Aplicación de Machine Learning a Predicción de Demanda

Cuarto Informe de Aprendizaje No Supervisado

02 DE OCTUBRE DE 2021

Integrantes:
Gutiérrez Montecino, Denise
Nieva, Sofía
Rodríguez, Alfredo Manuel

Contenido

1 INTRODUCCIÓN

En el siguiente informe se detallan los resultados obtenidos del aprendizaje no supervisado de la base de datos suministrada como parte de la Mentoría ***Predicciones de Demanda de Producto*** de la Diplomatura de Ciencia de Datos de la Facultad de Matemáticas, Astronomía y Física (FaMAF) de la Universidad Nacional de Córdoba, Argentina.

El presente entregable contiene documentación del aprendizaje no supervisado, utilizando como base el dataset obtenido del entregable de análisis y visualización filtrado para las seis categorías más vendidas 07 - SABORES COMUNES, 08 - SABORES ESPECIALES, 17 - POTE 1 LTS, 16 - FAMILIAR, 10 - PALITOS CREMA / FRUTAL / BOMBON y 09 - SABORES PREMIUM y las tres provincias que más venden, Buenos Aires, Santa Fe y Córdoba.

El objetivo de este informe es aplicar los pasos de aprendizaje no supervisado para obtener patrones sistematizables y así poder agrupar los productos que son comparables entre sí.

2 RESPUESTAS DEL PRÁCTICO

En esta sección se desarrollarán las consignas dadas para el informe con sus respectivas observaciones y comentarios de los resultados obtenidos así como también de las decisiones que se fueron realizando, si fuese el caso. A continuación se describe cada una de ellas.

1. Preparar el dataset

En esta sección se obtiene el dataset que se utilizará para desarrollar las consignas del presente informe. Para ello, se filtró la información por las tres provincias con más ventas, estas son: Buenos Aires, Santa Fe y Córdoba. También se seleccionaron las primeras seis categorías de productos que presentaron una mayor venta según lo obtenido de los análisis realizados anteriormente, estas son: 07 - SABORES COMUNES, 08 - SABORES ESPECIALES, 17 - POTE 1 LTS, 16 - FAMILIAR, 10 - PALITOS CREMA / FRUTAL / BOMBON y 09 - SABORES PREMIUM. El dataset final tiene una columna por cada categoría y las filas son cada una de las localidades de las tres provincias. En cada celda el valor es la suma total de la cantidad pedida en la localidad y categoría correspondiente.

2. Separar datos de muestra del Dataset

En esta sección se realiza una separación aleatoria de los datos de la muestra obtenida en la sección anterior. Esta separación se realiza dado que aplicaremos a continuación la metodología de aprendizaje no supervisado, para luego verificar si alguno de los datos separados corresponden o no a algún cluster encontrado.

En este caso se decidió separar 10 filas aleatoriamente y fueron eliminadas del dataset original con el que se continuará trabajando.

3. Eliminar las columnas no relevantes del Dataset

Se procede a eliminar las columnas de provincias y localidades dado que no serán relevantes para la identificación de clusters.

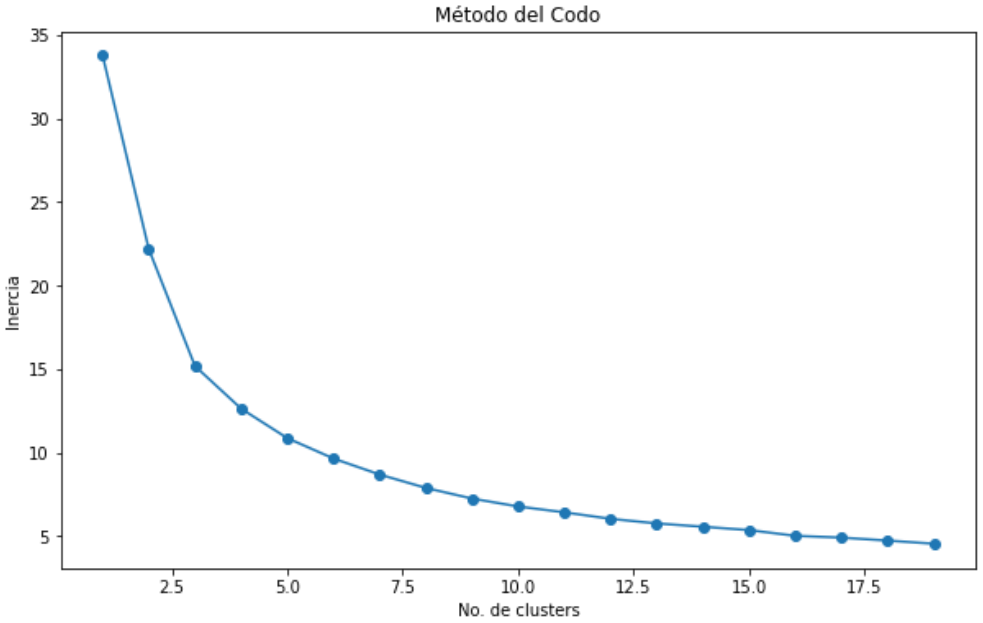
4. Normalización de Datos

En esta sección se normalizan los datos obtenidos tras los tratamientos realizados en los puntos anteriores. El objetivo es obtener un conjunto de datos apropiado para la aplicación de los algoritmos de clustering.

5. Encontramos Clusters

Con el conjunto de datos ya filtrado y normalizado, a través de la metodología de K-means se procede a identificar los clusters de aquellas categorías de productos que se venden juntos.

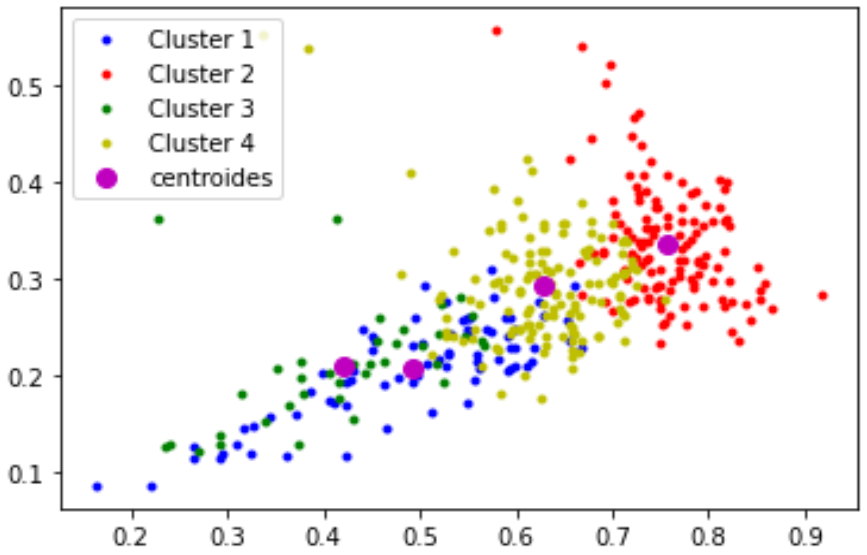
Para encontrar el valor de K se utiliza el método del codo. El método del codo consiste en calcular y graficar la suma de cuadrado en cada número de clústeres para luego buscar un cambio de pendiente de empinada a poca profundidad para determinar el número óptimo de clústeres. El gráfico que se obtuvo fue el siguiente:



Como se puede observar, el cambio de pendiente se produce aproximadamente en k = 4, por lo tanto, ese es el número de cluster que se eligió.

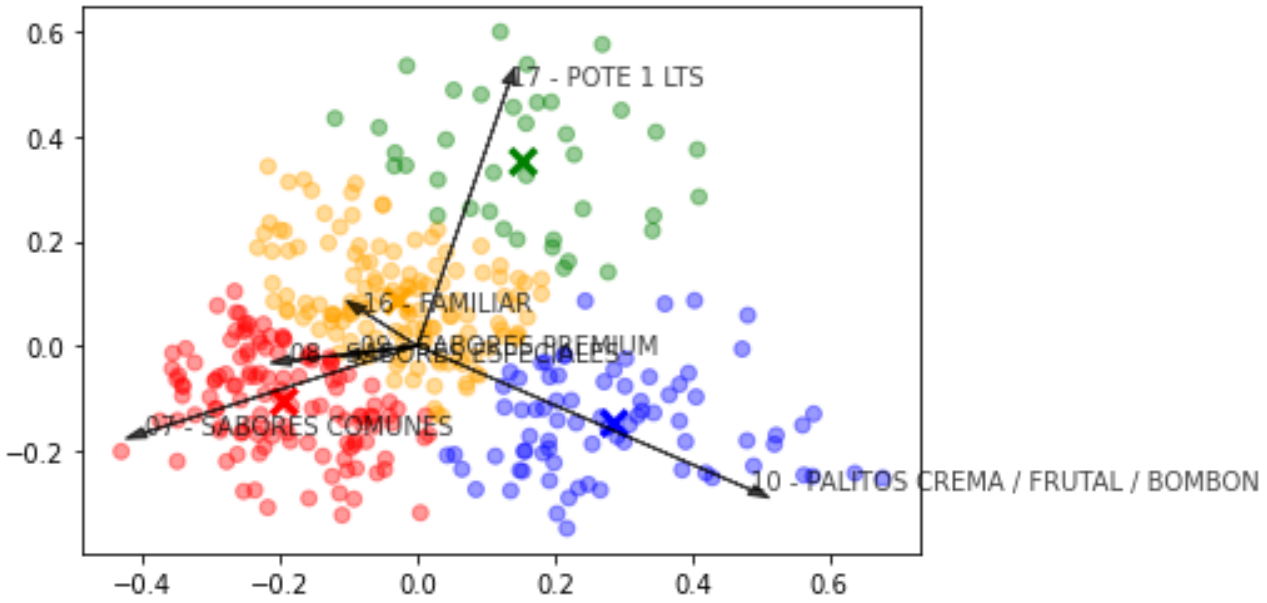
6. Graficamos los Clusters encontrados

Los cluster encontrados con k = 4, graficados usando las categorías 07 - SABORES COMUNES y 08 - SABORES ESPECIALES como ejes se pueden observar en la siguiente figura. También se marcaron los centroides de cada cluster.



7. Utilizamos PCA y graficamos

Se aplica PCA a los datos para reducir la dimensionalidad de 6 a 2 y se grafican los puntos tomando como ejes las dos componentes principales y coloreando según el cluster al que corresponden. En la gráfica se pueden visualizar también una serie de flechas que indican la proyección de cada característica en el eje principal del componente. Estas flechas representan el nivel de importancia de cada característica en la escala multidimensional. También se pueden observar si hay grupos bien definidos o hay puntos intercalados lo cual indicaría que son categorías que se venden en conjunto.



3 PRINCIPALES CONCLUSIONES

Luego de la aplicación del algoritmo de K-means se obtuvieron 4 clusters. En el primer gráfico están un poco mezclados pero esto se debe a que se están proyectando los puntos en el espacio generado por las primeras dos categorías que no necesariamente son las más representativas. En cambio, en el segundo gráfico, al proyectar los puntos en el espacio generado por las dos primeras componentes obtenidas con PCA, que son las que explican la mayor variabilidad en los datos, sí se puede ver que los cluster están bastante bien separados y definidos.

Por otro lado, se puede observar que el cluster azul se corresponde bastante bien con la categoría 10 y el cluster verde, con la categoría 17. Esto significa que probablemente las localidades pertenecientes a esos clusters piden en general más cantidad de productos de las categorías 10 y 17, respectivamente. Las flechas de las otras cuatro categorías están más cerca entre sí, con lo cual es más dificil sacar conclusiones, pero parece que el cluster amarillo se corresponde aproximadamente con la categoría 16, y el cluster rojo con las categorias 07, 08 y 09, lo que indicaria que estas tres categorías suelen venderse juntas.