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


# Laboratorio 3 - Reglas de Asociación

| | |
|---|---|
| Integrantes: | Gustavo Aguilar Morita |
| | Sandra Hernández Yamunaqué |
| Curso: | Inteligencia Computacional |
| Programa: | Magíster en Ingeniería Informática |
| Profesor Cátedra: | Max Chacón Pacheco |
| Profesor Laboratorio: | Héctor Rojas Pescio |

5 de Julio de 2021

# Tabla de contenidos

# 1. Introducción

A medida que pasa el tiempo se generan cada vez más datos, los cuales al ser procesados utilizando diversos métodos permiten obtener información, lo que resulta útil para la toma de decisiones. Una de las técnicas usadas para poder tomar decisiones corresponde a las reglas de asociación, técnica que busca mediante el reconocimiento de patrones descubrir hechos que ocurren frecuentemente sobre un conjunto de datos, y así poder apoyar en la toma de decisiones y encontrar el mejor resultado de acuerdo a un objetivo específico.

En la siguiente entrega se presenta el desarrollo del tercer laboratorio del curso de Inteligencia Computacional, el cual consiste en continuar con el análisis de la base de datos hepatitis.

## 1.1. Objetivos

Entre los objetivos a cumplir se encuentra:

1. Implementar diversas reglas de asociación sobre la base de datos hepatitis.

2. Utilizar las reglas de asociación como clasificador del dataset mencionado anteriormente, utilizando el lenguaje de programación R.

3. Hacer uso del package: arulesViz, y utilizar métodos para la visualización y selección de reglas interesantes.

4. Realizar un análisis comparativo respecto a los laboratorios realizados en entregas anteriores (laboratorio 1 y 2).

# 2. Marco Teórico

## 2.1. Reglas de asociación

Las reglas de asociación son un tipo de aprendizaje no supervisado, por lo que no debe ser entrenado para su implementación y en el cual no es posible obtener el error en sus resultados. El principal objetivo de esta técnica es descubrir relaciones dentro de un conjunto de datos que tienden a ocurrir de manera conjunta. Con estas relaciones se generan patrones que permiten tomar decisiones a la hora de agrupar los datos.

De forma simplificada, las reglas de asociación se representan como conjuntos de antecedentes y consecuentes como se ejemplifica a continuación:

$$\{X, Y\} \Rightarrow \{Z\}$$

La regla anterior se puede representar en el lenguaje natural como "Cuando ocurren X y Y, también ocurre Z".

El conjunto de agrupaciones que se generan desde los datos que se tienen se conoce como transacciones . Para el caso anterior $\{X, Y\}$ es una transacción.

Un ejemplo de lo anterior es cuando un supermercado decide hacer un pack de productos, y genera reglas de asociación para ver de qué forma generar dichos packs. Con una regla, se podrían dar cuenta de que cada vez que la gente compra pan de completo, vienesas y tomate, también lleva palta. Esto se puede representar como la siguiente regla:

$$\{Vienesas, Pan, Tomate\} \Rightarrow \{Palta\}$$

Con esto, el supermercado sabe que le conviene hacer un combo con vienesas, pan de completo, tomates y paltas.

## 2.2. Soporte y Confianza

El soporte equivale al número de transacciones que contienen a un elemento X dividido entre el total de transacciones. Si se tiene $Sop(X \Rightarrow Y)$, se entiende como el número de transacciones en el conjunto S tal que X e Y son verdaderos simultáneamente.

El soporte también se puede normalizar, para esto se realiza la siguiente operación:

$$Sop(X \Rightarrow Y)n = (X \Rightarrow Y)n$$

Siendo n la cantidad total de transacciones del conjunto S. Por otra parte, la confianza es la probabilidad de que una transacción que contiene el o los elementos de X, también contenga el o los elementos de Y.

$$Confianza(X \Rightarrow Y) = \frac{soporte(X \cup Y)}{soporte(X)}$$

## 2.3. Principio de Monotocidad

El principio de monotonicidad permite descartar ítemsets que no se presentan en las transacciones para evitar crear reglas con ellos y ası no sobrecargar el procesamiento de los datos. Este principio consta de dos partes:

1. Si un itemset es frecuente, entonces todos los subgrupos de éste también son frecuentes
2. Si un itemset no es frecuente, entonces cualquier conjunto que contenga a este itemset tampoco lo será.

Con esto, se pueden descartar subconjuntos poco frecuentes, y así buscar solo los ítems frecuentes necesarios para generar las reglas de asociación.

Esto es útil debido a que el orden de generar todos los subconjuntos de ítems es de $2^n1$ siendo $n$ el número de ítems del conjunto. Los algoritmos conocidos como "Apriori" y "Eclat" para generar reglas de asociación utilizan este principio para optimizar la creación de reglas.

## 2.4. Medidas de calidad

A continuación se describen diversas métricas para determinar la calidad de las reglas generadas, así saber qué reglas son importantes o relevantes para cada clase según un determinado problema.

- **Lift:** Es una medida usada en las herramientas de mineria de datos de IBM. Representa una medida de independencia entre X e Y. Tiene su valor más bajo, cuando X e Y son completamente independientes.

$$Lift(X \Rightarrow Y) = \frac{n * conf(X \Rightarrow Y)}{sop(Y)}$$

Si $lift = 1$: El conjunto aparece una cantidad de veces acorde a lo esperado bajo condiciones de independencia.

Si $lift > 1$: El conjunto aparece una cantidad de veces mayor a lo esperado bajo condiciones de independencia (mayor evidencia de que la regla represente un patrón real).

Si $lift < 1$: El conjunto aparece una cantidad de veces inferior a lo esperado bajo condiciones de independencia.

- **Convicción:** Representa la independencia entre X e Y. Es similar a lift.

$$convicción = \frac{n - Sop(Y)}{n[1 - conf(X \Rightarrow Y)]}$$

Puede ser interpretado como la relación de la frecuencia esperada que X se produce sin Y (es decir, la frecuencia que la regla hace una predicción incorrecta)

- **LaPlace:** Estima la precisión de predicción de una regla. Esta estimación se realiza de la siguiente forma:

$$LaPlace = \frac{Sop(X \Rightarrow Y) + 1}{\frac{Sop(X \Rightarrow Y)}{c} + k}$$

Donde $k$ es el número de clases del problema y $c$ es $p(X \cap Y)$.

# 3. Obtención de la reglas

## 3.1. Balance de los datos

Antes de obtener las reglas de asociación se debe tener en cuenta que el balance de los datos sea similar antes de aplicar cualquier algoritmos, esto se debe a que se busca que las clases tengan la misma o parecida probabilidad de ser consecuentes de las reglas a generar bajo los parámetros de selección escogido (confianza, soporte, o de acuerdo a la calidad) y de esta manera obtener la mayor cantidad de reglas para las clases. Nuestro dataset cuenta con dos clases como se puede ver en la figura.1, las cuales no tienen una cantidad de datos similar por lo que no se encuentra balanceada nuestra data, luego para corregir esto se aplicaron diferentes técnicas de balanceo, las cuales son explicadas con detalle en Capitulo 4. Como se

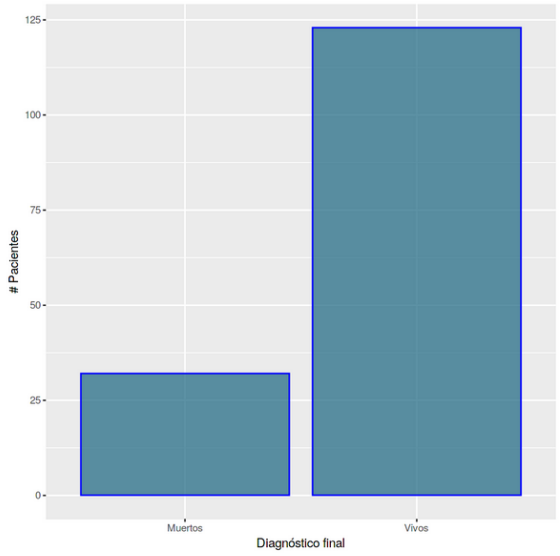
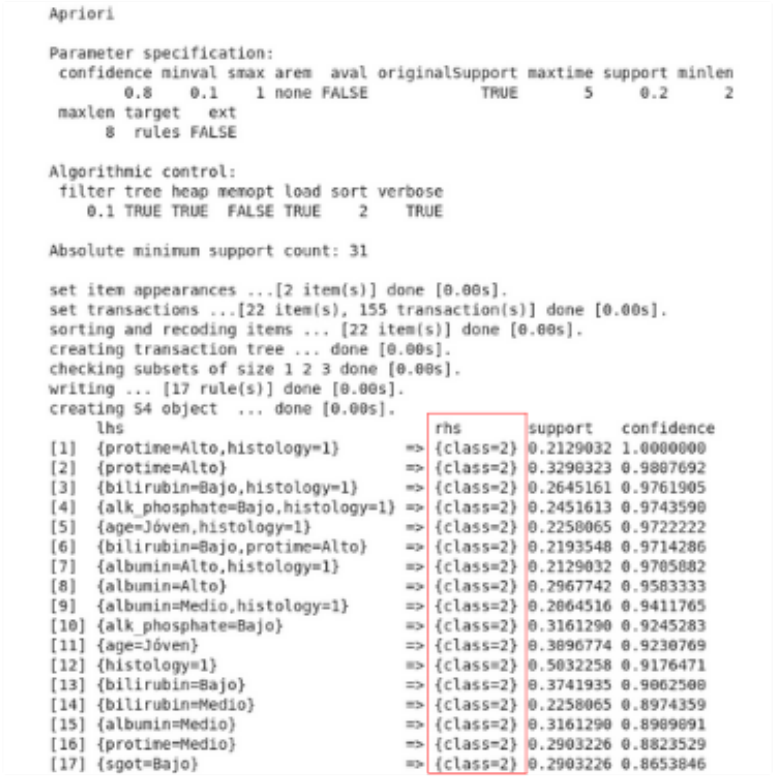



Figura 1: Histograma de la clase: siendo 1: muerto, 2: vivo

menciono anteriormente, si se utiliza una data desbalaceada( el cual es nuestro caso), para generar reglas de asociación se puede "perderreglas en el proceso e incluso no tener reglas que tengan como consecuencia una de las clases. En la figura.2 se puede observar este fenómeno.



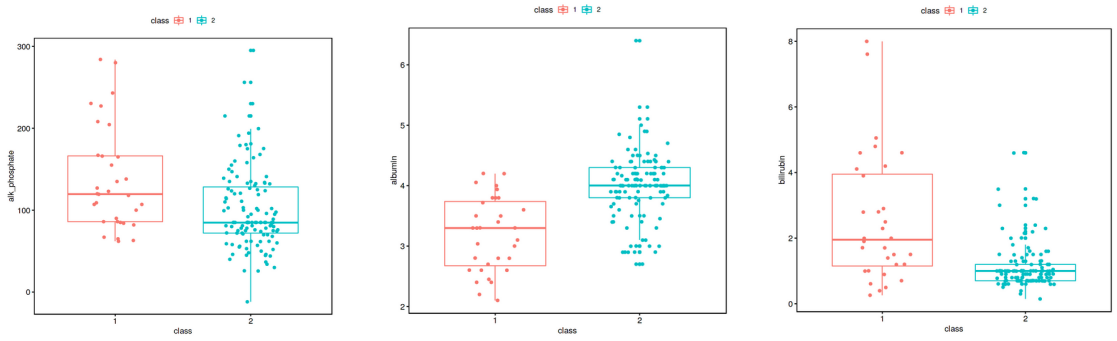
```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5     0.2      2
 maxlen target   ext
      8  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 31

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[22 item(s), 155 transaction(s)] done [0.00s].
sorting and recoding items ... [22 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [17 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs                                  rhs         support   confidence
[1]  {protime=Alto,histology=1}       => {class=2} 0.2129032 1.0000000
[2]  {protime=Alto}                   => {class=2} 0.3290323 0.9807692
[3]  {bilirubin=Bajo,histology=1}     => {class=2} 0.2645161 0.9761905
[4]  {alk_phosphate=Bajo,histology=1} => {class=2} 0.2451613 0.9743590
[5]  {age=Jóven,histology=1}          => {class=2} 0.2258065 0.9722222
[6]  {bilirubin=Bajo,protime=Alto}    => {class=2} 0.2193548 0.9714286
[7]  {albumin=Alto,histology=1}       => {class=2} 0.2129032 0.9705882
[8]  {albumin=Alto}                   => {class=2} 0.2967742 0.9583333
[9]  {albumin=Medio,histology=1}      => {class=2} 0.2064516 0.9411765
[10] {alk_phosphate=Bajo}             => {class=2} 0.3161290 0.9245283
[11] {age=Jóven}                      => {class=2} 0.3896774 0.9230769
[12] {histology=1}                    => {class=2} 0.5032258 0.9176471
[13] {bilirubin=Bajo}                 => {class=2} 0.3741935 0.9062500
[14] {bilirubin=Medio}                => {class=2} 0.2258065 0.8974359
[15] {albumin=Medio}                  => {class=2} 0.3161290 0.8909091
[16] {protime=Medio}                  => {class=2} 0.2903226 0.8823529
[17] {sgot=Bajo}                      => {class=2} 0.2903226 0.8653846
```

Figura 2: Obtención de reglas sin balancear dataset: hepatitis. Como se puede observar en el cuadro rojo solo tenemos reglas para la clase 2, lo que explica las consecuencias de usar no balanceados para la obtención de reglas

### 3.1.1. Análisis previo

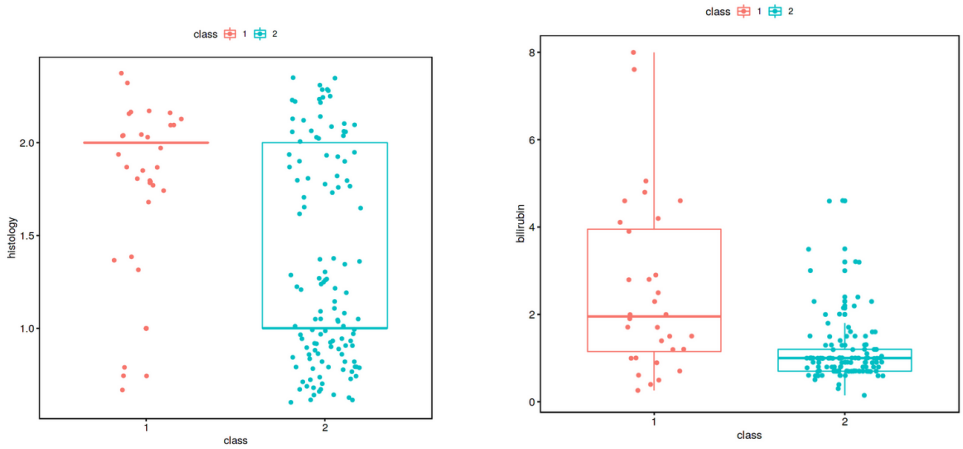
Se analizan aquellas variables de interés, estas son aquellas variables que han aparecido continuamente durante el desarrollo de las experiencias pasadas. En el Capitulo 4 se realiza un mayor análisis con respecto a la comparativa entre las diferentes experiencias y como las variables de repiten o no en las reglas obtenidas.

A continuación se muestran gráficos de caja y de histogramas de las variables de interés. A modo general se puede observar a simple viste que la mayoría de los gráficos nos indican que las variables no se distribuyen de manera normal, a excepción de protime. Dado que las distribuciones entre las variables son bastante variables es que se considero generar 3 categorías para este tipo de variables, la cual se detallan en Capitulo 3.1.2.




(a) alk-phophate (b) albumina (c) bilirrubina

Figura 3: Gráfico de cajas que muestra las distribuciones de los datos con respecto a cada una de las variables.

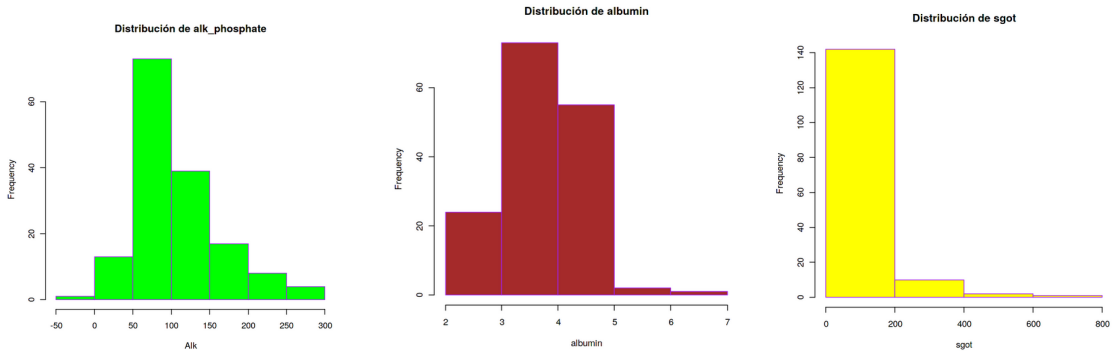



(a) histology (b) protime

Figura 4: Gráfico de cajas que muestra las distribuciones de los datos con respecto a cada una de las variables.

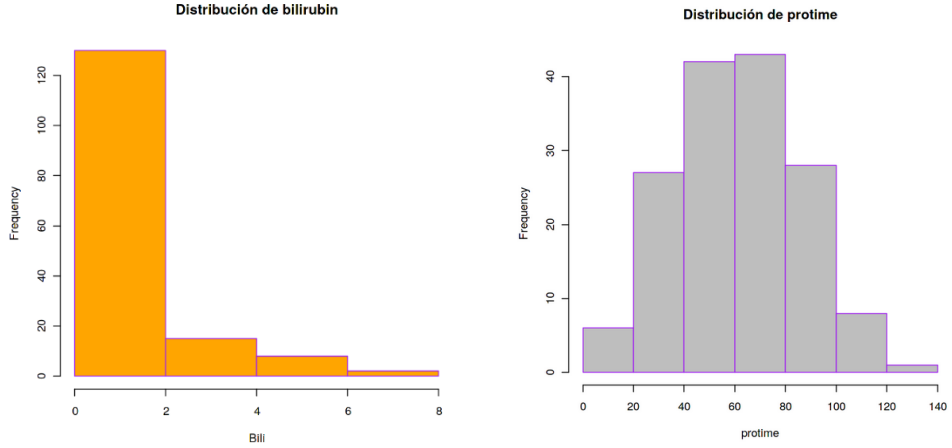
### 3.1.2. Modificación de los datos

Hay que tener en claro que para usar el algoritmo Apriori de R se debe trabajar con datos de tipo discreto por lo que se modificaron las variables de interés que eran de tipo continua. Para empezar se decidieron la cantidad de grupos o categorías se tendrían para cada una de las variables, para mantener la consistencia en los datos la cantidad seria igual para cada una de las variables.




(a) alk-phophate (b) albumina (c) sgot

Figura 5: Histogramas que muestra las distribuciones de los datos con respecto a cada una de las variables.




(a) bilirrubina (b) protime

Figura 6: Histogramas que muestra las distribuciones de los datos con respecto a cada una de las variables.

Luego de observar los gráficos de distribución de cada una de las varíales de interés y notar que existen diversos tipos de resultados es que se opta por agrupar en tres categorías, las cuales son bajo, medio y alto para cada una de las variables (bilirubin, alk-phosphate, sgot, albumin y protime).

Cabe destacar que los cortes o asignación de los datos a cada uno de las categorías se hicieron en función del percentil.

## 3.2. Medida de calidad

La medida de calidad para determinar cual de las reglas es la mejor corresponde al lift, esto debido a que entrega información sobre el grado de independencia de los ítems, es decir, indica si la regla aparece una cantidad mayor a lo esperado dentro del conjunto de datos, logrando así determinar si exite más evidencia de que efectivamente la regla representa un patrón real.

Además, en caso de que dos o más reglas tengan el mismo lift, el algoritmo apriori también utiliza la cobertura, la cual a través de la confianza y soporte determina la frecuencia con la que se puede aplicar la regla.

## 3.3. Obtención de reglas

Los parámetros escogidos para seleccionar y generar las reglas a considerar como relevantes para el desarrollo de esta experiencia son la siguientes:

- **Soporte= 0.2**: Se escoge este nivel de soporte dado que nos entrega una buena cantidad de reglas para ambas clases. Cabe destacar que al hacer pruebas con un soporte igual a 0.3 la cantidad de reglas generadas eran mucho menores a las obtenidas con un soporte igual a 0.2, además de no entregar reglas para amabas clases en algunos casos.

- **Confianza= 0.9**: Se considera que una confianza del 90 %, pues es el porcentaje mínimo para considerar resultados de buena calidad o en este caso reglas de buena calidad. Además con estudios médicos o relacionados al área de la salud no se puede ser tan poco exigente a la hora de buscar conclusiones pertinentes a la vida de una persona.

- **lift $\geq$ 1.8**: Un lift superior a 1, nos dice que un conjunto aparece más veces de lo esperado lo que nos da indicios de que la regla representa una mayor evidencia a ser un patrón real. El 0.8 extra que se exige es para filtrar de manera razonable la cantidad de reglas obtenidas.

A continuación se procede a mostrar aquellas reglas relevantes y que cumplen con los parámetros mencionados anteriormente:

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5     0.2      2
 maxlen target   ext
      8  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 49

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[26 item(s), 246 transaction(s)] done [0.00s].
sorting and recoding items ... [25 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [48 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs                    rhs         support confidence     lift count
[1]  {protime=Alto}      => {class=2} 0.2073171  1.0000000 2.000000    51
[2]  {liver_big=2,
      spiders=1,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2276423  0.9824561 1.964912    56
[3]  {liver_big=2,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2967480  0.9605263 1.921053    73
[4]  {liver_big=2,
      spiders=1,
      protime=Bajo,
      histology=2}       => {class=1} 0.2398374  0.9516129 1.903226    59
[5]  {liver_big=2,
      spiders=1,
      albumin=Bajo,
      histology=2}       => {class=1} 0.2601626  0.9411765 1.882353    64
[6]  {spiders=1,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2398374  0.9365079 1.873016    59
```

(a) Oversampling 1

```
[30] {albumin=Bajo,
      protime=Bajo}      => {class=1} 0.3495935  0.8600000 1.720000    86
[31] {spiders=1,
      albumin=Bajo}      => {class=1} 0.3170732  0.8571429 1.714286    78
[32] {bilirubin=Alto,
      albumin=Bajo}      => {class=1} 0.2886179  0.8554217 1.710843    71
[33] {liver_big=2,
      bilirubin=Alto,
      albumin=Bajo}      => {class=1} 0.2398374  0.8550725 1.710145    59
[34] {albumin=Bajo,
      histology=2}       => {class=1} 0.3536585  0.8529412 1.705882    87
[35] {liver_big=2,
      bilirubin=Alto,
      histology=2}       => {class=1} 0.2317073  0.8507463 1.701493    57
[36] {age=Adulto Mayor,
      liver_big=2,
      albumin=Bajo}      => {class=1} 0.2276423  0.8484848 1.696970    56
[37] {bilirubin=Alto,
      protime=Bajo}      => {class=1} 0.2723577  0.8481013 1.696203    67
[38] {spiders=1,
      bilirubin=Alto}    => {class=1} 0.2560976  0.8289474 1.657895    63
[39] {liver_big=2,
      histology=1}       => {class=2} 0.2723577  0.8271605 1.654321    67
[40] {age=Adulto Mayor,
      bilirubin=Alto}    => {class=1} 0.2113821  0.8253968 1.650794    52
[41] {bilirubin=Alto,
      histology=2}       => {class=1} 0.2682927  0.8250000 1.650000    66
[42] {age=Adulto Mayor,
      albumin=Bajo}      => {class=1} 0.2642276  0.8227848 1.645570    65
[43] {alk_phosphate=Alto,
      protime=Bajo}      => {class=1} 0.2032520  0.8196721 1.639344    50
[44] {age=Adulto Mayor,
      protime=Bajo}      => {class=1} 0.2398374  0.8194444 1.638889    59
[45] {liver_big=2,
      albumin=Bajo}      => {class=1} 0.3699187  0.8125000 1.625000    91
[46] {liver_big=2,
      protime=Bajo}      => {class=1} 0.3373984  0.8058252 1.611650    83
[47] {liver_big=2,
      spiders=1}         => {class=1} 0.3130081  0.8020833 1.604167    77
[48] {spiders=1,
      histology=2}       => {class=1} 0.3089431  0.8000000 1.600000    76
```

(b) Oversampling 2

Figura 7: Reglas obtenidas por Oversampling, con un soporte 0.2 y ordenadas por nivel de confianza descendente.

- **Reglas obtenidas con Oversamplig**: La regla relevante obtenida al aplicar esta técnica de balanceo fue la siguiente:

$$\{liver-big=2, spiders=1, albumin=Bajo, protime=Bajo\} => \{class=1\} \quad (1)$$

- **Reglas obtenidas con Undersampling**: Regla relevante obtenida luego de utilizar Undersampling fue la siguiente:

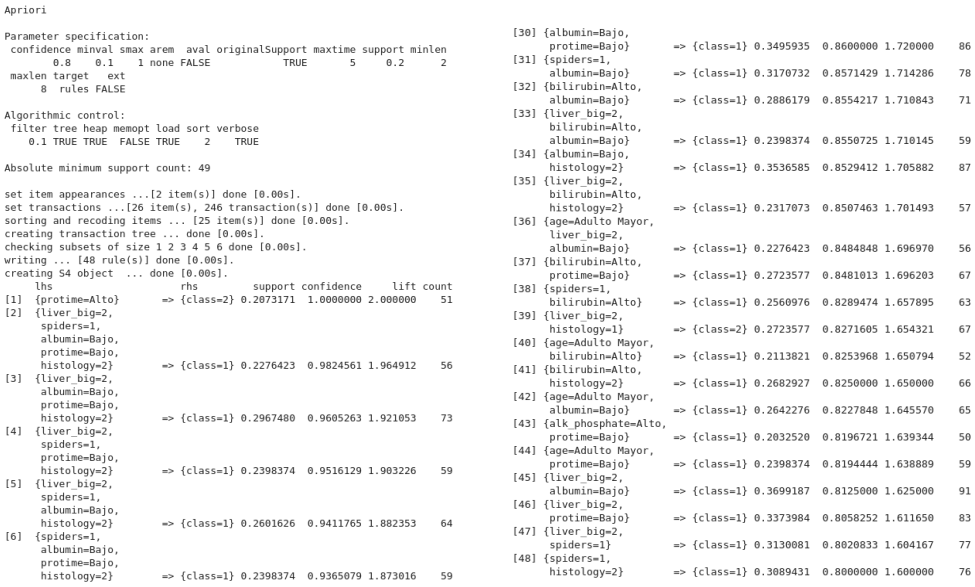
$$\{liver-big=2, bilirubin=Alto, albumin=Bajo\} => \{class=1\} \quad (2)$$

- **Reglas obtenidas con Dual Balance**:

Para este tipo de balance, se obtuvo que las reglas más relaventes, son:

$$\{alk-phosphate=Medio, protime=Bajo\} => \{class=1\} \quad (3)$$

$$\{sgot=Alto, protime=Bajo\} => \{class=1\} \quad (4)$$



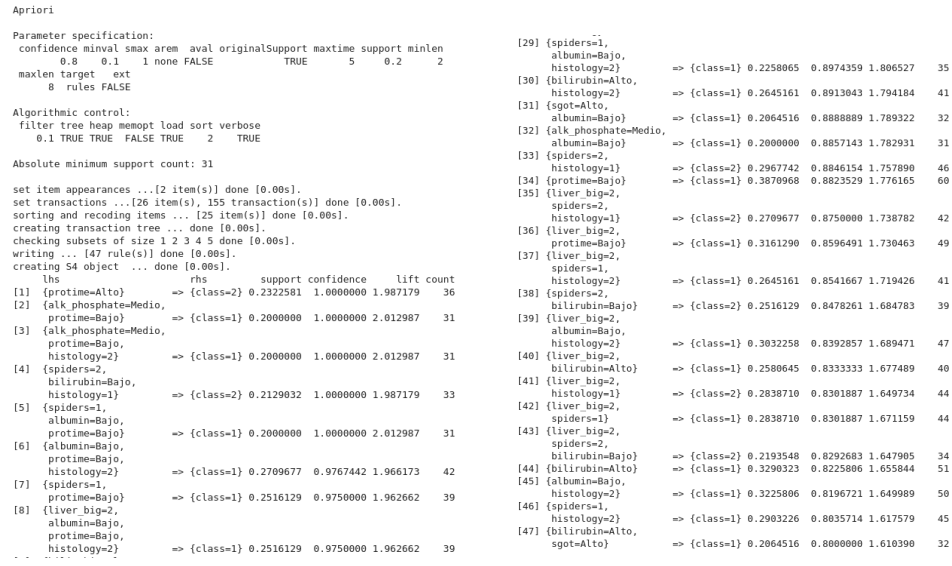
```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5     0.2      2
 maxlen target   ext
      8  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 49

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[26 item(s), 246 transaction(s)] done [0.00s].
sorting and recoding items ... [25 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [48 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs                    rhs         support confidence     lift count
[1]  {protime=Alto}      => {class=2} 0.2073171  1.0000000 2.000000    51
[2]  {liver_big=2,
      spiders=1,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2276423  0.9824561 1.964912    56
[3]  {liver_big=2,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2967480  0.9605263 1.921053    73
[4]  {liver_big=2,
      spiders=1,
      protime=Bajo,
      histology=2}       => {class=1} 0.2398374  0.9516129 1.903226    59
[5]  {liver_big=2,
      spiders=1,
      albumin=Bajo,
      histology=2}       => {class=1} 0.2601626  0.9411765 1.882353    64
[6]  {spiders=1,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2398374  0.9365079 1.873016    59
```

(a) Undersampling 1

```
[30] {albumin=Bajo,
      protime=Bajo}      => {class=1} 0.3495935  0.8600000 1.720000    86
[31] {spiders=1,
      albumin=Bajo}      => {class=1} 0.3170732  0.8571429 1.714286    78
[32] {bilirubin=Alto,
      albumin=Bajo}      => {class=1} 0.2886179  0.8554217 1.710843    71
[33] {liver_big=2,
      bilirubin=Alto,
      albumin=Bajo}      => {class=1} 0.2398374  0.8550725 1.710145    59
[34] {albumin=Bajo,
      histology=2}       => {class=1} 0.3536585  0.8529412 1.705882    87
[35] {liver_big=2,
      bilirubin=Alto,
      histology=2}       => {class=1} 0.2317073  0.8507463 1.701493    57
[36] {age=Adulto Mayor,
      liver_big=2,
      albumin=Bajo}      => {class=1} 0.2276423  0.8484848 1.696970    56
[37] {bilirubin=Alto,
      protime=Bajo}      => {class=1} 0.2723577  0.8481013 1.696203    67
[38] {spiders=1,
      bilirubin=Alto}    => {class=1} 0.2560976  0.8289474 1.657895    63
[39] {liver_big=2,
      histology=1}       => {class=2} 0.2723577  0.8271605 1.654321    67
[40] {age=Adulto Mayor,
      bilirubin=Alto}    => {class=1} 0.2113821  0.8253968 1.650794    52
[41] {bilirubin=Alto,
      histology=2}       => {class=1} 0.2682927  0.8250000 1.650000    66
[42] {age=Adulto Mayor,
      albumin=Bajo}      => {class=1} 0.2642276  0.8227848 1.645570    65
[43] {alk_phosphate=Alto,
      protime=Bajo}      => {class=1} 0.2032520  0.8196721 1.639344    50
[44] {age=Adulto Mayor,
      protime=Bajo}      => {class=1} 0.2398374  0.8194444 1.638889    59
[45] {liver_big=2,
      albumin=Bajo}      => {class=1} 0.3699187  0.8125000 1.625000    91
[46] {liver_big=2,
      protime=Bajo}      => {class=1} 0.3373984  0.8058252 1.611650    83
[47] {liver_big=2,
      spiders=1}         => {class=1} 0.3130081  0.8020833 1.604167    77
[48] {spiders=1,
      histology=2}       => {class=1} 0.3089431  0.8000000 1.600000    76
```

(b) Undersampling 2

Figura 8: Reglas obtenidas por Undersampling, con un soporte 0.2 y ordenadas por nivel de confianza descendente.



```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5     0.2      2
 maxlen target   ext
      8  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 31

set item appearances ...[2 item(s)] done [0.00s].
set transactions ...[26 item(s), 155 transaction(s)] done [0.00s].
sorting and recoding items ... [25 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [47 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
     lhs                    rhs         support confidence     lift count
[1]  {protime=Alto}      => {class=2} 0.2322581  1.0000000 1.987179    36
[2]  {alk_phosphate=Medio,
      protime=Bajo}      => {class=1} 0.2000000  1.0000000 2.012987    31
[3]  {alk_phosphate=Medio,
      protime=Bajo,
      histology=2}       => {class=1} 0.2000000  1.0000000 2.012987    31
[4]  {spiders=2,
      bilirubin=Bajo,
      histology=1}       => {class=2} 0.2129032  1.0000000 1.987179    33
[5]  {spiders=1,
      albumin=Bajo,
      protime=Bajo}      => {class=1} 0.2000000  1.0000000 2.012987    31
[6]  {albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2709677  0.9767442 1.966173    42
[7]  {spiders=1,
      protime=Bajo}      => {class=1} 0.2516129  0.9750000 1.962662    39
[8]  {liver_big=2,
      albumin=Bajo,
      protime=Bajo,
      histology=2}       => {class=1} 0.2516129  0.9750000 1.962662    39
```

(a) Dual Balance 1

```
[29] {spiders=1,
      albumin=Bajo,
      histology=2}       => {class=1} 0.2258065  0.8974359 1.806527    35
[30] {bilirubin=Alto,
      histology=2}       => {class=1} 0.2645161  0.8913043 1.794184    41
[31] {sgot=Alto,
      albumin=Bajo}      => {class=1} 0.2064516  0.8888889 1.789322    32
[32] {alk_phosphate=Medio,
      albumin=Bajo}      => {class=1} 0.2000000  0.8857143 1.782931    31
[33] {spiders=2,
      histology=1}       => {class=2} 0.2967742  0.8846154 1.757890    46
[34] {protime=Bajo}      => {class=1} 0.3870968  0.8823529 1.776165    60
[35] {liver_big=2,
      spiders=2,
      histology=1}       => {class=2} 0.2709677  0.8750000 1.738782    42
[36] {liver_big=2,
      protime=Bajo}      => {class=1} 0.3161290  0.8596491 1.730463    49
[37] {liver_big=2,
      spiders=1,
      histology=2}       => {class=1} 0.2645161  0.8541667 1.719426    41
[38] {spiders=2,
      bilirubin=Bajo}    => {class=2} 0.2516129  0.8478261 1.684783    39
[39] {liver_big=2,
      albumin=Bajo,
      histology=2}       => {class=1} 0.3032258  0.8392857 1.689471    47
[40] {liver_big=2,
      bilirubin=Alto}    => {class=1} 0.2580645  0.8333333 1.677489    40
[41] {liver_big=2,
      histology=1}       => {class=2} 0.2838710  0.8301887 1.649734    44
[42] {liver_big=2,
      spiders=1}         => {class=1} 0.2838710  0.8301887 1.671159    44
[43] {liver_big=2,
      spiders=2,
      bilirubin=Bajo}    => {class=2} 0.2193548  0.8292683 1.647905    34
[44] {bilirubin=Alto}    => {class=1} 0.3290323  0.8225806 1.655844    51
[45] {albumin=Bajo,
      histology=2}       => {class=1} 0.3225806  0.8196721 1.649989    50
[46] {spiders=1,
      histology=2}       => {class=1} 0.2903226  0.8035714 1.617579    45
[47] {bilirubin=Alto,
      sgot=Alto}         => {class=1} 0.2064516  0.8000000 1.610390    32
```

(b) Dual Balance 2

Figura 9: Reglas obtenidas por dual balance, con un soporte 0.2 y ordenadas por nivel de confianza descendente.

En el siguiente capitulo se explica detalladamente las reglas destacadas luego de aplicar los diferentes tipos de balances.

# 4. Análisis de los resultados

Una de las condiciones necesarias para poder trabajar las reglas de asociación, es balance de datos respecto a la clase, correspondiente pacientes fallecidos y sobrevivientes, como se puede ver en la figura (1), existe una proporción aproximada de $1:4$ respectivamente. Si no se realiza el tratamiento previo, vemos que la $class = 1$, no se hace presente dentro de las asociaciones, por lo que no se tendrían reglas que permitan asociar a pacientes fallecidos, como se puede ver en la figura (2). Dado esto, es que se descarta como parte del proceso de análisis.

Se considerará como relevantes las reglas que se presentarán, de acuerdo a niveles de confianza $> 90\,\%$ y con un $lift > 1{,}8$, con soporte entorno al $\sim 0{,}2$.

## 4.1. Oversampling

De acuerdo a las asociaciones presentados anteriormente, bajo este tipo de técnica de balance, se pueden observar reglas que contienen la $class = 1$ y $class = 2$.

Al ordenar las reglas asociativas por nivel de confianza, vemos que [5] es la primera regla con un soporte de $\sim 0{,}207$.

$$\{protime = Alto\} => \{class = 2\} \tag{5}$$

Lo que indica que existe un grupo de paciente relevantes que asocia a su alto nivel de coagulación en la sangre, y que han sobrevivido en la atención. Esto es semejante a lo obtenido en el estudio realizado por clusterización por k-means (1), sin embargo, es contradictorio con la investigaciones en (2) y (3).

La siguiente regla interesante, vemos que aparecen dos atributos que en los estudios anteriores no habían sido considerados, que son $liver - big$ y $spiders$.

$$\{liver - big = 2, spiders = 1, albumin = Bajo, protime = Bajo\} => \{class = 1\} \tag{6}$$

Por una parte, vemos que los atributos $albumin = Bajo$ y $protime = Bajo$ se relaciona con lo evidenciado en las distribuciones individuales en las figuras (3) y (4), y con los pacientes fallecidos, pero no así con la literatura, para el caso del *protime*. Sin embargo, bajos niveles de *albumin*, indican una cierta gravedad considerable de cirrosis, lo que implica un estado

muy avanzado de deterioro del hígado e incluso podría estar relacionado con otras patologías de riesgo que involucran los riñones (4).

También se puede recalcar la presencia de la $histology = 2$, como parte de las reglas frecuentes y que tengan como consecuencia la muerte del paciente. Recordar que este atributo, corresponde a si el paciente se ha realizado algún estudio histológico del tejido hepático.

Se logra observar además, que la regla presentada en [7], la *bilirubin* se hace presente para niveles altos, lo que también concuerda de alguna manera con lo evidenciado en los laboratorios anteriores y con la distribución en la figura (3).

$$\{liver - big = 2, bilirubin = Alto, albumin = Bajo, histology = 2\} => \{class = 1\} \quad (7)$$

Ahora bien, $liver - big$ en ambas reglas [6] y [7] indican que tiene un tamaño grande. En la literatura esto no es necesariamente una consecuencia de alguna enfermedad, de hecho existe una relación natural respecto a la edad del paciente y/o proporcionalidad física. Sin embargo, existen tendencias que en enfermedades; como la hepatitis, cirrosis, hígado graso, atresia, entre otros (5), el hígado comprende un tamaño mayor a la media de los pacientes en similares condiciones.

Por otra parte, para las reglas en la que se obtenga como consecuencia $class = 2$, vemos que se tienen pocas opciones del conjunto obtenido, y se encuentran bajo la condición de confianza $> 0{,}9$, aún así, se puede notar que la $histology = 1$ es parte de varias reglas que implican la supervivencia del paciente.

## 4.2. Undersampling

Para este tipo de balance, se puede ver que hay ciertas variaciones que son de interés respecto a las reglas generadas con el Oversampling. Por ejemplo en [8] vemos que *albumin* aparece como la consecuencia del paciente como sobreviviente para valores altos, que de cierta forma, podría ser algo intuitivo de acuerdo a literatura, donde un paciente que no ha tenido estudios previos, de alguna manera, no ha sido expuesto a situaciones de gravedad que requieran una extracción del tejido.

$$\{spiders = 2, albumin = Alto\} => \{class = 2\} \quad (8)$$

Luego, si se observa la regla en [9], notamos que tenemos una regla similar en [7], por lo que ya se comienzan a obtener ciertos hallazgos de atributos que pudiesen ser relevantes para una eventual predicción.

$$\{liver - big = 2, bilirubin = Alto, albumin = Bajo\} => \{class = 1\} \tag{9}$$

También se observa, al igual que en balance por Oversampling, que la $histology = 2$ es fuertemente frecuente en las reglas, y que tengan como consecuente la muerte del paciente, con niveles de confianza superiores a 0,9. Otras reglas interesantes, es que se repite $\{protime = Alto\}$, como en [5], y además aparecen los níveles bajo de *bilirubin*, como en [10], que de cierta forma, se espera que exista la asociación inversa a lo expuesto anteriormente.

$$\{bilirubin = Bajo, histology = 1\} => \{class = 2\} \tag{10}$$

## 4.3. Dual Balance

Para este tipo de balance, se puede observar otros atributos que entran en consideración para generar asociaciones, como lo son $alk - phosphate$ y *sgot*. Por ejemplo en [11], vemos que el $alk - phosphate$ es de niveles medio, donde según diferentes estudios, este índice no necesariamente se puede atribuir a un daño hepático, pero si puede ir acompañado, cuando existe hepatitis de tipo C en los pacientes, por lo que su valor tiende a bajar. Ahora bien, este valor puede estar influenciado por otras patologías respecto a una enfermedad (o incluso cancer) del sistea linfático, problemas cardíacos, entre otros (6), por lo que no se puede contrastar con el dataset del trabajo, ya que carece de estos elementos de análisis.

$$\{alk - phosphate = Medio, protime = Bajo\} => \{class = 1\} \tag{11}$$

Respecto al *sgot*, notamos que existe una regla [12], cuyo valor indica que para valores alto, podría implicar en la muerte del paciente, en conjunto con $protime = Bajo$.

$$\{sgot = Alto, protime = Bajo\} => \{class = 1\} \tag{12}$$

En diversos documentos respecto a este atributo, mencionan que es bastante frecuente encontrar pacientes con hepatitis altos níveles de aspartato (*sgot*), pero no es concluyente

de acuerdo a (7). Aún así, cuando se tienen niveles sobre los 50 unidades por litro de suero, es un indicador de preocupación por posible daño hepático (frecuente para hepatitis A y B).

Es relevante indicar que bajo los 3 métodos de balance, se pudo obtener que hay ciertos atributos que 'afloran' de un modo más pronunciado que en otros, o bien con mayor calidad de la asociación, como por ejemplo bajo Oversampling, tenemos que el tamaño del hígado es bastante frecuente en las reglas, en Undersampling la albumina, y para el dual, se tiene que la fosfatasa alcalina, toma un protagonismo relevante en las asociaciones. También notamos, que la histología es un atributo presente en muchas reglas, independiente del tipo de balance realizado, por lo que se puede asociar claramente a que el paciente, ya había tenido algún tipo de condición de cierta gravedad, respecto al hígado.
Para *protime*, sigue siendo inconsecuente con la literatura en los 3 procesos, pero se obtienen resultados similares a los laboratorios anteriores.
En general la reglas obtenidas, son bastante generales en su asociación, en cuanto a que se obtienen entre 3 a 4 atributos, que podrían explicar apriori, si el paciente puede fallecer o sobrevivir. Por una parte podría ser bastante rápido para discriminar si un paciente tiene riesgo alto de fallecer, como por ejemplo cuestionar si ha tenido anteriormente alguna intervención de muestra de tejido hepático, o un chequeo rápido de 1 o dos condiciones bioquímicas, que permitan evaluar la gravedad del paciente de manera ágil, en caso no haber tenido anteriormente algún síntoma.

# 5. Conclusión

En el presente laboratorio, se ha realizado un estudio de reglas de asociación, en donde se presentó que el dataset es desbalanceado respecto a su clase de predicción (paciente vivo o fallecido), por lo que para poder efectuar una correcta generación de conjunto de reglas, se tuvo que balancear en diferentes estrategias como; Oversampling, Undersampling y Dual Balance.

Se probó que sin balance de clase, las reglas de asociación fueron insatisfactorias, por lo que se recurre a balancear con los métodos señalados, obteniéndose reglas con atributos presentes en un método pero en otros no, como lo fue $liver - big$, $albumin$ y $alk - phosphate$, lo que se pudo explorar más reglas que asociación la clase de interés. En general las reglas tenían un común denominador, que es la $histology = 2$, así como también el $protime = Alto$. Las asociaciones obtenidas fueron relativamente satisfactoria, logrando evidenciar reglas que se relacionan con la literatura, como por ejemplo los niveles alto de bilirubina, en conjunto con nivel bajo de albumina, podría indicar un riesgo de daño hepático severo.

También se logra encontrar reglas que vinculan el tamaño del hígado, algo que en anteriores laboratorios no se logra evidenciar, y que de acuerdo con la investigación realizada, sí podría tener cierta relevancia frente a un daño importante del órgano.

De todo lo anterior, se concluye que mediante las reglas obtenidas, no es trivial concluir mediante 2, 3 o 4 atributos para indicar si el paciente tiene alto riesgo de fallecer, pero si podría contribuir en una toma rápida toma de decisión, en caso de que el paciente ya tenga alguna prexistencia, dado que la histología es un atributo, fuertemente presente en las asociaciones encontradas, con un nivel de confianza $> 0{,}9$ y un lift $> 1{,}8$.

# Bibliografía


[1] G. Aguilar and S. Hernández, *Laboratorio 2 - Clustering. https://github.com/naotoam/magister/blob/master/Inteligencia Computacional/Informe 2 - Laboratorio.pdf*. 2021.

[2] C. Y. K. S. K. D. e. a. Lee M, Kim W, *Spontaneous Evolution in Bilirubin Levels Predicts LiverRelated Mortality in Patients with Alcoholic Hepatitis*. 2014.

[3] N. S. E.-F. Myron J. Tong and M. I. Grew, *Clinical Manifestations of Hepatitis A: Recent Experience in a Community Teaching Hospital*. 1995.

[4] U. D. of Veteran Affairs, "Viral hepatitis and liver disease. https://www.hepatitis.va.gov/hcv/patient/diagnosis/labtests-albumin.asp," 2006.

[5] I. Franks, "What does liver size say about my health https://www.healthline.com/health/normal-liver-size," 2020.

[6] Q. Cheng, *Alkaline Phosphatase (ALP) https://labtestsonline.org/tests/alkaline-phosphatase-alp*. 2020.

[7] A. Gotter, "Sgot test. https://www.healthline.com/health/sgot-test," 2018.