Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Equipo 4

Deep Learning | 7 - Feature Selection

Integrantes:

Eliezer Cavazos Rochin A00835194

Facundo Colasurdo Caldironi A01198015

Saul Francisco Vázquez del Río A01198261

José Carlos Sänchez Gómez A01174050

Campus Monterrey **Fecha: 6/11/20**

1. Selección de Features

Nuestro equipo decidió crear clusters para poder manejar a los grupos de clientes, de esa manera consiguiendo obtener aquellos a los cuales se considera como público objetivo, es decir, quienes les interesan los productos de lanzamiento.

Para poder determinar los clusters, se tuvieron que utilizar diversas fuentes de productos, en el primero, se decidió utilizar 8 features de la información, siendo estos: La marca del producto, las categorías del producto, si es retornable, el sabor, la categoría del producto, el tamaño y contenedor del mismo. Todos estas features fueron necesarias para crear el primer filtrado del producto, el cual es la categoría en la que cae el producto de lanzamiento.

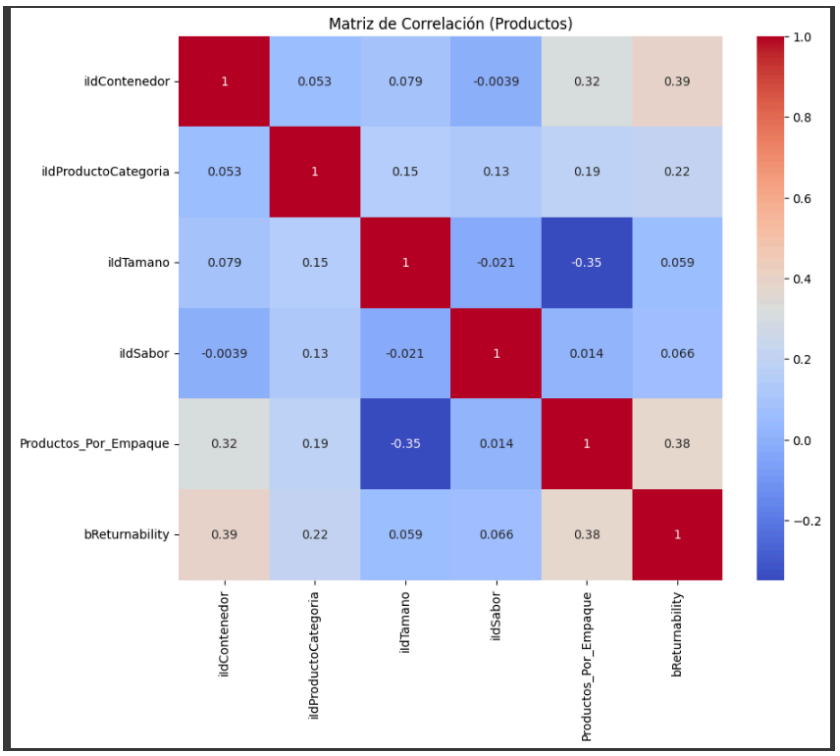
Por otra parte, para determinar a los clientes, se decidieron usar distintos aspectos que vienen prevenidos en la base de datos, siendo estas features el nivel socioeconómico, la cantidad de personas de ese nivel socioeconómico y las zonas alrededor de los mismos, cada uno de estas permitieron obtener el segundo y tercer filtro de productos, los cuales serán otros clusters respectivamente.

Todo lo anterior nos permite obtener el tipo de producto, el cual se pasará al segundo cluster, en donde se obtiene el nivel socioeconómico el cual se considera más apropiado para poder lograr vender este producto y conseguir el mayor número de ventas, con lo que finalmente se podrá pasar a los sub clusters, los cuales cada uno de ellos contiene su respectivo cluster de cada nivel socioeconómico para poder obtener los patrones de cobra de cada nivel, lo cual nos permite definir qué patrón de compra se adapta mejor a qué nivel socioeconómico del producto que se considera como producto de lanzamiento

2. Métodos y Técnicas Utilizadas:

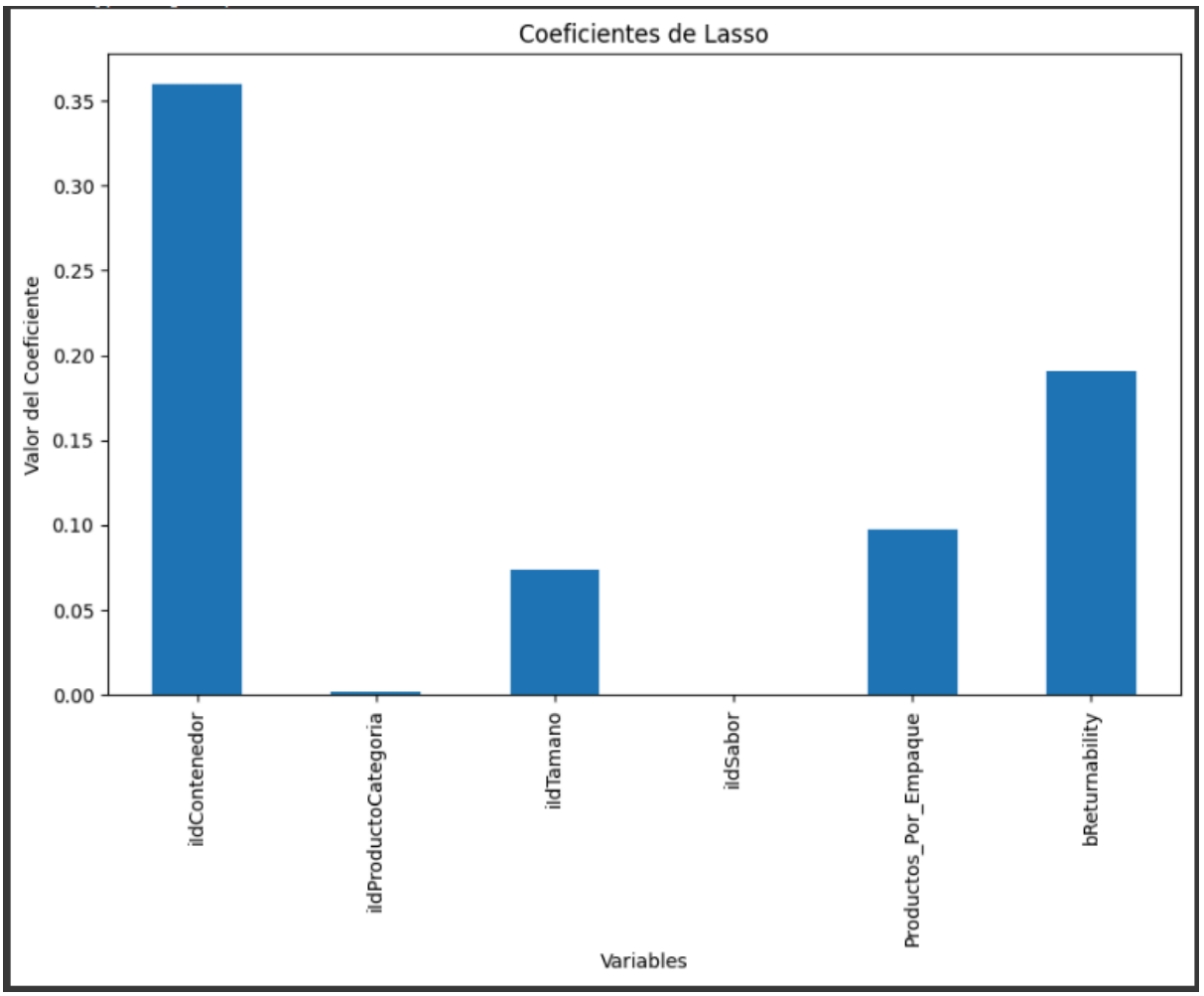
Para poder lograr seleccionar las mejores features, se decidió utilizar el método Intrínseco para la selección de las features, ya que este método nos permite poder identificar cuales se consideran como las más relevantes, por medio de uso de filtros más regulaciones para reducir los errores y mejorar la precisión de los modelos predictivos, se utilizó este método para asegurarse que todas las variables que fueron seleccionadas fueran consideradas de valor, para ser incluidas en el proceso de segmentación.

El primer paso que se realizó, fue el realizar un filtro inicial, por medio de un analisis de correlacion para poder detectar aquellas variables que no eran de importancia y eliminarlas, un ejemplo de esto se vio al analizar los datos de los productos, con variables como categoría y tamaño tienen un gran impacto con el tipo de producto.



A continuación, se utilizó la técnica de eliminación de características recursivas, junto a los Random forest, para poder seleccionar las features que son importantes para los clusters de los clientes y de los productos, la razón de por qué esta técnica es usada es sencilla, esto debido a que permite evaluar la importancia de cada feature con la relación del modelo objetivo, logrando eliminar las características menores relevantes, esto se vio cuando se evaluaron las características como el nivel socioeconómico y la densidad de población en las zonas circundantes, para determinar cuáles impactan significativamente en los patrones de de compra.

Finalmente se aplicó la regularización de lasso como paso final para esta selección, generando un umbral para asegurarse que las features menos relevantes no afecten a las que se consideran importantes, especialmente cuando se realiza clustering, ya que eliminar el ruido que se genera y asegura que solo queda la información significante,lo cual nos ayudó en gran medida para identificar los patrones de compra por cada nivel socioeconómico.




```
Características finales seleccionadas: ['iIdContenedor', 'bReturnability', 'Productos_Por_Empaque']
```

Todo este proceso fue importante para poder asegurarse que los clusters creados sean solo aquellos que se consideren relevantes y precisos, lo cual nos permitió identificar quienes eran los clientes objetivos de una manera eficiente y el detectar qué productos serán categorizados como productos de lanzamiento, ya que no solo se optimiza el uso de los datos, sino que también permite comprender los patrones de compra como las características de los productos se alinean con cada segmentación de los clientes.

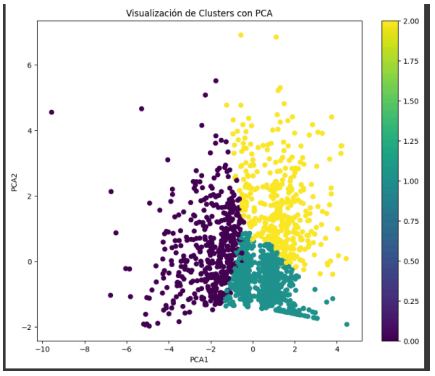
3. Métodos y Técnicas Utilizadas:

Además de los métodos mencionados, también se usaron otros criterios adicionales para probar otros enfoques en la resolución del proyecto, esto se vio en el análisis de productos, donde se seleccionaron variables como contenedor por que se pudo identificar que esta variable puede afectar en las preferencias de los clientes, ya que estos pueden variar dependiendo de las necesidades del cliente, por otra parte, para los clientes, se usó la densidad de población por nivel socioeconómico en zonas cercanas, pues determinamos que esos features tenían importancia en nuestro análisis para determinar el público objetivo.

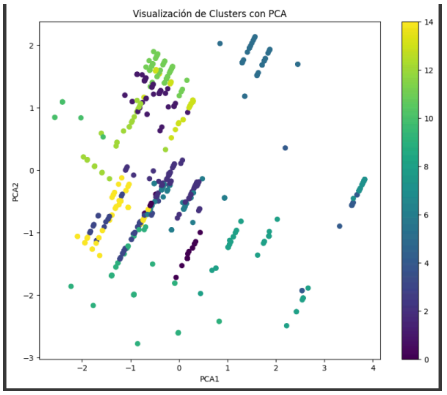
Estos criterios son importantes porque permiten una selección de features más realista, asegurando que el modelo no solo esté correcto estadísticamente, sino también, que sea útil para poder obtener los datos necesarios para la resolución del proyecto.

4. Resultados de Iteración

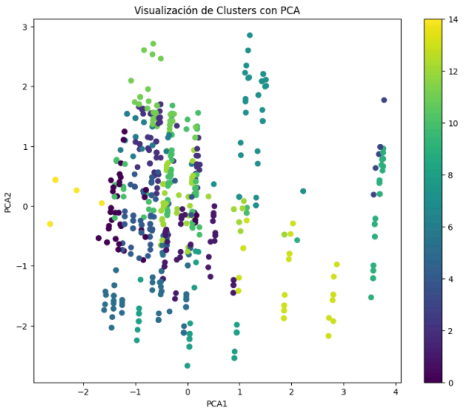
Visualización de clientes por nivel socioeconómico separados en 3 clusters



Clusters de Productos por características mencionadas anteriormente



Si probamos con otro set de features cómo cambiar la categoría del producto por la marca nos saldrá algo de esta maneraque muchos de los datos se pegan más entre sí



Al final el set de features en productos si tendra un mayor impacto al momento de relacionar los productos y puede tener un impacto negativo en nuestros clusters de productos para poder identificar parecidos que convenga más para poder recomendar a diferentes clientes