# Dataset : Los medicamentos al dia

*Luis Manuel Pérez Geraldino y Sergi Ramírez Mitjans*

*3 de abril de 2019*

## 1. Contexto

Recientemente han salido en las noticias informaciones que hablaban de que el ministerio de salud ha dejado de financiar medicamentos debido al mal uso que la población española estaba haciendo de ellos. Al ver dichas noticias, fue tal la necesidad de estudiar la estructura monetaria de los precios de los medicamentos que nos planteamos realizar un estudio de ello mediante el scraping de la información en la web.

A demás también teniamos conocimiento que muchas de las fundaciones que se dedican a la investigación en los hospitales dedican parte de sus recursos para investigar en la ciencia de la farmacoeconomia (ciencia que estudia el gasto de los pacientes en su medicación para poder mejorar el riesgo de mejora en calidad de vida versus gastos en esos tratamientos).

Por lo tanto al final el trabajo giraba a 3 objetivos:

- Conocer la metodologia de scraping para obtener información de internet
- Poder conocer la estructura de los medicamentos financiados por el ministerio de salud
- Generar un dataset donde contenga información de los medicamentos (tanto características como economica) para facilitarla a las fundaciones hospitalarias (en la medida que fuera posible) para realizar esos estudios farmacoeconomicos

## 2. Descripción del dataset

El conjunto de datos obtenido recoge los medicamentos con más facturación en el mundo farmacéutico. A la vez también esta incorporado los medicamentos más consumidos en España. Este dataset es un dataset variable ya que se tendrá tantos medicamentos como el que use el código demande. Esto es posible ya que el código esta preparado para hacer de buscador y extractor de aquellos medicamentos que se pasen por parámetros. Algunas variables de las que encontramos en el dataset seria el precio, nombre del medicamente, generico o url del vademecum.

## 3. Representación gráfica

### 3.1 Imagen del dataset

### 3.2 Esquema del dataset

## 4. Contenido

### 4.1 Variables del dataset

El proceso que hemos generado consta de un archivo *.csv* final llamado *result.csv*.

Dicho archivo contiene las variables obtenidas através de dos procesos de scraping de los datos. En el archivo final (*result.csv*) no contienen ningún tipo de indicador que nos permita diferenciar a partir de que proceso se ha obtenido las variables pero a continuación os detallaremos el significado de cada variable y de que proceso de scraping procede.

**Figura 1:** Imagenes de la dispensación de medicamentos en las farmacias

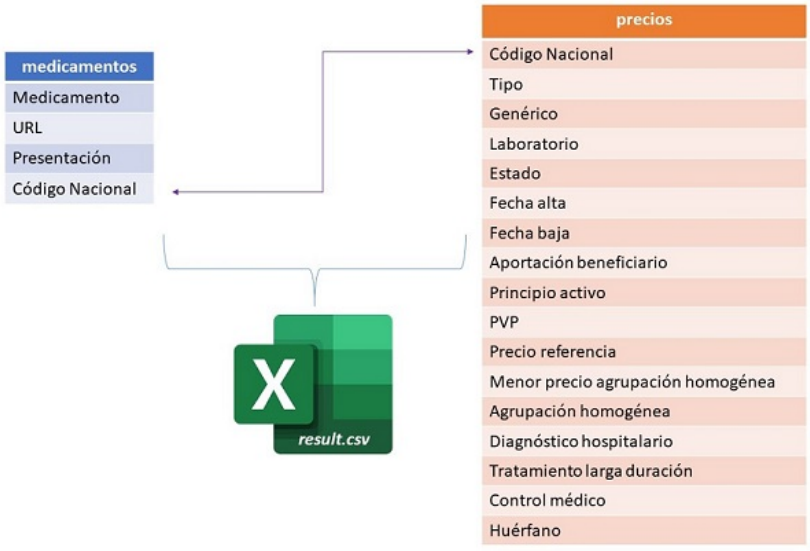


**Figura 2:** Esquema de integración de las bases de datos de medicamentos y precios para conseguir el archivo resultado (result.csv)

También podremos extraer una variable del nombre mediante la búsqueda dentro de la misma de las iniciales EFG (Equivalente Farmacéutico Genérico).

#### *4.1.1 Vademecum*

A través de la página web vademecum.es obtenemos las siguientes variables:

| | Variable | Tipo | Descripción |
|---|---|---|---|
| 1 | Medicamento | categórica | Nombre del medicamento. |
| 2 | URL | url | Dirección web donde se encuentra la información del medicamento especificado en la variable Medicamento. |
| 3 | Presentación | categórica | Tipo de envase como se facilita el medicamento. |
| 4 | Código Nacional | id | Código del registro nacional de medicamentos. Todos los medicamentos han de tener un Código Nacional |

#### *4.1.2 Sistema Nacional de Salud*

A través del formulario del sistema nacional de salud obtenemos las siguientes variables:

| | Variable | Tipo | Descripción |
|---|---|---|---|
| 1 | Tipo | categórica | Tipo de fármaco. |
| 2 | Genérico | categórica | Nombre genérico enfecto y accesorio del producto |
| 3 | Laboratorio | categórica | Código y nombre del laboratorio que elabora el medicamento |
| 4 | Estado | categórica | Estado en el cúal se encuentra el medicamento. |
| 5 | Fecha alta | fecha | Fecha en la que el medicamento se dio de alta en la base de datos del ministerio |
| 6 | Fecha baja | fecha | Fecha en la que el medicamento se dio de baja en la base de datos del ministerio |
| 7 | Aportación beneficiario | categórica | Aportacion del beneficiario. Porcentaje que ha de pagar el usuario según su clase de aportación. |
| 8 | Principio activo | categórica | Principio activo o asociación de principios activos |
| 9 | PVP | numérica | Precio de venta al público con IVA |
| 10 | Precio referencia | numérica | Precio de referencia |
| 11 | Menor precio agrupación homogénea | numérica | Menor precio de la agrupación homogénea del producto sanitario |
| 12 | Agrupación homogénea | categórica | Código y nombre de la agrupación homogénea del producto sanitario |
| 13 | Diagnóstico hospitalario | binaria | Nos indica si es necesario un diagnóstico hospitalaria o no para obtener el producto |
| 14 | Tratamiento larga duración | binaria | Nos indica si es un tratamiento de larga duración o no |
| 15 | Control médico | binaria | Nos indica si se necesita un control medico o no |
| 16 | Huérfano | binaria | Nos indica si el medicamento es un medicamento huerfano (productos medicinales destinados al diagnóstico, prevención o tratamiento de enfermedades que ponen en riesgo la vida, o muy graves o enfermedades que son raras) o no |

### 4.2 Periodo de tiempo de los datos

La web del vademecum recoge todos los medicamentos disponibles en la red de farmacias y medicamentos que hay en el mundo. Esta información la recogen de las industrias farmacéuticas encargadas de generar los medicamentos. Por lo tanto en dicha *web* no existe una ventana temporal destacable.

Por la contra, la página web del buscador del ministerio si que incorpora información monetaria de los medicamentos que estan financiados por el propio ministerio. Esta base de datos esta actualizada a los nuevos cambios que se han originado en el mes de Marzo. Por lo tanto la base de datos esta actualizada para Abril de 2019.

### 4.3 Proceso de colección de los datos

Para llegar al dataset que se puede encontrar en la carpeta result, se ha necesitado seguir un proceso de colección de los datos que acontinuación detallaré.

Primeramente realizamos un estudio del archivo robots.txt para comprobar si nuestro user-agent esta permitido en la web para la descarga. Si no fuera así, el proceso incorpora un user-agent aleatorio para que podamos seguir con el proceso de descarga.

Después de la comprobación del robot.txt iniciamos la búsqueda del medicamento o medicamentos que hemos pasado por parámetro y lo buscamos en la web del vademecum. Una vez encontrado nos guardamos la url del medicamento, el código nacional, el nombre del medicamento y en que envase se distribuye el medicamento.

Una vez tengamos el código nacional del medicamento, vamos a la página oficial del ministerio de salud donde encontramos una base de datos de los precios subvencionados de los medicamentos. A continuación incorporamos en el formulario dicho código y obtenemos una serie de variables a parte del precio (la lista de las variables se encuentra en el apartado 4.1.2 de este documento).

## 5. Agradecimientos

Los datos han sido extraidos de las web vademecum y del ministerio de sanidad, política social e igualdad de España. Estos datos se extraen a partir de técnicas de Web Scraping que nos permite descargar información de páginas web sin la necesidad de ir página a página copiando la información ya que seria una alta carga de tiempo para el usuario. Por lo tanto realizando el Web Scraping nos “curamos en salud” y convertimos las horas de tiempo de persona en horas de tiempo de máquina. Toda la programación se ha realizado mediante Python.

## 6. Inspiración

Los datos obtenidos son unos datos muy especificos que sólo podriamos usarlos en el mundo de la sanidad y la economia. Gracias a los datos obtenidos podriamos llegar a realizar un estudio del gasto que genera cada paciente en los tratamientos que necesita para curar una enfermedad (farmacoeconomia). Para ello una vez obtenida la tabla de datos donde tuvieramos características de los medicamentos, precio de los medicamentos e información de los pacientes podriamos ajustar estos tratamientos mediante modelos de Machine Learning dependiendo de la renta de cada paciente.

Estos estudios nos permitirian trabajar tanto para hospitales como para instituciones públicas a realizar campañas e incluso inversiones en según que tipo de medicamentos que su coste es mas alto del que se pueda esperar.

## 7. Licencia

Barajando las múltiples posibilidades de licencia a escoger, para nuestra publicación de datos hemos escogido la **licencia CC BY-SA 4.0 License**. Esta licencia se define por las siguientes características:

- Primero de todo deben incluir los nombres de los creadores. Utilizando esta clásula nos permetiria que se nos identificara como creadores originales de la base de datos y por lo tanto deberian de notificar cualquier cambio realizado en la base de datos, que el autor original somos nosotros.
- Las contribuciones que puedan realizar a nuestra base de datos tendran la misma licencia que la que actualmente estamos definiendo. Esta cláusula es así debido a ue nosotros queremos que no sólo nuestra autoria intelectual se hereden en futuros desarrollos de nuestra base de datos sino también las condiciones de distribucion se debera regir bajo la misma licencia.Con esta decision, nosotros comprometemos a que la base de datos se podra seguir usando sin restricciones y además veneficiamos al Open Data (Datos abiertos).

Pareceria raro que si nuestro planteamiento inicial era para fomentar el estudio farmacoeconomico de hospitales al final acabamos diciendo que también sirve para comercializar con ella. Es verdad que nuestro produto no sólo va destinado a la investigación de costes de tratamientos sino también permetiria a las empresas ver precios de la competencia, necesidades, tipos de medicamentos que hay en el mercado e incluso la falta de medicamentos o poca competencia de algun sector del genérico.

- Se permite un uso comercial. Como hemos decidido que se nos permite el uso comercial, se nos podria abrir proyectos en empresas donde podrian hacer uso de los datos generados para que pudiesen tanto vender los medicamentos como por ejemplo hacer estudios de viabilidad. Al final también podriamos incluir una licencia de acceso a los datos mediante un web service para gestionar licencias de mantenimiento de la base de datos.

Para acabar decir que una forma de escoger el tipo de licencia creative commons seria hacer uso de su propia web. En ella podemos encontrar un formulario que te permite identificar la licencia que más de asjusta a los parámetros que queremos. El enlace se encuentra en la siguiente ruta https://creativecommons.org/choose/.

## 8. Tabla de contribuciones

A continuación se muestra la tabla de contribución de dicho proyecto.

| **Contribuciones** | **Firmas** |
|---|---|
| Investigación previa | LMPG, SRM |
| Redacción de las respuestas | LMPG, SRM |
| Desarrollo código | LMPG, SRM |

## 9. Futuras lineas de trabajo

En cuanto a linias futuras de trabajo podemos explicar 3:

- La primera trataria de perfilar la forma de automatizar los posibles baneos que nos pudieran ocurrir. Como en las dos web que consultamos no tenemos restricciones de descargas, no podemos desarrollar unos métodos automáticos de detección de baneos pero si entendemos que en posibles descarga de datos más complejas pudieramos necesitar de estos métodos.
- Desarrollar más el scraping a realizar en la web de vademecum. Dicha web es muy compleja e incorpora mucha información que en este trabajo no hemos podido abarcar.
- Aprender de otras herramientas de Web Scraping como podria ser el Selenium.

## 10. Código

El código del proyecto se encuentra en el repositorio libre de Github. Enlace.

## 11. Bibliografia


**Webs:**

1. Vademecum. Enlace
2. Medicamentos genéricos. Ministerio de sanidad, política social e igualdad de España. Enlace
3. Nomenclátor de Facturación de Abril - 2019. Enlace

**Libros:**

4. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data
5. Masip, D. El lenguaje Python. Editorial UOC.
6. Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.
7. Munzert, S., Rubba, C., Meiner, P., Nyhuis, D. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
8. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

**Tutoriales:**

9. Tutorial de Github

**Prensa Web:**

10. El consumo de medicamentos genéricos cae en Castilla y León por debajo de cifras de hace cinco años (6 de Abril de 2019). El norte de Castilla. Enlace
11. Martín, P. Sanidad frena el incremento de precio del Fortasec (22 de Marzo de 2019). El Periodico. Enlace
12. Moreno, J. Los 10 medicamentos más vendidos en España (25 de Junio de 2019). Enlace
13. Nafria, I. Los 15 medicamentos genéricos más consumidos en España (13 de Mayo de 2015). Enlace
14. Rivera, M y Negrete, B. Comisión de precios: 'OK' a financiar 12 fármacos, dos de ellos innovadores (26 de Marzo de 2019). Redaccion Médica. Enlace
15. Sanidad veta por primera vez la subida de precios de fármacos no financiados (22 de Marzo de 2019). La Vanguardia. Enlace
16. Vigario, A. Nolotil, Enantyum y Adiro, los medicamentos más vendidos en las farmacias españolas (18 de Marzo de 2019). Enlace