TAREA 3 ANÁLISIS DE DATOS PRÁCTICO UTILIZANDO PYTHON (E/E)

A. Ferreira, W. Gómez, E. Castro
Departamento de Ingeniería Matemática
Universidad de La Frontera

09 de Junio de 2023

La Base de Datos *desafio3_1.xlsx* tiene 2 hojas. La primera hoja (OfferInformation) contiene 7 columnas con información sobre: ofertas de vino (Offer #), el mes de la oferta (Campaign), la variedad de vino (Varietal) y el origen (Origin), entre otras. La segunda hoja (Transactions) contiene 324 transacciones asociadas a compras de vinos. Los datos se componen de dos columnas, que se describen a continuación:

1. **Customer Last Name:** Nombre del cliente que realizó la transacción.
2. **Offer:** Etiqueta asociada a una oferta de vinos que el cliente adquirió.

Pregunta 1 (2 pts)

Cargar la tabla de la segunda hoja como un *DataFrame* de Pandas y realizar las siguientes actividades:

1. ¿Cuántos clientes distintos se contabilizan en los datos? ¿Cuántos tipos distintos de ofertas de vinos se pusieron a disposición de los clientes?
2. Construir una matriz de componentes binarias ($1's$ y $0's$) que describa a los clientes y las ofertas de vino que compraron. En este caso, las filas de la matriz corresponden a los clientes y las columnas a las ofertas de vinos. Cada valor igual a 1 en la matriz significa que el cliente adquirió la oferta de vinos correspondiente.

Pregunta 2 (1,5 pts)

Agrupar a los clientes en base a las compras realizadas. Para esto:

1. Correr el algoritmo K-means para agrupar a los clientes (utilizando la matriz obtenida en el item 1.2). Justificar el número de clusters K utilizado.

2. Caracterizar los clusters en base a las tres ofertas más compradas por los clientes de cada cluster, para los resultados entregados por cada algoritmo. Utilizar la primera hoja y las variables mencionadas en el encabezado de la Tarea para la caracterización.

Por otra parte, la Base de Datos *desafio3_2.csv* contiene 10 columnas con información sobre precios (en miles de dólares) de automóviles (columna *price*). Las columnas *carat, depth, table, x, y, z* representan variables continuas, mientras que las columnas *cut, color, clarity* representan variables categóricas.

Pregunta 3 (2,5 pts)

Ajustar un Modelo de Regresión para predecir el precio de los automóviles. Para esto:

1. Leer la base de datos y almacenarla como un objeto *DataFrame*. Mostrar las estadísticas descriptivas de todas las variables. Transformar las variables categóricas a variables dummies y generar un conjunto (X, y), con y la variable precio y X la matriz formada por el resto de las variables. ¿Cuál es la dimensión de X?

2. Ajustar un modelo de regresión con el algoritmo XGBoost sobre el conjunto de datos (X, y) utilizando el método 10-fold cross validation (CV). Describir los hiperparámetros utilizados en el algoritmo. Por cada fold, mostrar el valor del error cuadrático medio (MSE) y el coeficiente de determinación (R^2) en los respectivos conjuntos de entrenamiento y validación.

3. ¿Cómo es el comportamiento promedio de los modelos ajustados con la metodología CV? ¿Existe overfitting o underfitting? Interprete las métricas. ¿Qué puede concluir?

Cada estudiante debe desarrollar la tarea de forma individual y preparar un Notebook denominado $Nombre\_Apellido\_Tarea3.ipynb$. El Notebook debe incluir la explicación de cada paso implementado, los códigos correspondientes y los resultados obtenidos. En caso de haber copia, se sancionará a los estudiantes implicados con la nota mínima. Cada estudiante deberá subir su solución al campus virtual antes del día viernes 23 de junio a las 23:30 hrs.