CODER HOUSE

**Curso de Data Analytics Flex - Comisión 52000**

**Camila Martegani**

**27 de agosto de 2023**

Camila Martegani
Data Analytics Flex - Comisión 52000

# PROYECTO FINAL

## Power BI dashboard

Tabla de contenido

# I. Introducción

## i. Descripción de la temática:

La pandemia de COVID-19 vino a romper los esquemas y nos llevó a cuestionar nuestra visión en una gran cantidad de aspectos, entre ellas, la forma en que nos alimentamos. De pronto, la idea de consumir alimentos ultraprocesados pareció el apocalipsis y la búsqueda de los productos orgánicos, nuestra luz al final del túnel.

Las grandes cadenas de supermercados han identificado esta nueva necesidad y han intentado actuar acorde a ello, pero, ¿han logrado "convencer" al público de que venden productos más sanos?

En el presente proyecto se analizará una base de datos proveniente de la cadena de supermercados Philip Morrison proveniente de Reino Unido que decide ofrecer beneficios al consumo de productos orgánicos para los clientes que forman parte de su programa de fidelidad.

## ii. Objetivo:

- Visualizar y representar con claridad los distintos atributos que caracterizan a la cartera de clientes por medio de un dashboard.
- Segmentar la cartera de clientes de acuerdo a los atributos más relevantes y buscar un algoritmo que permita clasificar eficazmente nuevos clientes que se sumen a la cartera de la empresa (esto se complementará con el resultado del modelo generado en el curso de Data Science de Coderhouse en 2022, disponible en: https://github.com/martca14/Datascience_Coderhouse )

## iii. Usuario final y nivel de impacto:

Este análisis será realizado a nivel estratégico, de gran utilidad para el área de Category Management y/o Analytics tanto de Retailers como de Manufacturers.

Dado que la información recopilada proviene de Reino Unido, es posible que los resultados obtenidos tengan un pequeño sesgo regional/cultural.

iv. Tecnología a utilizar: Microsoft Power BI para el modelado y la generación del dashboard, Python para el modelo de machine learning y Adobe Photoshop para la edición de imágenes.

v. Glosario:
- *Machine Learning (ML)*: disciplina dentro de la Inteligencia Artificial que, a través de algoritmos, brinda a las computadoras la capacidad de identificar patrones en datos masivos y elaborar predicciones.
- *Cluster*: agrupación de conceptos, elementos u objetos que tienen algo en común.
- *Branding:* conjunto de estrategias, iniciativas y acciones orientadas a la gestión de una marca y su repercusión en la sociedad

vi. Paleta de colores elegida:

#8031A8 Color de marca

Paleta elegida en Power BI para el tablero (gráficos, cuadros de texto, menús, botones, etcétera)

#7732A6 #9B638F #AB2FBD #B32D8F #342DB3 #C2A0D9 #E7D5F2 #F2F2F2

vii. Tabla de versionado

| VERSIÓN | ACCIONES |
|---|---|
| V01 | Definición de la temática bajo la metodología SMART, elección de la base de datos y presentación de la documentación inicial |

| | |
|---|---|
| V02 | Definición del DER y documentación respaldatoria |
| V03 | Primer prototipo de tablero en Power BI, transformación y modelado de los datos. Tabla calendario y de medidas calculadas |
| Versión final | Definición de la paleta de colores y el branding de la compañía. Agregado de filtros y botones. Redistribución de gráficos en las distintas pestañas. Adición de KPIs. |

## II. Base de datos

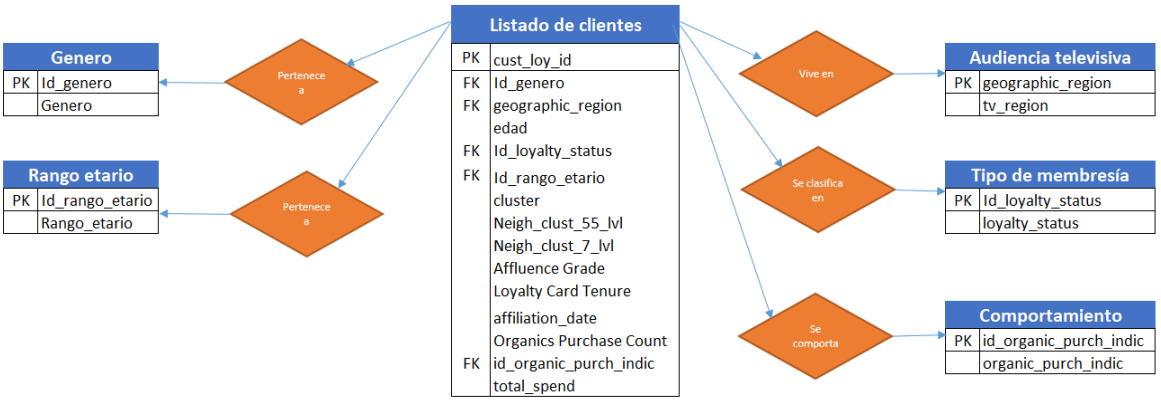
### i. Descripción del Dataset:

Como se mencionó anteriormente, en el presente proyecto se analiza una base de datos proveniente de una cadena de supermercados de Reino Unido llamada Philip Morrison. El dataset fue extraído del repositorio Kaggle .

A los fines prácticos de este curso, se generaron id 's adicionales para las siguientes variables: género, edad, nivel de fidelidad (loyalty_status) e indicador de compra de orgánicos (organic_purchase_indic), y se separó la base en varias tablas para poder linkearlas entre sí y armar el modelo de datos.

También se utilizó la clasificación de clientes (cluster) generada en el curso de Data Science mencionado más arriba.

ii. Diagrama de entidad - relación (DER):




iii. Listado de tablas:

a. Listado de clientes: contiene la información elemental de cada cliente
   i. pk: cust_loy_id
   ii. fk: Id_genero
   iii. fk: geographic_region
   iv. fk: Id_loyalty_status
   v. fk: Id_rango_etario
   vi. fk: id_organic_purch_indic

b. Comportamiento: indica si el cliente consume productos orgánicos o no
   i. pk: id_organic_purch_indic

c. Género: permite identificar el género de cada cliente
   i. pk: Id_genero

d. Rango etario: contiene los distintos rangos de edad comprendidos en la base de datos
   i. pk: Id_rango_etario

e. Audiencia televisiva: vincula la región geográfica del cliente con su respectiva región audiovisual (útil para campañas publicitarias regionales)
   i. pk: geographic_region

f. Tipo de membresía: identifica los distintos niveles de afiliación al programa de beneficios del supermercado
   i. pk: Id_loyalty_status

iv. Detalle de columnas:

a. Listado de clientes:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| cust_loy_id | código de identificación del cliente | int | PK |
| affiliation_date | año de afiliación al programa de lealtad de clientes | int | |
| Id_genero | código de identificación de cada género | int | FK |
| geographic_region | región geógrafica donde habita el cliente | varchar(10) | FK |
| Id_loyalty_status | código de identificación del tipo de miembro del programa de lealtad | int | FK |
| Neigh_clust_55_lvl | grupo vecindario al que pertenece el cliente | float | |
| Neigh_clust_7_lvl | tipo de barrio al que pertenece el cliente | varchar (1) | |
| tv_region | región televisiva a la que pertenece el cliente | varchar(12) | |
| Id_rango_etario | código de identificación del rango etario al que pertenece el cliente | int | FK |
| Edad | edad del cliente | int | |
| Loyalty Card Tenure | cantidad en meses como miembro del programa de lealtad | int | |
| Organics Purchase Count | cantidad de productos orgánicos adquiridos por el cliente | int | |
| id_organic_purch_indic | código de identificación de acuerdo a si el cliente compra o no productos orgánicos | int | FK |
| total_spend | monto total gastado en la tienda | int | |
| cluster | segmento al que pertenece el cliente (luego de clasificarlo por medio de un k-means) | int | |

b. Comportamiento:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| id_organic_purch_indic | código de identificación de acuerdo a si el cliente compra o no productos orgánicos | int | PK |
| organic_purch_indic | indicador de compra de productos orgánicos | varchar (19) | |

c. Género:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| Id_genero | código de identificación de cada género | int | PK |
| Genero | género de cada cliente | varchar(9) | |

d. Rango etario:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| Id_rango_etario | código de identificación del rango etario al que pertenece el cliente | int | PK |
| Rango_etario | rango etario al que pertenece el cliente | varchar(9) | |

e. Audiencia televisiva:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| geographic_region | región geógrafica donde habita el cliente | varchar(10) | PK |
| tv_region | región televisiva a la que pertenece el cliente | varchar(12) | |

f. Tipo de membresía:

| VARIABLE | DESCRIPCIÓN | Tipo de dato | OBS |
|---|---|---|---|
| Id_loyalty_status | código de identificación del tipo de miembro del programa de lealtad | int | PK |
| loyalty_status | tipo de miembro del programa de lealtad | varchar(9) | |

## III. Visualización

### i. Estructura inicial del dato

Se cargan todos los archivos de datos en formato csv a la plataforma Power BI y por medio de Power Query se hacen las transformaciones necesarias.
En todas las tablas se establece la primera fila como encabezado. Luego se realizan las siguientes transformaciones sobre la tabla de **Listado_de_clientes**:

- *Modifico tipo de columna de texto a número* = Table.TransformColumnTypes(#"Encabezados promovidos",{{"cust_loy_id", Int64.Type}, {"affiliation_date", Int64.Type}, {"Id_genero", Int64.Type}, {"geographic_region", type text}, {"Id_loyalty_status", Int64.Type}, {"Neigh_clust_55_lvl", Int64.Type}, {"Neigh_clust_7_lvl", type text}, {"tv_region", type text}, {"Id_rango_etario", Int64.Type}, {"Edad", Int64.Type}, {"Loyalty Card Tenure", Int64.Type}, {"Organics Purchase Count", Int64.Type}, {"id_organic_purch_indic", Int64.Type}, {"total_spend", type number}, {"cluster", Int64.Type}}),
- *Defino affiliation_date como texto* = Table.TransformColumnTypes(#"Tipo cambiado",{{"affiliation_date", type text}}),
- *Defino affiliation_date como fecha* = Table.TransformColumnTypes(#"Tipo cambiado1",{{"affiliation_date", type date}}),
- *Creo una columna nueva referente al país* = Table.AddColumn(#"Tipo cambiado2", "Literal", each "UK", type text),
- *Renombro la nueva columna creada*= Table.RenameColumns(#"Literal insertado",{{"Literal", "Country"}}),
- *Reordeno columnas* = Table.ReorderColumns(#"Columnas con nombre cambiado",{"cust_loy_id", "affiliation_date", "Id_genero", "geographic_region", "Country", "Id_loyalty_status", "Neigh_clust_55_lvl", "Neigh_clust_7_lvl", "tv_region", "Id_rango_etario", "Edad", "Loyalty Card Tenure", "Organics Purchase Count", "id_organic_purch_indic", "total_spend", "cluster"}),

- *Modifico el tipo de algunas columnas* = Table.TransformColumnTypes(#"Columnas reordenadas",{{"total_spend", Currency.Type}, {"affiliation_date", type date}})

ii. Medidas y columnas calculadas

- Creación de la tabla calendario:

```
Calendario =
calendar(
        MIN(Listado_de_clientes[affiliation_date]),
        max(Listado_de_clientes[affiliation_date]
        ))
```

- Creación de columnas calculadas:
  1. Se añade la columna Antiguedad a la tabla Listado_de_clientes.

```
Antiguedad = DATEDIFF(Listado_de_clientes[affiliation_date],TODAY(),YEAR)
```

- Creación de la tabla de medidas:
  1. Recuento del total de clientes de la base

```
Cantidad_total_clientes =
        DISTINCTCOUNT(Listado_de_clientes[cust_loy_id])
```

  2. Cálculo de la edad promedio de los clientes

```
Edad_Promedio = AVERAGE((Listado_de_clientes[Edad]))
```

  3. Cálculo del gasto promedio de los clientes

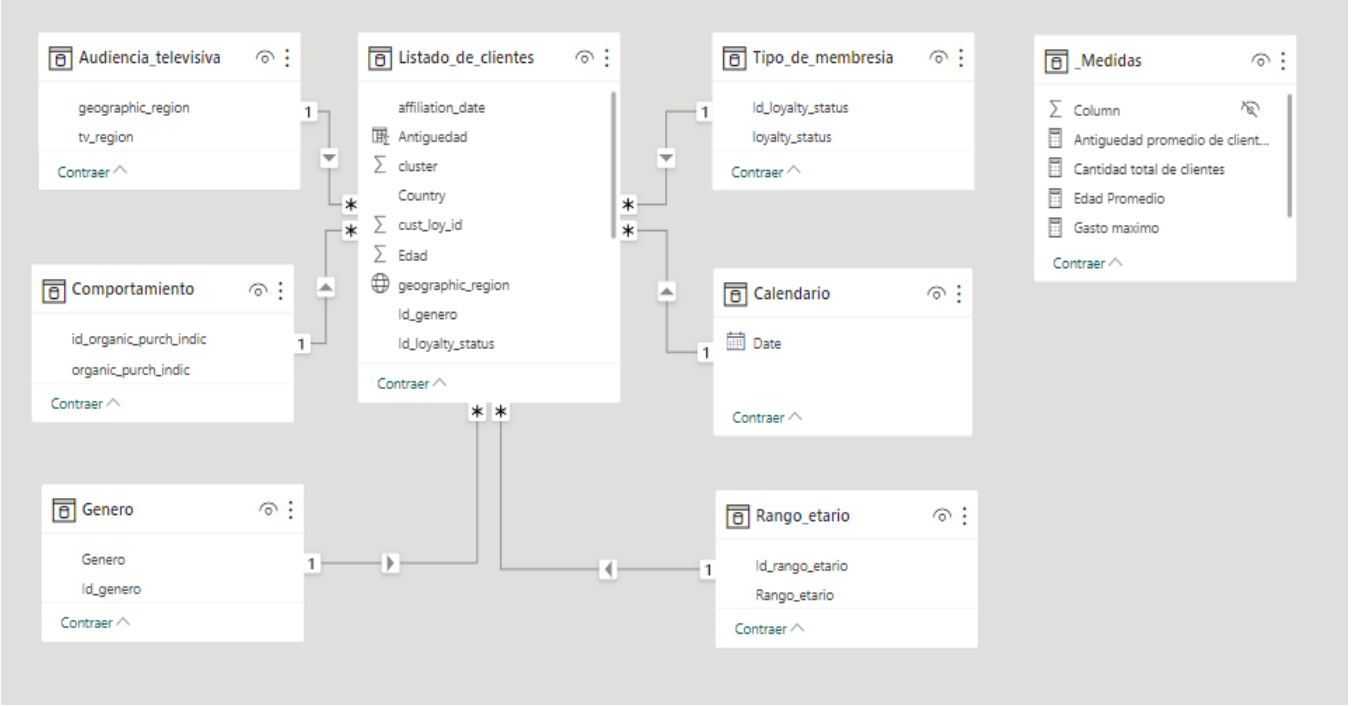
```
Gasto_promedio= AVERAGE(listado_de_Clientes[total_spend])
```

  4. Cálculo de la antigüedad promedio de los clientes expresado en años

```
Antiguedad promedio de clientes = AVERAGE(Listado_de_clientes[Antiguedad])
```

5. Cálculo del gasto máximo de los clientes
   Gasto maximo = max(Listado_de_clientes[total_spend])

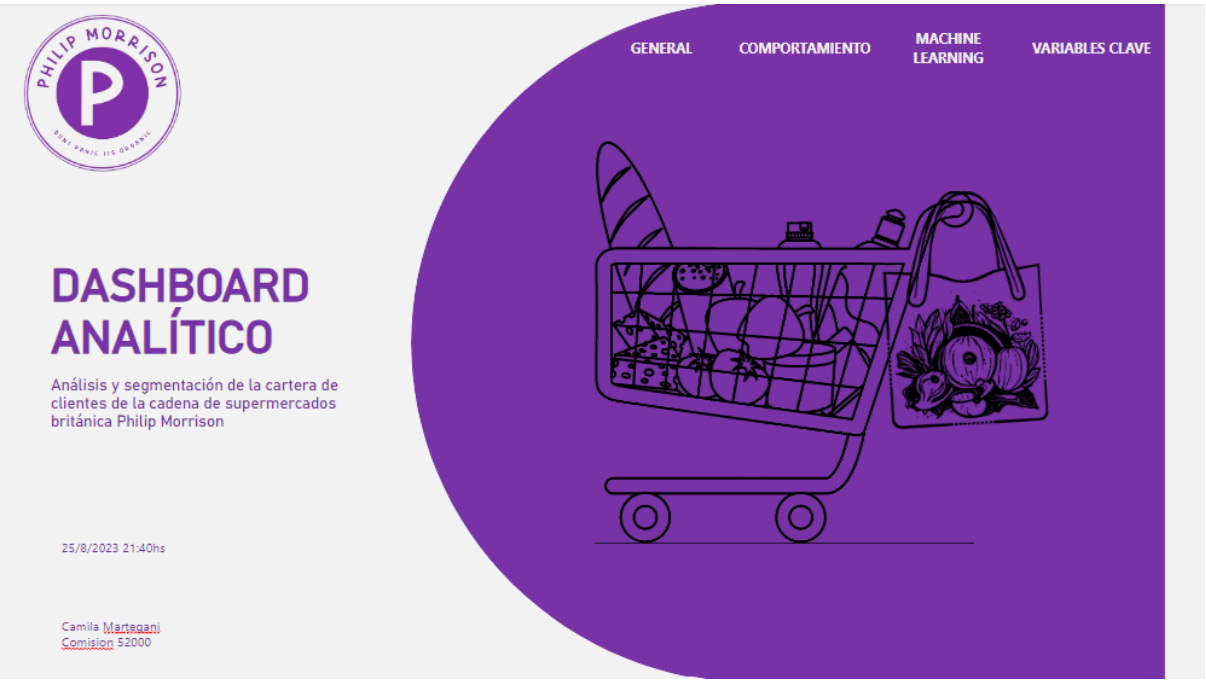
iii. Diagrama de entidad - relación (DER) en Power BI



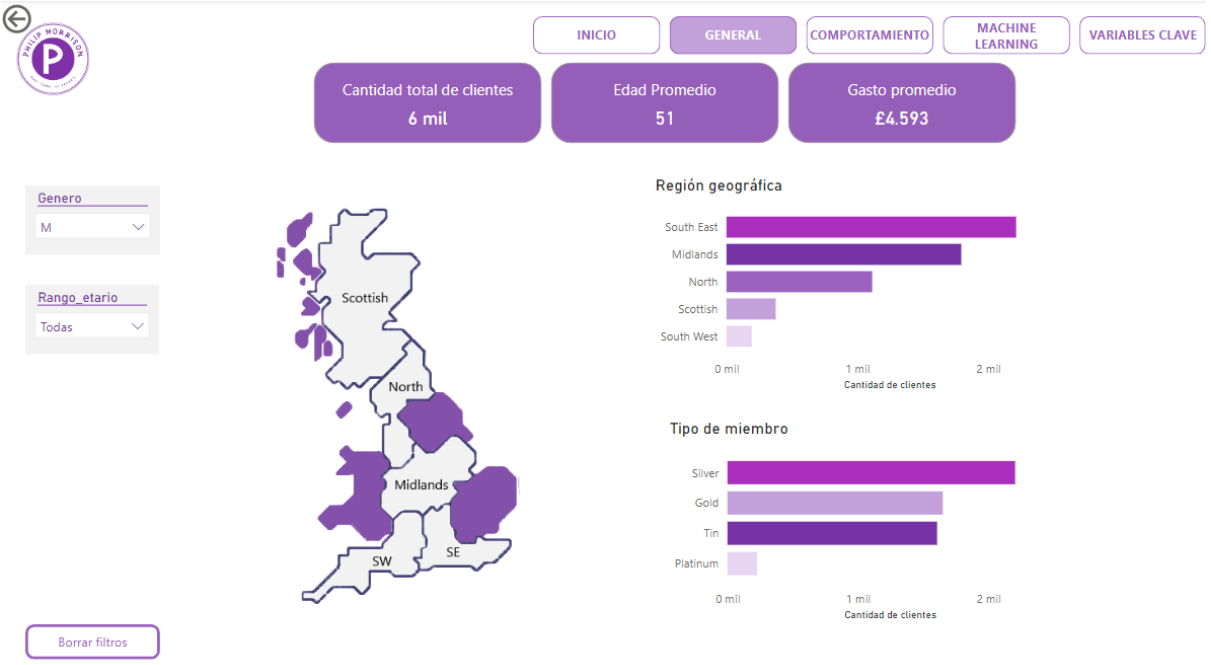

a. Listado de clientes: contiene la información elemental de cada cliente
b. Comportamiento: indica si el cliente consume productos orgánicos o no
c. Género: permite identificar el género de cada cliente
d. Rango etario: contiene los distintos rangos de edad comprendidos en la base de datos
e. Audiencia televisiva: vincula la región geográfica del cliente con su respectiva región audiovisual
f. Tipo de membresía: identifica los distintos niveles de afiliación al programa de beneficios del supermercado
g. _Medidas: Tabla de medidas calculadas en base a algunas variables del dataset, que se utilizará luego para mostrar indicadores clave en el tablero
h. Calendario: tabla de calendario generada a partir de la fecha de afiliación de los clientes

iv. Dashboard

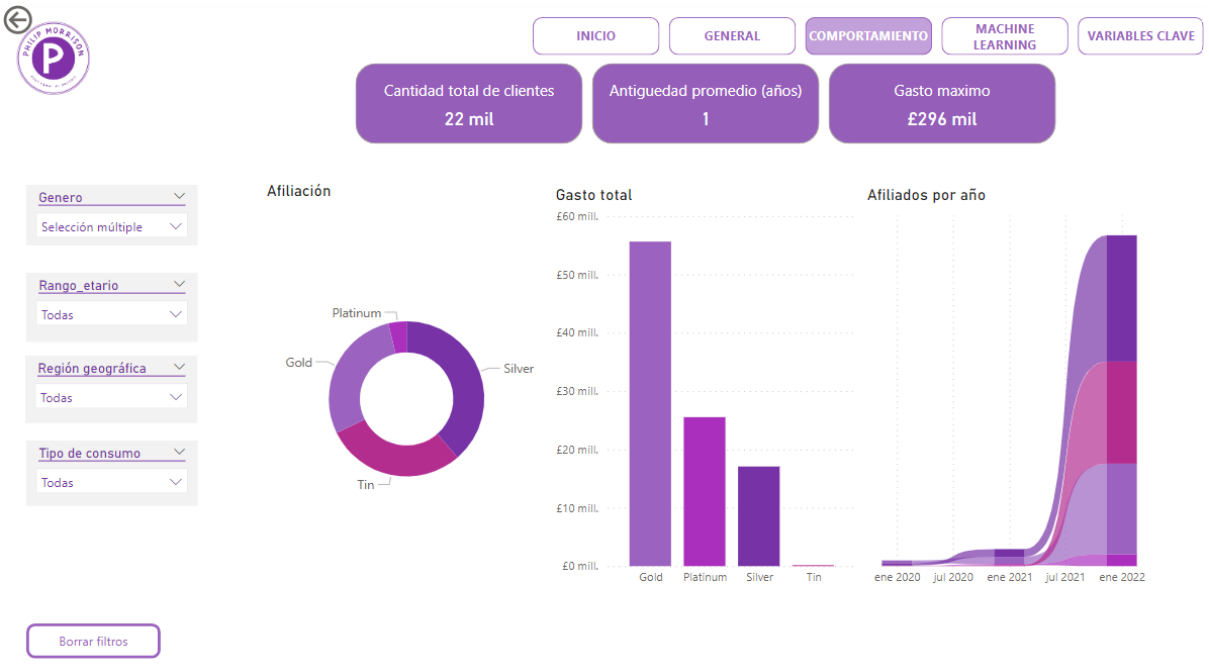
1. Índice: en esta pestaña se detallan los datos del alumno, logo y branding de la empresa. También se encuentran los botones a cada pestaña del dashboard y la fecha de la última actualización del mismo.



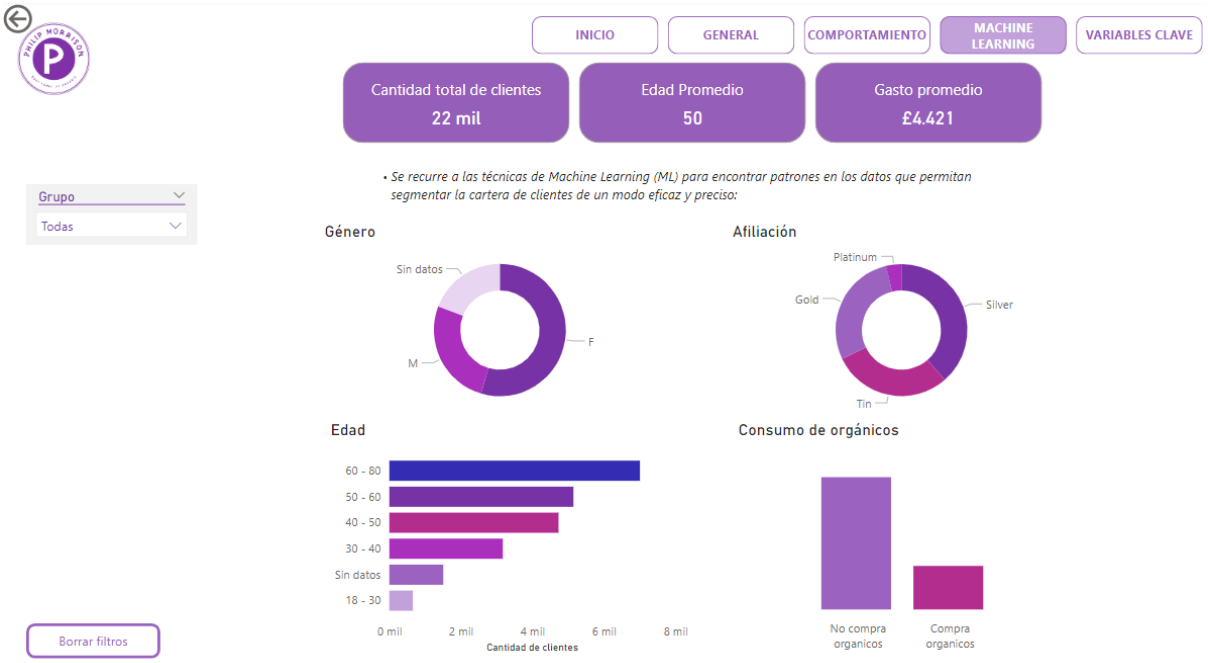

2. General: aquí se encuentran los KPIs principales sobre el dataset así como algunos gráficos de barra que muestran un pantallazo general de las características de los clientes. Un mapa acompaña a los gráficos para poder ilustrar más claramente la distribución geográfica de los clientes.
   También es posible filtrar los gráficos por género y rango etario.




3. Comportamiento: aquí se pueden ver más detalladamente las características de los clientes, cómo se distribuyen de acuerdo a lo que gastan, al nivel de afiliación y, gracias a los segmentadores de la izquierda, es posible filtrar los datos para entender las diferencias entre los que consumen productos orgánicos y los que no lo hacen. También se puede filtrar por ubicación geográfica.




4. Machine Learning: en esta pestaña se utiliza como único segmentador la columna cluster, la cual fue generada en un paso previo que consistió en correr un modelo de clasificación de datos llamado K-Means. Este modelo encontró patrones en la data que permitieron diferenciar a los clientes y reagruparlos en 3 segmentos distintos.
   Los gráficos muestran los principales atributos de la base.



5. Variables clave: luego del primer modelo de ML, se corrió un segundo paso (un modelo predictivo denominado Random Forest) en el que se pudo definir cuáles son las variables más relevantes a la hora de clasificar a los clientes dentro de un grupo u otro, y también clasificar a los nuevos clientes que se afilien al programa. El gráfico muestra la importancia de dichas variables aunque en este caso **NO** es interactivo. Por cuestiones de alcance del proyecto, no fue posible añadir a Power BI el dataset con los resultados del segundo modelo. Para futuras implementaciones, sería óptimo poder agregar dicho resultado y extender el análisis.


v. Conclusiones:

- ✓ Es posible encontrar patrones que distinguen a los clientes de la cartera, los cuales permiten segmentarla de una manera eficiente y con una gran exactitud.
- ✓ Las variables utilizadas para definir cada grupo son fáciles de obtener y no representan mayores costos para el detallista.
- ✓ Con los resultados obtenidos podrían ejecutarse acciones de marketing enfocadas en las características de cada uno de los segmentos, que luego podrían traducirse en mayores ingresos para el negocio y mayor visibilidad de la empresa.

vi. Futuras líneas

Con el fin de mejorar la calidad y profundidad del análisis, sería ideal recopilar otras variables que complementen el dataset, como por ejemplo, un detalle de los ítems que compra cada cliente,para poder construir un panel de datos de hogar y entender cómo se comportan los hogares. De este modo, se podrían enfocar las campañas de marketing más eficientemente.

También sería útil contar con información de campañas publicitarias en tv, redes, etcétera, para poder estimar ROI y otras variables relevantes que ayuden en la toma de decisiones estratégicas de la compañía.