



MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS  
(DATA SCIENCE)

# PRÁCTICA 1: WEB SCRAPING BOLSA DE MADRID

*Web Scraping para la extracción de los valores  
históricos de la bolsa e información sobre los activos*

*M2.851 - Tipología y ciclo de vida de los datos*

Autores:

Yosry Elsayed

Laura Pastor

Profesor/a:

Mireia Calvo González



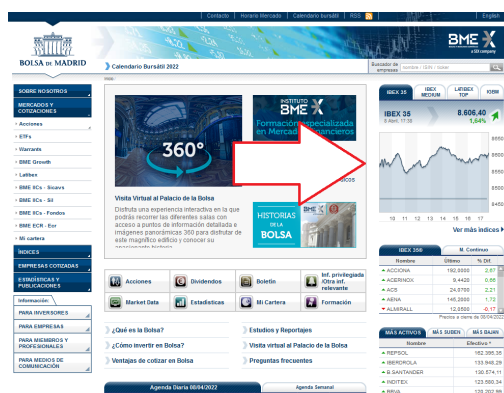
# Índice

<b>1. Introducción</b>	<b>1</b>
<b>2. Datasets</b>	<b>3</b>
2.1. DIM_Miembro.CSV . . . . .	3
2.2. FACT_IBEX35.CSV . . . . .	3
2.3. Carpeta: miembros . . . . .	4
<b>3. Metodología</b>	<b>5</b>
3.1. Preparación Web Scraping . . . . .	5
3.2. Código de Web Scraping . . . . .	6
3.3. Visualización . . . . .	9
<b>4. Aspectos claves a mejorar</b>	<b>11</b>
<b>5. Contribución al trabajo</b>	<b>11</b>

## 1. Introducción

El mercado financiero está en cambio constante, por lo que resulta complicado intentar analizarlo a tiempo real extrayendo el dato mediante la descarga de ficheros tipo Excel de manera continua. Por ello, es interesante extraer los diferentes valores del mercado mediante técnicas de web scraping para así conseguir crear un histórico a tiempo real que pueda alimentar un modelo predictivo cuyo objetivo es optimizar la cantidad de dinero a invertir y cuando, y visualizar dichos valores.

Los gráficos de la propia web no tienen cierta flexibilidad u opciones para modificar sus ejes y parámetros (como podemos observar en la siguiente imagen), algo que sí que podemos manipular y diseñar según nuestras necesidades en Power BI.



**Figura 1.** Gráfico ejemplo de la página web de Bolsa de Madrid

Resumidamente, en esta práctica solo nos centraremos en una extracción puntual del histórico de los diferentes valores financieros de las empresas que componen el índice bursátil IBEX35 desde la página web oficial de la Bolsa de Madrid, y presentarlos en un report de visualizaciones usando Power BI.

La Bolsa de Madrid es el principal mercado de valores de España, cuyo propietario es el BME que es el operador de todos los mercados de valores y sistemas financieros en España. El presidente actual del BME es Johannes Bernardus Dijsselhof.

El titular y prestador de la página web en cuestión se presenta como "Sociedad Rectora de la Bolsa de Valores de Madrid, S.A. Sociedad Unipersonal", como se puede ver en el apartado de "Aviso Legal".

En el mismo apartado de "Aviso Legal" de la página web nos encontramos con el subapartado "Uso de la Página Web" donde se señala que los contenidos de la página pueden ser descargados, copiados e impresos para uso personal y no comercial. Por lo que es legal la extracción y análisis de los diferentes valores de la página usando técnicas de Web Scraping e usarlos con fines estudiantiles.

Por el otro lado, El IBEX 35 es el principal índice bursátil de la bolsa española elaborado por BME. Está formado por un total de 35 empresas las cual extraeremos un histórico sus valores de precio de cierre, e información descriptiva propia de cada empresa (por ejemplo, dirección, teléfono, logo, etc.).

## 2. Datasets

La información extraída se dividirá en dos tablas (ficheros CSV que tienen como separador la comma ',') y una carpeta que contendrá el logo de cada empresa (formato gif y jpg).

### 2.1. DIM\_Miembro.CSV

Este fichero contiene información descriptiva de los diferentes miembros/empresas que forman parte del IBEX35. Es una tabla de 35 observaciones con los siguientes campos:

- **Id:** Identificador o clave primaria de cada uno de los miembros.
- **miembro:** El nombre de la empresa.
- **dirección:** La dirección de la central de la empresa.
- **postal:** El código postal junto a la ciudad donde se encuentra la central.
- **teléfono:** Teléfono de contacto.
- **logo:** La carpeta y el nombre del logo de cada empresa.

Si pensamos en la estructura de un modelo de datos, esta tabla se considera como una dimensión dado que contiene datos descriptivos de cada empresa y no incluye ninguna medida numérica a analizar.

### 2.2. FACT\_IBEX35.CSV

Este fichero contiene un histórico los valores financieros diarios de cada miembros/empresas que forman parte del IBEX35 por día. Es una tabla de 805 observaciones con los siguientes campos:

- **Id:** Identificador o clave primaria de cada observación.
- **fecha:** La fecha del registro/observación.
- **cierre:** El precio de cierre de cada empresa para cada día.
- **Referencia:** Código identificativo de un valor en el sistema de representación de Bolsa de Madrid.
- **volumen:** El número de títulos o contratos que se negocian de un activo financiero en un mercado.
- **Efectivo:** El efectivo negociado de una operación es el resultado de multiplicar el número de acciones negociado por el precio de cada acción.
- **último:** Último precio de acción, que coincide con el precio de cierre.

- **máximo:** Máximo precio alcanzado.
- **mínimo:** Mínimo precio alcanzado.
- **medio:** Precio medio de acción a lo largo del día.
- **activo:** Nombre de la empresa/miembro del IBEX35.

Si pensamos en la estructura de un modelo de datos, esta tabla se considera como una tabla de hechos dado que contiene las medidas numéricas que nos interesa analizar y no incluye información descriptiva en exceso.

Por lo tanto, nuestro modelo de datos se construirá por una tabla de hechos y una dimensión.

### 2.3. Carpeta: miembros

Se trata de una carpeta que incluye iconos o logos de cada empresa cuyos formatos varían entre gif y jpg. Se puede saber el nombre del icono correspondiente a cada empresa mediante la columna 'logo' en el fichero 'DIM\_Miembro.CSV'.

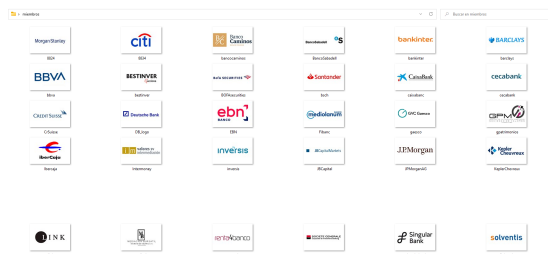


Figura 2. Los logos de los miembros del IBEX35

Nuestros datasets tendría una licencia ODbL. Una licencia Open Database License permite a los usuarios compartir, modificar y usar libremente las bases de datos en cuestión, sin temor a los derechos de autor o cuestiones de propiedad del dato. Este tipo de licencia encaja con nuestros datasets dado que se tratan de datos públicos informativos sobre los cual no tenemos ninguna propiedad, y se pueden usar libremente por cualquier persona ya que nuestro repositorio será público.

Los ficheros CSV se encuentran tanto en el repositorio GitHub (el repositorio es Yoyazoooo20/DBT-Training) como en Zenodo (el DOI de Zenodo es 10.5281/zenodo.6428376).

## 3. Metodología

### 3.1. Preparación Web Scraping

Antes de pasar a los pasos realizados para realizar Web Scraping a la página de la Bolsa de Madrid realizamos una evaluación inicial de la web.

En primer lugar buscamos el archivo de *robots.txt* el cual no existe para este dominio lo cual indica que esta permitido rastrear la página web en su totalidad.

En segundo lugar analizamos el código fuente para identificar los elementos que queríamos procesar con nuestro programa. Tal y como se muestra en la Figura 3



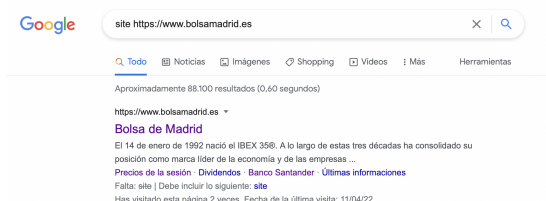
```

<DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html>
<head>
<meta charset="utf-8" />
<title>Bolsa de Madrid</title>
<meta name="description" content="Bolsa de Madrid. Noticias, cotizaciones, precios de acciones, divisas, futuros, opciones, etc." />
<meta name="keywords" content="Bolsa de Madrid, cotizaciones, precios, acciones, divisas, futuros, opciones" />
</head>
<body>
<div id="header">
<div id="logo">
<img alt="Logo of Bolsa de Madrid" data-bbox="331 335 450 350"/>
</div>
<div id="menu">
<ul>
<li>Inicio</li>
<li>Noticias</li>
<li>Cotizaciones</li>
<li>Precios</li>
<li>Divisas</li>
<li>Futuros</li>
<li>Opciones</li>
<li>Más</li>
</ul>
</div>
</div>
<div id="content">
<div id="main">
<div id="title">
<h2>Bolsa de Madrid</h2>
</div>
<div id="text">
<p>El 14 de enero de 1992 nació el IBEX 35®. A lo largo de estas tres décadas ha consolidado su posición como marca líder de la economía y de las empresas ...</p>
</div>
</div>
<div id="sidebar">
<div id="title">
<h3>Precios de la sesión</h3>
</div>
<div id="text">
<p>Dividendos · Banco Santander · Últimas informaciones</p>
</div>
</div>
</div>
<div id="footer">
<div id="text">
<p>Falta: site | Debe incluir lo siguiente: site</p>
<p>Has visitado esta página 2 veces. Fecha de la última visita: 11/04/22.</p>
</div>
</div>
</body>
</html>

```

**Figura 3.** Información en formato HTML de la página de Bolsa de Madrid

En tercer lugar analizamos el tamaño para ello escribimos en el buscador de Google lo siguiente "site <https://www.bolsamadrid.es>". Como resultado vimos que este dominio tiene unos 88,100 enlaces como se demuestra en la Figura 4.



**Figura 4.** Número de enlaces relacionados con la web Bolsa de Madrid

En cuarto lugar analizamos la tecnología del sitio, para ello utilizamos la librería de Python *"builtwith"*. El sitio utiliza la librería de JavaScript *"jQuery"* para facilitar el desarrollo de su web.

Y finalmente buscamos al propietario. Para ello utilizamos la librería de Python *"whois"* que nos da información acerca del mismo. En este caso el propietario no ha dado información acerca de si mismo en el código de la página, pero de todas formas sabemos quien es ya que ellos lo indican en la propia página tal y como comentamos en el Apartado 1.



```

In [2]: import builtwith
In [3]: builtwith.builtwith("https://www.bolsamadrid.es")
Out[3]: {'javascript-frameworks': ['jQuery']}
In [4]:

```

**Figura 5.** Análisis de la tecnología de la Bolsa de Madrid

```

In [7]: import whois
In [8]: whois.whois("https://www.bolsamadrid.es")
Out[8]:
{'domain_name': None,
 'registrar': None,
 'whois_server': None,
 'referral_url': None,
 'updated_date': None,
 'creation_date': None,
 'expiration_date': None,
 'name_servers': None,
 'status': None,
 'emails': None,
 'dnssec': None,
 'name': None,
 'org': None,
 'address': None,
 'city': None,
 'state': None,
 'zipcode': None,
 'country': None}

```

**Figura 6.** Información del propietario de la página <https://www.bolsamadrid.es>

### 3.2. Código de Web Scraping

Presentaremos como funciona el algoritmo, creado en Python, y su navegación en la página web para extraer los datos correspondientes a cada miembro que forma parte del IBEX35.

1. Entra en la página web <https://www.bolsamadrid.es> y extrae todos sus datos en formato HTML. Para ello utilizamos las librerías de *"request"* y *"BeautifulSoup"*. Cabe mencionar que a la hora de realizar la petición realizamos un cambio del *user agent*. Tanto en la utilización de las librerías como en el cambio del *user agent* estamos aplicando técnicas de buenas prácticas para la realización del Web Scraping. El resto de peticiones que realicemos en el resto de la práctica se van a realizar en las mismas condiciones que esta misma.



**Figura 7.** Interfaz de la página web de Bolsa de Madrid

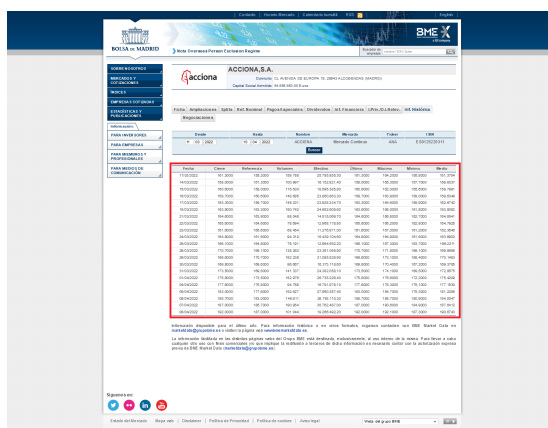
2. Extrae toda la información que contiene la tabla señalada, mediante el id

de la tabla que se encuentra en el HTML (Vease Figura 3).



**Figura 8.** La tabla de resumen del índice bursátil del IBEX35 que nos interesa para llegar al detalle de cada empresa que lo forma

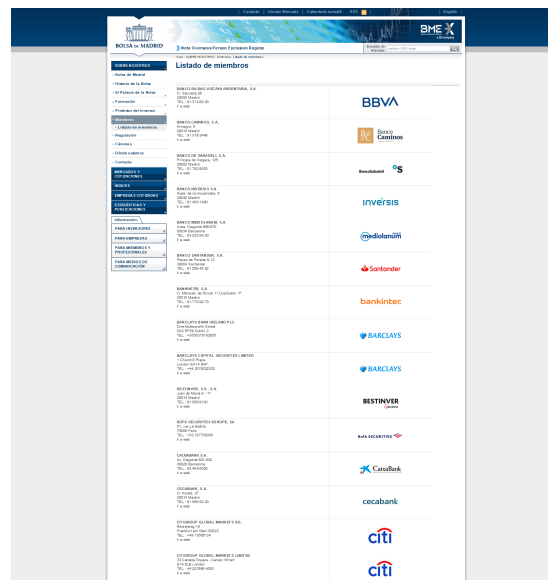
3. Guarda en una lista el nombre del activo y el enlace a su página web propia dentro de la bolsa de Madrid.
4. Se vuelve a realizar web scraping en cada una de los enlaces anteriormente guardados.
5. Las peticiones de cada uno de los activos se extrae la información histórica de los valores del último mes y se añade a una lista.



**Figura 9.** Histórico del último mes del miembro ACCIONA

6. Convertimos la lista en un conjunto de datos el que posteriormente guardamos en un archivo .csv.

Dentro de este código también se navega a través de esta web para acceder a la información de los miembros de la bolsa de Madrid a los que se le extrae la información y sus logos.



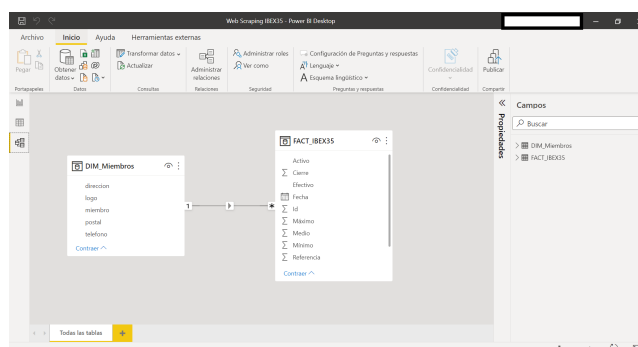
**Figura 10.** Listado de miembros disponible en la web de Bolsa de Madrid

1. Realizamos la petición para acceder al enlace de los miembros de la bolsa de Madrid.
2. Extraemos los enlaces de todas las imágenes que aparecen en la web.
3. Seleccionamos las imágenes que se corresponden con los logos. Y nos guardamos sus enlaces.
4. Mediante un bucle realizamos unas nuevas peticiones tomando ahora como enlace el de cada una de las imágenes. Y guardamos la respuesta de la petición en un archivo con el formato de la imagen. Las ubicamos en una carpeta creada anteriormente denominada "Miembros".
5. Obtenemos los datos de los miembros que se encuentran en una tabla ( A partir de la primera petición de esta parte del código). Y guardamos los datos extraídos y los directorios de los logos en un archivo con formato *.csv*.

### 3.3. Visualización

Para visualizar y analizar los datos se usará la herramienta Power BI. Se trata de un servicio de análisis de datos de Microsoft orientado a proporcionar visualizaciones interactivas, la cual se puede conectar a cualquier base de datos, página web o API. En nuestro caso Power BI se conectará directamente con los archivos CSV que guardaremos en un directorio GitHub público, así mismo cualquier persona que descargue el fichero de visualizaciones '.pbix' puede interactuar con los datos sin la necesidad de tener los ficheros CSV.

Una vez conectados ambos ficheros CSV a Power BI, construimos el siguiente modelo de datos:

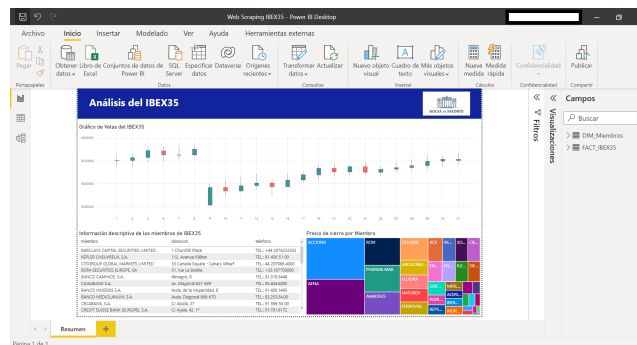


**Figura 11.** Modelo de los datos

Se tiene una relación 1:N entre la tabla dimensión y la tabla de hechos. Es decir, cada registro de la tabla de hechos está relacionada únicamente con un registro de la tabla de dimensión, pero cada registro de la tabla de dimensión puede estar relacionada con más de un registro de la tabla de hechos.

Las visualización que nos encontramos en el informe son:

1. Un gráfico clásico de velas del índice bursátil del IBEX35 (se usan las distintas métricas de cierre, medio, mínimo y máximo de precio por activo).
2. Una tabla de información descriptiva de cada miembro del IBEX35.
3. un gráficos de rectángulos de los miembros del IBEX35, donde el tamaño de cada rectángulo/empresa va relacionado con su precio de cierre. Es decir, cuanto más grande es el rectángulo de una empresa, mayor es su precio de cierre a comparación del resto de empresas.



**Figura 12.** Report con las distintas visualizaciones descritas anteriormente

Como se puede observar, a partir de los datos extraídos se pueden personalizar visualizaciones interactivas dependiendo de nuestras necesidades. Otra ventaja con Power BI es que pueden publicar los reports de manera que sean fáciles de visualizar, analizar y compartir en distintos dispositivos (tablet, móvil, etc). A partir de estas visualizaciones podemos ver la tendencia del IBEX35, por lo que si se observa una tendencia alcista podemos comprar acciones suponiendo que esa tendencia se mantendrá. También podemos ver las empresas con mayor precio de cierre de acciones dentro del IBEX35 en todo momento (se puede aplicar filtros por día, mes y/o año). Por último, también se pueden consultar información propia de cada empresa.

## 4. Aspectos claves a mejorar

Una de las mejoras que se podrían aplicar en este trabajo es aplicar el proceso de Web Scraping directamente en la interfaz de Power BI, así evitamos la necesidad de extraer el dato en un archivo CSV, y guardarlo en un repositorio GitHub para que posteriormente sea ingestado dentro de visualizaciones. Power BI actualmente tiene la posibilidad de incorporar código de Python o R, pero todavía esta en fase de mejora por lo que todavía no es óptimo ni fácil de usar.

## 5. Contribución al trabajo

Una tabla donde se muestra que cada uno de los autores/integrantes han participado en cada uno de los apartados del trabajo presentado.

Contribuciones	Firma
Investigación previa	Laura y Yosry
Redacción de las respuestas	Laura y Yosry
Desarrollo del código	Laura y Yosry