UNIVERSIDAD ANDRÉS BELLO
Facultad de Ingeniería
Escuela de Informática

HITO 1 PROYECTO: ANÁLISIS Y PREPROCESAMIENTO DE DATOS

Proyecto – T1: Ciencia de Datos
Ingeniería Civil Informática

Alumnos:
Paulo Jiménez Cisternas
Francisco Miranda Urrutia
Constanza Pérez Pizarro
Carlos Tobar Olivo

Profesor: Billy Peralta

Santiago – Chile
Mayo, 2019

1. Definición del proyecto

1.1. Contexto

> “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on April 26, 2019.

Instacart es una aplicación de pedido y entrega de comestibles que permite comprar en línea en tiendas locales. Los clientes seleccionan comestibles a través de una aplicación web de varios minoristas y el pedido es entregado por un comprador personal.

Instacart dispone de un conjunto de datos de compras de comestibles en línea de 2017 en línea para uso no comercial, el cual será utilizado para este proyecto. Este conjunto de datos contiene una muestra de más de 3 millones de pedidos de más de 200,000 usuarios dentro del año 2017.

Dada la estructura del dataset, es posible emplear algoritmos de predicción. Uno de los trabajos más conocidos en este dataset fue la competencia de análisis de canasta de mercado hecha por Kaggle, comunidad dedicada a la ciencia de datos, donde el objetivo era predecir qué producto comprará de nuevo un consumidor de Instacart y cuándo.

1.2. Descripción del proyecto

El proyecto consiste en encontrar itemsets frecuentes dentro del conjunto de datos utilizando el algoritmo *FP-Growth*. Se implementarán estrategias para mejorar la velocidad de éste por medio otras técnicas, como particiones aleatorias y *K-Means*, donde ésta última se aplicará dos veces: una con el algoritmo sin modificar y la otra aplicando el algoritmo de forma repetida, de forma tal que en cada iteración trabaje con el *cluster* de mayor varianza, removiendo los demás datos y volviendo a aplicar *K-Means* con K = K - 1, hasta que queden K = 2 *clusters*.

Como primera instancia, el presente informe mostrará un análisis numérico del comportamiento general de las variables, tanto como para ellas por sí solas como para las relaciones entre pares de éstas, así como también el detalle del preprocesamiento de los datos, definiendo los criterios utilizados para realizar las modificaciones.

La ejecución del programa fue realizada en sistema operativo Windows.

2. Descripción de los datos

La base de datos consta de seis tablas descritas a continuación:

- *orders*: pedidos (3.4m de filas, 206k usuarios).
    - *order_id*: identificador del pedido.
    - *user_id*: identificador del cliente.
    - *eval_set*: en qué set de evaluación pertenece este pedido (ver SET descrito más adelante).
    - *order_number*: número de secuencia de pedido para este usuario (1=primero, n=n-ésimo).
    - *order_dow*: el día de la semana en que se realizó el pedido (0=Sunday, 1=Monday, …, 6=Saturday).
    - *order_hour_of_day*: la hora del día en que se realizó el pedido.
    - *days_since_prior*: días desde el último pedido, con un límite máximo de 30 (con NA para order_number = 1).

- *products*: productos (50k filas).
    - *product_id*: identificador del producto.
    - *product_name*: nombre del producto.
    - *aisle_id*: llave foránea (identificador del pasillo).
    - *department_id*: llave foránea (identificador del departamento).

- *aisles*: pasillos (134 filas).
    - *aisle_id*: identificador del pasillo.
    - *aisle*: nombre de pasillo.

- *departments*: departamentos (21 filas).
    - *department_id*: identificador del departamento.
    - *department*: nombre del departamento.

- *order_products__SET*: productos por orden (dividido en tablas *order_products__PRIOR* y *order_products__TRAIN*) (30m+ filas).
    - *order_id*: llave foránea (identificador del pedido).
    - *product_id*: llave foránea (identificador del producto).
    - *add_to_cart_order*: orden secuencial en que el producto fue agregado al carro de compras.
    - *reordered*: 1 si este producto ha sido solicitado por este usuario en el pasado, 0 de lo contrario.

donde *SET* es es uno de los tres conjuntos de evaluación siguientes (*eval_set* en *orders*):

- *"prior"*: pedidos anteriores al pedido más reciente de los usuarios. (~3.2m pedidos)
- *"train"*: datos de entrenamiento suministrados a los participantes. (~131k pedidos)
- *"test"*: datos de prueba reservados para competencias de aprendizaje automático. (~75k pedidos)

El modelo entidad-relación correspondiente a este dataset es el siguiente (ver Figura 1):

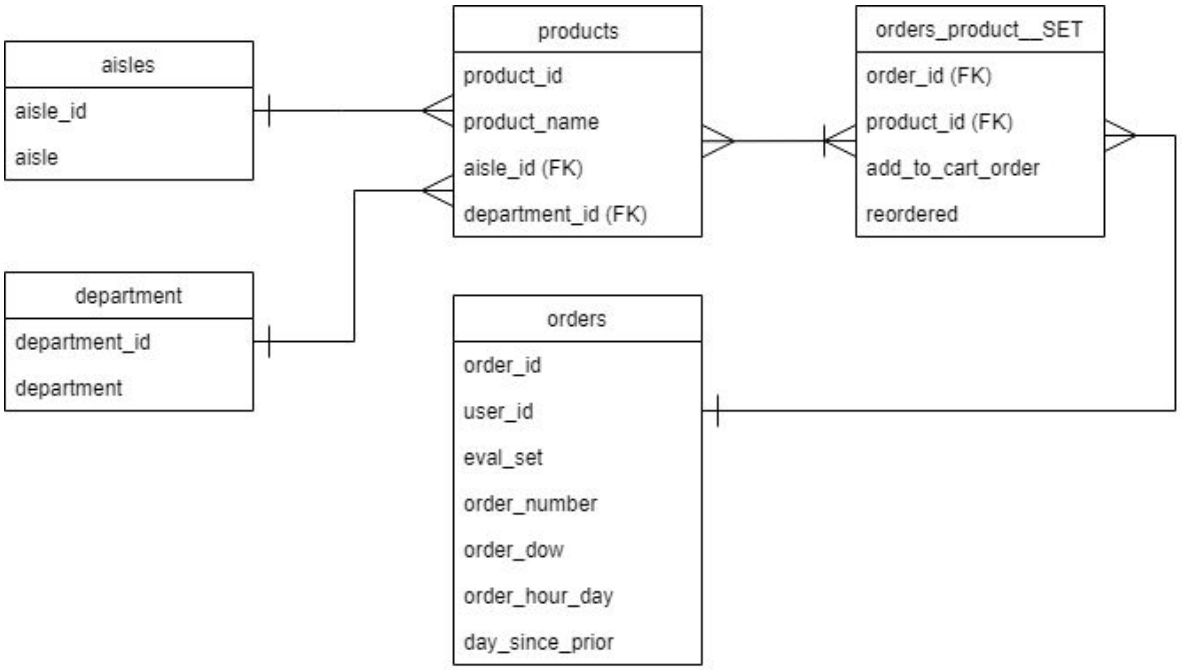

Figura 1. Modelo entidad-relación del conjunto de datos de compras de comestibles en línea de Instacart 2017.

Para términos de este trabajo, se utilizó sólo algunas tablas y se eliminaron algunas columnas, lo cual se detalla más en el apartado de preprocesamiento.

3. Análisis 1D y 2D de datos

Una análisis de los datos permite una mejor comprensión del contexto y situación en que se sitúa la entidad. Para esta sección, se tomó en consideración el conjunto de datos resultante luego del preprocesamiento.

3.1. Análisis 1D

El análisis 1D se manifiesta como una representación numérica o categórica del comportamiento general de una variable, donde se puede utilizar métricas tales como media, mediana, moda, etc. según sea el caso.

3.1.1. Moda de los productos

Del conjunto total de productos se muestran (ver Figura 2 y 3) la moda y los diez productos más vendidos son los siguientes:

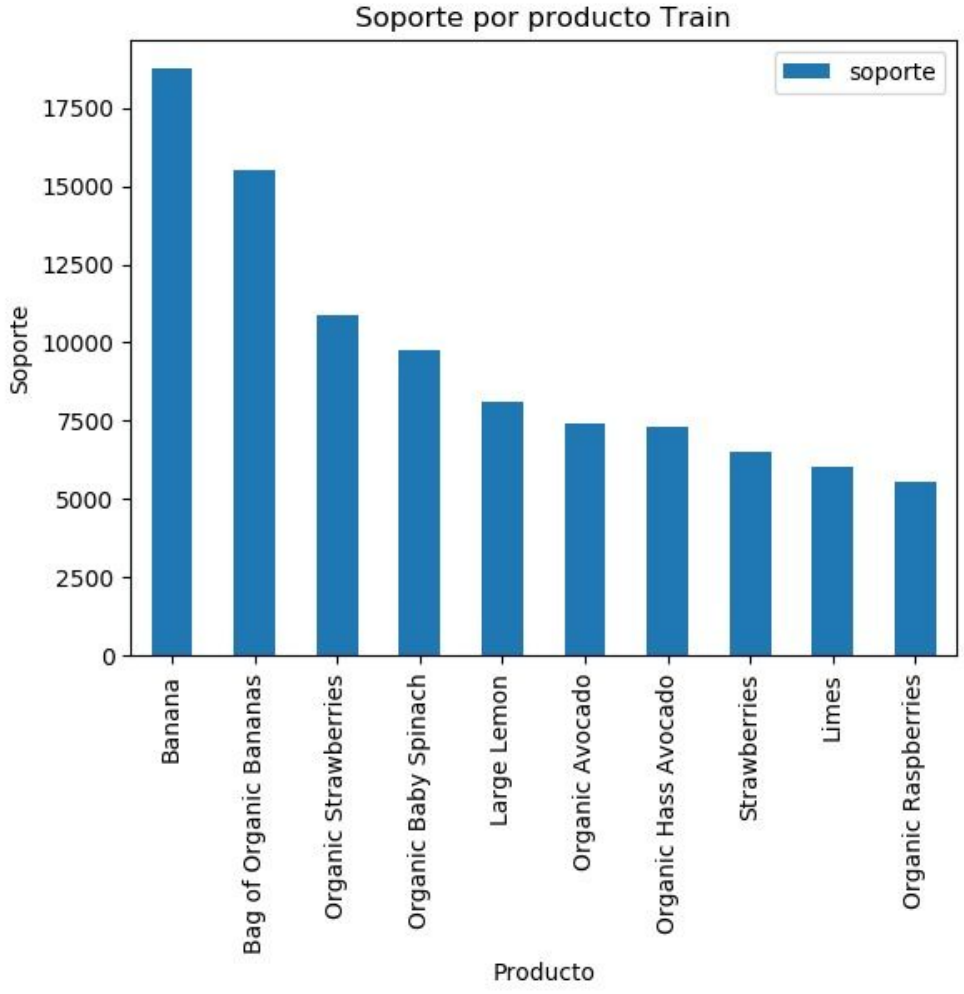

Figura 2. Histograma de los diez productos más comprados en tabla *order_products__TRAIN*.

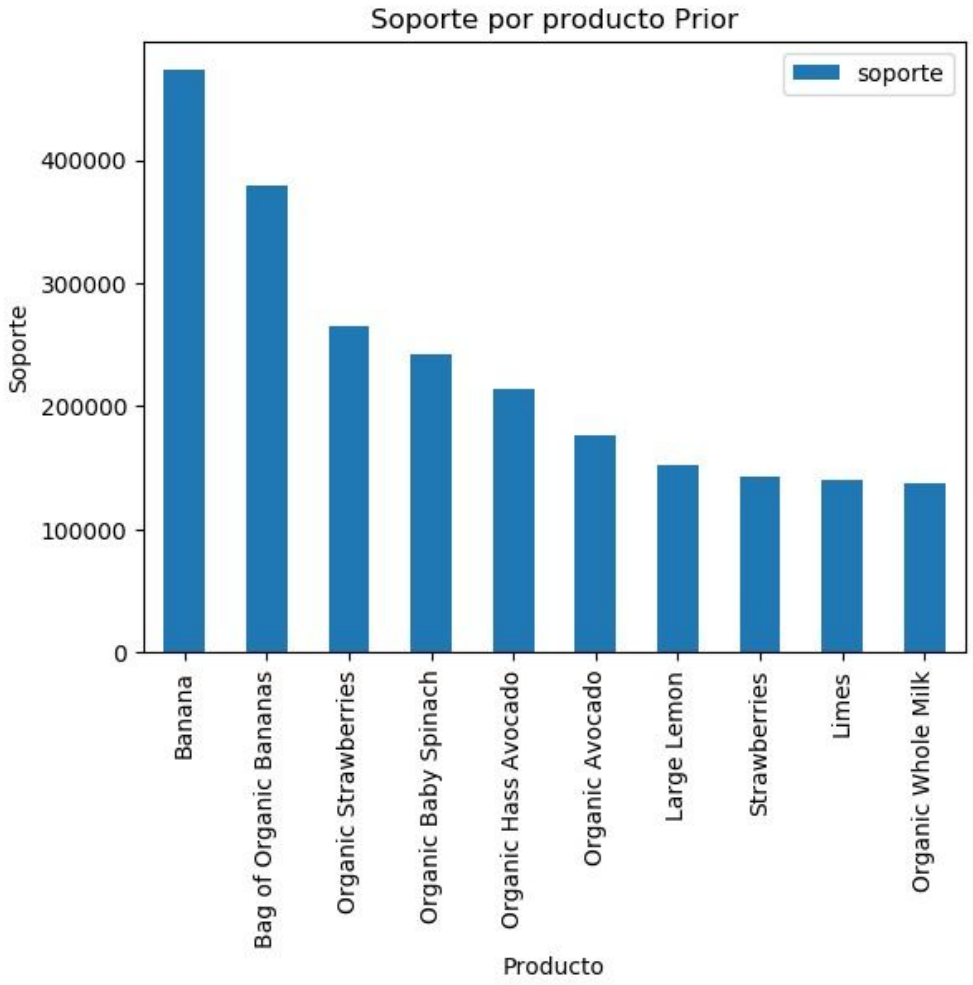

Figura 3. Histograma de los diez productos más comprados en tabla *order_products__PRIOR*.

De lo anterior se puede observar que el producto Banana es el que más se compra para ambos conjuntos de datos, donde el valor de *soporte* de Banana para *TRAIN* y *PRIOR* respectivamente es **18726** y **472565**.

De las tablas *products* y *order_products__TRAIN/PRIOR* se realizó un *merge* a través de la columna *id_producto*, presente en ambas, de modo que se muestre con los nombres de los productos. A este nuevo dataframe se le aplicó la función *value_counts()* de *pandas* en la columna que contiene los nombres de los productos.

3.1.2. Porcentaje de productos comprados y no comprados

El siguiente análisis comprende la comparativa de productos comprados y nunca comprados del conjunto total de productos (ver Figura 4 y 5):

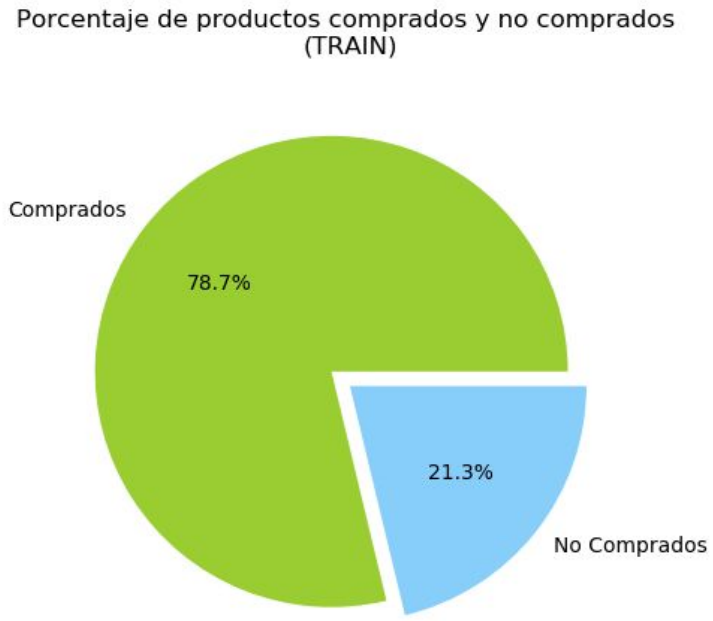

Figura 4. Gráfico de porcentaje de productos comprados y no comprados en tabla *order_products__TRAIN* con respecto al conjunto total de productos.

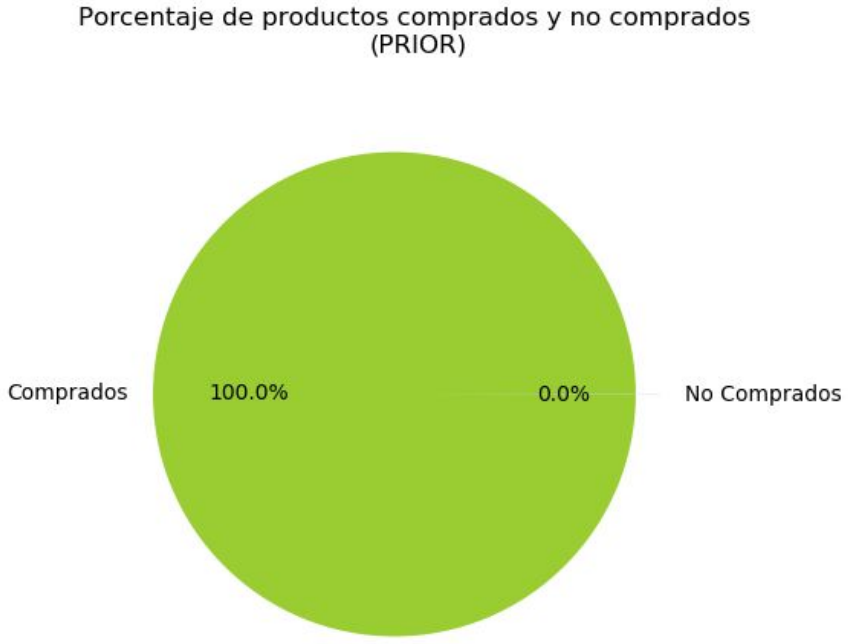

Figura 5. Gráfico de porcentaje de productos comprados y no comprados en tabla *order_products__PRIOR* con respecto al conjunto total de productos.

Donde para ambos *dataset*, *TRAIN* y *PRIOR*, se obtuvo respectivamente: **39123** comprados y **10565** no comprados; **491677** comprados y **11** no comprados. Es por esto último que el gráfico de torta para *prior* muestra un 0,0% para no comprados, dada la gran diferencia entre los casos comprado y no comprados.

Para llegar a estos valores se utilizó la función *isin()* de *pandas*, comparando la columna *id_producto* de la tabla *products* con la columna *id_producto* de la tabla

orders_products__TRAIN/PRIOR, de modo de obtener un booleano para cada producto y aplicar un conteo con *value_counts()*.

3.2. Análisis 2D

3.2.1. Tablas de contingencia

Una tabla de contingencia muestra la relación entre una regla de asociación X → Y, donde, en este caso, es la relación entre dos productos. La tabla de contingencia muestra el *soporte* para el caso en que el pedido contiene ambos productos, uno sí y el otro no (viceversa) y cuando no contiene ninguno. Esto se puede ver representado en la siguiente tabla (ver Tabla 1):

	Y	~Y	
X	F11	F10	F11 + F10
~X	F01	F00	F01 + F00
	F11 + F01	F10 + F00	Total de instancias

Tabla 1. Tabla de contingencia para dos elementos.

Donde:

F11: soporte de X e Y

F10: soporte de X e ~Y

F01: soporte de ~X e Y

F00: soporte de ~X e ~Y

La cantidad de combinaciones posibles entre dos productos está dada por:

$$\sum_{i=0}^{n-1}(n-1)-i$$

con n la cantidad de productos. Como existen 49688 productos, el número de tablas de contingencia generadas será 1234473516, valor que se escapa con creces al poder de procesamiento de las computadoras a disposición. Es por esto que se decidió realizar un conteo de soporte de cada producto y trabajar con los que presenten mayor número, en este caso los 10 más comprados (45 combinaciones). Las tablas de contingencia mostradas en este informe serán las cinco que presenten un mayor coeficiente de correlación phi (*ø-coefficient*), es decir, los que tengan valores más cercanos a 1. El cálculo de coeficiente de correlación phi está dado por:

$$ø - coefficient = \frac{P(X,Y) - P(X)*P(Y)}{\sqrt{P(X)*[1-P(X)]*P(Y)*[1-P(Y)]}}$$

El tiempo de procesamiento sólo para obtener las tablas de contingencia para la tabla order_products__TRAIN fue de 1 hora 45 minutos aproximadamente con 1.384.617

registros. Como la magnitud de registros en las tablas *order_products__PRIOR* es 32.434.489, más de 20 veces la magnitud de *TRAIN*, por lo que se optó sólo por generar las tabla de contingencia para *TRAIN*.

La siguientes tablas (ver Tabla 2 a 6) muestran las cinco tablas de contingencia para dos productos de mayor coeficiente phi:

	Limes	~Limes
Large Lemons	1595	6540
~Large Lemons	4438	118636

$$ø - coefficient = 0.18423$$

Tabla 2. Tabla de contingencia de Large Lemons, Limes y su soporte.

	Organic Raspberries	~Organic Raspberries
Organic Strawberries	1670	9224
~Organic Strawberries	3876	116439

$$ø - coefficient = 0.166$$

Tabla 3. Tabla de contingencia de Organic Strawberries, Organic Raspberries y su soporte.

	Organic Hass Avocado	~Organic Hass Avocado
Bag of Organic Bananas	2420	13060
~Bag of Organic Bananas	4873	110856

$$ø - coefficient = 0.16082$$

Tabla 4. Tabla de contingencia de Bad of Organic Bananas, Organic Hass Avocado y su soporte.

	Organic Strawberries	~Organic Strawberries
Bag of Organic Bananas	3074	12406
~Bag of Organic Bananas	7820	107909

$$ø - coefficient = 0.15316$$

Tabla 5. Tabla de contingencia de Bag of Organic Bananas, Organic Strawberries y su soporte.

	Organic Raspberries	~Organic Raspberries
Bag of Organic Bananas	1780	13700
~Bag of Organic Bananas	3766	111963

$$ø - coefficient = 0.13218$$

Tabla 6. Tabla de contingencia de Bag of Organic Bananas, Organic Raspberries y su *soporte*.

3.2.2. Soporte, confianza y lift para pares de productos

Otra forma de encontrar una relación entre productos es calcular *soporte*, *confianza* y *lift* para un *itemset* de dos elementos. Para esto, se aplicó algoritmo *apriori* por medio de la librería *apryori*, de modo de encontrar los *itemset* frecuentes según un valor de *soporte* mínimo, *confianza* mínima, *lift* mínimo, que para este caso fue de 0.0045 para cada parámetro, además de un tamaño de *itemset*, lo cual para este caso son dos elementos. El *soporte* está expresado como el porcentaje con respecto al total de pedidos.

La siguiente tabla (ver Tabla 1) muestra los 10 pares de productos comprados con mayor valor de *soporte*:

Itemset (ID)	Soporte	Confianza	Lift
{13176, 21137}	0,02342827	0,19857881	2,39171354
{13176, 47209}	0,01844386	0,15633075	2,81256017
{13176, 21903}	0,01704151	0,14444444	1,93708208
{24852, 47766}	0,01688909	0,11833814	2,09569833
{21137, 24852}	0,01656899	0,19955939	1,39826915
{47626, 24852}	0,01644704	0,11524084	1,85871366
{24852, 21903}	0,01524286	0,20441537	1,43229395
{24852, 16797}	0,01484654	0,2999692	2,10181881
{13176, 27966}	0,01356614	0,11498708	2,72040025
{21137, 27966}	0,01272779	0,15329539	3,62671026

Tabla 7. Tabla de los 10 *itemsets* frecuentes de dos productos de mayor soporte con su respectivo valor de *soporte*, *confianza* y *lift*.

4. Preprocesamiento

Como el objetivo del trabajo consiste en encontrar las reglas de asociación relativas a *itemset* frecuentes, existen datos dentro del dataset que no serán utilizados. Las tablas que se apartarán del análisis y del proceso son:

- *aisles:* no es relevante para la obtención de itemset de productos frecuentes.
- *departments:* no es relevante para la obtención de itemset productos frecuentes.
- *orders:* si bien contiene información de los pedidos, no son relevantes para los algoritmos.

Dada la magnitud de los datos, es importante acotarse a los datos que realmente aportan al algoritmo y al objetivo del trabajo, de modo de no realizar un sobrecosto de memoria. Junto con esto, es necesario eliminar los atributos no utilizados:

- *aisle_id* y *department_id* en tabla *products*: no hay tabla a la cual puedan referenciar.
- *add_to_cart_order en tabla order_products__TRAIN/PRIOR*: el orden secuencial es irrelevante para los algoritmos a utilizar.
- *reordered en tabla order_products__TRAIN/PRIOR*: dato irrelevante para el caso. Es útil en algoritmos de predicción.

Por consecuencia, el modelo entidad-relación resultante es el siguiente (ver Figura 6):

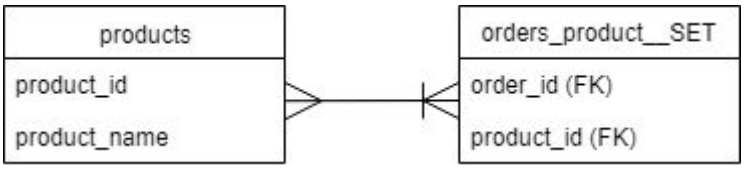

Figura 6. Modelo entidad-relación del conjunto de datos de compras de comestibles en línea de Instacart 2017 luego de la fase de preprocesamiento.

De las tablas *order_products__TRAIN/PRIOR* se revisó si existen atributos con datos faltantes y se determinó que ninguna de estas contiene datos faltantes.