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 7
6.4.-Preprocesado y transformación para el dataset 5 de Cajamar 7
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 8
7.3.- Tarjeta de datos del dataset Precio-Consumo nacional en relación al COVID 8
8.-Líneas de trabajo 9
8.1.-Líneas de trabajo de la hipótesis 1 9
8.2.-Líneas de trabajo de la hipótesis 2 11
8.3.-Líneas de trabajo de la hipótesis 3 12
8.4.-Líneas de trabajo de la hipótesis 4 13

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 muestras 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 la pandemia y las importaciones/exportaciones en Europa.
2. El número de ventas aumenta en zonas con mayor densidad de población cuanto menor sea el precio.
3. Durante una pandemia se produce un incremento del precio de los productos agrícolas.
4. El porcentaje de penetración de un alimento aumenta en los hogares según la fecha, la región, si el volumen del producto es mayor y su precio menor.

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 (variante 1)

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 son de datos nacionales. Aunque ya tengamos otro dataset con datos de regiones en este queremos usarlo para estudiar la variable de penetración en los hogares con todas las regiones y se adecuaba más al objetivo mencionado. 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.

Por último, en este apartado vamos a cambiar las variables en forma de texto de las columnas “mes”, “producto” y “CCAA” a variables en forma de número para que sea más cómodo de usar en un futuro con los algoritmos necesarios. Además, se cambia el tipo de las variables a unas más adecuadas.

6.2.-Preprocesado y transformación para el dataset 1 de Cajamar (variante 2)

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.

6.3.-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. Tras la limpieza de los nulos, debemos modificar el formato del dataframe con el fin de tener una versión más simplificada de este a través de la función *melt().* Una vez hecho esto lo transformamos en un fichero csv para su descarga.

6.4.-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 ingles a español para facilitarnos su posterior uso con otros dataset.

6.5.-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 opto 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 especifico.
- **Muertes mes:** suma de las defunciones por coronavirus a lo largo de un mes en un territorio especifico.
- **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 especifico.
- **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:** muestra el conjunto del mes y el año de los datos.
- **Región:** muestra la región de donde provienen los datos.
- **Precio_euro_Kg:** muestra el precio medio del producto para las distintas comunidades.
- **Valor_miles:** muestra el valor total de las ventas realizadas en las comunidades para cada uno de los productos.
- **Volumen_miles:** muestra el volumen total de las ventas realizadas en las comunidades para cada uno de los productos.
- **Dens_pon:** muestra el nivel de la densidad de la población para cada comunidad según el 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.

7.4.- Tarjeta de datos del dataset de consumo alimentario por CCAA

Este dataset se ha construido limpiando el dataset 1. Como resultado tenemos las siguientes columnas:

- **año:** es el año en el que se están mostrando los datos de la fila.
- **mes:** es el mes en el que se están mostrando los datos de la fila. Los valores se muestran en forma de número con una equivalencia con el valor en texto. Esto se especifica en profundidad en el colab.
- **CCAA:** es la comunidad autónoma en la que se muestran los datos. Los valores se muestran en forma de número con una equivalencia con el valor en texto. Esto se especifica en profundidad en el colab.
- **Producto:** es el nombre de la fruta u hortaliza. Los valores se muestran en forma de número con una equivalencia con el valor en texto. Esto se especifica en profundidad en el colab.
- **Volumen_kg:** es la cantidad total, en kilogramos, del producto.
- **Precio_medio_kg:** es el precio medio de la fruta u hortaliza ese mes.

- **Penetracion_%:** porcentajes de hogares que compran ese producto.

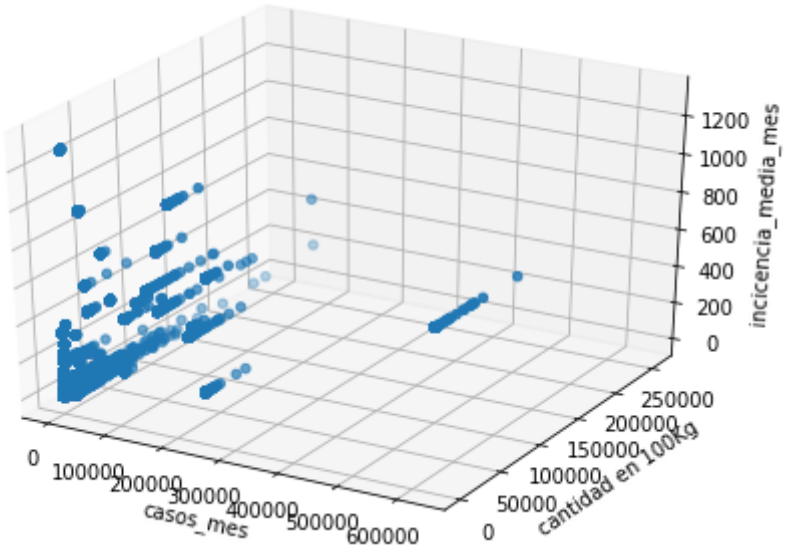
8.-Líneas de trabajo

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 la densidad 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 uso de regresión para poder ver la relación entre las diferentes variables y poder inferir un porcentaje de penetración dados unos datos de entrada. Haremos uso de regresión debido a que el valor que queremos inferir es un valor continuo.

8.1.-Línea de trabajo de la hipótesis 1

Para analizar la hipótesis 1 se ha decidido el uso de técnicas de clustering, entre las cuales se ha elegido K-means,. El porqué de elegir este algoritmo recae en que es un algoritmo rápido y fácil de usar que produce buenos resultados en muchos casos, y asignar a los datos clústers, puede ser útil para este análisis en específico. Durante el proceso y tras analizar cuidadosamente los datos del relacionados con la hipótesis se observó, que datos están distribuidos de manera relativamente equilibrada y no hay muchos valores atípicos (después de haber procesado y normalizado los datos).



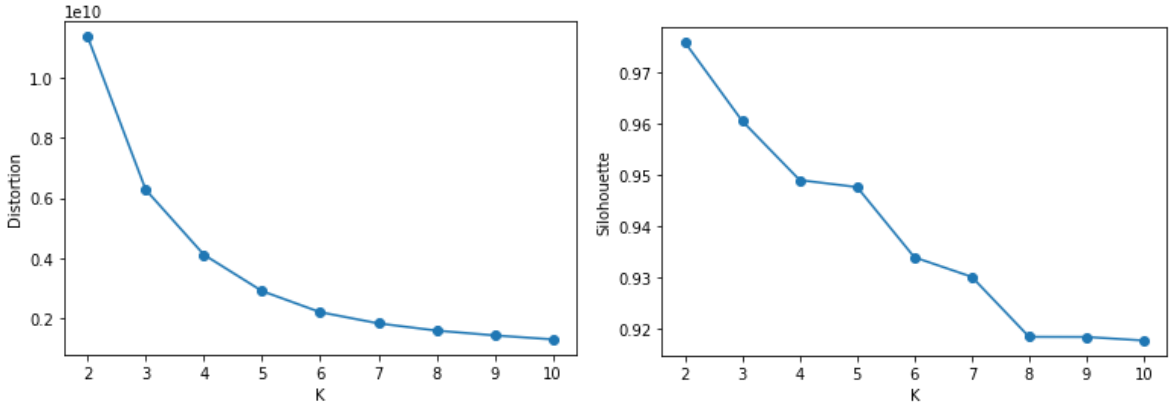
Se tomaron, de la tarjeta de datos, los valores de referentes a:

- La cantidad de casos registrados en un mes.

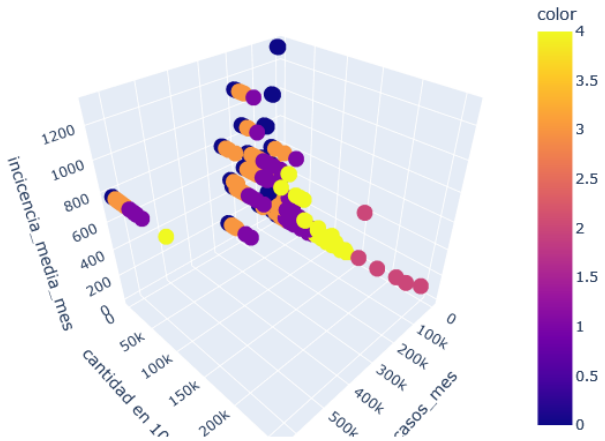
- La montó total de productos (indistinto que producto) transportado.
- La incidencia acumulada media de personas infectadas en un mes.

Los demás valores fueron desechados, por lo menos de las gráficas, por carecer de aporte relevante para el test.

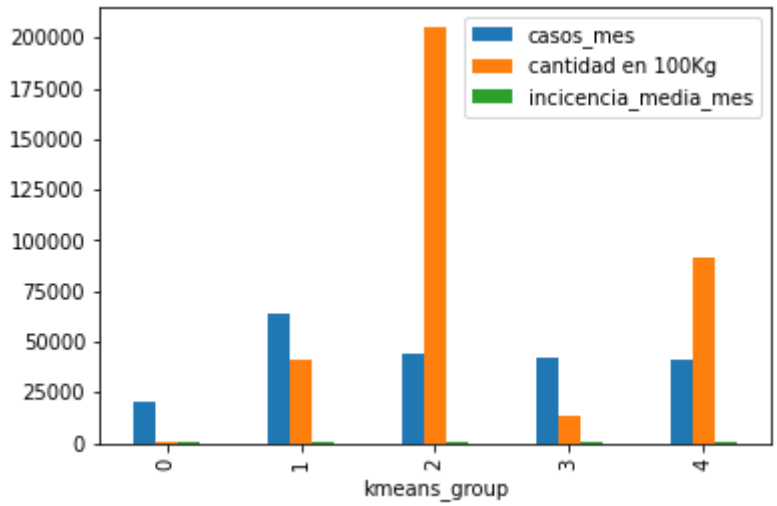
Para trabajar con los datos, estos fueron normalizados para que le algoritmo pudiera trabajar con todos sin ningún problema.



Se realizo el análisis de la distorsión y la silueta respectivamente, junto al uso del algoritmo de maximización GaussianMixture, para determinar la cantidad de clúster óptimo para ejecutar el algoritmo, el cual no fue incluido en el Colab, resulto ser 5 en vez de 3 como se pensó en un primer momento.



Dentro de la imagen anterior podemos observar los 5 clústeres (por la inversion de los ejes se ve de esta manera). Depues de esto, ejecutamos de nuevo K-means y mostramos la gráfica con los resultados finales, dentro de la cuál podemos observar que la hipótesis 1 es errónea



Por lo cual podemos decir que **No existe** una correlación entre el número de casos de enfermos de Corona Virus durante la pandemia y las importaciones/exportaciones en el continente Europeo.

8.2.-Línea de trabajo de la hipótesis 2

Como se puede ver en las líneas de trabajo, inicialmente se decidió emplear regresión lineal para nuestra segunda hipótesis. Sin embargo, como se puede ver en el anexo (archivo Colab compartido) de forma más detallada, el nivel de correlación entre las dos variables que se usarían para comprobar nuestra hipótesis (Valor_miles y Dens_pob) es demasiado bajo como para poder usar regresión lineal. Por ello, preferimos emplear otro método de regresión no lineal como es el caso de los árboles de decisión.

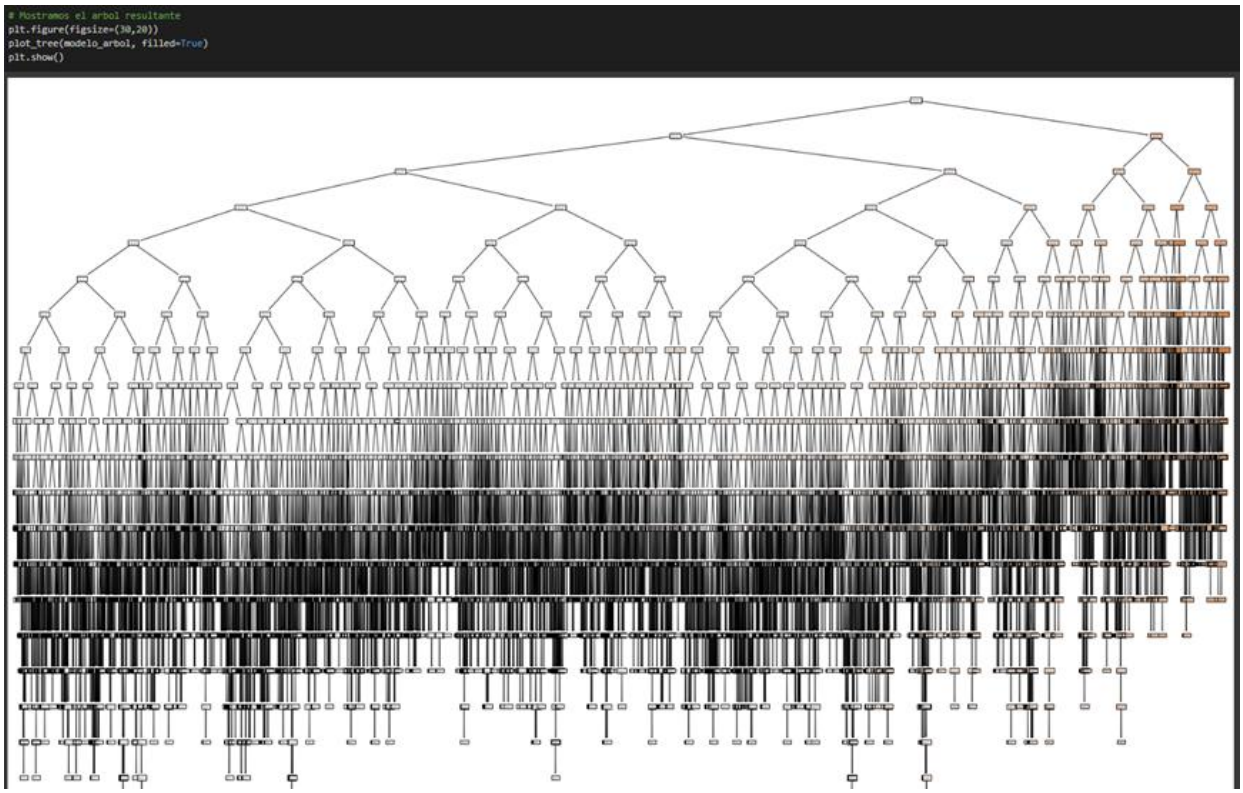
Los árboles de decisión son una técnica de aprendizaje supervisado con el objetivo de predecir valores de respuestas mediante el aprendizaje de decisión derivadas de características. Este algoritmo puede ser empleado para casos de clasificación como de regresión dependiendo de la variable que se use para las predicciones. Como en nuestro caso la variable a predecir es numérica y es un valor continuo, podemos usar los árboles de decisión regresivos.

Dentro de las razones que hemos tenido a la hora de decantarnos por este modelo serían que este modelo es muy útil para explorar los datos de una manera rápida con el fin de identificar las variables más significativas y las relaciones entre estas. Gracias a esto, se puede obtener una predicción mejor para la variable dependiente. Junto a esto, también cabe destacar la facilidad de crear y usar un modelo de árboles de decisión.

Para crear este modelo, usamos en un Colab la librería de sklearn para los métodos de DecisionTreeRegressor y los distintos métodos contenidos en la librería para poder evaluar nuestro modelo. Para la división de los datos contenidos en la tarjeta de datos, nos hemos ayudado de la herramienta BigML para poder dividirlo en subconjuntos de entrenamiento y de testing. Para ello, se han dividido de forma aleatoria ya que no están ordenados según la fecha de los datos, si no que están ordenados según la región (muestra los datos de las regiones una a una ordenada por la fecha) por lo que si se dividían de forma lineal no podíamos obtener un buen resultado a la hora de predecir los valores.

En cuanto a los resultados obtenidos por este modelo no hemos podido mostrarlo gráficamente en el fichero Colab final, posiblemente por la sobrecarga de información contenida en el propio fichero, pero en un fichero aparte sí que logramos mostrar el gráfico del árbol resultante, aunque no se puede

llegar a ampliar lo suficiente como para poder ver el contenido de los nodos, por lo que pedimos disculpas por ello.



Junto a esto, hemos obtenido los valores del coeficiente de determinación (R2) y el coeficiente del error cuadrático medio (MSE), cuyos valores son de 0'99 y 66060'21 respectivamente. Si los comparamos con los resultados que se podrían obtener con regresión lineal (obtenidos con BigML y son de MSE=10511621'27 y R2=0'03), vemos que hay una mejora sustancial por la que creemos que, a pesar de obtener un valor para el MSE relativamente alto, está a un nivel bajo si analizamos los datos que tenemos, indicando que el modelo tiene un rendimiento lo suficientemente alto como para realizar predicciones correctamente.

Para hacer una demostración de la predicción del modelo, la primera predicción que hacemos en el Colab vemos que obtiene un valor de 1936.12, que si lo buscamos dentro de la propia tabla completa vemos que el valor es de 1742'42, lo cual es un resultado que se acerca bastante.

```
Instancia de entrada:
Producto            13.00
Precio_euro_Kg       0.78
Volumen_miles     2245.92
Dens_pob            95.71
Name: 0, dtype: float64

Valor resultante para la entrada 1: [1936.12]
```

Producto	Fecha	Region	Precio_eur	Valor_mile	Volumen_r	Dens_pob
ZANAHORIAS	01/01/2018	AND	0,78	1742,42	2245,92	95,71

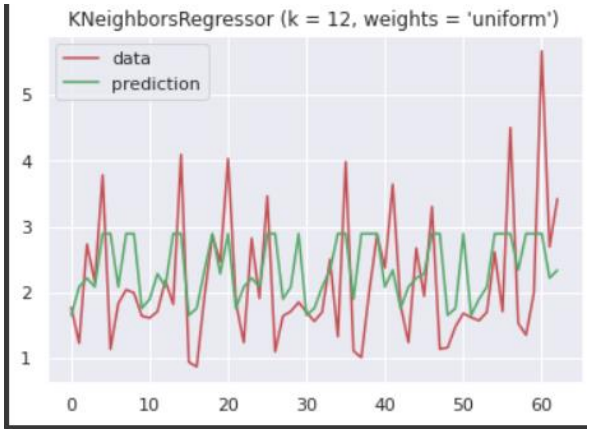
8.3.-Línea de trabajo de la hipótesis 3

En un principio decidimos utilizar PCA para esta hipótesis, pero cuando realmente intentamos resolver el problema vimos que no era la mejor decisión. Debido a esto decidimos usar un algoritmo de regresión para así predecir el precio de los productos en un futuro. Para ello hemos usado KNN, que

se usa para estimar la relación entre una variable dependiente y una o más variables independientes. El método KKN utiliza una técnica de agrupamiento llamada K-means para dividir los datos en k grupos, donde k es un número especificado por el usuario. A continuación, se ajusta un modelo de regresión a cada grupo de datos. Finalmente, se combinan los resultados de cada modelo de regresión para obtener la estimación final.

Hemos elegido KNN porque tiene buen rendimiento en conjuntos de datos pequeños, ya que utiliza la información de todas las observaciones para hacer predicciones y porque es fácil de implementar (el algoritmo KNN es sencillo de entender y de implementar, lo que lo hace adecuado para principiantes en el aprendizaje automático).

Como ya hemos comentado, el algoritmo es sencillo y los problemas vienen al separar los datos de entrenamiento y test. Primero consideramos separar el 80% de los datos para entrenamiento y el 20% restante para test. Una vez que ejecutamos y vimos los resultados, vimos que la predicción no tenía nada que ver con los datos originales porque todavía no había tenido suficiente impacto sobre los precios de los productos el aumento de casos de covid ya que la separación de datos se encontraba en marzo de 2020. Por ello, decidimos ampliar el conjunto de datos de entrenamiento hasta agosto-septiembre y la predicción mejora considerablemente. Con esto, comprobamos que hay un impacto relacionado entre los casos de covid y el aumento del precio.



Como apreciamos en la figura, los casos en los que los picos no son muy altos y distinguidos, el modelo se asemeja bastante a los datos reales.

8.4.-Línea de trabajo de la hipótesis 4

Para esta línea de trabajo hemos decidido usar regresión puesto que queremos predecir un valor continuo. Vamos a usar la tarjeta de datos 4 sobre el consumo alimentario por comunidades autónomas.

Lo primero que hemos hecho es verificar si usaremos regresión lineal o no lineal. Para ello hemos empezado comprobando la correlación de las diferentes variables de la tarjeta de datos para descartar o no la regresión lineal.

La regresión lineal es una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal. Como nos dimos cuenta de que si comprobamos la correlación de las columnas de primeras se falsearían los datos porque está mezclando datos de diferentes alimentos lo que hicimos fue crear un bucle que permite analizar la correlación entre las columnas centrado siempre en un único alimento. El resultado de esto fue que ninguna columna tenía la suficiente correlación con ninguna otra excepto en dos

alimentos que cada uno tenía una correlación, lo cual lo hace inútil y arbitrario. Esto nos llevó a descartar la regresión lineal e ir a por un enfoque centrado en la regresión no lineal.

Dentro de la regresión no lineal decidimos que usaríamos Random Forest Regression porque permite tratar con relaciones de datos complejos y proporciona más precisión que el método de árboles de decisión. Random Forest es un conjunto de árboles de decisión combinados con bagging. Al usar bagging, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor que un sólo árbol. Lo que hicimos fue separar la tarjeta de datos en dos conjuntos de datos para entrenamiento y para testeo usando random select porque en este caso nos convenía más siendo que estaba el covid en las últimas fechas y eso a lo mejor alteraría los datos en parte. A continuación, entrenamos el modelo con el dataset de entrenamiento usando 100 árboles de decisión y al hacer una predicción y compararla con el dataset de testeo obtuvimos las siguientes métricas:

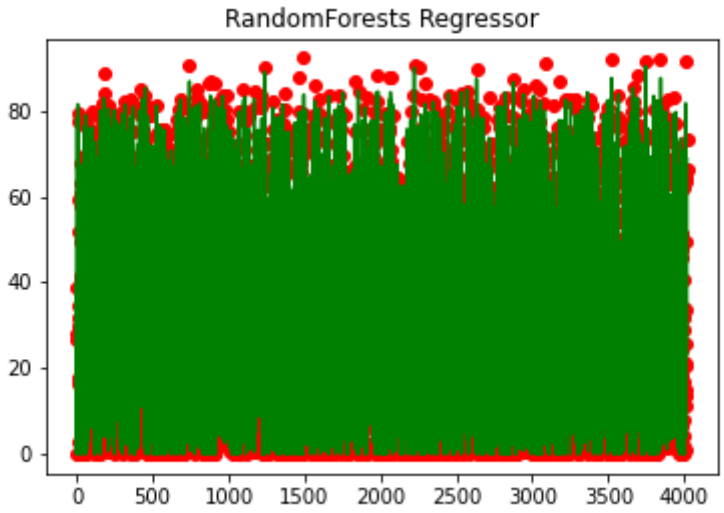
```
MSE: 26.716945626546405
R2: 0.9566176686608688
```

Como podemos ver las métricas son muy bajas y el coeficiente de determinación es alto, por lo que parece que va a ajustarse bien el modelo a los datos. Por si acaso hicimos una comprobación y sacamos una gráfica para ilustrar el modelo:

- Para la comprobación hicimos una predicción usando los siguientes datos:
  - Año: 2018
  - Mes: abril
  - Comunidad autónoma: Madrid
  - Producto: KIWI
  - Volumen en kg: 1784.81
  - Precio medio por kilogramo: 2.84

  La predicción resultante es 39,4677 y el dato real es 38,61. Como podemos ver es un número cercano.

- La gráfica que sacamos es la siguiente:



En esta imagen los puntos rojos son los datos reales y las líneas verdes son las predicciones. Como podemos ver están cerca unos de otros, por lo que efectivamente podemos comprobar nuevamente que es un buen modelo.