UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERIA

DEPARTAMENTO DE INGENIERIA EN INFORMATICA

Análisis de Arboles de Decisión para la Base de Datos ZOO Utilizando la Herramienta de Software “R”

LUIS ORELLANA

1 Resumen

Este trabajo de investigación consiste en estudiar e interpretar mediante arboles de decisión, las clasificaciones y relaciones con el etiquetado originales (atributo experto) de 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 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 el marco teórico, la base de datos ZOO, su normalización y la presentación de las técnicas aplicadas al análisis del algoritmo C5.0 de J. R. Quinlan (1997), 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, algoritmo C5.0, ganancia e información de atributos, base de datos Zoo, software "R", poda de árbol de decisión, matriz de confusión.

2 Objetivos

2.1 Objetivos Generales

El objetivo general de este trabajo de investigación es la recuperación, análisis y comparación de conocimiento por medio de árboles de decisiones con respecto a reglas de asociación visto en el laboratorio anterior. Con el objetivo de recuperar información desde el dataset "Zoo", se ha empleado el paquete "C50" del software R. Por otra parte, es necesario mencionar que por medio de árboles de decisión se busca clasificar a animales por medio de sus características morfológicas y de esta, forma contrastar tales clasificaciones predictivas con las reales clasificaciones del muestreo de animales.

2.1 Objetivos Específicos

Los objetivos específicos planteados para este trabajo de investigación son:

- Analizar y normalizar la base de datos ZOO.

- Investigar y estudiar el algoritmo C5.0 para confeccionar arboles de decisión y reglas de asociación, como también las funciones de evaluación de métodos de agrupamiento, matriz de confusión y su implementación con el software "R".

- Realizar análisis deductivo a partir de los resultados del árbol de decisión y reglas de asociación obtenidas, de esta forma contrastar tal análisis con las hipótesis y/o problema planteado a solucionar en el desarrollo del estudio, además de la comparación en relación a los resultados del análisis de reglas de asociación del laboratorio 2.

3 Descripción del Problema

3.1 Motivación

El reino animal, con su infinidad de formas, características y cualidades es un área que ha sido de gran interés para diferentes ciencias durante la historia de la humanidad. Sin embargo es a la vez un lugar al cual queda mucho por conocer. Por tal motivo, por medio de este laboratorio se intentara realizar clasificaciones de animales que van más allá de lo obvio y banal. Por tal motivo, se busca clasificar animales por características que por medio de análisis no estadísticos sería una tarea muy compleja, sino infructuosa. A través del análisis de árboles de decisiones, como también reglas de asociación se pretende lograr obtención de información no banal, es decir, características de animales las cuales tienen en común una o varias especies de forma transversal, las cuales serán analizadas a fondo a medida que el informe prosigue.

3.2 Literatura Relevante

Tal como se ha mencionado anteriormente, para la obtención de conocimiento relevante relacionada al dataset "Zoo", se empleara un algoritmo propuesto por Quinlan (1997), el cual fue llamado "C5.0". Este modelo se basó principalmente en los trabajos de Hoveland y Hunt en 1950, el cual inspiro su trabajo en modelos psicológicos de la manera en que se efectúa el proceso de aprendizaje. Por otra parte, el análisis del conocimiento adquirido en este laboratorio, y contrastado con laboratorios anteriores es fundamentado en artículos científicos tales como el afamado libro "el origen delas especies" de Darwin C. (1859), o el trabajo de Bickler P. (2007) el cual menciona caracteristicas morfologicas entre las especies reptil, anfibios y pesces. De esta forma, se intenta contrastar el conocimiento recuperado, y hallazgos cientificos en el area de las ciencies de la biologia como tambien documentos de mineria de datos que avalan y da soporte a tal conocimiento y los procedimientos para conseguirlo.

3.3 Definición del Problema

Este laboratorio se centra en abalar la hipótesis de clasificación de animales por medio de las características morfológicas de las especies. Ante esto cabe destacar que el problema en cuestión radica en encontrar y demostrar un modelo óptimo para realizar tal clasificación. Por otra parte, el conocimiento que se recupere del dataset "Zoo" debe ser consecuente con la información bibliográfica que se ha encontrado. Para el caso de este laboratorio, se hará uso del algoritmo C5.0, sin embargo es necesario variar los parámetros de manera de obtener de la manera más asertiva, las características más relevantes de clasificación de especies.

Así pues, sería de mucho interés responder algunas preguntas como:

- ¿Cuáles son las características principales que permiten clasificar un animal dentro de una determinada especie u otra?
- ¿Cuáles son las reglas más representativas para avalar y probar la hipótesis?
- ¿Los parámetros escogidos las l algoritmo C5.0 son los óptimos para encontrar morfologías relevantes?

4 Descripción de la Solución Propuesta

Con la finalidad de dar solución al problema propuesto (el cual consiste en clasificar animales según sus características morfológicas en las respectivas especies a las cuales pertenecen), se ha establecido la utilización de árboles de decisión y reglas de asociación obtenidas de este árbol, empleando para ello el software R y el algoritmo C5.0 el cual permite realizar .

Para la utilización de este algoritmo, se ha utilizado los paquetes del software R "C5.0", "PartyKit" y "GModels" los cuales permiten crear arboles de decisión, validar el modelo por medio de la bondad visualizada en matriz de confusión y la visualización grafica de tanto los arboles como de las reglas de relación recogidas desde el árbol.

4.1 Características de la Solución.

La solución contempla el uso del algoritmo "C5.0" y la experimentación de las medidas de parámetros relacionados con este algoritmo (tales como el porcentaje de datos contenidos en el grupo de prueba y el de entrenamiento, la cantidad de iteraciones de cálculo de información y ganancia de atributos) para luego medir la bondad del modelo por medio de matriz de confusión. Debido a que la muestra de animales para el dataset "Zoo" no es grande, se ha escogido la validación cruzada de k iteraciones. De esta forma, se pretende cubrir el problema planteado y por consiguiente extraer conocimiento que no es posible inferir o extraer de forma natural.

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 los atributos más relevantes del dataset para caracterizar 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 2 referentes a las reglas de asociación y contrastarlas con las reglas obtenidas desde el árbol de decisión.

4.3 Alcances y Limitaciones de la Solución.

Este laboratorio contempla sólo el análisis y extracción de reglas desde el árbol de decisión como también el análisis de la bondad del modelo para la base de datos ZOO utilizando la herramienta de software “R” y su librería “C5.0” con sus métodos integrados. El laboratorio no contempla realizar análisis a la base de dato ZOO con otros métodos de análisis de datos y/o minería de datos.

5 Metodología, Herramientas y Experimentación

5.1 Marco Teórico

Validación Cruzada

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar cómo de preciso es un modelo que se llevará a cabo a la práctica. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados.

Validación Cruzada de K Iteraciones

En la validación cruzada de K iteraciones o K-fold cross-validation los datos de muestra se dividen en K subconjuntos. Tal como se muestra en ilustración 1, uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos.

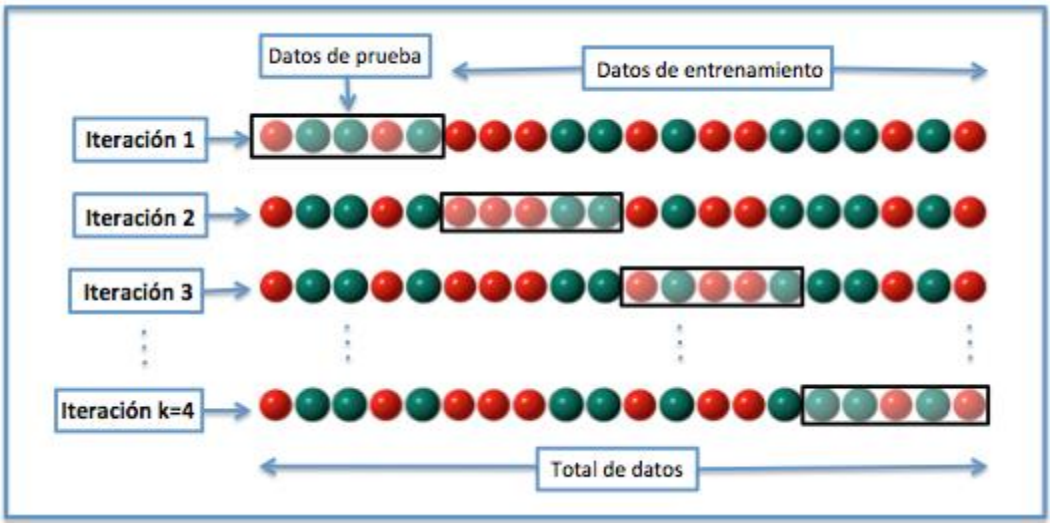

Ilustración 1. Método de validación cruzada para k iteraciones.

Los errores del modelo se calculando según la ponderación de la media aritmética de las sumas de los errores de cada iteración, tal que:

$$\mathrm{E} = \frac{1}{K}\sum_{k=1}^{n} E_i$$

Arboles de decisión

Entre los diversos tipos de aprendizaje que es posible diferenciar, destaca el aprendizaje inductivo, que se basa en el descubrimiento de patrones a partir de ejemplos. Desde el punto de vista de la clasificación, la tarea general consiste en lo siguiente:

Se dispone de N ejemplos observados en el mundo real, {e1,…,eN}{e1,…,eN}, que vienen definidos a partir de un conjunto de atributos (propiedades), ei=(pi1,…,pim)ei=(pi1,…,pim), y para cada uno de ellos tenemos una clasificación observada, cici. La tarea del aprendizaje inductivo consiste en inducir de los datos anteriores un mecanismo que nos permita inferir las clasificaciones de cada uno de los ejemplos a partir únicamente de las propiedades. Si esto es posible, se podría utilizar

este mecanismo para deducir la clasificación de nuevos ejemplos habiendo observado únicamente sus propiedades.

Entre todos los posibles mecanismos para obtener estas predicciones de manera fiable, una de las que más destaca es la creación de árboles de decisión, que proporcionan un conjunto de reglas que se van aplicando sobre los ejemplos nuevos para decidir qué clasificación es la más adecuada a los atributos.

Ilustración 2. Ejemplo árbol de condiciones para categorizar una instancia como posibilidad de jugar o no al golf.

Un árbol de decisión está formado por un conjunto de nodos de decisión (interiores) y de nodos-respuesta (hojas):

- Un nodo de decisión está asociado a uno de los atributos y tiene 2 o más ramas que salen de él, cada una de ellas representando los posibles valores que puede tomar el atributo asociado. De alguna forma, un nodo de decisión es como una pregunta que se le hace al ejemplo analizado, y dependiendo de la respuesta que dé, el flujo tomará una de las ramas salientes.
- Un nodo-respuesta está asociado a la clasificación que se quiere proporcionar, y nos devuelve la decisión del árbol con respecto al ejemplo de entrada.

En la ilustración 2 se muestra la tabla de ejemplos que hemos observado a lo largo del tiempo respecto a las condiciones meteorológicas y la posibilidad de jugar al golf o no. El

árbol de la derecha muestra un posible mecanismo aprendido para poder tomar decisiones para esta tarea de clasificación. Observa en el árbol está formado únicamente por los nodos azul oscuro (nodos de decisión) y los de color amarillo (nodos-respuesta), mientras que los rectángulos azules claro son simplemente las etiquetas de las ramas de salida de cada nodo de decisión, indicando cuál es la opción que verifica nuestro ejemplo.

Es evidente que no siempre será posible conseguir un árbol de decisión que sea capaz de predecir los ejemplos con una fiabilidad del 100%, pero cuanto mejor sea la batería de ejemplos de los que se disponen (por ejemplo, que no haya contradicciones entre clasificaciones), mejor se comportará el árbol que es construido a partir de ellos.

Para determinar cuan relevante es un atributo estudiante en comparación con el atributo experto, se deben calcular las probabilidades de cada una de las instancias del atributo estudiante en relación a la clase y luego determinar su información o entropía la cual se expresa como:

$$\text{Entropia}(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

Dónde:
S: es una colección de objetos
Pi : es la probabilidad de los posibles valores
i: las posibles respuestas de los objetos

Por otro lado, también de debe determinar la ganancia del atributo estudiante en relación al atributo experto, donde se expresa como la resta de la información o entropía del atributo experto menos la información del atributo estudiante:

$$\text{Gan Inf}(S, A) = \text{Entropia}(S) - \sum_{v \in V(A)} \frac{|Sv|}{|S|} \text{Entropia}(Sv)$$

Dónde:

S : es una colección de objetos

A : son los atributos de los objetos

V(A) : Conjunto de valores que A puede tomar

Algoritmo C5.0

C5.0 construye árboles de decisión desde un grupo de datos de entrenamiento, usando el concepto de entropía de información. Los datos de entrenamiento son un grupo S = s1, s2, … de ejemplos ya clasificados. Cada ejemplo si = x1, x2,… es un vector donde x1, x2,… representan los atributos o características del ejemplo. Los datos de entrenamiento son aumentados con un vector C=c1, c2,… donde c1, c2,… representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C5.0 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. El algoritmo C5.0 divide recursivamente en sublistas más pequeñas.

Este algoritmo tiene unos pocos casos base.

- Todas las muestras en la lista pertenecen a la misma clase. Cuando esto sucede, simplemente crea un nodo de hoja para el árbol de decisión diciendo que elija esa clase.

- Ninguna de las características proporciona ninguna ganancia de información. En este caso, C5.0 crea un nodo de decisión más arriba del árbol utilizando el valor esperado de la clase.
- Instancia de la clase previamente no vista encontrada. Una vez más, C5.0 crea un nodo de decisión más arriba en el árbol con el valor esperado.

En pseudocódigo, el algoritmo general para construir árboles de decisión es:2

1. Comprobar los casos base
2. Para cada atributo a
    1. Encontrar la ganancia de información normalizada de la división de a
3. Dejar que a_best sea el atributo con la ganancia de información normalizada más alta
4. Crear un nodo de decisión que divida a_best
5. Repetir en las sublistas obtenidas por división de a_best, y agregar estos nodos como hijos de nodo

Matriz de Confusión

Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

Si en los datos de entrada el número de muestras de clases diferentes cambia mucho la tasa de error del clasificador no es representativa de lo bien que realiza la tarea el clasificador. Si por ejemplo hay 990 muestras de la clase 1 y sólo 10 de la clase 2, el clasificador puede tener fácilmente un sesgo hacia la clase 1. Si el clasificador clasifica todas las muestras

como clase 1 su precisión será del 99%. Esto no significa que sea un buen clasificador, pues tuvo un 100% de error en la clasificación de las muestras de la clase 2.

5.2 Descripción de la Base de Datos ZOO (Dataset ZOO)

El Dataset ZOO cuenta con 101 animales extraídos desde una colección Zoológica. Hay 23 variables con una serie de rasgos que describen a los animales. Se definen 7 Tipos de Clases: Mamíferos, Aves, Reptiles, Peces, Anfibios, Insectos (voladores) e Invertebrados.

Originalmente, las características disponibles para esta dataset era un conjunto de 16 morfologías, las cuales son dicotómicas, excepto la cantidad de patas las que pueden ir desde 0 hasta 8 patas. Sin embargo, con el propósito de poder disponer de datos normalizados, se divide esta variable en 6 distintas variables, las que al igual que las demás variables de este dataset, son binarias.

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-arbol.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 son:

- 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_0 ∈ {0, 1}: No posee patas (si, no)
- legs_2 ∈ {0, 1}: Posee dos patas (si, no)
- legs_4 ∈ {0, 1}: Posee cuatro patas (si, no)
- legs_5 ∈ {0, 1}: Posee cinco patas (si, no)
- legs_6 ∈ {0, 1}: Posee seis patas (si, no)
- legs_8 ∈ {0, 1}: Posee ocho patas (si, no)
- 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 (volador)
- 7 = Invertebrado

5.2.1 Herramientas de Software

La herramienta de software utilizada para realizar los experimentos presentados en este documento es “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 paquetes de software dentro de “R”, se utilizó “C50” y “Caret” , los cuales permiten crear arboles binarios, reglas, y evaluar los árboles de decisión generados con la finalidad de realizar inferencias de características comunes a cada clasificación, además, estos paquetes permiten la visualización de resultados en reportes, gráficos y tablas.

5.3 Funciones de los Paquetes de Software R y Pre procesamiento

Funciones de los Paquetes de Software

Los siguientes comandos en "R" fueron utilizados para realizar el análisis de los árboles de decisión generados por medio de los modelos basados en reglas y árboles de decisión C50:

- 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_tree)

Esta función despliega resúmenes detallados para los modelos C5.0. Los nodos terminales tienen texto que indica el número de muestras cubiertas por el nodo y los números que fueron clasificados incorrectamente.

- data_tree <- C.5(type~.,data=data_zoo)

Esta función extiende los algoritmos de clasificación C4.5 descritos en Quinlan (1992), adopta la forma de un árbol de decisión o una colección de reglas.

- C5imp(data_tree)

Esta función calcula la importancia de la variable durante la división del árbol de decisión (también conocido como atributo) para los modelos C5.0.

- Plot(data_tree)

Esta función traza un árbol de decisión

- Predict(data_tree, newdata = NULL, trials = object$trials["Actual"], type = "class", na.action = na.pass, ...)

Esta función produce clases predictivas o valores de confianza de un modelo C5.0.

Pre procesamiento

Tal como se ha mencionado anteriormente, el dataset ZOO consta de un total de 101 animales con 16 características morfológicas. Todas esas características son descriptivas, es decir todos los valores disponibles son binarios, excepto la variable “patas” con valores desde 0 hasta 8.

Con la finalidad de poder normalizar todos los datos de la base de datos a un tipo en particular (binarias para este dataset), se separó la variable “patas” en 6 distintas variables, las cuales son transformadas a descriptivas, de esta forma, tener una base de datos con solo valores descriptivos.

Así pues, se eliminó las variable "animal_name", las cual cumple la función de individualizar a cada uno de los miembros (animales) del dataset:

De esta forma, solo se dispondrá de una matriz apropiada para aplicar el algoritmo C50 según sea el caso, las cuales cuentan con 22 atributos estudiantes y 1 atributo experto (class_type) que serán analizadas más a fondo en la siguiente sección.

Experimentación y Análisis de Resultados

Conforme a los resultados obtenidos del algoritmo C5.0, se pudo determinar en la tabla 1 que muestra las variables que caracterizan mejor a las clases de especie. Así pues, se puede inferir que el atributo milk=si (amantan) es la que mejor caracteriza a la clase 1 (mamíferos), además es la de mayor peso (cantidad de miembros) y confianza 100% (% uso del atributo), por lo tanto, es considerada la raíz del árbol. Cabe destacar que este atributo en particular obtuvo un mayor índice de ganancia con respecto a los demás atributos estudiantes. Por otro lado, el atributo Feathers=si (poseen plumas) caracteriza de mejor forma a la clase 2 (aves) con 20 miembros y un porcentaje de uso del atributo cercano al 60%, así este atributo encabeza la rama derecha del árbol. La variable Backbone=si (posee espina dorsal) es la que mejor caracteriza la clase 3 (reptil) con 5 miembros y un porcentaje del uso del atributo cercano al 40%,, también es caracterizada la clase en menor medida por la variable Tail= (posee cola) con 5 miembros y un porcentaje de uso del atributo cercano al 9%. La variable Fins=si (posee aletas) es la que mejor caracteriza la clase 4 (pez) con 13 miembros y un porcentaje del uso del atributo cercano al 22%. La variable Tail =1 (posee cola) es la que mejor caracteriza la clase 5 (anfibios) con 3 miembros y un porcentaje de uso del atributo cercano al 9%. La variable airbone=si (puede volar) es la que mejor caracteriza la clase 6 (insectos) con 6 miembros y un porcentaje de uso del atributo cercano al 18%. La variable predator=si (es predador) es la que mejor caracteriza la clase 7 (invertebrados) con 8 miembros y un porcentaje cercano al 12% de uso del atributo.

Tabla 1. Variables que caracterizan a la clase en el árbol de decisión.

Variables	Descripción	Miembros	Clase	% Uso del Atributo
Milk=1	Amamantan	41	Mamíferos	100%
Feathers=1	Poseen plumas	20	Aves	59.41%
Backbone=1	Posee espina dorsal	5	Reptil	39.60%
Fins=1	Posee aletas	13	Pez	21,78%
Airbone=1	Puede Volar	6	Insecto	17.82%
Predator=1	Es predador	10	Invertebrado	11.88%
Tail = 1	posee cola	5	Reptil	8.91%

La evaluación sobre el entrenamiento de los datos (101 casos) indica que el tamaño del árbol es de 9 hojas, contiene 1 caso mal clasificado y un porcentaje de error de mala clasificación del 1%, lo cual infiere en primera instancia que los atributos que mejor representan a las clases son los más representativos de acuerdo a lo entregado por el modelo.

Para analizar los tipos de errores se verifica la matriz de confusión indicada en la tabla 2. La cual muestra que la predicción para más del 99% de los casos corresponde a la clasificación correcta, el caso en donde no existe 100% de predicción es para la clase 5 (anfibio), en donde de un total de 4 casos uno no fue bien clasificado y correspondería a la clase reptil, tal es el caso de la rana, pues tiene en su mayoría las características de un reptil, pero no posee cola como en el caso de los anfibios.

Este margen de error en el que algunos de los integrantes de las especies reptil y anfibio fueron mal clasificados, puede ser entendido por medio de la investigación realizada por Randall(1981), el que sugiere que muchas de las características en común que

poseen estas dos especies, son debido a que la secuencia de evolución fue bastante directa desde los peces a los anfibios, ya que desde que el primer pez anfibio que se aventuró en ambiente no acuático (rhipidistian crossopterygians) hace 350 millones de año, señalo un punto de quiebre entre estas dos especies, lo cual sin embargo represento un punto de unión entre estas dos especies las que es posible ver incluso en la actualidad. Es por tal razón que estas dos clasificaciones están muy unidas entre sí.

Tabla 2. Matriz de confusión. Error del 1% para la clase anfibio.

(a)	(b)	(c)	(d)	(e)	(f)	(g)	← Clasificado como
41							(a): clase 1
	20						(b): clase 2
		5					(c): clase3
			13				(d): clase 4
		1		3			(e): clase 5
					8		(f): clase 6
						10	(g): clase 7

Conforme al método aplicado, se pudo construir el árbol de decisión general presentado en la ilustración 1, desde un conjunto de 21 clasificadores que consisten de 101 ejemplos, pertenecientes a 7 diferentes clases {Mamíferos, Aves, Reptiles, Peces, Anfibios, Insectos, Invertebrados} indicadas por el atributo experto type_class (especie). De los resultados presentados por el árbol de decisión, se puede inferir, al igual que lo visto anteriormente que el atributo amantar es el más predomínate y, por lo tanto, la raíz del árbol. Por otro lado, de la ilustración 2, se puede inferir los miembros de cada atributo y el peso de cada atributo sobre la clase correspondiente.

Ilustración 1: Árbol de Decisión general

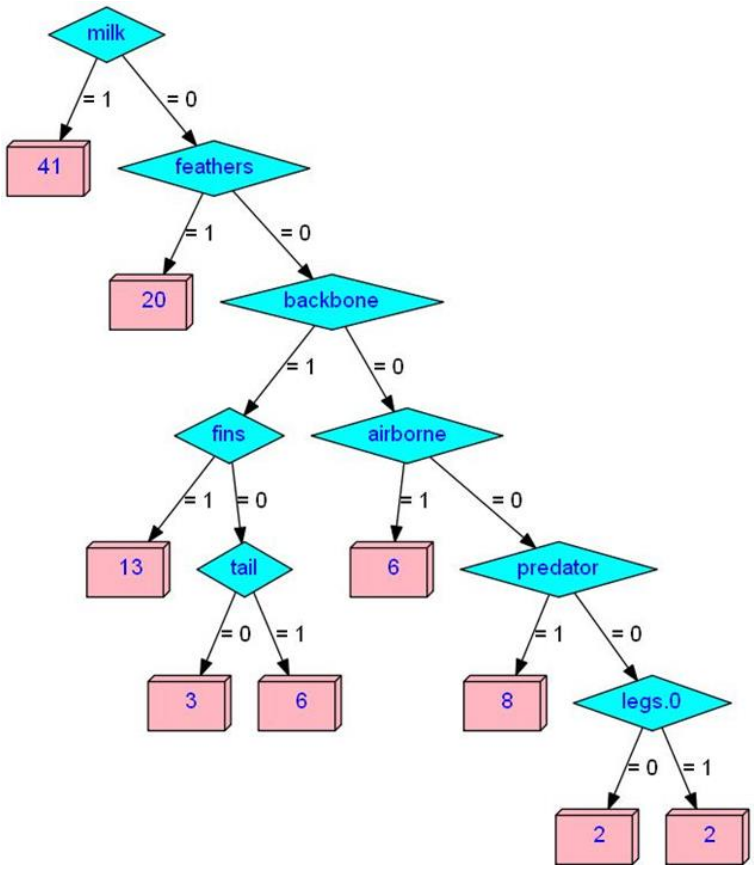

Ilustración 2: Arbol de Decisión y atributos característicos

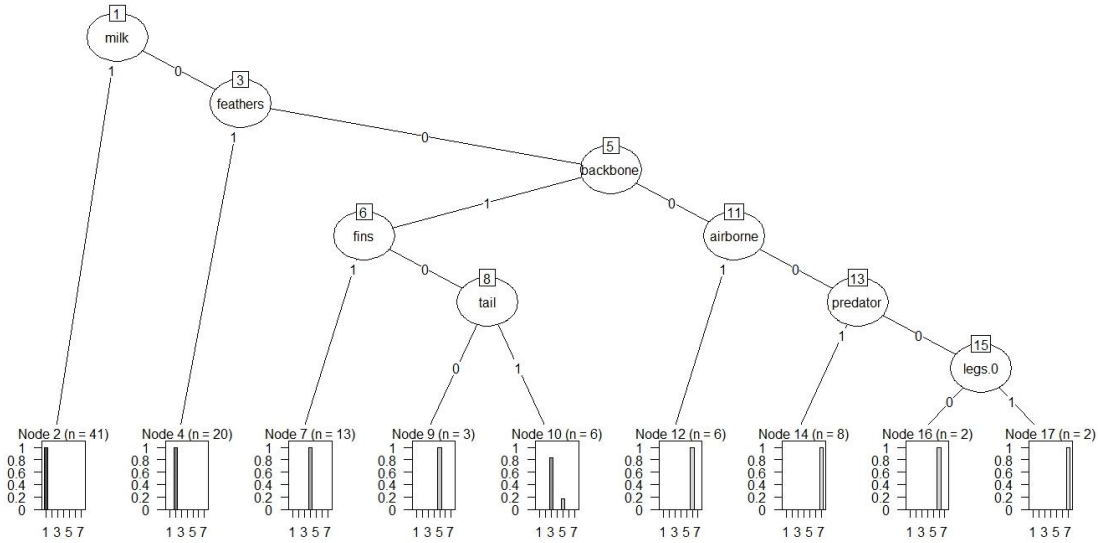

Tabla 3. Uso del atributo.

Atributo	Porcentaje
milk	100.00%
feathers	59.41%
backbone	39.60%
fins	21.78%
airborne	17.82%
predator	11.88%
tail	8.91%
legs.0	3.96%

Para este trabajo, especialmente por la cantidad de atributos, es útil saber cómo contribuyen los atributos individuales al clasificador. Esto se muestra en la tabla 3, en donde el atributo amamantar (100%) es la más importante variable de división del árbol de decisión, seguido en menor grado de la variable plumas (59,4%) y en tercer orden la variable posee espina dorsal (39,6%), luego en cuarta posición la variable posee aletas (21,78%) y con medidas cercanas la variable poder volar (17,82%). Todas las anteriores variable no sólo permite la división directa de árbol sino también son la que representan por sí solas a la clase de especie mamíferos, aves, reptil, pez e insecto. Por otro lado, las variables que también permite dividir el árbol están ser predador (11,88%), posee cola (8,91%) y tener o no patas (3,96%)

Reglas

Modelo Basado en Reglas

Conforme al método aplicado se pudo construir un conjunto de reglas desde el árbol de decisión que son presentadas más abajo. Desde un conjunto de 21 predictores que consiste de 101 ejemplos, se pudieron generar 8 reglas, las cuales se analizan a continuación:

Regla 1: (41, lift 2.4)

{milk = 1 => clase 1 (mamíferos) }

Para esta regla se tiene 41 casos de entrenamiento cubiertos por la regla, con confianza = 0,977 y Lift = 2.4, esto implica, que la confianza es casi 1 y muestra que prácticamente todos los casos pertenecen a la clase predicha por la regla, por lo anterior, si una especie amamanta implica que pertenece a la clase mamíferos.

Regla 2: (20, lift 4.8)

{ feathers = 1 => clase 2 (aves) }

Para esta regla se tiene 20 casos de entrenamiento cubiertos por la regla, con confianza = 0,955 y Lift = 4.8, esto implica, que la confianza es 1 (alto) y muestra que todos los casos pertenecen a la clase predicha por la regla, por lo anterior, si una especie tiene plumas implica que pertenece a la clase aves.

Regla 3: (6/1, lift 15.1)

{ feathers = 0, milk = 0, backbone = 1, fins = 0, tail = 1 => clase 3 (Reptil) }

Para esta regla se tiene 6 casos de entrenamiento cubiertos por la regla y 1 caso no pertenece a la clase predicha por la regla, con confianza = 0,750 y Lift = 15.1, esto implica,

que la confianza cercana a 0,76 (relativamente alto) y muestra que de 6 casos predichos por la regla, un caso no pertenece a la clase predicha por la regla , por lo anterior, si una especie no tiene plumas, no amamanta, tiene espina dorsal, posee aletas y posee cola implica que la probabilidad de pertenencia a la clase reptil es alta. La excepción es la rana que cumple con todas las condiciones excepto que no posee cola. Se destaca que en la base de datos ZOO existen 2 instancia de rana, cuya diferencia es poseer o no cola, por ello es que para el modelo es clasificar la rana a una u otra clase.

Regla 4: (13, lift 7.3)

{ milk = 0, fins = 1 => clase 4 (Pez) }

Para esta regla se tiene 13 casos de entrenamiento cubiertos por la regla, con confianza = 0,933 y Lift = 7.3, esto implica, que la confianza es 1 (muy alta). Por lo tanto, si una especie no amamanta y tiene alas implica que pertenece a la clase aves.

Rule 5: (3, lift 20.2)

{ milk = 0, backbone = 1, tail=0 => clase 5 (Anfibio) }

Para esta regla se tiene 3 casos de entrenamiento cubiertos por la regla, con confianza = 0,8 y Lift = 20.2, esto implica, que la confianza es alta aun cuando su soporte no lo es. Para estos casos si una especie no amamanta, tiene espina dorsal y no posee cola implica que la pertenencia a la clase anfibio es muy alta.

Rule 6: (7, lift 11.2)

{ predator = 0, backbone = 0, legs.0=0 => clase 6 (Insecto) }

Para esta regla se tiene 7 casos de entrenamiento cubiertos por la regla, con confianza = 0,88 y Lift = 20.2, esto implica, que la confianza es alta aun cuando su soporte no lo es. Para estos casos si una especie no es predador, no posee espina dorsal y no posee patas implica que la pertenencia a la clase insecto es muy alta. Ahora bien, en esta regla la única característica cierta para un insecto es que no posee espina dorsal, pues todos los insectos posee patas e incluso hay algunos insecto que son depredadores, por lo tanto, a través del conocimiento previo de la clasificación insecto esta regla no aplicaría. Para analizar esta regla en más profundidad se realizó un comparativo con la regla 7, que también posee condicionantes para clasificar la clase insecto.

Rule 7: (6, lift 11.0)

{ feathers = 0, milk = 0, airbone=1 => clase 6 (Insecto) }

Para esta regla se tiene 6 casos de entrenamiento cubiertos por la regla, con confianza = 0,87 y Lift = 11, esto implica, que la confianza es alta. Para estos casos si una especie no posee plumas, no amamanta y puede volar implica que la pertenencia a la clase insecto es muy alta. Ahora bien, esta regla al igual que la regla 6 está relacionada con la clase 6. Para determinar cual es más representativa para la clase 6, creemos una forma sería a través del análisis del árbol de decisión visto anteriormente, esto es, utilizar como base la característica airbone (capacidad de volar), que para nuestro análisis la regla 7 la posee, por lo tanto, la regla 7 sería la más representativa para determinar las condiciones que se deben de dar con la clase 6.

Rule 8: (18/8, lift 5.6)

backbone = 0

-> class 7 [0.550]

{ backbone = 0 => clase 7 (Invertebrado) }

Para esta regla se tiene 18 casos de entrenamiento cubiertos por la regla, con confianza = 0,55 y Lift = 5.6, esto implica, que la confianza es baja. Por lo anterior, según esta regla, si especie no posee espina dorsal implica que la pertenencia a la clase invertebrado es media. La baja confianza se puede interpretar a que existe otra especie que no posee espina dorsal, tal es el caso de los insectos

En resumen considerando las reglas obtenidas del árbol de decisión se tiene que la regla de mayor importancia, considerando como parámetro la de mayor confianza, es la regla uno y que está relacionada con el atributo amamantar y mamíferos, la cual también es la raíz del árbol de decisión.

Reglas de Asociación del Laboratorio 2.

Las 15 reglas de asociación presentadas en la tabla 3, generadas en el laboratorio 2, representan a las distintas especies Para ello se puso énfasis en recuperar reglas representativas de cada especie y que fueran predominantes por medio del soporte y la confianza de estas, y por último la no menos importante el lift.

Tabla 3. Conjunto de reglas (15) según soporte, confianza y lift.

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

En la tabla 4 están categorizadas las diferentes reglas con sus respectivas clases o especies de animales, las cuales pueden categorizar animales según especies. Cada una de las reglas pueden aplicar tanto a una sola especie, como a un conjunto teniendo en cuenta que se ha considerado el conocimiento adquirido en el primer laboratorio, el cual indicaba la predominancia de características morfológicas para algunas especies por medio del análisis de la estadística descriptiva.

Tabla 4. Conjunto de reglas categorizadas por especie.

Reglas	Mamífero	Ave	Reptil	Pez	Anfibio	Invertebrado	Insecto
{hair=1,feathers=0,eggs=0, airborne=0,breathes=1}=> {milk=1}	X						
{hair=1,milk=1,breathes=1, legs=4} => {class_type=1}	X						
{backbone=1,venomous=0,milk=1}=> {eggs=0}	X						
{hair=0,eggs=1,milk=0}=> {domestic=0}		X	X	X	X	X	X
{hair=0, eggs=1, milk=0} => {feathers=0}			X	X	X	X	X
{hair=0, eggs=1,feathers =0} => {breathes=1}				X	X	X	X
{eggs=1, aquatic=1} => {hair=0}				X	X		
{feathers=0,eggs=1, airborne=0, catsize=0} => {milk=0}			X	X	X	X	
{eggs=1, backbone=1, tail=1} => {fins=0}		X	X				
{eggs=1,fins=0} => {feathers=0}			X		X	X	X

{eggs=1, milk=0, fins=0} => {backbone=1}	X	X				
{hair=0,feathers=0, airborne=0}=> {backbone=0}					X	X
{eggs=1, backbone=0} => {aquatic=0}					X	X
{eggs=1, backbone=0, airborne=1} => {aquatic=0}						X
{eggs=1, backbone=1} => {milk=0}	X					
{eggs=1, backbone=0, airborne=0} => {aquatic=0}					X	

Comparación Reglas del Árbol de Decisión V/S Reglas de Asociación Laboratorio 2.

El análisis comparativo y/o relación entre las reglas generadas por las reglas de asociación del laboratorio 2 y las regles generadas mediante el árbol de decisión, fueron analizadas para cada clase de especie de animal, y se detallan a continuación.

Clase Mamíferos:

Regla Árbol de Decisión:

- {milk = 1 => clase 1 (mamíferos) }

Reglas de Asociación:

- {hair=1,feathers=0,eggs=0, airborne=0,breathes=1}=> {milk=1}
- {backbone=1,venomous=0,milk=1}=> {eggs=0}
- {hair=1,milk=1,breathes=1, legs=4} => {class_type=1}
- {eggs=1,fins=0} => {feathers=0}

Se aprecia que las reglas de ambos métodos coinciden en la variable más predomínate “milk”, esto es, amamantar. Si bien la regla del árbol de decisión clasifica en forma directa a la especie mamíferos a través de la variable amantar, no así las reglas de asociación, pues inicialmente toman las características de las otras variables para concluir cuando un animal amamanta, luego otra regla infiere cuando una especie no nace por huevos, y finalmente recién la última regla clasifica a la especie mamífero considerando las variables amamanta=si, tiene pelos=si, pulmones y poseen 4 patas=si.

Clase Aves:

Regla Árbol de Decisión:

- { feathers = 1 => clase 2 (aves) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0} => {domestic=0}
- {eggs=1, backbone=1, tail=1} => {fins=0}
- {eggs=1, milk=0, fins=0} => {backbone=1}
- {eggs=1, backbone=1} => {milk=0}

Se aprecia que la regla del árbol de decisión clasifica en forma directa a la especie aves a través del atributo feathers (poseen plumas), no así las reglas de asociación, pues en base a los consecuentes en su conjunto podrían definir la clase, esto es, no amantan, tienen espina dorsal, no poseen aletas y no son domésticos, además, del antecedente que todos nacen de huevos. En resumen, no se menciona en ninguna regla de asociación la variable poseen plumas tanto en el antecedente como consecuente, la cual por si sólo podría definir la clase ave.

Clase Reptil:

Regla Árbol de Decisión:

- { feathers = 0, milk = 0, backbone = 1, fins = 0, tail = 1 => clase 3 (Reptil) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0}=> {domestic=0}
- {hair=0, eggs=1, milk=0} => {feathers=0}
- {feathers=0,eggs=1, airborne=0, catsize=0} => {milk=0}
- {eggs=1, backbone=1, tail=1} => {fins=0}
- {eggs=1,fins=0} => {feathers=0}

- {eggs=1, milk=0, fins=0} => {backbone=1}

Se aprecia que la regla del árbol de decisión asocia a la especie reptil con sus antecedentes, tales como poseen plumas=no, amamanta=no, tiene espina dorsal=si, posee aletas=no y posee cola=si. Por otro lado la relación de la regla del árbol de decisión con las reglas de asociación están muy ligados con los resultantes de las reglas de asociación, esto es, si unimos en una sola regla todas reglas de asociación por sus consecuentes, se podría dar una regla igual o similar a la regla del árbol de decisión. Por lo anterior, todas las reglas están asociadas y todas clasifican bien a la especie reptil. Es necesario recordar que la clasificación de la especie reptil en laboratorios anteriores, fue particularmente de gran dificultad. Tal fue el caso al crear clusters, ya que la especie reptil y anfibio comparte muchas similitudes, lo cual se explica por la investigación realizada por Randall(1981).

Clase Pez:

Regla Árbol de Decisión:

- { milk = 0, fins = 1 => clase 4 (Pez) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0}=> {domestic=0}
- {hair=0, eggs=1, milk=0} => {feathers=0}
- {hair=0, eggs=1,feathers =0} => {breathes=1}
- {eggs=1, aquatic=1} => {hair=0}
- {feathers=0,eggs=1, airborne=0, catsize=0} => {milk=0}

Se aprecia que la regla del árbol de decisión asocia directo a la especie pez con sus antecedentes, esto es amamantan=no y poseen aletas=si. Por otro lado los resultantes de las reglas de asociación en su conjunto podrían estar relacionados con la clase pez a través de

todas las reglas en su conjunto considerando sólo los resultados. Así pues, la única variable que coincide para ambos modelos de reglas es que los peces no amamantan

Clase Anfibio:

Regla Árbol de Decisión:

- { milk = 0, backbone = 1, tail=0 => clase 5 (Anfibio) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0}=> {domestic=0}
- {hair=0, eggs=1, milk=0} => {feathers=0}
- {hair=0, eggs=1,feathers =0} => {breathes=1}
- {eggs=1, aquatic=1} => {hair=0}
- {feathers=0,eggs=1, airborne=0, catsize=0} => {milk=0}
- {eggs=1,fins=0} => {feathers=0}

Se aprecia que la regla del árbol de decisión asocia directo a la especie anfibio y sus 3 antecedentes, esto es, para que se cumpla que existe relación con la clase anfibio, los 3 atributos deben ser: amamantan=no , posee espina dorsal=si y poseen cola=no. Por otro lado los resultados de las reglas de asociación en su mayoría no están relacionados con la regla del árbol de decisión, pues casi todas en su antecedente tienen la variable ponen huevos=si y en su resultados la variable si poseen pulmones=no y una regla se da como resultado posee pulmones =si y amamantan=no. Sin embargo, por medio del conjunto delas reglas de asociación es posible clasificar de manera casi exacta la especie anfibio.

Clase Insecto:

Regla Árbol de Decisión:

- { predator = 0, backbone = 0, legs.0=0 => clase 6 (Insecto) }
- { feathers = 0, milk = 0, airbone=1 => clase 6 (Insecto) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0}=> {domestic=0}
- {hair=0, eggs=1, milk=0} => {feathers=0}
- {hair=0, eggs=1,feathers =0} => {breathes=1}
- {eggs=1, backbone=0} => {aquatic=0}
- {eggs=1, backbone=0, airborne=1} => {aquatic=0}

Se aprecia que son 2 reglas las cuales clasifican a la especie insecto. La primera clasifica a la especie insecto con la condición que no sea predador, no posee espina dorsal ni posee patas, pero todos los insectos clasificados poseen patas. Por lo anterior, se dejó como regla del árbol de decisión más representativa la segunda regla (indicada más arriba), la cual asocia a la clase insecto con los atributos de no amamantar, no poseer plumas y poder volar. Por otro lado, se pudieron determinar 6 reglas de asociación relacionadas con la clase insecto, en donde no es posible determinar a simple vista las variables antecesoras o resultantes que dan mayor peso a caracterizar la clase insecto. La única regla de asociación que se asemeja a la regla de decisión es aquella en donde sus antecesores indican que nacen de huevos, no tienen espina dorsal y pueden volar, lo cual implica que no son acuáticos. Por otro lado, por conocimiento general se sabe que los insectos son una subcategoría de invertebrados, los cuales se diferencian principalmente (según las características morfológicas disponibles en el dataset "Zoo") de los demás invertebrados en el dataset, es que tienen la capacidad de volar, lo cual se refleja por el ítem "airborne=1".

Clase Invertebrados:

Regla Árbol de Decisión:

- { backbone = 0 => clase 7 (Invertebrado) }

Reglas de Asociación:

- {hair=0,eggs=1,milk=0}=> {domestic=0}
- {hair=0, eggs=1, milk=0} => {feathers=0}
- {hair=0, eggs=1,feathers =0} => {breathes=1}
- {feathers=0,eggs=1, airborne=0, catsize=0} => {milk=0}
- {eggs=1,fins=0} => {feathers=0}
- {hair=0,feathers=0, airborne=0}=> {backbone=0}
- {eggs=1, backbone=0} => {aquatic=0}
- {eggs=1, backbone=0, airborne=0} => {aquatic=0}

Se aprecia que la regla del árbol de decisión asocia directamente a la especie invertebrado con el atributo no posee espina dorsal. Por otro, lado una regla de asociación entrega como resultado que no posee columna vertebral en base a las variables que no posee pelos, no vuela ni posee plumas. Luego, la variable espina dorsal es utilizada en 2 reglas para concluir que la especie no es acuática, lo cual indica que la clasificación es de invertebrados, sin embargo la variable que en definitiva hace la diferencia entre invertebrados e insectos es "airborne=0" lo cual indica que para los animales que no pueden volar que además no poseen espina dorsal, pertenecen a la clasificación de invertebrados y no a la de insectos.

Tal como menciona Baran Mandal (2012), la especie "insecto" es una subcategoría de invertebrados, con lo cual se entiende porque las reglas que clasifican a animales dentro de la especie "invertebrados" están muy bien relacionados con las reglas clasificadoras de "insectos". Otra característica muy interesante que presentan los insectos, lo cual es

mencionado por James Gould (1986), es la transversalidad de una característica que se aplica a la gran mayoría de insectos, la cuales la capacidad de realizar mapas marcando rutas por medo de feromonas, lo cual es una de las grandes diferencias entre “insectos” y otros animales pertenecientes al grupo de los “invertebrados”.

Validación Cruzada de K-Iteraciones (K-fold) para entrenar y probar el modelo de Árbol de Decisión

La validación cruzada de K-fold es un método para asegurar una estimación de error robusto en un modelo de clasificación entrenado. En nuestro experimento se entrenó un modelo predictivo dividiendo datos de entrenamiento y datos y un conjunto de datos de prueba. De manera similar, se realizó en conjunto la validación cruzada k veces, dividiendo los datos en k Sub muestras igualmente repartidas. Luego, para cada una de las k particiones, mantenemos la i-ésima partición y entrenamos nuestro modelo en las otras particiones k-1 y probamos en la iª partición. Finalmente, se promedia el error sobre los resultados de las pruebas de todas nuestras rondas de entrenamiento y prueba.

Así pues, de esta forma, vamos a usar la validación cruzada K-fold con 10 iteraciones para obtener un modelo más generalizable para el algoritmo de árbol de decisión C.50 para la base de datos ZOO, así se utilizó el 90% de los datos como entrenamiento y el 10% para prueba en cada iteración.

El algoritmo utilizado para determinar el promedio final de error para las predicciones y validación cruzada fue:

```
form <- "class_type ~
hair+feathers+eggs+milk+airborne+aquatic+predator+toothed+backbo
ne+breathes+venomous+fins+legs0+legs2+legs4+legs5+legs6+legs8
+tail+domestic+catsize"


folds <- split(data_zoo , cut(sample(1:nrow(data_zoo)),10))

errs.c50 <- rep(NA, length(folds))

for (i in 1:length(folds)) {

test <- ldply(folds[i], data.frame)

train <- ldply(folds[-i], data.frame)

tmp.model <- C5.0(as.formula(form), train)

tmp.predict <- predict(tmp.model, newdata=test)

conf.mat <- table(test$class_type , tmp.predict)

errs.c50[i] <- 1 - sum(diag(conf.mat))/sum(conf.mat)

 }

print(sprintf("El promedio de error usando la validación cruzada k-fold y el algoritmo de arbol
de decision  C5.0 es: %.3f percent", 100*mean(errs.c50)))
```

Finalmente, el resultado del promedio de los errores fue del 8,9%. Con esto se pudo inferir que el error aumentó cuando se usó la validación cruzada k-fold, pues el error con todos los datos sin entrenamiento fue originalmente del 1%, lo que indica que puede haber habido sesgo introducido por la superposición del modelo al realizar muchas interacciones de entrenamiento y pruebas.

9 Conclusiones

Al final de las experimentaciones y análisis de los resultados se pudo resolver gran parte del problema en la búsqueda de atributos que mejor caracterizan a una clase. Así pues, se pudo analizar y verifica que el atributo de división del árbol de decisión más importante "es amamantar" y que es la que mejor caracteriza la clase mamíferos. Seguido en menor grado de importancia, y que divide otra rama del árbol, está el atributo "posee plumas" que caracteriza mejor a las aves, luego en tercer orden está el atributo "no posee espina dorsal" que caracteriza mejor a los reptiles. En cuarta posición de importancia y dividiendo también en hojas el árbol está el atributo "poseen aletas" que caracteriza mejor a los peces. Continuando con la división del árbol está el atributo "poder volar" y que caracteriza mejor a los insectos. Todos los anteriores atributos sin duda caracterizan mejor a sus respectivas clases, pero por otro lado las clases reptil, anfibio e invertebrado tienen atributos que posiblemente no sean los más representativos de su clase, así se tiene que el atributo "ser depredador" aunque es el que mejor caracteriza a los invertebrado, está presente en muchas especies de animales distintas a los invertebrados. Por otro lado, el atributo "posee espina dorsal" es la que mejor caracteriza a la clase reptil, aun cuando existen muchas especies de animales que poseen espina dorsal y no son clasificados como reptil. También, es interesante mencionar que existen 2 instancias de ranas en la base de datos ZOO, y que una de ellas fue clasificada como reptil y la otra como anfibio, y que la única diferencia entre ellas es que una posee cola y la otra no.

En lo que respecta a las reglas generadas por el árbol de decisión y la comparación con las reglas de asociación, se pudo inferir que para la clase mamíferos tanto en los antecedentes como resultados de cada regla está presente la variable amamantar=si, la cual es considerada por el método como la más importante y refleja claramente que es la de mayor peso para las reglas y que se puede inferir en las reglas de ambos métodos como la condicionante y resultante con mayor certeza.

Respecto al uso de los modelo C.50, se pudo comprobar que el árbol de decisión generado fue fácil de interpretar, sumado a que sólo se generaron 9 hojas y que interpretan

en gran parte la clasificación de las especies, esto es, se produjo un clasificador casi preciso referente a las variables que fueron más importantes en la clasificación y se calcularon los prototipos que dan información sobre la relación entre las variables y la clasificación. En lo que respecta a las reglas generadas por este método, están cubren todas las clases, aun cuando existen reglas inconsistentes como las generadas para la clase insecto en donde si una especie no posee patas implica que es un insecto, esto se explica principalmente porque la generación de las reglas es continua por las hojas y no es posible saltar el centro del árbol para iniciar una evaluación nueva de regla.

8 Referencias

1. Tao li ,T.L. (2010). A General Model for Clustering Binary Data. *School of Computer Science, Florida International University*.

2. Quinlan J.R. (1997) C5.0 and see 5: illustrative examples. RuleQuest Res.

3. Bickler P. & Buck L. (2007). *Hypoxia Tolerance in Reptiles, Amphibians, and Fishes: Life with Variable Oxygen Availability.* Annual Review of Physiology.

4. Darwin C. (1859) *The Origin of Species by Means of Natural Selection*. London, p 44.

5. Randall D.J., Burggren W.W., Farrell A.P. & Haswell M.S. (1981) *The Evolution of Air Breathing in Vertebrates*. Cambridge University Press, p 133.

6. Baran Mandal F. (2012) *Invertebrate Zoology*. Bankura, Weste: Department of Zoology, Bankura Chistian College.

7. Gould, James L. (1986) *The locale map of honey bees: do insects have cognitive maps?.* Science, vol. 232, p. 861.

9 Apéndice

Script

#ARBOL BINARIO

```
#Carga dataset ZOO
data_zoo  =  read.csv("zoo-tree1.csv")
#Discretiza datos
data_zoo["animal_name"] <- NULL
data_zoo <- data.frame(sapply(data_zoo,as.factor))
# Clasifica los modelos de árbol usando el algoritmo de Quinlan C5.0
c50_result<-C5.0(class_type~.,data=data_zoo)
# Despliega resumen detallado para los modelos C.50
summary(c50_result)
# calcula y despliega la importancia de la variable para el modelo C5.0
 importanceC5imp(c50_result,metric='usage')
```

#REGLAS.

```
# Generar reglas a partir del arcol de decisión de los modelos C.50
ruleModel <- C5.0(type ~ ., data = data_zoo, rules = TRUE)
ruleModel
# Despliega resumen detallado para las reglas de los modelos C.50
summary(ruleModel)
```

#VALIDACION CRUZADA.

```
library(arulesViz)
install.packages('partykit', dep = TRUE)
install.packages('gmodels', dep = TRUE)
library(partykit)
library(C50)
```

```
library(gmodels)

 library(plyr)
data_zoo  =  read.csv("zoo-tree1.csv") #Carga dataset
data_zoo["animal_name"] <- NULL
data_zoo <- data.frame(sapply(data_zoo,as.factor)) #Discretiza datos


 form <- "class_type  ~
hair+feathers+eggs+milk+airborne+aquatic+predator+toothed+backbone+
breathes+venomous+fins+legs0+legs2+legs4+legs5+legs6+legs8+tail+do
mestic+catsize"


folds <- split(data_zoo , cut(sample(1:nrow(data_zoo)),10))
length(folds)
folds
errs.c50 <- rep(NA, length(folds))
for (i in 1:length(folds)) {
test <- ldply(folds[i], data.frame)
train <- ldply(folds[-i], data.frame)
tmp.model <- C5.0(as.formula(form), train)
tmp.predict <- predict(tmp.model, newdata=test)
conf.mat <- table(test$class_type , tmp.predict)
errs.c50[i] <- 1 - sum(diag(conf.mat))/sum(conf.mat)
 }
print(sprintf("El promedio de error usando la validación cruzada k-fold y el algoritmo de
arbol de decision  C5.0 es: %.3f percent", 100*mean(errs.c50)))
```