INFORME DE CALIDAD DE LOS DATOS

MANUEL RODRÍGUEZ VILLEGAS

Para la práctica se proporcionan cinco Datasets con información acerca de las ventas de una pizzería. El objetivo es analizar los datos para exclarecer qué ingredientes comprar cada semana buscando, lógicamente, aumentar al máximo los beneficios.

El primero de los Datasets es un diccionario donde se describen los otros cuatro. Dicho Dataset será utiliado simplemente para hacernos una idea del contenido.

El segundo Dataset, denominado *order_details.csv* hace referencia a los detalles de todos los pedidos registrados en la pizzería durante 2015. Está compuesto de cuatro columnas: *order_details_id*, *order_id*, *pizza_id* y *quantity*. Cada fila es un pizza particular de cada pedido, es decir, la columna *order_details_id* guarda enteros haciendo referencia los indicadores de todas las pizzas que se vendieron durante ese año. La siguiente columna, *order_id*, guarda el número del pedido particular. No es igual que la anterior porque en un mismo pedido, es decir, para un mismo valor de *order_id*, podemos tener tantos valores como pizzas diferentes se incluyeron en ese pedido. La columna *pizza_id* guarda strings compuestos por el nombre de la pizza y el tamaño unidos por un guion bajo. Finalmente, la columna *quantity* guarda enteros que hacen referencia a la cantidad de pizzas particulares se piddieron en cada pedido.

Continuando con los Datasets, *orders.csv* guarda la información relativa al momento en el que se hicieorn todos los pedidos del año. Está compuesto por tres columnas: *order_id*, *date* y *time*. Cada fila hace referencia a un pedido, de manera que la primera columna, *order_id*, almacena enteros referentes al identificador de cada producto. La columna *date* guarda la fecha en la que se realizó cada pedido, en formato %d/%m/%y. Finalmente, la columna *time* guarda la hora a la que se pidió.

El siguiente Dataset es *pizza_types.csv*, formado por cuatro columnas que hacen referencia al nombre e ingredientes de cada pizza que se ofrece en la carta de la pizzería. La primera columna, *pizza_type_id*, está formada por strings que indican el tipo de pizza al que se hace referencia en cada línea. La segunda columna, *name*, es el nombre de la pizza en particular. La tercera columna es la categoría a la que corresponde cada pizza y, por último, la cuarta fila son strings compuestos por los ingredientes que contiene cada pizza (separados por comas).

Por último, el Dataset *pizzas.csv* almacena información sobre el precio de cada pizza en función de su tamaño. La primera columna, *pizza_id*, está formada por strings que indican la pizza a la que corresponde cada fila y su tamaño, nuevamente separados por un guion bajo. La columna *pizza_type_id* es similar a la primera solo que no se muestra el tamaño de la pizza, solo el tipo. La columna *size* hace referencia al tamaño de cada pizza, que puede ser S, L o M, salvo en el caso de la pizza griega que también puede ser XL o incluso XXL. Finalmente, la última columna es la del precio de cada pizza particular, teniendo en cuenta su tamaño.

La calidad de los datos es excelente. En ninguna columna encontramos un Nan o un None, todas tienen el formato adecuado y el único problema que ha surgido ha sido con el nombre de una pizza en *pizza_types.csv*, que ha necesitado *encoding=latin1*.