sgot = Sgot Alta	0.15862	0.92000	1.4988

Cuadro 5: Reglas de asociación con consecuente Sgot Alta

3.2.4. Albumin

Bajos niveles de Albumin en el cuerpo pueden ser un signo de enfermedades como: Enfermedades al hígado incluyendo cirrosis severa, hepatitis y enfermedad del hígado graso. Es por esto que es importante reconocer cuáles son los factores que implican bajos niveles de esta enzima. (MedlinePlus, 2022)

Para la obtención de reglas de asociación con consecuente "Albumin baja" se consideró un soporte de 0.1 dada la baja presencia del consecuente en la base de datos. También se tuvo en cuenta una confianza de 0.8.

El valor de *lift* en cada una de las reglas obtenidas es superior a 4, lo cual señala que existe una asociación real entre las variables y no es una coincidencia.

La Tabla 6 muestra las reglas obtenidas con mayor *lift*.

Antecedente	Consecuente	Soporte	Confianza	Lift
Liver_Big=No Liver Big, Ascites=Yes Ascites	Albumin=Albumin baja	0.117241	0.89473	4.3245
Fatigue=YesFatigue, Ascites=Yes Ascites	Albumin=Albumin baja	0.117241	0.85000	4.1083
Antivirals=No Antivirals, Ascites=Yes Ascites	Albumin=Albumin baja	0.117241	0.85000	4.1083
Sex=Hombre, Ascites=Yes Ascites	Albumin=Albumin baja	0.117241	0.85000	4.1083
Ascites=Yes Ascites, Histology=No Histology	Albumin=Albumin baja	0.10344	0.83333	4.0277

Cuadro 6: Reglas de asociación con consecuente Albumin Baja

4. Análisis de resultados y comparación

En el capítulo 3 se obtuvieron las mejores reglas de la base de datos (transacciones), tanto para la clase como para otras variables importantes consideradas como consecuentes. A continuación, los resultados estudiados serán analizados de acuerdo a la información presente en la literatura. Posteriormente, se compararán los resultados con aquellos obtenidos en el laboratorio de agrupamientos (*clustering*).

4.1. Análisis de reglas con consecuente "Clase = Muerto"

Se tiene que la clase *muerto* contempla el 19.3 % de los datos totales de la base de datos de *Hepatitis*, por lo tanto, se espera que la frecuencia de reglas que puedan aparecer luego del proceso de minería, contemplando a esta clase como consecuente, sea baja. Es por esto que se definió un soporte mínimo del 8 %. Sin embargo, cada una de las reglas presentes en la Tabla 3 son consideradas importantes, debido a que el *lift* es 4 veces superior al umbral. Además, cada una de las reglas tienen una confianza del 80 % por lo que, a pesar de su baja frecuencia, otorgan información relevante.

La primera regla que cuenta con la clase "Muerto" en su consecuente: {Yes Ascites, Albumin baja, No Histology ⇒ Clase=Muerto }, otorga información sobre que, cuando el paciente cuenta con ascitis, los niveles de *albumin* bajos y no se ha realizado estudios de tejidos y células (no *histology*), entonces el paciente muere. Y, de acuerdo a la literatura, si existe una relación, dado que la hypoalbuminemia (ocurre cuando hay niveles bajos de la proteína *albumin* en el cuerpo) puede poner en riesgo al paciente de padecer de una variedad de enfermedades, entre ellas *ascites* (Jewell, 2018).

Si bien la *ascites* no es una causa de muerte por si sola, las infecciones encubiertas o manifiestas pueden conducir a un empeoramiento de la vasodilatación, que provocará insuficiencia renal y finalmente, la muerte (Garcia-Tsao, 2017). Y considerando que la variable *Histology* en la primera regla se encuentra configurada con la respuesta *No* entonces se puede suponer, que no se descubrió a tiempo los efectos de la *ascites* en los pacientes.

La segunda regla, {Hombre, Yes Fatigue, Bilirubin alta, Albumin baja ⇒ Clase=Muerto}, señala que aquellos pacientes hombres que presentaron fatiga, altos niveles de

bilirrubina e hypoalbuminemia, finalmente murieron. Y, de acuerdo a un estudio realizado por Chen and Lin (2017), se encontró que de un total de 84 pacientes con HBV (hepatitis B), 41 de ellos fallecieron. De los pacientes fallecidos, 34 eran hombres (83 % de los pacientes muertos), presentaron hypoalbuminemia y una bilirrubina total elevada. Por lo que, se puede afirmar que las investigaciones respaldan los resultados obtenidos al menos para la segunda regla de asociación generada.

	Non-surviving patients (n = 41)	Surviving patients (n = 43)	*P*
Age, y	47.1 ± 11.2	45.0 ± 9.8	.345
Gender (male/female)	34/7	36/7	.922
Total protein, g/L	58.2 ± 7.0	59.6 ± 8.3	.310
Albumin, g/L	30.1 ± 4.0	33.8 ± 4.4	<.001
ALT, U/L	115.0 (66.8–272.0)	77.0 (39.5–175.0)	.060
Total bilirubin, μmol/L	369.0 (276.0–493.3)	288.0 (221.3–374.0)	.017
INR	2.22 ± 0.61	1.93 ± 0.46	.017
MELD score	24.0 (21.3–26.8)	19.60 (18.1–21.4)	.001
ALBI score	−0.87 (−1.06 to −0.67)	−1.24 (−1.53 to −0.96)	<.001
Child–Pugh score	9.7 ± 0.61	7.8 ± 1.5	<.001

Data are expressed as n, mean ± SD, or median (interquartile range).
ALBI score = albumin-bilirubin score, ALT = alanine aminotransferase, INR = International normalized ratio, MELD score = model for end-stage liver disease score.

Figura 2: Comparación de características clínicas entre pacientes que sobrevivieron y no sobrevivieron con insuficiencia hepática aguda crónica (Chen and Lin, 2017)

.

En cuanto a la fatiga (Fatigue), según Cleveland Clinic (2022), las personas con el síndrome de Gilbert heredan un gen mutado que afecta la capacidad del hígado para procesar la bilirrubina. Las personas con síndrome de Gilbert no producen suficientes enzimas hepáticas para mantener la bilirrubina a un nivel normal. Como resultado de esto, se acumula un exceso de bilirrubina en el cuerpo (hiperbilirrubinemia).

Dentro de los diferentes síntomas que produce el síndrome de Gilbert, se encuentra la fatiga. Es por esto, que se da la relación encontrada por la regla de asociación.

La tercera regla, {Yes Malaise, No Liver Big, Alk Phosphate normal, No Histology $\Rightarrow$ Clase=Muerto}, indica que aquellos pacientes que presentaron malestar y no se realizaron estudios de células y tejidos, pero no tuvieron un hígado agrandado y además sus niveles

de *alk phosphate* estaban en niveles normales, murieron. Información que parece no aportar demasiado, sin embargo, la confianza es alta. Los pacientes que cumplen con las características de la tercera regla, es probable que hayan muerto por no tratar la enfermedad. Y como comenta Clinic (2022b), el malestar o la sensación general de no sentirse bien, es un síntoma de tener algún problema asociado al hígado. Sumando que no estudiaron el estado interno de sus órganos, la probabilidad de morir por no tratar una enfermedad en estado avanzado, es alta.

Por último, la última regla asocia a los pacientes que son adultos, hombres, con fatiga (yes fatigue), hígado no agrandado, con spiders y niveles normales de alk phosphate, con la clase *muerto*. Trust (2022) menciona que los angiomas de araña (spiders) son un síntoma temprano de un mal funcionamiento del hígado, y la fatiga un síntoma tardío (o cuando la enfermedad está más avanzada) de que el hígado está teniendo dificultades en su correcto funcionamiento. Por lo tanto, estas dos características juntas, permiten concluir que el paciente no trato la enfermedad a tiempo y por lo tanto, un empeoramiento en su estado de salud pudo haber concluido en su posterior deceso.

4.2. Análisis de reglas con consecuente "Clase = Vivo"

Como ya se adelantaba en la sección 3.2.2, vemos que las reglas poseen una confianza muy cercana a 1, lo que podría parecer muy bueno. Aun así, el *lift* es solamente décimas, mayor a 1. Esto es que, si bien, la mayoría de las veces los atributos aparecen juntos en las transacciones (base de datos), estos están débilmente correlacionados (positivamente). Teniendo esto en consideración, se detalla cada una de las reglas:

- Primero, tenemos la regla de antecedente {No Spiders, No Varices}. Como es sabido de la experiencia número 1, por un lado, las arañas o *spiders* son un síntoma principal de hepatitis de tipo C según el Recovery Care Blog (2021) y se presentan visiblemente en la piel, mientras que las varices, en este contexto, son esofágicas, y pueden ser producidas por afecciones hepáticas como la cirrosis producida por la hepatitis, pero también por el consumo de alcohol en exceso (Top Doctors INC, s. f.). Así, vemos que ambos datan de una grave condición, haciendo sentido que su ausencia implique una alta posibilidad

de supervivencia.

- En segundo lugar, tenemos la regla con antecedente {No Varices, Bilirubin Normal}. La variable *Bilirubin Normal* representa un rango de la bilirrubina definido como normal (recordar el preprocesamiento de datos inicial), y la falta de varices se detalla en la regla anterior, por lo que la presenta de falta de varices y el nivel normal de bilirrubina muestra concordancia.

- Con la tercera regla con antecedente {No Spleen Palpable, No Ascites, No Varices}, se observa que tenemos la ausencia de varices una tercera vez, pudiendo concluir que influye altamente al momento de la predicción de la clase *Vivo*. También se observa la falta de la condición de *ascites*, obtenida por infección crónica de hepatitis C o B (MedlinePlus, s. f.), junto a un bazo no palpable (*spleen palpable*). La presencia de un bazo palpable puede ser síntomas de esplenomegalia, justamente relacionado con hepatitis (crónica, B o C, según Vargas Viveros et al. (2013)), lo que da indicios teóricos de la correcta asociación de la falta de *ascites* y *spleen palpable* unidos en el antecedente, para determinar que se sobrevive (clase Vivo).

- {No Spleen Palpable, Albumin Normal}: Nuevamente nos encontramos con que un bazo no palpable implica la clase Vivo, ahora junto a niveles normales de *albumin* en el antecedente. En la sección 4.1, se detallan los problemas de la *hypoalbuminemia*, por lo que es lógico que la albumina, en niveles normales, ayude a concluir la clase *Vivo*.

4.3. Análisis de otras reglas interesantes

Adicionalmente, se estudiaron otras variables como consecuente de las reglas de asociación, para determinar que factores inciden directamente en su comportamiento. En primer lugar, la Tabla 5 muestra con un 92 % de confianza las siguientes asociaciones mas relevantes:

- {Yes Fatigue, Alk Phosphate alta} implican un nivel elevado de Sgot.

- {Yes Malaise, Bilirubin alta} implican un nivel elevado de Sgot.

La primera regla, asocia elevados niveles de la enzima ALP (Alk phosphate) y fatiga con elevados niveles de Sgot. Clinic (2022a) afirma que una de las principales causas de tener ALP (Alk phosphate) alta es producto de daño hepático debido a la inflamación del hígado, es decir, hepatitis. Y dentro de las pocas condiciones que se deben dar para que los niveles de Sgot se eleven se encuentra la Hepatitis C Aggarwal (2022).

La segunda regla, afirma que aquellos pacientes con malestar (Yes Malaise) y bilirrubina alta, tuvieron elevados niveles de Sgot. Cuando los niveles de bilirrubina se encuentran altos, se producen dolores en el pecho del paciente y debilidad, lo cual podría explicar el malestar (*Malaise*). Este malestar y altos niveles de bilirrubina son una causa directa de algún problema en el hígado, como por ejemplo la infección por hepatitis Rossiaky (2022). Y, por lo tanto, tal como indica Health (2022), la hepatitis viral aguda A, B y C son una de la causa de que la *Sgot* se encuentre elevada. Es decir, tener malestar y la bilirrubina alta implica tener algún problema al hígado como la hepatitis, y, tener hepatitis implica tener los niveles de *Sgot* elevados.

Por otro lado, también es interesante estudiar la *Albumin*, como se detalla en experiencias anteriores y en la sección 3.2.4, que en la teoría está relacionada con variedad de enfermedades al hígado, en este caso, cuando es baja. Tomando la regla con mayor *lift*, tenemos $\{\text{No Liver Big, Yes Ascites}\} \Rightarrow \{\text{Albumin baja}\}$. En este caso, se destaca que el top de las mejores reglas con consecuente *albumin baja* (tabla 6) posee la presencia de *Ascites* en su antecedente. Según Facciorusso et al. (2011), como la albumina constituye aproximadamente la mitad de las proteínas en el plasma, "it is commonly employed in cirrhotic patients in association with diuretics for the treatment of ascites". Así, si bien la regla no contribuye directamente al problema de la hepatitis, sí lo hace entre las variables en su conjunto.

4.4. Comparación experiencias anteriores

Para tener una mejor perspectiva del problema, es decir, el análisis de la enfermedad y base de datos de la hepatitis, se comparan los resultados obtenidos de la minería de reglas de asociación con los resultados del proceso de agrupamiento utilizando el método de los k-medioides, recordando que tal agrupamiento era no supervisado, es decir, no era necesario entregar las etiquetas de las clases, ya que el algoritmo se encarga de agrupar

según la matriz de distancias provistas de antemano, con el fin de descubrir si tenemos más conocimiento a través de las reglas o saber si el agrupamiento fue suficiente para el análisis.

4.4.1. Grupo 1: Vivos

El grupo número 1, en la experiencia anterior, fue determinado en gran cantidad por un gran porcentaje (88 %) de pacientes etiquetados con la clase *vivo*, concluyendo que este grupo contenía a aquellos pacientes capaces de superar los cuadros más graves de la enfermedad, o los sobrevivientes. En este grupo, se había detallado que *bilirubin*, *alk phosphate* y *sgot* estaban en un nivel normal, bajo en comparación a las mediciones para el grupo 2, para variables continuas; para aquellas binarias, se identificaba *histology* (presente en su mayoría) y *liver firm* (no presente mayormente). Comparando con las reglas, vemos que la primera, {No Spiders, No Varices} aporta dos atributos que fueron omitidos en el agrupamiento. La segunda regla {No Varices, Bilirubin Normal} indica *Bilirubin* en un nivel normal como ya se tenía conocimiento con el *clustering*, agregando la falta de *varices*.

Por otro lado, con la tercera y cuarta regla para un consecuente de la clase *vivo*,

- {No Spleen Palpable, No Ascites, No Varices}
- {No Spleen Palpable, Albumin Normal},

vemos el aporte de *Spleen Palpable* = 0, o su ausencia junto con la de la variable *Ascites* y un nivel normal de *Albumin* (recordando que fue categorizada).

4.4.2. Grupo 2: Muertos

Las características del grupo 2 obtenido a través del algoritmo de *clustering* desarrollado en el laboratorio anterior, permitieron concluir que en este grupo fueron clasificados aquellos pacientes con un alto riesgo de morir. Es por esto que, las reglas de asociación obtenidas que contemplan en su estructura, a la clase *muerto* en el consecuente, son las más adecuadas para realizar el análisis comparativo.

Altos niveles de *Alk Phosphate*, *Bilirubin* y *Sgot* diferencian a los pacientes del grupo 2, de aquellos del grupo 1. Además, se logró reconocer que los pacientes (en su mayoría hombres) presentaron en mayor medida hígado firme (liver firm) y ningún estudio de células

y tejidos. Al comparar estos resultados con la nueva información obtenida de las reglas de asociación (Tabla 3), basándonos en lo descrito en la sección 4.1, se logra corroborar que aquellos pacientes con mayor riesgo de muerte (o muertos directamente), no realizaron estudios acabados de células y tejidos. De acuerdo a Thompson (2017), muchas de las enfermedades que afectan al cuerpo humano ocurren en el ámbito de los tejidos, como las consecuencias de la hepatitis. La histología puede ayudar a reconocer que tratamiento es mejor utilizar para atacar a la enfermedad en el paciente. Sin embargo, de no realizarse se ignora una herramienta potente para asegurar la integridad del paciente.

La *Bilirubin* alta, en ambos estudios, se ha reconocido como una potencial causa de muerte combinada con otros nuevos factores, reconocidos gracias a la minería de reglas de asociación. Entre estos nuevos factores se encuentran, la *Ascites*, *Albumin* baja y la fatiga (en la sección 4.1 se detallan los efectos de estos atributos en los pacientes). Es importante reconocer que existen otros factores que inciden en la muerte de un paciente, dado que de ignorarse es posible llevar a cabo por parte del especialista médico un tratamiento equivocado, o incluso un diagnóstico erróneo por desconocimiento de la incidencia de la ascitis, niveles bajos de albúmina y fatiga en el riesgo de muerte de pacientes con hepatitis.

A diferencia del algoritmo de *clustering*, las reglas de asociación no reconocieron como causas diferenciadoras o causas que impliquen un efecto mortal en los pacientes, a los elevados niveles de *Sgot* y *Alk Phosphate*. Sin embargo, si bien tener elevados niveles de *Sgot* podría no ser mortal, si acorta la esperanza de vida en hombres en aproximadamente 17.2 años, cuando los niveles de *Sgot* superan las 70 unidades internacionales por litro.

5. Conclusiones

A partir de la minería de reglas de asociación, se obtuvieron 4 reglas tanto para la clase *vivo* y *muerto*. Las reglas para la clase *muerto* se obtuvieron fijando un soporte de 0.08 (debido al desbalance de clase, del que se tiene cuenta desde la primera experiencia), y resultaron con un buen nivel de *lift* (mayor a 4) y una confianza de 0.8. Así mismo, las reglas para la clase *vivo* se obtuvieron fijando un soporte mayor, dado que la base de datos posee más de esta clase, resultando con una buena confianza, aunque poco *lift*, pero mayor que 1.

Se destaca en los resultados obtenidos el descubrimiento de nuevas variables que aportan información relevante para entender la muerte de los pacientes. Como por ejemplo, las reglas 1 y 2 de la tabla 3, que incluyen dentro de sus antecedentes a la variable *Albumin baja* que también se conoce como hypoalbuminemia. Jewell (2018) aportó información importante al señalar que este padecimiento es una causa potencial de adquirir *ascites*, y tal como lo comenta la primera regla, la combinación de estos dos factores en los antecedentes es una potencial causa de muerte en pacientes que cuentan con hepatitis.

Demographic information on patients, n=5451.

Variable	Overall n=5894	Low (0–34 g/L), n=742	Normal (35–44 g/L), n=3840	High (>44 g/L), n=869	Albumin missing, n=443
Female, n (%)	2950 (50.1%)	370 (49.9%)	1910 (49.7%)	443 (51.0%)	227 (51.2%)
Age, median years (IQR)	65 (49–77)	74 (64–83)	67 (53–77)	49 (34–64)	60 (41–74)
Length of stay, median days (IQR)	2 (1–6)	6 (2–12)	2 (1–6)	1 (0–2)	1 (0–3)
Charlson comorbidity score, median score (IQR)	1 (0–3)	2 (1–4)	1 (0–3)	0 (0–1)	1 (0–3)
WPS, median score (IQR)	1 (0–2)	2 (1–4)	1 (0–2)	0 (0–2)	1 (0–2)
30-day mortality, n (%)	332 (5.6%)	121 (16.3%)	165 (4.3%)	14 (1.6%)	32 (7.2%)

Figura 3: Porcentaje de mortalidad a los 30 días de acuerdo a los niveles de albumin (Jellinge et al., 2014).

Además, como se observa en la Figura 3, en un estudio realizado por Jellinge et al. (2014) con 5.894 pacientes se concluyó que la *hypoalbuminemia* está asociada con

la mortalidad a los 30 días en pacientes médicos ingresados de forma aguda. Por lo que, la minería de reglas de asociación estableció correctamente la relación entre la muerte del paciente y la causa.

En forma adicional, otro punto a destacar es que las variables más representativas de los grupos entregados por k-medioides no se repitieron tanto como se esperaba en los conjuntos de reglas. Se sugiere que una de las razones de esto es porque la función que ejecuta el algoritmo *A priori* en R permite buscar exactamente aquellas reglas con el consecuente basado en la clase dada, mientras que, sin supervisión, el algoritmo de *clustering* agrupa en cada paso de forma golosa o como un algoritmo *greedy*, cuyas soluciones son locales, y también debido a que los grupos contenían algunas observaciones mezcladas de clases (sin separación al 100 %), resultando en diferentes atributos representativos para las clases, mostrando que el desbalance de clases de la base de datos afecta más en el agrupamiento que en las reglas de asociación.

Por otro lado, se destaca que, a través del proceso de obtención de reglas de asociación, también se evidencian las diferencias en funcionamiento con algoritmos de *clustering*. Mientras que el segundo agrupa basado en medidas de similaridad o distancias, pudiendo tener muchos *clusters* distintos, dependiendo del autor de la investigación —y su conocimiento en el tema asociado— la elección del mejor, en las reglas de asociación solo obtenemos una descripción en profundidad de nuestra base de datos mediante *reglas* que relacionan antecedentes y consecuentes (constituidos por atributos) de forma lógica, utilizando los *itemsets* frecuentes, por lo que aplicaríamos *clustering* para explorar nuestros datos en forma inicial, sin sacar fuertes conclusiones antes de investigar, y familiarizarnos con ellos, en tanto las reglas de asociación actúan como una herramienta en un nivel profesional (como se demostró en el marco teórico, muchas medidas pueden ser reformuladas en términos de probabilidades, gracias al teorema de Bayes).

Así mismo, como en el estudio realizado con las herramientas de *clustering*, se destaca en esta instancia la importancia de estudiar con mayor énfasis las cualidades o factores que inciden en el empeoramiento de los pacientes que cuentan con hepatitis.

Bibliografía

Aggarwal, K. (2022). What causes increase in sgot and sgpt level? https://www.medtalks.in/articles/what-causes-of-sgot-and-sgpt-indicator-increase.

Chen, B. and Lin, S. (2017). Albumin-bilirubin (albi) score at admission predicts possible outcomes in patients with acute-on-chronic liver failure. *Medicine*, 96(24).

Cleveland Clinic (2022). Gilbert's syndrome: Symptoms, causes, tests & treatment. https://my.clevelandclinic.org/health/diseases/17661-gilberts-syndrome.

Clinic, C. (2022a). Alkaline phosphatase (alp): What it is, causes & treatment. https://my.clevelandclinic.org/health/diagnostics/22029-alkaline-phosphatase-alp.

Clinic, M. (2022b). Symptoms of liver disease. https://britishlivertrust.org.uk/information-and-support/liver-health-2/symptoms-of-liver-disease/.

Facciorusso, A., Nacchiero, M., Rosania, R., Laonigro, G., Longo, N., Panella, C., and Ierardi, E. (2011). The use of human albumin for the treatment of ascites in patients with liver cirrhosis: item of safety, facts, controversies and perspectives. *Current Drug Safety*, 6(4).

Garcia-Tsao, G. (2017). Ascites. *Liver Pathophysiology*, page 475–484.

Health, I. (2022). Lower sgpt & sgot level naturally: Avoid liver damage.

Jellinge, M. E., Henriksen, D. P., Hallas, P., and Brabrand, M. (2014). Hypoalbuminemia is a strong predictor of 30-day all-cause mortality in acutely admitted medical patients: A prospective, observational, cohort study. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141840/.

Jewell, T. (2018). Hypoalbuminemia: Causes, treatment, and more. https://www.healthline.com/health/hypoalbuminemia#complications.

KidsHealth (2022). Análisis de sangre: aspartato aminotransferasa (Ast o SGOT). https://kidshealth.org/es/parents/test-ast.html#:~:text=La%20AST%20tambin%20se%20llama,rebosan%20desde%20las%20clulas%20hepticas.

Martinez, C. (2020). Reglas de asociacion. .

MedlinePlus (2022). Albumin blood test: Medlineplus medical test. https://medlineplus.gov/lab-tests/albumin-blood-test/.

MedlinePlus (s. f.). Ascitis: MedlinePlus, enciclopedia médica. https://medlineplus.gov/spanish/ency/article/000286.htm.

Recovery Care Blog (2021). Signs and Symptoms of Hepatitis C. https://recovery.care/what-are-signs-and-symptoms-of-hepatitis-c/.

RevereHealth (2016). How your body systems are connected - revere health: Live better. https://reverehealth.com/live-better/how-body-systems-connected/.

Rodrigo, J. A. (2018). Reglas de asociación y algoritmo apriori con r. https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion#:~:text=Una%20regla%20de%20asociacin%20se,hand-side%20(RHS).

Rossiaky, D. (2022). High bilirubin levels: Symptoms, causes, and treatment.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Pearson.

Tan, S. C. (2018). Improving association rule mining using clustering-based discretization of numerical data. *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*.

Thompson, V. (2017). What are the branches of physiology? https://education.seattlepi.com/branches-physiology-7043.html.

Top Doctors INC (s. f.). Varices esofágicas: qué es, síntomas y tratamientos. https://www.topdoctors.es/diccionario-medico/varices-esofagicas.

Trust, B. L. (2022). Symptoms of liver disease. https://britishlivertrust.org.uk/information-and-support/liver-health-2/symptoms-of-liver-disease/.

Vargas Viveros, P., Hurtado Monroy, R., and Villalobos Alva, J. A. (2013). Esplenomegalia. *Revista de la Facultad de Medicina (México)*, 56(2).