-

Estudio exploratorio: Análisis de Correspondencia Múltiple y de Clúster aplicado a las importaciones argentinas en el año 2017

Álvarez Eugenia
Laguna Nicolás
Mosconi Florencia
Quispe Santiago

Tutor: Dr. David Giuliodori

Universidad Nacional de Córdoba

Diplomatura en Ciencias de Datos, Inteligencia Artificial y sus Aplicaciones en Economía y Negocios

2 de diciembre de 2022

1. Introducción y planteamiento del problema

El objetivo del presente trabajo es, mediante un estudio exploratorio aplicando dos métodos multivariados, determinar si existe una relación entre la rama de actividad, el medio de transporte utilizado, el peso y la procedencia de las importaciones realizadas por diversas empresas localizadas en la República Argentina durante el año 2017.

Se aplicará un Análisis de Correspondencia Múltiple para determinar las relaciones entre algunas de las variables mencionadas y, por otro lado, a través de un Análisis de Clúster, se realizará el armado de grupos de observaciones que presenten semejanzas, para luego proceder a la caracterización de cada uno de ellos.

Este estudio pretende brindar información útil para empresas logísticas y aseguradoras de transporte/empaques que presten sus servicios, y a industrias de índole privada orientadas a importaciones que deseen planificar sus futuras compras y decisiones de transporte.

2. Datos utilizados y análisis exploratorio de datos

Para este estudio, se dispuso de una base de datos de importaciones de productos de empresas argentinas para el año 2017, que cuenta con 436.439 observaciones y 28 variables, de tipo categóricas y numéricas.

Posterior al análisis exploratorio de datos y previo a la limpieza definitiva, se cruzó información para decodificar nomenclaturas con distintos entes como la Administración Federal de Ingresos Públicos (AFIP), el Ministerio de Relaciones Exteriores y Culto, entre otros.

Para los objetivos de nuestro trabajo, se realizaron las siguientes adecuaciones a la base:

- Se eliminó la variable *'year'* por presentar un valor único (2017).
- Entre las variables de medidas se escogió *'kilos'* por ser comparable entre observaciones.
- Utilizamos la variable *'codprov'* para conectar con la tabla de Códigos de Provincias y así determinar a qué provincia pertenece cada observación.
- Unimos la tabla con el nomenclador de actividades de AFIP a partir de la variable *'clae'* para obtener las distintas actividades importadoras.
- Corregimos errores ortográficos dentro de la variable *'procedencia'* (Haití y España).
- Eliminamos 6265 valores nulos que representan el 1.48% del total de la base.
- Analizamos las variables '*origen'* y '*procedencia'*. Se acordó trabajar con los datos de '*procedencia*', ya que para nuestro análisis sobre su costo y transporte no es de interés el país donde se fabricó el producto, sino de donde se importó.

Integrantes: Álvarez Eugenia, Laguna Nicolás, Mosconi Florencia, Quispe Santiago

- A partir de la variable *'posic_sim'* cruzamos datos con la Nomenclatura Común del Mercosur para clasificar e individualizar las mercaderías comercializadas.

3. Creación de nueva base de datos.

Con el objetivo de obtener una base idónea para su análisis se gestionaron las siguientes modificaciones.

En un primer momento, transformamos la variable *'peso'* en tipo categórica generando agrupamientos en intervalos conceptuales según su valor de Kg.

- Peso menor a 1 kg - Muy liviano
- Peso mayor a 1 kg, menor a 5 kg - Liviano
- Peso mayor a 5 kg, menor a 35 kg - Mediano
- Peso mayor a 35 kg, menor a 350 kg - Pesado
- Peso mayor a 350 kg - Muy Pesado

Previamente, se realizó un estudio separando en Quantiles para conocer las proporcionalidades de los pesos incluidos.

Luego de conocer los conceptos de las actividades mediante unión de la base de AFIP utilizando el código *'clae'* determinamos una gran variante en dicha variable, por lo que se agrupó en categorías generalizadas para facilitar su análisis.

Por otro lado, se decidió agrupar la variable *'procedencia'* por continente para poder facilitar su interpretación.

Por último, mediante la codificación *'.sample'* se generó un entorno reducido de 101490 (25%) registros aleatorios utilizando como propiedad una semilla (random_state) de valor 1 para no alterar los resultados en cada estudio, para agilizar la ejecución.

A continuación, se presentan las variables escogidas para los análisis multivariados.

Variable	Descripción
peso	Valor de pesaje del producto expresado en Kilogramos
medio_tra	Modo de traslado del producto importado (Acuática, Avión, Camión, Ferrocarril, etc.)
region	Continente de exportación del producto obtenidos a partir de los países de procedencia
categorias	Actividad general de la empresa importadora
aduana	Entidad reguladora y controladora por donde se registra el ingreso del producto

A fines informativos, se analizó y ubicó las aduanas nacionales con la librería *folium* a donde llegaban las importaciones de la muestra estudiada, las cuales quedaron ubicadas en el mapa que se observa a continuación.

Los porcentajes se calcularon mediante frecuencias, determinando que la mayoría de las transacciones se concentran en dos aduanas finales: Buenos Aires (Capital) con el 42% y Ezeiza con el 25% de importaciones recibidas.

Integrantes: Álvarez Eugenia, Laguna Nicolás, Mosconi Florencia, Quispe Santiago

4. Modelos Utilizados

Análisis de Correspondencia Múltiple

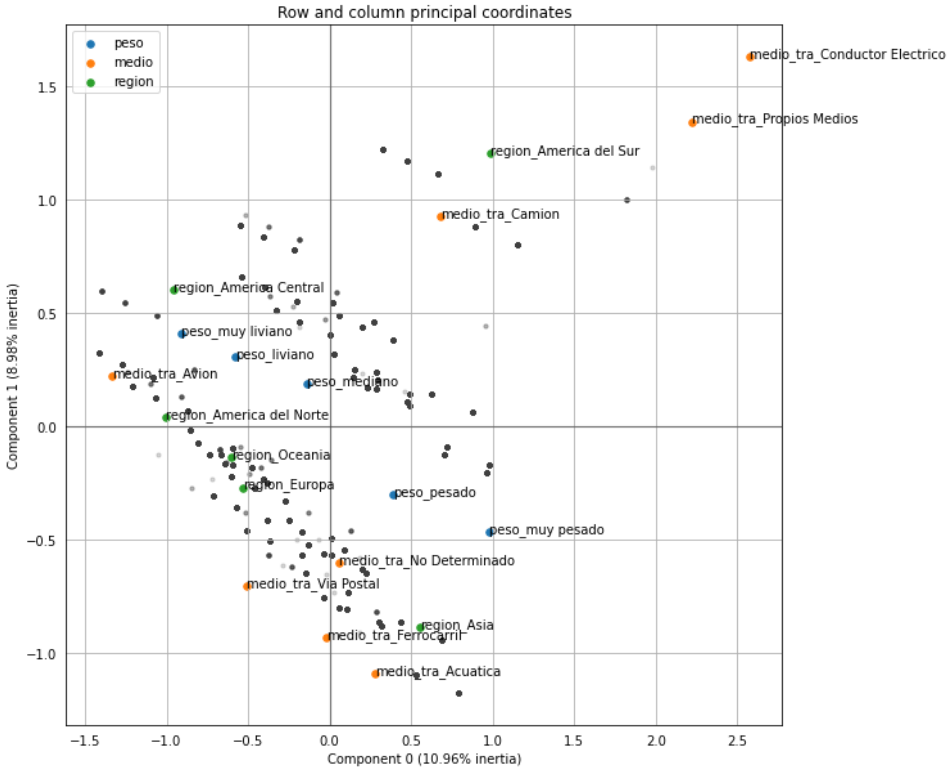
El análisis de correspondencia múltiple (ACM) es una ampliación del análisis de correspondencia simple para resumir y visualizar una tabla de datos que contiene más de dos variables categóricas. También puede considerarse como una generalización del análisis de componentes principales cuando las variables a analizar son categóricas en lugar de cuantitativas. Permite analizar el patrón de relación y las asociaciones entre varias variables categóricas correlacionadas. (ver *"Multiple Correspondence Analysis" – Hervé Abdi & Dominique Valentin*)

El siguiente gráfico muestra mediante una nube de puntos a los individuos y las variables, en el espacio de las primeras dos dimensiones, que son las que mayor porcentaje de inercia explican, acumulando aproximadamente el 19% de la inercia total de la muestra. Así se reduce la complejidad del problema y se facilita la representación gráfica, el cual se construye a partir de las puntuaciones o coordenadas en dos dimensiones.

Las categorías de peso de muy liviano y liviano y medio de transporte avión se ubican cercanas, por lo que sobre el total de transacciones que se caracterizan por al menos uno de estos pesos o medio de transporte, la proporción de que los que se caracterizan por ambos es elevada. Caso contrario podemos observar que las categorías de peso pesado y muy pesado y transporte acuático son lejanas, por lo tanto sobre el total de transacciones que se caracterizan por al menos de una de ellas, la proporción que se caracterizan por ambas es mucho menor.

El componente 0 opone en el semieje positivo a las transacciones realizadas mediante los siguientes medios de transporte, conductor eléctrico, propios medios, camión, y en menor medida acuático; que se producen en las regiones de América del Sur y Asia; y que son pesados. Contrariamente en el semieje negativo las transacciones transportadas mediante avión y vía postal; producidas en América del Norte y Central, Oceanía y Europa; y cuyo peso es liviano.

En el semieje positivo del componente 1 se observan aquellas transacciones realizadas mediante conductor eléctrico, propios medios, camión y avión (menos representativa; producidas en América del Sur y Central; y livianas. En el semieje negativo las transacciones transportadas mediante ferrocarril, por vía marítima y postal; provenientes de Asia y Europa; y pesadas.



Integrantes: Álvarez Eugenia, Laguna Nicolás, Mosconi Florencia, Quispe Santiago

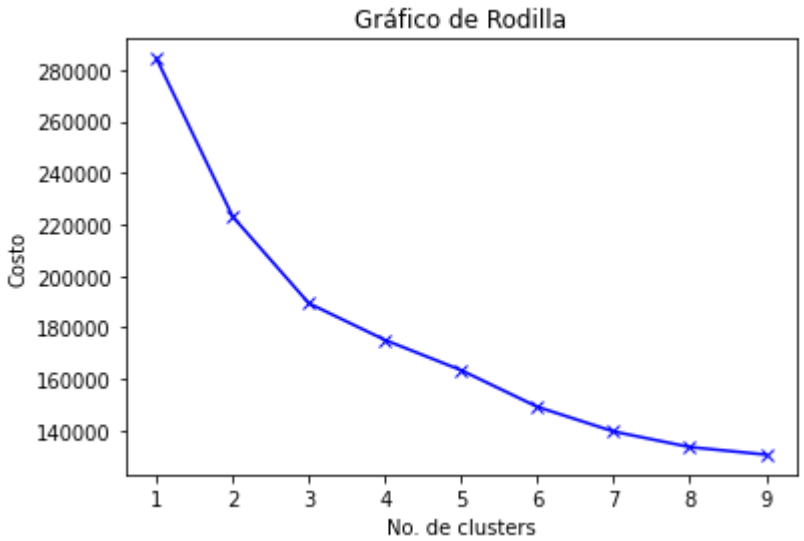
Análisis de Clúster

Kmodes es un algoritmo no supervisado de Machine Learning que se utiliza para clusterizar variables categóricas. A diferencia de Kmeans, este método utiliza las disimilitudes (total de diferencias) entre los datos, mientras menores sean, más similares son nuestras observaciones. Para determinar los centroides, utiliza la medida descriptiva moda (modes). (*"The k-modes as Clustering Algorithm for Categorical Data Type" - Adhi Aprilliant*)

Para empezar, elegimos las variables a incluir en este estudio las cuales son *peso* que representa cuán pesado es el producto importado*; región de procedencia* u origen de fabricación del bien/materia prima; *medio de transporte* por el cual llegan a su último destino desde la aduana argentina; y *categoría* que refiere a la actividad para la que se utilizará la importación.

Comenzamos determinando el número de clusters óptimos mediante el *método de rodilla*, cuyo gráfico representa linealmente el coste para un rango de valores de K (suma de las disimilitudes entre los clusters).

Para la muestra observada, como se observa en el siguiente gráfico, la forma curva de rodilla o codo, que representa un costo menor, se produce al seleccionar 7 clusters, por lo que ésta será la cantidad elegida a estudiar.



Lo siguiente es ajustar el algoritmo y ejecutarlo, para luego determinar los centroides de los clusters, quienes van a representar las modas de las observaciones estudiadas en cada grupo.

Una vez obtenidos los centroides de los 7 grupos, se obtuvo que el método realizó 2 iteraciones (cantidad de veces que se ejecutó el algoritmo) con un costo de 144.215, que representa el costo del del número de clusters elegidos en el gráfico de rodilla.

Finalmente, se obtuvieron los siguientes clústeres, descritos por sus centroides.

index	medio_tra	peso	region	categoria
Primer Cluster	Camion	muy pesado	Asia	Fabricación de vehículos automotores, remolques y semirremolques
Segundo Cluster	Avion	mediano	Europa	Fabricación de otros equipos y productos
Tercer Cluster	Camion	muy liviano	America del Sur	Comercio al por mayor y menor excepto autos y motos
Cuarto Cluster	Avion	muy liviano	America del Norte	Comercio al por mayor y menor excepto autos y motos
Quinto Cluster	Acuatica	liviano	Europa	Fabricación de vehículos automotores, remolques y semirremolques
Sexto Cluster	Acuatica	mediano	America del Norte	Comercio al por mayor y menor excepto autos y motos
Séptimo Cluster	Acuatica	pesado	Asia	Fabricación de otros equipos y productos

Primer Clúster: bienes con procedencia asiática destinados a la fabricación de autos, remolques y semirremolques, de un peso mayor a 350 kg y los cuales llegaron a destino nacional mediante camiones, lo cual es lógico considerando su posible tamaño y peso.

Segundo Cluster: bienes de procedencia europea, con un peso entre 5 kg y 35 kg, utilizados para la fabricación de otros equipos y productos (maquinarias, equipos informáticos, muebles, manufacturación, etc.) que fueron transportados en avión por sus dimensiones.

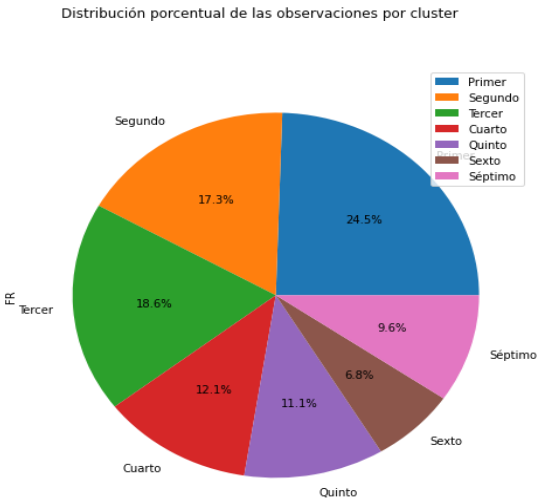
Tercer Cluster: bienes provenientes de América del Sur, destinados directamente al comercio por mayor y por menor general (no automotriz) de un peso menor a 1 kg y transportados en camión.

Cuarto Cluster: similar al anterior en cuanto a peso y categoría de actividad de la importación, pero la procedencia es de América del Norte y el medio de transporte a destino es el avión.

Quinto Cluster: importaciones de procedencia europea, destinados a la fabricación automotriz, de entre 1kg y 5 kg, transportado por medios acuáticos.

Sexto Cluster: bienes destinados como productos finales a la comercialización por mayor y por menor (excluyendo automóviles y motos), con un peso entre 5 kg y 35 kg y proveniente de América del Norte y que llegaron a su destino final en medios acuáticos.

Séptimo Cluster: importaciones procedentes de Asia, cuyo peso es de entre 35 kg y 350 kg, cuya finalidad es la fabricación de otros equipos y productos varios y que han sido transportados mediante medios acuáticos.



Por último, se obtuvo que el primer cluster es el que representa el mayor número de observaciones totales (casi un 25%), seguidos por el segundo al quinto clúster (entre el 17% y el 11%) y por último los el sexto y séptimo clúster que conforman la minoría de observaciones (menos del 10% cada uno).

Integrantes: Álvarez Eugenia, Laguna Nicolás, Mosconi Florencia, Quispe Santiago

6. Conclusiones Finales

La detección de relaciones entre variables mediante el análisis de datos masivos requiere del uso de métodos multivariantes como los aplicados en este informe. La reducción de la dimensionalidad y la agrupación de transacciones según sus semejanzas facilita y agiliza el análisis de información y no presenta grandes dificultades en su aplicación.

La muestra utilizada presenta un sesgo sobre el destino de las importaciones observadas, ya que el 67% de éstas se observaron en solo dos aduanas sobre un total de 48, por lo que los resultados serán mas representativos para éstas aduanas.

La relación entre las variables escogidas (peso, región, medio de transporte) resultantes del Análisis de Correspondencia Múltiple presentan tres agrupaciones, dos principales que poseen una dispersión lineal paralela, uno caracterizado por el medio de transporte y el origen, y el otro por el peso de la importación. Por otro lado, un tercer grupo se ubica alejado, probablemente presenta características menos predominantes.

En busca de una mejor caracterización, se realizó el análisis de clúster, sumando la variable actividad a las anteriores nombradas, que concluyó en la construcción de 7 grupos que se distinguen principalmente por la actividad y el peso. A un mismo medio de transporte, le corresponden distintos pesos y origen, lo cual podría responder al cambio de actividad dada la diferencia en lo analizado en el estudio anterior.

Como propuesta de mejora, para obtener una caracterización con un número mayor de variables, se podría aplicar análisis de clúster con distancia por atributos. En una segunda instancia, otro método aplicable es el análisis de grafos que permite representar gráficamente múltiples relaciones entre variables mediante la formación de nodos y vértices, permitiendo una mayor detección de relaciones y mejor claridad en la composición de estas relaciones.

Integrantes: Álvarez Eugenia, Laguna Nicolás, Mosconi Florencia, Quispe Santiago