



PRA 1

John Mauricio Olivos Dominguez

Máster universitario en Ciencia de Datos

Tipología y ciclo de vida de los datos

Abril de 2023

Tabla de Contenido

1.	Contexto.	3
2.	Título.....	3
3.	Descripción del dataset.....	3
4.	Representación gráfica.	3
5.	Contenido.	4
6.	Propietario.	5
7.	Inspiración.	6
8.	Licencia.	6
9.	Código.	6
10.	Dataset.....	7
11.	Vídeo.....	7
12.	Bibliografía	8

1. Contexto.

Para la mayoría de las personas la compra de vivienda es un hito de gran importancia para sus vidas, razón por la cual deben poner en consideración varios factores al momento de realizar la compra, teniendo en cuenta que actualmente la oferta de vivienda es muy variada muchas veces la evaluación de dichos factores se hace exhaustiva al no existir herramientas que faciliten una evaluación rápida. En este ejercicio se pretende extraer un dataset a partir de la información extraída del sitio web <https://fincaraiz.com.co/>, este es un sitio en el cual diariamente se publican anuncios para la venta y/o arriendo de inmuebles en Colombia.

2. Título.

“Precios de referencia compraventa inmuebles”

3. Descripción del dataset.

El dataset resultante de este ejercicio contiene información detallada acerca de los inmuebles que a la fecha de ejecución del código (20/04/2023) se encontraban publicados para venta en la página web <https://fincaraiz.com.co/>. Este dataset permite consultar atributos de diversos inmuebles que se encuentran en venta (para este ejercicio inmuebles ubicados en la localidad de Engativá, Bogotá, Colombia), de igual manera los datos allí registrados permiten hacer diversas comparativas para determinar precio por metro cuadrado, precio del inmueble, entre otras.

4. Representación gráfica.

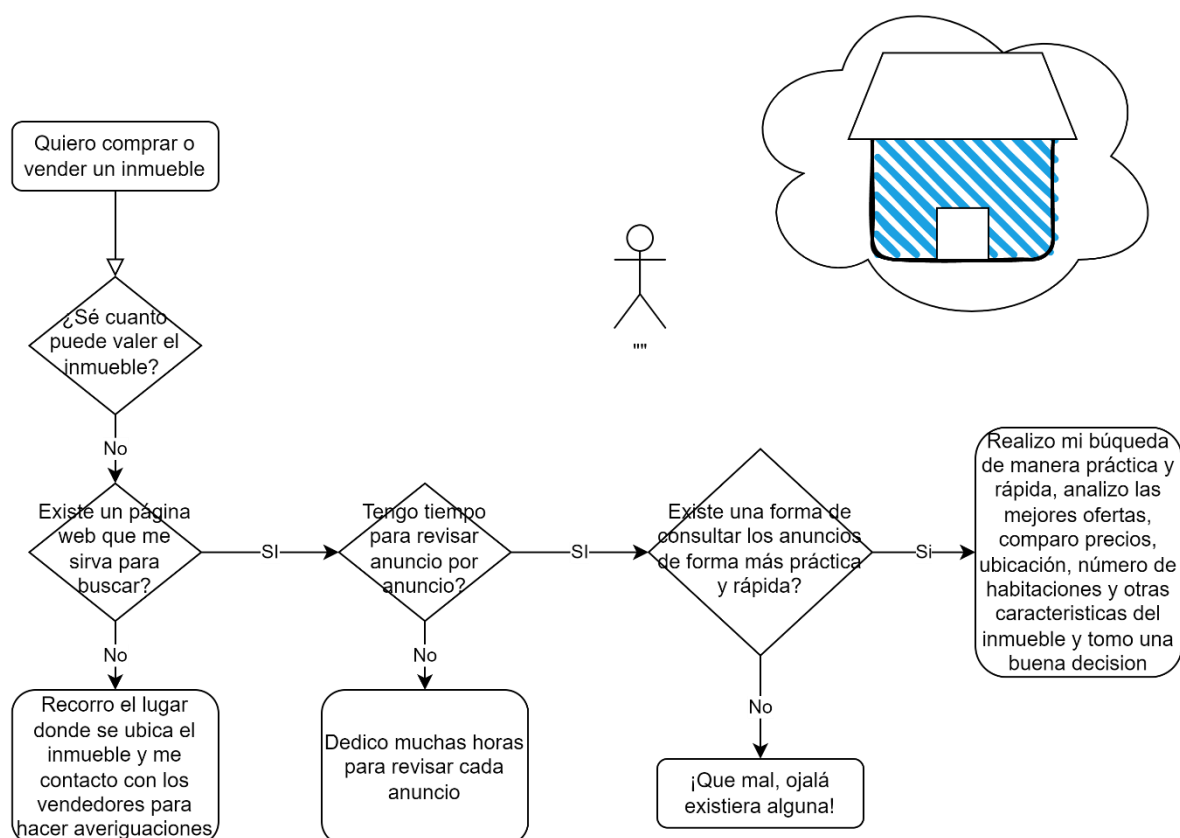


Ilustración 1 Diagrama Problema Solución

5. Contenido.

Tabla 1 Campos Dataset Precios de referencia compraventa inmuebles

NOMBRE DEL CAMPO	TIPO DE DATO	DESCRIPCIÓN
TITULO	str	Título del anuncio, facilita la identificación del tipo de inmueble
CODIGO INMUEBLE	str	Identificador asignado por la página al inmueble
ÁREA PRIVADA	str	Área privada del inmueble
ÁREA CONSTRUÍDA	float	Área construida del inmueble
PRECIO POR M ²	int	Valor por cada m ² construido
ESTRATO	str	Estrato social de la zona en la que se encuentra ubicado el inmueble
ANTIGUEDAD	str	Tiempo transcurrido desde el año de construcción del inmueble
HABITACIONES	int	Número de habitaciones construidas en el inmueble
BAÑOS	int	Número de baños construidos en el inmueble
PARQUEADEROS	int	Número de zonas de parqueo con las que cuenta el inmueble
BARRIO	str	Barrio en el cual se encuentra ubicado el inmueble
DIRECCIÓN	str	Dirección de ubicación del inmueble
PRECIO	int	Precio total del inmueble

La información contenida en este Dataset corresponde a la que se encontraba publicada en la página web el día 20 de abril de 2023.

6. Propietario.

La página web www.fincaraiz.com.co como los datos presentes en este dataset son propiedad de EDITORA URBANA LTDA, identificada con NIT 800.229.663-1 y con domicilio principal en la ciudad de Bogotá, Colombia.

La idea de realizar este trabajo surge de una necesidad propia, sin embargo, al investigar acerca de estudios o análisis similares que se hayan realizado con anterioridad, se encuentra que según Londoño *“Actualmente no existen conjuntos de datos públicos que permitan a un investigador monitorear el estado y la evolución del mercado de vivienda en Colombia. La falta de una estructura de información organizada limita la capacidad de entender y estudiar sus dinámicas¹”*. Razón por la cual se desarrollado algunos trabajos e investigaciones acerca de este tema, entre las cuales se encuentran:

- **PRECIO DE LA VIVIENDA EN COLOMBIA** GitHub - CamiloAguilar/HousePricing: The main objective of this project is to predict the price of housing in Colombia using different models and computational techniques framed in Machine Learning El objetivo central del presente proyecto consiste en predecir el precio de la vivienda en Colombia haciendo uso de diferentes modelos y técnicas computacionales enmarcadas en el Machine Learning. El suministro de información necesario para aplicar estas técnicas, resulta de la actual disponibilidad de consulta en decenas de aplicaciones web que ofrecen servicios centrados en brindar espacios adecuados en la red, para que cualquier persona o empresa pueda comprar o vender cualquier tipo de inmueble, presentando características específicas de cada uno, tales como la ciudad, el tipo de vivienda, barrio, estrato socioeconómico, número de habitaciones, entre muchas otras².

Adicionalmente se han desarrollado alguna herramientas que permiten hacer webScraping a personas que no necesariamente tienen conocimiento acerca de cómo hacerlo, un ejemplo es la plataforma ofrecida por Octoparse, la cual se puede consultar a través del siguiente enlace: [Extraer datos de inmuebles de Fincaraiz – Web Scraping Herramienta | Octoparse](#)

Por último, para dar cumplimiento a los principios éticos y legales, además de las condiciones de uso establecidas por el propietario de los datos, se toman las siguientes precauciones:

- El propietario de los datos en la página web de Términos y condiciones (https://fincaraiz.com.co/informacion#terminos_y_condiciones) , numeral 7.4, literal a, dicta lo siguiente: *“No es permitido realizar Crawling y scraping de las páginas web de fincaraiz.com.co y todos sus subdominios, del portal web y de las aplicaciones (Android e IOS) por medio de robots o spiders que no estén explícitamente autorizados en el archivo www.fincaraiz.com.co/robots.txt”*. Se consulta la página robots referenciada pero no se encuentra información alguna.
- El presente trabajo genera un dataset con fines netamente académicos por lo cual no será utilizado con fines comerciales, ni se hará captura de datos de carácter privado o sensible.

¹ Tomado de: Londoño Botero, J. (2021). Construcción de un conjunto de microdatos del mercado de vivienda en Colombia a partir de información de anuncios web (Publicación n.º 10784/29875) [Trabajo de grado, Universidad EAFIT]. repository.eafit.edu.co/bitstream/handle/10784/29875/Jose_LondonoBotero_2021.pdf?sequence=2&isAllowed=y

² Tomado de: GitHub - CamiloAguilar/HousePricing: The main objective of this project is to predict the price of housing in Colombia using different models and computational techniques framed in Machine Learning

7. Inspiración.

Durante los últimos años el mercado de bienes raíces ha crecido de manera significativa. Se decide realizar webscraping sobre este sitio teniendo en cuenta que el mercado de compraventa de bienes raíces es un mercado que diariamente presenta una gran dinámica, motivo por el cual en ocasiones se hace complicado evaluar rápidamente aquellas opciones que son más favorables para el comprador. La generación de este dataset está pensado para facilitar las siguientes actividades:

- Identificación rápida por parte del comprador de aquellos inmuebles que se ajusten a su presupuesto, localización.
- Identificación rápida de inmuebles que presenten valores favorables por metro cuadrado construido.
- Identificación de rango de precios de venta para quien requiera vender su inmueble.

La idea de creación de este dataset como se mencionó en el punto anterior surge de una necesidad personal, sin embargo, se evidencian estudios anteriores en los cuales la data obtenida en el presente trabajo podría contribuir a complementar investigaciones en curso o futuras acerca de dinámicas comerciales y avalúo de predios en Colombia.

8. Licencia.

Teniendo en cuenta las restricciones indicadas por el propietario de los datos y a la información que puede aportar este dataset a otras investigaciones, la licencia adecuada para el dataset resultante corresponde a la CC BY-NC-SA (Atribución-NoComercial-CompartirIgual). Esta licencia permite a otros remezclar, adaptar y construir sobre su trabajo de manera no comercial, siempre y cuando le acrediten y licencien sus nuevas creaciones bajo los mismos términos.

9. Código.

La construcción y ejecución del código para la extracción de la información presente en esta práctica se trabajó en lenguaje Python mediante el uso del entorno proporcionado por Jupiter Notebook. El código completo se encuentra en la ruta `source/PRA1_webscraping.ipynb`, en esta misma carpeta se encuentra el archivo `requirements.txt` el cual contiene las librerías utilizadas en el entorno de trabajo en el cual se desarrolló este proyecto.

El script construido ejecuta las siguientes instrucciones:

1. Importa las librerías requeridas para realizar webscraping
2. Define Google Chrome el navegador a utilizar
3. Captura un área de interés para la búsqueda de inmuebles, posteriormente ingresa a la página <https://www.fincaraiz.com.co> simulando dicha búsqueda, recorre todas las páginas de resultado y recopila las url de todos los inmuebles que cumplen con la condición especificada, y genera una lista y un dataframe con dichas url. Este proceso lo realiza accediendo a la información contenida en el elemento 'aria-label' contenido en 'ul',{'class':'MuiPagination-ul'}. Para este ejercicio se hace la búsqueda para la localidad de Engativá, que es una de las más grandes en la ciudad de Bogotá.
4. Crea función encargada de recorrer cada url de referencia, extraer los datos más relevantes. Esta función recorre página a página, extrae la información contenida en los elementos 'div', {'id': 'general'} y 'div',{'id':'location'} en donde se encuentran los datos más relevantes del inmueble; habitaciones, baños, parqueadero, etc... y los

- datos de ubicación geográfica del inmueble. Estos datos son limpiados y almacenados en un diccionario de donde se tomarán los datos finales.
5. Aplica la función creada en el punto 4 y la aplica a cada elemento alojado en el dataframe creado en el paso 2, los resultados recolectados son guardados en un dataframe final (df_inmuebles).
 6. Exporta los datos recolectados en el dataframe final a un archivo .csv nombrado "InmueblesVentaEngativa.csv"

Los principales retos para extraer la información de este sitio web se presentaron al momento de identificar los elementos en los cuales se encontraba contenida la información de interés, puesto que en el material dispuesto por la UOC no se encuentra explicado el método a utilizar cuando la página web no utiliza un nombre único para los elementos en común de cada artículo (por ejemplo, en el inmueble 1 la clase en la que se encuentra el dato de número de habitaciones es el "js66 1", pero al consultar el inmueble 2 el dato correspondiente al mismo atributo se encuentra en la clase "js66 2"). Para solucionar esta dificultad consulté varias páginas web y vi muchos vídeos en Youtube, sin embargo, no encontré la solución, así que analicé un poco mejor los datos que hasta el momento me arrojaba el script y me percaté de que podía generar un diccionario para almacenar los diferentes atributos del inmueble con su respectivo valor. Esta solución me permitió avanzar para lograr la captura y posterior consulta de los atributos que requería.

10. Dataset.

El dataset resultante de este ejercicio se puede consultar en el siguiente enlace: <https://doi.org/10.5281/zenodo.7854063>

11. Vídeo.


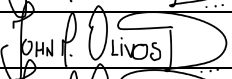
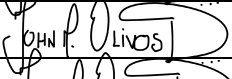
El vídeo explicativo de este proyecto se puede consultar en el siguiente enlace: https://drive.google.com/file/d/1SevogEM7vFW8mJaIEOoExLVCQZqo4 - v/view?usp=share_link

12. Bibliografía

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Masip, D. (2019). El lenguaje Python. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.
- Tutorial de GitHub <https://guides.github.com/activities/hello-world>.
- GitHub - CamiloAguilar/HousePricing: The main objective of this project is to predict the price of housing in Colombia using different models and computational techniques framed in Machine Learning.
- Londoño Botero, J. (2021). Construcción de un conjunto de microdatos del mercado de vivienda en Colombia a partir de información de anuncios web (Publicación n.º 10784/29875) [Trabajo de grado, Universidad EAFIT]. repository.eafit.edu.co/bitstream/handle/.
https://repository.eafit.edu.co/bitstream/handle/10784/29875/Jose_LondonoBotero_2021.pdf?sequence=2&isAllowed=y

Contribuciones

Tabla 2 Tabla de contribuciones

Contribuciones	Firma
Investigación previa	
Redacción de las respuestas	
Desarrollo del código	
Participación en el vídeo	