

Definición del problema

Minería de datos



Adrián Ruiz Esteban

Mario Martín Alonso

Anderson Marroquín Rivas

Pablo Valle Nieto

Índice

1.-Descripción de los datos originales	3
1.1.-Dataset 1 de Cajamar	3
1.2.-Dataset Consumo Comunidades	3
1.3.-Dataset 4 de Cajamar	3
1.4.-Dataset 5 de Cajamar	3
2.-Antecedentes similares	4
3.-Hipótesis	4
4.-Posibilidades de enriquecimiento de los datos	4
5.-Selección de los datos. Integración de varias fuentes. Selección de la información básica	5
6.-Preprocesado y transformación de los datos	5
6.1.-Preprocesado y transformación para el dataset 1 de Cajamar	6
6.2.-Preprocesado y transformación para el dataset de las comunidades	6
6.3.-Preprocesado y transformación para el dataset 4 de Cajamar	6
6.4.-Preprocesado y transformación para el dataset 5 de Cajamar	6
7.-Tarjetas de datos	7
7.1.-Tarjeta de datos del dataset Relación entre enfermos y las importaciones/exportaciones en Europa	7
7.2.-Tarjeta de datos del dataset Consumo Andalucía-Madrid	7
7.3.- Tarjeta de datos del dataset Precio-Consumo nacional en relación al COVID	8
8.-Líneas de trabajo	8

1.-Descripción de los datos originales

Hemos decidido trabajar con los datos del reto de agro analysis de Cajamar. Vamos a utilizar los datasets 1, 2, 4 y 5 de los que se nos han facilitado inicialmente en la misma página para resolver el problema.

1.1.-Dataset 1 de Cajamar

El primer dataset recoge, de mano del Ministerio de Agricultura, Pesca y Alimentación, la evolución mensual desde enero de 2018 a noviembre de 2020 del consumo de frutas y hortalizas en España por comunidades autónomas. Específicamente habla, además del año y mes de la recogida de información, de la comunidad autónoma, en la cual también está incluido el dato del total nacional; del producto, en miles de kg; del valor, en miles de euros; el precio medio del kg, en euros; la penetración, que es el porcentaje de hogares que compran el producto en cuestión; el consumo per capita, que es el consumo total dividido del número de habitantes en la región y mes tratados; y el gasto per capita, que es el gasto total dividido del número de habitantes en la región y mes tratados.

1.2.-Dataset Consumo Comunidades

El segundo dataset que usaremos en nuestro proyecto proviene de varios archivos del Ministerio de Agricultura, Pesca y Alimentación. Dentro de la página oficial del ministerio, podemos acceder a un subapartado en el que podemos consultar series de datos de consumo alimentario en hogares ([enlace](#)). Dentro de este enlace, los datos que nos interesan se encontrarían en formato anual en archivos Excel (la columna "Mensual CCAA"). En estos archivos se muestra la información de distintos productos alimenticios que se han ido vendiendo en España, recopilados de forma mensual y separados por comunidades autónomas. Para cada comunidad, tenemos los valores de: Consumo x Cápita, Gasto x Cápita, Porcentaje de Penetración, Precio medio en Kg, Valor en miles de € y Volumen en miles de Kg.

Junto a estos datos, debemos añadir el dataset conteniendo la información de la población por comunidad autónoma. Este dataset contiene la población total en los años de 2018 a 2020 para comunidad española.

1.3.-Dataset 4 de Cajamar

El dataset número 4 nos detalla los registros de importaciones y exportaciones de frutas y hortalizas realizados por parte de España desde enero de 2018 hasta noviembre de 2020, ambos inclusive. También incorpora a modo de resumen, una cantidad total de los datos de los años 2018 y 2019 de cada elemento. Entrando en detalle, en este dataset nos encontramos primero con la fecha del dato, a continuación, el país extranjero que importa o exporta, seguido de España, el producto que se está tratando, si se requiere de una operación de importación o exportación y por último un valor cuantitativo que se refiere a la cantidad de producto en valor monetario o la cantidad de producto en peso. Respecto a esta última característica hemos observado que se requieren dos líneas completamente iguales sobre el mismo producto con la única diferencia de que en la primera línea encontramos el valor en euros y en la segunda su peso. Encontramos 360.976 líneas de datos distintos en este dataset. Este conjunto de datos ha sido obtenido a través de Eurostat.

1.4.-Dataset 5 de Cajamar

Los datos contenidos en este dataset se corresponden a los casos de Coronavirus, registrados internacionalmente agrupados por país, en el lapso de un año desde la fecha de 31 de diciembre de 2019 al 29 de diciembre de 2020. Los presentes datos muestran por los nuevos casos registrados por cada día de la pandemia, perteneciente al periodo previamente mencionado. También las defunciones, y el total de casos acumulados durante un periodo de dos semanas. Las columnas que

presenta el dataset, en el orden el que se muestran, son: fecha (formato dd/mm/aaaa), día, mes, año, casos registrados en el día, muertes registradas en el día, país, abreviatura del país, código del país, cantidad de población en 2019, continente, y número total de casos acumulados durante las dos últimas semanas.

2.-Antecedentes similares

En la propia página de Cajamar vienen varios estudios que han sido realizados previamente por estudiantes de distintas universidades, de los cuales mencionaremos algunos a continuación.

Uno de los casos de ejemplo que vienen es el del equipo The Data Masters formado por estudiantes de la escuela Mondragon Unibertsitatea, el cual proporciona una página web en la que podemos ver los resultados de los análisis que realizaron sobre los datos. Dentro de estos resultados podemos ver varios análisis de forma visual sobre los estudios realizados, como es el caso de un mapa interactivo en el que podemos ver de forma directa cómo varían los productos en las distintas comunidades autónomas proveyendo la fecha y el alimento a medir, lo cual es una herramienta muy útil para poder comparar los resultados que vayamos obteniendo con los de este estudio. Junto a esto podemos ver el impacto que tuvo el Covid en la venta como la cantidad de productos vendidos junto al precio de estos, y su consumo con unos gráficos sencillos. Al igual que con el anterior, nos puede llegar a ser útil para poder evaluar los resultados que obtendremos junto con ideas de cómo poder mostrar los datos.

Otro ejemplo muy útil es el del equipo Datacrop compuesto por los estudiantes de la Escuela de Organización Industrial. Con este estudio podemos sacar más contexto de cómo emplear la librería Pandas junto a otras librerías útiles para la representación de forma visual. Junto a esto, también podemos ver de forma visual los resultados obtenidos con la herramienta PowerBi, lo cual también podemos aprovechar para tomarlo como idea para futuras implementaciones.

3.-Hipótesis

Hemos creado un total de cuatro hipótesis para trabajar con los datos:

1. Existe una correlación entre el número de casos de enfermos en una pandemia y las importaciones/exportaciones en Europa.
2. Las zonas con mayor número de población tienen un mayor número de ventas de productos agrícolas.
3. Durante una pandemia se produce un incremento del precio de los productos agrícolas.
4. Durante una pandemia se produce un decremento del consumo de los productos agrícolas.

Estas hipótesis distan de un problema estadístico o mera visualización ya que las respuestas a estas preguntas no están en los datos en bruto originales y necesitan de profundización en ellos y procesado de estos mismos encontrando reglas asociativas, clasificaciones, etc.

Además, las hipótesis tienen un componente inteligente debido a que la conclusión a la que llegaremos será a partir del patrón generado a partir del procesado de los datos del que hemos hablado previamente, no simplemente describiendo los propios datos.

4.-Posibilidades de enriquecimiento de los datos

En nuestro caso dentro de la página de Cajamar vienen múltiples páginas que pueden ser de ayuda a la hora de encontrar más datasets. Dentro de estas páginas recomendadas se encuentran el Eurostat, oficina de estadística de la Unión Europea; la página oficial del Ministerio de Agricultura, Pesca y

Alimentación (MAPA); o el Instituto Nacional de Estadística. Todas ellas son fuentes con gran renombre y de todas ellas podemos acceder a grandes cantidades de datos y de estudios públicos por parte del gobierno que pueden ayudar bastante a expandir los datos que poseemos.

En nuestro caso, en un principio queríamos usar el segundo dataset proporcionado por el propio reto de Cajamar. Sin embargo, tras avanzar con el proyecto y revisar las hipótesis de este, se decidió ampliar los datos con un dataset nuevo, el cual proviene del MAPA. Este cambio de planes se debe a que, tras cambiar las hipótesis, no veíamos viable usar el dataset de los datos de ventas y precios de la comunidad de Andalucía de forma exclusiva, por lo que quisimos ampliar el rango de los datos y poder comparar múltiples comunidades autónomas al mismo tiempo.

Para ello escogimos un nuevo dataset que, como bien hemos explicado previamente, contiene información anual dividida por meses de los distintos productos que se han llegado a vender en las distintas comunidades españolas en varios años. Para nuestro estudio únicamente hemos usado los datos entre 2018 y 2020, ya que la mayoría de nuestros datasets se sitúan en dichas fechas.

Junto a estos datos, hemos añadido otro dataset que proviene del Instituto Nacional de Estadística que contiene la información del nivel de población en cada comunidad autónoma. Esta información nos es relevante a la hora de comparar los datos de las ventas con el nivel de población.

5.-Selección de los datos. Integración de varias fuentes. Selección de la información básica

En el caso del dataset con los datos de las comunidades, los datos principales provienen del MAPA, como bien explicamos al describir la información base de este dataset. Para indicar los datos que queríamos filtrar de los datos originales, analizamos los distintos valores que contenían las columnas de los datos, y decidimos que lo óptimo sería comparar directamente las ventas y precios de los productos agrícolas en varias comunidades, para poder comprobar la segunda hipótesis. Dichas comunidades serían Andalucía, Madrid, Castilla y León, Castilla-La Mancha y Cataluña.

Junto a estos datasets, hemos añadido otro obtenido del Instituto Nacional de Estadística (INE) que contiene la información del nivel de población en cada comunidad autónoma para los años 2018, 2019 y 2020. Estos datos vienen en un formato CSV con el pequeño inconveniente de que los datos vienen dados al revés en cuanto al orden de las fechas, empezando por la más reciente a la más antigua, pero no es nada grave que suponga una complejidad grande.

Algo negativo que tenemos que comentar es el mal formato que tienen inicialmente los datos del MAPA, pues al usar un formato Excel, los datos de cada mes vienen dados en distintas hojas del Excel. Esto ha llegado a complicar un poco la correcta lectura de los datos y la extracción de los mismos. Además, dentro de los datos originales, en los últimos meses de 2019 en adelante, añaden más productos al listado original. Esto complicó de nuevo la lectura correcta de los productos alimentarios, teniendo que guardar el primer listado de productos agrícolas en una lista e ir comprobando mes a mes los productos que se fueran leyendo, para así guardar únicamente los que nos resultaban de interés.

6.-Preprocesado y transformación de los datos

A pesar de que todo el proceso de limpieza de los datos viene explicado en el Colab que les hemos adjuntado, en este apartado vamos a hablar brevemente de los cambios realizados para el proceso de transformación de los datos.

6.1.-Preprocesado y transformación para el dataset 1 de Cajamar

Lo primero que hemos hecho es limpiar el número de columnas eliminando las que no vamos a necesitar. A continuación, se han renombrado el nombre de las columnas para quitar espacios y poner todo en minúscula con el objetivo de facilitar el posterior trabajo con el dataset. Por último, hemos comprobado si había nulos y, efectivamente, no los había.

A la hora de limpiar las filas hemos eliminado todas las que no son de datos nacionales ya que queremos que este dataset sólo contenga información nacional. Además, a continuación, hemos eliminado los productos que sobraban ya sea porque fuesen muy raros de encontrar en los datos o porque fuesen el conjunto de otros productos ya representados en el dataset.

Finalmente se unen las columnas de “año” y “mes” en una sola llamada “fecha” para poder mezclar más adelante este dataset con otro. Además, se cambia el tipo de las variables a unas más adecuadas.

6.2.-Preprocesado y transformación para el dataset de las comunidades

Los procesos realizados en este apartado para este caso sería filtrar los datos de los productos agrícolas contenidos en los archivos del MAPA. Estos datos están contenidos en un listado dentro de cada mes. Dentro de lo que son los productos agrícolas, decidimos emplear únicamente los datos de los productos frescos, pues también hay datos de productos de conserva y congelados. Para esto, los filtramos en una lista al leer el primer archivo por primera vez para así poder comparar archivo a archivo, mes a mes, con el fin de guardar sólo los datos de los productos frescos. Para ello usamos la función *merge()* que contiene la librería pandas, uniendo el listado con los nombres más los datos de las ventas en cada comunidad.

Además de esto, almacenamos en varias columnas los datos provenientes del INE, iterando con el índice *i* para guardar los datos de cada año.

Toda la información que se recoge de cada mes se va almacenando en un objeto *Dataframe* adicional (*dataset_comunidades*) que es el que empleamos para almacenar los datos de todos los ficheros y usarlo como dataset en el que basarnos para terminar de limpiar datos nulos y convertir los tipos de las columnas. El Dataframe final (*ds_clean*) se transforma al finalizar el código en un fichero CSV descargable.

6.3.-Preprocesado y transformación para el dataset 4 de Cajamar

Hemos reducido el tamaño del dataset a la mitad debido a que observamos la redundancia de filas. Para ello hemos creado dos nuevas características que nos indican el valor en euros y la cantidad en peso del producto al que nos referimos. Con la introducción de estas nuevas características hemos eliminado las columnas INDICATORS y VALUE. Estas dos nuevas columnas las hemos transformado de string a entero, sumándose así a la anterior transformación de FLOW y REPORTER que ambas son categorías.

A continuación, hemos eliminado la columna PARTNER porque solo adquiría el valor ES refiriéndose a España, por lo que hemos visto innecesario su uso.

Por último, hemos traducido la columna PRODUCT desde su idioma original inglés a español para facilitarnos su posterior uso con otros dataset.

6.4.-Preprocesado y transformación para el dataset 5 de Cajamar

Se ha reducido el tamaño del dataset eliminando las filas correspondientes a países que no forman parte de Europa, lo cual indirectamente elimino gran parte de datos nulos. También se han eliminado datos inútiles y datos parciales para el estudio.

Se han extraído los datos más relevantes para en estudio, como serían los casos de contagio por coronavirus al igual que las muertes que este provocó. Estos se procesaron para que reflejasen las cifras de cada país por mes, e lugar de hacerlo por día como sucedía en el dataset original.

En la columna relacionada con los datos de la incidencia acumulada a lo largo de 14 días se encontraron numerosos nulos, puesto que la infección empezó de forma asíncrona entre los países. En el caso de datos de inicio no se reflejó valor pues no había casos, y a finales en algunos países valor era mucho menores a 0. Se optó por sustituir dicho nulos por un 0.0 (float).

Para la obtención de los datos de la incidencia por país en un mes en específico, se optó por hacer la media de los valores de la incidencia de cada uno de los meses.

7.-Tarjetas de datos

En este apartado se explicarán las tarjetas de datos de esta práctica.

7.1.-Tarjeta de datos del dataset Relación entre enfermos y las importaciones/exportaciones en Europa.

El dataset final es la conjunción del dataset 5, casos de coronavirus, y el dataset , con los datos de importaciones y exportaciones en Europa.

- País: identifica los distintos países necesarios para el estudio.
- Fecha: describe la tupla 'mes/año' sobre el que se realizara el estudio, se optó por acotar el tiempo dentro del rango de mes.
- Casos mes: suma de los casos de coronavirus a lo largo de un mes en un territorio específico.
- Muertes mes: suma de las defunciones por coronavirus a lo largo de un mes en un territorio específico.
- Incidencia media mes: media geométrica de la incidencia acumulada de casos de contagio de coronavirus en los últimos 14 días a lo largo de un mes en un territorio específico.
- Producto: el producto en cuestión (banana, fresa, ...) que se está importando o exportando.
- Flow: nos indica que acción de transporte (importación o exportación) se está realizando sobre el producto.
- Valor en euros: valor en euros de la mercancía transportada en la operación.
- Cantidad en 100 kg: cantidad de mercancía transportada en la operación, calculo sobre 100 kg.

7.2.-Tarjeta de datos del dataset Consumo Andalucía-Madrid

En el dataset final, se juntó los tres datasets de los distintos años, que obtuvimos inicialmente del MAPA y del INE, tras filtrar su contenido y limpiar los datos que contenían. El resultado final contiene la comparativa entre las ventas de las distintas comunidades con las siguientes columnas:

- Producto: muestra la lista de frutas y hortalizas frescas del dataset.
- Fecha: conjunto del mes y el año de los datos.
- Precio (€/Kg): muestra el precio medio del producto para Andalucía y Madrid.
- Valor (miles €): muestra el valor total de las ventas realizadas en las comunidades para cada uno de los productos.

- Volumen (miles Kg): muestra el volumen total de las ventas realizadas en las comunidades para cada uno de los productos.
- Población: muestra el nivel de población total de las comunidades para cada año.

7.3.- Tarjeta de datos del dataset Precio-Consumo nacional en relación al COVID

Este dataset se ha construido mezclando el dataset 1 con el dataset 5 ya limpios. Como resultado tenemos las siguientes columnas:

- Fecha: es el mes y año en el que se están mostrando los datos de la fila. Está en formato "Mes/Año".
- Producto: es el nombre de la fruta u hortaliza.
- Precio_medio_kg: es el precio medio de la fruta u hortaliza ese mes.
- Consumo_per_capita: es la relación entre el total de los habitantes del país y la cantidad de producto que han consumido ese mes.
- Casos_mes: suma de los casos de coronavirus a lo largo del mes en España.

8.-Líneas de trabajo

Tendríamos una línea de trabajo por hipótesis:

- **Línea de trabajo de la hipótesis 1:** se va a utilizar k-means para agrupar los diferentes países en clústeres en función del número de casos de la enfermedad de la pandemia. A continuación, se va a analizar el número de importaciones/exportaciones en cada clúster para ver como varía, permitiéndonos encontrar patrones en los datos que muestren como el número de estos últimos varía en función del número de casos de la pandemia.
- **Línea de trabajo de la hipótesis 2:** aquí haremos uso de una regresión lineal para que podamos estimar la relación entre el número de ventas de cada producto y el número de población de la comunidad en específico.
- **Línea de trabajo de la hipótesis 3:** en esta línea de trabajo vamos a utilizar el algoritmo PCA en datos de precios de diferentes productos en diferentes momentos para poder identificar patrones en la evolución de los precios y localizar puntos de inflexión o cambios en la tendencia de los mismos.
- **Línea de trabajo de la hipótesis 4:** en esta última línea de trabajo haremos lo mismo que en la tercera, pero en vez de con el precio de los productos lo haremos con el consumo de determinados productos agrícolas a lo largo del tiempo.