Práctica de Web Scraping

Tipología y ciclo de vida de los datos

Autores:

Tania Gualli Culqui (TGC) y Fernando Meza Ibarra (FMI)

Enlace GITHUB

https://github.com/fmezaibarra/WebScraping

https://github.com/taniagdn/WebScraping

Abril 2020

Tabla de contenido

Introducción:

La revista Mundo Constructor [1], en el newsletter de octubre de 2019, presentó un estudio de la oferta del sector inmobiliario, en las tres ciudades más importantes de Ecuador: Quito, Guayaquil y Cuenca. Para conocer las características de las edificaciones más demandadas en estas ciudades, Mundo Constructor tuvo acceso a los informes del sector inmobiliario que prepara **Properati**, una plataforma que gestiona y genera informes a partir de las búsquedas de los propios usuarios en su portal. Los datos de este estudio se obtuvieron en un rango de fechas entre marzo y agosto de 2019.

A continuación se citan algunos aspectos importantes de un estudio realizado en Quito, Guayaquil y Cuenca:

- El 64 % de la oferta de construcciones, está enfocado en la venta, mientras que el 36 % es para alquiler. Por otra parte, el 52 % de la demanda de inmuebles, por parte del consumidor, es para alquiler, mientras que el 48 % busca bienes en venta, con lo cual se puede concluir, que existe un exceso de oferta en la venta de casas y departamentos en las ciudades más importantes del país.

- En Quito, el 70% de la oferta de inmuebles corresponde a departamentos y tan solo el 30% a casas. En Guayaquil esta tendencia es más equilibrada, mientras, en Cuenca la oferta de casas es mayor a la de departamentos.

- El valor por metro cuadrado de los bienes inmuebles también presenta una diferencia importante. Así en Quito, en promedio el valor del metro cuadrado se ubica en USD 1.318, mientras en Guayaquil el precio por metro cuadrado es de USD 1.142.

- Los valores del precio por m² por sector se obtuvieron tras un análisis de relocalización de las propiedades listadas en Properati.

Conocer estos datos, sin duda, aporta en la decisiones de vendedores y compradores, razón por la que se ha identificado la necesidad de contar con información detallada de las características de las propiedades inmobiliarias y su valoración, de manera automatizada, con el fin de contar con un conjunto de datos actualizados con los que se pueda generar análisis similares a los presentados por la revista Mundo Constructor de forma inmediata y por otra parte, mediante el uso de algoritmos especializados de machine learning o deep learning se puedan construir modelos predictivos y de clasificación que sean capaces de determinar el valor del bien inmueble y preferencias (terreno, casa, departamentos, suite), con base a ciertas características como el área, sector, número de habitaciones, servicios adicionales, etc.

Para ello, se propone realizar un proceso de Web Scraping en el portal inmobiliario PROPERATI para obtener la información indicada.

Este portal es una plataforma web, de alta afluencia, que conecta compradores y vendedores de bienes raíces, facilitando las tareas de quienes buscan un bien inmueble y de las inmobiliarias que ofrecen propiedades. [2].

El ámbito de acción de Properati se extiende a varios países, como se puede observar en el siguiente gráfico:

Figura 1 – Ámbito de acción de Properati
Tomado de [2]

Actualmente, el sitio de e-commerce OLX[i] realizó la compra de la plataforma Properati, con lo cual alcanzará más usuarios en el mercado inmobiliario. El acuerdo es válido para la Argentina y para otros 13 países de América latina. [3]

Figura 2 – Filtro aplicado en la consulta de datos
Tomado de [3]

Properati también dispone de una aplicación móvil que facilita a los usuarios acceder a las ofertas de bienes inmuebles.

Desarrollo:

1. Contexto

Para el proceso de Web Scraping se ha elegido el portal inmobiliario de "*Properati*", porque es uno de los sitios más frecuentados en el país para la búsqueda de bienes inmuebles, además, opera en varios países de Latinoamérica. Es un sitio sobrio, bien organizado y de acuerdo al análisis realizado, recolecta gran cantidad de características de las propiedades.

El conjunto de datos se recopilará realizando un filtrado únicamente para la ciudad de Quito.

El dataset resultante tendrá almacenado la fluctuación de precios y demás características de cada bien inmueble. Siendo fundamentalmente el precio y el sector, indicadores muy importantes para una consideración de analítica de datos.

2. Título de dataset

Valoración de Inmuebles Quito

3. Descripción del dataset

El dataset es una recopilación de características de bienes inmuebles que se encuentran publicados en el portal de "***Properati***", cuyas variables más importantes son la valoración del bien inmueble y el sector donde se encuentra. Además, en este portal, de manera permanente, se dan de alta nuevas publicaciones de propiedades y asimismo las dan de baja cuando un comprador ha tenido éxito y se ha cerrado la operación de compra- venta.

En una segunda fase, el dataset requerirá de un proceso de limpieza y depuración de datos, para corregir valores tales como NA en los precios y en el área de construcción, quitar símbolos de moneda y de metros cuadrados y acondicionamiento de la fecha principalmente.

Como un avance futuro a esta práctica de web Scraping se puede considerar obtener otra información adicional como:

- Detalle del bien inmueble
- Número de baños
- Número de estacionamientos

- Número de fotos
- Contactos (teléfonos y correos electrónicos)
- Fotografías del bien inmueble
- Precio por m²

La URL que se considera para la extracción del dataset es la siguiente:

https://www.properati.com.ec/quito/venta/

La consulta obtenida responde al filtrado de todos los bienes inmuebles ofertados en el portal para todos los sectores de la ciudad de Quito.

La figura siguiente muestra el filtrado elegido:

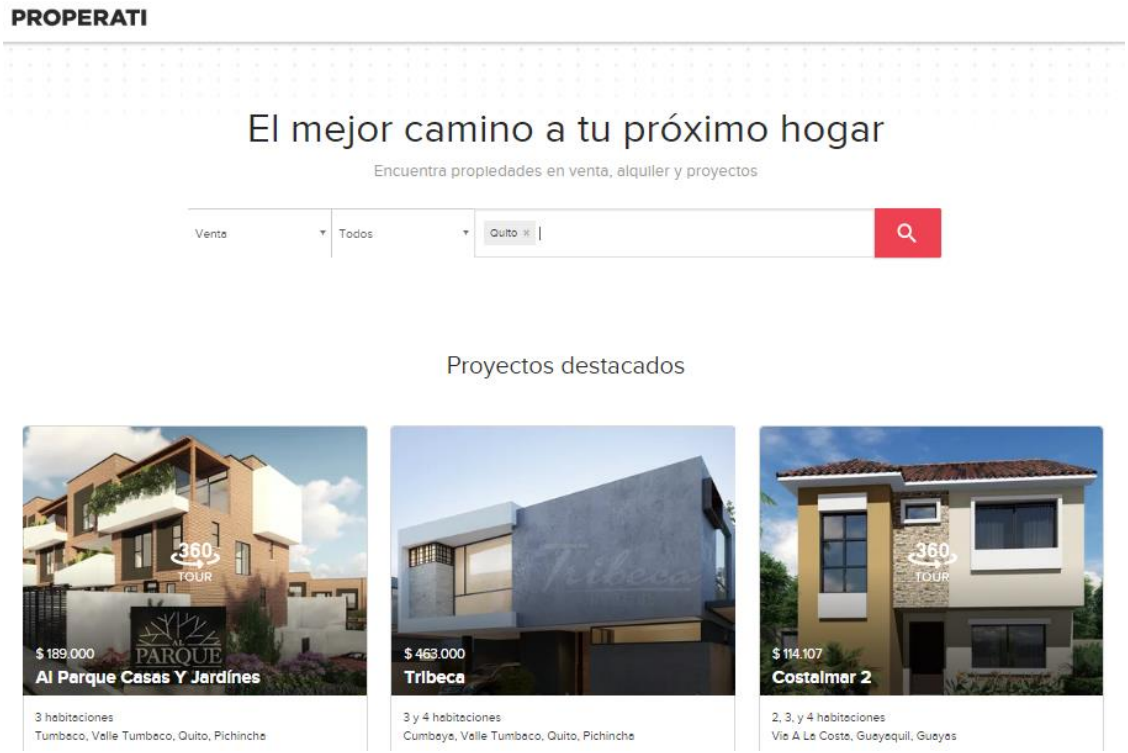

Figura 3 – Filtro aplicado en la consulta de datos
Tomado de [2]

4. Representación gráfica

A continuación, se presenta una imagen o esquema que identifica al dataset:

Figura 4 – Imagen que representa el dataset de propiedades inmobiliarias
Tomado de [2]

5. Contenido

El conjunto de datos fue recolectado de todo el resultado obtenido en la búsqueda con el filtrado de todos los ítems que están a la venta en los diferentes sectores de la ciudad de Quito.

El portal presenta este resultado en páginas de 20 ítems cada una, al momento de realizar el proceso de Scraping se alcanzó las 50 páginas. El dataset tendrá alrededor de 1000 registros.

A continuación, se realiza una descripción de los campos del dataset:

- ✓ **descripcion_conjunto**: Muestra el nombre del conjunto o urbanización, en muchos casos va el nombre de la inmobiliaria o constructora.
- ✓ **tipo**: Nombre del bien (casa, departamento, suite, terreno / lote)
- ✓ **precio**: Es el precio de venta del bien inmueble (hay que considerar que este valor incluye un valor de comisión a favor de los agentes inmobiliarios)
- ✓ **localización**: Muestra el nombre del sector, parroquia, barrio principalmente.
- ✓ **fecha_publicacion**: Indica la fecha de publicación, cuando salió al aire.
- ✓ **area**: Indica el área de construcción (cubierta) del bien o dimensión del terreno / lote.
- ✓ **num_habitaciones**: cuando existe, indica de cuantas habitaciones está compuesta la vivienda.

- ✓ **proy**: Muestra el nombre del proyecto. En Ecuador, en lo que respecta a las viviendas existen proyectos que son financiados por las entidades bancarias y por el banco del Instituto Ecuatoriano de Seguridad Social (BIESS), y de acuerdo a su monto han sido catalogados como proyectos de Vivienda de Interés Social (VIS), Vivienda de Interés Social Prioritaria (VIP) y proyectos inmobiliarios. Todos ellos diferenciándose por el precio, área fundamentalmente. Lo que se intenta es que estas viviendas estén al alcance de personas de todos estratos socio-económicos.

Los datos registran fechas desde hace un año atrás.

Como primer paso se ha realizado la descarga del sitio web de interés. Para ello se utilizan las librerías ***urlopen*** y ***BeautifulSoup***.

Con el objeto ***BeautifulSoup***, accedemos al documento en su conjunto. Generalmente puede tratarse como un objeto de tipo ***tag***, por lo que soporta el uso de la mayoría de operaciones que permiten navegar por la estructura anidada de una página.
Se realizó un Análisis Horizontal, accediendo a los elementos hermanos (siblings) del documento HTML, es decir aquellos que se encuentran al mismo nivel, es decir, que son hijos de una misma etiqueta.

Para esta práctica se ha revisado varias páginas web, con el objetivo de evitar scrapear información que no esté permitida por el webmaster.

Se permite acceso desde todos los navegadores y se ha dispuesto un tiempo de 30 segundos entre peticiones a las páginas.

La carpeta de “investigación anterior”, contiene información en los siguientes archivos:

Builtwith_whois: Contiene información sobre la tecnología utilizada en la creación de la página web y sobre el propietario de esta.

robots_txt: El fichero robots.txt, contiene las restricciones de rastreo que ha establecido el webmaster.

site_size: Estimación del tamaño de la página web.

who_is: Indica el propietario e información del registro del dominio properati.com

sitemap: https://www.properati.com.ec/static/EC/sitemap-EC.xml.

Terms_and_conditions: Contiene los términos y condiciones de uso de la página web de Properati.

Estructura de carpetas (GITHUB):

Carpeta	Descripción
Investigación anterior:	Contiene información recogida en la fase previa a la práctica además de información inherente al sitio properati.
scr:	Programa en lenguaje Python que realiza el scraper.
doc:	Contiene el desarrollo de la práctica de Web Scraping.
csv:	Contiene el dataset final y el archivo en formato excel.
README.md:	Famoso “léeme” con una descripción rápida del proyecto de Scraping.

6. Agradecimientos

Los autores agradecemos al portal web de Properati, del cual se recopilaron los registros de los bienes inmuebles para el presente proyecto.

7. Inspiración

El dataset obtenido en el presente proyecto, es muy interesante porque tiene un gran potencial ser aplicado en diversas situaciones u oportunidades de negocio, a continuación, se presentan varios ejemplos del uso de estos datos inmobiliarios:

Los tesistas Santiago y Forte [4] desarrollaron un estudio para estimar los valores de las propiedades en la Ciudad de Buenos Aires, analizando cuales son los determinantes del precio, es decir que variables son las que influyen sobre el precio de la propiedad y con qué magnitud, con el objetivo de complementar el trabajo de los tasadores. Para ello analizaron las variaciones económicas, como la demanda, oferta y relaciones con la situación económica en general y a su vez, también, factores de la propiedad que afecten a su valor, como lo son: ambientes, barrio, ubicación dentro del barrio, tipo de propiedad, ubicación dentro del edificio, cercanía de subterráneo y de vías del tren, y todos los posibles elementos que puedan tener alguna relación con el valor de la propiedad. Estos datos los obtuvieron de varias fuentes, una de ellas, el portal Properati.

Por otra parte, en la Universidad San Francisco de Quito, se desarrolló un estudio para determinar la viabilidad del proyecto inmobiliario “La Victoria” [5], ubicado en el valle de Tumbaco, Ecuador, para lo cual se recopiló datos de fuentes

primarias, entre ellas, portales inmobiliarios como Plusvalía y Properati, con lo cual se pudo realizar, un análisis objetivo de la demanda y también un análisis comparativo de las características y posicionamiento del proyecto en comparación con la competencia. Entre la información levantada para cada proyecto se buscó: la ubicación del proyecto, tipología de producto, promotor del proyecto, número de unidades, etc. Al final, el estudio permitió proponer la redefinición del proyecto.

De igual forma, en la Universidad San Andrés de Argentina, los tesistas Pierro y Ricciardelli [6] generaron una propuesta innovadora de servicios, para la venta de propiedades en la ciudad de Buenos Aires, mediante la implementación de nuevas tecnologías. El principal objetivo fue conectar la demanda existente, con la gran oferta producida por los grandes constructores, para lo cual, entre otras cosas, recabaron datos inmobiliarios del portal Properati. Con esta data y mediante otras fuentes, pudieron proponer un servicio personalizado para cada cliente, mediante la identificación de sus intereses y gustos de compra, a tal punto que a la hora de explorar la propiedad en realidad virtual, esta podría estar decorada según el estilo del cliente.

También, en la Universidad de las Fuerzas Armadas del Ecuador, el tesista Borja [7] realizó un estudio para valorar la contaminación del aire en los precios de las viviendas mediante el método de los precios hedónicos de las parroquias de Calderón, Belisario Quevedo y Guamaní. Para cumplir con dicho objetivo se muestreó un número representativo de casas mediante buscadores web de inmobiliarias que recogen las ofertas de casas, uno de ellos, el portal Properati. También se recopiló información ambiental de las concentraciones de los contaminantes. A través de este estudio, se desarrollaron algunas propuestas de políticas ambientales orientadas al mejoramiento de la salud pública de la ciudad.

Como se ha observado, el uso de estos datos es bastante diverso, sin embargo, la extracción de los datos se realizó de manera puntual y en un periodo específico, de modo que si se necesitan datos actualizados se deberá volver a repetir el proceso de obtención de los mismos. En nuestro caso, el valor agregado es la implementación de un proceso automatizado de la extracción, transformación y almacenamiento de los datos de tal forma que se cuente con información actualizada y lista para la tarea de análisis.
Además, en trabajos futuros se podría ir complementando con la extracción de información de otros portales inmobiliarios existentes en Ecuador como Plusvalía.

El Dataset utilizado en algoritmos de Machine Learning propende dar solución a las siguientes necesidades:

- ✓ Machine Learning para tasar una promoción inmobiliaria, mediante el uso de algoritmos que realizan los cálculos de cuánto puede valer un inmueble, según los datos recolectados, teniendo en cuenta múltiples variables como antigüedad del inmueble, barrio, superficie, categoría,

entre otras. Introduciendo estos datos, machine learning pueden predecir cuánto costará un bien raíz en un tiempo determinado, la evolución de precios en la zona, a cuanto podemos vender nuestros inmuebles… O de la parte del comprador, su capacidad de endeudamiento y el valor aproximado de su hipoteca.

- ✓ Machine Learning para determinar el estilo de vida de los compradores en función de sus preferencias.
- ✓ Machine Learning para que las instituciones financieras o cualquier institución puedan decidir más inteligentemente cómo actuar en el caso de invertir en construcciones o de dar forma a una hipoteca.
- ✓ Machine Learning para que las compañías de seguros también disfrutan de las bondades del Big Data aplicado al mundo inmobiliario. Gracias a él, pueden calcular la póliza más apropiada para cada edificación en función de información relacionada con variables de localización, así como de otras variables adicionales tales como; el índice de criminalidad o de incendios en una región especial determinada, o por la disponibilidad de servicios como centros educativos, hospitales o facilidades de transporte.

8. Licencia

La licencia seleccionada es:

Released Under CC BY-NC-SA 4.0 License
Se usa esta licencia porque según sus cláusulas:
El beneficiario tiene que reconocer y citar la obra y el autor del conjunto de datos generado, e indicar las obras derivadas, por lo que se reconoce la obra original y las contribuciones posteriores.
Se puede copiar, distribuir, representar la obra y hacer modificaciones para fines no comerciales. Por último, las obras derivadas sobre el trabajo que se publica bajo la licencia que regula la obra original deben distribuirse bajo la misma, así nos aseguramos además de que cumplimos con las estricciones de las páginas web que rastreamos.

9. Código

Se anexa el código fuente del Scraper desarrollado en Python, ver el enlace GIHUB en la siguiente url:
https://github.com/fmezaibarra/WebScraping/tree/master/src
La versión de Python es la 3.7.

10. Dataset

El Dataset resultante del Scraper se encuentra publicado en Zenodo [8], como se observa en la siguiente imagen:

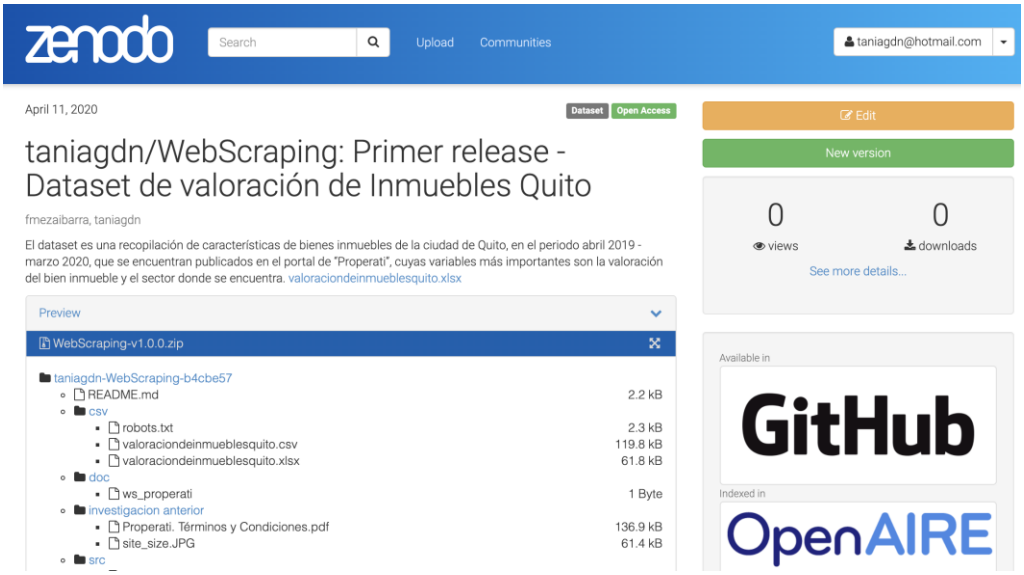

Figura 5 – Publicación del dataset en Zenodo
Tomado de [9]

11. Entrega

A continuación, se indica el DOI del dataset:

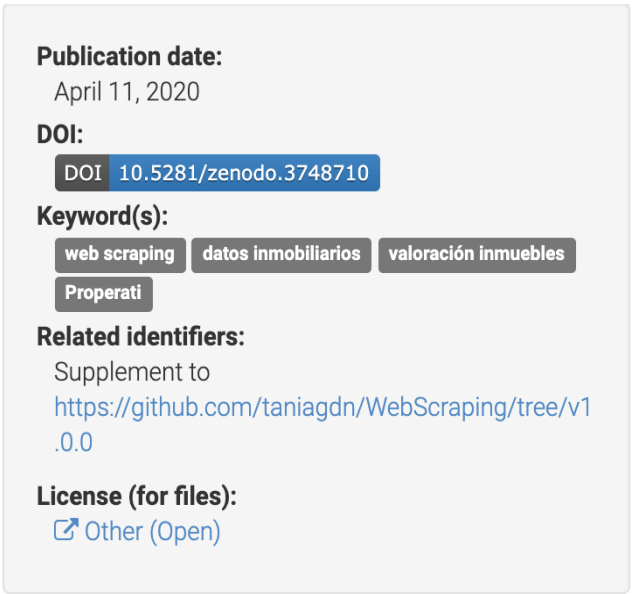

Figura 6 – DOI del dataset
Tomado de [9]

Y el repositorio GitHub con el nuevo DOI:

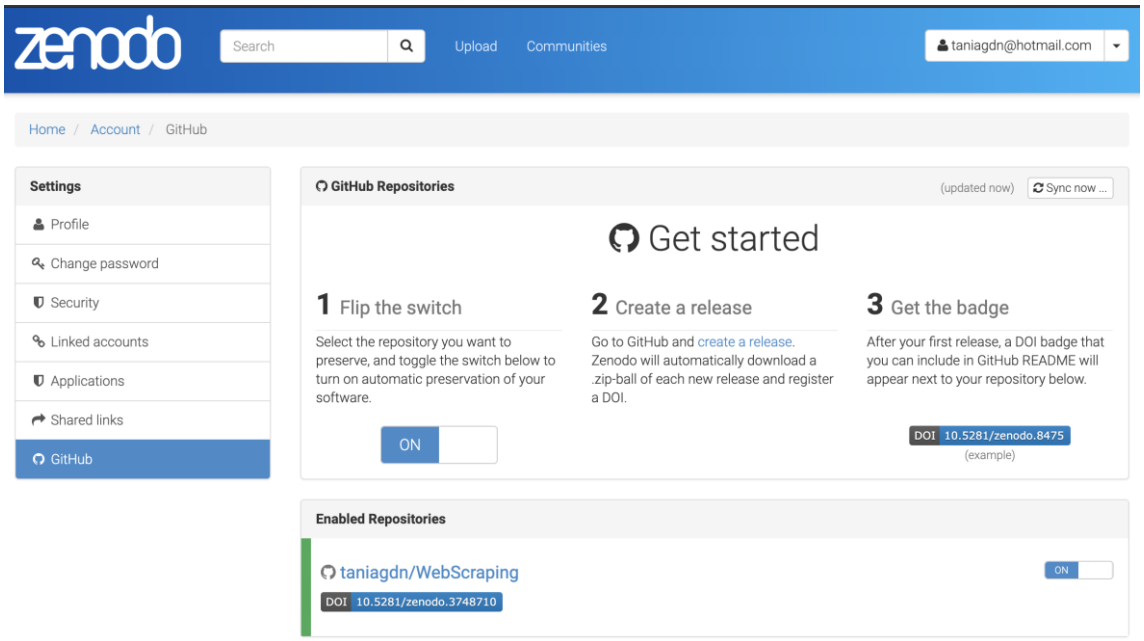

Figura 7 – Repositorio del dataset con el nuevo DOI
Tomado de [9]

12. Referencias

[1] M. Constructor, “El mercado inmobiliario ecuatoriano se desacelera pese al crecimiento del crédito en la economía,” 2019. [Online]. Available: https://www.mundoconstructor.com.ec/el-mercado-inmobiliario-ecuatoriano-se-desacelera-pese-al-crecimiento-del-credito-en-la-economia/.

[2] Properati, “Inmuebles en Venta en Quito,” 2020. [Online]. Available: https://www.properati.com.ec/nf/inmuebles/?address=&keywords=&operation=sell&operation_id=2&place_ids=%5B1701%5D&place_parent_ids=&point=.

[3] IProfesional, “La firma de e-commerce llegó a un acuerdo con la plataforma especializada en el segmento inmobiliario para operar en Latinoamérica,” 2018. [Online]. Available: https://www.iprofesional.com/notas/270419-OLX-compro-Properati-y-se-fortalece-para-competir-con-Mercado-Libre.

[4] C. Santiago and F. Forte, “Desarrollo de un estimador del valor de propiedades en la ciudad de Buenos Aires,” 2013.

[5] H. Maldonado Cardona, “Plan de Negocio : Conjunto La Victoria Harold Alberto Cardona Maldonado,” 2019.

[6] T. Pierro and F. Ricciardelli, “Propuesta Plan de Negocios Developers Club,” 2018.
[7] S. Borja Urbano, “Valoración Económica de la Contaminación del aire en los precios de las viviendas de las parroquias de Calderón, Belisario Quevedo y Guamaní,” 2019.
[8] GitHub, “Making Your Code Citable,” 2016. [Online]. Available: https://guides.github.com/activities/citable-code/.
[9] CERN, “Zenodo,” 2020. .

Contribuciones	Firma
Investigación previa	TGC, FMI
Redacción de las respuestas	TGC, FMI
Desarrollo código	TGC, FMI

Fin de Documento
Abril 2020