**INSTITUTO POLITÉCNCO 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: |
|  |
|  |

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.

1. **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 |

1. **Diccionario de datos.**

El conjunto de datos cuenta con 18 atributos, los cuales de 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.

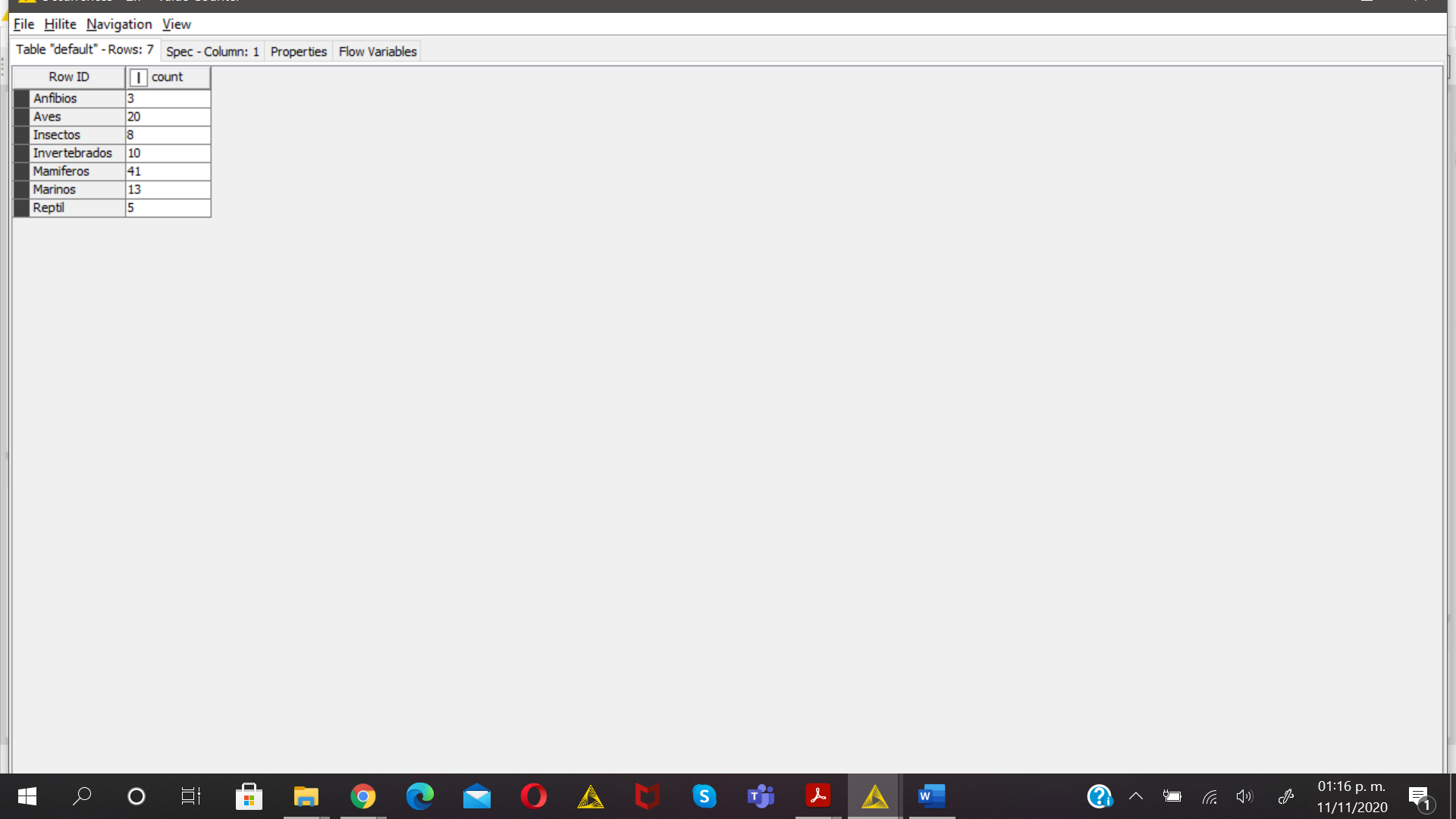
1. **Consideraciones encontradas en el conjunto de datos**

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

1. **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)



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



Ilustración 1. Flujo de trabajo.

1. **Resultados**

Presente los resultados considerando lo siguiente:

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