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

Laboratorio N°3

Reglas de Asociación

Integrantes:	Felipe González
	Carlos Pérez
Curso:	Análisis de Datos
Sección	0-A-1
Profesor(a):	Dr. Max Chacón Pacheco
Ayudante:	Javier Arredondo Contreras

29 de Diciembre de 2020

Tabla de contenidos

Índice de cuadros

Índice de figuras

1. Introducción

El cáncer es la segunda causa de muerte en el mundo, esto según la Organización Mundial de la Salud (2018). Como se trata de una enfermedad que al pasar el tiempo va evolucionando, su tratamiento también va cambiando. A esto, si se le suma que existe una gran variedad de tipos, resulta crucial tener métodos de detección eficientes.

Se le llama cáncer solo a los tumores que son capaces de realizar metástasis, es decir, aquellos que son capaces de extenderse a los tejidos cercanos o de desprender células cancerosas que viajan a través del sistema circulatorio o linfático para llegar a producir nuevos tumores.

Cuando un cáncer produce metástasis en otra parte del cuerpo, el nuevo tumor posee las características de su origen. Por ejemplo, cuando un cáncer de mama se disemina a los pulmones, el nuevo tumor que se genera tiene las características de un cáncer de mama, e incluso, se le llama cáncer metastásico de mama y no cáncer de pulmón. Esto hace que la clasificación de un tumor recién hallado sea aún mas compleja.

En específico, en este informe se estudia el cáncer de mama. Usualmente, un tumor en el seno puede ser visto a través de un examen de rayos x y se puede sentir. Pero como se explicó anteriormente, no todo cáncer hallado en la zona corresponde a cáncer de mama, pues puede corresponder a un linfoma o sarcoma.

Es crucial estudiar este tipo de cáncer, puesto que "en las mujeres, el cáncer de mama es el segundo tipo de cáncer más común"(NIC, 2020). Además, posee una mortalidad del 10 %.

En esta investigación se realizó un estudio de *Reglas de asoiciación* mediante el uso del algoritmo *Apriori*, usando el dataset *Breast Cancer Wisconsin*. Uno de los objetivos de este laboratorio es determinar las reglas de mayor peso aplicando distintas medidas de calidad. Además, se analiza el significado de cada una con respecto a la clasificación de los tumores.

2. Marco Teórico

2.1. Reglas de asociación

Las Reglas de asociación se usan para encontrar la relación que hay entre los atributos (items) de un conjunto de datos (itemset) respecto a sus ocurrencias (transacciones). Por ejemplo, dentro de las compras realizadas en un supermercado, los items corresponderían a pan, leche, mantequilla, entre otros, las transacciones son las distintas canastas de compras realizadas durante el mismo día y un itemset pordría ser la compra de pan y leche. En este caso, resulta interesante estudiar qué items son los que se suelen comprar juntos, en otras palabras, asociar los productos.

Una regla de asociación se lee como una implicación: "si ocurre A, entonces B", donde a la parte izquierda (A) se le llama antecedente y a al derecha (B) consecuente. En la ecuación 1 se aprecia matemáticamente una regla de asociación.

$$A \Rightarrow B \tag{1}$$

Existen distintos algoritmos para realizar estas reglas de asociación, en esta experiencia se trabaja con el algoritmo *Apriori*. El algoritmo consta de dos partes: primero se identifican los itemset frecuentes, y segundo, transformar estos casos en reglas de asociación.

2.2. Medidas de calidad y confianza

2.2.1. Support (Soporte)

Es una medida de calidad que determina la cantidad de veces que está presente un itemset en las transacciones. En la ecuación 2 se define.

$$\sup(A) = P(A) = \frac{\#\text{Número de ocurrencias de A}}{\#\text{Casos totales}} \tag{2}$$

También puede ser aplicada a una regla de asociación y se interpreta como la probabilidad de que ocurra tanto el antecedente como el consecuente. Matemáticamente queda como se muestra en la ecuación 3.

$$\sup(A \Rightarrow B) = \sup(A \cup B) = \frac{\#\text{Numero de ocurrencias de} A \cup B}{\#\text{Casos totales}} \tag{3}$$

2.2.2. Confidence (Confianza)

Esta medida permite determinar si una regla de asociación tiene el suficiente peso como para ser considerada. Se interpreta como la probabilidad de que ocurra el consecuente dado su antecedente. En la ecuación 4 se presenta su expresión matemática.

$$\text{conf}(A \Rightarrow B) = \frac{\sup(A \Rightarrow B)}{\sup(A)} \tag{4}$$

2.3. Monotonicidad

Como se mencionó anteriormente, este algoritmo se basa en encontrar los itemset frecuentes. Esto implica que la complejidad del algoritmo aumente de forma exponencial debido a la cantidad de combinaciones posibles de itemset que puedan existir $(2^n - 1)$.

Para reducir esta cantidad, se debe considerar que si un itemset es frecuente, entonces todos sus subgrupos también los son. De forma análoga, si un itemset no es frecuente, entonces cualquier conjunto que contenga a este itemset, tampoco lo es (B. Buitrago, 2020).

De esta forma, al buscar reglas que, por ejemplo, cumplan con un valor de soporte mínimo, no sea necesario realizar las $2^n - 1$ iteraciones, sino que permite descartar los casos que ya se sabe que no cumplen con el criterio (realizar poda).

Para poder aplicar este concepto a una medida, hay que comprobar que la medida sea monótona, es decir que al generalizar una regla (agregar items al antecedente), el valor de la medida solo aumente o solo disminuya. El soporte es monótono y la confianza no.

2.4. Algoritmo Apriori

Para utilizar este algoritmo primero se necesitan las transacciones y definir un valor mínimo para los soportes y la confianza. Su funcionamiento es el siguiente

- Primero se generan todos los conjuntos de largo k = 1 de items. Luego, se descartan los itemset que no superen el soporte mínimo previamente establecido. Se repite este

paso aumentando el valor de k y considerando el principio de monotonicidad. Se itera hasta que el algoritmo no pueda seguir.

- En segundo lugar y a partir de los conjuntos obtenidos en el punto anterior, se generan todas las reglas de asociación posibles y se descartan las que no alcancen el umbral de confianza mínima previamente establecida.

El resultado de este algoritmo es un conjunto de reglas que cumplen con un cierto valor de soporte mínimo y una confianza mínima, a estas se les llama reglas interesantes.

2.5. Medidas de calidad

Luego de tener las reglas interesantes, se debe estudiar cuales son las que tienen mayor impacto en el caso de estudio. Para esto se ordenan las reglas siguiendo un criterio basado en alguna medida de calidad.

2.5.1. Lift

Medida de calidad usada para medir la frecuencia que tiene el antecedente y consecuente de aparecer en la misma transacción. Para esta medida se asume la independencia de los datos. En la ecuación 5 se expresa su formalización.

$$\text{lift}(A \Rightarrow B) = \frac{\text{conf}(A \Rightarrow B)}{\text{sup}(B)} \tag{5}$$

Se analizan tres casos:

- lift < 1: Indica que hay mas posibilidades que ocurra el consecuente cuando no ocurre su antecedente y viceversa.
- lift $= 1$: Indica que la ocurrencia del antecedente no tiene ningún efecto sobre el consecuente y son independientes entre sí.
- lift > 1: Indica que hay mas posibilidad que ocurra el consecuente cuando ocurre su antecedente y viceversa.

2.5.2. Conviction

Esta medida sirve para analizar la fuerza que tiene la implicación de una regla de asociación. Mientras su valor esté mas lejos de 1, entonces la regla es mas interesante. En la siguiente ecuación (6) se define su expresión matemática.

$$\text{conv}(A \Rightarrow B) = \frac{1 - \text{sup}(B)}{1 - \text{conf}(A \Rightarrow B)} \tag{6}$$

3. Obtención de Reglas

Las reglas de Asociación en el contexto de la Minería de Datos, al igual que el algoritmo de *K-medias* visto en experiencias anteriores, cae dentro de las denominadas *Técnicas de Aprendizaje No Supervisado*, recordando que son aquellas técnicas que trabajan con conjuntos de datos que no cuentan con una variable objetivo o *explicativa*, ya que precisamente busca establecer relaciones entre las variables a partir de patrones que tienden a repetirse frecuentemente entre las variables involucradas. Entonces, para ver como se ve afectado esto en el problema en estudio, es necesario recordar las características involucradas.

Nuevamente, el conjunto de datos corresponde a muestras de líquido tumoral de pacientes con protuberancias sólidas en su pecho, recabadas en distintas fechas y que en total suman 699 observaciones, las cuales fueron reducidas a 683 producto de la eliminación de 16 instancias que presentaban un único *missing value* para la variable denominada como *Bare Nuclei*. Inicialmente el conjunto cuenta con 11 variables **discretas** (ver cuadro 1), de las cuales 9 corresponden a características propias de la observación, mas una variable que denota la clase a la cual pertenece, y el código identificador.

Atributo	Tipo	Dominio
Clump Thickness	Numérico	1 - 10
Uniformity of Cell Size	Numérico	1 - 10
Uniformity of Cell Shape	Numérico	1 - 10
Marginal Adhesion	Numérico	1 - 10
Single Epithelial Cell Size	Numérico	1 - 10
Bare Nuclei	Numérico	1 - 10
Bland Chromatin	Numérico	1 - 10
Normal Nucleoli	Numérico	1 - 10
Mitoses	Numérico	1 - 10
Class	Nominal	2 , 4

Cuadro 1: Variables de Estudio

Para esta experiencia en particular y considerando la naturaleza de las Reglas de Asociación, mas allá del caracter clasificador de la variable *class*, esta se trata como cualquier otra variable del conjunto, sin embargo toma un papel fundamental en la generación de las distintas reglas. Considerando que el campo bajo el cual se desarrollo el estudio y se recaban las distintas observaciones, corresponde al del análisis de determinadas características propias de las células, y su posible influencia en la especialización anómala de estas, una pregunta que resulta de gran interés responder considerando la naturaleza del estudio es si, existe alguna configuración entre estas características que pueda influir o determinar si la característica *class* toma el caracter de *Benigno* (valor 2) o *maligno* (valor 4), y si es así, cual de estas configuraciones es la que tiene una mayor fuerza. Dicho en términos de las Reglas de Asociación, para cada una de las reglas generadas se considera como consecuente principal a la variable *class*.

3.1. Seleccion de Confianza y Soporte

Como ya debe ser de conocimiento, unas de las primera variantes que deben ser consideradas antes de ejecutar el algoritmo de Reglas de Asociación y que tiene una gran influencia en el resultado y en la calidad de estas, son la *Confianza* y el *Soporte*, y es que el cálculo de cada una de estas depende directamente de la relación entre ambas.

Tanto el Soporte como la Confianza determinan cuan *interesante* es una regla, parámetro el cual es medido en base a un umbral de interés, entonces mientras mas cerca estén estos valores de ese umbral, mas útil es la regla para el cliente (en el contexto de la canasta de mercado). Una regla **interesante** es aquella que es frecuente y confiable a la vez, donde:

- El conjunto de reglas **confiables** se entiende que es el conjunto de todas las reglas que cumplen con un umbral de confianza mínima *'minconf'*
- El conjunto de reglas **frecuentes** se entiende que es el conjunto de todas las reglas que cumplen con un umbral de soporte mínimo *'minsop'*

Ahora bien, definir estos valores mínimos puede resultar no ser tan trivial, ya que dependen directamente de la naturaleza de los datos, como por ejemplo, escoger un valor

muy bajo para el soporte puede concluir en la generación de un número extremadamente alto de reglas, lo cual en primera instancia puede considerarse bueno, pero también hay que tomar en cuenta que gran parte de estas reglas son redundantes, es decir, forman parte de subconjuntos mas grandes. Escoger un valor muy alto puede provocar que no se generen reglas para un determinado valor del *consecuente*. Es por esto que para buscar la mejor combinación de valores *Soporte-Confianza*, se realizan una serie de pruebas, registrando los datos obtenidos. En cuanto a los valores iniciales, se consideran los valores que trae por defecto la función *apriori* de R para ambos parámetros. (ver cuadro 2)

C	Soporte	Confianza	#*RHS* B	#*RHS* M	Total
1	0.2	0.8	135	0	135
2	0.1	0.8	644	2	646
3	0.05	0.8	1218	18	1236
4	0.01	0.8	4854	507	5361
5	0.001	0.8	43837	97994	141831

Cuadro 2: Distribución del *RHS* en las Reglas

Dado que la *confianza* no tiene un impacto significante en el número de reglas obtenidas, se decidió evaluar solo la variación del *soporte*. Como primera impresión se puede decir que para valores altos de *soporte*, el consecuente tiende a ser *Benigno*, mientras que para valores mas bajos tiende a *Maligno*.

En conclusión y considerando los resultado obtenidos en el cuadro anterior, se realizan las siguientes acotaciones:

- La primera configuración se descarta inmediatamente, a pesar de que se tiene un numero considerable de reglas correspondientes al consecuente *Benigno*, no existe ninguna regla que incorpore como consecuente al valor *Maligno*, lo que imposibilita realizar conclusiones respecto a esta ultima.

- La segunda ya incorpora reglas con consecuente *Maligno*, aun así se considera que solo tomar en cuenta dos reglas para el análisis, es un numero demasiado bajo.

- La tercera incorpora mas reglas con consecuente *Maligno*, aun así es demasiado bajo, considerando que aun deben descartarse las reglas redundantes.

- Disminuyendo el soporte un valor de *0.01*, el numero de reglas con consecuente *Maligno* toma un valor mucho mas considerable, a pesar de que las reglas con consecuente *Benigno* siguen en aumento y existe una gran diferencia, considerando nuevamente que hay que descartar reglas redundantes, se toma como una buena distribución para trabajar.

- Para la ultima configuración, el numero de reglas aumenta considerablemente para ambos consecuentes, que incluso las reglas con consecuente *Maligno* superan a las con consecuente *Benigno*, el problema es que tal como se menciona anteriormente el algoritmo comienza a demorarse en términos de computo.

Finalmente, considerando las reflexiones anteriores, a configuración que se decide tomar para proceder a la generación de reglas es la numero 4 que considera un valor de **0.01** para el *Soporte* y uno de **0.8** para la *Confianza*.

3.2. Eliminación de Reglas Redundantes

De acuerdo al punto anterior, considerando los valores escogidos para el *Soporte* y la *Confianza*, se tiene un total de 4854 reglas con consecuente *Benigno*, y un total de 507 reglas con consecuente *Maligno*, valores que siguen siendo demasiado grandes si se desea encontrar las reglas mas interesantes del conjunto. Es por esto que antes de pasar al directo análisis de las reglas mas interesantes, es necesario eliminar aquellas reglas denominadas como *'Reglas Redundantes'*, que corresponden a aquellas reglas que son un subconjunto de alguna otra regla, y que por ende conservarlas no tiene mucho sentido.

LHS + RHS	Original		Sin Redundancia	
	Benigno	Maligno	Benigno	Maligno
2	16	50	16	50
3	185	272	3	8
4	661	155	X	X
5	1178	27	X	X
6	1295	3	X	X
7	933	X	X	X
8	444	X	X	X
9	126	X	X	X
10	16	X	X	X
Total	4854	507	19	58

Cuadro 3: Distribución del largo (LHS + RHS) de la Regla

Como se puede observar en el cuadro 3, luego de eliminar las reglas redundantes del conjunto de reglas original, se obtiene un valor mucho mas reducido, 19 reglas para el caso del consecuente *Benigno*, y 58 para el caso del consecuente *Maligno*, ambos casos distribuidas para antecedentes de largo 2 y 3. Otro punto que vale la pena mencionar, es que normalmente cuando el antecedente tiende a aumentar de tamaño, como es el caso de 3 o 4 en adelante, las reglas son mucho mas especializadas, y su soporte también tiende a bajar mas, pues la probabilidad de que estas ocurran también lo hace.

3.3. Obtención de Reglas Interesantes

Ahora que el conjunto de reglas esta mucho mas acotado en comparación con el conjunto original, se puede proceder a determinar cual de estas reglas es la mas *interesante*, desde luego considerando que se entiende por 'interesante' para el dominio del problema. Para lograr esto, es necesario acotar aun mas el rango de reglas que se dispone, por lo que se procede a generar un *Top 10*, de aquellas reglas mas *confiables* y *frecuentes*, es decir aquellas que tienen un valor

mas elevado tanto para el soporte como para la confianza.

La medida de calidad considerada para el estudio de las reglas con un consecuente *Benigno* es *Conviction*, porque mide la probabilidad que en una transacción aparezca el antecedente sin su consecuente, es decir, un tumor aparenta ser benigno cuando no lo es. Si esto llegase a ocurrir, no se le diagnosticaría correctamente a un paciente que presenta un tumor canceroso. El resultado se puede ver en la tabla 4

RHS Benigno			
Regla	Soporte	Confianza	Convicción
1	0.5724744	0.9050926	3.687034
2	0.5666179	0.9626866	9.378038
3	0.5402635	0.9892761	32.630673
4	0.5314788	0.9236641	4.584041
5	0.5197657	0.9441489	6.265356
6	0.5036603	0.9942197	60.537335
7	0.2240117	0.9562500	7.998327
8	0.2166911	0.9866667	26.244510
9	0.1991215	0.9784173	16.213275
10	0.1786237	0.8531469	2.382835

Cuadro 4: Resumen del resultado del caso de los tumores benignos

Por otro lado, para el caso de las reglas que presentan como consecuente *Maligno*, se utiliza la medida *Lift*, porque se prioriza las veces que tanto el antecedente como el consecuente estén presentes en las transacciones a la vez. De esta forma, se prioriza las veces que la regla se cumple. El resultado se aprecia en la tabla 5

RHS Maligno			
Regla	Soporte	Confianza	Lift
1	0.18887262	0.9772727	2.792792
2	0.10102489	1.0000000	2.857741
3	0.09809663	1.0000000	2.857741
4	0.09516837	0.9154930	2.616241
5	0.08784773	1.0000000	2.857741
6	0.08491947	1.0000000	2.857741
7	0.07906296	0.9818182	2.805782
8	0.06002928	0.8541667	2.440987
9	0.05856515	0.9090909	2.597946
10	0.05710102	0.9750000	2.786297

Cuadro 5: Resumen del resultado del caso de los tumores cancerosos

Finalmente, las reglas interesantes corresponden a las expuestas en la ecuación 7 y 8

$$\begin{aligned}
(\text{normalNucleoli} = 1) &\Rightarrow \text{Benigno} \\
(\text{bareNuclei} = 1) &\Rightarrow \text{Benigno} \\
(\text{unifCellSize} = 1) &\Rightarrow \text{Benigno} \\
(\text{marginalAdhesion} = 1) &\Rightarrow \text{Benigno} \\
(\text{epithCellSize} = 2) &\Rightarrow \text{Benigno} \\
(\text{unifCellShape} = 1) &\Rightarrow \text{Benigno} \\
(\text{blandChromatin} = 2) &\Rightarrow \text{Benigno} \\
(\text{blandChormatin} = 1) &\Rightarrow \text{Benigno} \\
(\text{clumpThickness} = 1) &\Rightarrow \text{Benigno} \\
(\text{blandChromatin} = 3, \text{mitoses} = 1) &\Rightarrow \text{Benigno}
\end{aligned} \tag{7}$$

$$
\begin{aligned}
(\text{bareNuclei} = 10) &\Rightarrow \text{Maligno} \\
(\text{clumpThickness} = 10) &\Rightarrow \text{Maligno} \\
(\text{unifCellSize} = 10) &\Rightarrow \text{Maligno} \\
(\text{blandChromatin} = 7) &\Rightarrow \text{Maligno} \\
(\text{normalNucleoli} = 10) &\Rightarrow \text{Maligno} \\
(\text{unifCellShape} = 10) &\Rightarrow \text{Maligno} \\
(\text{marginalAdhesion} = 10) &\Rightarrow \text{Maligno} \\
(\text{epithCellSize} = 4) &\Rightarrow \text{Maligno} \\
(\text{clumpThickness} = 8) &\Rightarrow \text{Maligno} \\
(\text{epithCellSize} = 6) &\Rightarrow \text{Maligno}
\end{aligned}
\tag{8}
$$

4. Analisis de Resultados y Comparación

4.0.1. Análisis respecto al dominio del problema

Para el caso de los tumores cancerosos, las reglas que destacan por tener el lift mas alto (2.857741) son clumpThickness, unifCellSize, normalNucleoli y unifCellShape (reglas 2, 3, 5 y 6 para los tumores malignos) con un valor de 10. Además, clumpThickness aparece en dos reglas (2 y 9) al igual que epithCellSize (8 y 10) con valores de lift considerablemente altos.

Ahora, analizando el caso de los tumores benignos, el valor que pose la convicción más alta (60.537335) es unifCellShape (regla 6 para los tumores benignos) cuando presenta un valor de 1, le sigue unifCellSize (regla 3 para los tumores benignos) con la segunda convicción mas alta (32.6360673) cuando tiene valor 1. Se destaca blandChromatin que aparece en tres reglas (7, 8 y 10).

Para comprender de mejor manera que es lo que podrían estar tratando de comunicar cada una de las relaciones obtenidas anteriormente, en relación al dominio del problema, se puede resumir cada una de estas en una frase, que considere los antecedentes y consecuentes de la regla en cuestión, y el significado del valor correspondiente a cada una de las características involucradas en la regla. Si se consideran cada una de las reglas obtenidas anteriormente, se pueden determinar las siguientes conclusiones.

En cuanto a las conclusiones respecto a los consecuentes *Benignos*, se tiene lo siguiente:

- Si las células pertenecientes al liquido tumoral de un paciente son completamente uniformes, entonces se puede decir con un 100 % de certeza que el tumor es Benigno.
- Si las células pertenecientes al liquido tumoral de un paciente presentan un tamaño completamente normal, entonces se puede decir con un 100 % de certeza que el tumor es Benigno.
- Si la cromatina de una célula perteneciente al liquido tumoral de un paciente es ya sea, de textura completamente fina, o solo un 20 % de la cromatina es gruesa, o el 30$ de la cromatina es gruesa y además la actividad mitótica es completamente normal, entonces se puede decir con un 95,62&, 98,66 % y 85,31 % de certeza que el tumor es Benigno, respectivamente.

Ahora, en cuanto a las conclusiones basadas en las reglas mas interesantes que incorporan consecuentes *Malignos*, se tiene lo siguiente:

- Si las células pertenecientes al liquido tumoral de un paciente son completamente multicapa, o solo en un 80 % mono-capa, entonces se puede decir con un 100 % y un 90 % de certeza que el tumor es Maligno, respectivamente.
- Si las células epiteliales pertenecientes al liquido tumoral de un paciente son o un 40 %, o un 60 % mas grandes de lo habitual, entonces se puede decir con un 85 % y un 100 % de certeza, que el tumor es Maligno, respectivamente.
- Si las células pertenecientes al liquido tumoral de un paciente presentan un tamaño inconsistente, entonces se puede decir con un 100 % de certeza que el tumor es Maligno.
- Si los nucléolos de las células pertenecientes al liquido tumoral de un paciente son normales en su totalidad, entonces se puede decir con un 100 % de certeza, que el tumor es Maligno.
- Si las células pertenecientes al liquido tumoral de un paciente presentan un tamaño completamente inconsistentes con su uniformidad, entonces se puede decir con un 100 % de certeza que el tumor es Maligno.

4.0.2. Análisis comparativo respecto a los laboratorios 1 y 2

Recordar que en la primera experiencia se concluyó que existe una correlación positiva de todas las variables con respecto a class, donde se destacó las variables unifCellShape y unifCellSize por tener un valor mas alto en comparación a las demás.

En cuanto a la segunda actividad de laboratorio, los tumores cancerígenos presentaron valores altos en comparación a con los benignos para unifCellShape, unifCellSize y normalNucleoli. Además, los tumores cancerosos no requirieron valores altos de ephitCellSize y bareNuclei, de hecho, la media de estos fueron 5 y 3 respectivamente. Finalemente, clumpThickness tuvo las medias mas altas tanto para los tumores malignos como para los benignos.

Al igual que las experiencias anteriores, los valores altos de las variables indican que el tumor corresponde a uno canceroso y, de forma análoga, los valores bajos a uno benigno.

Se repiten las variables unifCellShape, unifCellSize y normalNucleoli como indicadores de interés para determinar si un tumor es canceroso, debido a que poseen el lift mayor. Cabe destacar que, para este método, también sobresale la medida clumpThickness.

También se destaca que, al igual que en la experiencia pasada, epithCellSize es la medida que tiene el menor valor para considerar que un tumor es canceroso (4). Sin embargo, pasa lo contrario con la variable bareNuclei, que solo corresponde a un indicador de tumor canceroso cuando tiene su máximo valor (10).

5. Conclusión

A modo de retrospectiva y considerando el objetivo general de esta experiencia, se logró realizar una implementación del algoritmo de reglas de asociación *Apriori* aplicado al conjunto de datos *Breast Cancer Wisconsin*, analizando distintos factores que influyen en la calidad de la agrupación final, llegando a las siguientes conclusiones.

- La razón de ser de este estudio, tiene principal relación con el carácter **no supervisado** del método de reglas de asociación, pues se busca determinar si las reglas formuladas a través de las transacciones presentes en los datos, pueden de alguna forma entregar información que permita explicar algún fenómeno. En específico, considerando la naturaleza del problema, determinar las reglas con suficiente peso, para poder identificar los conjuntos de características de mayor interés.

- De la mano de lo anterior, y considerando los resultados obtenidos, se logra aplicar el método para un determinado rango de valores de *minsop*, que corresponde al valor mínimo de soporte permitido para considerar una regla como regla de interés y, a partir de esto, elegir las mejores según sus *lift* o *coverage*.

- Las reglas de asociación generadas, además de depender del valor de *minsop* y *minconf*, dependen del preprocesamiento de los datos donde se considera el tratamiento de los *missing values* que, dependiendo de la naturaleza del problema, se pueden tratar de una forma u otra. En este caso, el conjunto de datos se basa en un estudio de diagnostico de células cancerosas, prescindir o tratar de predecir el valor de estos valores puede no ser algo que se tome tan a la ligera, por lo que dependiendo de la magnitud de la perdida de información, se puede tomar una opción o la otra.

- Conocer de antemano la variable explicativa de los datos, como es el caso de la variable 'class', puede influenciar en cierta forma la interpretación de los resultados finales, ya que en el proceso se puede esperar que los resultados arrojen el número de clases provistos por el conjunto de datos.

- Comparando con las entregas anteriores, los resultados no fueron totalmente sorprendentes, los valores altos de las variables siguen indicando una correspondencia con

tumores malignos. Además, siguen destacándose las mismas variables para considerar los casos cancerosos: *unifCellSize*, *unifCellShape*, *clumpThickness* y *normalNucleoli*. Sin embargo, la principal diferencia radica en la variable *bareNuclei* que, en este caso, requiere el valor mas alto para considerarlo como un factor de riesgo de cáncer.

- Finalmente, de esta experiencia se puede afirmar que los patrones observados en las transacciones dicen mas sobre los tumores benignos que de los cancerosos, donde se destacan los casos de *unifCellSize* y *unifCellShape*. Además, si se contrasta con el punto anterior, corresponden a las variables que mayor impacto tuvieron para determinar si un tumor resulta ser canceroso o no.

En cuanto a lo que queda pendiente, se puede considerar realizar una selección aún mas exhaustiva de medidas de calidad para elegir las reglas con mayor peso, esto con el objetivo de asegurar un conjunto de reglas mas adecuado. Además, resulta interesante estudiar el caso de reglas mas específicas para analizar el caso de síntomas tipo para un tumor canceroso. Finalmente, se puede considerar un mejor método para tratar los *missing values*.

Bibliografía

Amat, J. (2018). Reglas de asociación y algoritmo apriori con r. [Online] https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion.

Borges, L. (2015). Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection.

Buitrago, B. (2020). Data mining overview i. [Online] https://medium.com/iwannabedatadriven/data-mining-overview-i-c8546764d86f.

Contreras, J. A. (2020). Laboratorio 3 - análisis estadístico. [Online] https://www.overleaf.com/project/5d9d1e607432db0001bb293a.

Hahsler, M. (2004). A probabilistic comparison of commonly used interest measures for association rules. [Online] https://michael.hahsler.net/research/association_rules/measures.html#coverage.

Institute, N. C. (2020a). Breast cancer—patient version. [Online] https://www.cancer.gov/espanol/tipos/seno.

Institute, N. C. (2020b). Cancer stat facts: Female breast cancer. [Online] https://seer.cancer.gov/statfacts/html/breast.html.

Institute, N. C. (2020c). ¿qué es el cáncer? [Online] https://www.cancer.gov/espanol/cancer/naturaleza/que-es.

MERZOUKI, R. (2017). User manualbreast cancer diagnosis web user interface. [Online] https://www.rai-light.com/docs/BCD_User_Manual_v01.pdf.

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

OMS (2018). Cancer. [Online] https://www.who.int/news-room/fact-sheets/detail/cancer.

Salom, E. V. (2017). ¿qué son las células epiteliales? [Online] https://cienciatoday.com/que-son-celulas-epiteliales/.

Society, A. C. (2020a). Tipos de cáncer de seno. [Online] https://www.cancer.org/es/cancer/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/tipos-de-cancer-de-seno.html.

Society, A. C. (2020b). What is cancer? [Online] https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html.

Society, A. C. (2020c). What is cancer? [Online] https://medlineplus.gov/spanish/cancer.html.

UCI (2020). Breast cancer wisconsin (original) data set. [Online] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).