

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



Laboratorio 3 - Reglas de Asociacion

Integrantes: Felipe Cornejo
Bastían Loyola
Curso: Análisis de Datos
Sección A-1
Profesor: Max Chacón Pacheco

30 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. Medidas de Calidad y Confianza	2
2.2.1. Soporte	2
2.2.2. Confianza	3
2.2.3. <i>Lift</i>	3
2.3. Monotonicidad	3
2.4. Propiedades de las Medidas	4
2.5. Algoritmo Apriori	4
2.6. Regla de Sturges	4
3. Obtención de reglas	6
3.1. Reglas de asociacion	7
4. Análisis de resultados y comparación	10
4.1. Reglas sin restricción	10
4.2. Reglas con consecuente 'y'='yes'	11
4.3. Reglas con consecuente 'y'='no'	12
4.4. Comparación	13
5. Conclusión	15
Bibliografía	17

1. Introducción

Tal como en la experiencia anterior donde se conoció acerca de la agrupación de nubes de datos, por medio de la técnica de *clustering* sobre una base de datos de una campaña de telemarketing realizado por un banco de Portugal, proporcionada por Moro et al. (2014), donde se logra identificar los distintos grupos y características de los individuos que llegan a aceptar el crédito propuesto por el banco, la cantidad óptima de grupos los cuales deberían de haber. Esta nueva experiencia ahondará en las reglas de asociación relacionadas con la misma base de datos, con el objetivo principal de analizar los resultados y compararlos con la experiencia anterior, utilizando la herramienta *arulesViz*, un paquete de R, así como menciona en la documentación de *RPubs*, se hace interesante esta técnica a la hora de descubrir patrones de objetos o atributos en la base de datos, permitiendo así establecer relaciones entre variables (Gil, 2020). Este documento presenta a primera mano las definiciones importantes a utilizar a lo largo del desarrollo, seguido por la obtención de los resultados, los cuales en este caso serán las reglas de asociación junto con sus características, el análisis ya mencionado haciendo alusión a la comparación con la experiencia anterior y finalmente un resumen y observación del cumplimiento de los objetivos y sus obstáculos.

1.1. Objetivos

1. Conocer acerca las reglas de asociación implícitas en los datos.
2. Implementar las reglas de asociación en el caso propuesto.
3. Extraer medidas de calidad para el caso.
4. Analizar los resultados y compararlos con las experiencias anteriores.
5. Comparar y corroborar resultados con literatura encontrada.

2. Marco Teórico

A continuación se expondrán los distintos conceptos relevantes para el entendimiento de este documento para la futura utilización de estos y así llegar a una mejor comunicación con el lector.

2.1. Reglas de Asociación

Las reglas de Asociación, tal como el nombre lo indica, relacionan una determinada conclusión (por ejemplo, la compra de un producto dado) con un conjunto de condiciones (IBM, 2021). Siguiendo por una definición lógica de estas, guiado por Amat (2018), se define una regla de asociación como una implicación del tipo “si X entonces Y” ($X \Rightarrow Y$), donde X e Y son *itemsets* o *items* individuales. El lado izquierdo de la regla recibe el nombre de antecedente o *left-hand-side* (LHS) y el lado derecho el nombre de consecuente o *right-hand-side* (RHS). Por ejemplo, la regla $A, B \Rightarrow C$ significa que, cuando ocurren A y B, también ocurre C.

Además existen varios algoritmos diseñados para identificar *items* frecuentes y reglas de asociación que se definirán más adelante.

2.2. Medidas de Calidad y Confianza

Las medidas de calidad son ratios que indicarán la probabilidad en que estas transacciones ocurran en casos reales. Las medidas a utilizar serán las siguientes:

2.2.1. Soporte

El soporte de un *item* es el número de transacciones (u observaciones) que contienen a dicho *item* sobre el total de transacciones.

$$\text{Soporte}(X \Rightarrow Y) = \text{Soporte}(XUY) = \frac{\text{count}(XUY)}{N} \quad (1)$$

Donde N es la cantidad total de observaciones y $\text{count}(XUY)$ es el número de transacciones que contienen todos los *items* X e Y en la ecuación 1 (Gil, 2020). Este valor, como se puede ver, puede tomar valores entre 0 y 1.

2.2.2. Confianza

Con lo anterior conocido, la confianza se puede definir como la probabilidad que una relación que contenga un *item* X, tambien contenga un *item* Y, este concepto puede ser interpretado de manera matemática como se ve a continuación.

$$Confianza(X \Rightarrow Y) = \frac{Soporte(XUY)}{Soporte(X)} \quad (2)$$

Como se puede ver en la ecuación 2 tiene mucha similitud con la probabilidad condicional de Y dado X, es decir $P(Y|X)$, por ello considera valores entre 0 y 1.

2.2.3. Lift

Lift es un estadístico el cual indica que tanto se acerca la transacción a un patrón real según la situación estudiada o si está marcado por una situación azarosa.

$$Lift = \frac{Soporte(XUY)}{Soporte(X) * Soporte(Y)} \quad (3)$$

El cual un valor cercano a 1 indicaría que los *items* de la transacción son totalmente independiente entre sí. Al contrario un resultado alejado de 1 indicará que los *itemsets* conforman una regla que representa un patrón real. Si el valor es mayor a 1 indica que están relacionados positivamente y por el contrario indicaría que están correlacionados negativamente (Gil, 2020).

2.3. Monotonidad

El principio de Monotonidad permite descartar *items* que no se presentan en las observaciones o transacciones, para así evitar crear reglas con ellos y sobrecargar el procesamiento del análisis. Este principio se puede dividir en dos aseveraciones como menciona Buitrago (2021):

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

2.4. Propiedades de las Medidas

Para hallar reglas de asociación, es necesario establecer mínimos, lo cual normalmente es indicado por el usuario, en niveles de confianza y soporte. Por esto mismo, es pertinente que se cumplan ambos niveles mínimos para que la regla sea interesante.

2.5. Algoritmo Apriori

El Algoritmo Apriori, es una serie de pasos para la búsqueda de reglas de asociación la cual se divide en dos etapas:

1. Identificar todos los *items* por encima de una frecuencia mínima dada.
2. Convertir los *itemsets* frecuentes en reglas de asociación.

La primera parte de este algoritmo es obstaculizado por la cantidad de combinaciones posibles que se podrían generar, sin embargo, una vez identificados se obtiene directamente al aplicar la ecuación 2. Este método abraza la afirmación de "Si un *itemset* no es frecuente, entonces cualquier conjunto que contenga a este itemset tampoco lo será." Por ello, reduce la cantidad de observaciones a revisar.

Luego de obtener los *itemsets* **frecuentes**, se procede a generar otra gran cantidad de combinaciones utilizando los *items* encontrados anteriormente contra *subsets* que serán cualquier combinación de elementos del conjunto siguiendo la siguiente regla:

$$S \Rightarrow (I - S) \quad (4)$$

Donde S, son los *Subsets* generados e I, los *Items* frecuentes.

Finalmente se obtienen todas las reglas que superen el mínimo de confianza otorgado para el análisis.

2.6. Regla de Sturges

La regla de Sturges es una regla la cual, por medio de una expresión matemática se obtiene el número de clases que debe considerarse para la elaboración de un histograma

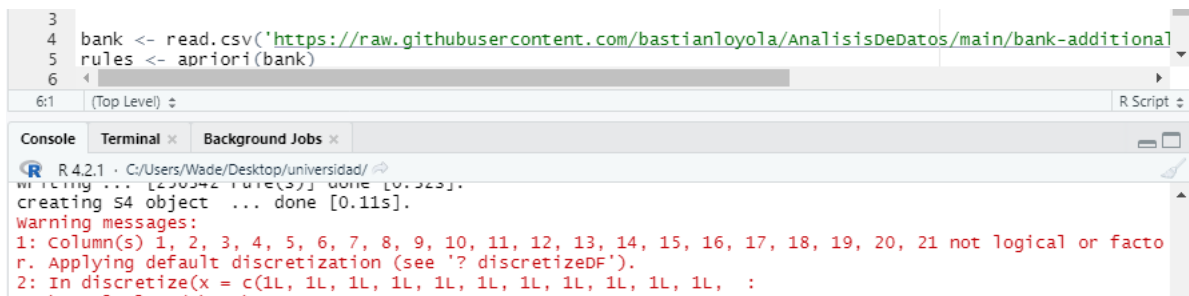
Hyndman (1995). La cual a su vez sirve para poder discretizar variables numéricas. Este número resultante es definido de la siguiente manera:

$$k = 1 + \log_2 n \tag{5}$$

Donde k es el número de clases ya mencionado.

3. Obtención de reglas

En la siguiente sección se hablará sobre la obtención de reglas de asociación para la base de datos que se ha trabajado en la línea de trabajo, dicha base de datos tiene datos acerca de una campaña de telemarketing de un banco de Portugal, la cual consta con 21 variables y 41188 observaciones. Para la obtención de reglas se hará uso de lenguaje de programación R y el método 'apriori' de la librería 'arulesViz', inicialmente se eliminaron datos duplicados similar a las otras experiencias dejándonos con 41176 observaciones, a continuación se utilizó el método 'apriori' en los datos; sin embargo, genero el error que se puede observar en la Figura 1, ya este método solo funciona sobre variables de tipo lógicas o factor. En consideración de



```
3
4 bank <- read.csv('https://raw.githubusercontent.com/bastianloyola/AnalisisDeDatos/main/bank-additional')
5 rules <- apriori(bank)
6

6:1 (Top Level)
R Script

Console Terminal Background Jobs
R 4.2.1 · C:/Users/Wade/Desktop/universidad/
writing ... [230.7421016(5)] done [0.923].
creating 54 object ... done [0.11s].
warning messages:
1: Column(s) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 not logical or factor. Applying default discretization (see '? discretizeDF').
2: In discretize(x = c(1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, :
```

Figura 1: Error en 'apriori' sin discretización

lo anterior se deberá llevar las variables a factor, por el lado de las variables categóricas se deberá aplicar directamente la función 'factor()'; sin embargo, para las variables numéricas se deberá realizar un proceso para llevar los valores continuos a valores discretos y luego poder factorizarlos, dicho proceso recibe el nombre de discretización, en donde se deberán establecer intervalos de valores que hagan referencia a un valor categórico, el proceso utilizado se llama 'binning', que nos permite generar 'k' intervalos de ancho igual a la diferencia entre el valor máximo y mínimo de dicha variable numérica dividido en la cantidad de intervalos, la cantidad de intervalos está definido por la ecuación 5 llamada regla de Sturges según Hyndman (1995), donde n es 41176 y dando como resultado aproximado que $k = 16$.

Con toda la información previamente mencionada se realizó las debidas transformaciones y procesos obteniendo datos de la Figura 2

Una vez ya teniendo los tipos de datos correctos para lograr aplicar la función 'apriori', esta función permite establecer valores mínimos de confianza y soporte, ya que


```

> str(bank_unique)
'data.frame':  41176 obs. of  21 variables:
 $ age      : Factor w/ 16 levels "(17,22.1]", "(22.1,27.1]",...: 8 8 4 5 8 6 9 5 2 2 ...
 $ job      : Factor w/ 12 levels "admin.", "blue-collar",...: 4 8 8 1 8 8 1 2 10 8 ...
 $ marital  : Factor w/ 4 levels "divorced", "married",...: 2 2 2 2 2 2 2 2 3 3 ...
 $ education : Factor w/ 8 levels "basic.4y", "basic.6y",...: 1 4 4 2 4 3 6 8 6 4 ...
 $ default  : Factor w/ 3 levels "no", "unknown",...: 1 2 1 1 1 2 1 2 1 1 ...
 $ housing  : Factor w/ 3 levels "no", "unknown",...: 1 1 3 1 1 1 1 1 3 3 ...
 $ loan     : Factor w/ 3 levels "no", "unknown",...: 1 1 1 1 3 1 1 1 1 1 ...
 $ contact  : Factor w/ 2 levels "cellular", "telephone": 2 2 2 2 2 2 2 2 2 2 ...
 $ month    : Factor w/ 10 levels "apr", "aug", "dec",...: 7 7 7 7 7 7 7 7 7 7 ...
 $ day_of_week : Factor w/ 5 levels "fri", "mon", "thu",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ duration : Factor w/ 16 levels "(0,307]", "(307,615]",...: 1 1 1 1 1 1 1 1 2 1 ...
 $ campaign : Factor w/ 16 levels "(1,4.44]", "(4.44,7.88]",...: NA NA NA NA NA NA NA NA ...
 $ pdays   : Factor w/ 16 levels "(0,62.4]", "(62.4,125]",...: 16 16 16 16 16 16 16 16 16 16 ...
 $ previous : Factor w/ 16 levels "(0,0.438]", "(0.438,0.875]",...: NA NA NA NA NA NA NA NA ...
 $ poutcome : Factor w/ 3 levels "failure", "nonexistent",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ emp.var.rate : Factor w/ 16 levels "(-3.4,-3.1]",...: 15 15 15 15 15 15 15 15 15 15 ...
 $ cons.price.idx : Factor w/ 16 levels "(92.2,92.4]",...: 12 12 12 12 12 12 12 12 12 12 ...
 $ cons.conf.idx : Factor w/ 16 levels "(-50.8,-49.3]",...: 10 10 10 10 10 10 10 10 10 10 ...
 $ euribor3m    : Factor w/ 16 levels "(0.634,0.91]",...: 16 16 16 16 16 16 16 16 16 16 ...
 $ nr.employed  : Factor w/ 16 levels "(4.96e+03,4.98e+03]",...: 14 14 14 14 14 14 14 14 14 14 ...
 $ y            : Factor w/ 2 levels "no", "yes": 1 1 1 1 1 1 1 1 1 1 ...

```

Figura 2: Resultados de la discretización

esta función no generará todas las posibles reglas, dichos valores por defecto son 0.8 y 0.1 respectivamente; sin embargo, utilizar los parámetros predeterminados no siempre serán la mejor opción, otro parámetro es 'minlen' que nos permite definir la cantidad mínima de antecedentes y consecuentes, con su valor predeterminado se pueden generar reglas que no tengan ningún antecedente, generando reglas que no proporcionen información relevante. Por otro lado, se deberá definir los valores mínimos de confianza y soporte, como menciona Fournier-Viger et al. (2012) el nivel de soporte depende de las características de la base de datos, por lo que se utilizará el valor predeterminado de la función 'apriori', a menos que dicho valor sea mayor al máximo, en dicho caso se reducirá el valor hasta que sea menor al máximo; Por el lado del nivel de confianza se espera destacar aquellas reglas que tengan el mayor valor, ya que nos aseguran la probabilidad con que una relación contiene a los antecedentes y consecuentes a la vez, según Fournier-Viger et al. (2012) el valor mínimo de confianza es más fácil de establecer por lo que se utilizará el valor predeterminado.

3.1. Reglas de asociación

Con todos los parámetros ya establecidos se obtuvieron reglas de asociación, las cuales ordenamos de mayor a menor acorde a tres valores: lift, soporte y confianza; la medida de lift fue principalmente elegida porque esta describe su cercanía a un patrón real, además

es una variable bastante utilizada y que logra dar información relevante para evaluar reglas tal como dice Lin et al. (2002). En la Figura 3 se muestran las 3 principales reglas para cada orden.

```
> inspect(head(sort(rules, by='supp'),3))
  lhs      rhs      support  confidence coverage  lift    count
[1] {y=no}    => {pdays=(937,999]} 0.8740286 0.9850015 0.8873373 1.022627 35989
[2] {pdays=(937,999]} => {y=no} 0.8740286 0.9074153 0.9632067 1.022627 35989
[3] {poutcome=nonexistent} => {pdays=(937,999]} 0.8633913 1.0000000 0.8633913 1.038199 35551
> inspect(head(sort(rules, by='conf'),3))
  lhs      rhs      support  confidence coverage  lift    count
[1] {nr.employed=(5.18e+03,5.2e+03]} => {emp.var.rate=(0.8,1.1]} 0.1885079 1 0.1885079 5.304818 7762
[2] {emp.var.rate=(0.8,1.1]} => {nr.employed=(5.18e+03,5.2e+03]} 0.1885079 1 0.1885079 5.304818 7762
[3] {nr.employed=(5.18e+03,5.2e+03]} => {cons.price.idx=(94.4,94.1]} 0.1885079 1 0.1885079 5.006809 7762
> inspect(head(sort(rules, by='lift'),3))
  lhs      rhs      support  confidence coverage  lift    count
[1] {emp.var.rate=(1.1,1.4],
  cons.conf.idx=(-41.8,-40.3]} => {cons.price.idx=(94.4,94.6]} 0.1062269 1 0.1062269 8.994321 4374
[2] {cons.conf.idx=(-41.8,-40.3],
  nr.employed=(5.21e+03,5.23e+03]} => {cons.price.idx=(94.4,94.6]} 0.1062269 1 0.1062269 8.994321 4374
[3] {month=jun,
  emp.var.rate=(1.1,1.4]} => {cons.price.idx=(94.4,94.6]} 0.1062269 1 0.1062269 8.994321 4374
```

Figura 3: Reglas con los mayores soporte, confianza y lift respectivamente

Además de parámetros como confianza y soportes mínimos que evitan una gran combinatoria en las reglas, también es posible restringirlas acorde a las variables en el antecedente o en el consecuente, unas reglas destacables que son interesantes de analizar son aquellas que contengan como consecuente los dos valores de la clase 'y' para determinar si existe un patrón o existen reglas con buena confianza para intentar predecir 'y' con ciertas variables, es por ello que se realizó nuevamente el método 'apriori' para encontrar las reglas mencionadas; se utilizarán los mismos parámetros para ordenarlos de mayor a menor y lograr obtener las primeras 3 reglas acorde a lift, soporte y confianza.

Para el caso donde 'y' es yes, se obtuvo una confianza máxima de 0.006314358, por lo tanto, utilizar el parámetro predeterminado de 0.1 para el soporte mínimo es infactible, ya que no entregará ninguna regla, por lo tanto, se redujo el soporte mínimo a 0.001, en el caso de 'y' = 'no' se puede utilizar el valor predeterminado.

Las reglas obtenidas que tienen como consecuente los valores de la clase 'yes' y 'no' son respectivamente las Figuras 4 y 5.

```

> inspect(head(sort(rulesyes, by='lift'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {month=jul,                        => {y=yes} 0.001044298  0.9772727 0.001068584 8.674322   43
     duration=(307,615],
     pdays=(0,62.4]}
[2] {default=no,                      => {y=yes} 0.001044298  0.9772727 0.001068584 8.674322   43
     month=jul,
     duration=(307,615],
     pdays=(0,62.4]}
[3] {duration=(307,615],
     previous=(1.75,2.19],
     poutcome=success,
     emp.var.rate=(-1.9,-1.6],
     euribor3m=(0.634,0.91]} => {y=yes} 0.001020012  0.9767442 0.001044298 8.669631   42
> inspect(head(sort(rulesyes, by='supp'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {loan=no,                         => {y=yes} 0.006314358  0.8074534 0.007820089 7.166998   260
     contact=cellular,
     poutcome=success,
     emp.var.rate=(-1.9,-1.6],
     euribor3m=(0.634,0.91]}
[2] {duration=(307,615],
     pdays=(0,62.4],
     euribor3m=(0.634,0.91]} => {y=yes} 0.005707208  0.8545455 0.006678648 7.584989   235
[3] {default=no,
     duration=(307,615],
     pdays=(0,62.4],
     euribor3m=(0.634,0.91]} => {y=yes} 0.005561492  0.8576779 0.006484360 7.612793   229
> inspect(head(sort(rulesyes, by='conf'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {month=jul,                        => {y=yes} 0.001044298  0.9772727 0.001068584 8.674322   43
     duration=(307,615],
     pdays=(0,62.4]}
[2] {default=no,                      => {y=yes} 0.001044298  0.9772727 0.001068584 8.674322   43
     month=jul,
     duration=(307,615],
     pdays=(0,62.4]}
[3] {duration=(307,615],
     previous=(1.75,2.19],
     poutcome=success,
     emp.var.rate=(-1.9,-1.6],
     euribor3m=(0.634,0.91]} => {y=yes} 0.001020012  0.9767442 0.001044298 8.669631   42
> |

```

Figura 4: Reglas con los mayores lift, soporte y confianza con consecuente 'y'='yes'

```

> inspect(head(sort(rulesno, by='lift'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {duration=(0,307], nr.employed=(5.18e+03,5.2e+03]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
[2] {duration=(0,307], emp.var.rate=(0.8,1.1]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
[3] {duration=(0,307], emp.var.rate=(0.8,1.1], nr.employed=(5.18e+03,5.2e+03]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
> inspect(head(sort(rulesno, by='supp'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {pdays=(937,999]} => {y=no} 0.8740286  0.9074153 0.9632067 1.022627 35989
[2] {poutcome=nonexistent} => {y=no} 0.7871333  0.9116762 0.8633913 1.027429 32411
[3] {pdays=(937,999], poutcome=nonexistent} => {y=no} 0.7871333  0.9116762 0.8633913 1.027429 32411
> inspect(head(sort(rulesno, by='conf'),3))
  lhs                                rhs      support confidence   coverage   lift count
[1] {duration=(0,307], nr.employed=(5.18e+03,5.2e+03]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
[2] {duration=(0,307], emp.var.rate=(0.8,1.1]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
[3] {duration=(0,307], emp.var.rate=(0.8,1.1], nr.employed=(5.18e+03,5.2e+03]} => {y=no} 0.1372887  0.9996463 0.1373373 1.126569 5653
> |

```

Figura 5: Reglas con los mayores lift, soporte y confianza con consecuente 'y'='no'

4. Análisis de resultados y comparación

A continuación se analizarán las distintas reglas extraídas previamente, siendo las reglas sin restricción en el consecuente y las reglas que contengan en el consecuente a la clase 'y'; además de analizar sus variables de calidad

4.1. Reglas sin restricción

En la Figura 3 se puede observar en el primer apartado que las reglas con mayor soporte son aquellas que contienen al último intervalo definido para la variable 'pdays', tanto en antecedente como en consecuente, esto debido a que dicha variable en la base de datos tiene una frecuencia muy grande en los valores de dicho intervalo, si bien corresponde a un gran porcentaje de las reglas no se puede destacar nada significativo. Con respecto a las reglas ordenadas por valor de confianza, se puede apreciar que dicho valor es 1 por lo que el 100 % de las veces que se contenga al antecedente se contendrá al consecuente, dichos elementos son variables sociales, principalmente como es nr.employed en el intervalo $[5.18e+03, 5.2e+03]$ y emp.var.rate en el intervalo de valores $[0.8, 1.1]$, dicha regla tienen sentido, ya que entre esas dos variables existe una relación estrecha. Como ultima variable de calidad considerada se tiene el lift que nos indicará que tan cercano a un patrón real es cierta regla, las reglas con mayor valor de lift son aquellas que tienen como consecuente a cons.price.idx con el valor $[94.4, 94.6]$ y como antecedente se tiene emp.var.rate, cons.conf.idx y month en los valores $[1.1, 1.4]$, $[-41.8, -40.3]$ y jun respectivamente, estas reglas indicarán que en la existencia de los antecedentes en dichos valores, el incremento del costo de vida y canasta básica está en el intervalo $[94.4, 94.6]$, si estas reglas tienen un lift de 8.994321 indicará que se asemeja a un patrón real, con esto se puede deducir que el mes del año, la tasa de empleo y la confianza del consumir afectan en el valor del costo de vida y la canasta básica. Esto se puede confirmar en la bibliografía con respecto al mes (fecha) y el precio de una canasta básica tienen una relación porque en cada mes se modifica la composición alimentos, tal como se menciona en Bejarano-Roncancio and Rivera-Torres (2014), desde vegetales que no se pueden consumir en determinada época o incremento en ciertos productos por festividades o celebraciones. Sobre la tasa de empleo y la calidad de vida, es bien sabido que estos dos tienen relación

en la vida, son indicadores económicos bastante utilizados. La tasa de empleo indicará que tantas personas tienen empleo, por ende tienen recursos económicos para comprar productos, a mayor demanda los precios de la canasta básica incrementarán (González Aguilar, 2020).

4.2. Reglas con consecuente 'y'='yes'

Para estas reglas en particular, como fue mencionado, fue necesario disminuir el valor mínimo de soporte, esto dado que no era posible encontrar reglas con un soporte más grande, esto debe haber sucedido principalmente por la poca frecuencia de registros en la base de datos que indicaban que un cliente se suscribió a un depósito a largo plazo, otro elemento a destacar frente a los otros conjuntos de reglas generados es que estas reglas tienen una mayor cantidad de elementos en los antecedentes, esto nos puede brindar un mayor grado de detalle sobre qué elementos deben estar para que la clase 'y' sea 'yes'; sin embargo, estas reglas mantienen un bajo nivel de soporte indicando que muy poco porcentaje de las transacciones total contienen a los ítems de la regla, siendo en específico el mayor de todos un soporte de 0.006314 (0.6 % aprox.), siendo poco representativo con respecto al total de valores. Con respecto a las reglas ordenadas por valor de confianza y el lift se hablarán a la vez, ya que en ambos órdenes las tres primeras reglas son las mismas, primero cabe destacar que a pesar de ser reglas con bajo valor de soporte, la cantidad de veces que aparecen te brindan una gran confianza de que se cumplirá y además de que se asemejarán mucho a un patrón real, esto debe ser principalmente por lo específicas que son las reglas al incluir tantos antecedentes y un consecuente con poca frecuencia en la base de datos. Con respecto a comparación a bibliografía, se encontró un caso de estudio sobre la misma base de datos extrayendo reglas de asociación (Abdulazeez, 2019), en dicho en la cual se obtuvieron valores de soporte, confianza y lift menores a los vistos en la Figura 4, esto puede ser principalmente por el valor 'k' utilizado en el proceso de discretización, este afecta en los valores del antecedente de la regla, siendo más específicos a un mayor k, como se mencionó en este documento se utilizó para las reglas en la Figura 4 'k' era igual a 16; sin embargo, en Abdulazeez (2019) se utilizó un 'k' igual a 3, siendo menor y por ende generando intervalos de valores menos específicos. Por último, se debe destacar los principales elementos en los antecedentes, siendo los siguientes:

1. La tasa euribor está índica la tasa de interés de préstamos en banco europeos, que se debe tener un alto impacto en la decisión de subscribirse a un credito en personas interesadas.
2. El valor 'duration' se encuentra en el intervalo $]307,615]$ lo que corresponde a una llamada entre 5 y 10 minutos, si bien esta variable no sirve para un modelo predictorio, sé puede establecer que si una persona está en una llamada de telemarketing superior de 5 minutos con un 97 % de confianza esa persona se subscribirá
3. El valor de 'pdays' esta variable se encuentra en el intervalo $]0,62.4]$, lo que significa que a personas contactadas previamente en menos de 62 días y que cumplan con los otros antecedentes, existirá un 97 % de confianza aproximado que la persona se subscriba al crédito
4. La Tasa de empleo tiene un valor negativo, lo que puede identificar que cuando hay más desempleo o despido y cumpliendo los otros antecedentes, también existirá 'y' con el valor 'yes'.

4.3. Reglas con consecuente 'y'='no'

Hablando sobre el conjunto de reglas de la Figura 5 se tienen los órdenes decrecientes por las tres medidas estipuladas: lift, soporte y confianza. Con respecto al lift de las reglas, se puede apreciar que incluso el mayor siendo este 1.126569 es bastante cercano a uno, por lo que se esperaría que las reglas que tengan como consecuente a 'y=no', no se asemejen a ningún patrón real. Por otro lado, se tiene el orden por soporte, se tienen valores altos, por lo que se esperaría que dichas reglas son un gran porcentaje del total; sin embargo, a pesar de tener una gran presencia, no es posible obtener una información muy relevante dada la naturaleza las variables ubicadas en los antecedentes, siendo estos 'pdays' y 'poutcome' dichos variables tenían una alta frecuencia en los valores '999' y 'nonexistent' respectivamente en la base de datos. Acerca de las tres primeras reglas ordenadas por confianza, se tienen las variables 'duration', 'nr.employed'] y 'emp.var.rate' en los intervalos $]0,307]$, $]5.18e+03,5.2e+03]$ y $]0.8,1.1]$ respectivamente, tal como se mencionó la variable 'duration' no debería ser usada

para un modelo predictivo; sin embargo, es interesante analizar que si en una llamada de tele-mercadeo existe una duración menor a 5 minutos, dicha llamada al menos en un 99 % de confianza el valor de la clase 'y' será 'no'. Por otro lado, analizando la variable emp.var.rate se puede determinar que la tasa de empleo en el intervalo positivo]0.2,1.1] junto con los otros antecedentes tendrán un 99 % aprox. de confianza a 'y=no', similar caso para nr.employed, por lo que a pesar de que estas reglas tengan un soporte bajo del 13 % aproximadamente, cada vez que aparezcan asegurarán que la persona no se suscribirá a un crédito a largo plazo.

4.4. Comparación

En el documento N.º 2 de la línea de documentos referentes a la base de datos sobre el banco de Portugal se hablaba principalmente sobre algoritmos de agrupamiento, más en específico sobre el algoritmo K-medoids, que nos permitió encontrar dos clústeres que agruparon valores según la distancia Gower entre un valor y el medoide de un grupo, dicho medoide era calculado como el valor real más cercano al promedio de los valores de dicho grupo, por lo que era representativo del grupo y robusto, de esta manera se extrajeron los medoides de cada grupo y las frecuencias para realizar un análisis, de manera resumida las principales diferencias entre ambos clúster fueron:

1. Duration: Siendo mayor en el primer clúster, en ambos un valor positivo
2. emp.var.rate: Siendo negativo en el primer clúster y positivo en el segundo.
3. nr.employed y euríbor3m: Siendo mayor para ambos en el segundo clúster y a su vez ambos eran positivos en los dos clústeres.

Además, cabe recalcar que en el primer clúster existía un mayor porcentaje de los elementos que contaba con 'y' en el valor de 'yes' que en el segundo. Considerando lo anterior y las reglas obtenidas, se esperaría que las reglas descritas en la sección 4.2 tengan mayor representación en el primer clúster dado el mayor porcentaje de elementos con 'y'='yes' y para el clúster 2 se tengan las reglas descritas en la sección 4.3. Evaluando lo mencionado se tiene que los antecedentes de la Figura 4 que son principalmente las variables que se diferenciaban entre

los clústeres, salvo por 'nr.employed' la cual no se encuentra presente, en el caso de los antecedentes de la Figura 5, se encuentran principalmente las variables duration y emp.var.rate, de manera más específica el valor de duration describe un tiempo de llamada inferior a 5 minutos que es lo contrario al conjunto de reglas con el consecuente 'y=no', donde se presenta el antecedente duration en el intervalo entre 5 a 10 minutos. Hablando de la variable 'emp.var.rate' se mencionó que en el primer clúster era un valor negativo y en el segundo positivo, esto se relaciona con los valores de los antecedentes para conjunto de reglas. En las reglas de la sección 4.2 se tiene a la variable en un intervalo negativo, por lo tanto, las reglas que cumplan con dicho antecedente se cumplan en el primer clúster; En el otro lado se tiene a las reglas de la sección 4.3 que tiene a la variable en un intervalo positivo, porque también se deberán las reglas con dicho antecedente en el segundo clúster.

5. Conclusión

Siguiendo la línea de trabajos sobre la base de datos de 'Bank Marketing' (Moro et al., 2014), en el presente documento se logra establecer reglas de asociación para poder obtener patrones de comportamiento en la población en frente a la elección de un crédito bancario por su campaña de *telemarketing*. En el cual a su vez se compara con experiencias pasadas de esta línea de trabajos para comparar con 'clusters' obtenidos anteriormente e identificar si tienen alguna relación las nubes de datos separadas por las clases respectivas con las reglas obtenidas.

Para las reglas identificadas en esta base de datos se ha separado en tres grupos restringidos por el consecuente. Para las más interesantes son las cuales el consecuente es 'y' = 'yes' e 'y' = 'no', la cual para el primer filtro se obtienen reglas las cuales tienen un soporte bajo, esta situación es causada principalmente por la baja cantidad de observaciones en la base de datos la cual la variable 'y' toma aquel valor. No obstante los resultados de confianza y *lift* toman valores muy reconfortantes. Por otro lado, para las reglas 'y' = 'no', los datos de confianza y *lift* no son muy buenos, en especial para este último el cual indica que son valores con una tendencia a ser reglas independientes o azarosas.

En cuanto a la comparación con los clusters, solo se ha podido observar tres relaciones con las reglas las cuales destaca la variable de 'duration' con que la nube de datos que representa el resultado 'no', este tenía un valor menor a 5 minutos, la cual no sirve para un modelo predictivo si se requiere al caso, no obstante podría servirle el dato a la empresa por si da capacitación a sus empleados acerca de estrategias a optar en las llamadas.

Acerca de las dificultades presentes en la elaboración de esta investigación se puede decir que la obtención de la cantidad de categorías al hacer la discretización de las variables, por medio de la regla de Sturges, se ha encontrado que ha dejado de tener soporte para grandes cantidades de observaciones, no obstante con los resultados obtenidos se ha encontrado que a pesar de aquello han sido válidos los resultados de la discretización. Por otro lado, lo que respecta a la variable 'duration' junto con su regla de llamadas por superior a los 5 minutos para que 'y' sea 'yes' con un 97% de confianza, no se tiene información el proceso de la llamada. Esto se refiere a si un individuo en un tiempo en minutos inferior a 5 ya

acepta el crédito, no se sabe cuanto dura el proceso despues de aceptar, por lo que obtención de otros datos o protocolos demoren la llamada, haciendo que realmente la regla cumpla para tiempos menores que 5. Finalmente, en cuanto a la última dificultad, es la elección de limites de soporte y confianza para la obtención de reglas, las cuales deberían ser otorgadas por profesionales del campo de la banca y no supuestas. Por lo que los datos podrían estar ensuciados por aquellas suposiciones.

Finalmente, los objetivos propuestos se han cumplido con satisfacción, siguiendo el proceso de idenfificar, implementar y obtener medidas de calidad de las reglas con mayor interés implícitas en esta base de datos. Seguidos por un análisis separado por el filtro sobre la clase 'y' y una comparación con experiencias pasadas en terminos de identificar patrones en *clusters* y reglas.

Bibliografía

- Abdulazeez, T. (2019). Bank marketing association rule mining.
- Amat, J. (2018). Reglas de asociación y algoritmo Apriori con R.
- Bejarano-Roncancio, J. J. and Rivera-Torres, E. A. (2014). Determinación de la canasta básica de alimentos de la fundación banco arquidiocesano de alimentos de bogotá. *Revista de la Facultad de Medicina*, 62:11–17.
- Buitrago, B. (2021). Data Mining Overview I. - iWannaBeDataDriven.
- Fournier-Viger, P., Wu, C.-W., and Tseng, V. S. (2012). Mining top-k association rules. In *Canadian Conference on Artificial Intelligence*, pages 61–73. Springer.
- Gil, C. (2020). https://rpubs.com/cristina_gil/reglas_de_asociacion.
- González Aguilar, J. A. (2020). Intensidad de apoyos, salud mental, empleo y su relación con resultados de calidad de vida.
- Hyndman, R. J. (1995). The problem with sturges’ rule for constructing histograms. *Monash University*, pages 1–2.
- IBM (2021). Reglas de asociación.
- Lin, W.-Y., Tseng, M.-C., and Su, J.-H. (2002). A confidence-lift support specification for interesting associations mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 148–158. Springer.
- Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.

Anexos