Examen práctico de programación Python para selección de candidato a puesto "Programador Científico Jr".

A este documento se adjuntó un archivo CSV "adult_data" con 12 columnas: id, age, workclass, education, marital-status, occupation, relationship, race,sex, hours-per-week, native-country, income".

Se requiere un programa en lenguaje de programación Python que lea el archivo de datos, aplique un clasificador Naive Bayes simple (que será descrito más adelante) y en la salida cree un archivo csv con los pesos (score) para cada una de las variables y sus valores. Se contará con 5 hrs para realizar el ejercicio.

Flujo del programa:

1. Leer datos
2. Crear bins para las variables continuas: se requieren 5 bins (quintiles) para cada variable continua, bins se refiere a transformar una variable continua en una categórica con 5 divisiones cuyo numero de elementos es igual o similar.

   Ejemplo:

Intervalo	Bin	Elementos
0-40	1	20
42-60	2	20
60-71	3	19
75-90	4	20
90-200	5	19

3. Calcular los pesos de cada una de las variables (columnas), incluyendo las transformadas e ignorando las continuas y la columna "id" , tomando como clase la columna "income > 50K", a través de la siguiente ecuación:

$$ScoreXi = log\left(\frac{\frac{NxiC}{NC}}{\frac{Nxi\sim C}{N\sim C}}\right)$$

   Donde:
   NxiC: es el número de elementos de la variable x con valor i, que pertenecen a la clase.
   NC: es el número total de elementos en la clase
   Nxi~C: es el número de elementos de la variable x con valor i, que no están en la clase
   N~C: es el número total de elementos en la no clase

Ejemplo:

Para calcular el valor de la variable “Sex” y valor “Male” para la clase “income = >50K”

NC = N(>50K)= 7841 elementos con valor de la clase
NxiC = N (Sex = Male y income = > 50K) = 6662 elementos con sex = male y income = >50K
N~C = N(<=50K) = 24720 elementos con el valor de la no clase
Nxi~C = N (Sex = Male y income = <= 50K) = 15128 elementos con sex = male y income = <=50K

$$Score\ (sex = male) = log10\left(\frac{\frac{6662}{7841}}{\frac{15128}{24720}}\right)$$

4. Guardar salida con todas las variables, valores y su score en un archivo csv con nombre “scores.csv”

   Ejemplo:

Variable	**Valor**	**Score**
Sex	Male	0.142
Sex	Female	….
Race	..	..
…	..	..