INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO

MATERIAL EDUCATIVO PARA LA UNIDAD DE APRENDIZAJE DE MINERÍA DE DATOS.

2022-2

Grupo 5CDM1

PRACTICA DE ÁRBOLES DE DECISIÓN

Nombres:

Angeles Lomeli Felipe Alberto

García Rodríguez Diana Itzel

De Luna Ocampo Yanina

Medina Barreras Daniel Ivan

1. Descripción del conjunto de datos.

Autores del conjunto de datos:
Creador: Richard Forsyth
Donante: Richard S. Forsyth
Disponible en: https://archive.ics.uci.edu/ml/datasets/zoo
El conjunto de datos original tiene registros repetidos.

2. Objetivo de la práctica.

Realizar un árbol de decisión (clasificación) para identificar las características distintivas de las siete clases de animales:

Los tipos de clases de animales se presentan en la tabla número 1.

Tabla 1. *Categorías de animales*

Numero	Tipo
1	Mamiferos
2	Aves
3	Reptiles
4	Marinos
5	Anfibios

6	Insectos
7	Invertebrados

3. Diccionario de datos.

El conjunto de datos cuenta con 18 atributos, los cuales se describen en la tabla número 2.

Tabla 2. *Diccionario de datos.*

No	Nombre	Tipo Significado
1	animal name: Unique for each instance	Nombre del animal
2	Hair: Boolean	Pelo
3	feathers: Boolean	Plumas
4	eggs: Boolean	Huevos
5	milk: Boolean	Leche
6	airborne: Boolean	Vuela
7	aquatic: Boolean	Acuatico
8	predator: Boolean	Depredador
9	toothed: Boolean	Dentado
10	backbone: Boolean	Columna vertebral
11	breathes: Boolean	Respira
12	venomous: Boolean	Venenoso
13	fins: Boolean	Aletas
14	legs: Numeric (set of values: {0,2,4,5,6,8})	Piernas. Numérico
15	tail: Boolean	Cola
16	domestic: Boolean	Domestico
17	catsize: Boolean	Tamano de gato
18	type: Numeric (integer values in range [1,7])	Tipo. Categórico

Nota. Podría categorizar el atributo *legs* en dicotómico.

4. Consideraciones encontradas en el conjunto de datos

Contiene 17 atributos con valores booleanos. Es inusual que haya 2 instancias de "rana" y una de "niña" (Portal del conjunto de datos)

5. Flujo de los datos.

En la tabla número 3 se describen las instancias de cada una de los conjuntos de datos.

Tabla 3. *Frecuencias de los conjuntos de datos (campo: tipo de animal)*

Row ID	count
Anfibios	3
Aves	20
Insectos	8
Invertebrados	10
Mamiferos	41
Marinos	13
Reptil	5

Una posibilidad de flujo de trabajo para resolver este problema se presenta en la ilustración número 1.

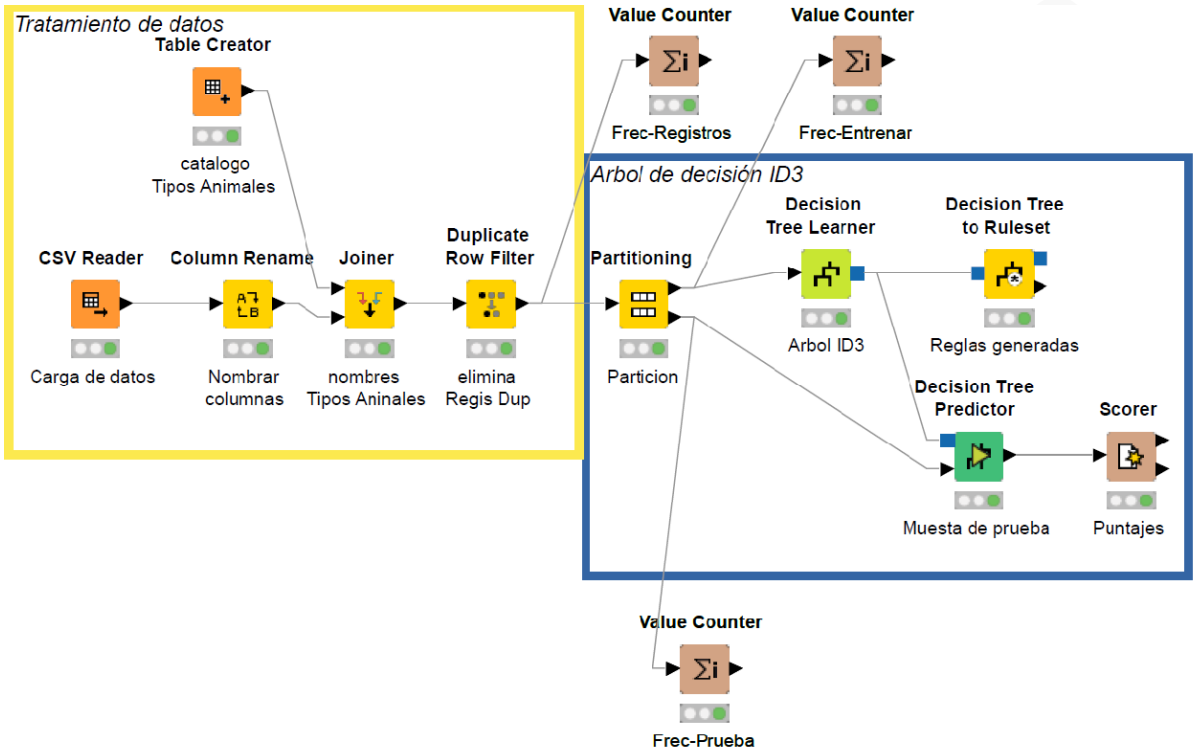


Ilustración 1. *Flujo de trabajo.*

6. Resultados

Presente los resultados considerando lo siguiente:

1. Describa las medidas generadas a partir de la matriz de confusión (archivo anexo) en cada tipo de animal
2. Analice este comportamiento en función la cantidad de elementos de cada tipo que existen en el conjunto de datos
3. ¿Qué significa el índice de Cohen´s Kappa?
4. Anexe el modelo y las reglas generadas.
5. Describa las medidas generadas a partir de la matriz de confusión (archivo anexo) en cada tipo de animal

1. Describa las medidas generadas a partir de la matriz de confusión (archivo anexo) en cada tipo de animal

nomTipo \ ...	Mamiferos	Aves	Reptiles	Marinos	Insectos	Invertebra...	Anfibios
Mamiferos	12	0	0	0	0	0	0
Aves	0	6	0	0	0	0	0
Reptiles	0	0	2	0	0	0	0
Marinos	0	0	0	4	0	0	0
Insectos	0	0	0	0	2	0	0
Invertebrados	0	0	0	0	2	1	0
Anfibios	0	0	1	0	0	0	0

Correct classified: 27 Wrong classified: 3

Accuracy: 90% Error: 10%

Cohen's kappa (κ): 0.869%

2. Analice este comportamiento en función la cantidad de elementos de cada tipo que existen en el conjunto de datos

El conjunto de datos tiene 100 filas y solo tiene 1 registro para anfibios, por lo que el entrenamiento está sesgado, al momento que esto sucede la parte de la prueba sufre errores a la hora de predecir de manera esperada, de manera similar para los invertebrados y los insectos.

3. ¿Qué significa el índice de Cohen 's Kappa?

- El índice kappa (κ) se usa para evaluar la concordancia o reproducibilidad de instrumentos de medida cuyo resultado es categórico (2 o más categorías).

- El índice kappa (κ) representa la proporción de acuerdos observados más allá del azar respecto del máximo acuerdo posible más allá del azar.

- En la interpretación del índice kappa (κ) hay que tener en cuenta que el índice depende del acuerdo observado, pero también de la prevalencia del carácter estudiado y de la simetría de los totales marginales.

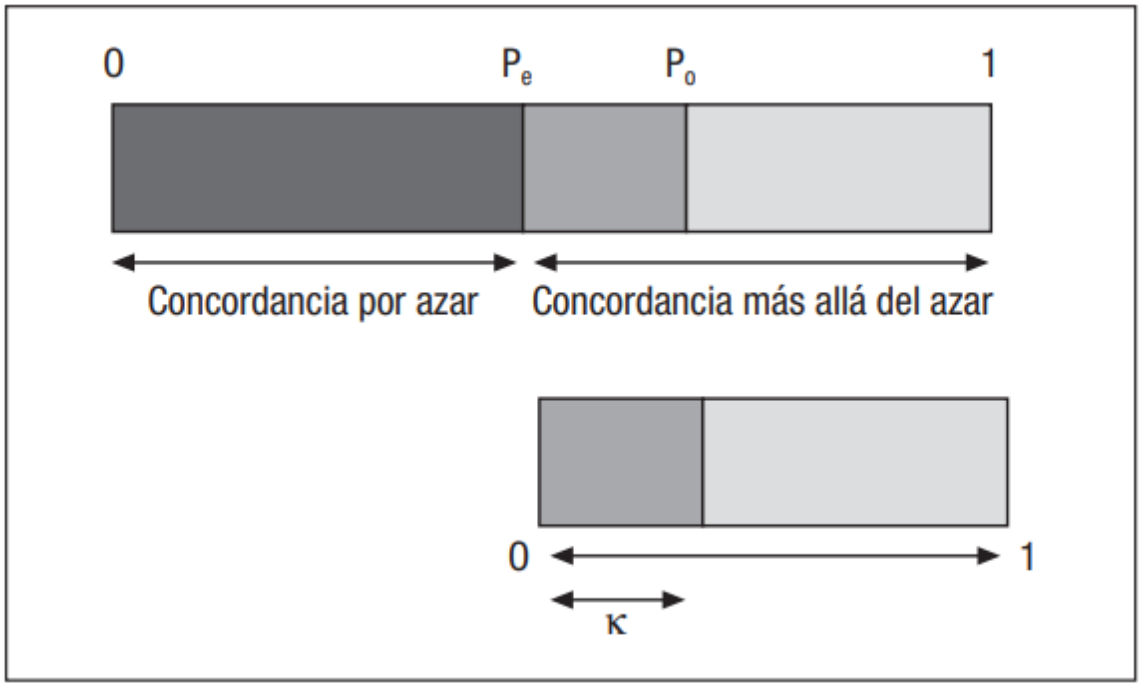
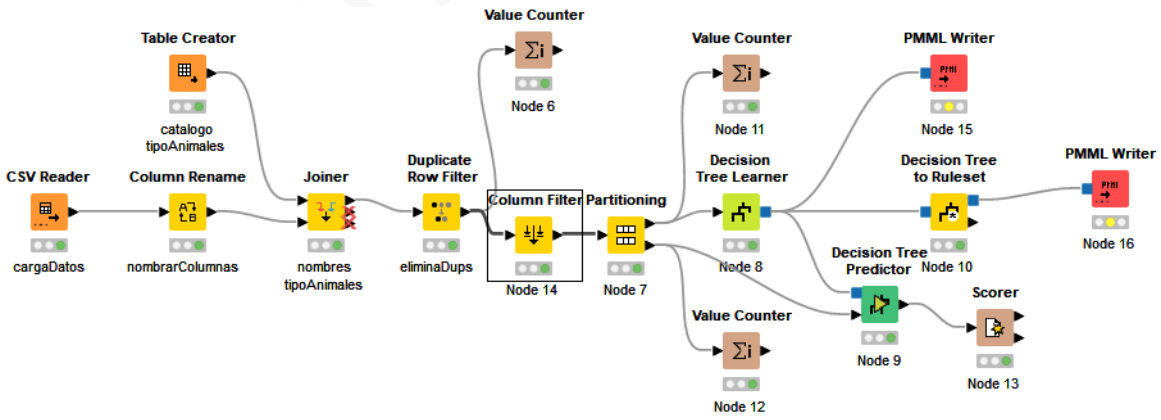


Figura 1. Representación gráfica del índice kappa.

4. Anexe el modelo y las reglas generadas.



rule set

S Rule	D Record ...	D Number...
$piernas$ <= 5.5 AND $columna vertebral$ <= 0.5 AND $aletas$ <= 0.5 AND $plumas$ <= 0.5 AND $leche$ <= 0.5 => "Invertebrados"	5	5
$piernas$ <= 7.0 AND $piernas$ > 5.5 AND $columna vertebral$ <= 0.5 AND $aletas$ <= 0.5 AND $plumas$ <= 0.5 AND $leche$ <= ...	6	6
$piernas$ > 7.0 AND $piernas$ > 5.5 AND $columna vertebral$ <= 0.5 AND $aletas$ <= 0.5 AND $plumas$ <= 0.5 AND $leche$ <= 0....	2	2
$cola$ <= 0.5 AND $columna vertebral$ > 0.5 AND $aletas$ <= 0.5 AND $plumas$ <= 0.5 AND $leche$ <= 0.5 => "Anfibios"	2	2
$cola$ > 0.5 AND $columna vertebral$ > 0.5 AND $aletas$ <= 0.5 AND $plumas$ <= 0.5 AND $leche$ <= 0.5 => "Reptiles"	3	3
$aletas$ > 0.5 AND $plumas$ <= 0.5 AND $leche$ <= 0.5 => "Marinos"	9	9
$plumas$ > 0.5 AND $leche$ <= 0.5 => "Aves"	14	14
$leche$ > 0.5 AND TRUE => "Mamiferos"	29	29

5. Describa las medidas generadas a partir de la matriz de confusión (archivo anexo) en cada tipo de animal

nomTipo \ ...	Mamiferos	Aves	Reptiles	Marinos	Insectos
Mamiferos	12	0	0	0	0
Aves	0	6	0	0	0
Reptiles	0	0	2	0	0
Marinos	0	0	0	4	0
Insectos	0	0	0	0	2
Invertebrados	0	0	0	0	2
Anfibios	0	0	1	0	0

Correct classified: 27 Wrong classified: 3

Accuracy: 90% Error: 10%

Cohen's kappa (κ): 0.869%

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
Mamiferos	12	0	18	0	1	1	1	1	1	?	?
Aves	6	0	24	0	1	1	1	1	1	?	?
Reptiles	2	1	27	0	1	0.667	1	0.964	0.8	?	?
Marinos	4	0	26	0	1	1	1	1	1	?	?
Insectos	2	2	26	0	1	0.5	1	0.929	0.667	?	?
Invertebrados	1	0	27	2	0.333	1	0.333	1	0.5	?	?
Anfibios	0	0	29	1	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.9	0.869

Observamos que tenemos algunos positivos verdaderos, lo que implica que predijo los valores “verdaderos”, omitiendo el sesgo mencionado. Asimismo, vemos que tenemos negativos verdaderos, el error del sesgo se ve en esta parte ya que salieron demasiados valores en donde se equivocó prediciendo.