



Instituto Politécnico Nacional

Escuela Superior de Cómputo

# Práctica 1 - Proceso para descubrir conocimiento en bases de datos (KDD) aplicando árboles de decisión.

Unidad de aprendizaje: Data Mining

Grupo: 3CV6

*Alumno(a):*

Ramos Diaz Enrique

*Profesor(a):*

Ocampo Botello Fabiola

17 de septiembre 2019

# Índice

<b>1</b>	<b>Introducción</b>	<b>2</b>
<b>2</b>	<b>Objetivo</b>	<b>2</b>
<b>3</b>	<b>Desarrollo: Proceso KDD</b>	<b>3</b>
3.1	Carga de la base de datos	3
3.2	Limpieza de datos	5
3.3	Integración de los datos	5
3.3.1	Carga del catálogo de clases (especie animal)	5
3.4	Selección de los datos	6
3.5	Transformación de los datos	7
3.5.1	Reemplazar registros de los campos restantes	7
3.5.2	Reemplazar clase/especie animal	11
3.6	Minería de datos	13
3.6.1	Partición de los datos	13
3.6.2	Árbol de decisión: Aprendizaje	15
3.6.3	Árbol de decisión: Predicción	17
3.7	Evaluación de patrones	19
3.7.1	Matriz de confusión	19
3.8	Representación del conocimiento	21
<b>4</b>	<b>Diagrama final KNIME</b>	<b>22</b>
<b>5</b>	<b>Bibliografía</b>	<b>22</b>

## 1. Introducción

La minería de datos es el uso de técnicas automatizadas de análisis de datos para descubrir relaciones previamente no detectadas entre elementos de datos. La minería de datos a menudo implica el análisis de datos almacenados en un almacén de datos.

Data Mining, también conocido popularmente como Knowledge Discovery in Databases (KDD), se refiere a la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de datos en bases de datos. Si bien la minería de datos y el descubrimiento de conocimiento en bases de datos (o KDD) se tratan con frecuencia como sinónimos, la minería de datos es en realidad parte del proceso KDD.

El proceso KDD esta conformado por:

1. **Limpieza de datos:** Fase en la que los datos con ruido y los datos irrelevantes se eliminan de la colección.
2. **Integración de datos:** Se pueden combinar múltiples fuentes de datos, a menudo heterogéneas, en una fuente común.
3. **Selección de datos:** Los datos relevantes para el análisis se eligen y se filtran de la recopilación de datos.
4. **Transformación de datos:** Los datos seleccionados se transforman en formas apropiadas para el procedimiento de minería.
5. **Minería de datos:** Se aplican técnicas inteligentes para extraer patrones potencialmente útiles.
6. **Evaluación de patrones:** Se identifican patrones estrictamente interesantes que representan el conocimiento en base a medidas dadas.
7. **Representación del conocimiento:** El conocimiento descubierto se representa visualmente para el usuario.

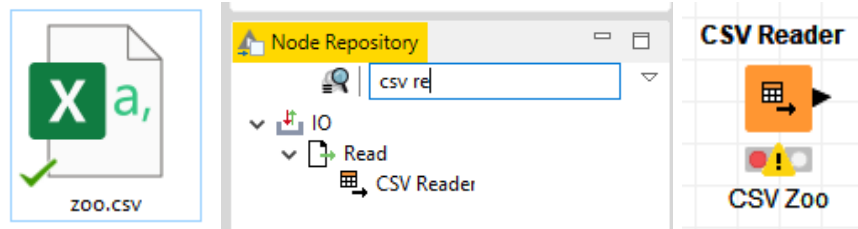
## 2. Objetivo

Generar un modelo previamente entrenado para predecir a que especie o clase corresponde un animal, con base a una serie características del conjunto de datos manejado, utilizando algoritmos supervisados como lo son los árboles de decisiones.

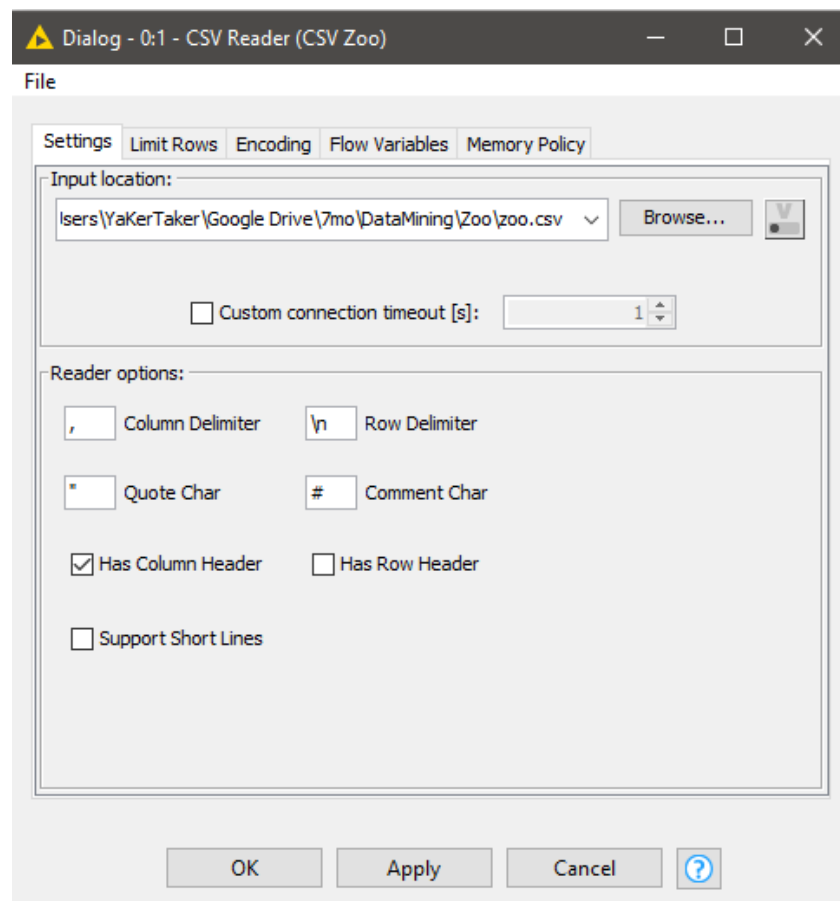
## 3. Desarrollo: Proceso KDD

### 3.1. Carga de la base de datos

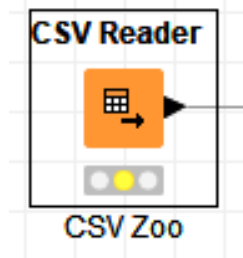
La base de datos a utilizar tiene como nombre *zoo.csv* y se trata de un archivo en formato CSV. Dentro de KNIME, buscamos **CSV Reader** en el repositorio de nodos y lo arrastramos a nuestro espacio de trabajo. Cada nodo tiene una especie de semáforo en la parte inferior, de momento estará en color rojo y con un signo de admiración.



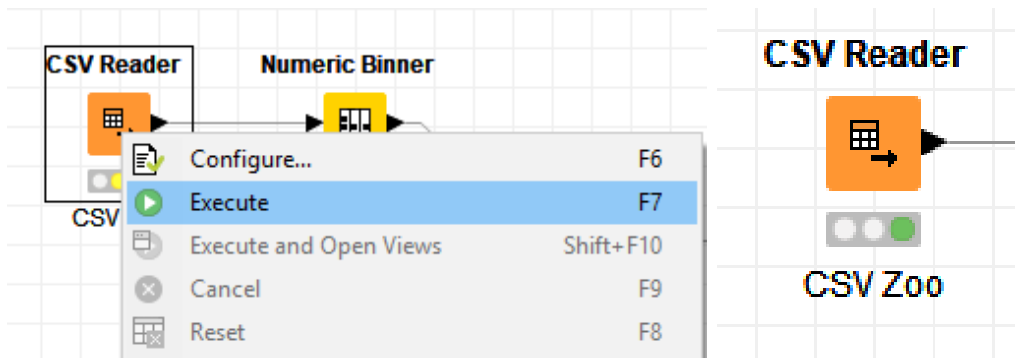
Al hacer doble clic en él, aparecerá la pantalla de configuración. Damos clic en el botón *Browse* para navegar por nuestros directorios hasta encontrar y seleccionar el archivo CSV. Es muy importante desmarcar la opción llamada *Has Row Header*, para evitar que el nodo añada un identificador a cada registro como un nuevo campo. A continuación se muestra la configuración de esta pantalla para el propósito de esta práctica:



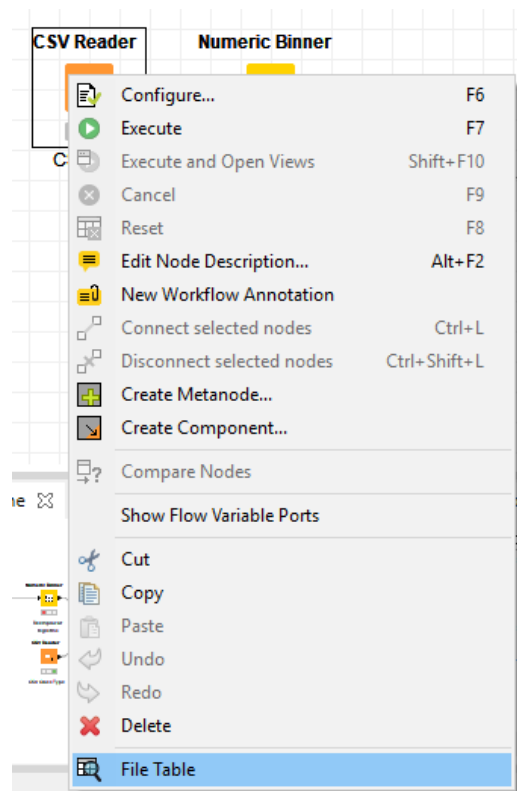
El semáforo cambiará a color amarillo una vez damos clic en el botón *OK* de la pantalla anterior.



Ahora damos clic derecho en el nodo y luego en *Execute*. El semáforo cambiará a color verde.



Podemos verificar que se ha cargado correctamente la base de datos nuevamente dando clic derecho y seleccionando la opción *File Table*:



File Table - 0:1 - CSV Reader (CSV Zoo)

File Hilite Navigation View

Table "zoo.csv" - Rows: 101 Spec - Columns: 18 Properties Flow Variables

Row ID	S animal_...	I hair	I feathers	I eggs	I milk	I airborne	I aquatic	I predator	I toothed	I backbone	I
Row0	aardvark	1	0	0	1	0	0	1	1	1	1
Row1	antelope	1	0	0	1	0	0	0	1	1	1
Row2	bass	0	0	1	0	0	1	1	1	1	0
Row3	bear	1	0	0	1	0	0	1	1	1	1
Row4	boar	1	0	0	1	0	0	1	1	1	1
Row5	buffalo	1	0	0	1	0	0	0	1	1	1
Row6	calf	1	0	0	1	0	0	0	1	1	1
Row7	carp	0	0	1	0	0	1	0	1	1	0
Row8	catfish	0	0	1	0	0	1	1	1	1	0
Row9	cavy	1	0	0	1	0	0	0	1	1	1
Row10	cheetah	1	0	0	1	0	0	1	1	1	1
Row11	chicken	0	1	1	0	1	0	0	0	1	1
Row12	chub	0	0	1	0	0	1	1	1	1	0
Row13	clam	0	0	1	0	0	0	1	0	0	0
Row14	crab	0	0	1	0	0	1	1	0	0	0
Row15	crayfish	0	0	1	0	0	1	1	0	0	0
Row16	crow	0	1	1	0	1	0	1	0	1	1
Row17	deer	1	0	0	1	0	0	0	1	1	1
Row18	dogfish	0	0	1	0	0	1	1	1	1	0
Row19	dolphin	0	0	0	1	0	1	1	1	1	1
Row20	dove	0	1	1	0	1	0	0	0	1	1
Row21	duck	0	1	1	0	1	1	0	0	1	1
Row22	elephant	1	0	0	1	0	0	0	1	1	1
Row23	flamingo	0	1	1	0	1	0	0	0	1	1

Todos los nodos poseen este semáforo y se comportan de la misma manera. En caso de error, este se tornará color rojo y mostrará un signo de admiración amarillo.

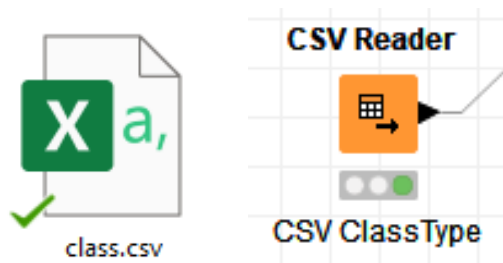
## 3.2. Limpieza de datos

La base de datos utilizada en esta práctica ya tiene los datos limpios y tratados para nuestro propósito. No es necesario realizar ninguna acción ni aplicar algún nodo para este caso.

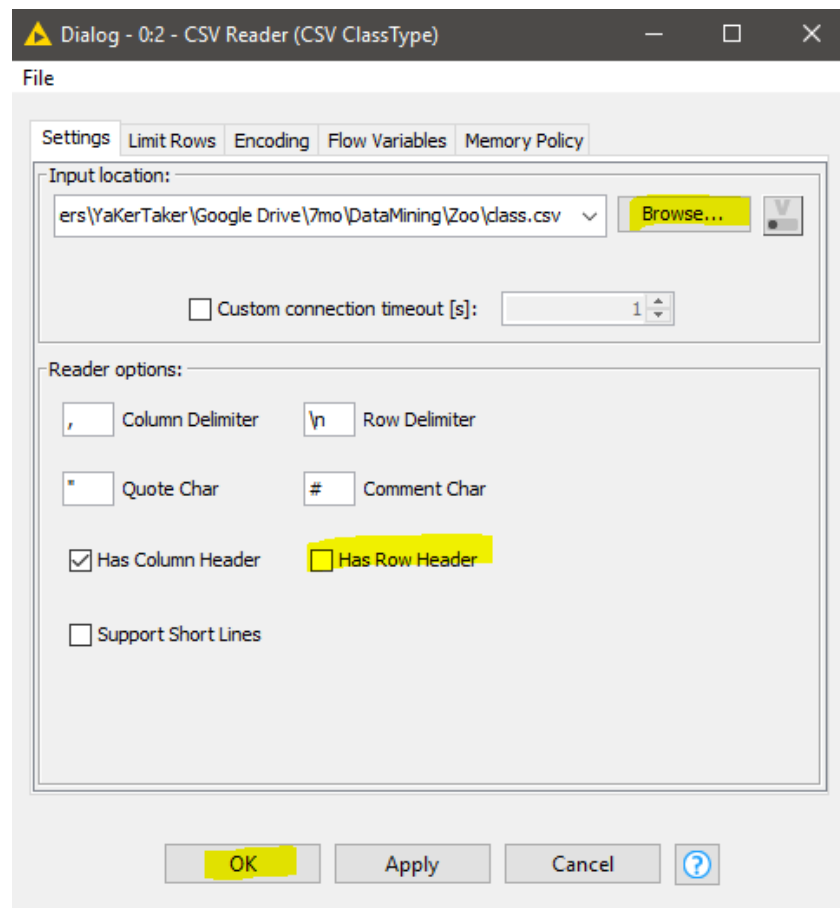
## 3.3. Integración de los datos

### 3.3.1. Carga del catálogo de clases (especie animal)

Es necesario cargar el catálogo de clases, que es un archivo CSV llamado *class.csv* el cual contiene el código de la especie a la que pertenece cada animal registrado en la base de datos principal *zoo.csv*. Esto nos servirá en la etapa de **Transformación de los datos**, que más adelante detallaremos.



El método de carga es el mismo que el de la base de datos principal, haciendo uso del nodo **CSV Reader** y dejando la misma configuración. Éste catálogo no necesita de una limpieza de datos.



File Table - 0:2 - CSV Reader (CSV ClassType)				
File Hilite Navigation View				
Table "class.csv" - Rows: 7 Spec - Columns: 4 Properties Flow Variables				
Row ID	I Class_...	I Number...	S Class_T...	S Animal_Names
Row0	1	41	Mammal	aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, g...
Row1	2	20	Bird	chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, sku...
Row2	3	5	Reptile	pitviper, seasnake, slowworm, tortoise, tuatara
Row3	4	13	Fish	bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna
Row4	5	4	Amphibian	frog, frog, newt, toad
Row5	6	8	Bug	flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
Row6	7	10	Invertebrate	clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

### 3.4. Selección de los datos

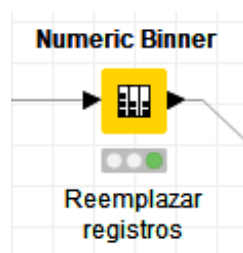
Para el propósito de ésta práctica, se utilizarán todos los atributos disponibles en la base de datos, así como todos sus registros, por lo que no es necesario hacer alguna selección específica.

## 3.5. Transformación de los datos

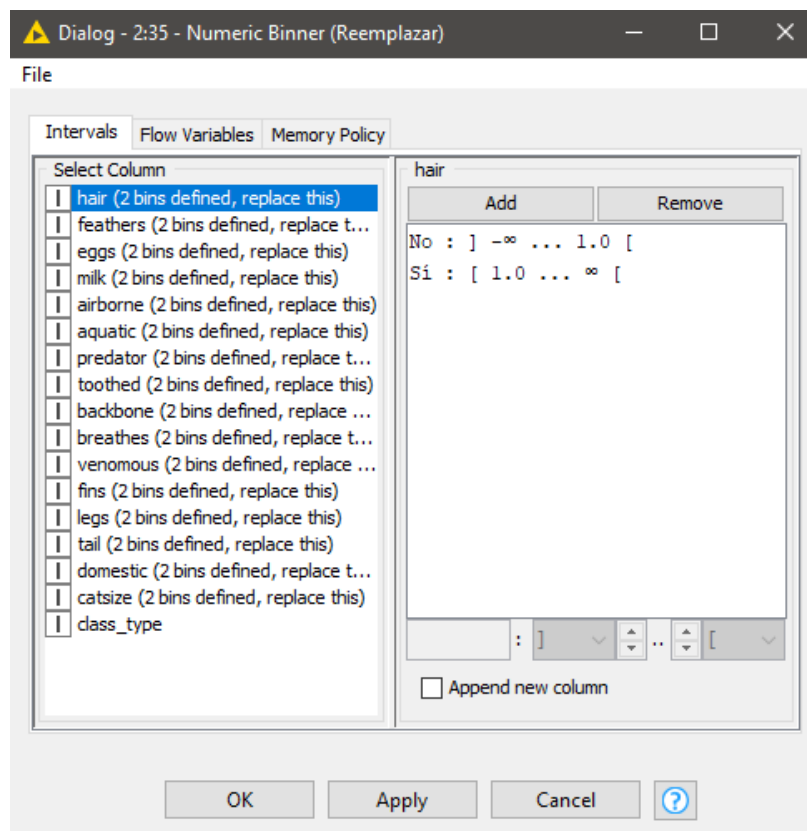
### 3.5.1. Reemplazar registros de los campos restantes

En todos los atributos, los registros vienen como datos numéricos, específicamente en binario, siendo 1 un "Sí" y 0 un "No". Es necesario reemplazar estos datos numéricos para darles un formato nominal, y así los árboles puedan clasificar el conjunto de datos de una forma correcta. Hay un caso especial, y es el del atributo **legs**, pues aquí se indica el total de patas que tiene cada animal.

Buscamos y agregamos al espacio de trabajo el nodo **Numeric Binner**. Este nodo reemplaza todos los valores contenidos en rangos que nosotros le indiquemos, por una cadena. La salida del nodo **CSV Reader** de la base de datos va conectada a la entrada de este.



Al abrir la pantalla de configuración nos topamos con todos los atributos de la base de datos listados en la parte izquierda, y en la derecha dos botones: *Add* y *Remove*.





Para reemplazar los datos por Sí/No, debemos seleccionar atributo por atributo, dar clic 2 veces en el botón *Add*. Se deben añadir dos rangos.

Debemos dar clic en el primero, y en el campo inferior ingresamos un "1.00". Con esto le indicamos que el primer rango es de menos infinito a uno.

hair

Add Remove

No : ]  $-\infty$  ... 1.0 [  
Sí : [ 1.0 ...  $\infty$  [  
  
No : ]  $-\infty$  .. 1.0 [   
☐ Append new column hair\_binned

En este intervalo entran los registros con un "0", por lo que escribimos "No" en el siguiente campo:

hair

Add Remove

No : ]  $-\infty$  ... 1.0 [  
Sí : [ 1.0 ...  $\infty$  [  
  
No : ]  $-\infty$  .. 1.0 [   
☐ Append new column hair\_binned

Si revisamos el siguiente rango veremos que éste va de 1 a más infinito; en este intervalo entran los registros con un "1", por lo que escribimos "Sí".

hair

Add Remove

No : ]  $-\infty$  ... 1.0 [

Sí : [ 1.0 ...  $\infty$  [

Sí : [  $\infty$  ... 1.0 [

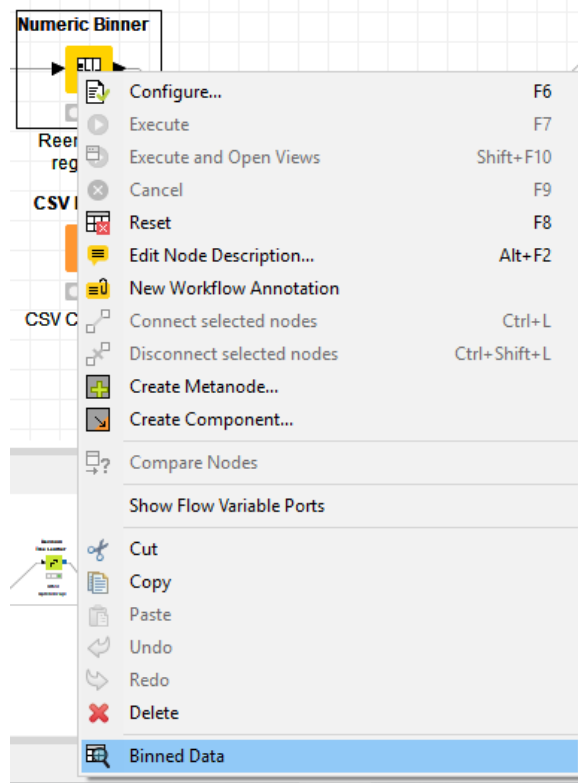
☐ Append new column hair\_binned

Este método funciona para todos los atributos, incluso para **legs**, pues convertimos esos valores numéricos a nominales; es decir, mientras el animal tenga más de 1 pata, se etiqueta como "Sí". Es muy importante omitir el atributo **class\_type**.

Select Column

- ☐ hair (2 bins defined, replace this)
- ☐ feathers (2 bins defined, replace this)
- ☐ eggs (2 bins defined, replace this)
- ☐ milk (2 bins defined, replace this)
- ☐ airborne (2 bins defined, replace this)
- ☐ aquatic (2 bins defined, replace this)
- ☐ predator (2 bins defined, replace this)
- ☐ toothed (2 bins defined, replace this)
- ☐ backbone (2 bins defined, replace this)
- ☐ breathes (2 bins defined, replace this)
- ☐ venomous (2 bins defined, replace this)
- ☐ fins (2 bins defined, replace this)
- ☐ legs (2 bins defined, replace this)
- ☐ tail (2 bins defined, replace this)
- ☐ domestic (2 bins defined, replace this)
- ☐ catsize (2 bins defined, replace this)
- ☐ class\_type

Para revisar la tabla resultante, damos clic derecho en el nodo y seleccionamos la opción *Binned Data*.



**Binned Data - 2:35 - Numeric Binner (Reemplazar)**

File Hilite Navigation View

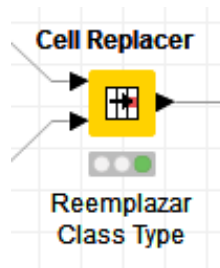
Table "default" - Rows: 101 Spec - Columns: 18 Properties Flow Variables

Row ID	animal_...	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	
Row0	aardvark	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row1	antelope	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row2	bass	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row3	bear	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row4	boar	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row5	buffalo	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row6	calf	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row7	carp	No	No	Sí	No	No	Sí	No	Sí	Sí	No
Row8	catfish	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row9	cavy	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row10	cheetah	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row11	chicken	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row12	chub	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row13	clam	No	No	Sí	No	No	No	Sí	No	No	No
Row14	crab	No	No	Sí	No	No	Sí	Sí	No	No	No
Row15	crayfish	No	No	Sí	No	No	Sí	Sí	No	No	No
Row16	crow	No	Sí	Sí	No	Sí	No	Sí	No	Sí	Sí
Row17	deer	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row18	dogfish	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row19	dolphin	No	No	No	Sí	No	Sí	Sí	Sí	Sí	Sí
Row20	dove	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row21	duck	No	Sí	Sí	No	Sí	Sí	No	No	Sí	Sí
Row22	elephant	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row23	flamingo	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí

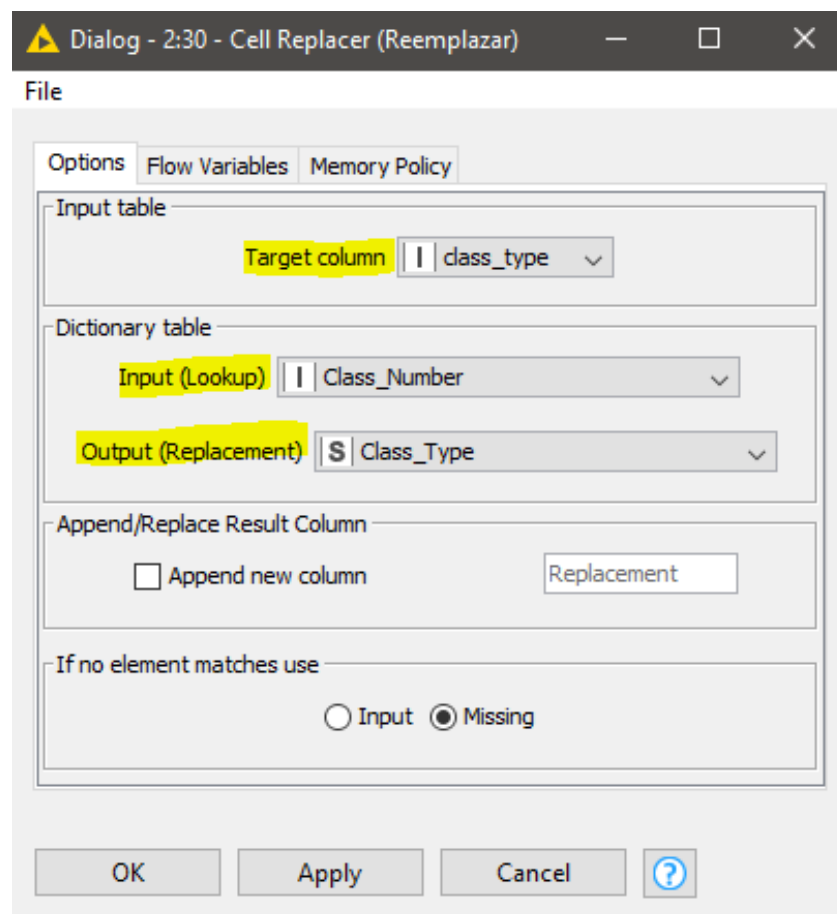
### 3.5.2. Reemplazar clase/especie animal

Por último, nos falta reemplazar el atributo **class\_type**.

Haremos uso del nodo **Cell Replacer**. Este nodo reemplaza todos los registros de una columna (atributo) en un conjunto de datos por otra de otro conjunto de datos. Tiene dos entradas: en la primera irá conectada la salida del **Numeric Binner** del paso anterior, y en la otra la salida del **CSV Reader** con el catálogo de clases/especies cargado.



Al abrir la pantalla de configuración existen 3 campos importantes para nuestro propósito: *Target column*, que son los registros que deseamos reemplazar; *Input (Lookup)*, que es la columna en donde buscará coincidencias con los registros a reemplazar; y *Output (Replacement)*, que es el nuevo dato a reemplazar según las coincidencias encontradas en las columnas de los campos anteriores. Al final la pantalla debe quedar como se muestra a continuación:



Para comprobar que todo ha salido correctamente, revisamos la tabla de salida dando clic derecho en el nodo y seleccionando la opción *Table with replaced column*:

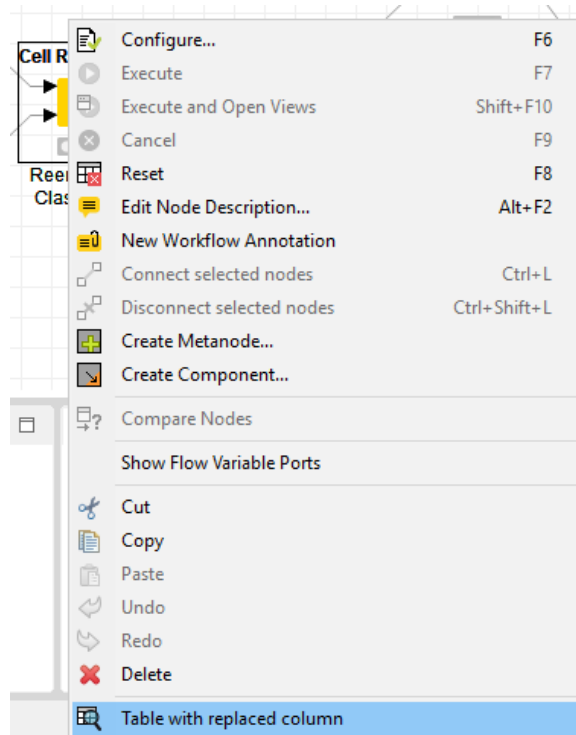
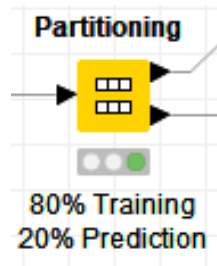


Table "default" - Rows: 101 Spec - Columns: 18 Properties Flow Variables											
Row ID	r	S toothed	S backbone	S breathes	S venomous	S fins	S legs	S tail	S domestic	S catsize	S class_t...
Row0		Sí	Sí	Sí	No	No	Sí	No	No	Sí	Mammal
Row1		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row2		Sí	Sí	No	No	Sí	No	Sí	No	No	Fish
Row3		Sí	Sí	Sí	No	No	Sí	No	No	Sí	Mammal
Row4		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row5		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row6		Sí	Sí	Sí	No	No	Sí	Sí	Sí	Sí	Mammal
Row7		Sí	Sí	No	No	No	Sí	No	Sí	No	Fish
Row8		Sí	Sí	No	No	Sí	No	Sí	No	No	Fish
Row9		Sí	Sí	Sí	No	No	Sí	No	Sí	No	Mammal
Row10		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row11		No	Sí	Sí	No	No	Sí	Sí	Sí	No	Bird
Row12		Sí	Sí	No	No	Sí	No	Sí	No	No	Fish
Row13		No	No	No	No	No	No	No	No	No	Invertebrate
Row14		No	No	No	No	No	Sí	No	No	No	Invertebrate
Row15		No	No	No	No	No	Sí	No	No	No	Invertebrate
Row16		No	Sí	Sí	No	No	Sí	Sí	No	No	Bird
Row17		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row18		Sí	Sí	No	No	Sí	No	Sí	No	Sí	Fish
Row19		Sí	Sí	Sí	No	Sí	No	Sí	No	Sí	Mammal
Row20		No	Sí	Sí	No	No	Sí	Sí	Sí	No	Bird
Row21		No	Sí	Sí	No	No	Sí	Sí	No	No	Bird
Row22		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal
Row23		No	Sí	Sí	No	No	Sí	Sí	No	Sí	Bird

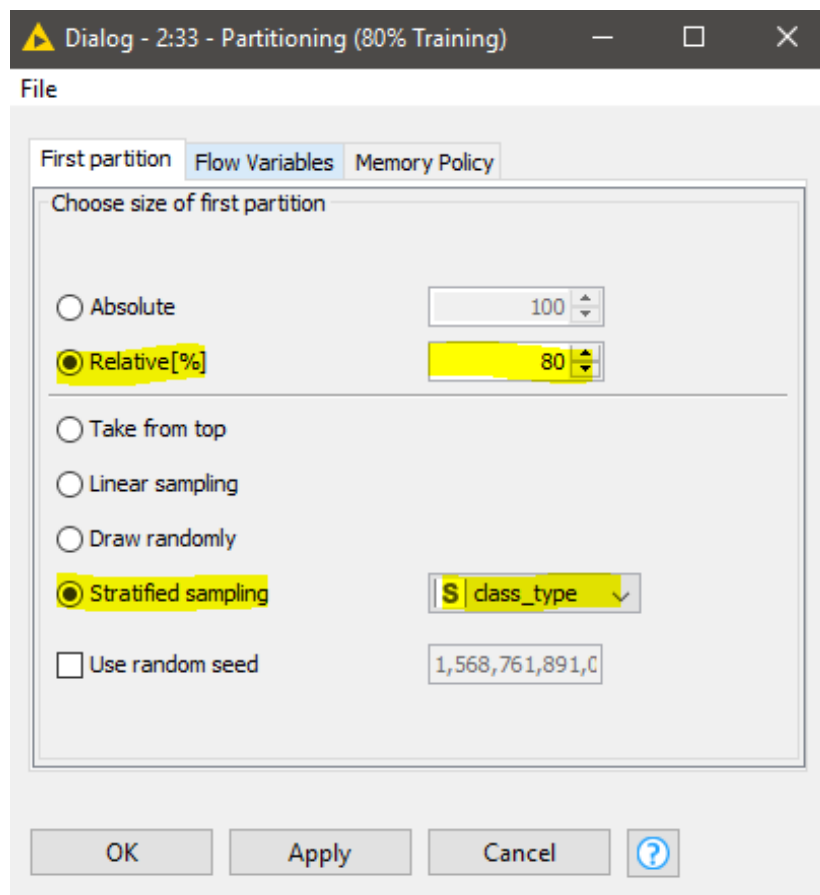
## 3.6. Minería de datos

### 3.6.1. Partición de los datos

Al ser los árboles de decisiones algoritmos de aprendizaje supervisados, es necesario dividir el conjunto de datos para generar el modelo. Una parte será para el entrenamiento o aprendizaje, que será el 80 % de todos los datos. La otra parte servirá para ser evaluada por el árbol de predicción, con base al entrenamiento del porcentaje anterior de datos. Esta parte corresponde al 20 % restante.



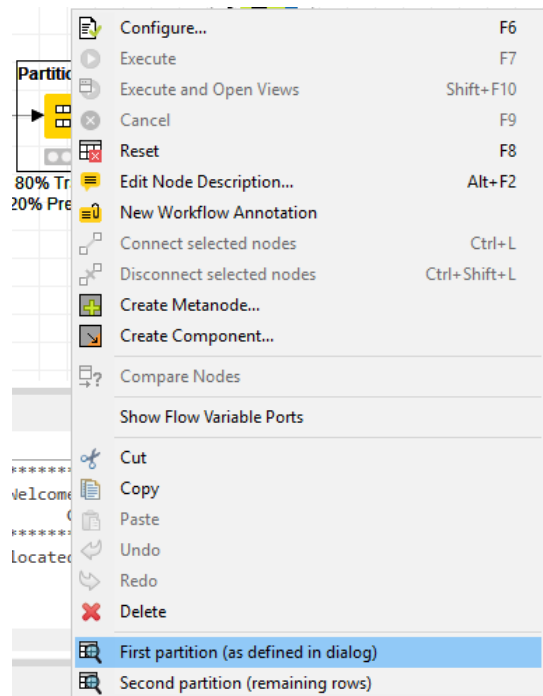
Para este fin utilizamos el nodo **Partitioning**. Tendrá como entrada la salida del nodo **Cell Replacer** con los registros ya reemplazados. Tiene dos salidas: la primera de arriba corresponde a la partición más grande (80 %), y la de abajo al resto (20 %). La pantalla de configuración deberá quedar como se muestra a continuación:



El campo *Relative*[ %] indica el porcentaje de la parte más grande en que será particionado el conjunto de datos.

Se selecciona la opción *Stratified sampling* y se selecciona el atributo de interés para la futura clasificación, en este caso **class\_type**.

Una vez configurado y ejecutado el nodo, podemos ver ambas particiones dando doble clic sobre él:



First partition (as defined in dialog) - 2:33 - Partitioning (80% Training)

File Hilite Navigation View

Table "default" - Rows: 80 Spec - Columns: 18 Properties Flow Variables

Row ID	S animal_...	S hair	S feathers	S eggs	S milk	S airborne	S aquatic	S predator	S toothed	S backbone	S
Row0	aardvark	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row1	antelope	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row2	bass	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row3	bear	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row4	boar	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row5	buffalo	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row6	calf	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row8	catfish	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row9	cavy	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row11	chicken	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row12	chub	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row14	crab	No	No	Sí	No	No	Sí	Sí	No	No	No
Row15	crayfish	No	No	Sí	No	No	Sí	Sí	No	No	No
Row16	crow	No	Sí	Sí	No	Sí	No	Sí	No	Sí	Sí
Row17	deer	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row18	dogfish	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row19	dolphin	No	No	No	Sí	No	Sí	Sí	Sí	Sí	Sí
Row20	dove	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row21	duck	No	Sí	Sí	No	Sí	Sí	No	No	Sí	Sí
Row22	elephant	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row24	flea	No	No	Sí	No	No	No	No	No	No	Sí
Row25	frog	No	No	Sí	No	No	Sí	Sí	Sí	Sí	Sí
Row26	frog	No	No	Sí	No	No	Sí	Sí	Sí	Sí	Sí
Row27	fruitbat	Sí	No	No	Sí	Sí	No	No	Sí	Sí	Sí

Second partition (remaining rows) - 2:33 - Partitioning (80% Training)

File Hilite Navigation View

Table "default" - Rows: 21 Spec - Columns: 18 Properties Flow Variables

Row ID	S animal_...	S hair	S feathers	S eggs	S milk	S airborne	S aquatic	S predator	S toothed	S backbone	S br
Row7	carp	No	No	Sí	No	No	Sí	No	Sí	Sí	No
Row10	cheetah	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row13	clam	No	No	Sí	No	No	No	Sí	No	No	No
Row23	flamingo	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row29	girl	Sí	No	No	Sí	No	No	Sí	Sí	Sí	Sí
Row31	goat	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row32	gorilla	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row35	hamster	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row36	hare	Sí	No	No	Sí	No	No	No	Sí	Sí	Sí
Row42	ladybird	No	No	Sí	No	Sí	No	Sí	No	No	Sí
Row43	lark	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row61	piranha	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row63	platypus	Sí	No	Sí	Sí	No	Sí	Sí	No	Sí	Sí
Row75	sealion	Sí	No	No	Sí	No	Sí	Sí	Sí	Sí	Sí
Row80	slowworm	No	No	Sí	No	No	No	Sí	Sí	Sí	Sí
Row83	sparrow	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí
Row85	starfish	No	No	Sí	No	No	Sí	Sí	No	No	No
Row86	stingray	No	No	Sí	No	No	Sí	Sí	Sí	Sí	No
Row88	termite	No	No	Sí	No	No	No	No	No	No	Sí
Row89	toad	No	No	Sí	No	No	Sí	No	Sí	Sí	Sí
Row100	wren	No	Sí	Sí	No	Sí	No	No	No	Sí	Sí

### 3.6.2. Árbol de decisión: Aprendizaje

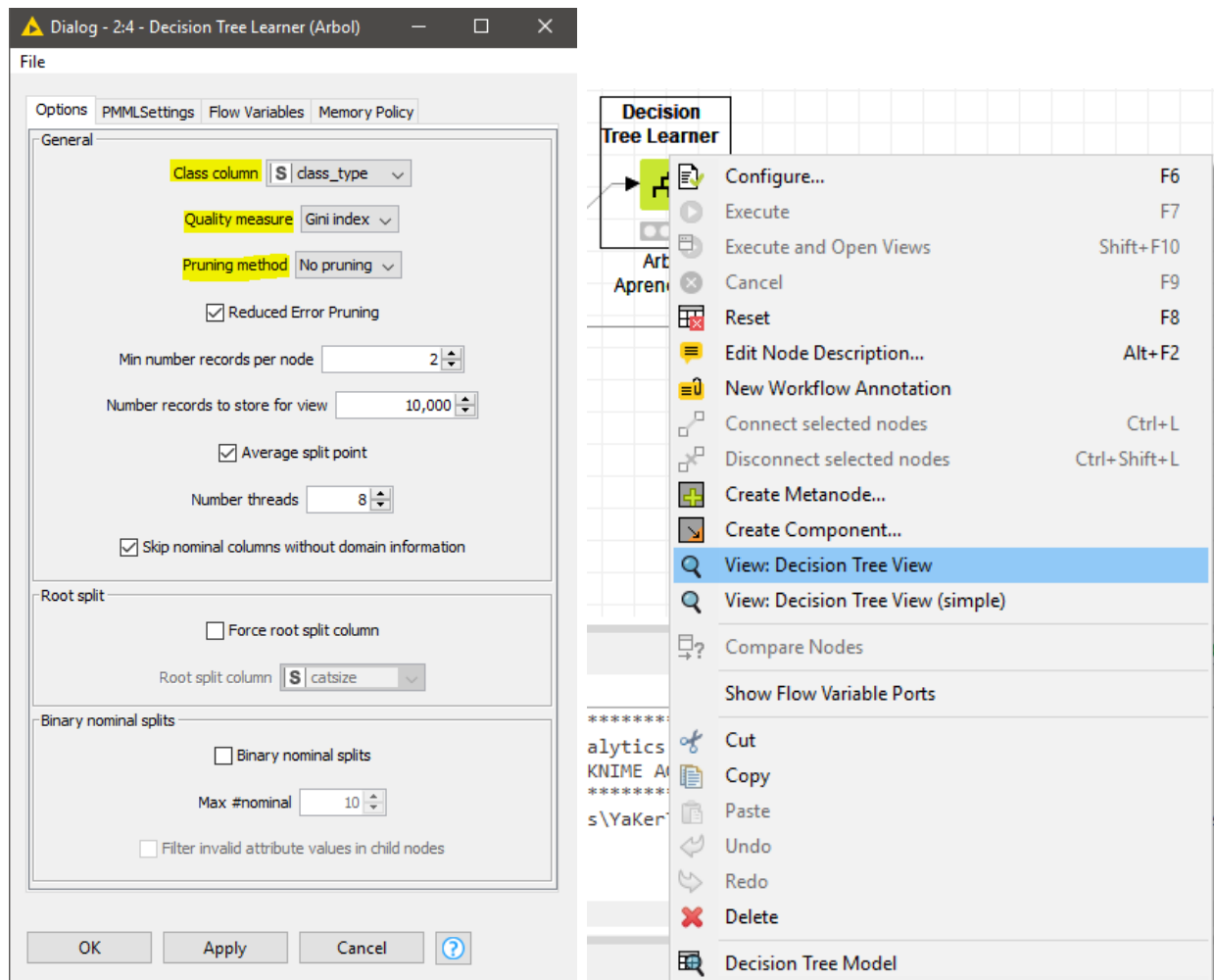
Primero es necesario entrenar al árbol de aprendizaje con un conjunto de datos. Posteriormente, éste será usado como referencia por el árbol de predicción para clasificar los datos entrantes (desconocidos).

En el repositorio de nodos buscamos y agregamos al espacio de trabajo el nodo **Decision Tree Learner**. A la entrada conectamos la salida del nodo Partitioning correspondiente al 80 % del total de registros de la base de datos.



La pantalla de configuración queda de la siguiente manera:





Los campos relevantes de esta pantalla son los siguientes: *Class column*, que es el atributo que se requiere para predecir; *Quality measure*, que determina a través de qué medición se harán los particionamientos del árbol; y *Pruning method*, que nos permite controlar el sobreajuste del algoritmo.

Para visualizar el modelo del árbol de decisión generado, damos doble clic sobre el nodo, y seleccionamos la opción *View: Decision Tree View*.

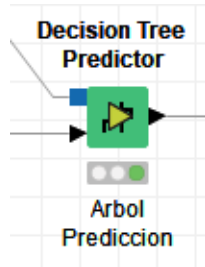


### 3.6.3. Árbol de decisión: Predicción

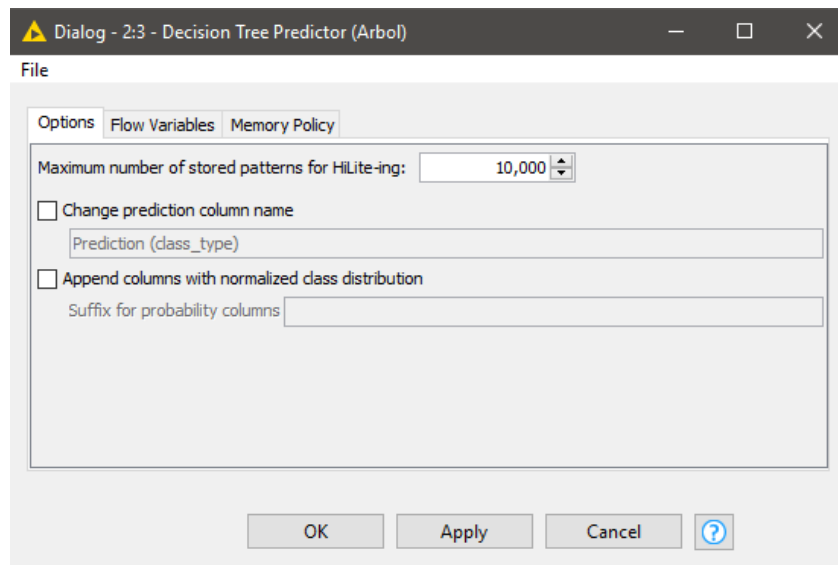
Ahora vamos a evaluar y clasificar el resto de datos con ayuda del modelo generado por el árbol de aprendizaje en el paso anterior.

Buscamos y añadimos al espacio de trabajo el nodo **Decision Tree Predictor**. Tiene dos

entradas: en la del cuadro azul va conectada la salida del nodo **Decision Tree Learner**, y en la otra va la salida con el 20 % de los datos particionados en el nodo **Partitioning**; recordemos que estos datos nunca fueron vistos durante el entrenamiento y, por ende, serán clasificados con base al modelo.



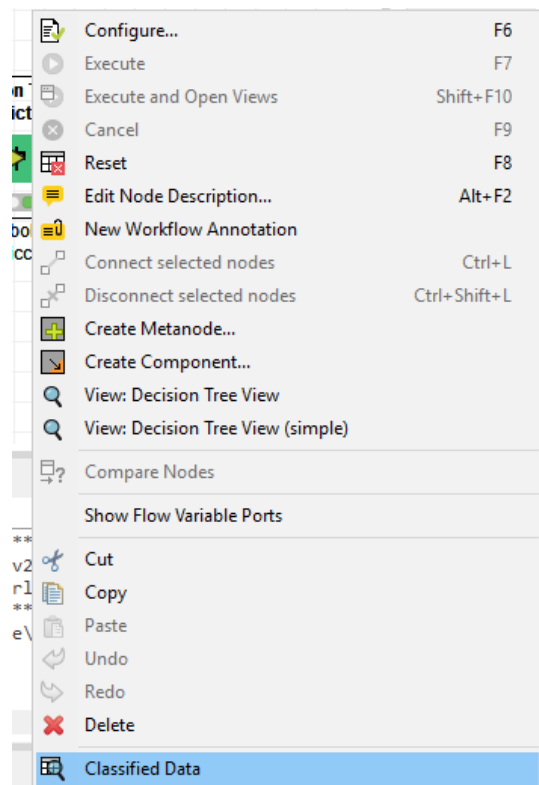
La pantalla de configuración debe lucir de la siguiente forma:



Como se puede notar, no es necesario modificar ningún campo.

A la salida del nodo se genera una columna nueva llamada *Prediction (class\_type)* que contiene el resultado del árbol de decisión en función del modelo generado anteriormente.

Podemos ver el árbol generado de manera similar al nodo anterior, o podemos revisar el resultado de la clasificación haciendo clic derecho sobre el nodo, y seleccionando la opción *Classified data*. La última columna corresponde a la predicción.



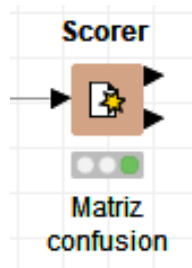
Classified Data - 2:3 - Decision Tree Predictor (Arbol)													
File Hilite Navigation View													
Table "default" - Rows: 21 Spec - Columns: 19 Properties Flow Variables													
Row ID	S predator	S toothed	S backbone	S breathes	S venomous	S fins	S legs	S tail	S domestic	S catsize	S class_t...	S Predicti...	
Row7	o	Sí	Sí	No	No	Sí	No	Sí	Sí	No	Sí	Fish	Fish
Row10		Sí	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal	Mammal	
Row13		No	No	No	No	No	No	No	No	No	Invertebrate	Invertebrate	
Row23	o	No	Sí	Sí	No	No	Sí	Sí	No	Sí	Bird	Bird	
Row29		Sí	Sí	Sí	No	No	Sí	No	Sí	Sí	Mammal	Mammal	
Row31	o	Sí	Sí	Sí	No	No	Sí	Sí	Sí	Sí	Mammal	Mammal	
Row32	o	Sí	Sí	Sí	No	No	Sí	No	No	Sí	Mammal	Mammal	
Row35	o	Sí	Sí	Sí	No	No	Sí	Sí	Sí	No	Mammal	Mammal	
Row36	o	Sí	Sí	Sí	No	No	Sí	Sí	No	No	Mammal	Mammal	
Row42		No	No	Sí	No	No	Sí	No	No	No	Bug	Bug	
Row43	o	No	Sí	Sí	No	No	Sí	Sí	No	No	Bird	Bird	
Row61		Sí	Sí	No	No	Sí	No	Sí	No	No	Fish	Fish	
Row63		No	Sí	Sí	No	No	Sí	Sí	No	Sí	Mammal	Mammal	
Row75		Sí	Sí	Sí	No	Sí	Sí	Sí	No	Sí	Mammal	Mammal	
Row80		Sí	Sí	Sí	No	No	No	Sí	No	No	Reptile	Reptile	
Row83	o	No	Sí	Sí	No	No	Sí	Sí	No	No	Bird	Bird	
Row85		No	No	No	No	No	Sí	No	No	No	Invertebrate	Invertebrate	
Row86		Sí	Sí	No	Sí	Sí	No	Sí	No	Sí	Fish	Fish	
Row88	o	No	No	Sí	No	No	Sí	No	No	No	Bug	Invertebrate	
Row89	o	Sí	Sí	Sí	No	No	Sí	No	No	No	Amphibian	Amphibian	
Row100	o	No	Sí	Sí	No	No	Sí	Sí	No	No	Bird	Bird	

## 3.7. Evaluación de patrones

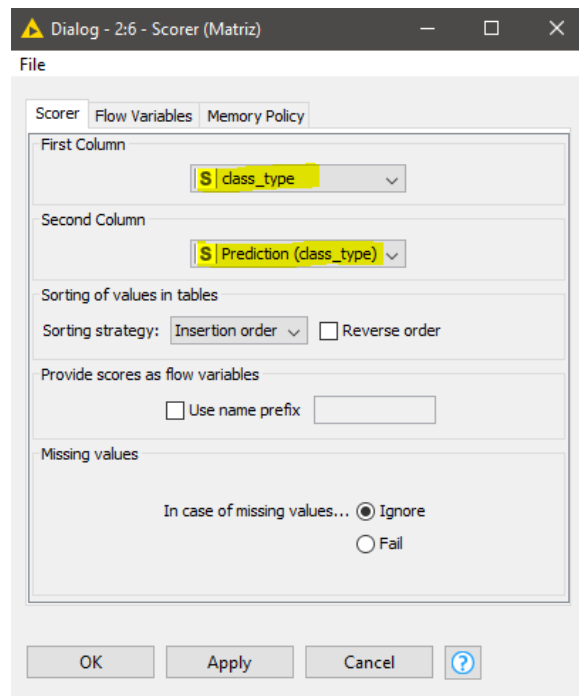
### 3.7.1. Matriz de confusión

Por último, nos queda evaluar la precisión de la clasificación de los datos desconocidos de entrada por el árbol de predicción. Para este fin haremos uso de la *matriz de confusión*, una forma de medir el error, exactitud, falsos-verdaderos, falsos-positivos, positivos - positivos y falsos-falsos.

Seleccionamos y añadimos el nodo **Scorer**; este nodo compara dos columnas o atributos y crea la matriz de confusión. A su entrada va conectada la salida del nodo **Decision Tree Predictor**.

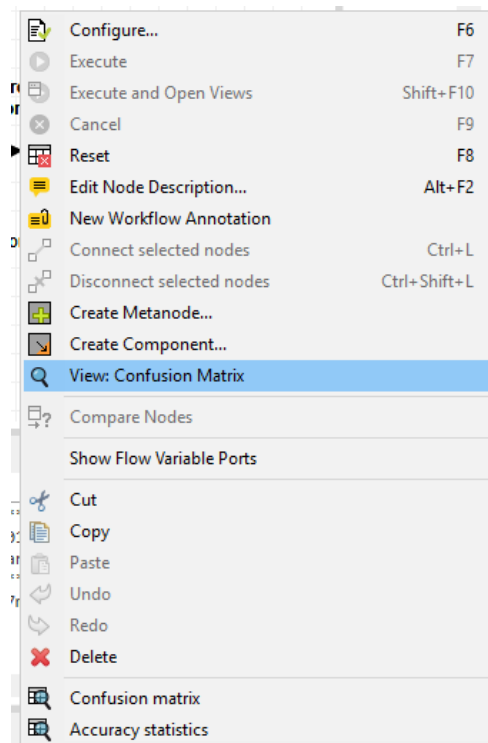



La pantalla de configuración debe quedar de la siguiente manera:



El campo *First Column* será la columna que contiene los datos correctos - originales de cada registro. Por otro lado, en el campo *Second Column* irá la columna con los datos predichos por el árbol de predicción.

Para visualizar la matriz de confusión generada, damos clic derecho en el nodo y seleccionamos la opción *View: Confussion matrix*. Aquí aparecen todas las medidas descritas al principio de esta sección.





Confusion Matrix - 2:6 - Scorer (Matriz)

—

□

✕

File Hilite

class_type...	Mammal	Fish	Bird	Invertebrate	Bug	Amphibian	Reptile
Mammal	8	0	0	0	0	0	0
Fish	0	3	0	0	0	0	0
Bird	0	0	4	0	0	0	0
Invertebrate	0	0	0	2	0	0	0
Bug	0	0	0	1	1	0	0
Amphibian	0	0	0	0	0	1	0
Reptile	0	0	0	0	0	0	1

Correct classified: 20

Accuracy: 95.238 %

Cohen's kappa ( $\kappa$ ) 0.939

Wrong classified: 1

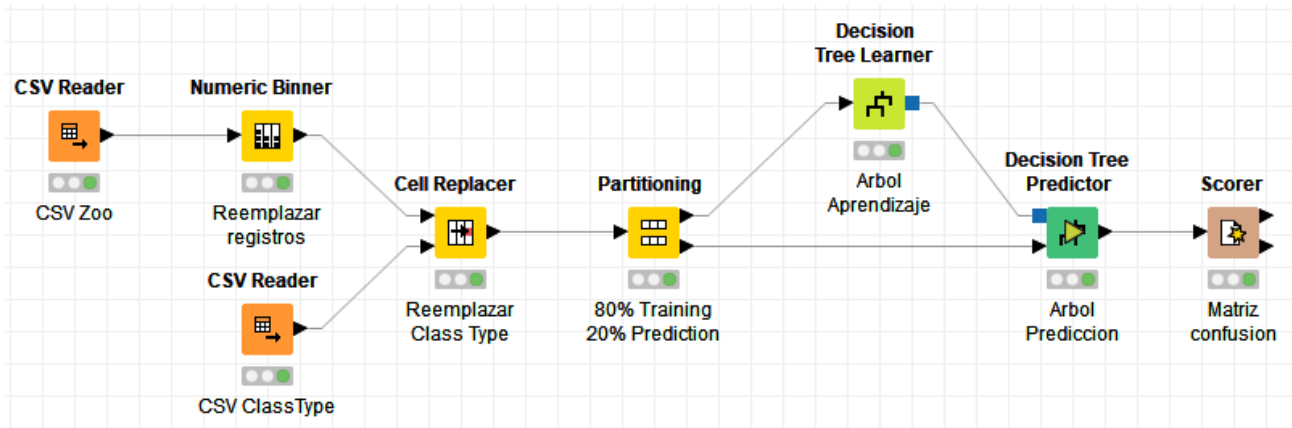
Error: 4.762 %

### 3.8. Representación del conocimiento

Formalmente ningún tipo de conocimiento se está presentado con ayuda de alguna interfaz gráfica de usuario. Sin embargo, KNIME nos ofrece los diagramas de las clasificaciones generadas por los árboles de aprendizaje y predicción.

Así como la representación tabular de la matriz de confusión.

## 4. Diagrama final KNIME



## 5. Bibliografía

- H. Sahu, S. Shrma y S. Gondhalakar, *A Brief Overview on Data Mining Survey*, International Journal of Computer Technology and Electronics Engineering (IJCTEE), vol. 1, no. 3, pp. 114 - 115, 2011
- NodePit for KNIME. [Online] Disponible en: <https://nodepit.com/nodepit-for-knime>