Actividad 4. Clasificación multiclase con Scikit Learn y TensorFlow sobre *dataset diamonds*

Objetivos

* Aprender el proceso de CRISP-DM de forma práctica con las librerías de Pandas, Seaborn, Scikit Learn y TensorFlow.
* Aprender a realizar un EDA univariante, bivariante, multivariante.
* Aprender a realizar una limpieza y preparación de datos.
* Aprender a implementar modelos de clasificación multiclase de Machine Learning y Deep Learning.

Pautas de elaboración

El ejercicio consiste en crear un modelo predictivo con scikit learn para predecir la variable *cut* o corte de un diamante. El *dataset* será el de *diamonds* y está disponible públicamente en la librería *seaborn* directamente desde su enlace: <https://github.com/mwaskom/seaborn-data/blob/master/diamonds.csv>

Crear el código Python utilizando las librerías de Pandas, Seaborn, Scikit-Learn donde se cumplan los siguientes apartados:

* Cargar el *dataset* y estadísticas descriptivas.
* Limpieza de datos.
* EDAS univariante, bivariante y multivariante.
* Preparación de datos para el modelado.
* Modelado con Scikit Learn.

La columna para predecir es la columa *cut*, que es una variable categórica con más de 2 valores, por tanto, se trata de un ejercicio de clasificación multiclase. No se busca la máxima precisión para este ejercicio, se busca practicar los pasos comunes a la hora de crear modelos con scikit-learn y tensorflow.

Extensión y formato

* Entregar un archivo notebook (ipynb) con el siguiente formato:

**diamonds\_nombre\_apellido.ipynb**

* Si se detecta plagio en el *notebook* que copie parte de las respuestas de otro alumno, todos los alumnos involucrados obtendrán una calificación para la actividad de 0 puntos.
* Si se detecta contenido generado con modelos LLM como los de OpenAI o Anthropic se obtendrá una calificación para la actividad de 0 puntos.

Rúbrica

|  |  |  |  |
| --- | --- | --- | --- |
| Proyecto clasificación multiclase | Descripción | Puntuación máxima  (puntos) | Peso % |
| Criterio 1 | Cargar el dataset diamonds.csv con Pandas. Se puede añadir nulos manualmente y errores al CSV para poder practicar la limpieza de nulos y valores erróneos y duplicados. | 1.50 | 15% |
| Criterio 2 | EDAs univariante, bivariante, multivariante para entender los datos, analizarlos también estandarizados. | 3.00 | 30% |
| Criterio 3 | Preprocesados: encoding, escalado, creación de nuevas características. | 2.00 | 20% |
| Criterio 4 | Modelados: usar modelos de scikit learn para realizar una clasificación multiclase sobre la columna cut | 3.50 | 35% |
|  |  | 10 | 100 % |