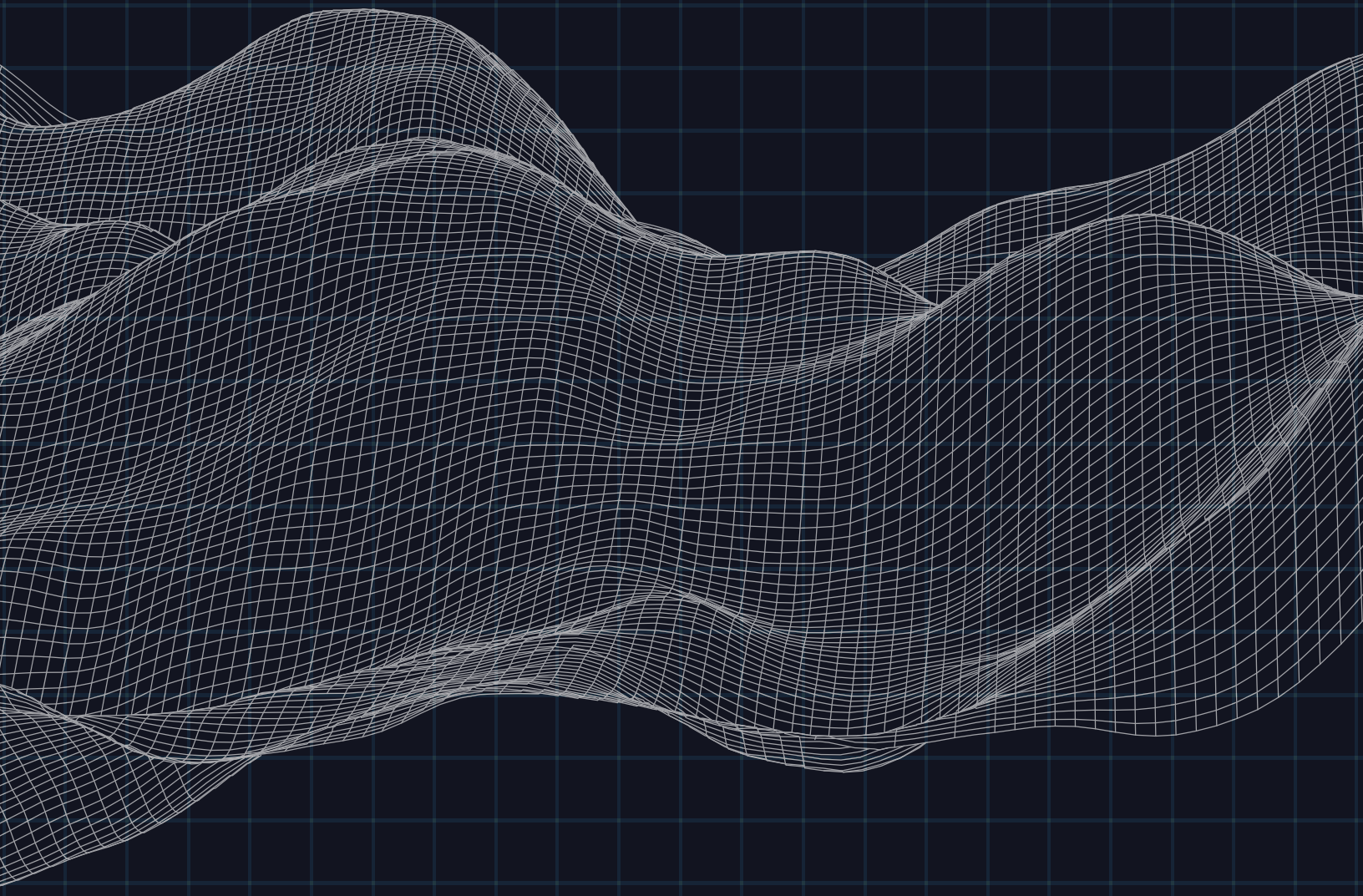
Gobierno de Colombia

IGAC
INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI

Boletín de Analítica N°1



Dirección de
INVESTIGACIÓN
y Prospectiva

Boletín de analítica N°1

El Boletín de Analítica de Datos del OIC fue creado por iniciativa del Equipo Directivo del IGAC, con el objetivo de conocer y difundir los aportes significativos en analítica de datos realizados por los equipos de la Dirección de Investigación y Prospectiva (DIP) y del Observatorio Inmobiliario Catastral (OIC).

A menudo, los desarrollos en analítica de datos realizados en el marco de proyectos específicos, no se comunican y por ende se restringe la posibilidad de aprovecharlos en otros proyectos; este Boletín busca incentivar la difusión de los trabajos técnicos que se desarrollan en el IGAC, pues en la actualidad todo lo que se realiza en la dirección de investigación y prospectiva está permeado por la analítica, la ciencia de datos y la inteligencia artificial y los logros alcanzados tienen una expresión en la formación y en la gestión y apropiación del conocimiento.

En esta primera entrega se presenta procesamiento de unos datos de transacciones jurídicas registradas en folios de matrícula inmobiliaria suministrados por la Superintendencia de Notariado y Registro – SNR, el cual fue realizado por el equipo del OIC.

Estos datos son cruciales para la conformación de la línea base de transacciones que a su vez permiten dimensionar y comprender el mercado inmobiliario en Colombia y servir de insumo para los procesos de formalización y actualización catastral.

Contexto de las acciones de la DIP

Fuente: DIP – IGAC, (2024)

Requísitos para difundir aportes en el boletín de analítica del IGAC

La DIP del IGAC, a través del OIC invita a los diferentes equipos de trabajo a difundir sus iniciativas en analítica de datos, para lo que se solicita el cumplimiento de los siguientes requisitos:

1. Identificar y describir el problema que se enfrenta y su relación con la analítica, ciencia de datos o inteligencia artificial.
2. Definir y presentar claramente la necesidad de tratamiento de datos en ese contexto.
3. La solución al problema se atiende usando: Inteligencia Artificial – IA, Modelos de Lenguaje – LLM, Analítica de datos mediante lenguaje Python, y visualización de datos.
4. Compartir el código empleado.
5. Describir de manera clara y concisa el proceso.

Autores:
Oscar Romero
Juan Pablo Gómez
Luis Andrés Campos
Mariana Ríos

Tratamiento de datos de la SNR

La estrategia que el OIC ha implementado para procesar, integrar y analizar los datos de las transacciones registradas en la SNR refleja el avance técnico del IGAC en el manejo y utilización de información relevante. Este enfoque es fundamental para comprender la dinámica inmobiliaria nacional y ha permitido consolidar una línea base para el período 2015–2022, la cual facilita la identificación de patrones y tendencias en los niveles departamental, municipal, y en las zonas urbano-rurales.

Desarrollo de la contribución

La SNR tiene como propósito dar publicidad a los actos jurídicos registrados. Esto lo realiza mediante anotaciones cronológicas en los folios de matrícula inmobiliaria, las cuales se gestionan a través de dos sistemas de información: FOLIO y SIR. Cada uno de estos sistemas genera un registro por cada interviniente en la transacción. Por ejemplo, si se realiza la compraventa de una casa de 2 vendedores a 3 compradores, la SNR efectúa 5 anotaciones, cada una con su respectivo número de documento.

El objetivo principal del preprocesamiento de los datos de la SNR es establecer y continuar alimentando la línea base de transacciones inmobiliarias, garantizando precisión, integridad y uniformidad, facilitando así análisis posteriores y decisiones basadas en información confiable y actualizada.

Este proceso implica una serie de pasos críticos, que van desde la limpieza inicial de los datos hasta técnicas más complejas de normalización y transformación, de tal manera que los datos no solo sean utilizables sino también óptimos para los análisis requeridos por las diversas partes interesadas, en este caso, el ejercicio es insumo para atender las funciones del OIC relativas al seguimiento del mercado inmobiliario del país, así como elaborar investigaciones y estudios asociados a la dinámica inmobiliaria (art 20, Decreto 846 de 2021).

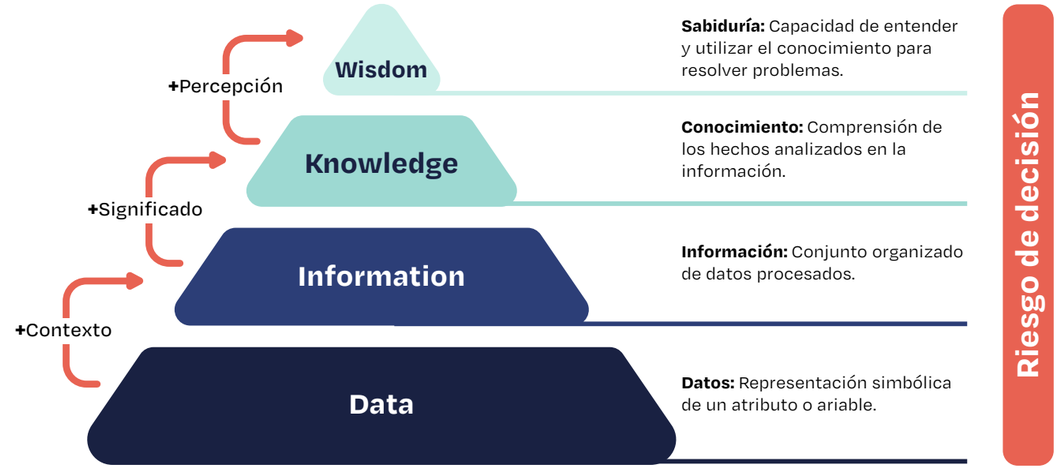
Uno de los principales desafíos en el preprocesamiento de los datos de la SNR es asegurar que el tratamiento de los datos conduzca a resultados reproducibles (obtención de resultados consistentes usando el mismo método en el mismo conjunto de datos) y replicables (resultados consistentes usando el mismo método en diferentes conjuntos de datos).

Atendiendo a lo anterior, el desarrollo realizado atiende al modelo de Jerarquía de Datos, Información, Conocimiento y Sabiduría (DIKW), que se presenta como un marco teórico esencial para abordar la transformación de datos brutos en conocimiento más complejo y aplicable (Ackoff, 1989). Esta perspectiva es conveniente para el IGAC que requiere la conversión eficiente de datos en información y luego en conocimiento operable para el catastro multipropósito y la toma de decisiones que le permitan alcanzar sus objetivos institucionales.

El proceso es complejo e implica pasos críticos que se exponen a continuación, los cuales van desde la limpieza inicial de los datos, hasta técnicas avanzadas de normalización y transformación.

2. Enriquecimiento y Segmentación de Datos:

Posteriormente, los datos fueron enriquecidos con información adicional utilizando scripts de Python, lo que incluyó códigos estandarizados y atributos esenciales para los análisis registrales. En la segmentación y estandarización, se aplicaron múltiples métodos de programación orientada a objetos creados específicamente para el procesamiento de esta información. Estos métodos procesaron los datos para grupos específicos, aplicando transformaciones y estandarizaciones que facilitaron la organización de los datos por ubicación municipal en formatos estructurados. El resultado del proceso de depuración y enriquecimiento permitió la identificación de 28.785.000 millones de registros únicos, que dan cuenta del número de transacciones diferentes.



Fuente: *OIC – IGAC. (2024) adaptado de: Ackoff, R. L. (1989). From data to wisdom. In R. L. Ackoff (Ed.), Theories of communication: A collection of readings.*

1. Preparación y Normalización de Datos

En esta fase inicial, los datos provistos por la SNR fueron recibidos en varios formatos de Excel, sumando aproximadamente 63 millones de registros y ocupando 7,5 GB de espacio en disco. Se utilizó un proceso de preprocesamiento para convertir y optimizar estos datos en el formato Parquet, facilitando su manejo debido a su estructura columnar que mejora la velocidad de acceso y procesamiento. Durante esta etapa, se realizaron tareas de limpieza y estandarización, incluyendo la normalización de nombres de municipios y departamentos según los estándares del DANE (Departamento Nacional de Estadística) y la asignación de códigos DIVIPOLA (Códigos de división político-administrativa) correctos para cada entrada, lo cual es crucial para la precisión en la vinculación con registros catastrales y la representación cartográfica.

3. Evaluación, Validación y Optimización del Proceso:

A lo largo de todo el proceso, se realizaron comprobaciones de integridad para asegurar que no se introdujeran errores o inconsistencias. La fase de evaluación y validación verificó que el conjunto de datos transformado y enriquecido mantuviera su integridad. Además, la conversión a Parquet y el uso de métodos optimizados subrayaron la importancia de las herramientas auxiliares en el proceso de estandarización de datos. La colaboración entre diferentes scripts y métodos contribuyó significativamente al éxito del proyecto, preparando los datos para enfrentar desafíos analíticos complejos y decisiones basadas en datos confiables y bien estructurados.

Bibliotecas Utilizadas en el Proceso:

El procesamiento y la estandarización de datos en la clase PreprocessingSNR se llevaron a cabo utilizando una variedad de poderosas bibliotecas de Python, cada una

aportando funcionalidades específicas que son cruciales para manejar eficientemente grandes volúmenes de datos y realizar operaciones complejas. Las bibliotecas utilizadas incluyen:

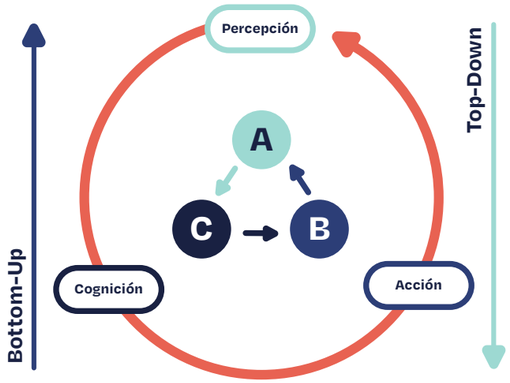
- **Pandas:** Utilizada para la creación y manipulación de DataFrames, permitiendo consolidar, transformar y limpiar datos eficazmente. Es fundamental para la manipulación y el análisis de datos, ofreciendo estructuras de datos y operaciones para tablas numéricas y series temporales.
- **NumPy**: Empleada en conjunto con Pandas para realizar cálculos numéricos eficientes, proporciona soporte para arrays y matrices grandes y multidimensionales.
- **PyArrow**: Esencial para leer y escribir archivos en formato Parquet, optimizando el almacenamiento y acceso de grandes conjuntos de datos.
- re (Regular Expressions): Utilizada en métodos de limpieza para remover caracteres especiales y realizar normalizaciones de texto, proporciona operaciones de coincidencia de patrones de texto mediante expresiones regulares.
- **json:** Clave para manejar configuraciones y datos estructurados en formato JSON, como los códigos DIVIPOLA y configuraciones del proyecto.
- **os y pathlib:** Cruciales para la gestión de rutas de archivos y directorios, permiten un manejo flexible y robusto de la estructura de archivos del proyecto.

Usos y aprovechamiento:

Tras un riguroso proceso de preprocesamiento de los datos de la SNR, estos se han convertido en un recurso invaluable para dos proyectos de investigación de la DIP. El primero enfocado en detectar posibles cambios físicos tanto en predios rurales como urbanos. El segundo aborda el tema de la dinámica inmobiliaria basada tanto en el número de transacciones, así como en su tipo y la cantidad de predios, todo a nivel municipal. Sobre estos resultados, se adelantan desarrollos para su visualización que permitan mejorar la transparencia y faciliten la interpretación de la información, de tal manera que resulte una herramienta efectiva para apoyar la toma de decisiones y la planeación estratégica.

Para lograrlo, se siguen las pautas de la metodología DAR (Mendoza Rivilla, J. 2023), mediante la cual se definen los elementos clave para el diseño y desarrollo de tableros de visualización de datos.

Esta metodología permite que el OIC sea capaz de presentar información compleja de una manera clara y fácilmente interpretable, lo que es crucial para la gestión efectiva de grandes volúmenes de datos catastrales, D.A.R. se construye con visión de lo general al detalle (Top-Down) donde el creador guía a los usuarios con su experiencia para entregar una explicación, esto ayuda a que el Usuario perciba de la información de abajo hacia arriba (Bottom-Up) guiándolo a una lectura que le brinde una respuesta.



Fuente: *(OIC - IGAC, 2024): adaptado de Qlikview Technical Brief. (2013) Dashboard, Analysis, Reporting (D.A.R.)*

Retos identificados

El OIC reconoce en el avance de las tecnologías una innegable necesidad de acoger sus contribuciones y ponerlas al servicio del país, por lo que se plantea mejorar los procesos teniendo en cuenta las siguientes oportunidades.

Evolución de la analítica de datos: Del ábaco a la IA

Fuente: *Romero, O. (2024). [Ilustración digital, empleando DALLE].*

Expansión de capacidades analíticas: El OIC busca ampliar sus capacidades analíticas para incluir predicciones avanzadas sobre tendencias del mercado inmobiliario, evaluación de impactos de políticas y planificación estratégica. Esto aumentará la relevancia del OIC como herramienta de planificación y mejorará su capacidad para influir y guiar las políticas públicas. también mejorará su capacidad para influir y guiar políticas públicas.

Automatización de la generación de insights: El OIC está trabajando en el uso de modelos predictivos y algoritmos de aprendizaje automático para automatizar la generación de insights. Esto apoyará la toma de decisiones y permitirá reacciones más rápidas a las condiciones cambiantes del mercado inmobiliario, además de mejorar la anticipación a las necesidades de planificación urbana y rural.

Personalización de la Visualización de Datos: El OIC planea desarrollar interfaces de usuario personalizables que permitan a ciudadanos, funcionarios e investigadores visualizar los datos según sus necesidades específicas.

Disposición de la información: El OIC está gestionando la publicación de los datos en el micrositio del OIC y en datos abiertos.

Mejorar las tasas de georreferenciación: El OIC busca mejorar las tasas de georreferenciación, actualmente bajas debido a la desactualización del catastro. Esta mejora es esencial para poder desarrollar análisis geoestadísticos

Conclusiones

El OIC goza de una posición privilegiada pues por una parte el Decreto 148 de 2020 le facilita el acceso a la información y por otra parte dispone de capital humano calificado y acceso a recursos tecnológicos y capacidad de cómputo que le permite aprovechar los avances en la analítica, la ciencia de datos y la inteligencia artificial y aplicarlos en el procesamiento de datos como los de la SNR. Esto permite que el país disponga de datos valiosos para analizar las dinámicas del mercado inmobiliario y apoyar grandes proyectos transformadores del gobierno nacional como son la reforma agraria, la reducción del rezago de los avalúos catastrales y los estudios de prefactibilidad para proyectos de infraestructura en todo el país.

La implementación de técnicas avanzadas y la integración de múltiples fases de procesamiento reflejaron un avance significativo en el tratamiento de los datos, este proceso no solo consolidó los datos en un formato eficiente, también los organizó de manera que reflejaron la estructura municipal precisa, lo cual es crucial para análisis posteriores. Los resultados obtenidos, son apenas el inicio de muchas posibilidades de generación de información y de conocimiento que le permiten al OIC estar mejor equipado para responder a los desafíos que le demanda el servicio público catastral y el monitoreo de la dinámica inmobiliaria, como instrumentos de apoyo a la planificación, el ordenamiento y el desarrollo territorial.

Referentes

-**Ackoff, R. L.** (1989). *From data to wisdom. In R. L. Ackoff (Ed.), Theories of communication: A collection of readings (pp. 3-9). New York: John Wiley & Sons.*

-**Baskarada, S., & Koronios, A.** (2013). *Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension.*

-**Departamento Nacional de Estadística,** (2021). *Decreto 846 de 2021. Bogotá, D.C.*

-**García, S. R., & RÍOS-INSÚA, S. I. X. T. O.** (1998). *La teoría de la decisión de Pascal a Von Neumann. Historia de la Matemática.*

-**QlikTech International** (2013) *Dashboard, Analysis, Reporting (D.A.R.). Qlikview Technical Brief, Simplifying Analysis for Everyone.*

OBSERVATORIO INMOBILIARIO CATASTRAL

Si tienes dudas o quieres ampliar sobre el contenido presentado en esta entrega del Boletín de analítica escríbenos
obs_inmobiliario@igac.gov.co
visita nuestro sito web
https://www.igac.gov.co/index.php/el-igac/areas-estrategicas/direccion-de-investigacion-y-prospectiva/observatorio-inmobiliario-catastral

Para participar en el Boletín escríbenos al correo
obs_inmobiliario@igac.gov.co
indicando:
Nombre, Área, Tema y entregando la propuesta