IMF Business School
Universidad Camilo José Cela

Máster en
Big Data & Business Analytics

Trabajo Fin de Máster

Análisis de productos químicos en cosméticos

Autor:
José María
Sánchez Salas

Tutor:
Juan Manuel
Moreno Lamparero

Marzo 2019

Una de las cosas más fascinantes de los programadores es que no puedes saber si están trabajando o no sólo con mirarlos. A menudo están sentados aparentemente tomando café, chismorreando o mirando a las nubes. Sin embargo, es posible que estén poniendo en orden todas las ideas individuales y sin relación que pululan por su mente.

Charles M. Strauss

Como dirían en el cole:
no hay mejor consejo que el de ser tú mismo,
a pesar de que al resto no les mole.

ZPU

Agradecimientos

Resumen

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y antecedentes

1.1. Introducción

La preocupación del ser humano por su aspecto y su cuidado es algo que se ha ido manteniendo a lo largo de los siglos, pues aunque hoy en día la industria cosmética parezca algo tecnológico y sofisticado, las primeras civilizaciones ya se preocupaban por ello. Precisamente, en el Antiguo Egipto se encuentran los primeros vestigios de la elaboración y utilización de diferentes productos cosméticos, utilizando para ello productos naturales, como plantas aromáticas.

La cosmética es uno de los sectores que mayor auge ha vivido durante las últimas décadas. Entre los siglos XVI al XVIII se produjo un gran desarrollo de los cosméticos y se introdujeron numerosos productos nuevos, aún fabricados principalmente a base de plantas. Pero fue a partir de principios del siglo XX cuando los cosméticos se popularizarían, hasta convertirse hoy en un producto casi imprescindible en la mayoría de los hogares.

Hoy en día, los cosméticos han vuelto a incorporar productos químicos dentro de sus fórmulas y se utilizan miles de estos compuestos a los que se le atribuyen multitud de propiedades. Sin embargo, varios cientificos y organizaciones han levantado la voz de alerta sobre el impacto de estos compuestos, pues en un alto porcentaje no han sido analizados para saber el daño de estos productos sobre las personas.

1.2. Antecedentes

Este trabajo se embarca dentro del contexto de la aplicación de técnicas Big Data e Inteligencia Artificial al sector de la cosmética. Existen varios productos y prototipos de productos que aplican al sector cosmético este tipo de técnicas, como por ejemplo el servicio Identité, que utiliza técnicas de Inteligencia Artificial y Big Data para “Crear paquetes de belleza y productos para el cuidado de la piel muy personalizados.”.

1.3. Estructura de la memoria

Capítulo 2

Hipótesis de trabajo y objetivos

2.1. Introducción

En este capítulo se va a exponer la hipótesis de este trabajo así como los objetivos propuestos para el mismo.

2.2. Hipótesis de trabajo

Los cosméticos son unos de los productos que más se consumen en los hogares, hasta tal punto que en algunos casos se han vuelto imprescindibles. Muchos productos químicos diferentes se utilizan en la fabricación de cosméticos. Los grupos de defensa de consumidores y trabajadores están preocupados porque algunos productos cosméticos contienen sustancias químicas que se sabe o se sospecha que causan cáncer, defectos de nacimiento o daños al sistema reproductivo. Aquellos que trabajan con cosméticos, incluidos los peluqueros, los estilistas y el cuidado de la piel, el cuidado del cuerpo y los trabajadores de salones de uñas, pueden ser más vulnerables a los efectos adversos para la salud que presentan estos productos porque manejan con mayor frecuencia grandes cantidades de cosméticos (State of California, 2019).

Con esta idea en mente, la hipótesis de este trabajo es aplicar técnicas de Inteligencia Artificial y Big Data al sector de la cosmética para poder encontrar patrones y relaciones entre los cosméticos y los productos químicos que contienen, así como encontrar una clasificación de dichos productos químicos para poder tener un mayor control sobre de qué están compuestos los cosméticos que consumimos.

2.3. Objetivos

Los objetivos principales de este trabajo son:

- Encontrar una clasificación de los productos químicos presentes en los cosméticos.
- Encontrar qué productos químicos son los más frecuentes en los cosméticos, así como los cosméticos que mayor productos químicos presentan.
- Obtener una predicción de la cantidad de productos químicos dañinos que contendrán los futuros cosméticos.

Capítulo 3

Material y métodos

3.1. Introducción

En este capítulo se detallará, en primer lugar, el dataset utilizado para el desarrollo de este trabajo, la descripción de sus características y la selección de las características más importantes. En segundo lugar, se detallarán los métodos y técnicas que se han aplicado al dataset para conseguir los objetivos de este trabajo.

3.2. Material

Para la realización de este trabajo se ha hecho uso del dataset público **Chemicals in Cosmetics** (State of California, 2019) proporcionado por **HealthData** (U.S. Department of Health & Human Services, 2019a).

Para todos los productos cosméticos vendidos en California, la Ley de Cosméticos Seguros de California requiere que el fabricante, empacador y/o distribuidor nombrado en la etiqueta del producto proporcione una lista al Programa de Cosméticos Seguros de California *(California Safe Cosmetics Program (CSCP))* perteneciente al Departamento de Salud Pública de California *(California Department of Public Health (CDPH))* de todos los productos cosméticos que contengan cualquier ingrediente conocido o sospechoso de causar cáncer, defectos de nacimiento u otros daños al desarrollo o reproductivos. El CSCP mantiene una lista de ingredientes "reportables". Las compañías con ingredientes reportables en sus productos deben enviar información al Programa de Cosméticos Seguros de California si la compañía:

- Tiene ventas anuales agregadas de productos cosméticos de un millón de dólares o más, y
- Ha vendido productos cosméticos en California a partir del 1 de enero de 2007.

El CSCP mantiene un sistema de informes *online* para que las empresas informen sobre productos e ingredientes reportables, generando así el dataset utilizado para este trabajo. Los datos reflejan información que ha sido reportada al CSCP. No se incluyen todos los productos que contengan carcinógenos o tóxicos para el desarrollo o la reproducción, debido a que las compañías no los informan.

3.2.1. Descripción del dataset

El dataset proporcionado por el CSCP es un dataset en formato CSV con un histórico desde el año 2009 y que se va actualizando cada 3 o 4 días. En las fechas de realización de este trabajo (última fecha reportada: 21/02/2019), el dataset consta de:

- 97.760 registros.

- 22 columnas (características).

En la Tabla 3.2 se describen cada una de las características. La agrupación de las características `CDPHId`, `CSFId`, `SubCategoryId` y `ChemicalId` forma la clave primaria de este dataset.

3.2.2. Selección de características importantes

No todas las características descritas en la Tabla 3.2 son necesarias para aplicar las técnicas descritas en la sección 3.3 ya que algunas de ellas son informativas o no son relevantes a la hora de aplicar dichas técnicas.

Se van a tener en cuenta todas las características de formato Fecha, exceptuando las características `DiscontinuedDate` y `ChemicalDateRemoved`, ya que solamente interesan aquellos productos que no hayan sido eliminados. Además, se van a utilizar las características `SubCategoryId` y `CASId`, pues son los identificadores de los cosméticos y los productos químicos, respectivamente. Se utiliza `SubCategoryId` en vez de `PrimaryCategoryId` debido a que la primera es más específica que la segunda y a partir de la primera se puede sacar la segunda (pues está contenido en ella). Y por supuesto, `ChemicalCount`, que indica la cantidad de productos químicos.

Así pues, la Tabla 3.1 muestra las características que se van a tener en cuenta:

Nombre
`SubCategoryId`
`CASId`
`InitialDateReported`
`MostRecentDateReported`
`ChemicalCreatedAt`
`ChemicalUpdatedAt`
`ChemicalCount`

TABLA 3.1: Selección de características importantes.

Nombre	Formato	Definición
CDPHId	Texto	Número identificativo interno del CDPH para el producto.
ProductName	Texto	Nombre del producto introducido por el fabricante, empacador y/o distribuidor.
CSFId	Texto	Número identificativo interno del CDPH para el color/aroma/sabor.
CSF	Texto	Color, aroma y/o sabor introducido por el fabricante, empacador y/o distribuidor.
CompanyId	Texto	Número identificativo interno del CDPH para la compañía.
CompanyName	Texto	Nombre de la compañía introducido por el fabricante, empacador y/o distribuidor.
BrandName	Texto	Nombre de la marca introducido por el fabricante, empacador y/o distribuidor.
PrimaryCategoryId	Texto	Número identificativo interno del CDPH para la categoría.
PrimaryCategory	Texto	Tipo de producto (13 categorías primarias).
SubCategoryId	Texto	Número identificativo interno del CDPH para la subcategoría.
SubCategory	Texto	Tipo de producto dentro de una de las categorías primarias.
CASId	Texto	Número identificativo interno del CDPH para el producto químico.
CasNumber	Texto	Número identificativo del producto químico seleccionado por el fabricante, empacador y/o distribuidor.
ChemicalId	Texto	Número identificativo interno del CDPH para el registro específico del producto químico en ese cosmético.
ChemicalName	Texto	Nombre del producto químico seleccionado por el fabricante, empacador y/o distribuidor.
InitialDateReported	Fecha	Fecha en la que el cosmético fue reportado por primera vez al CDPH.
MostRecentDateReported	Fecha	Fecha de la última modificación del perfil del cosmético por el fabricante, empacador y/o distribuidor.
DiscontinuedDate	Fecha	Si aplica, fecha en la que el cosmético fue interrumpido.
ChemicalCreatedAt	Fecha	Fecha en la que el producto químico fue reportado al CDPH por primera vez en ese cosmético.
ChemicalUpdatedAt	Fecha	Fecha de la última modificación del perfil del producto químico por el fabricante, empacador y/o distribuidor.
ChemicalDateRemoved	Fecha	Si aplica, fecha en la que el producto químico fue eliminado del cosmético.
ChemicalCount	Número	Número total de productos químicos reportados en ese cosmético

TABLA 3.2: Descripción de las características del dataset.

3.3. Métodos

Este trabajo se va a realizar utilizando el lenguaje de programación Python y como entorno de desarrollo Jupyter Notebook. Los métodos que se van a utilizar para la consecución de los objetivos descritos en la sección 2.3 son:

3.3.1. Obtención del dataset

Como se ha comentado en la sección 3.2, el dataset se haya ubicado en la plataforma HealthData y además es incremental (State of California, 2019), por lo que para obtener dicho dataset se va a hacer uso de la API (U.S. Department of Health & Human Services, 2019b) que proporciona HealhData.

3.3.2. Clustering

Para poder obtener la clasificación de los productos químicos presentes en los cosméticos, se van a aplicar técnicas de Clustering. Concretamente, se va a utilizar la implementación del algoritmo K-Means proporcionada por Scikit-learn (Pedregosa et al., 2011).

3.3.3. Data Analysis

Para poder encontrar qué productos químicos son más frecuentes, así como aquellos cosméticos que presentan mayor número de productos químicos, se van a aplicar técnicas de Data Science como el Análisis Exploratorio de Datos *(Exploratory Data Analysis (EDA))* (NIST/SEMATECH, 2012). La aplicación de EDA se va a realizar haciendo uso de los resultados obtenidos tras la clusterización.

3.3.4. Forecasting

Para poder obtener la predicción de la cantidad de productos químicos que contendrán los futuros cosméticos, se van a aplicar técnicas de Forecasting. Concretamente, se va a utilizar la implementación del modelo ARIMA proporcionada por StatsModels (StatsModels, 2019).

Capítulo 4

Aplicación de Técnicas y Resultados

4.1. Introducción

En este capítulo se realizará la explicación de todas las técnicas aplicadas sobre el dataset. En primer lugar, se detallará cómo se ha realizado la obtención del dataset; en segundo lugar, cómo se ha realizado la aplicación de las técnicas de clustering; en tercer lugar, los resultados obtenidos tras la aplicación del análisis exploratorio de los datos; y en último lugar, cómo se ha realizado la aplicación de las técnicas de forecasting.

El código asociado a este capítulo se encuentra en el repositorio público (Salas, 2019).

4.2. Obtención del dataset

El dataset empleado para este trabajo es un dataset incremental actualizado cada 3 o 4 días, por lo que solo sería necesario descargar el dataset cuando haya sido actualizado. Sin embargo, el volumen de registros añadidos en cada actualización no son significativos para la aplicación de las técnicas de clustering y forecasting, de tal manera que la descarga de los datos se realiza si han pasado, como mínimo, 15 días desde la última descarga del dataset (el número de días es un parámetro configurable). Así, cada vez que se descarga el dataset, el volumen de registros nuevos que sí es significativo para la aplicación de las técnicas de clustering y forecasting.

El dataset en formato CSV se almacena dentro de la carpeta `src/data/` mientras que la información que viene asociada (State of California, 2019) se almacena en la carpeta `src/data/info/`. Por último, junto al dataset se almacena el fichero `local_data_date` con la marca de tiempo en el que se descargó el dataset. Cada vez que se descarga el dataset, el antiguo dataset (y su información asociada) queda almacenado en la carpeta `src/data_backup/`.

Todo esto se encuentra implementado en el notebook `src/download_data.ipynb` (Salas, 2019).

4.3. Clustering

La aplicación de técnicas de clustering permite obtener agrupaciones y relaciones en los datos que a simple vista no se pueden obtener. En esta sección se va a realizar la aplicación del algoritmo de clustering K-Means (Pedregosa et al., 2011), realizando primero un preprocesamiento de los datos y posteriormente la obtención de algunas métricas sobre los clusters obtenidos. Todo esto se encuentra implementado en el notebook `src/clustering-and-data-analysis.ipynb` (Salas, 2019).

4.3.1. Preprocesamiento

Antes de poder aplicar el algoritmo de clustering K-Means (Pedregosa et al., 2011), el dataset debe ser preprocesado. El preprocesamiento de este dataset ha consistido en los siguientes pasos:

- **Rellenar valores nulos**. En el dataset se han encontrado tres tipos de valores nulos:
  - **Valores de formato fecha**. Han sido rellenados con el valor `01/01/1990`.
  - **Valores de formato texto**. Han sido rellenados con el valor de la cadena vacía.
  - **Valores de formato texto asociados a identificadores**. Han sido rellenados con el valor `-1`.
- **Eliminar de los registros** con valor distinto de `01/01/1990` en los campos `DiscontinuedDate` y `ChemicalDateRemoved`, pues solo se precisan de aquellos registros de cosméticos que tengan productos químicos y no hayan sido retirados del mercado.
- **Seleccionar las características** expuestas en la Tabla 3.1.
- **Agrupar por año y mes** de cada una de las características de formato fecha, sumando los valores del campo `ChemicalCount`.
- **Seleccionar características que presenten multicolinealidad** en el dataset aplicando el Factor de Inflación de la Varianza *(Variance Inflation Factor (VIF))* (Statistics How To, 2015).

Tras aplicar este preprocesamiento, el dataset resultante se compone de 7.487 registros y 7 características, las cuales se muestran en la Tabla 4.1, donde `_Year` y `_Month` indican el año y el mes del campo asociado, respectivamente.

Nombre
`InitialDateReported_Year`
`InitialDateReported_Month`
`MostRecentDateReported_Year`
`MostRecentDateReported_Month`
`SubCategoryId`
`CasId`
`ChemicalCount`

TABLA 4.1: Características del dataset después del preprocesamiento.

4.3.2. Aplicación del algoritmo K-Means

Una vez aplicado el preprocesamiento, tenemos el dataset preparado para poder aplicar el algoritmo de clustering K-Means (Pedregosa et al., 2011). Sin embargo, este algoritmo necesita que se le proporcione el número de clusters en los que dividir el dataset. Con las características que presenta el dataset no se puede saber el número óptimo de clusters sin aplicar alguna técnica que nos proporcione este número.

Para la obtención del número óptimo de clusters, se han utilizado los métodos de la Silueta *(Silhouette Method)* (Pedregosa et al., 2011) y del Codo *(Elbow Method)* (RicardoMoya,

2016). En la Figura 4.1 se muestran las dos gráficas de la aplicación de los métodos anteriores al dataset. Estas gráficas nos indican que el número óptimo de clusters en los que se divide el dataset es 3.

Una vez obtenido el número óptimo de clusters, se puede aplicar el algoritmo K-Means indicándole que el número de clusters es 3. Las Tablas A.1, A.2, A.3 y A.4 muestran la distribución de los productos químicos en los tres clusters tras aplicar el algoritmo.

La Figura 4.2 muestra la distribución de los cluster obtenidos según la relación entre los productos químicos `CasId` y los cosméticos `SubCategoryId`. Mientras que la Tabla 4.2 muestra la distribución de los clusters de manera numérica.

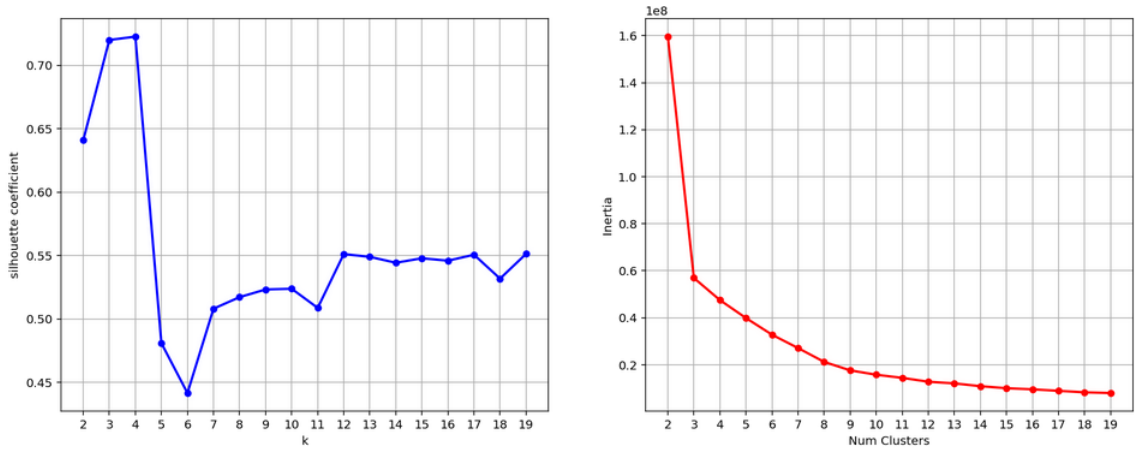

FIGURA 4.1: Aplicación de los métodos *Silhouette* (izquierda) y *Elbow* (derecha) para la obtención del número óptimo de clusters.

FIGURA 4.2: Distribución de los clusters según la relación entre los productos químicos `CasId` y los cosméticos `SubCategoryId`.

Cluster 0	Cluster 1	Cluster 2
5.532	1.476	479

TABLA 4.2: Distribución de los clusters según el número de registros de cada cluster.

4.3.3. Métricas del clustering

Tras aplicar el clustering, se van a calcular ciertas métricas sobre los clusters obtenidos para ofrecer los resultados del clustering de manera más precisa. Las métricas que han sido utilizadas son las siguientes:

- `Average Within` (Lamparero, 2018). Mide la distancia media dentro de las observaciones de cada cluster (distancia intra-cluster). Viene definido por la ecuación 4.1, siendo K el número de clusters, C_i el conjunto de elementos del cluster i y $centroid_i$ el centroide del cluster i:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(centroid_i, x)^2 \tag{4.1}$$

- `Dunn Index` (Lamparero, 2018). Define el ratio entre la distancia mínima inter-cluster y la máxima distancia intra-cluster. Viene definido por la ecuación 4.2, siendo C el conjunto de todos los clusters.

$$D(C) = \frac{\min\limits_{C_k, C_l \in C, C_k \neq C_l} \left(\min\limits_{i,j \in C_k, C_l} dist(i,j) \right)}{\max\limits_{C_m \in C} diam(C_m)} \tag{4.2}$$

Así pues, las métricas obtenidas se muestran en la Tabla 4.3:

`Average Within`	`Dunn Index`
0,048633	$5{,}692750 \cdot 10^7$

TABLA 4.3: Valores obtenidos para las métricas `Average Within` y `Dunn Index`.

4.4. Data Analysis

El Análisis Exploratorio de Datos *(Exploratory Data Analysis (EDA))* (NIST/SEMATECH, 2012) es muy importante para poder obtener información y conocimiento acerca del dataset. Así pues, en esta sección se van a aplicar distintas técnicas para poder obtener qué productos químicos son los más frecuentes, qué cosméticos son los que presentan mayor número de productos químicos, cómo se distribuyen los datos en función de ciertos campos y cómo se distribuyen por cluster, la cantidad de productos químicos totales. Todo esto se encuentra implementado en el notebook `src/clustering-and-data-analysis.ipynb` (Salas, 2019).

4.4.1. Obtención de los productos químicos más frecuentes en los cosméticos

La obtención de los productos químicos más frecuentes en los cosméticos se va a realizar realizando histogramas sobre el campo `CasId` del dataset. La Figura 4.3 muestra el histograma sobre todo el dataset, donde se puede observar que entre los valores 600 y 700 hay un gran volumen.

Concretamente, este volumen se encuentra entre los valores 656 and 658. La Figura 4.4 muestra el histograma entre dichos valores, donde se puede observar que el volumen se encuentra en el producto químico con `CasId` 656, cuyo nombre es:

- `656 - Titanium dioxide`.

y pertenece al Cluster 0.

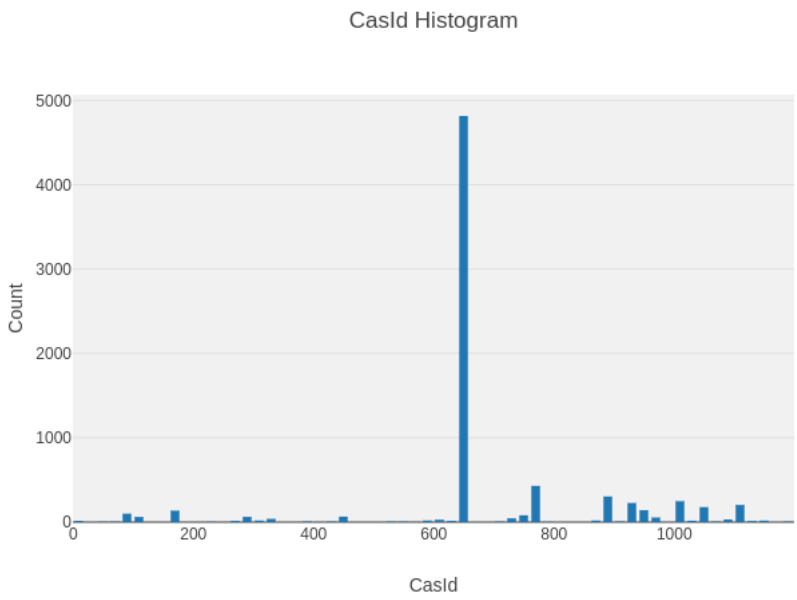

FIGURA 4.3: Histograma sobre el campo `CasId`.

Sin embargo, como podemos observar en la Figura 4.3, la diferencia de volumen es muy grande entre el `CasId` 656 y el resto. Por lo que se va a realizar los mismos pasos anteriores, pero quitando el `CasId` 656 de los datos.

La Figura 4.5 muestra el histograma sobre el campo `CasId` del dataset sin el `CasId` 656 y, además, diferenciando por cluster, donde se puede observar que sigue habiendo un gran volumen entre los valores 700 y 800 y que pertenecen al Cluster 0.

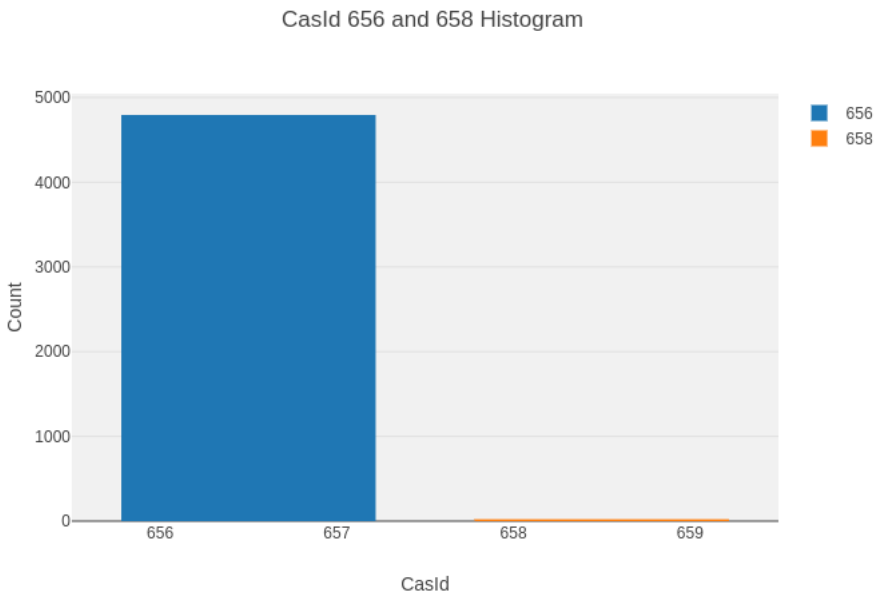

FIGURA 4.4: Histograma de los valores 656 y 658 del campo `CasId`.

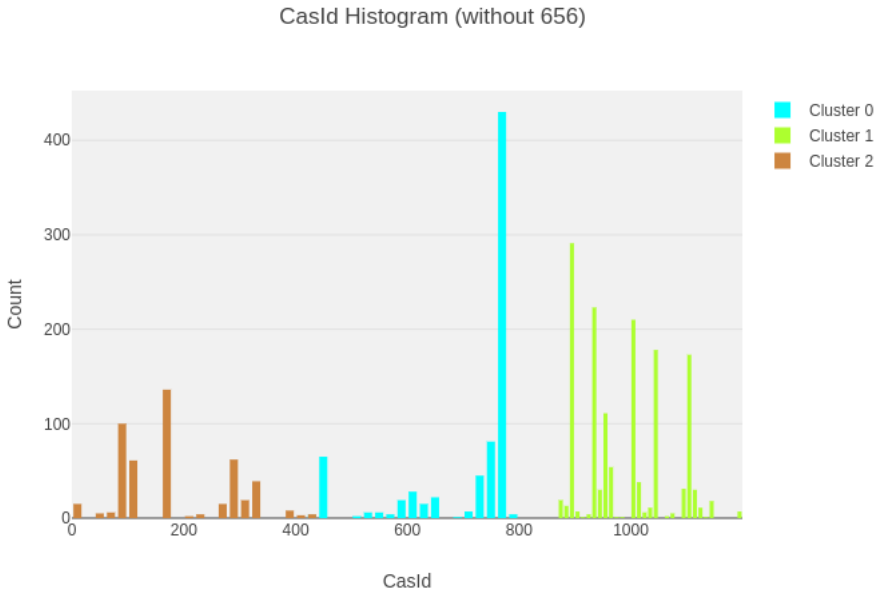

FIGURA 4.5: Histograma sobre el campo `CasId` sin el `CasId` 656, diferenciando por cluster.

Concretamente, este volumen se encuentra entre los valores 773 y 776. La Figura 4.6 muestra el histograma de dichos valores, donde se puede apreciar que el `CasId` 773 tiene mayor volumen. El nombre de cada uno de los productos químicos es:

- `773 - Retinol/retinyl esters, when in daily dosages in excess of 10,000 IU, or 3,000 retinol equivalents.`
- `776 - Silica, crystalline (airborne particles of respirable size).`

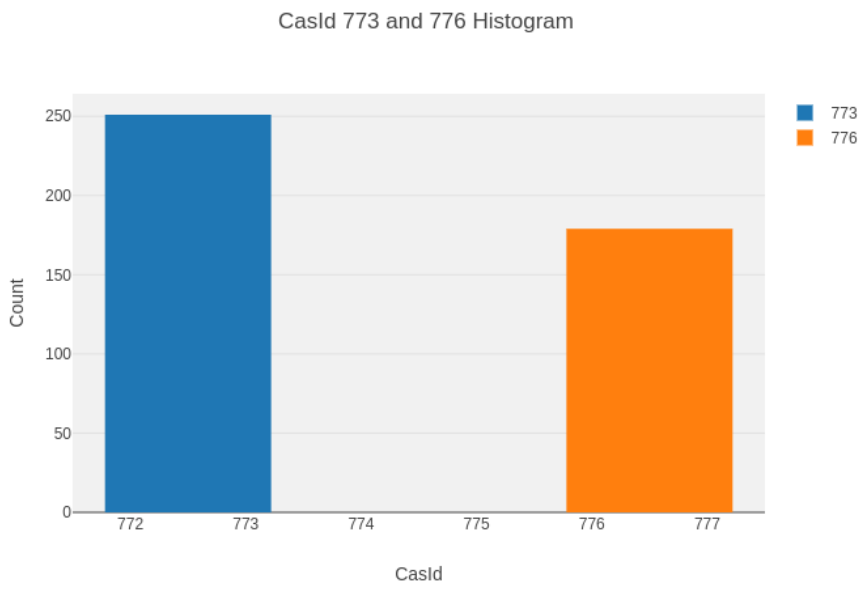

FIGURA 4.6: Histograma de los valores 773 y 776 del campo `CasId`.

4.4.2. Obtención de los cosméticos con mayor número de productos químicos

Con la misma filosofía que en la sección 4.4.1, se van a realizar histogramas sobre el campo `SubCategoryId` para obtener los cosméticos que presentan mayor número de productos químicos. La Figura 4.7 muestra el histograma sobre el campo `SubCategoryId` de todo el dataset diferenciando por cluster.

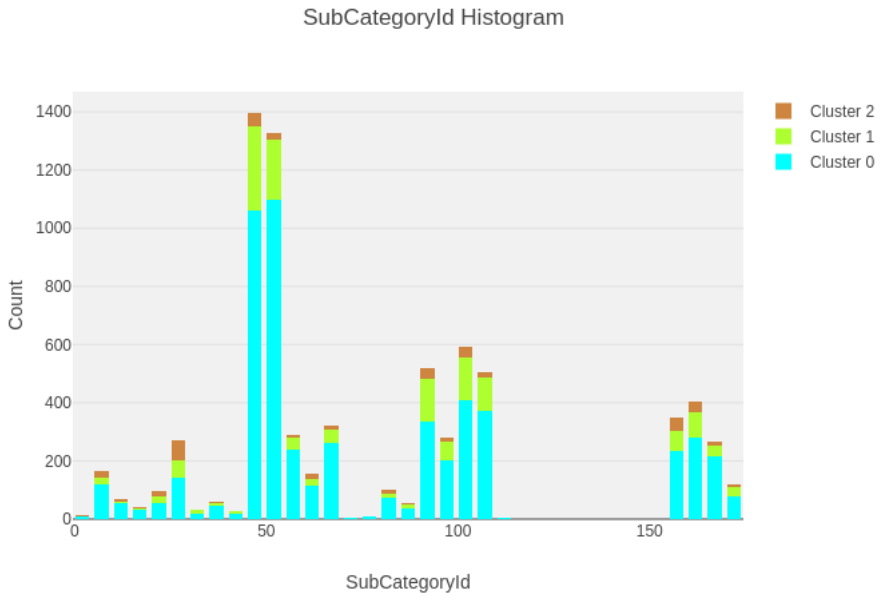

FIGURA 4.7: Histograma sobre el campo `SubCategoryId` diferenciando por cluster.

Como se puede observar, hay un gran volumen de registros que pertenecen al Cluster 0, esto es debido al volumen que presenta el `CasId` 656 en todo el dataset. Por lo tanto, para poder hacer un estudio más detallado, se va a eliminar el `CasId` 656. Así pues, la Figura 4.8 muestra el mismo histograma que en la Figura 4.7 pero sin el `CasId` 656.

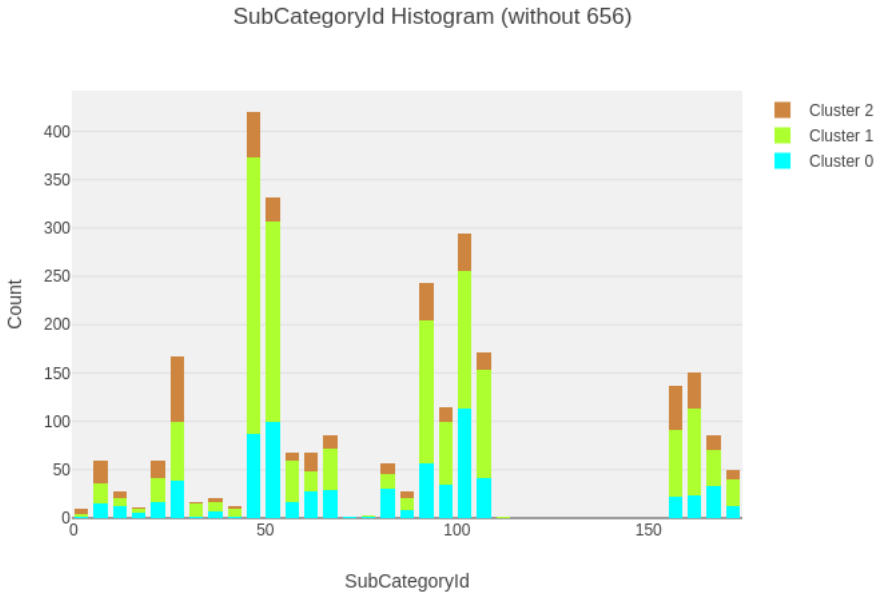

FIGURA 4.8: Histograma sobre el campo `SubCategoryId` diferenciando por cluster, sin el `CasId` 656.

Al igual que ocurrió con el campo `CasId`, se puede observar que entre los valores 40 y 50 del campo `SubCategoryId` se encuentra un volumen superior al resto. Concretamente, se encuentra entre los valores 45, 46, 48 y 49. La Figura 4.9 muestra el histograma de dichos valores, donde se puede apreciar que el valor `SubCategoryId` 48 es el que tiene un volumen mayor. Además, también se puede observar que la gran mayoría de los registros pertenecen al Cluster 1.

El nombre de cada uno de los cosméticos es:

- `45 - Blushes.`
- `46 - Eyeliner/Eyebrow Pencils.`
- `48 - Eye Shadow.`
- `49 - Face Powders.`

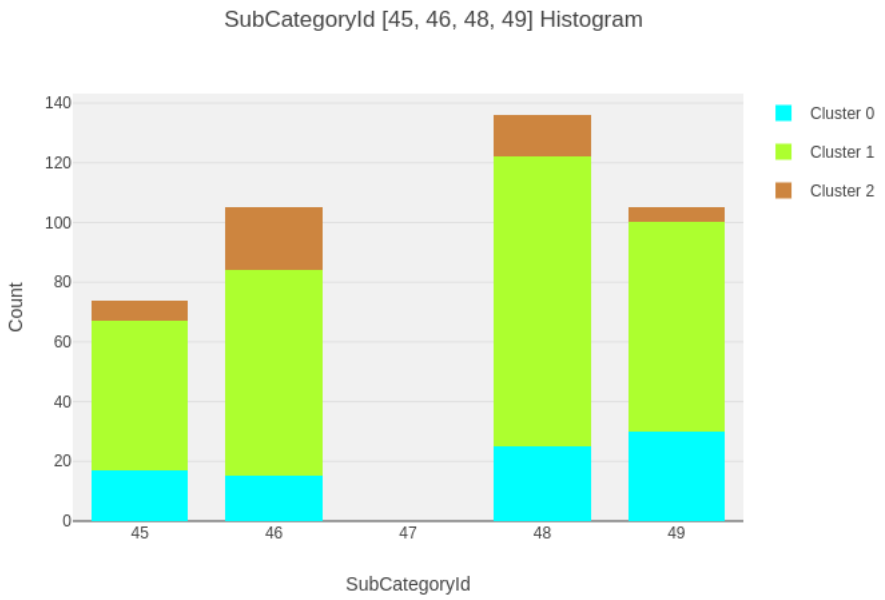

FIGURA 4.9: Histograma de los valores 45, 46, 48 y 49 del campo `SubCategoryId` diferenciando por cluster, sin el `CasId` 656.

4.4.3. Distribución del dataset

En la Figura 4.10 se muestra la distribución del dataset en función de los campos `CasId`, `InitialDateReported_Year`, `MostRecentDateReported_Year` y `SubCategoryId`, diferenciando por cluster, en el que se puede observar que en los primeros años (2009 y 2010) fue cuando se reportaron la gran mayoría de los productos químicos y que en los años siguientes han ido reportándose de una manera muy equilibrada.

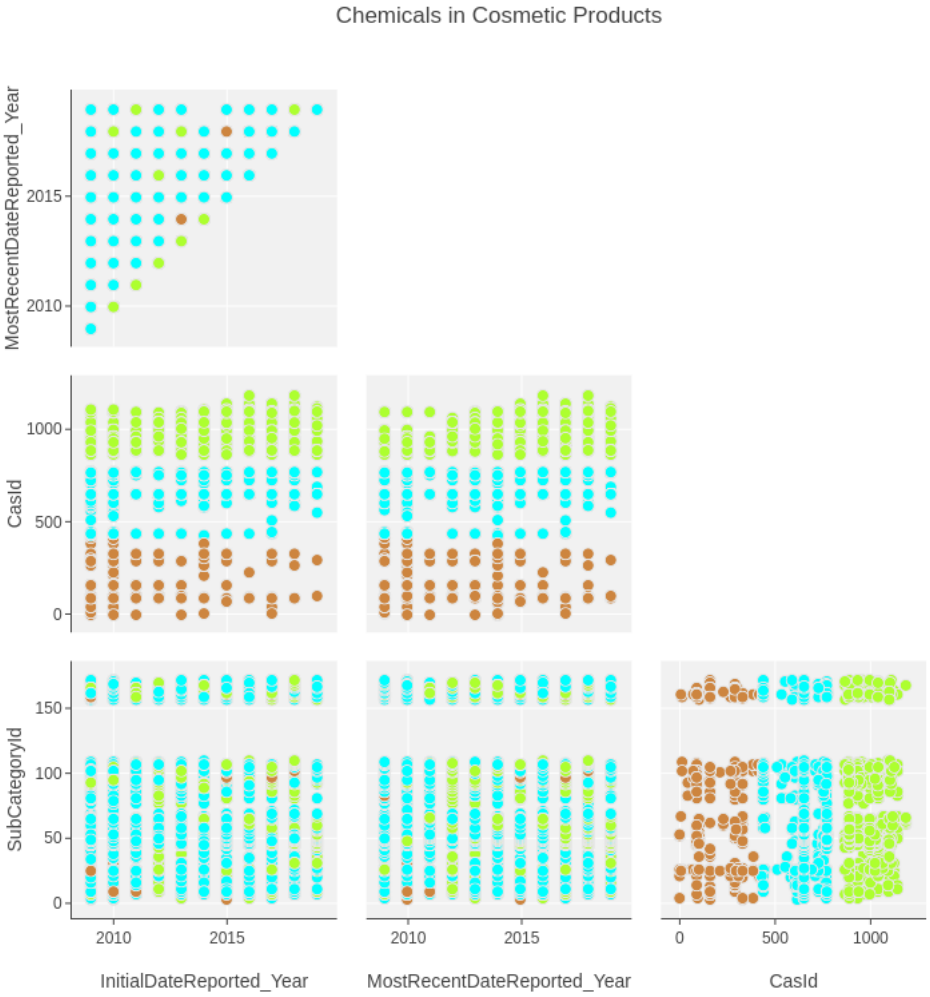

FIGURA 4.10: Distribución del dataset en función de los campos `InitialDateReported_Year`, `MostRecentDateReported_Year`, `SubCategoryId` y `CasId`.

4.4.4. Distribución de la cantidad de productos químicos por cluster

Hasta ahora se han estudiado la frecuencia en la que aparecen los productos químicos y los cosméticos en el dataset. En este punto, se va a estudiar la distribución de la cantidad de productos químicos en cada uno de los clusters, esto es: la suma del campo `ChemicalCount` distribuido en cada uno de los clusters. La Figura 4.11 muestra esta distribución.

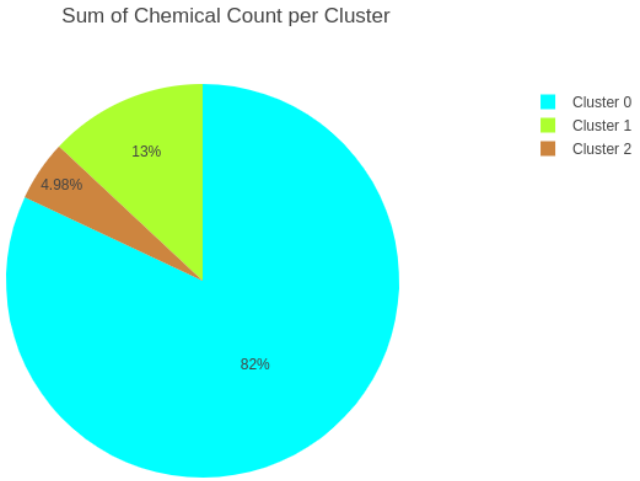

FIGURA 4.11: Distribución de la suma del campo `ChemicalCount` por cada cluster.

Al igual que pasaba en la sección 4.4.2, el alto porcentaje del Cluster 0 se debe al `CasId` 656. Por lo que, la Figura 4.12 muestra la misma distribución eliminando el `CasId` 656 del dataset. Donde se puede observar que más del 50 % de los productos químicos se encuentran en el Cluster 1.

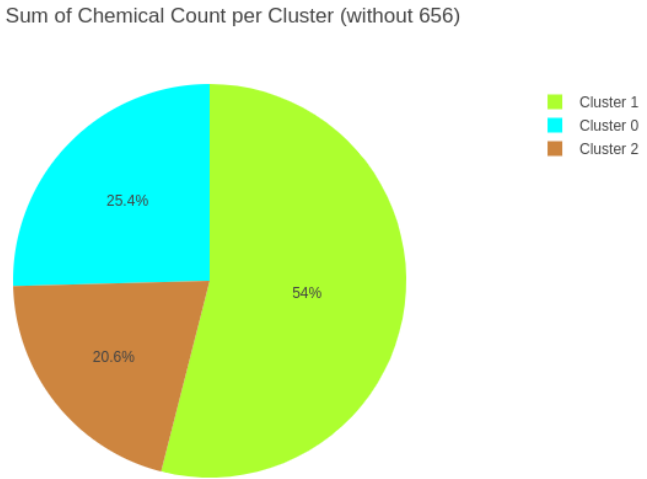

FIGURA 4.12: Distribución de la suma del campo `ChemicalCount` por cada cluster, sin el `CasId` 656.

4.5. Forecasting

El forecasting permite realizar predicciones futuras basadas en un histórico de datos. Dado que el dataset (State of California, 2019) contiene un histórico desde el año 2009, es idóneo para aplicar este tipo de técnicas. En esta sección se van a detallar los pasos necesarios para la aplicación del algoritmo de forecasting ARIMA (StatsModels, 2019), realizando primero un preprocesamiento de los datos, seguido de la obtención de los conjuntos de datos para el entrenamiento y la validación del algoritmo. Por último, se aplicará el algoritmo ARIMA sobre 4 agrupaciones distintas de los datasets.

Todo esto se encuentra implementado en el notebook `src/forecasting.ipynb` (Salas, 2019).

4.5.1. Preprocesamiento

Antes de poder aplicar el algoritmo de forecasting ARIMA (StatsModels, 2019), el dataset debe ser preprocesado, al igual que se realizó con el clustering en la sección 4.3.1. Sin embargo, no se ha realizado el mismo preprocesamiento. A continuación se detallan los pasos realizados en este preprocesamiento:

- **Rellenar valores nulos**. En el dataset se han encontrado tres tipos de valores nulos:
  - **Valores de formato fecha**. Han sido rellenados con el valor `01/01/1990`.
  - **Valores de formato texto**. Han sido rellenados con el valor de la cadena vacía.
  - **Valores de formato texto asociados a identificadores**. Han sido rellenados con el valor `-1`.
- **Eliminar de los registros** con valor distinto de `01/01/1990` en los campos `DiscontinuedDate` y `ChemicalDateRemoved`, pues solo se precisan de aquellos registros de cosméticos que tengan productos químicos y no hayan sido retirados del mercado.
- **Agrupar por el campo** `InitialDateReported` sumando los valores del campo `ChemicalCount`. Ya que el objetivo de la aplicación de forecasting es poder obtener una predicción de la cantidad de productos químicos que serán reportados en el futuro.

Tras aplicar este preprocesamiento nos queda un dataset con 1.863 registros y 2 características: `InitialDateReported` y `ChemicalCount`.

4.5.2. Obtención de los datasets de entrenamiento y validación

Para poder realizar el forecasting, se necesita tener datos de entrenamiento y datos de validación (en adelante, `dataset` y `validation`, respectivamente). La obtención de estos datasets está ligada a que el dataset (State of California, 2019) es incremental y se tienen almacenados las siguientes dos versiones del dataset, como se ha comentado en la sección 4.2:

- Carpeta `src/data/` (Salas, 2019), donde se almacena la versión más reciente del dataset.
- Carpeta `src/data_backup/` (Salas, 2019), donde se almacena la versión anterior del dataset.

Con estas dos versiones, la obtención de los datasets `dataset` y `validation` se realiza de la siguiente manera:

- `dataset`. Contendrá todos los datos hasta los 5 meses previos al último mes de la versión `src/data_backup/`. Es decir, suponiendo que el último mes de la versión `src/data_backup/` es el 02/19 (Febrero del 2019), este dataset contendrá todos los datos antes del mes 09/18 (Septiembre del 2018): la última fecha de este dataset será el 31/08/2018 (31 de Agosto del 2018).
- `validation`. Contendrá el resto de datos. Es decir, siguiendo el mismo ejemplo, contendría los datos desde el 01/09/2018 (01 de Septiembre del 2018) hasta los datos de la última versión `src/data/`.

Así, se tiene un dataset de entrenamiento `dataset` con 1.749 registros y un dataset de validación `validation` con 114 registros.

4.5.3. Obtención de los parámetros `(p,d,q)`

Al algoritmo de forecasting ARIMA (StatsModels, 2019) hay que proporcionarle tres parámetros `(p,d,q)` para el número de parámetros AR, las diferencias y los parámetros MA que definen el modelo.

Para saber cuáles son los valores de estos tres parámetros, se ha realizado una búsqueda "en grid" con cada uno de los siguientes valores:

- `p`: desde 0 hasta 3.
- `d`: desde 0 hasta 2.
- `q`: desde 0 hasta 3.

teniendo así $4 * 3 * 4 = 48$ posibles modelos. La combinación `(p,d,q)` que dé el modelo ARIMA con menor error, será la combinación elegida para realizar el forecasting. El error se ha obtenido calculando la Error Cuadrático Medio *(Root-Mean-Square Deviation (RMSE))* definido por la ecucación 4.3, siendo n el número total de observaciones, T el conjunto de valores reales y P el conjunto de valores predichos:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(T_i - P_i)^2}{n}} \tag{4.3}$$

4.5.4. Aplicación del algoritmo ARIMA

Los datasets obtenidos en la sección 4.5.2 contienen datos agrupados por días. Sin embargo, esto puede hacer que el modelo ARIMA sea muy complejo. Por lo que, para buscar el modelo que mejor se ajuste a los datos, se va a aplicar el modelo a las siguientes agrupaciones de los datasets:

4.5.4.1. Agrupando por mes

Los datasets `dataset` y `validation` han sido agrupados por mes, de tal manera que los datasets se reducen a 110 y 6 registros, respectivamente. La Tabla 4.4 muestra la configuración del modelo ARIMA aplicado a estos datasets, así como los RMSE obtenidos tanto en el entrenamiento como en la validación. La Figura 4.13 muestra la comparativa de los valores reales con los predichos por el modelo.

(p,d,q)	$RMSE_{dataset}$	$RMSE_{validation}$
(1,1,0)	18.822	15.212

TABLA 4.4: Configuración y valores RMSE obtenidos por el modelo ARIMA agrupando por mes.

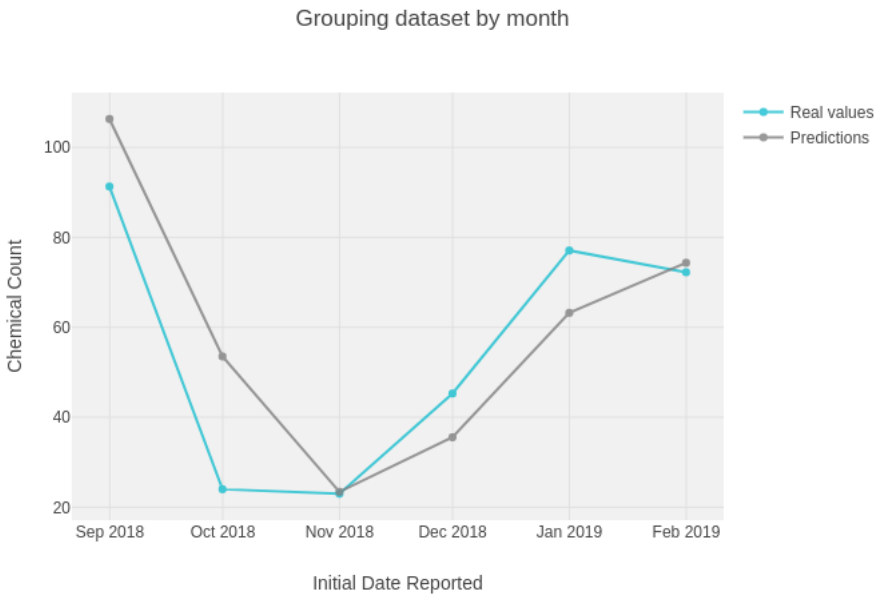

FIGURA 4.13: Comparación entre los valores reales y los predichos por el modelo ARIMA agrupando por mes.

4.5.4.2. Agrupando cada 15 días

Los datasets `dataset` y `validation` han sido agrupados cada 15 días, de tal manera que los datasets se reducen a 223 y 12 registros, respectivamente. La Tabla 4.5 muestra la configuración del modelo ARIMA aplicado a estos datasets, así como los RMSE obtenidos tanto en el entrenamiento como en la validación. La Figura 4.14 muestra la comparativa de los valores reales con los predichos por el modelo.

(p,d,q)	$RMSE_{dataset}$	$RMSE_{validation}$
(1,1,0)	15.903	19.856

TABLA 4.5: Configuración y valores RMSE obtenidos por el modelo ARIMA agrupando por cada 15 días.

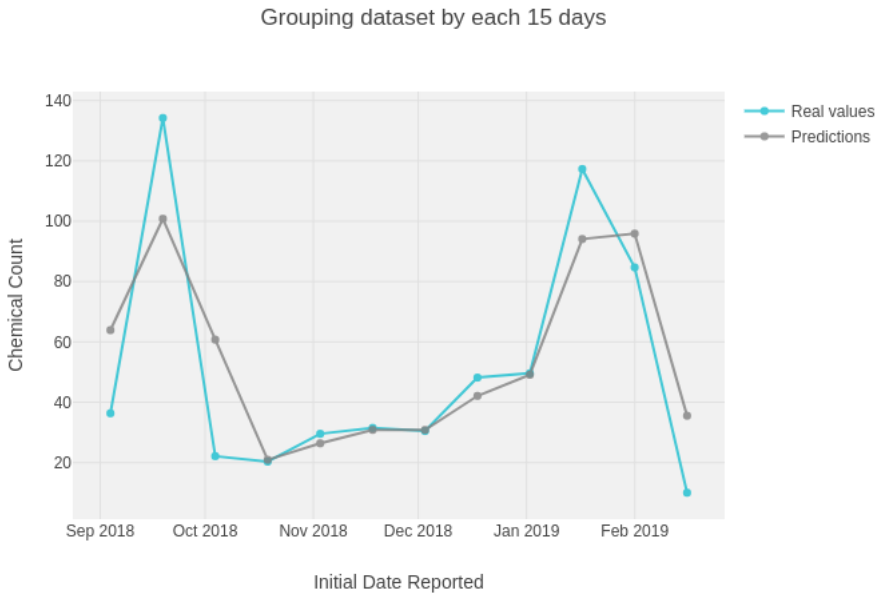

FIGURA 4.14: Comparación entre los valores reales y los predichos por el modelo ARIMA agrupando por cada 15 días.

4.5.4.3. Agrupando cada 7 días

Los datasets `dataset` y `validation` han sido agrupados cada 7 días, de tal manera que los datasets se reducen a 467 y 25 registros, respectivamente. La Tabla 4.6 muestra la configuración del modelo ARIMA aplicado a estos datasets, así como los RMSE obtenidos tanto en el entrenamiento como en la validación. La Figura 4.15 muestra la comparativa de los valores reales con los predichos por el modelo.

(p,d,q)	$RMSE_{dataset}$	$RMSE_{validation}$
(1,1,0)	15.064	17.986

TABLA 4.6: Configuración y valores RMSE obtenidos por el modelo ARIMA agrupando por cada 7 días.

4.5.4.4. Todo el dataset

Los datasets `dataset` y `validation` no han sido agrupados, de tal manera que los datasets contienen 1.749 y 114 registros, respectivamente. La Tabla 4.7 muestra la configuración del modelo ARIMA aplicado a estos datasets, así como los RMSE obtenidos tanto en el entrenamiento como en la validación. La Figura 4.16 muestra la comparativa de los valores reales con los predichos por el modelo.

(p,d,q)	$RMSE_{dataset}$	$RMSE_{validation}$
(1,1,0)	12.905	23.571

TABLA 4.7: Configuración y valores RMSE obtenidos por el modelo ARIMA sin agrupación.

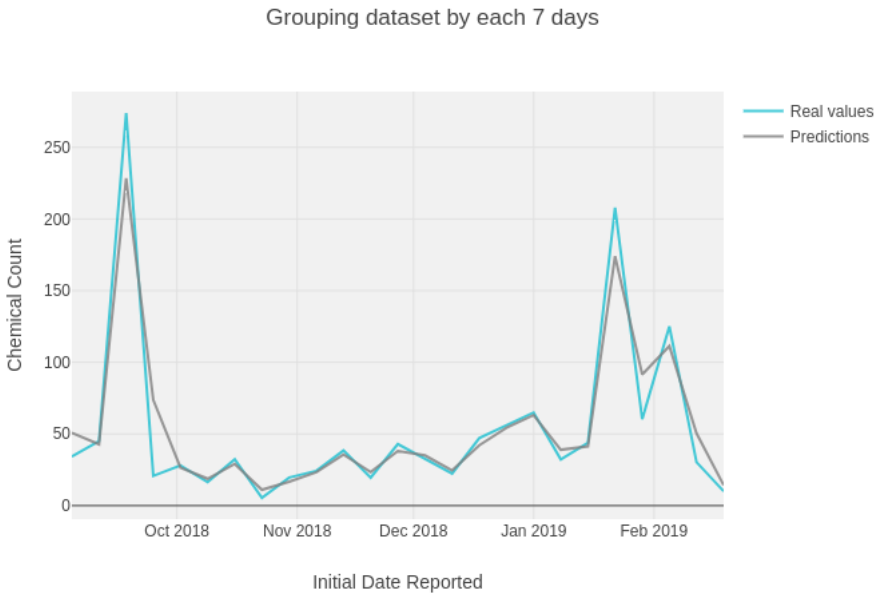

FIGURA 4.15: Comparación entre los valores reales y los predichos por el modelo ARIMA agrupando por cada 7 días.

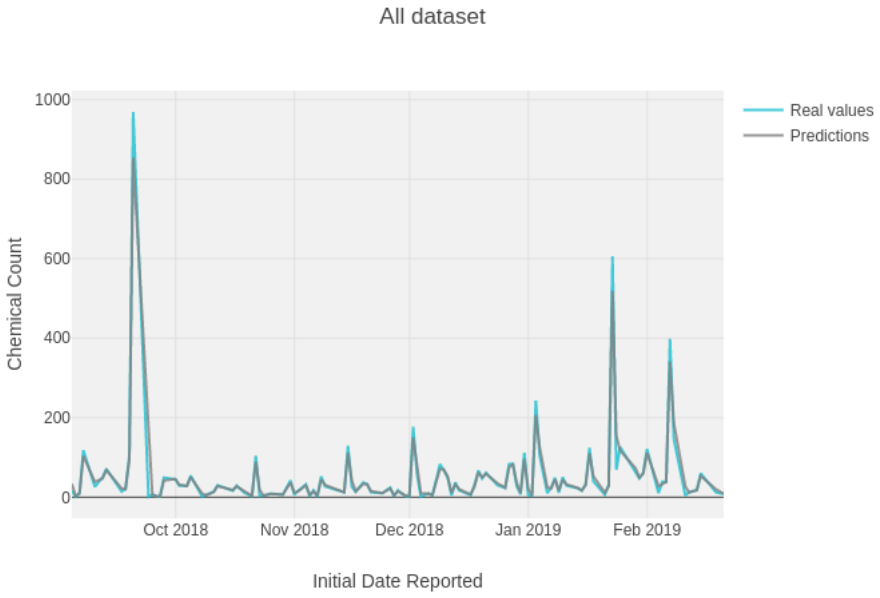

FIGURA 4.16: Comparación entre los valores reales y los predichos por el modelo ARIMA sin agrupación.

Capítulo 5

Discusión

5.1. Introducción

5.2. Clustering

5.3. Data Analysis

5.4. Forecasting

Capítulo 6

Conclusiones

6.1. Main Section 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

6.1.1. Subsection 1

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

6.1.2. Subsection 2

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

6.2. Main Section 2

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Apéndice A

Distribución de clusters

Cluster 0 | CasId - ChemicalName
433 - Methyl chloride
442 - Methyleugenol
451 - N-Methylpyrrolidone
514 - N-Nitrosodiethanolamine
516 - N-Nitrosodimethylamine
538 - Oil Orange SS
555 - Permethrin
556 - Phenacemide
557 - Phenacetin
566 - o-Phenylenediamine and its salts
570 - o-Phenylphenol
586 - Progesterone
593 - Propylene glycol mono-t-butyl ether
594 - Propylene oxide
601 - Quinoline and its strong acid salts
608 - Safrole
610 - Selenium sulfide
620 - Styrene
656 - Titanium dioxide
658 - Toluene
696 - Vinyl acetate
715 - Arsenic (inorganic arsenic compounds)
716 - Arsenic (inorganic oxides)
726 - Carbon-black extracts
730 - Chromium (hexavalent compounds)
731 - Coal tars
732 - Coffee
750 - Genistein (purified)
759 - Mineral oils, untreated and mildly treated
773 - Retinol/retinyl esters, when in daily dosages in excess of 10,000 IU, or 3,000 retinol equivalents.
776 - Silica, crystalline (airborne particles of respirable size)

TABLA A.1: Productos químicos pertenecientes al Cluster 0.

Cluster 1 | CasId – ChemicalName
871 – Titanium dioxide
872 – Titanium dioxide
888 – Coal Tar
889 – Distillates (coal tar)
890 – Coffea arabica extract
892 – Acetic acid, retinyl ester
893 – Vitamin A palmitate
894 – Talc
895 – Coal tar extract
898 – Coal tar solution
908 – Coffee bean extract
909 – Coffee extract
911 – Extract of coffee bean
927 – Retinyl acetate
933 – Retinol palmitate
938 – Retinyl palmitate
943 – Cosmetic talc
953 – Talc
955 – Talc (powder)
957 – Vitamin A
958 – Retinol
960 – Quartz
967 – Cocamide diethanolamine
968 – Cocamide diethanolamine
969 – Cocamide DEA
973 – Cocamide
987 – Lauramide DEA

TABLA A.2: Productos químicos pertenecientes al Cluster 1 (primera parte).

Cluster 1 | CasId - ChemicalName
1001 - Mica
1002 - Musk xylene
1003 - p-Aminodiphenylamine
1004 - Sodium Bromate
1005 - Titanium dioxide
1006 - Methylene glycol
1007 - Cocamide diethanolamine (DEA)
1011 - Benzophenone
1029 - Ethanol in alcoholic beverages
1032 - Titanium dioxide (airborne, unbound particles of respirable size)
1044 - Cocamide MEA
1045 - Diethanolamine
1046 - Triethanolamine
1048 - Lauramide DEA
1068 - Formaldehyde
1069 - Formaldehyde
1070 - Methylene glycol
1075 - Formaldehyde solution
1097 - Caffeine
1098 - Benzophenone-2
1099 - Benzophenone-3
1101 - Benzophenone-4
1102 - Trade Secret
1103 - Avobenzone
1104 - Carbon black
1108 - Aloe vera, whole leaf extract
1115 - Ginkgo biloba extract
1122 - N,N-Dimethyl-p-toluidine
1129 - Pulegone
1131 - Trichloroacetic acid
1147 - beta-Myrcene
1191 - Isopropyl alcohol manufacture using strong acids

TABLA A.3: Productos químicos pertenecientes al Cluster 1 (segunda parte).

Cluster 2 | CasId - ChemicalName
2 - Acetaldehyde
9 - Acrylamide
15 - All-trans retinoic acid
45 - Aspirin
61 - Benzene
71 - Benzyl chloride
74 - 2,2-Bis(bromomethyl)-1,3-propanediol
92 - Butylated hydroxyanisole
97 - Cadmium and cadmium compounds
98 - Caffeic acid
104 - Carbon black (airborne, unbound particles of respirable size)
162 - Cocamide diethanolamine
214 - Dichloroacetic acid
232 - Di-n-butyl phthalate (DBP)
270 - 1,4-Dioxane
293 - Estragole
299 - Ethyl acrylate
305 - Ethylene glycol
311 - Ethylene oxide
333 - Formaldehyde (gas)
388 - Lead
390 - Lead acetate
415 - Mercury and mercury compounds
421 - Methanol

TABLA A.4: Productos químicos pertenecientes al Cluster 2.

Bibliografía

J. M. M. Lamparero. LMA - Estudio comparativo de algoritmos de clustering para segmentación de clientes en entidades bancarias. 2018.

NIST/SEMATECH. e-Handbook of Statistical Methods, 2012. URL http://www.itl.nist.gov/div898/handbook/. (última visita 09/03/2019).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

RicardoMoya. OptimalNumClusters/Elbow_Method, 2016. URL https://github.com/RicardoMoya/OptimalNumClusters/blob/master/Elbow_Method.py. (última visita 09/03/2019).

J. M. S. Salas. Master's Thesis Repository, 2019. URL https://github.com/jmssalas/master-bigdata-businessanalytics/tree/master/09-masters-thesis. (última visita 09/03/2019).

State of California. Chemicals in cosmetics, 2019. URL https://healthdata.gov/dataset/chemicals-cosmetics. (última visita 09/03/2019).

Statistics How To. Variance Inflation Factor, 2015. URL https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/. (última visita 09/03/2019).

StatsModels. ARIMA algorithm, 2019. URL https://www.statsmodels.org/dev/generated/statsmodels.tsa.arima_model.ARIMA.html. (última visita 09/03/2019).

U.S. Department of Health & Human Services. Healthdata.gov, 2019a. URL https://healthdata.gov/. (última visita 09/03/2019).

U.S. Department of Health & Human Services. Healthdata.gov API, 2019b. URL https://healthdata.gov/api. (última visita 09/03/2019).