Daniel Gonzalez Bernal
Código: 202023300010

Juan Felipe Martinez Bedoya
Código: 201916750010

ENTREGA #2

Para el proyecto se usará el algoritmo CART, se encontró que es uno de los algoritmos mejor documentados y es el algoritmo ideal para problemas de clasificación y datasets con datos atípicos, estos atributos son esenciales para el manejo del caso propuesto en el proyecto. Como otras opciones de algoritmos se tiene en cuenta el algoritmo ID3 y el algoritmo C4.5, que es una versión mejorada del ID3 mencionado anteriormente.

	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible to outliers
CART	Towing Criteria	Handles both Categorical & Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical & Numeric value	Handle missing values.	Error Based pruning is used	Susceptible to outliers

La carga de datos no tuvo ningún cambio adicional, se seguirá utilizando el panda.

Como métodos auxiliares se está implementando primeramente una función de limpieza de base de datos, se eliminarán las columnas redundantes tales como “Nombre de Colegio” y “Código de colegio” o “Nombre del departamento” y “Código del departamento”, en estos casos los datos otorgados nos denotan la misma información, estos datos repetidos además de no tener relevancia pueden generar problemas a futuro.

Los próximos métodos auxiliares que se pretenden implementar son:

- Método recursivo para clasificar la relevancia de cada pregunta según el Gini.
- Método para calcular el Gini.
- Método para identificar el tipo de dato.
- Método para crear rangos en los datos cuantitativos.

- Método para validar la exactitud del árbol.
- Método para mostrar individualmente el éxito de una persona especifica (La entrada será por medio del ID del estudiante).

Estos y futuros métodos se irán analizando a medida que se realiza el trabajo y las necesidades puntuales que encontremos para el desarrollo del problema.

```
import pandas as pd
import numpy as np


file = pd.read_csv('TEST lite.csv', sep=';')


def separador(x):
    matriz = np.matrix(x.values)
    for i in range(len(matriz)):
        s = matriz[i, 18]
        s = s[-4:]
        p = matriz[i, 19]
        p = p // 10
        matriz[i, 18] = int(p) - int(s)
        if matriz[i, 1] == 'NO':
            matriz[i, 1] = 00

    matriz1 = np.delete(matriz, [0, 11, 12, 22, 24, 48, 50, 57, 58, 62,
64], 1)
    return matriz1


print(separador(file)[:, -1].transpose().tolist()[0])
print(separador(file)[:, :-1].tolist())

x = separador(file)[:, :-1].tolist()
y = separador(file)[:, -1].transpose().tolist()[0]
```