**TÉCNICAS DE INTELIGENCIA ARTIFICIAL**

Laboratorio: **Árboles de decisión, reglas y ensemble learning**

Prof. **Dr. Oscar García**

Autor: **Carlos Luis Rodríguez Niama**

**UNIR - 2022**

Laboratorio: Árboles de decisión, reglas y ensemble learning

**DESCRIPCIÓN DE LOS DATOS (dataset)**

**Número de clases, indicando que representan dichas clases y el tipo de valor que toman:**

El dataset a analizar cuenta con 7 clases, de las cuales seis (6) se consideran atributos y una (1) clase; la clase toma los siguientes valores, unacc (inaceptable), acc (aceptable), good (bueno), Vgood (muy bueno).

**Numero de instancias en total:**

El dataset en análisis cuenta con 1.728 instancias.

**Numero de instancias pertenecientes a cada clase:**

El dataset a analizar cuenta con 4 elementos:

|  |  |
| --- | --- |
| **Clase** | **Cantidad de instancias** |
| Acc | 390 |
| Good | 75 |
| Unacc | 1.215 |
| Vgood | 70 |
| **Total** | **1.750** |

**Número de atributos de entrada, su significado y tipo:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Atributo** | **Significado** | **Valor** | **Observación** |
| Buying | Rango de Valor de venta. | vhigh, high, med, low | Ejecutando procesos EDA se evidencia que son valores únicos, y de tipo categóricas. |
| Maintenance | Rango de Costo de mantenimiento. | vhigh, high, med, low |
| Doors | Característica de numero de puertas. | 2, 3, 4, 5more |
| Person | Característica de número de personas. | 2, 4, more |
| Lug\_boot | Rango de capacidad de la cajuela. | big, med, small |
| Safety | Característica del nivel de seguridad. | high, med, low |

**¿Hay algún valor de atributo desconocido?**

No existe ningún valor con atributos desconocidos

De acuerdo al análisis EDA inicial se determina lo siguiente:

* El atributo “unacc” (inaceptable), gestiona la cantidad de valores más alto en comparación con los otros atributos que gestiona a clase (acc, good, vgood), lo cual nos lleva a determinar que este atributo del dataset esta desbalanceado a comparación a los otros atributos de la clase.
* Con esta información se evidencia que este es un problema de clasificación multiclase desequilibrado.
* Finalmente se comenta que todos los atributos son categóricos, por lo que para poder ejecutar los algoritmos de sklearn debemos convertir estos datos categóricos en valores enteros.

Una vez ejecutado el proceso de transformación de valores categóricos a numéricos, procedemos a dividir el dataset resultante en datos de entrenamiento y datos de test, una práctica común es 80% de los datos como entrenamiento y el 20% restante como datos de test, esto evitará problemas en los que nuestro algoritmo pueda fallar por sobregenerar conocimiento.

Una vez generados estos datos se procede a ejecutar los algoritmos