**UNIVERSIDAD DE SANTIAGO DE CHILE**

**FACULTAD DE INGENIERIA**

**DEPARTAMENTO DE INGENIERIA EN INFORMATICA**


# Análisis de Reglas de Asociación de la Base de Datos ZOO.

**AUTOR:**

LUIS ORELLANA ALTAMIRANO

**UNIVERSIDAD DE SANTIAGO DE CHILE**

**FACULTAD DE INGENIERIA**

**DEPARTAMENTO DE INGENIERIA EN INFORMATICA**

# TABLA DE CONTENIDO

# 1 Resumen

Este trabajo de investigación consiste en estudiar e interpretar mediante el análisis de reglas de asociación, los datos correspondientes a la base de datos ZOO. Inicialmente, se describen los objetivos generales y específicos de este trabajo de investigación. Luego, se describe el problema, describiendo la motivación que inspira este trabajo, el dominio del problema y la descripción del problema a resolver. A continuación, se presenta una descripción de la solución propuesta indicando las características, propósitos, alcances y limitaciones de la solución a implementar. Luego, en la metodología, herramientas y experimentación se describe la base de datos ZOO y se presentan las técnicas aplicadas de análisis de reglas de asociación, tales como: método de minimización de reglas por medio de algoritmo apriori, mínimo tanto de soporte como de confianza, utilización de técnicas de visualización tales como matriz de asociación, diagrama de caja, agrupación de reglas, diagrama de vértices, diagrama de paralelas, entre otras, haciendo uso del software "R" V3.3. A continuación, se muestran tablas y gráficos resultantes y su respectivo análisis detallado. Y finalmente, se entregan las conclusiones respecto al problema presentado.

*Palabras clave*: estadística descriptiva, análisis de reglas de asociación, algoritmo Apriori, base de datos Zoo, software "R"

# 2 Objetivos

## 2.1 Objetivos Generales

El objetivo general de este trabajo de investigación es realizar un análisis de reglas de asociación de la Base de Datos ZOO utilizando la herramienta de software "R" y así poder visualizar las diferentes reglas de asociación relacionadas con las características y clasificaciones de las diferentes especies de animales.

### 2.1 Objetivos Específicos

Los objetivos específicos planteados para este trabajo de investigación son:

- Revisar y normalizar la base de datos ZOO.

- Investigar y estudiar los métodos relacionados con reglas de asociación y su implementación en el software "R".

- Realizar análisis deductivo a partir de los datos encontrados y las hipótesis y/o problema planteado a solucionar en el desarrollo del estudio, además de la relación con los resultados del análisis de estadística descriptiva del laboratorio 1.

## 3 Descripción del Problema

### 3.1 Motivación

La principal motivación de este trabajo de investigación es encontrar regularidades inherentes en la base de datos ZOO y que tienen relación con sus características y clasificación, así pues sería de mucho interés responder algunas preguntas como:

- ¿Qué especies y características están frecuentemente juntos?
- ¿Qué característica o características de las especies son concluyentes para esperar el resultado de otra variable?

## 3.2 Literatura Relevante

Si bien en el primer informe del laboratorio 1 se menciona la literatura relacionada con artículos escritos que han hecho uso de la base de datos ZOO, es importante destacar otros artículos que también hacen uso de la base de datos ZOO, pero que están relacionados con las reglas de asociación, tal es el caso de lo escrito por Giraud-Carrieriu y Martinez (1995), quienes presentan un modelo de aprendizaje incremental, PDL2, que tiene la capacidad para codificar conocimientos previos en forma de preceptos y así obtener reglas de asociación más precisas. Por otro lado, lo presentado por Deshpandel y Karypis (2002) realizan una serie de experimentos con algoritmos de clasificación que usan Itemset frecuentes para expandir el espacio de funciones y evaluar una variedad de esquemas y así seleccionar características discriminantes en sus Dataset, así ellos en sus experimentos pudieron demostrar poder reducir sustancialmente el número de características compuestas y mejorar la precisión de la clasificación.

## 3.3 Definición del Problema

El problema principal planteado radica en encontrar reglas de asociación dentro de la base de datos ZOO, esto es para poder determinar patrones entre los datos que permitan las agrupaciones de especies según sus características morfológicas, y así afirmar y abalar las hipótesis demostradas de clasificaciones de animales en el anterior laboratorio, y resolver preguntas como:

• Las asociaciones más frecuentes están relacionadas a los mamíferos ?.

• Las características más concluyentes son referentes a la agrupación de animales para las especies “mamíferos”, “aves” y “peces” ?.

• Características tales como pelaje, número de patas o la capacidad de amamantar son concluyentes al momento de la clasificación de animales dentro de la especie “mamíferos”.

También, se hace necesario aclarar algunas asociaciones implícitas en el laboratorio 1 que no eran completamente claras al momento de utilizar técnicas de estadística descriptiva tales como moda, media, medidas de dispersión, etc. Por tal motivo, se presenta el problema de clasificar especies de forma más precisa por medio de analizar reglas de asociación entre características morfológicas, y así obtener conocimiento por medio de relaciones relevantes entre variables que son presentadas en el dataset para cada animal.

# 4 Descripción de la Solución Propuesta

La solución para el problema planteado será la investigación e implementación de búsqueda y análisis de reglas de asociación para la base de datos ZOO, y que permita aplicar métodos y/o técnicas de reglas de asociación haciendo uso del software "R". Para tal escenario, se ha planteado hacer uso del paquete "arulesViz", el cual permite visualización, obtención y por consiguiente el análisis de dichas reglas encontradas.

## 4.1 Características de la Solución.

La solución contempla la búsqueda de la más eficiente técnica de análisis y extracción de reglas de asociación para la base de datos ZOO, que no es posible inferir o extraer de forma natural. Para ello, se utilizarán los métodos contenidos y/o incorporados dentro del paquete de software llamado "ArulesViz" y que es utilizado por el software "R" para entrega y visualización de resultados en reportes, gráficos y tablas.

### 4.2 Propósitos de la Solución.

Al termino del trabajo de investigación y una vez analizados los resultados obtenidos de cada experimento se espera encontrar reglas de asociación dentro de la base de datos ZOO y así determinar patrones entre las especies de animales y sus agrupaciones según sus características morfológicas, y también poder realizar un comparativo con las conclusiones dadas en el laboratorio 1 referentes a los análisis de estadísticas descriptivas.

### 4.3 Alcances y Limitaciones de la Solución.

Este laboratorio contempla sólo el análisis y extracción de reglas de asociación para la base de datos ZOO y haciendo uso de la herramienta de software "R" y su librería "ArulesViz" con sus métodos integrados. El laboratorio no contempla realizar análisis a la base de dato ZOO con otros métodos de análisis datos y/o minería de datos.

## 5 Metodología, Herramientas y Experimentación

### 5.1 Descripción de la Base de Datos ZOO (Dataset ZOO)

El Dataset ZOO cuenta con 101 animales extraídos desde una colección Zoológica. Hay 16 variables con una serie de rasgos que describen a los animales. Se definen 7 Tipos de Clases: Mamíferos, Aves, Reptiles, Peces, Anfibios, Insectos e Invertebrados.

Con la finalidad de disponer del contenido del Dataset ZOO, y hacer más fácil su análisis en interpretación de estos, se dispone del siguiente archivo, el cual se detalla a continuación:

Zoo.csv: Corresponde a una muestra de 101 animales. Por otro lado, el primer registro que se encuentra en este archivo es el encabezado, el cual corresponde a las características más

relevantes de cada especie, pues cobra sentido al momento de agrupar y clasificar cada animal según sus especies por medio de estas características morfológicas.

Las características morfológicas disponibles para cada animal presente en el dataset se presentan a continuación:

- hair ∈ {0, 1}: Posee pelaje (si, no)
- feathers ∈ {0, 1}: Posee plumas (si, no).
- eggs ∈ {0, 1}: Nace por medio de huevos (si, no)
- milk ∈ {0, 1}: Capacidad de amamantar (si, no)
- airborne ∈ {0, 1}: Capacidad de volar (si, no).
- aquatic ∈ {0, 1}: Vive en medio acuático (si, no)
- predator ∈ {0, 1}: Es depredador (si, no)
- toothed ∈ {0, 1}: Posee dentadura (si, no)
- backbone ∈ {0, 1}: Posee columna vertebral (si, no)
- breathes ∈ {0, 1}: Es pulmonado (si, no)
- venomous ∈ {0, 1}: Es venenoso (si, no)
- fins ∈ {0, 1}: Posee aletas natatorias (si, no)
- legs Numeric ∈ {0,2,4,5,6,8}: Cantidad de patas que posee
- tail ∈ {0, 1}: Posee cola (si, no)
- domestic ∈ {0, 1}: Puede ser domesticado (si, no)
- catsize ∈ {0, 1}: Posee el tamaño de un gato doméstico (si, no)
- class_type ∈ {1, 2, 3, 4, 5, 6, 7}: Clasificación del animal, donde:

- 1 = Mamífero
- 2 = Ave
- 3 = Reptil
- 4 = Pez
- 5 = Anfibio
- 6 = Insecto
- 7 = Invertebrado

## 5.2 Metodología y Herramientas.

Como herramienta de software utilizado para realizar los experimentos presentados en este documento fue "R" versión (3.3.3). Esta herramienta está bajo licencia GNU y cuenta con su propio lenguaje de programación con un enfoque al análisis estadístico.
"R" entre sus características principales permite: modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de clasificación y agrupamiento, generar gráficos, etc.
Como paquete de software dentro de "R", se utilizó "ArulesViz" y que hace uso de algoritmos de generación de reglas de asociación como apriori y otros, además de entrega y visualización de resultados en reportes, gráficos y tablas.

## 5.3 Experimentación.

Los siguientes comandos en "R" fueron utilizados para realizar el análisis de las reglas de asociación:

- `library('')`

Permite Agregar nuevas librerías, las cuales incluyen funciones no disponibles de forma nativa al repositorio local del Software “R”.

- `Nombre_Variable = read.csv("")`

Lee en memoria volátil un documento con extensión “csv”, y lo almacena en la variable “Nombre_Variable”.

- `summary(data)`

Entrega detalles globales estadísticos de los datos (media, moda, etcétera).

- `data.frame(sapply(data,as.factor))`

Permite discretizar las columnas en las que se aplicara algoritmo apriori.

- `apriori(data,parameter=list(support=#, confidence=#))`

Se obtienen las reglas de asociación indicando el soporte y confianza mínima.

- `inspect(head(sort(rules, by ="lift"),3))`

Permite visualizar las primeras tres reglas obtenidas, ordenándolas de forma descendente según lift.

- `rules[quality(rules)$lift > 1.46]`

Se obtiene un subgrupo de reglas con un lift mínimo.

- `plot(subrules,method="",measure="")`

Visualización de un gráfico determinado según “method” indicando parámetro de medición “mesure”.

**Extracción de conocimiento del DataSet ZOO**

**Reglas de asociación obtenidas.**

Para el proceso de obtención de reglas, se ha utilizado el algoritmo apriori, el cual hace uso del concepto o propiedad de anti-monotonía. Esta propiedad es expresada de la siguiente manera:

$$\forall X,Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

El principio fundamental de esta propiedad radica en el hecho de que no es posible que el sustento de determinado itemset sea mayor al sustento de sus subconjuntos de ítems. Ante esto, es posible determinar de antemano, las reglas que tendrán un menor sustento al establecer un soporte mínimo.

Como se ha visto anteriormente, el dataset “Zoo” posee un total de 16 propiedades o características morfológicas y 101 animales, por lo que el soporte mínimo estará sustentado con un rango muy reducido de animales a los cuales aplicar tales reglas de asociación.

Se sabe que la cantidad de reglas que es posible obtener por medio del algoritmo apriori puede ser completamente inmanejable y muy difícil de analizar en su conjunto sin las

herramientas precisas, por tal motivo es necesario acotar en gran medida la cantidad de tales reglas que son obtenidas en este dataset. Con la finalidad de obtener en lo posible una cantidad reducida de reglas, el soporte mínimo esperado es de:

- MinSupp: El soporte mínimo escogido para la obtención de reglas relevantes es de 0,37. Cabe destacar que el soporte mínimo se aplica tanto en la reducción de dimensiones previa confección de reglas, como reducción de reglas encontradas después del proceso anteriormente mencionado.

  Es necesario mencionar que el soporte mínimo es crucial en el proceso de obtención de reglas, debido a que determinara la cantidad de reglas a analizar. Por tal motivo, cabe destacar que reglas con muy bajo soporte en su mayoría, no revisten una gran importancia en el proceso de obtener relaciones significativas que permitan inferir información relevante. Debido a esto se busca obtener la mayor dispersión o concentración de reglas relevantes en el menor número de datos posibles. Con tal premisa, al obtener reglas con un soporte mínimo de 0,2 se obtienen más de 400.000 reglas lo cual claramente se vuelve insostenible y no viable para el proceso de análisis. Por otra parte, al obtener reglas por sobre un soporte de 0,35 se obtiene una cantidad por sobre las 45.000 reglas. Sin embargo, lo que se pretende es obtener una cantidad de relaciones que permita realzar análisis y conseguir conocimiento relevante para la clasificación de animales. El número de reglas que se busca obtener, es en el orden de 10.000. Esa cantidad de asociaciones se consigue por medio de un soporte mínimo de 0,37.

- MinConf: La confianza mínima aceptada para la obtención de reglas de asociación es de 0,5 lo que implica que el 50% de los items que contienen los antecedentes, tendrán una ocurrencia sobre los ítems contenidos en el consecuente de tales reglas encontradas. Este porcentaje es escogido debido a que tan solo son 16 características que posee este dataset, a diferencia de ejemplos ocurrentes en la clasificación de reglas, los cuales hablan acerca de la aplicación de esta técnica sobre supermercados, los cuales pueden contener miles de productos distintos.

Con los criterios de búsqueda anteriormente mencionados, la cantidad de reglas encontradas son de 5.736 dentro de las cuales, es posible destacar las siguientes reglas (tabla 1):

*Tabla 1. Reglas con mayor soporte y lift para cada una de las especies en dataset Zoo.*

| Reglas | Soporte | Confianza | Lift |
|---|---|---|---|
| {hair=1,feathers=0,eggs=0,airborne=0, breathes=1}=>{milk=1} | 0.3564356 | 1 | 2.463415 |
| {hair=1,milk=1,breathes=1,legs=4} => {class_type=1} | 0.3069307 | 1 | 2.463415 |
| {backbone=1,venomous=0,milk=1}=> {eggs=0} | 0.3960396 | 0.975609 | 2.346109 |
| {hair=0,eggs=1,milk=0}=> {domestic=0} | 0.4950495 | 0.925925 | 1.062710 |
| {hair=0, eggs=1, milk=0} => {feathers=0} | 0.3366337 | 0.629629 | 0.785093 |
| {hair=0, eggs=1, =0} => {breathes=1} | 0.3069307 | 0.620000 | 0.782750 |
| {eggs=1, aquatic=1} => {hair=0} | 0.2871287 | 0.966666 | 1.683333 |
| {feathers=0, eggs=1,airborne=0, catsize=0} => {milk=0} | 0.2574257 | 1 | 1.683333 |
| {eggs=1, backbone=1, tail=1} => {fins=0} | 0.2574257 | 0.666666 | 0.801587 |
| {eggs=1,fins=0} => {feathers=0} | 0.2574257 | 0.565217 | 0.704777 |
| {eggs=1, milk=0, fins=0} => {backbone=1} | 0.2772277 | 0.622222 | 0.757162 |
| {hair=0,feathers=0, airborne=0}=> {backbone=0} | 0.2574257 | 0.448275 | 0.696551 |
| {eggs=1, backbone=0, airborne=1} => {aquatic=0} | 0.2574257 | 0.472727 | 0.734545 |
| {eggs=1, backbone=1} => {milk=0} | 0.2434485 | 0.635827 | 1.457835 |
| {eggs=1, backbone=0, airborne=0} => {aquatic=0} | 0.2573432 | 0.472727 | 0.734545 |

# 6 Análisis de Resultados de las Reglas de Asociación

Debido a la gran cantidad de reglas de asociación obtenidas (5.736), es necesario ir a través de las reglas con el objetivo de encontrar las más interesantes. Sin embargo, es imposible realizar esta tarea analizando y comparando regla por regla. Por tal motivo, la visualización tiene una larga historia de permitir que los grandes conjuntos de datos sean más accesibles utilizando técnicas como la selección y el zoom. En este documento se ha utilizado el paquete de extensión R "arulesViz", el cual implementa varias técnicas de visualizaciones conocidas y novedosas para explorar reglas de asociación. Por medio de este paquete se muestra cómo estas técnicas de visualización pueden ser utilizadas para analizar un conjunto de datos. Es claro que analizar e ir a través de 5.736 no es una forma valida de seleccionar las que sean más interesantes. Por tal motivo, a continuación se presentas distintas perspectivas y métodos para analizar las reglas obtenidas.

**Diagrama de dispersión.**

Una visualización directa de las reglas de asociación es usar un diagrama de dispersión con dos medidas de interés sobre los ejes. Tal presentación se puede encontrar ya en un artículo anterior de Bayardo Jr. y Agrawal cuando discuten reglas sc-óptimas. Por consiguiente, es posible visualizar y analizar de forma global, muchas reglas a la vez, lo cual puede dar indicios de la utilidad y cuan precisas pueden ser tales reglas en su conjunto.

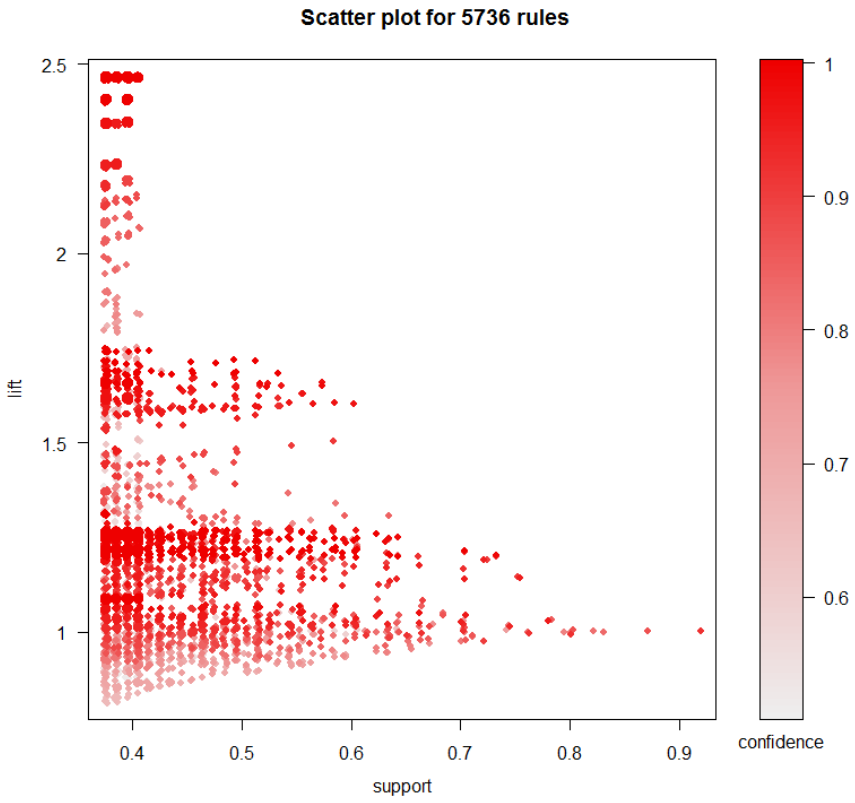


*Ilustración 1. Diagrama de dispersión para 5.736 reglas de forma simultánea.*

Por medio de la Ilustración 1, es posible observar que la mayor dispersión de reglas se encuentra por debajo del lift 1.5, lo cual indica que la probabilidad de la mayoría de las reglas de ocurrencia de sus respectivos antecedentes y consecuentes, son independientes entre sí ya que están muy próximos a lift 1.

Por otra parte, es posible detectar que la mayor concentración de reglas se encuentra cercana al soporte de 0,4, lo que indica que la probabilidad de que esas reglas se cumplan a lo largo del dataset es superior al 40%, lo cual sin duda es señal que se está ante un volumen alto de reglas de las cuales tienen una gran ocurrencia dentro del dataset. También es indicio que sin duda hay bastantes relaciones y patrones a tener en cuenta al momento de seguir el análisis de las reglas.

Por último, es posible notar que las reglas obtenidas tienen una gran confianza, lo que es indicio que la probabilidad de implicancia entre el antecedente y el consecuente es alta, lo cual hace referencia a que entre los itemset existe una gran afinidad.

**Análisis de resultados obtenidos por medio de diagrama de dispersión con respecto a conclusiones de laboratorio anterior.**

Por medio del diagrama de dispersión visto anteriormente, es posible inferir que en su mayoría, las reglas obtenidas son confiables desde la perspectiva de los indicadores "confianza" y "soporte", lo que indica que las probabilidades de ocurrencia de las reglas encontradas en el dataset son altas. Ante esto, es posible afirmar que en definitiva, hay indicios de patrones marcados en las reglas, sin embargo, estos serán tratados más a fondo en los siguientes segmentos de este documento.

Como fue posible determinar en el laboratorio anterior, había características morfológicas muy marcadas que indicaban de forma decisiva la especie a la cual podría corresponder un animal determinado. Tal es el caso de la característica "plumas", las cuales ningún animal además de las aves posee. Otro ejemplo es la característica "amamantar", la cual está asociada por completo a los mamíferos. Es decir, sin duda hay patrones en la secuencia y reglas implícitas dentro de los datos que permiten realizar agrupaciones según morfologías. Tal como se pudo observar por medio del grafico de dispersión, hay reglas muy recurrentes en el dataset.

**Diagrama de matriz de asociación de reglas.**

Las técnicas de visualización basadas en matriz, organizan los conjuntos de elementos antecedentes y consecuentes X y Y respectivamente. Una medida de interés seleccionada se muestra en la intersección del antecedente y consecuente de una regla dada. Si no hay ninguna regla disponible para un antecedente/consecuente según su combinación el área de intersección se dejara en blanco.

Se debe tener en cuenta que el grafico de matriz fue diseñado para el análisis de número limitado de reglas con la finalidad de facilitar su comprensión. Para tal caso, se seleccionó un compendio de un total de 1.928 reglas, las cuales tienen un lift igual o superior a 1,46.

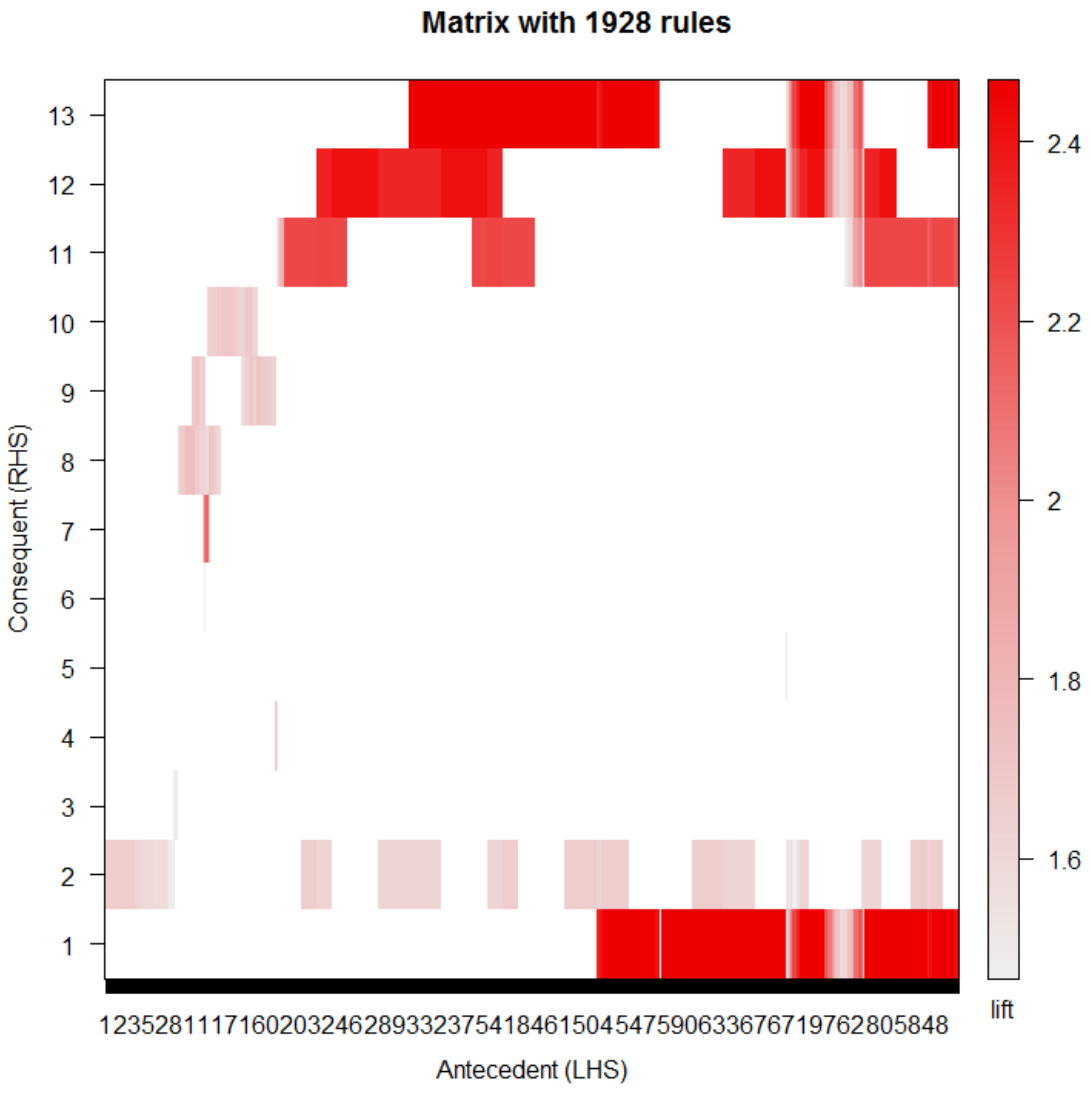

*Ilustración 3. Diagrama de matriz de asociación de reglas según lift.*

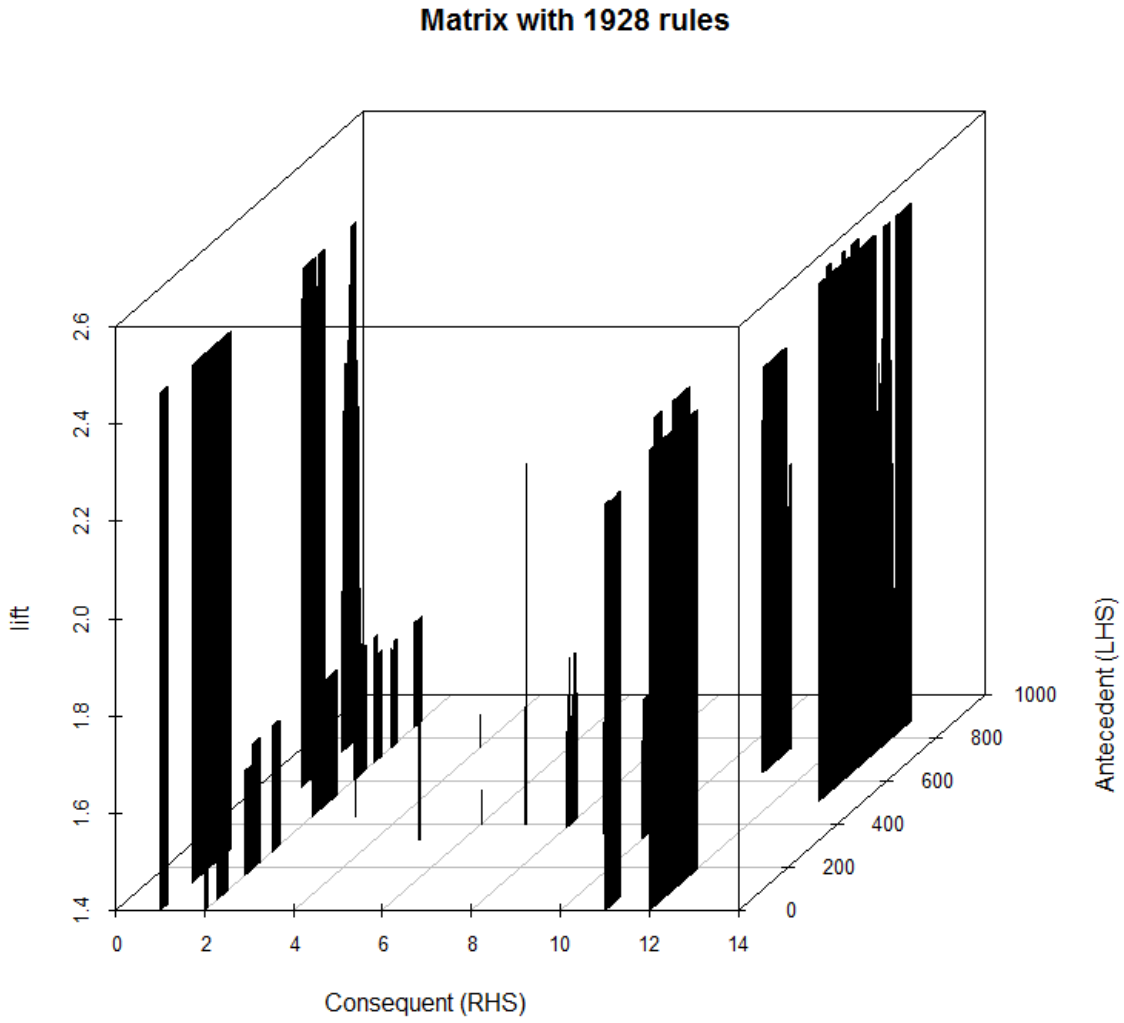

*Ilustración 4. Diagrama de caja de asociación de reglas.*

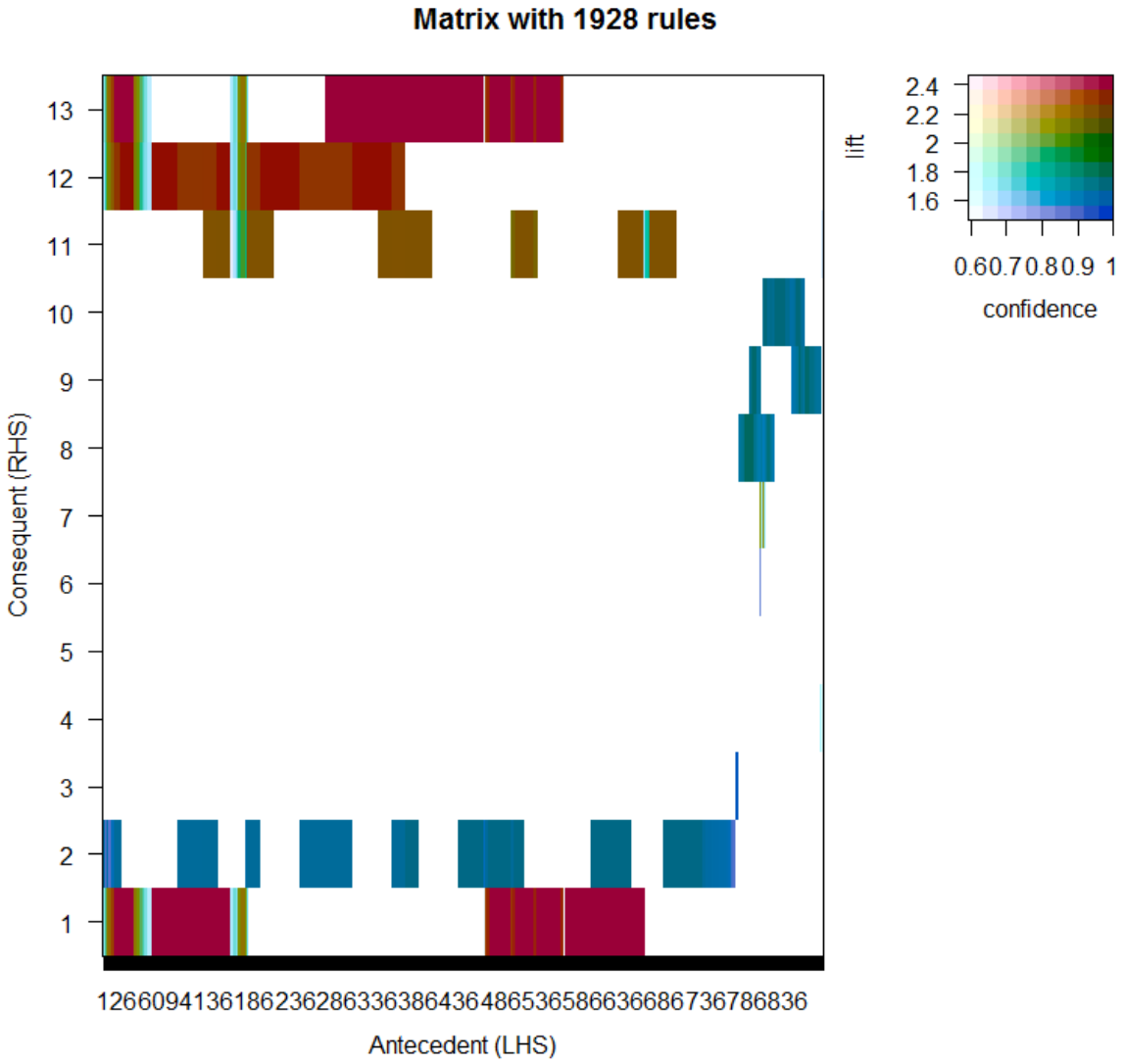


*Ilustración 5. Diagrama de matriz de asociación de reglas según lift y confianza.*

*Tabla 2. Lista de consecuentes de matriz de asociación de reglas*

| | | | |
|---|---|---|---|
| [1] | {milk=1} | [2] | {toothed=1} |
| [3] | {aquatic=0} | [4] | {legs=4} |
| [5] | {catsize=1} | [6] | {catsize=0} |
| [7] | {toothed=0} | [8] | {hair=0} |
| [9] | {eggs=1} | [10] | {milk=0} |
| [11] | {hair=1} | [12] | {eggs=0} |
| [13] | {class_type=1} | | |

Por medio de las gráficas anteriores (Ilustración 3 y 4), es posible inferir que los consecuentes que tienen una mayor asociación con sus antecedentes son el 1, 11, 12 y 13 (Tabla 2). Esto indica que una característica morfológica decisiva en la agrupación de muchos animales presentes en el dataset es que dan de amamantar. Otra característica que tiene gran afinidad con sus antecedentes son: Poseen “Pelaje”, no nacen por medio de “huevos” y por ultimo si son “mamíferos”. Por consiguiente es posible afirmar que las relaciones más fuertes entre características morfológicas de los animales presentes en el dataset, corresponden a los antecedentes anteriormente mencionados. Sin embargo, al considerar la Ilustración 5, es posible inferir que la confianza del consecuente “Pelaje”, tiene una menor ocurrencia en combinación con sus antecedentes, con respecto al resto de consecuentes analizados por medio de las Ilustración 3 y 4. Esto quiere decir que la probabilidad de encontrar en el dataset que el consecuente “Pelaje” junto a sus antecedentes, es menor que la probabilidad de los consecuentes 1, 12 y 13 (Tabla 2). Es decir, “Pelaje” no es una morfología decisiva al momento de categorizar a un determinado animal.

**Análisis de resultados obtenidos por medio de matrices de relaciones con respecto a conclusiones de laboratorio anterior.**

Durante el anterior laboratorio, fue efectivamente posible determinar que la mayoría de los animales presentes en el dataset tienen como característica principal el "amamantar" tal como lo menciona el consecuente 1 de la Tabla 2. Por otra parte, otra característica decisiva para la clasificación de otras especies que no pertenecen a mamíferos, es que no "amamantan", lo que nuevamente es indicado por el consecuente 10. Adicionalmente, tanto la característica "pelaje" y "clasificación 1" poseen gran lift, lo cual indica la gran afinidad que poseen con sus consecuentes y de esta forma reafirman la información recogida en el primer laboratorio que indica la presencia de características decisivas para la especie mamífero.

**Diagrama de matriz de agrupación según soporte y lift.**

La visualización basada en matriz está limitada en el número de reglas que se puede visualizar con eficacia desde grandes conjuntos de reglas. También suelen tener grandes conjuntos de antecedentes/consecuentes únicos. Se analizará una nueva técnica de visualización que mejora las técnicas anteriores basadas en matriz, mediante el agrupamiento de reglas para manejar un mayor número de relaciones e implicancias. Las reglas agrupadas se presentan como un agregado en la matriz y se pueden explorar de manera más eficiente.

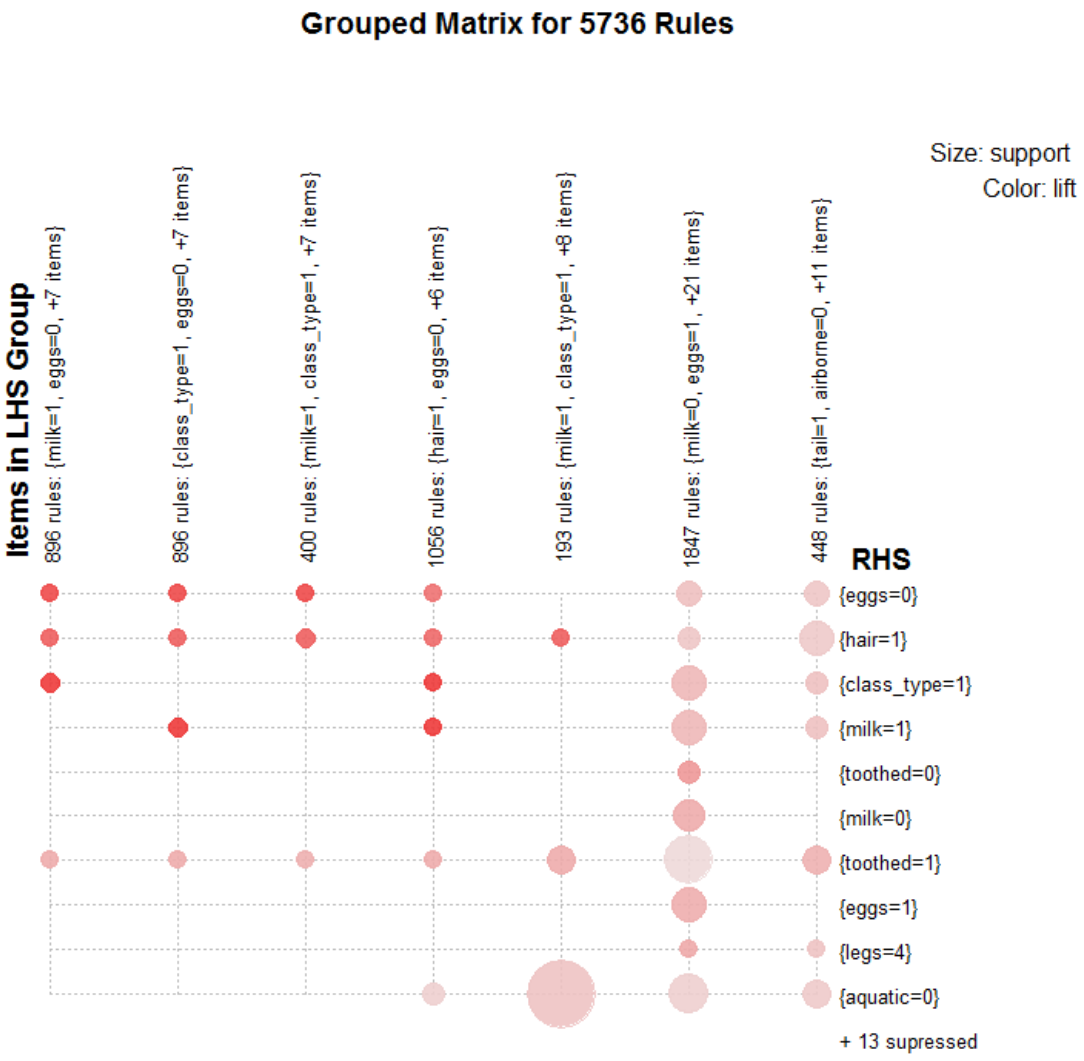


*Ilustración 6. Diagrama de agrupación de reglas según lift y soporte.*

Por medio de la Ilustración 6, es posible inferir que los ítemset más frecuente (soporte) dentro del dataset son:

- {dan de amamantar, son mamíferos} => {no son acuáticos}

  Tal como se ha mencionado durante este documento y el laboratorio anterior, en su gran mayoría, el dataset Zoo se compone de animales categorizados como mamíferos, lo cual es sin duda reafirmado por el conjunto de reglas con mayor soporte (tamaño de la esfera) dentro del diagrama de agrupación. Por consiguiente, es posible mencionar la posibilidad en la relación de la especie "mamíferos" con "amamantar" y en su mayoría, no ser animales "acuáticos".

- {no dan de amamantar, ponen huevos} => {son dentados}

  Otra regla que tiene gran frecuencia es la presentada anteriormente. Por consiguiente, otra mayoría dentro del dataset son especies que dentro de sus características está el no dar de "amamantar", nacer por medio de "huevos", lo cual implica que sí son "dentados". Ante esto, se puede inferir que la otra mayoría dentro de los datos corresponde a animales categorizados como "anfibios", "reptiles" y "peces", los cuales no poseen las características morfológicas mencionadas anteriormente.

Otro criterio de medición para la gráfica de agrupación de reglas es lift, el cual indica cuan significativa es la implicancia entre antecedente y consecuente. A continuación se muestran las agrupaciones más significativas según este criterio de medición:

- {mamífero, no ponen huevos} => {amamanta}
- {mamífero, posee pelaje} => {amamanta}

Con estas agrupaciones de reglas se confirma que la clase mamíferos está estrechamente

asociada y es completamente dependiente a la característica “Pelaje” y “Huevos”, al igual que su implicancia, la cual es “Amamantar”.

**Análisis de resultados obtenidos por medio de matrices de agrupaciones con respecto a conclusiones de laboratorio anterior.**

Por medio del análisis de la gráfica de agrupación de reglas, fue posible determinar asociaciones significativas o representativas de características morfológicas de la especie “mamíferos”.

En el laboratorio anterior fue posible determinar patrones que seguía tal especie. Por ejemplo, una de sus características más significativas, la cual no posee ninguna otra especie es la capacidad de “amamantar”, lo cual fue corroborado en el análisis de esta sección. Una característica prácticamente excluyente para la especie “mamíferos” que fue mencionada en el primer laboratorio y que es corroborada en esta sección, es el atributo “nacer por medio de huevos”.

**Gráfico basado en visualización de vértices.**

Las técnicas basadas en gráficos de vértices, permiten visualizar reglas de asociación que utilizan vértices y bordes, donde los vértices representan típicamente agrupaciones o conjuntos de elementos y los bordes indican la relación entre las reglas. Las medidas de interés se añaden típicamente a la etiquetas en los bordes o por color o anchura de las flechas que muestran los bordes. La visualización basada en gráficos de vértice ofrece una representación muy clara de las reglas, pero tienden a volverse confusas y por lo tanto sólo son viables para conjuntos muy pequeños de reglas. Para las siguientes gráficas, fueron seleccionadas solo 10 reglas con lift más alta.

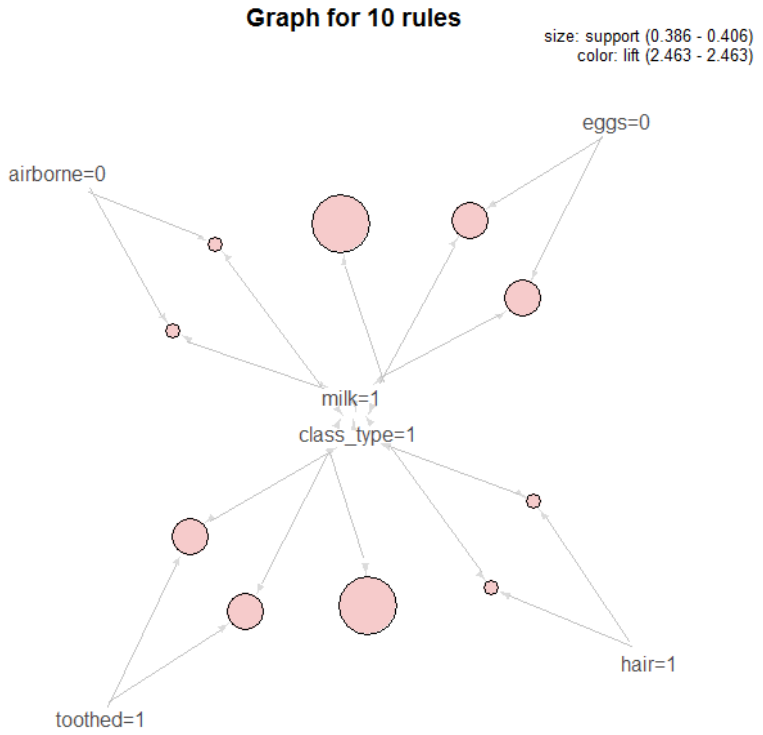


*Ilustración 7. Diagrama de visualización de vértices según lift y soporte.*

Por medio de la gráfica, es posible inferir que las reglas con mayor lift representan nuevamente a la clase “mamíferos”. Tanto los ítemsets {airborne=0, eggs=0} como {toothed, hair=1} poseen prácticamente el mismo soporte.

Según la gráfica, las dos características más representativas para los “mamíferos” son “dentados” y “no nacen por medio de huevos”. Por otra parte, las características menos significativas al momento de clasificar “mamíferos” son “no aéreos” y “poseen pelos”. Esto es posible entenderlo por medio de un animal particular. Para la propiedad “no aéreos” el cual quedarían fuera de la agrupación “mamíferos”, es el animal murciélago el que se presenta como una minoría dentro del dataset. Por otra parte, para la morfología “pelo”, un ejemplar que quedaría fuera de la clasificación “mamíferos” es el topo, el cual si pertenece a esta clasificación, sin embargo no reúne esta característica, por tal motivo al igual que

murciélagos, es una minoría dentro de la muestra disponible.

**Análisis de resultados obtenidos por medio de visualización de vértices con respecto a conclusiones de laboratorio anterior.**

Para la propiedad “no aéreos” quedaría fuera de la agrupación “mamíferos”, el animal murciélago. Por otra parte, para la morfología “pelo”, quedaría fuera de la clasificación “mamíferos” el topo, el cual si pertenece a la agrupación “mamíferos”. Por tal motivo, como se pudo analizar en el laboratorio anterior, para la especie “mamíferos” hay algunas excepciones minoritarias, debido a esto, las características anteriormente mencionadas no tienen un lift tan elevado como la característica “amamantar”, el cual si es una característica decisiva para la agrupación de la especie “mamíferos”, tal como se pudo observar en graficas de agrupación morfológica en el documento previo. Es posible visualizar la minoría de animales en que se aplica el itemset encontrado {airborne=0, class_type=1} por medio de la Ilustración 8 el que pertenece al laboratorio anterior.

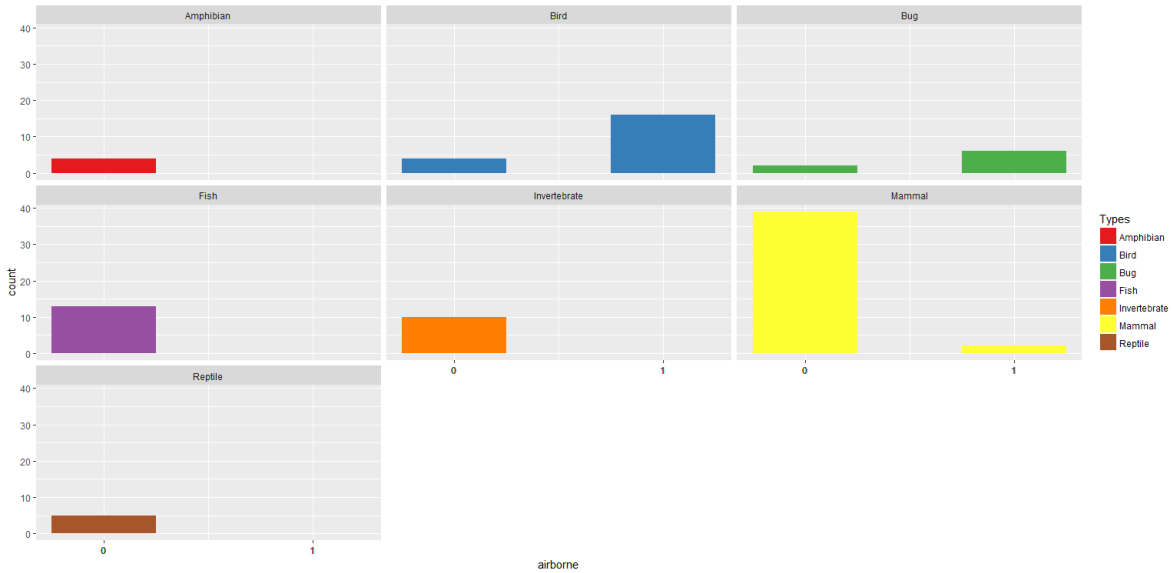


*Ilustración 8. Agrupación de animales según característica “aéreos”.*

**Grafico basado en coordenadas paralelas para 10 reglas.**

Este tipo de grafica está diseñada para visualizar datos multidimensionales donde cada dimensión se muestra por separado en el eje X y el eje Y. Cada punto de datos es representado por una línea que conecta los valores para cada dimensión. El grafico basado en coordenadas paralelas muestra los ítems en el eje Y como valores nominales y el eje X representa las posiciones en una regla, es decir, primer elemento, segundo elemento, etc. En lugar de una línea simple, se usa una flecha donde la cabeza apunta al elemento consiguiente. Las flechas solamente abarcan suficientes posiciones en el eje X para representar todos los elementos de la regla, es decir, reglas con menos elementos son flechas más cortas.

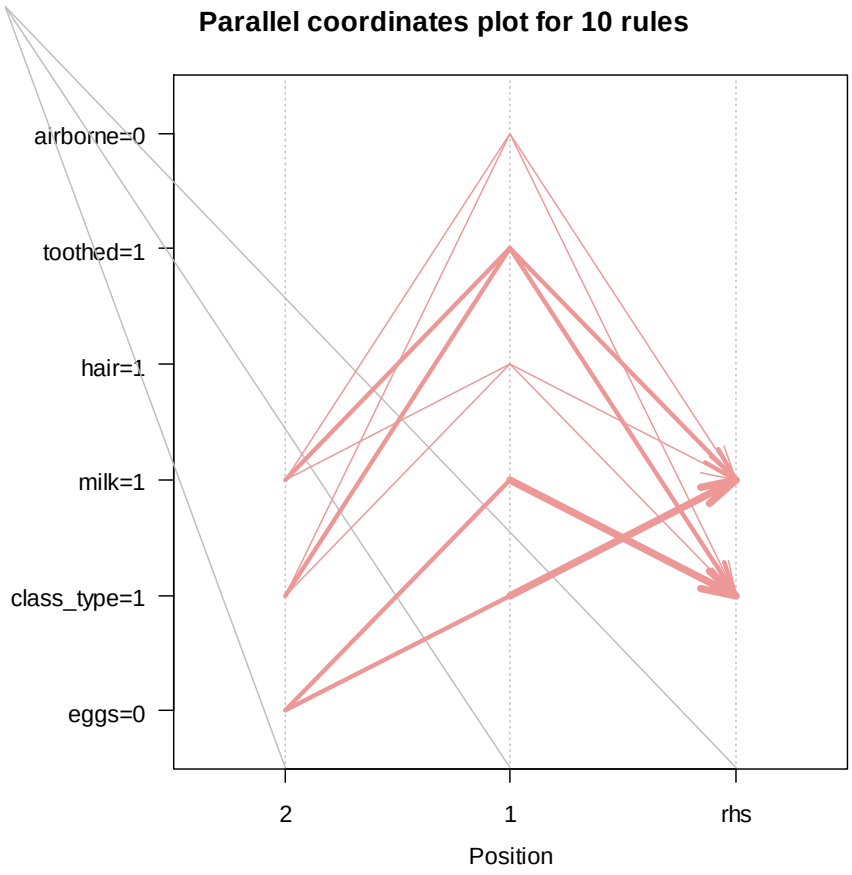


*Ilustración 9. Diagrama paralela según confianza (color) y soporte (ancho).*

Por medio de la Ilustración 9 es posible visualizar que tal como se ha visto durante este documento, la asociación más fuerte dentro del dataset {“no pone huevos”, “mamíferos”, “amamanta”} es la que más ocurrencias tiene entre los datos. Por otra parte, una asociación menos significativa entre los datos es {“mamíferos”, “aéreo”}, debido a que es una minoría dentro de los animales presentes en el dataset. Otro itemset que es posible destacar por su soporte es {“mamífero”, “dentado”}, el cual al igual que los itemset anteriores, determina si un animal pertenece a la especie “mamíferos”.

**Análisis de resultados obtenidos por medio de coordenadas paralelas con respecto a conclusiones de laboratorio anterior.**

De la misma forma que indicaba el laboratorio anterior en cuanto a la relevancia y ocurrencia de datos, las reglas de asociación más significativas son referentes a “mamíferos”. De esta forma, citando uno de los gráficos propuestos para el análisis estadístico e inferencial (Ilustración 10), la característica “eggs = 0” es casi en su totalidad excluyente en la categorización de los animales para la especie “mamíferos”. De esta misma manera, el dataset {“no pone huevos”, “mamífero”, “amamanta”} encontrado a través de la Ilustración 9 propone que tiene una gran relevancia la dimensión “huevo” en la clasificación “mamíferos”. Sin embargo, solo hay tan solo una excepción en el reino animal la cual si está incluida en el dataset. Este animal es el ornitorrinco y es por tal razón que en la gráfica de paralelas es posible visualizar que {eggs=0, class_type=1} hay un leve adelgazamiento de la flecha (soporte) en comparación con el resto de la misma relación de ítems.

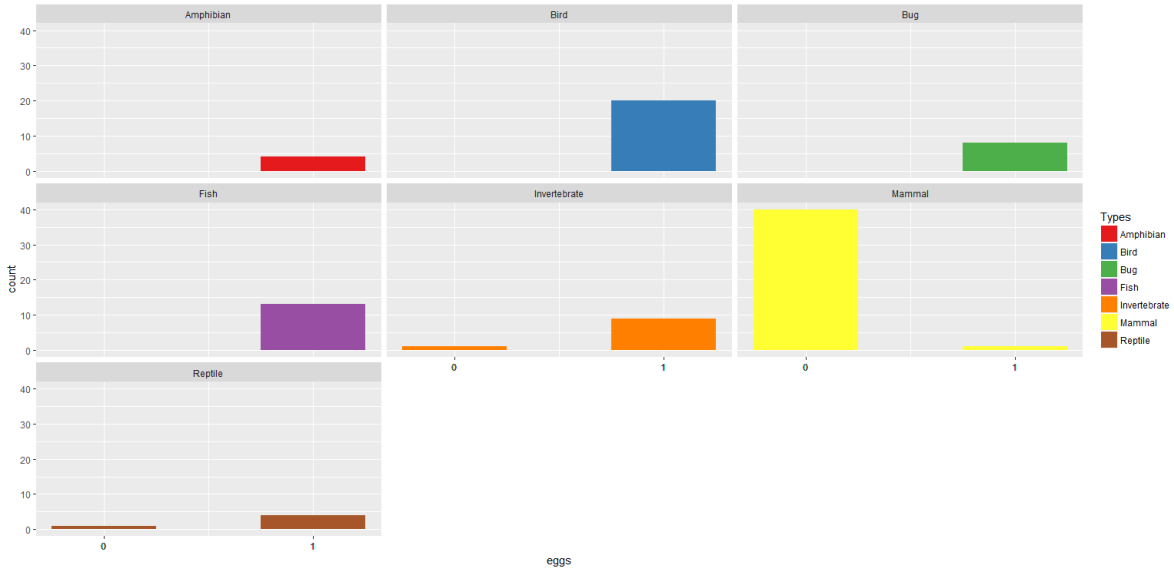


*Ilustración 10. Grafico de agrupación para la dimensión "Huevo".*

**Análisis general.**

Durante el transcurso de los distintos análisis realizados en este laboratorio por medio de una variedad de técnicas de visuales tales como matriz de asociación, diagrama de caja, agrupación de reglas, diagrama de vértices, diagrama de paralelas, entre otras, fue posible determinar algunas relaciones que corroboran información extraída en el laboratorio anterior. Gracias a tales técnicas de visualización, fue posible determinar reglas predominantes por sobre otras las cuales son:

1. {dan de amamantar, mamífero} => {no son acuáticos}
2. {no dan de amamantar, ponen huevos} => {son dentados}
3. {mamífero, no ponen huevos} => {amamanta}
4. {mamífero, posee pelaje} => {amamanta}

Es posible inferir que las reglas más importantes y que tienen un mayor número de ocurrencias y mayor implicancia entre antecedente y consecuente son las relacionadas con los "mamíferos". Esto se puede entender por el número de animales pertenecientes a esa

especie presentes en el dataset (40 mamíferos de un total de 101 animales => 40% probabilidad de que un animal sea mamífero).

Una excepción a la regla de mamíferos la cual es su contraparte o su excluyente, el cual está dada por la regla número 2, en la que se incluyen especies tales como “reptiles”, “anfibios” y “peces”, lo cual se entiende como la probabilidad de que un animal sea de una de estas especies no supera la probabilidad de 60%.

Una regla muy interesante encontrada, la cual tan solo se presenta solo una vez, por lo que su lift bordea el mínimo, es la siguiente:

- {dan de amamantar, pone huevos} => { mamífero }

Lo cual es una excepción muy llamativa. Sin embargo, es posible entenderla por medio de tan solo un animal que reúne esas características, el cual está presente en el dataset. Tal animal es el ornitorrinco, el cual al momento de su clasificación efectuada por George Shaw en 1799, hubieron grandes complicaciones.

# 7 Conclusiones

Al final de las experimentaciones y análisis de los resultados se pudo encontrar reglas de asociación dentro de la base de datos ZOO y también patrones entre los datos y agrupaciones, así pues fue posible determinar las siguientes reglas predominantes:

1. {dan de amamantar, mamífero} => {no son acuáticos}
2. {no dan de amamantar, ponen huevos} => {son dentados}
3. {mamífero, no ponen huevos} => {amamanta}
4. {mamífero, posee pelaje} => {amamanta}

Con lo anterior es posible inferir que las reglas más importantes y que tienen un mayor número de ocurrencias y mayor implicancia entre antecedente y consecuente son las relacionadas con los "mamíferos". Esto se puede entender por el número de animales pertenecientes a esa especie presentes en el dataset (40 mamíferos de un total de 101 animales => 40% es la probabilidad de que un animal sea mamífero).

Una excepción a la regla de mamíferos, la cual es su contraparte o su excluyente, es la dada en la regla 2, en la que se incluyen especies tales como "reptiles", "anfibios" y "peces", esto se entiende como la probabilidad de que un animal sea de una de estas especies no supera el 60%. Esta regla se presenta solo una vez, por lo que su lift bordea el mínimo y se puede entender porque está relacionada a un animal que reúne esas características y que está presente en el dataset, tal animal es el ornitorrinco, el cual al momento de su clasificación efectuada por George Shaw en 1799 tuvo grande complicaciones, mencionado en el artículo de QI Daily (2016).

Por otro lado se pudo afirmar ciertas conclusiones del laboratorio 1 y que están relacionadas con las reglas de asociación obtenidas en este laboratorio. Así pues, en su gran mayoría, el dataset Zoo se compone de animales categorizados como mamíferos, lo cual es sin duda reafirmado por el conjunto de reglas con mayor soporte (tamaño de la esfera) dentro

del diagrama de agrupación. Por consiguiente, es posible mencionar la posibilidad en la relación de la especie “mamíferos” con “amamantar” y en su mayoría, no ser animales “acuáticos”.

# 8 Referencias

1. Giraud-Carrieriu C. & Martinez T. (1995). An Incremental Learning Model For Commosense Reasoning. *Journal of Artificial Intelligence Research*, vol (2) pp. 134-141.

2. Michael H. & Sudheer C. (2012). Visualizing Association Rules: Introduction to the R-extension Package arulesViz. *Journal of Artificial Intelligence Research*.

3. Deshpandel H. & Karypis G. (2002). Using Conjunction of Attribute Values for Classification. 11th Conference of Information and Knowledge Management. pp. 356 – 364.

4. QI Daily (2016). The duck-billed platypus seemed on its first discovery to be a creature just as wonderful as any mermaid. *http://qi.com/infocloud/platypus*

# 9 Apéndice

## Script

```
setwd(choose.dir())#Entregar ruta de trabajo

#Instala paquete "aruleViz"
install.packages('arulesViz', dep = TRUE)
library(arulesViz)

data  =  read.csv("zoo.csv") #Carga dataset

summary(data) #Resumen estadístico

data <- data.frame(sapply(data,as.factor)) #Discretiza datos

rules <- apriori(data, parameter=list(support=0.37, confidence=0.5, maxlen=50))#5736 Encontradas

inspect(head(sort(rules, by ="lift"),3))#Tres primeras reglas

plot(rules, measure=c("support", "lift"), shading="confidence")#Diagrama de dispersión

inspect(head(sort(rules, by ="lift"),260)) #Visualizar reglas para seleccionar


subrules <- rules[quality(rules)$lift > 1.46]#Visualizar reglas que tengan más de "1.46" lift

subrules #Número de reglas obtenidas: 1928

#Matrices de reglas
plot(subrules, method="matrix", measure="lift", control=list(reorder=TRUE))
plot(subrules, method="matrix3D", measure="lift", control=list(reorder=TRUE))
plot(subrules, method="matrix3D", measure="support", control=list(reorder=TRUE))
plot(subrules, method="matrix", measure=c("lift", "confidence"),control=list(reorder=TRUE))

#Matriz de agrupación (5.736 reglas). Siete agrupadores de reglas en eje "X".
plot(rules, method="grouped", control=list(k=7))

subrules2 <- head(sort(rules, by="lift"), 10)#Sub Set de 10 reglas.

plot(subrules2, method="graph")#Grafico de vértices para diez reglas.

#Grafico de coordenadas paralelas para diez reglas.
plot(subrules2, method="paracoord", control=list(reorder=TRUE))
```