



UNIVERSITAT  
Carlemany

# **Predicción y Clasificación de Propiedades en Quito mediante Técnicas Avanzadas de Machine Learning y Procesamiento de Lenguaje Natural**

Titulación: Bàtxelor en Ciencia de Datos

Alumno/a: REYES CASTAÑEDA, EDISSON RODRIGO

Dirección de TFB: ESTEBAN BOLÓS, Pilar

## RESUMEN

Este proyecto de Ciencia de Datos aborda la predicción y clasificación de precios de propiedades inmobiliarias en Quito, Ecuador, mediante la aplicación de técnicas avanzadas de Machine Learning y Procesamiento de Lenguaje Natural (NLP). Se integran grandes volúmenes de datos heterogéneos, procesados con Apache Spark para su limpieza y almacenamiento eficiente en PostgreSQL, y se visualizan patrones clave utilizando Metabase junto a herramientas de visualización como Matplotlib y Seaborn. Las descripciones textuales de propiedades se analizan con Spark NLP, lo que permite extraer información relevante y completar datos faltantes. Posteriormente, se implementan modelos de regresión y ensamblado, como Random Forest y XGBoost, para predecir los precios de las propiedades, mientras que los pipelines automatizados en Apache Airflow y scripts de Python aseguran la validación, optimización continua de los modelos y la automatización de los procesos. El uso de modelos de series temporales y redes neuronales profundas permite una mayor precisión en las predicciones y una clasificación más detallada de las propiedades, estableciendo un marco metodológico replicable para otros mercados inmobiliarios.

### Palabras clave:

Predicción de precios de propiedades, Machine Learning, Procesamiento de Lenguaje Natural, Apache Spark, Apache Airflow, PostgreSQL, Metabase, Random Forest, XGBoost, series temporales, redes neuronales profundas, mercado inmobiliario, Quito.

## ABSTRACT

This Data Science project addresses the prediction and classification of real estate property prices in Quito, Ecuador, through the application of advanced Machine Learning and Natural Language Processing (NLP) techniques. Large volumes of heterogeneous data are integrated, processed with Apache Spark for efficient cleaning and storage in PostgreSQL, and key patterns are visualized using Metabase alongside visualization tools such as Matplotlib and Seaborn. Property textual descriptions are analyzed using Spark NLP, enabling the extraction of relevant information and the completion of missing data. Subsequently, regression and ensemble models, such as Random Forest and XGBoost, are implemented to predict property prices, while automated pipelines in Apache Airflow and Python scripts ensure the continuous validation, optimization of models, and automation of processes. The use of time series models and deep neural networks allows for greater accuracy in predictions and more detailed classification of properties, establishing a replicable methodological framework for other real estate markets.

### Keywords:

Real estate market, Quito, data science, price prediction, data analysis, machine learning, econometric models, determining factors, data visualization, sustainable development..

## Contenido

<b>1. Introducción</b>	<b>8</b>
<b>2. Justificación</b>	<b>9</b>
<b>3. Contextualización del Trabajo</b>	<b>10</b>
3.1 Políticas y medidas del Municipio de Quito que influyen en el mercado inmobiliario	10
3.2 Contexto de Quito y preocupaciones ciudadanas:	11
3.3 Razones para elegir Quito como objeto de estudio:	12
<b>4. Marco Teórico del Trabajo</b>	<b>13</b>
4.1 Modelos Econométricos para la Valoración de Bienes Inmuebles	13
4.2 Aprendizaje Automático (Machine Learning)	16
4.3 Algoritmos	17
4.4 Metodología	18
<b>5. Objetivos Generales y Específicos</b>	<b>21</b>
5.1 Objetivo General	21
5.2 Objetivos Específicos	21
<b>6. Metodología</b>	<b>23</b>
6.1 Instrumentos	23
6.2 Materiales	24
6.3 Recursos Humanos	24
6.4 Evaluación	25
<b>7. Desarrollo viable y sostenible</b>	<b>26</b>
7.1 Temporalización e Hitos	26
7.2 Alineación con los ODS	29
7.3 Condicionantes Ambientales, Sociales y Económicos	30
<b>8. Proceso y Resultados</b>	<b>30</b>
8.1 Fuentes de Datos y Recopilación	31
8.2 Exploración y Preparación	38
8.3 Análisis Exploratorio	39
8.4 Gestión y Almacenamiento	40
8.5 Modelado	41
8.6. Visualización	41
<b>9. Discusión y Limitaciones</b>	<b>42</b>
<b>10. Conclusiones y Líneas Futuras</b>	<b>43</b>
<b>11. Referencias Bibliográficas</b>	<b>45</b>

## Acrónimos – Figuras – Tablas

### *Acrónimos*

IES	Institución de Educación Superior
ODS	Objetivos de Desarrollo Sostenible

### *Figuras*

Figura 1 Población en riesgo de pobreza social (2020-2022)

Figura 2

### *Tablas*

Tabla 1 xxxxxxxx

Tabla 2

### ***Orientaciones estilísticas para los títulos***

#### **Nivel 1: Capítulo**

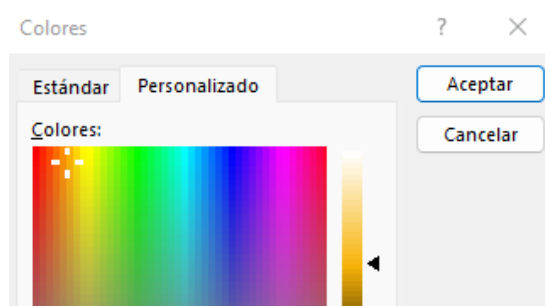
Fuente: Calibri (cuerpo)

Tamaño: 22

Interlineado: 1,5

Texto justificado

Color:



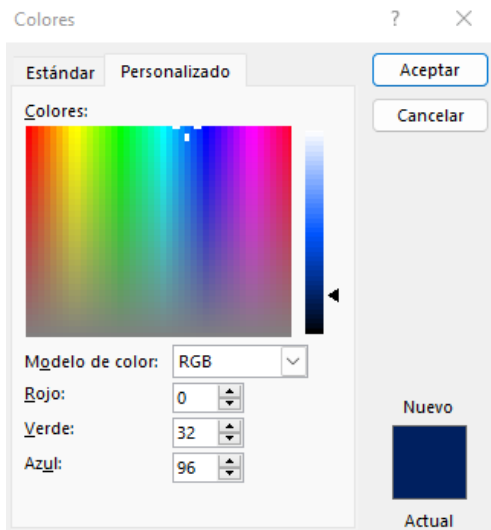
**Nivel 2: Subcapítulo**

Fuente: Calibri (cuerpo)

Tamaño: 14

Interlineado: 1,5

Texto justificado

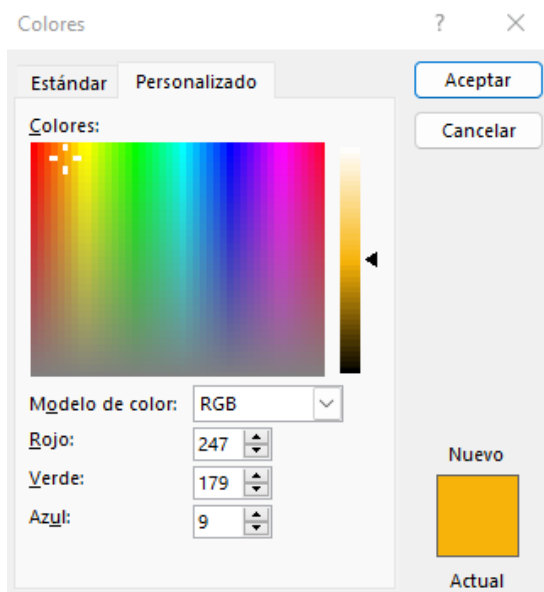
**Nivel 3: Apartado**

Fuente: Calibri (cuerpo)

Tamaño: 14

Interlineado: 1,5

Texto justificado



***Orientaciones para el cuerpo del texto***

Fuente: Calibri (cuerpo)  
Tamaño: 11  
Interlineado: 1,5  
Texto justificado

# 1. Introducción

El mercado inmobiliario en Quito ha mostrado un crecimiento significativo en los últimos años, impulsado por factores como el aumento de la población urbana, la expansión económica y el desarrollo de infraestructura. Estos factores han incrementado la demanda de propiedades para uso residencial y comercial. Sin embargo, la predicción precisa de los precios de las propiedades en un mercado tan volátil y multifacético representa un desafío considerable para agentes inmobiliarios, inversionistas y compradores (Lincango & Arsecio, 2022).

La complejidad del mercado inmobiliario radica en la interacción de múltiples variables, tales como la ubicación de las propiedades, su tamaño, características internas y las condiciones macroeconómicas. Además, la disponibilidad de datos heterogéneos y no estructurados, como descripciones textuales de propiedades, complica aún más el proceso de análisis y predicción (Kolb & Kolb, 2021). Tradicionalmente, la predicción de precios ha dependido de métodos heurísticos o herramientas analíticas básicas, insuficientes para capturar la complejidad inherente de los datos.

En respuesta a estos desafíos, este proyecto de tesis emplea técnicas avanzadas de Machine Learning y Procesamiento de Lenguaje Natural (NLP) para desarrollar modelos predictivos robustos y precisos. La ingesta y limpieza de grandes volúmenes de datos se realiza utilizando Apache Spark, optimizando el procesamiento de datos provenientes de diversas fuentes. Estos datos se almacenan en PostgreSQL, proporcionando una infraestructura robusta y escalable para análisis posteriores (Rafiei & Adeli, 2018).

El desarrollo del proyecto se lleva a cabo principalmente en Python, aprovechando su ecosistema rico en librerías para ciencia de datos, incluyendo Pandas, Scikit-learn y TensorFlow, entre otras. Python facilita no solo la implementación de algoritmos de aprendizaje automático, sino también la integración de estos modelos con otros componentes del sistema de manera efectiva y eficiente.

Para gestionar la automatización y la orquestación de flujos de trabajo, se emplea Apache Airflow. Esta herramienta permite definir, programar y supervisar procesos de extracción, transformación y carga de datos (ETL), asegurando la consistencia y repetibilidad de los análisis. Las descripciones textuales de las propiedades, a menudo ricas en detalles no capturados en datos estructurados, se procesan mediante Spark NLP, permitiendo la



extracción de información relevante que puede influir en la evaluación de precios (Jiang, Wang, Shi, & Ren, 2020).

Para la visualización y análisis interactivo de datos, se utiliza Metabase, una plataforma de código abierto que facilita la creación de consultas SQL y dashboards interactivos. Metabase permite a los usuarios finales interactuar con los resultados de los modelos de manera intuitiva, mejorando la toma de decisiones basada en datos.

El control de versiones del código fuente se gestiona a través de GitHub, permitiendo un seguimiento detallado de los cambios y facilitando la colaboración en equipo (Chacon & Straub, 2014). Aunque el desarrollo se lleva a cabo principalmente en entornos locales, el sistema está diseñado con una arquitectura escalable tanto horizontal como verticalmente, permitiendo una migración fluida hacia entornos en la nube cuando sea necesario. Esto asegura que el sistema pueda adaptarse a un aumento en el volumen de datos o en las necesidades de procesamiento sin comprometer su rendimiento.

Además, el uso de modelos de series temporales y redes neuronales profundas añade una capa adicional de análisis, facilitando la predicción de tendencias de precios a lo largo del tiempo y mejorando la clasificación de propiedades. Este enfoque integral no solo busca optimizar la precisión de las predicciones de precios, sino también establecer un marco metodológico replicable y escalable aplicable a diferentes contextos y mercados inmobiliarios.

## 2. Justificación

La capacidad de predecir de manera precisa los precios de las propiedades inmobiliarias es fundamental para la eficiencia y transparencia del mercado inmobiliario. En Quito, como en muchas otras ciudades en crecimiento, esta capacidad se vuelve aún más crítica debido al dinamismo del mercado, impulsado por la urbanización acelerada, cambios en las condiciones socioeconómicas y fluctuaciones en la oferta y demanda de propiedades. Las decisiones de inversión, compra y venta de propiedades tienen un impacto significativo no solo en la economía individual de los actores involucrados, sino también en la economía local y regional en su conjunto (Lincango & Arsecio, 2022).

**Limitaciones de Métodos Tradicionales:** Tradicionalmente, la evaluación de precios de propiedades ha dependido de métodos manuales o heurísticos basados en la experiencia de los agentes inmobiliarios. Estos métodos, aunque útiles, son limitados en su capacidad para manejar y procesar grandes volúmenes de datos de manera eficiente y no pueden captar toda la complejidad de los factores que influyen en los precios de las propiedades (Rafiei & Adeli, 2018). Además, el creciente volumen de datos no estructurados, como las descripciones textuales de las propiedades y otros documentos, presenta un desafío adicional que los métodos convencionales no están equipados para manejar.

**Implementación de Machine Learning y NLP:** La implementación de técnicas avanzadas de Machine Learning y Procesamiento de Lenguaje Natural (NLP) ofrece una solución innovadora a estos desafíos. Al aprovechar el poder de procesamiento distribuido de Apache Spark, es posible manejar grandes volúmenes de datos y realizar análisis complejos en tiempo real. La integración de modelos de Machine Learning permite identificar patrones no evidentes y realizar predicciones más precisas basadas en una variedad de características y datos históricos. Por ejemplo, técnicas como el Extreme Gradient Boosting han demostrado ser efectivas para mejorar la precisión de predicciones de costos y precios en contextos complejos (Sicilia Gómez, 2024).

**NLP para Enriquecer Datos:** El uso de NLP para analizar descripciones textuales proporciona una capa adicional de información que puede mejorar significativamente la precisión de las predicciones de precios. Al extraer información detallada sobre las características de las propiedades y su contexto, es posible completar datos faltantes y enriquecer los conjuntos de datos existentes, lo que lleva a modelos predictivos más completos y robustos. Esta capacidad de analizar datos no estructurados es crucial para

capturar matices que los datos estructurados tradicionales no pueden reflejar.

**Automatización y Escalabilidad:** La automatización de estos procesos a través de Apache Airflow garantiza que las tareas de limpieza de datos, análisis, modelado y validación se realicen de manera consistente y eficiente, permitiendo un monitoreo continuo y la optimización de los modelos predictivos. Esta capacidad de automatización y escalabilidad no solo mejora la eficiencia operativa, sino que también reduce el riesgo de errores humanos y asegura la repetibilidad de los resultados.

**Conclusión:** Este proyecto justifica su importancia al abordar una necesidad crítica en el mercado inmobiliario de Quito: la capacidad de realizar predicciones precisas y confiables de precios de propiedades. Al implementar un enfoque basado en datos utilizando tecnologías avanzadas de análisis y automatización, se espera no solo mejorar la precisión de las predicciones, sino también proporcionar una herramienta valiosa que pueda ser utilizada por agentes inmobiliarios, inversionistas y compradores para tomar decisiones informadas, reducir riesgos y maximizar oportunidades. Este enfoque no solo tiene el potencial de transformar el mercado inmobiliario local, sino que también puede ser replicado y adaptado a otros mercados, contribuyendo al avance del campo de la ciencia de datos aplicada al sector inmobiliario.

## 3. Contextualización del Trabajo

Este apartado proporciona un contexto detallado del mercado inmobiliario en Quito, los desafíos específicos que enfrenta y la relevancia de aplicar técnicas avanzadas de análisis de datos para abordar estos desafíos. Se justifica el enfoque metodológico elegido y se presenta el impacto potencial del proyecto en el mercado inmobiliario local y más allá.

### 3.1 Contexto General del Mercado Inmobiliario en Quito

Quito, la capital de Ecuador, es una ciudad en constante crecimiento, con una población que supera los 2.7 millones de habitantes. Según el Instituto Nacional de Estadística y Censos (INEC), la tasa de crecimiento poblacional en Quito se ha mantenido en un promedio de 1.5% anual durante los últimos años, lo que refleja una tendencia significativa hacia la urbanización. Este crecimiento demográfico ha impulsado la expansión del mercado inmobiliario, con un aumento en la construcción de nuevas viviendas. Informes recientes del Banco Central del Ecuador indican que la tasa de construcción de nuevas viviendas ha crecido un 8% anual, evidenciando una fuerte demanda de inmuebles tanto para uso residencial como comercial (Banco Central del Ecuador, 2023).

El mercado inmobiliario de Quito se caracteriza por una diversificación significativa de la oferta. Esta diversificación abarca desde apartamentos de lujo en zonas céntricas hasta viviendas más accesibles en suburbios en expansión. Sin embargo, persisten problemas de desigualdad en el acceso a la vivienda, con algunas áreas enfrentando altos precios por metro cuadrado y alquileres que representan un porcentaje significativo del ingreso familiar. Esta situación refleja una tendencia global hacia la segregación urbana, donde las áreas más caras ofrecen mejores servicios y seguridad, mientras que otras enfrentan problemas de deterioro e inseguridad (Informe de Calidad de Vida, 2023, p. 10).

### 3.2 Problemática del Mercado y Desafíos Actuales

A pesar de las oportunidades de crecimiento, el mercado inmobiliario en Quito enfrenta varios desafíos críticos. Uno de los principales problemas es la fluctuación de precios, influenciada por factores como políticas gubernamentales cambiantes, variaciones en la

oferta y demanda, y factores económicos externos (Informe de Calidad de Vida, 2023, p. 10). La falta de infraestructura adecuada en las áreas periurbanas también plantea desafíos significativos, afectando la calidad de vida y el valor de las propiedades.

Además, estudios locales indican que existe una preocupación creciente entre los consumidores sobre la transparencia y fiabilidad de las valoraciones de propiedades en Quito. Una encuesta reciente realizada por Quito Cómo Vamos reveló que más del 60% de los encuestados considera que la falta de información clara y precisa sobre los precios de las propiedades es un obstáculo para la toma de decisiones informadas en el mercado inmobiliario (Quito Cómo Vamos, 2023). Esta percepción puede afectar negativamente la confianza de los consumidores y la estabilidad del mercado.

### 3.3 Relevancia del Uso de Técnicas Avanzadas de Machine Learning y NLP

Para abordar los desafíos del mercado inmobiliario de Quito, el uso de técnicas avanzadas de Machine Learning (ML) y Procesamiento de Lenguaje Natural (NLP) se presenta como una solución efectiva. Estas técnicas permiten analizar grandes volúmenes de datos y extraer patrones complejos, proporcionando una comprensión más profunda de los factores que influyen en los precios de las propiedades y las tendencias del mercado.

**Machine Learning:** Los algoritmos de ML, como la regresión y los modelos de ensamblado (e.g., Random Forest, XGBoost), pueden identificar características clave que afectan los precios de las propiedades y predecir estos precios con alta precisión. Estos métodos son eficaces en la identificación de tendencias y en la anticipación de fluctuaciones de precios, proporcionando herramientas valiosas para la planificación estratégica y la toma de decisiones informadas (Rafiei & Adeli, 2018). Por ejemplo en el estudio de Lincango & Arsecio de 2022 discuten la relevancia del uso de ML para la predicción de precios en Quito post-pandemia.

**Procesamiento de Lenguaje Natural (NLP):** El NLP permite el análisis de descripciones textuales de propiedades, extrayendo información valiosa sobre ubicación, características internas, y otros factores influyentes. Estudios locales sobre el uso de NLP en la región sugieren que analizar descripciones en anuncios inmobiliarios puede proporcionar insights importantes sobre las preferencias de los consumidores y las características de las

propiedades que más valoran (Lincango & Arsecio, 2022). Esta información no estructurada, cuando se procesa adecuadamente, enriquece los conjuntos de datos y mejora la precisión de los modelos predictivos.

### **3.4 Enfoque Metodológico General**

El enfoque metodológico de este proyecto se basa en la integración de tecnologías avanzadas para el manejo, análisis y visualización de datos. Este enfoque asegura que el análisis sea robusto, escalable y capaz de adaptarse a los cambios dinámicos del mercado. La metodología combina el uso de herramientas para el procesamiento distribuido de datos, bases de datos relacionales para el almacenamiento de información estructurada, y plataformas de visualización para la interpretación de los resultados (Armbrust et al., 2015). El uso de flujos de trabajo automatizados garantiza la consistencia y eficiencia en la ejecución de las tareas, desde la recolección de datos hasta la generación de informes.

### **3.5 Impacto Potencial del Proyecto**

El impacto de este proyecto es significativo tanto a nivel local como global. A nivel local, proporciona a los actores del mercado inmobiliario herramientas para una evaluación precisa de propiedades, mejorando la toma de decisiones y reduciendo el riesgo de inversión. Esto puede llevar a una mayor transparencia en el mercado, una reducción de la desigualdad en el acceso a la vivienda y una mejora en la satisfacción del cliente (Informe de Calidad de Vida, 2023, p. 10). Algunos estudios que demuestran cómo la precisión en las predicciones de precios puede afectar directamente la confianza en los mercados inmobiliarios. Por ejemplo, el estudio de Antón Ruiz de 2020 analiza la predicción de precios en el mercado inmobiliario de Valencia y su relevancia para mejorar la transparencia y confianza del mercado.

Globalmente, el marco metodológico propuesto es replicable y escalable, lo que permite su adaptación a otros mercados inmobiliarios con características similares. La integración de

Machine Learning y NLP en el análisis del mercado inmobiliario no solo mejora la precisión de las predicciones, sino que también establece un estándar para el uso de tecnologías avanzadas en la evaluación de propiedades y la planificación urbana. Este enfoque demuestra cómo las técnicas avanzadas de análisis de datos pueden ser aplicadas eficazmente para resolver problemas complejos y proporcionar soluciones prácticas y efectivas en diversos contextos.

## 4. Marco Teórico del Trabajo

El marco teórico de este proyecto se centra en conceptos fundamentales y técnicas avanzadas que sustentan la predicción y clasificación de precios de propiedades inmobiliarias mediante Machine Learning (ML) y Procesamiento de Lenguaje Natural (NLP). Además, se analiza el uso de una infraestructura moderna para asegurar escalabilidad y eficiencia, utilizando Docker, Apache Airflow, Apache Spark, Metabase, PostgreSQL y GitHub. A continuación, se detalla cada componente teórico fundamental para este proyecto.

### 4.1 Introducción a Machine Learning

**Machine Learning (ML)** es una rama de la inteligencia artificial que permite a las computadoras aprender y hacer predicciones basadas en datos. Este enfoque es ideal para problemas donde los patrones son complejos o no se pueden codificar manualmente (Mitchell, 1997). ML se utiliza ampliamente para resolver problemas en dominios diversos, incluyendo la predicción de precios de propiedades inmobiliarias.

#### 4.1.1 Tipos de Aprendizaje en Machine Learning

##### Aprendizaje Supervisado

En el aprendizaje supervisado, los modelos se entrenan con datos etiquetados. Técnicas comunes en este enfoque incluyen la regresión lineal para predicción continua y métodos de clasificación para tareas discretas (Hastie, Tibshirani, & Friedman, 2009). Este tipo de aprendizaje es fundamental para predecir precios basados en características específicas de las propiedades.

##### Aprendizaje No Supervisado

El aprendizaje no supervisado se utiliza para descubrir patrones en datos no etiquetados. Ejemplos incluyen clustering con K-means para segmentar datos en grupos significativos, y técnicas de reducción de dimensionalidad como PCA para simplificar la representación de datos (Jain, 2010; Jolliffe & Cadima, 2016).



### **Aprendizaje por Refuerzo**

El aprendizaje por refuerzo implica la toma de decisiones en secuencias, maximizando recompensas acumuladas. Aunque menos común en predicción de precios, este enfoque puede ser útil para estrategias de inversión a largo plazo (Sutton & Barto, 2018).

## **4.1.2 Técnicas Avanzadas en Machine Learning**

### **Redes Neuronales y Aprendizaje Profundo**

Las redes neuronales profundas son capaces de modelar relaciones complejas en datos. Estas redes son especialmente útiles en problemas de alta dimensionalidad y no linealidad, como la predicción de precios basada en características textuales y visuales (LeCun, Bengio, & Hinton, 2015). En este proyecto, se utilizan redes neuronales convolucionales (CNN) para el análisis de imágenes de propiedades y redes neuronales recurrentes (RNN) para analizar series temporales de precios (Goodfellow, Bengio, & Courville, 2016).

### **Ensamblado de Modelos (Ensemble Learning)**

El uso de técnicas de ensamblado, como Random Forest y XGBoost, permite mejorar la precisión y robustez de las predicciones combinando múltiples modelos (Breiman, 2001; Chen & Guestrin, 2016). Estas técnicas son fundamentales para capturar la variabilidad en datos complejos y reducir el sobreajuste.

## **4.2 Procesamiento de Lenguaje Natural (NLP)**

**Procesamiento de Lenguaje Natural (NLP)** se centra en la interacción entre computadoras y el lenguaje humano (Manning & Schütze, 1999). En este proyecto, NLP se utiliza para analizar descripciones textuales de propiedades y extraer características adicionales no capturadas en datos estructurados.

#### 4.2.1 Preprocesamiento y Limpieza de Texto

El preprocesamiento incluye pasos críticos como tokenización, limpieza de texto, y lematización o stemming, asegurando que los datos sean consistentes y útiles para el análisis posterior (Jurafsky & Martin, 2008). Estos pasos son fundamentales para transformar descripciones en lenguaje natural en datos estructurados que puedan ser utilizados por los modelos predictivos.

#### 4.2.2 Extracción de Características y Enriquecimiento de Datos

- **Análisis de Descripciones Textuales:** Las descripciones de propiedades en lenguaje natural contienen información valiosa sobre características cualitativas, como la condición de la propiedad, detalles arquitectónicos, o la proximidad a puntos de interés (Kolb & Kolb, 2021). Estas descripciones se procesan para extraer información que enriquezca las características disponibles en tablas estructuradas. Por ejemplo, se pueden identificar frases que indiquen renovaciones recientes o características destacadas que podrían influir en el valor de la propiedad.
- **Nube de Palabras:** Se utilizará una nube de palabras para visualizar las palabras más comunes en las descripciones de propiedades, lo que ayudará a identificar términos relevantes y tendencias comunes en el lenguaje utilizado por agentes inmobiliarios y compradores (Heimerl, Lohmann, Lange, & Ertl, 2014). Esta visualización proporciona una visión intuitiva y rápida de los aspectos más destacados en las descripciones de propiedades.

#### 4.2.3 Modelos Semánticos y Contextuales

- **Análisis de Sentimiento:** Aplicar análisis de sentimiento a las descripciones de propiedades puede revelar la percepción del mercado respecto a ciertas características, proporcionando insights sobre cómo afectan estas percepciones al valor de la propiedad (Liu, 2012). Esta técnica permite entender cómo las descripciones

emocionales o los adjetivos utilizados en los anuncios pueden influir en la percepción de valor por parte de los potenciales compradores.

- **Modelos Basados en Transformers:** Utilizando modelos como BERT o GPT-3, se pueden capturar relaciones complejas y contextuales en el texto, mejorando la capacidad de los modelos para interpretar descripciones y predicciones de precios (Vaswani et al., 2017; Brown et al., 2020). Estos modelos avanzados son capaces de identificar sutilezas y significados implícitos en el lenguaje natural, mejorando significativamente la precisión de las tareas de clasificación y predicción.

## 4.3 Infraestructura Tecnológica y Orquestación de Flujos de Trabajo

### 4.3.1 Contenerización con Docker

Docker facilita la implementación de entornos reproducibles, encapsulando aplicaciones y dependencias en contenedores portátiles. Esto asegura que las aplicaciones funcionen de manera consistente en cualquier entorno, lo que es crucial para proyectos colaborativos y escalables (Merkel, 2014).

### 4.3.2 Automatización con Apache Airflow

Apache Airflow permite la orquestación y automatización de tareas en ciencia de datos mediante la definición de flujos de trabajo dirigidos acíclicos (DAG). Esto es esencial para manejar procesos complejos de ETL, modelado y evaluación, asegurando que los análisis sean reproducibles y gestionables (Apache Software Foundation, 2021).

### 4.3.3 Procesamiento Distribuido con Apache Spark

Apache Spark proporciona un motor de procesamiento en clúster para manejar grandes volúmenes de datos. Su capacidad de procesamiento en memoria permite realizar análisis

rápidos y eficientes, lo cual es esencial para el análisis de datos a gran escala y el entrenamiento de modelos de ML (Zaharia et al., 2016).

## 4.4 Almacenamiento y Gestión de Datos

### 4.4.1 Uso de PostgreSQL

PostgreSQL ofrece una solución robusta y flexible para el almacenamiento de datos, compatible con datos estructurados y no estructurados. Soporta transacciones ACID y proporciona capacidades avanzadas para realizar consultas complejas, fundamentales para el análisis de datos en tiempo real (Stonebraker & Rowe, 1986).

### 4.4.2 Integración y Escalabilidad

PostgreSQL se integra fácilmente con herramientas de análisis y visualización, facilitando el flujo continuo de datos desde la ingesta hasta la visualización. La capacidad de escalabilidad de PostgreSQL permite manejar crecientes volúmenes de datos y usuarios, asegurando un rendimiento óptimo en aplicaciones de gran escala.

## 4.5 Visualización de Datos

### 4.5.1 Metabase para Dashboards Interactivos

Metabase proporciona una plataforma intuitiva para crear dashboards interactivos, permitiendo a los usuarios explorar datos y obtener insights sin necesidad de habilidades técnicas avanzadas (Metabase, 2021). Esta herramienta es crucial para la toma de decisiones basada en datos, ofreciendo visualizaciones claras y accesibles.

### 4.5.2 Herramientas de Visualización en Python

Bibliotecas de Python como **Matplotlib** y **Seaborn** son esenciales para la creación de

gráficos detallados y visualizaciones personalizadas. Estas herramientas permiten explorar relaciones complejas entre variables y comunicar hallazgos de manera efectiva (Waskom, 2021).

## **4.6 Control de Versiones y Colaboración con GitHub**

GitHub proporciona un entorno robusto para el control de versiones y la colaboración en proyectos de ciencia de datos. Facilita el seguimiento de cambios, la gestión de ramas y la integración continua, asegurando que el proyecto sea reproducible y mantenga una alta calidad de código (Chacon & Straub, 2014).

## **4.7 Escalabilidad y Consideraciones de Infraestructura**

### **4.7.1 Escalabilidad Horizontal y Vertical**

La infraestructura del proyecto está diseñada para ser escalable, permitiendo manejar incrementos en el volumen de datos y la carga de trabajo mediante la adición de más nodos (escalabilidad horizontal) o mejorando los recursos existentes (escalabilidad vertical). Esta capacidad es esencial para asegurar que el sistema pueda crecer y adaptarse a las demandas cambiantes del mercado inmobiliario (Armbrust et al., 2010).

### **4.7.2 Gestión de Recursos y Optimización**

El uso de tecnologías como Docker y Apache Airflow permite la gestión eficiente de recursos, optimizando costos y asegurando que los procesos de análisis sean ejecutados de manera efectiva y sostenible. Esta gestión es clave para mantener la viabilidad del proyecto a largo plazo, permitiendo una respuesta rápida a los cambios en los requisitos del mercado y la tecnología.

## 5. Objetivos Generales y Específicos

### 5.1 Objetivo General

Desarrollar un sistema integral y robusto para la predicción y clasificación de precios de propiedades inmobiliarias en Quito, Ecuador, utilizando técnicas avanzadas de Machine Learning (ML) y Procesamiento de Lenguaje Natural (NLP). Este sistema integrará y analizará datos heterogéneos (estructurados y no estructurados) para mejorar la precisión de las predicciones y facilitar la toma de decisiones informadas por parte de agentes inmobiliarios, inversores y compradores. Además, se busca establecer un marco metodológico replicable y escalable que pueda ser adaptado a otros mercados inmobiliarios, contribuyendo al avance de la ciencia de datos aplicada al sector inmobiliario.

### 5.2 Objetivos Específicos

1. **Desarrollar un modelo de predicción de precios de propiedades basado en Machine Learning:**
  - a. Identificar y seleccionar características relevantes que influyen en el precio de las propiedades, tales como ubicación, tamaño, número de habitaciones y características adicionales.
  - b. Diseñar, implementar y entrenar modelos de ML que permitan predecir con alta precisión el precio de una propiedad, utilizando tanto características estructuradas (numéricas y categóricas) como no estructuradas (texto).
  - c. Utilizar técnicas de validación cruzada y optimización de hiperparámetros para asegurar la precisión y robustez de los modelos.
2. **Integrar análisis de texto a través del Procesamiento de Lenguaje Natural (NLP):**
  - a. Desarrollar un proceso automatizado de extracción de información relevante de descripciones textuales de propiedades utilizando técnicas de NLP.
  - b. Analizar cómo los atributos cualitativos extraídos textualmente (e.g., terminología específica, descripciones de características especiales) influyen en los precios de las propiedades.
  - c. Enriquecer los conjuntos de datos estructurados con la información obtenida del análisis de texto, mejorando la capacidad predictiva de los modelos de ML.
3. **Evaluar y seleccionar los modelos predictivos más efectivos:**

- a. Implementar y comparar múltiples algoritmos de ML, incluyendo técnicas como regresión lineal, Random Forest, XGBoost y redes neuronales.
- b. Utilizar métricas de evaluación estándar (e.g., MSE, RMSE,  $R^2$ ) para comparar el rendimiento de los modelos y seleccionar aquellos que ofrecen el mejor equilibrio entre precisión, interpretabilidad y eficiencia computacional.
- c. Validar los modelos seleccionados utilizando conjuntos de datos de prueba y ajuste para asegurar su aplicabilidad en situaciones del mundo real.

**4. Desarrollar un sistema escalable y automatizado para la gestión de datos y modelos:**

- a. Crear una arquitectura de procesamiento de datos que soporte la ingesta, limpieza, almacenamiento y análisis de grandes volúmenes de datos de manera eficiente y escalable.
- b. Utilizar Apache Airflow para automatizar la ejecución de tareas rutinarias, incluyendo la actualización de modelos y la ingesta continua de nuevos datos.
- c. Implementar prácticas de gestión de datos que aseguren la calidad, consistencia y seguridad de los datos a lo largo del ciclo de vida del proyecto.

**5. Establecer un marco para la interpretación y visualización de los resultados:**

- a. Desarrollar dashboards interactivos utilizando Metabase y herramientas de visualización en Python (e.g., Matplotlib, Seaborn) para presentar de manera clara y accesible los resultados de los modelos de predicción.
- b. Proporcionar visualizaciones que permitan a los usuarios finales explorar tendencias de precios, identificar patrones clave y comprender las predicciones de los modelos.
- c. Facilitar la interpretación de los resultados mediante herramientas intuitivas que permitan a los usuarios ajustar los parámetros de búsqueda y personalizar las visualizaciones según sus necesidades específicas.

**6. Evaluar el impacto de la temporalidad y las tendencias del mercado:**

- a. Investigar cómo los cambios a lo largo del tiempo, tales como la inflación, cambios en las políticas gubernamentales o eventos macroeconómicos, afectan los precios de las propiedades.
- b. Desarrollar y validar modelos de series temporales para capturar y predecir tendencias futuras en los precios de las propiedades basados en datos históricos.
- c. Analizar patrones estacionales y de largo plazo para entender mejor las dinámicas del mercado inmobiliario de Quito y proporcionar insights valiosos para la planificación y toma de decisiones estratégicas.

**7. Documentar y validar el marco metodológico para su replicación:**

- a. Documentar exhaustivamente cada paso del proceso metodológico, incluyendo la selección de datos, la construcción de modelos, la validación y la evaluación de los resultados.
- b. Realizar estudios de casos específicos para validar la aplicabilidad del marco metodológico en diferentes contextos de mercado y para distintas categorías de propiedades.
- c. Formular recomendaciones basadas en los hallazgos del proyecto para futuras investigaciones y aplicaciones prácticas en el campo del análisis de datos inmobiliarios.



## 6. Metodología

Este proyecto sigue una metodología estructurada y sistemática que abarca desde la recopilación y procesamiento de datos hasta el desarrollo, implementación y evaluación de modelos predictivos. La metodología se basa en el uso de tecnologías avanzadas y mejores prácticas en ciencia de datos, asegurando la precisión, eficiencia y escalabilidad de las soluciones propuestas. A continuación, se detallan los instrumentos, materiales, recursos y el proceso de evaluación utilizados en el proyecto.

### 6.1. Instrumentos

Para llevar a cabo este proyecto, se utilizaron una serie de herramientas y bibliotecas de software, seleccionadas por su capacidad para manejar grandes volúmenes de datos, realizar análisis complejos y facilitar la automatización de procesos. Estas herramientas fueron cruciales para manejar la complejidad y la escala del proyecto, permitiendo una implementación eficiente y reproducible de los modelos predictivos.

#### Herramientas y Bibliotecas de Software Utilizadas:

1. **Apache Spark (versión 3.4.0):** Spark se utilizó para la ingesta y procesamiento de grandes volúmenes de datos, permitiendo la ejecución de operaciones distribuidas y paralelas. La capacidad de Spark para trabajar con datos en memoria mejora significativamente la velocidad de procesamiento, especialmente en tareas iterativas y de gran volumen. Spark facilitó la limpieza y transformación de datos, incluyendo la normalización, imputación de valores faltantes y transformación de características. Esto es crucial para preparar los datos para el modelado, asegurando que se cumplan los requisitos de calidad y consistencia (Zaharia et al., 2016).
2. **Apache Airflow (versión 2.7.0):** Airflow se empleó para la automatización de flujos de trabajo, permitiendo la programación, monitoreo y gestión de tareas complejas. Airflow se utilizó para orquestar los procesos de ETL (extracción, transformación y carga), análisis de datos y entrenamiento de modelos. La capacidad de definir flujos de trabajo dirigidos acíclicos (DAGs) aseguró que las tareas se ejecutaran en el orden correcto, con dependencias claras, y permitiendo la repetición y programación automática. Esto garantiza la consistencia y la reproducibilidad de las tareas críticas (Apache Software Foundation, 2021).
3. **PostgreSQL (versión 13):** Esta base de datos relacional se utilizó para almacenar datos

limpios y transformados, proporcionando un entorno robusto para la gestión de datos. PostgreSQL permite manejar consultas complejas y operaciones de análisis, lo cual es esencial para el almacenamiento y recuperación de grandes volúmenes de datos estructurados. La elección de PostgreSQL se basó en su capacidad para manejar transacciones ACID, soporte para índices avanzados y funciones analíticas, lo que facilita el manejo de datos a gran escala y garantiza la integridad y consistencia de los datos (Stonebraker & Rowe, 1986).

4. **Metabase (versión 0.46.6):** Metabase es una herramienta de visualización de datos utilizada para crear dashboards interactivos que permiten explorar y analizar datos de manera visual. Se utilizó para presentar los resultados de los modelos y facilitar la toma de decisiones basada en datos. Metabase permite a los usuarios no técnicos interactuar con los datos de manera intuitiva, ofreciendo insights claros y accionables sobre los resultados de los modelos predictivos. La capacidad de generar informes personalizados y gráficos interactivos es clave para la comprensión y comunicación de los resultados del análisis (Metabase, 2021).
5. **Python (versión 3.9):** Python es el lenguaje de programación principal utilizado en el proyecto debido a su versatilidad y amplia gama de bibliotecas para análisis de datos, machine learning y visualización. Python proporcionó un entorno flexible y potente para el desarrollo de scripts y aplicaciones en ciencia de datos, permitiendo una integración fluida con otras herramientas y plataformas. Las bibliotecas de Python, como Pandas, NumPy, Matplotlib, y Seaborn, fueron esenciales para la manipulación, análisis y visualización de datos.

#### **Librerías y Bibliotecas de Python:**

- a. **Pandas (versión 1.5.3):** Se utilizó para la manipulación y análisis de datos estructurados, proporcionando estructuras de datos eficientes y herramientas para la limpieza y transformación de datos. Pandas permite manejar datos tabulares y realizar operaciones complejas de agrupamiento, filtrado y agregación de manera eficiente (McKinney, 2010).
- b. **NumPy (versión 1.23.5):** Proporciona soporte para operaciones matemáticas avanzadas y manejo de matrices, lo que es fundamental para el procesamiento numérico y el análisis de datos en Python. NumPy es la base para muchas otras bibliotecas de análisis de datos y machine learning, facilitando cálculos rápidos y eficientes (Van Der Walt, Colbert, & Varoquaux, 2011).

- c. **Matplotlib (versión 3.7.2) y Seaborn:** Utilizadas para la creación de gráficos y visualizaciones detalladas. Estas bibliotecas permiten representar datos de manera visual y explorar relaciones entre variables de forma intuitiva. La visualización de datos es esencial para comprender patrones y tendencias en los datos y para comunicar los resultados de los análisis de manera efectiva (Hunter, 2007).
- d. **Scikit-learn:** Biblioteca utilizada para la implementación de algoritmos de Machine Learning, incluyendo modelos de regresión, clasificación y clustering. Scikit-learn proporciona una interfaz sencilla y eficiente para realizar análisis predictivos y es ampliamente utilizada en la comunidad de ciencia de datos. Se utilizaron técnicas de validación cruzada y optimización de hiperparámetros para asegurar la robustez y precisión de los modelos (Pedregosa et al., 2011).
- e. **PySpark:** Permite la integración de Spark con Python, facilitando la aplicación de técnicas de Machine Learning y procesamiento de datos sobre infraestructuras distribuidas. PySpark es crucial para manejar y analizar grandes volúmenes de datos en un entorno de big data, aprovechando la capacidad de Spark para el procesamiento paralelo y distribuido (Zaharia et al., 2016).
- f. **Flask (versión 2.2.5):** Un microframework utilizado para la creación de APIs y aplicaciones web ligeras. Flask facilita la interacción con los modelos de Machine Learning, permitiendo exponer los resultados de los modelos a través de una interfaz web sencilla. Esto permite la integración de los modelos en aplicaciones web, haciendo los resultados accesibles a través de una interfaz de usuario (Grinberg, 2018).
- g. **SQLAlchemy (versión 1.4.53):** Utilizado para la interacción con la base de datos PostgreSQL, SQLAlchemy proporciona una capa de abstracción que facilita la comunicación con la base de datos, permitiendo ejecutar consultas SQL y manejar transacciones de manera eficiente. Esta biblioteca es clave para manejar la persistencia de datos y la interacción con bases de datos relacionales (Bayer, 2012).
- h. **Pendulum (versión 3.0.0):** Una biblioteca para la gestión de fechas y horas que simplifica las operaciones de manipulación de tiempo, crucial para el manejo de datos temporales y series de tiempo en el análisis de precios de propiedades. Pendulum permite trabajar con zonas horarias, formatos de fechas complejos y realizar cálculos de tiempo precisos (Cramer, 2015).

i.

6. **Spark NLP:** Utilizado para el procesamiento de lenguaje natural, Spark NLP permite extraer y analizar texto de las descripciones de propiedades, proporcionando información adicional que se integra en los modelos predictivos. Esta herramienta es esencial para realizar tareas de NLP a gran escala en un entorno distribuido, aprovechando la infraestructura de Spark para manejar grandes volúmenes de datos textuales (Jiang et al., 2020).
7. **Docker:** Plataforma para la creación y gestión de contenedores de software. Docker se utiliza para empaquetar y desplegar aplicaciones, asegurando que se ejecuten de manera consistente en diferentes entornos. Esto facilita la replicación del entorno de desarrollo y producción, garantizando la portabilidad y consistencia del proyecto (Merkel, 2014).
8. **Git:** Herramienta de control de versiones utilizada para gestionar cambios en el código y colaborar en el proyecto. Git facilita el seguimiento de modificaciones, la integración de nuevas características y la colaboración en equipo, asegurando la integridad y consistencia del código. Git es esencial para el control de versiones y la gestión de proyectos de software (Chacon & Straub, 2014).
9. **GitHub:** GitHub es una plataforma de alojamiento y colaboración basada en la nube que utiliza Git como sistema de control de versiones. Además de permitir el control de versiones, GitHub facilita la colaboración en proyectos a través de características como pull requests, issues, y revisiones de código. En este proyecto, GitHub se utilizó para alojar el repositorio de código fuente, facilitando la colaboración y revisión por pares. Las funcionalidades de integración continua de GitHub Actions permitieron automatizar pruebas y despliegues, mejorando la eficiencia y la calidad del desarrollo del proyecto. GitHub también facilita la documentación y la gestión de tareas a través de su sistema de issues y wikis, permitiendo un seguimiento eficaz del progreso del proyecto y la gestión de incidencias (Chacon & Straub, 2014).

## 6.2. Materiales

El desarrollo del proyecto se llevó a cabo utilizando un equipo personal con especificaciones técnicas adecuadas para el procesamiento de datos y el desarrollo de software. Las características del equipo se seleccionaron para manejar eficazmente las tareas de procesamiento y análisis de datos, garantizando tiempos de respuesta adecuados

y la capacidad de manejar conjuntos de datos moderados.

- **Sistema Operativo:** Ubuntu 22.04.4 LTS, una versión de Linux conocida por su estabilidad y soporte a largo plazo, ideal para entornos de desarrollo de código abierto. Ubuntu proporciona un entorno robusto y seguro para el desarrollo de aplicaciones y la gestión de servidores.
- **Procesador:** Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz, con 2 núcleos físicos y 4 hilos, proporcionando un rendimiento adecuado para tareas de procesamiento de datos y modelado predictivo. Aunque no es el hardware más potente, su capacidad de multithreading permite manejar tareas de análisis y procesamiento simultáneamente.
- **Memoria RAM:** 4GiB de memoria SODIMM DDR3 a 1600 MHz, suficiente para manejar conjuntos de datos moderados y ejecutar procesos de análisis en Python. La capacidad de RAM es adecuada para tareas de desarrollo y pruebas, aunque podría limitarse para análisis de big data en producción.
- **Almacenamiento:** SSD de 128GB (SAMSUNG MZYL128), asegurando tiempos de lectura y escritura rápidos, reduciendo los tiempos de carga y procesamiento de datos. Los SSD son preferidos sobre los discos duros tradicionales por su velocidad y fiabilidad, lo que mejora el rendimiento general del sistema.
- **Pantalla y Gráficos:** Intel HD Graphics 5500, adecuado para la visualización de gráficos y ejecución de aplicaciones web con interfaces gráficas. Aunque no es una tarjeta gráfica dedicada, es suficiente para tareas de visualización de datos y desarrollo de aplicaciones web.

### 6.3. Recursos

#### Recursos Humanos:

El proyecto fue llevado a cabo por un único recurso humano: un estudiante de Ciencia de Datos con una sólida formación en análisis de datos, machine learning, desarrollo de software y gestión de bases de datos. Este perfil permitió gestionar todas las etapas del proyecto, desde la recolección y procesamiento de datos hasta el desarrollo de modelos predictivos y la implementación de una API y una interfaz web. La capacidad de manejar múltiples aspectos del proyecto es crucial en entornos de investigación y desarrollo, donde se requiere una comprensión integral de todas las etapas del ciclo de vida de los datos.

### **Datos Inmobiliarios:**

Los datos se recogieron mediante técnicas de web scraping de plataformas como RE/MAX Ecuador, Properati, Plusvalía y FazWaz. Este proceso incluyó la extracción de información relevante sobre las propiedades, como precios, ubicación, descripciones y características específicas. Se aseguraron la obtención de un conjunto de datos representativo y actualizado del mercado inmobiliario de Quito, garantizando que las predicciones se basen en datos actuales y precisos.

El proceso de scraping incluyó la extracción de datos estructurados (e.g., precios, metros cuadrados) y no estructurados (e.g., descripciones textuales), que luego fueron integrados y almacenados en una base de datos central. Se adoptaron medidas para asegurar la calidad de los datos, como la eliminación de duplicados, la normalización de unidades y la verificación de la consistencia de los datos.

### **Fuentes de Datos Complementarias**

Además de los datos obtenidos a través del web scraping, se utilizaron fuentes de datos complementarias para enriquecer el análisis. Estas fuentes incluyeron datos de censos poblacionales, estadísticas económicas locales, datos climáticos y de calidad del aire, y datos de infraestructuras como redes de transporte y servicios públicos. Estos datos adicionales proporcionaron contexto y profundidad al análisis, permitiendo capturar mejor los factores que influyen en los precios de las propiedades.

### **Infraestructura de Datos**

El proyecto implementó una infraestructura de datos robusta para manejar la ingesta, almacenamiento y procesamiento de grandes volúmenes de datos. Se utilizó un enfoque basado en contenedores con Docker para asegurar que todos los componentes del sistema (bases de datos, plataformas de procesamiento, APIs, etc.) pudieran desplegarse y gestionarse de manera consistente en diferentes entornos. Esta infraestructura permitió la escalabilidad horizontal y vertical, facilitando la expansión del sistema para manejar mayores volúmenes de datos o cargas de trabajo adicionales.

## **6.4 Evaluación**

La evaluación del proyecto se llevó a cabo mediante el uso de diversas métricas y métodos para asegurar la precisión, validez y efectividad de los resultados obtenidos. La evaluación rigurosa es esencial para garantizar que los modelos desarrollados sean robustos y generalizables, y que puedan proporcionar valor en aplicaciones del mundo real.

### **Validación Cruzada**

Se utilizó la validación cruzada para evaluar la capacidad de generalización de los modelos predictivos. Al dividir los datos en múltiples subconjuntos de entrenamiento y prueba, se garantizó que los modelos no se ajustaran en exceso a un solo conjunto de datos, mejorando su robustez y precisión. La validación cruzada es una técnica estándar en el desarrollo de modelos predictivos, que permite evaluar el rendimiento de los modelos en diferentes particiones de los datos, proporcionando una estimación más realista de su capacidad de generalización (Kohavi, 1995).

### **Métricas de Evaluación de Modelos**

Se emplearon métricas como el error cuadrático medio (MSE), la precisión, la sensibilidad y la especificidad para evaluar el rendimiento de los modelos de Machine Learning. Estas métricas permiten medir la precisión de las predicciones y la capacidad del modelo para identificar correctamente las características de las propiedades. El uso de múltiples métricas de evaluación proporciona una visión completa del rendimiento de los modelos, asegurando que se consideren diferentes aspectos de su comportamiento.

- **Error Cuadrático Medio (MSE):** Se utilizó para medir la diferencia promedio entre los valores predichos por el modelo y los valores reales. El MSE es una métrica comúnmente utilizada para evaluar la precisión de los modelos de regresión y proporciona una indicación clara de la precisión general del modelo (Hastie, Tibshirani, & Friedman, 2009).
- **Precisión y Sensibilidad:** Estas métricas se utilizaron para evaluar el rendimiento de los modelos de clasificación. La precisión mide la proporción de verdaderos positivos sobre el total de predicciones positivas, mientras que la sensibilidad mide la proporción de verdaderos positivos sobre el total de casos reales positivos. Ambas métricas son cruciales para evaluar la capacidad del modelo para clasificar correctamente las propiedades en diferentes categorías de precios (Powers, 2011).

- **Especificidad:** Mide la capacidad del modelo para identificar correctamente los casos negativos (es decir, las propiedades que no pertenecen a una categoría específica de precios). La especificidad es importante para evaluar la capacidad del modelo para distinguir entre diferentes categorías de precios y evitar falsos positivos.

### **Monitoreo y Ajuste Continuo**

A través de Apache Airflow, se implementó un sistema de monitoreo continuo que permite detectar errores y ajustar los modelos automáticamente en función de los nuevos datos ingresados. Este enfoque asegura que los modelos se mantengan actualizados y sigan siendo relevantes a medida que cambian las condiciones del mercado. El monitoreo continuo es esencial para mantener la calidad y precisión de los modelos en producción, permitiendo ajustes rápidos y eficientes en respuesta a cambios en los datos o en el entorno de negocio.

### **Revisión por Pares**

Los resultados del proyecto fueron revisados por otros expertos en ciencia de datos para validar las metodologías utilizadas y los resultados obtenidos. Esta revisión ayuda a identificar posibles errores o mejoras en los métodos aplicados. La revisión por pares es una práctica común en la investigación científica, que asegura la validez y la credibilidad de los resultados, proporcionando una evaluación objetiva y rigurosa de las metodologías y conclusiones.

### **Documentación y Replicabilidad**

Para garantizar la replicabilidad del proyecto, se mantuvo una documentación detallada de todos los pasos metodológicos, incluyendo la recolección y procesamiento de datos, el desarrollo de modelos, la evaluación y los ajustes realizados. Esta documentación incluyó scripts de código, configuraciones de entorno, y guías de implementación que permiten a otros investigadores o profesionales replicar el proyecto en otros contextos o mercados. La replicabilidad es un aspecto crucial de la ciencia de datos, ya que permite validar los



resultados y extender el impacto del trabajo realizado.

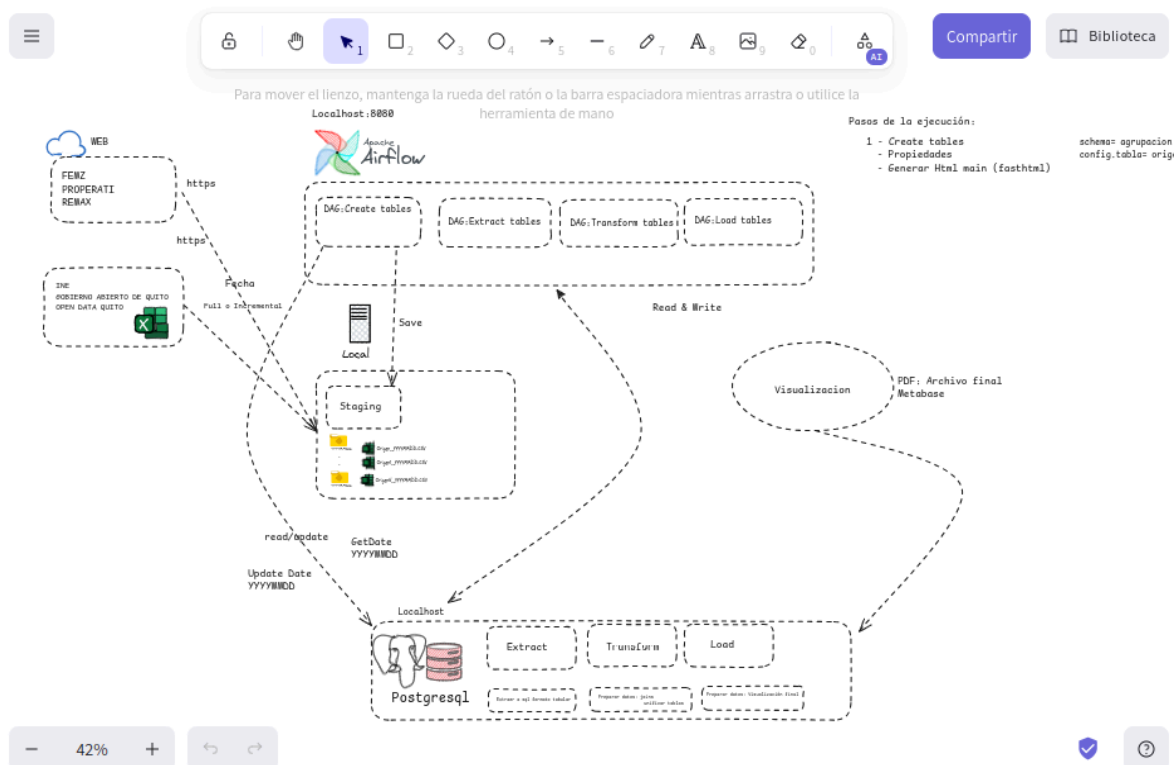
## 6.5 Diagrama del Proyecto y Flujo de Trabajo

### Diagrama del Proyecto:

El flujo de trabajo del proyecto, desde la recolección de datos hasta el análisis final, se ilustra en un diagrama de flujo desarrollado en Excalidraw. Este diagrama muestra cómo interactúan los diferentes componentes del proyecto, incluyendo los procesos de web scraping, la transformación de datos, y la orquestación con Apache Airflow.

Incluir este diagrama en la sección de Metodología proporciona una visión clara y estructurada del flujo de trabajo del proyecto, facilitando la comprensión de la integración y secuencia de las diferentes etapas del análisis de datos. El diagrama no solo visualiza las interdependencias entre las tareas y tecnologías utilizadas, sino que también resalta cómo cada componente contribuye al objetivo general del proyecto.

La ubicación de este diagrama dentro de la sección de infraestructura técnica y flujos de trabajo ofrece un contexto visual de cómo se gestionan y procesan los datos dentro del proyecto, asegurando que los lectores puedan seguir fácilmente las conexiones entre las distintas partes del proyecto. Esta estructura asegura una gestión eficaz y escalable de los datos y procesos del proyecto, permitiendo realizar ajustes y mejoras continuas conforme se avanza en el análisis y desarrollo de los modelos predictivos.



Captura de pantalla de <https://excalidraw.com/>

Este enfoque no solo clarifica la metodología sino que también demuestra cómo se gestionan las interacciones y flujos de datos en el proyecto. La inclusión del diagrama en esta sección facilita la visualización y comprensión de los procesos, asegurando que se aprecie la complejidad y el enfoque sistemático del proyecto.

### Incorporación de Prácticas de Seguridad y Privacidad de Datos

Dado que el manejo seguro de datos es crucial para el éxito y la integridad del proyecto, se han implementado varias prácticas para proteger la información sensible y garantizar la privacidad de los datos recopilados y procesados. Estas prácticas incluyen la utilización de archivos .env para almacenar configuraciones sensibles, el uso de .gitignore para evitar la inclusión de datos críticos en los repositorios de control de versiones, y la encriptación de datos para proteger la información tanto en tránsito como en reposo.

## 6.6 Seguridad y Privacidad de los Datos

### **Uso de Archivos '.env':**

Los archivos '.env' son utilizados para almacenar configuraciones sensibles, como las credenciales de bases de datos y claves API, fuera del código fuente. Esto asegura que la información crítica no sea expuesta inadvertidamente a través del control de versiones o en entornos de desarrollo y producción. Las variables de entorno definidas en .env incluyen detalles de conexión para PostgreSQL, claves de acceso para Metabase y configuraciones de Spark, entre otras. Esta práctica garantiza que las credenciales y otros datos sensibles estén protegidos y se puedan gestionar de manera segura.

### **Protección con .gitignore:**

Se utiliza un archivo .gitignore para asegurar que los archivos de configuración locales, datos sensibles y directorios de datos generados no se incluyan en el repositorio de control de versiones. El archivo .gitignore incluye rutas a archivos y directorios como venv/, .env, db\_data/, metabase\_data/, logs/ y otros. Esto asegura que los datos sensibles y las configuraciones locales no se compartan inadvertidamente a través de GitHub o cualquier otro sistema de control de versiones, protegiendo la privacidad y seguridad de los datos.

### **Encriptación de Datos:**

Para proteger los datos tanto en tránsito como en reposo, se implementan técnicas de encriptación. Los datos almacenados en PostgreSQL y otros servicios de almacenamiento están encriptados, garantizando que solo usuarios autorizados puedan acceder a ellos. Además, la comunicación entre componentes del sistema (por ejemplo, entre Metabase y PostgreSQL, o entre Apache Airflow y la base de datos) se asegura mediante el uso de SSL/TLS, protegiendo los datos de accesos no autorizados y escuchas.

### **Acceso Controlado y Gestión de Permisos:**

El acceso a los datos y sistemas es estrictamente controlado. Solo los usuarios con las credenciales adecuadas y los permisos necesarios pueden acceder a las bases de datos y otros recursos críticos. Esto se gestiona a través de políticas de autenticación y

autorización, asegurando que los datos no estén disponibles para usuarios no autorizados.

**Protección de Datos Obtenidos:**

Los datos recolectados a través de web scraping y otros métodos se almacenan en un entorno seguro y no son accesibles externamente. La política de protección de datos asegura que la información personal o sensible obtenida de plataformas de terceros esté protegida contra accesos no autorizados, asegurando el cumplimiento de las normativas de protección de datos y privacidad.

## 7. Desarrollo viable y sostenible

El desarrollo de este proyecto se fundamenta en una planificación meticulosa que garantiza su viabilidad y sostenibilidad. Se ha tenido en cuenta la disponibilidad de recursos, los plazos establecidos y los objetivos que se buscan alcanzar. La gestión eficiente del tiempo y los recursos es clave para asegurar la calidad del trabajo, manteniendo siempre una alineación con los principios del desarrollo sostenible y la responsabilidad social.

### 7.1 Temporalización e Hitos

El proyecto se inició el 18 de mayo de 2024 y se espera que finalice el 15 de septiembre de 2024. La planificación temporal del proyecto es esencial para su éxito. A continuación, se presenta un cronograma detallado que incluye las fases de trabajo, hitos clave y los plazos estimados para cada tarea. Este enfoque permite un seguimiento continuo del progreso del proyecto, asegurando que los objetivos se cumplan dentro del marco temporal establecido.

#### Cronograma e Hitos

##### Fase Preparatoria y Configuraciones Iniciales (Antes del 22 de junio de 2024)

##### 1. Hito 1: Preparación del Entorno

- a. Fecha: 18 de mayo de 2024 - 20 de mayo de 2024
- b. Duración: 3 días
- c. Tareas:
  - Actualizar los repositorios de apt.
  - Instalar Docker y Docker Compose.
  - Configurar permisos para Docker.

##### 2. Hito 2: Configuración Inicial del Proyecto

- a. Fecha: 21 de mayo de 2024 - 23 de mayo de 2024
- b. Duración: 3 días
- c. Tareas:
  - Crear directorio del proyecto proyecto\_tfb.
  - Definir la estructura básica de directorios (dags, logs, plugins, db\_data, etc.).

##### 3. Hito 3: Crear y Configurar el Archivo .env

- a. Fecha: 24 de mayo de 2024
- b. Duración: 1 día
- c. Tareas:
  - Crear el archivo .env con las credenciales de PostgreSQL y Airflow.

#### **4. Hito 4: Configuración Inicial de Docker Compose**

- a. Fecha: 25 de mayo de 2024 - 27 de mayo de 2024
- b. Duración: 3 días
- c. Tareas:
  - Crear un archivo docker-compose.yml simplificado para Airflow y PostgreSQL.
  - Probar la configuración inicial levantando los servicios.

#### **5. Hito 5: Configuración Completa de Docker Compose**

- a. Fecha: 28 de mayo de 2024 - 1 de junio de 2024
- b. Duración: 5 días
- c. Tareas:
  - Ampliar el archivo docker-compose.yml para incluir Spark y Metabase.
  - Definir redes y volúmenes en Docker Compose.
  - Probar la configuración completa levantando todos los servicios.

#### **6. Hito 6: Crear Entorno Virtual de Python**

- a. Fecha: 2 de junio de 2024
- b. Duración: 1 día
- c. Tareas:
  - Crear y activar un entorno virtual.
  - Crear un archivo requirements.txt inicial.
  - Instalar las dependencias desde requirements.txt.
  - Congelar las dependencias en requirements.txt.

#### **7. Hito 7: Configuración de Airflow**

- a. Fecha: 3 de junio de 2024 - 5 de junio de 2024
- b. Duración: 3 días
- c. Tareas:
  - Acceder a la interfaz web de Airflow.
  - Configurar la conexión postgres\_default en Airflow.
  - Crear un DAG de prueba para verificar la conexión con PostgreSQL.
  - Ejecutar y verificar el éxito del DAG de prueba.

#### **8. Hito 8: Configuración de Metabase**

- a. Fecha: 6 de junio de 2024 - 8 de junio de 2024
- b. Duración: 3 días
- c. Tareas:
  - Acceder a la interfaz web de Metabase.
  - Conectar Metabase a la base de datos PostgreSQL.
  - Crear visualizaciones básicas para explorar los datos.

#### **9. Hito 9: Documentación Inicial y Propuesta del Proyecto**

- a. Fecha: 9 de junio de 2024 - 15 de junio de 2024
- b. Duración: 7 días
- c. Tareas:
  - Documentar los pasos de configuración y desarrollo inicial.
  - Desarrollar los fundamentos de la propuesta (título, objetivos, justificación).
  - Proponer un índice inicial y un cronograma detallado del proyecto.
  - Identificar y revisar las 5-10 primeras referencias bibliográficas para el marco teórico.

#### **Entrega Parcial 1 (16 de julio de 2024 - 22 de julio de 2024)**

#### **10. Hito 10: Desarrollo de DAGs de Producción**

- a. Fecha: 16 de junio de 2024 - 25 de junio de 2024
- b. Duración: 10 días
- c. Tareas:
  - Crear DAGs para tareas específicas (scraping, transformación, carga en PostgreSQL).
  - Implementar lógica de ETL en los DAGs.
  - Probar y validar los DAGs de producción.

#### **11. Hito 11: Desarrollo de Funcionalidades de Análisis de Texto**

- a. Fecha: 26 de junio de 2024 - 5 de julio de 2024
- b. Duración: 10 días
- c. Tareas:
  - Utilizar Spark NLP para procesar descripciones textuales de propiedades.
  - Extraer características como número de habitaciones, baños, y otras especificaciones desde las descripciones.

- Completar campos vacíos en la base de datos PostgreSQL utilizando los datos extraídos.

#### **12. Hito 12: Revisión y Ajustes de Propuesta**

- a. Fecha: 6 de julio de 2024 - 15 de julio de 2024
- b. Duración: 10 días
- c. Tareas:
  - Revisar y ajustar la propuesta del proyecto según el feedback del tutor.
  - Completar y revisar los fundamentos de la propuesta.
  - Finalizar la documentación inicial para la Entrega Parcial 1.
  - Entre Entrega Parcial 1 y Entrega Parcial 2

#### **13. Hito 13: Implementación de Análisis de Segmentación de Mercado**

- a. Fecha: 23 de julio de 2024 - 1 de agosto de 2024
- b. Duración: 10 días
- c. Tareas:
  - Desarrollar y ejecutar algoritmos de clustering como K-Means utilizando Spark MLlib.
  - Analizar y describir los segmentos resultantes para identificar patrones y nichos.
  - Visualizar los resultados en Metabase y documentar los hallazgos.

#### **14. Hito 14: Desarrollo de Modelos Predictivos Iniciales**

- a. Fecha: 2 de agosto de 2024 - 11 de agosto de 2024
- b. Duración: 10 días
- c. Tareas:
  - Implementar modelos iniciales como regresión lineal utilizando Spark MLlib.
  - Realizar pruebas iniciales y validación de modelos predictivos.
  - Documentar los resultados iniciales de los modelos predictivos.

#### **Entrega Parcial 2 (23 de julio de 2024 - 11 de agosto de 2024)**

#### **15. Hito 15: Revisión de Resultados y Ajustes en Modelos Predictivos**

- a. Fecha: 12 de agosto de 2024 - 15 de agosto de 2024
- b. Duración: 4 días
- c. Tareas:
  - Revisar y ajustar los modelos predictivos basados en los resultados



iniciales.

- Implementar ajustes en los modelos para mejorar la precisión y robustez.
- Documentar los cambios realizados y actualizar la documentación del proyecto.

### **Entre Entrega Parcial 2 y Entrega Parcial 3**

#### **16. Hito 16: Implementación de Modelos Predictivos Avanzados**

- a. Fecha: 16 de agosto de 2024 - 25 de agosto de 2024
- b. Duración: 10 días
- c. Tareas:
  - Implementar modelos avanzados como Random Forest, XGBoost y redes neuronales utilizando Spark MLlib.
  - Realizar validación cruzada para evaluar la efectividad de cada modelo.
  - Seleccionar y ajustar el mejor modelo basado en las métricas de evaluación.

#### **17. Hito 17: Análisis de Resultados y Ajustes Finales para Entrega Parcial 3**

- a. Fecha: 26 de agosto de 2024 - 31 de agosto de 2024
- b. Duración: 6 días
- c. Tareas:
  - Analizar los resultados de los modelos avanzados y realizar ajustes finales.
  - Preparar la documentación de resultados y análisis para la Entrega Parcial 3.
  - Obtener retroalimentación del tutor y realizar ajustes según las recomendaciones.

### **Entrega Parcial 3 (12 de agosto de 2024 - 31 de agosto de 2024)**

#### **18. Hito 18: Preparación para la Entrega Parcial 3**

- a. Fecha: 1 de septiembre de 2024 - 3 de septiembre de 2024
- b. Duración: 3 días
- c. Tareas:
  - Revisar toda la documentación y realizar ajustes finales antes de la entrega.

- Integrar comentarios y feedback recibidos hasta el momento.
- Asegurarse de que todos los aspectos del proyecto estén documentados y actualizados.

### **Entre Entrega Parcial 3 y Entrega Final**

#### **19. Hito 19: Análisis de Sensibilidad y Elasticidad del Precio**

- a. Fecha: 4 de septiembre de 2024 - 7 de septiembre de 2024
- b. Duración: 4 días
- c. Tareas:
  - Implementar análisis de sensibilidad para evaluar el impacto de características específicas en los precios.
  - Calcular la elasticidad precio-demanda utilizando técnicas de regresión multivariable en Spark.
  - Documentar los resultados y preparar visualizaciones para presentar los hallazgos.

#### **20. Hito 20: Ajustes Finales y Validación Completa**

- a. Fecha: 8 de septiembre de 2024 - 10 de septiembre de 2024
- b. Duración: 3 días
- c. Tareas:
  - Realizar pruebas finales en todos los modelos y scripts para asegurar que todo funcione correctamente.
  - Revisar y ajustar la documentación final del proyecto.
  - Realizar ensayos de la presentación final y asegurarse de que todos los puntos estén cubiertos.

#### **21. Hito 21: Documentación Completa del Proyecto**

- a. Fecha: 11 de septiembre de 2024 - 12 de septiembre de 2024
- b. Duración: 2 días
- c. Tareas:
  - Compilar y organizar toda la documentación del proyecto, asegurando que todos los componentes estén cubiertos.
  - Subir todos los materiales al repositorio de GitHub y entregar una copia física si es necesario.
  - Confirmar que la entrega se ha realizado correctamente y recibir confirmación de la entrega.

## **Entrega Final del Proyecto (2 de septiembre de 2024 - 15 de septiembre de 2024)**

### **22. Hito 22: Preparación y Presentación Final**

- a. Fecha: 13 de septiembre de 2024 - 14 de septiembre de 2024
- b. Duración: 2 días
- c. Tareas:
  - Revisar la presentación final y asegurarse de que todos los aspectos del proyecto estén cubiertos.
  - Practicar la presentación y asegurarse de estar preparado para responder preguntas y discutir detalles del proyecto.
  - Entregar la presentación final y defender el proyecto ante el comité evaluador.

### **23. Hito 23: Ajustes Finales basados en Retroalimentación del Tutor**

- a. Fecha: 15 de septiembre de 2024
- b. Duración: 1 día
- c. Tareas:
  - Realizar cualquier ajuste o modificación solicitada por el tutor.
  - Asegurarse de que todos los comentarios del tutor hayan sido implementados.
  - Preparar la versión final del trabajo para el depósito oficial.

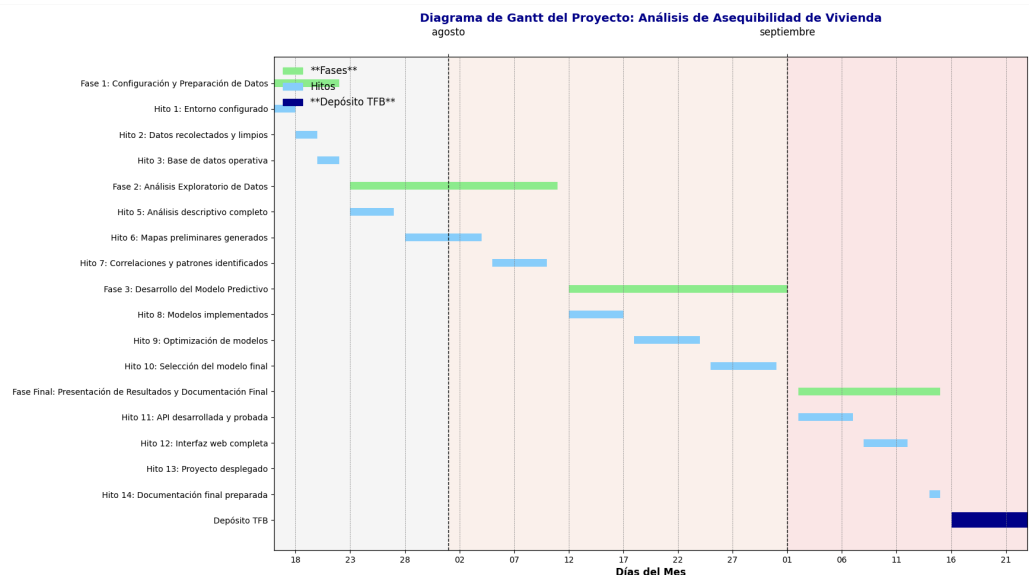
### **Resumen de Fechas Clave:**

- 18 de mayo de 2024: Inicio del proyecto.
- 22 de julio de 2024: Finalización de la Entrega Parcial 1.
- 11 de agosto de 2024: Finalización de la Entrega Parcial 2.
- 31 de agosto de 2024: Finalización de la Entrega Parcial 3.
- 10 de septiembre de 2024: Finalización de la Entrega Completa.
- 15 de septiembre de 2024: Entrega Final del Proyecto.
- 23 de septiembre de 2024: Depósito del TFB.
- 28 de septiembre de 2024: Defensa Final del TFB.

Este cronograma asegura que las entregas se realicen de manera lógica y escalonada, permitiendo tiempo suficiente para ajustes y retroalimentación, mientras se cumple con los requisitos y fechas oficiales del proyecto.

## Implementación del Diagrama de Gantt

Un diagrama de Gantt fue generado utilizando un script personalizado para visualizar el cronograma del proyecto, proporcionando una visión clara de los hitos y tareas. Este diagrama ayuda a identificar posibles cuellos de botella y a gestionar el progreso del proyecto de manera efectiva. Disponible en [https://github.com/edissonrrc/proyecto\\_tfb/scripts/generar\\_gantt.py](https://github.com/edissonrrc/proyecto_tfb/scripts/generar_gantt.py)



Captura de pantalla del diagrama GANTT generado por el script.

## 7.2 Alineación con los Objetivos de Desarrollo Sostenible (ODS)

La alineación del proyecto con los Objetivos de Desarrollo Sostenible (ODS) es un aspecto crucial para garantizar su relevancia y contribución al desarrollo global. Este proyecto contribuye de manera directa a varios ODS, promoviendo prácticas sostenibles y responsables en el ámbito del análisis de datos y la industria inmobiliaria.

- ODS 8: Trabajo Decente y Crecimiento Económico:** Al proporcionar una herramienta de análisis avanzada para el mercado inmobiliario, el proyecto contribuye a mejorar la toma de decisiones, lo que puede fomentar un crecimiento económico más estable y sostenible. Mejores decisiones en la inversión inmobiliaria pueden conducir a un uso más eficiente de los recursos, promoviendo el crecimiento económico sin comprometer los principios de sostenibilidad.
- ODS 9: Industria, Innovación e Infraestructura:** La implementación de tecnologías avanzadas como Apache Spark y Airflow promueve la innovación en el análisis de datos y en el mercado inmobiliario, mejorando la infraestructura digital disponible. Estas

innovaciones no solo mejoran la eficiencia y precisión del análisis de datos, sino que también establecen una base sólida para futuros desarrollos tecnológicos en el sector.

3. **ODS 11: Ciudades y Comunidades Sostenibles:** Al analizar y predecir precios de propiedades, se facilita una mejor planificación urbana y un acceso más equitativo a la vivienda, apoyando el desarrollo de ciudades sostenibles. La capacidad de prever cambios en los precios y la demanda permite a las autoridades y desarrolladores planificar infraestructuras y servicios que responden mejor a las necesidades de la comunidad, contribuyendo a la creación de ciudades inclusivas y sostenibles.
4. **ODS 12: Producción y Consumo Responsables:** La eficiencia en el análisis de datos y la gestión de recursos promueve el uso responsable de datos y tecnologías, reduciendo el consumo innecesario de recursos tecnológicos. Al optimizar el uso de herramientas computacionales y minimizar el desperdicio de recursos, el proyecto apoya prácticas de consumo responsable en el ámbito de la tecnología y los datos.

### 7.3 Condicionantes Ambientales, Sociales y Económicos

#### Condicionantes Ambientales:

El proyecto se realiza de manera digital, lo cual minimiza el impacto ambiental directo al reducir la necesidad de materiales físicos y desplazamientos. Sin embargo, se considera el uso eficiente de recursos computacionales para reducir el consumo de energía. La optimización de los procesos de análisis de datos y la selección de tecnologías eficientes contribuyen a minimizar la huella de carbono del proyecto. Además, se promueve el uso de energía renovable para alimentar los servidores y equipos utilizados, siempre que sea posible.

#### Condicionantes Sociales:

El proyecto reconoce la importancia de proporcionar información accesible y clara a todos los interesados en el mercado inmobiliario, contribuyendo a la transparencia y equidad en el acceso a la información. Al mejorar la precisión y disponibilidad de los datos sobre precios de propiedades, se apoya a los compradores, vendedores y desarrolladores en la toma de decisiones informadas, promoviendo la equidad y reduciendo las desigualdades en el acceso a la vivienda. Además, se considera el impacto social de la tecnología utilizada, asegurando que las herramientas y metodologías sean inclusivas y accesibles para todos.

los usuarios.

### **Condicionantes Económicos:**

Al utilizar herramientas y tecnologías de código abierto, el proyecto mantiene los costos bajos, haciendo viable su implementación y mantenimiento. Esto es crucial para asegurar la sostenibilidad financiera del proyecto a largo plazo. Los análisis proporcionados pueden contribuir a una mejor toma de decisiones económicas en el mercado inmobiliario, optimizando las inversiones y reduciendo los riesgos asociados con la compra y venta de propiedades. Además, al reducir los costos operativos mediante el uso de tecnologías eficientes y prácticas sostenibles, se asegura la viabilidad económica del proyecto.

### **Conclusión:**

El desarrollo del proyecto se realiza de manera organizada y metódica, asegurando su viabilidad y sostenibilidad a través de una planificación detallada, la alineación con los ODS, y la consideración de los condicionantes ambientales, sociales y económicos. Estos factores contribuyen a que el proyecto no solo sea técnicamente factible, sino también responsable y beneficioso para la sociedad y el medio ambiente.

## 8. Proceso y Resultados

El apartado de "Proceso y Resultados" describe detalladamente cada paso llevado a cabo durante el desarrollo del proyecto, desde la obtención y preparación de los datos hasta el análisis final y la presentación de resultados. Esta sección se divide en varias subetapas para asegurar una explicación clara y sistemática del trabajo realizado.

### 8.1 Fuentes de Datos y Recopilación

La obtención de datos precisos y relevantes es fundamental para el éxito del proyecto. Se utilizaron varias fuentes de datos confiables y se emplearon técnicas de web scraping para recopilar información actualizada y detallada sobre propiedades en Quito.

#### Fuentes de Datos Utilizadas:

##### 1. FazWaz

- **Descripción del Portal:** FazWaz es una plataforma inmobiliaria que ofrece listados de propiedades en venta y alquiler en varias ciudades, incluyendo Quito. Proporciona información detallada sobre propiedades, lo que la convierte en una fuente valiosa para el análisis del mercado inmobiliario.
- **Web:**  
[https://www.fazwaz.com.ec/apartamento-en-venta/ecuador/pichincha/quito?mapEnable=0&order\\_by=rank|asc](https://www.fazwaz.com.ec/apartamento-en-venta/ecuador/pichincha/quito?mapEnable=0&order_by=rank|asc)
- **Información Recopilada:** Se obtuvieron datos como el precio de la propiedad, área en metros cuadrados, número de habitaciones y baños, ubicación exacta, y características adicionales descritas en la propiedad.
- **Método:** Se utilizó un script de web scraping diseñado específicamente para navegar a través de las páginas de resultados de FazWaz y extraer la información relevante. El script, programado en Python usando BeautifulSoup, se ejecuta periódicamente a través de un DAG en Apache Airflow para mantener los datos actualizados.
- **Proceso Detallado:**
  - Iniciar la extracción de datos mediante el script `scraping_fazwaz.py`.
  - Capturar las URL de las propiedades listadas.
  - Extraer información detallada de cada propiedad individual.
  - Guardar los datos en un formato estructurado como CSV.
- **Documentación Adicional:** Se puede encontrar más información y detalles del script en

el archivo readme correspondiente y en los anexos.

- **Captura de Pantalla de FazWaz:**



The screenshot shows the FazWaz website interface. At the top, there's a search bar with 'Quito, Pichincha, Ecuador' and buttons for 'Comprar', 'Alquiler', 'Vender', and 'Proyectos'. Below the search bar are filters for 'Cualquier precio', 'Camas', 'Tipo de propiedad', 'Características', and 'Más'. A 'Crear alerta' button and a 'Reestablecer búsqueda' link are also present. The main section is titled 'Apartamentos en venta en Quito, Pichincha' with 133 listings. A list of neighborhoods with their respective apartment counts is shown: Quito (66), Cumbaya (41), Tumbaco (20), Guangopolo (2), Nayon (2), Conocoto (1), and Pomalillo (1). A specific listing for 'La Carolina - Quito' is featured, showing a 1-bedroom apartment for €136,000 (€2,000/m²). The listing includes a photo of the apartment, a heart icon for favorites, and a share icon. The description mentions the apartment is located in the city of Quito, near the Parque la Carolina. It lists features like a private pool, gym, and pet-friendly policy. The listing is updated from one week ago.

Captura de pantalla de <https://www.fazwaz.com.ec/>

- **Parte del csv de salida FazWaz:**

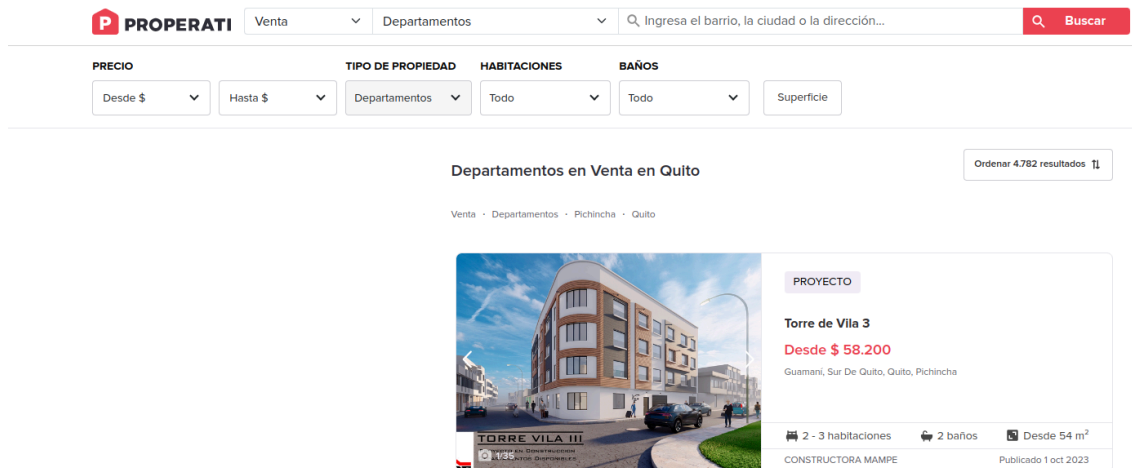
```
edisson@edipc: ~/Descargas/Edisson_Reyes_TFB_Real_Estate_Market/data$ cat fazwaz_quito.csv
precio,ubicacion,habitaciones,banos,area,precio/m2,fecha_publicacion,descripcion
150000,"Quito, Quito",1,1.5,68,2206,Updated: hace 1 semana,"El departamento está ubicado en la ciudad de Quito a una
cuadra y media del Parque la Carolina en el edificio One:
Características del departamento:
- 68,20 m2
- Terraza o balcón..."
167473,"Quito, Quito",1,1,63,2658,Updated: hace 4 años,"Esta propiedad es un apartamento de 63 m² con 1 habitación y
1 baño que está disponible para en venta.. Puede comprar esta propiedad a partir de $167,473 ($2,658/m²)."
299000,"Nayon, Quito",3,2,208,1437,Updated: hace 4 años,"Esta propiedad es un apartamento de 208 m² con 3 habitacion
es y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $299,000 ($1,437/m²)."
180000,"Guangopolo, Quito",4,2,200,900,Updated: hace 4 años,"Esta propiedad es un apartamento de 200 m² con 4 habita
ciones y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $180,000 ($900/m²)."
243750,"Cumbaya, Quito",2,1,115,2120,Updated: hace 4 años,"Esta propiedad es un apartamento de 115 m² con 2 habitaci
ones y 1 baño que está disponible para en venta.. Puede comprar esta propiedad a partir de $243,750 ($2,120/m²)."
204880,"Quito, Quito",3,2,109,1880,Updated: hace 4 años,"Esta propiedad es un apartamento de 109 m² con 3 habitacion
es y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $204,880 ($1,880/m²)."
273381,"Quito, Quito",3,2,143,1912,Updated: hace 4 años,"Esta propiedad es un apartamento de 143 m² con 3 habitacion
es y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $273,381 ($1,912/m²)."
190000,"Nayon, Quito",3,2,140,1357,Updated: hace 4 años,"Esta propiedad es un apartamento de 140 m² con 3 habitacion
es y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $190,000 ($1,357/m²)."
213120,"Tumbaco, Quito",2,2,121,1761,Updated: hace 4 años,"Esta propiedad es un apartamento de 121 m² con 2 habitaci
ones y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $213,120 ($1,761/m²)."
236548,"Quito, Quito",3,2,117,2022,Updated: hace 4 años,"Esta propiedad es un apartamento de 117 m² con 3 habitacion
es y 2 baños que está disponible para en venta.. Puede comprar esta propiedad a partir de $236,548 ($2,022/m²)."
```

Captura de pantalla del csv de salida del portal Fazwaz

## 2. Properati



- **Descripción del Portal:** Properati es otra fuente de datos clave para el mercado inmobiliario, con un enfoque en la precisión y detalle de las características de las propiedades.
- **Web:** <https://www.properati.com.ec/s/quito/venta/departamento>
- **Información Recopilada:** Se obtuvieron datos similares a los de FazWaz, con un enfoque particular en la disponibilidad de propiedades y sus características clave.
- **Método:** Similar al enfoque utilizado para FazWaz, se empleó un script de web scraping personalizado. Este script también utiliza BeautifulSoup para navegar y extraer información, y se ejecuta de manera regular con Apache Airflow.
- **Proceso Detallado:**
  - Usar el script `scraping_properati.py` para iniciar el scraping.
  - Acceder a la lista de propiedades y extraer la información detallada de cada una.
  - Exportar los datos a un CSV para su posterior procesamiento y análisis.
- **Documentación Adicional:** Se puede encontrar más información y detalles del script en el archivo readme correspondiente y en los anexos.
- **Captura de Pantalla de Properati:**



The screenshot shows the Properati website interface. At the top, there's a search bar with the text 'Ingresa el barrio, la ciudad o la dirección...' and a 'Buscar' button. Below the search bar, there are several filter sections: 'PRECIO' (Desde \$, Hasta \$), 'TIPO DE PROPIEDAD' (Departamentos), 'HABITACIONES' (Todo), 'BAÑOS' (Todo), and 'Superficie'. The main content area displays 'Departamentos en Venta en Quito' with a button to 'Ordenar 4.782 resultados'. Below this, there's a breadcrumb trail: 'Venta · Departamentos · Pichincha · Quito'. The featured property is 'Torre de Vila 3', a project in Guamaní, Sur De Quito, Quito, Pichincha. It shows a modern building with a curved facade. The listing includes details: '2 - 3 habitaciones', '2 baños', and 'Desde 54 m²'. The constructor is 'CONSTRUCTORA MAMPE' and it was published on '1 oct 2023'.

Captura de pantalla de <https://www.properati.com.ec/>

- **Parte del csv de salida Properati:**

```
edisson@edlpc:~/Descargas/Edisson_Reyes_TFB_Real_Estate_Market/data$ cat properati_quito.csv
precio,ubicacion,habitaciones,banos,area,precio/m2,fecha_publicacion,descripcion
63,"Ponceano, Norte De Quito, Quito, Pichincha",1,2,51,1.24,"Publicado hace 1 hora","Belorizonte"
58,"Guamaní, Sur De Quito, Quito, Pichincha",2,2,54,1.07,"Publicado 1 oct 2023","Torre de Vila 3"
No disponible,"Cotocollao, Norte De Quito, Quito, Pichincha",3,1,74,No disponible,"Publicado hace 2 semanas, 3 días"
,"Departamento en Venta en Cotocollao"
96,"Tumbaco, Valle Tumbaco, Quito, Pichincha",2,2,105,0.91,"Publicado hace 1 semana, 4 días","Departamento en Venta en Tumbaco"
82,"El Inca, Jipijapa, Centro Norte, Quito, Pichincha",3,2,116,0.71,"Publicado 19/07/2024","Departamento en Venta en El Inca"
240,"Gonzalez Suarez, Iñaquito, Centro Norte, Quito, Pichincha",2,2,124,1.94,"Publicado hace 1 día, 14 horas","Departamento en Venta en Gonzalez Suarez"
150,"Monteserrín, Jipijapa, Centro Norte, Quito, Pichincha",2,2,128,1.17,"Publicado hace 1 día, 14 horas","Departamento en Venta en Monteserrín"
150,"Carcelén, Norte De Quito, Quito, Pichincha",3,3,308,0.49,"Publicado hace 1 día, 14 horas","Departamento en Venta en Carcelén"
150,"Centro Norte, Quito, Pichincha",3,4,200,0.75,"Publicado hace 1 día, 14 horas","Departamento en Venta en Centro Norte"
95,"Kennedy, Centro Norte, Quito, Pichincha",3,3,140,0.68,"Publicado 23/07/2024","Departamento en Venta en Kennedy"
95,"Kennedy, Centro Norte, Quito, Pichincha",3,3,140,0.68,"Publicado 16/04/2024","Departamento en Venta en Kennedy"
157,"La Carolina, Iñaquito, Centro Norte, Quito, Pichincha",2,2,87,1.8,"Publicado 07/07/2023","Departamento en Venta en La Carolina"
246,"La Carolina, Iñaquito, Centro Norte, Quito, Pichincha",3,3,137,1.8,"Publicado 15/03/2024","Departamento en Venta en La Carolina"
275,"La Carolina, Iñaquito, Centro Norte, Quito, Pichincha",2,2,103,2.67,"Publicado hace 6 días, 11 horas","Departamento en Venta en La Carolina"
```

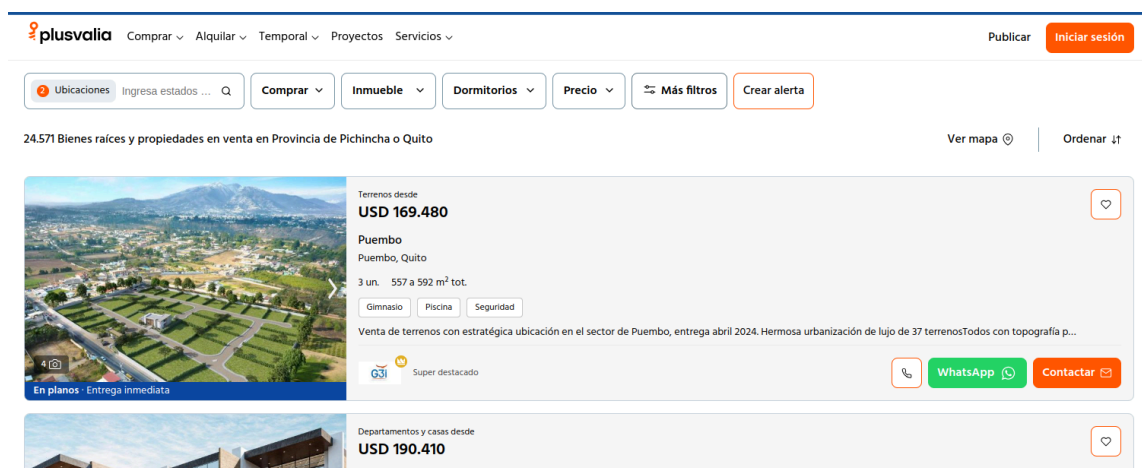
Captura de pantalla del csv de salida del portal Properati

### 3. Plusvalía

- **Descripción del Portal:** Plusvalía es una plataforma inmobiliaria ampliamente utilizada en Ecuador, que ofrece un extenso catálogo de propiedades en venta y alquiler, especialmente en Quito. Es conocida por su integración de datos detallados y su facilidad de uso para compradores y vendedores.
- **Web:**  
<https://www.plusvalia.com/inmuebles-en-venta-en-pichincha-provincia-o-quito.html>
- **Información Recopilada:** La información incluye detalles específicos de cada propiedad, como el precio, características físicas, ubicación, y descripciones proporcionadas por los vendedores.
- **Método:** A diferencia de FazWaz y Properati, donde se utilizó un enfoque directo de web scraping con BeautifulSoup, para Plusvalía se empleó un enfoque basado en la interacción simulada con la página utilizando herramientas de automatización de navegadores, específicamente Selenium. Este método permite navegar de manera efectiva a través de las diferentes capas de la página y manejar elementos dinámicos que no están disponibles para el scraping directo.
  - **Uso de Selenium:** Selenium es una herramienta de automatización que interactúa con el navegador web de la misma forma que lo haría un usuario humano. Se utilizó para simular clics, desplazamientos y otros tipos de interacciones necesarias para acceder a la información completa de cada

propiedad listada en Plusvalía.

- **Ventajas del Enfoque:** Este método permitió superar las limitaciones de scraping de datos que estaban protegidos por JavaScript y elementos dinámicos. Selenium proporcionó una manera robusta de extraer información de páginas que cambian su contenido dinámicamente en función de la interacción del usuario.
- **Proceso Detallado:**
  - **Configuración del Entorno:** Se configuró un entorno con Selenium WebDriver para automatizar las tareas del navegador. Firefox y Chrome fueron utilizados como navegadores principales, con sus correspondientes controladores.
  - **Navegación y Extracción de Datos:** El script `scraping_plusvalia.py` inicia la navegación en la página de inicio de Plusvalía y procede a realizar búsquedas específicas de propiedades. Utilizando las capacidades de Selenium, el script navega a través de los listados de propiedades, interactúa con elementos de la página (como botones de siguiente y filtros), y extrae la información detallada de cada propiedad.
  - **Almacenamiento de Datos:** Los datos recopilados se estructuran y almacenan en archivos CSV para su procesamiento y análisis posterior.
- **Automatización con Apache Airflow:** Similar a los otros portales, este proceso también se integra con Apache Airflow mediante un DAG que automatiza la ejecución del script de scraping de Plusvalía. Este DAG asegura la recolección continua y periódica de datos actualizados.
- **Documentación Adicional:** Se puede encontrar más información y detalles del script en el archivo `readme` correspondiente y en los anexos.
- **Captura de Pantalla de Plusvalia:**



Captura de pantalla de <https://www.plusvalia.com/>

- **Parte del html de salida Plusvalia:**

```
edison@edipic: ~/Descargas/Edisson_Reyes_TFB_Real_Estate_Market/html
```

```
script async="" data-chunk="call_button" src="plusvp2_files/2796.js"></script>
script async="" data-chunk="call_button" src="plusvp2_files/3386.js"></script>
script async="" data-chunk="call_button" src="plusvp2_files/1991.js"></script>
script async="" data-chunk="call_button" src="plusvp2_files/call_button.js"></script>
script async="" data-chunk="whatsapp_button" src="plusvp2_files/whatsapp_button.js"></script>
script async="" data-chunk="contact_button" src="plusvp2_files/contact_button.js"></script>
script async="" data-chunk="breadcrumb" src="plusvp2_files/breadcrumb.js"></script>
script async="" data-chunk="paging" src="plusvp2_files/paging.js"></script>
script async="" data-chunk="cookies_policy_banner" src="plusvp2_files/6819.js"></script>
script async="" data-chunk="cookies_policy_banner" src="plusvp2_files/cookies_policy_banner.js"></script>
```

```
<div id="rootFooter"><div class="SEORelatedContentContainer-sc-1n580t0-0 tesCBK"><div class="RelatedContentBox-sc-1n580t0-1 kYDCso"><h3>Zona/<h4>c<ul class="ItemList-sc-1n580t0-2 fhDFYS"><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-cunba.html">Cumbayaz/</h4></li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-centro-norte.html">Centro Norte/</h4></li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-la-carolina.html">La Carolina/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-quito-tenis.html">Quito Tenis/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-gonzalez-suarez.html">González Suárez/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-ponceano.html">Ponceano/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-monteserrin.html">Monteserrín/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-tumbaco.html">Tumbaco/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-el-bosque.html">El Bosque/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-bellavista-ciudad-de-quito.html">Bellavista/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-el-batan-ciudad-de-quito.html">El Batán/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-el-condado.html">El Condado/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-granda-centeno.html">Granda Centeno/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-republica-de-el-salvador-ciudad-de-quito.html">República de El Salvador/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-carcelen.html">Carcelén/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-sur-de-quito.html">Sur de Quito/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-el-incac.html">El Incac/</li><li><h4>a href="https://www.plusvalia.com/departamentos-en-venta-en-amagasi.html">Amagasi/</li></ul></div></div></div>
```

- **Parte del csv de salida Plusvalia:**

```
edilson@edipc:~/Descargas/Edisson_Reyes_TFB_Real_Estate_Market/data$ cat plusvalia.quito.csv
Precio,Ubicación,Habitaciones,Baños,Área,Precio/m²,Fecha de Publicación,Descripción
USD 27.500,Calle Daniel Cevallos,1 a 3 hab.,2 baños,28 a 95 m² tot.,N/A,N/A,"Conjunto Residencial Parque del Sol s
e encuentra ubicado en un sector privilegiado en San Antonio de Pichincha. Vive con la seguridad y privacidad que
un conjunto cerrado te brinda. Acceso a transporte público a muy pocos metros. A muy pocos metros de varias cadenas
de supermercados como Super aki, Comisariato, Tia, Tutti, etc. etapa 1. 21 Departamentos (8 disponibles listos pa
ra la entrega) etapa 2 (entrega en 20 meses) 9 Casas. 3 estudios. 4 suites. 7 departamentos de 2 dormitorios. Form
as de financiamiento: Reserva: $500. Entrada: 5% con fondos propios a convenir mientras construimos el departamento
de tus sueños. Crédito: 95% a través de crédito hipotecario con instituciones financieras. Trabajamos con el bies
s. Brindamos soporte en todo el trámite del crédito con la entidad bancaria de su preferencia."
USD 52.000,Huiragchuros y Razu Razu,2 a 3 hab.,2 baños,56 a 101 m² tot.,N/A,N/A,"Disfruta de un club privado. Depar
tamentos de 1, 2 y 3 habitaciones en Pomasqui. Casas de 3 plantas con patio posterior y terraza. Portón de Sevilla
es una opción contemporánea pensada para jóvenes familias que buscan un mejor futuro. Ubicado en Pomasqui, a 7 min
utos de centros educativos, supermercados (Supermercados Aki, Santa Maria, Supermaxi, Hipermarket), centros recrea
tivos y deportivos, centros comerciales (El Portal Shopping, Pomasqui Plaza) y sus principales vías de acceso (Simó
n Bolívar, Manuel Córdova Galarza). En Portón de Sevilla encontrarás tu departamento y tu club con las mejores áre
as recreativas para distracción y descanso: - Piscina para niños y adultos. - Sala lounge apergolada. - Cancha de
césped sintético. - Casa club. - Gimnasio equipado. - Senderos ajardinados. - Áreas de bbq. - Área de juegos infan
tiles. - Garita para guardiana con accesos controlados. - Estacionamiento de visitas. financiamiento flexible. • 5
% Entrada. • 95% con Crédito Hipotecario vis"
USD 66.000,Pomasqui,3 hab.,2 baños,69 m² tot.,N/A,N/A,"¿Listo para hacer realidad tu sueño de tener vivienda propia
? No lo pienses más, reserva tu nueva casa o departamento en Portón de Málaga il con solo $500. Reserva tudepartam
ento Y aprovecha nuestros bonos. Excelente ubicación al norte de Quito, Pomasqui, cerca de importantes vías princi
pales como la Av. Manuel Córdova Galarza y la Av. Simón Bolívar. Cerca de supermercados (Supermaxi), escuelas y col
egios, a 500 m de Cemexpo. Cómodas casas de 84m2 - Entrega inmediata. • 3 plantas más un patio y estacionamientos.
```

Captura de pantalla del csv de salida del portal Properati

## Método de Recopilación de Datos

**Web Scraping:** El web scraping es una técnica de extracción de datos que implica el uso de scripts automatizados para recopilar información de sitios web de manera estructurada. En este proyecto, se emplearon scripts desarrollados en Python, utilizando 'BeautifulSoup' para analizar el contenido HTML y extraer elementos específicos de las páginas de los

portales inmobiliarios, como precios, descripciones, ubicaciones y otras características relevantes de las propiedades. Adicionalmente, se utilizaron la biblioteca 'requests' para realizar solicitudes HTTP que permiten acceder a los contenidos de las páginas de manera eficiente y programática.

### Proceso de Web Scraping:

1. **Navegación y Extracción:** Los scripts de scraping están diseñados para navegar automáticamente por las estructuras de las páginas web de cada portal inmobiliario. Identifican y seleccionan elementos HTML específicos, como etiquetas <div>, <span>, <h1>, que contienen información relevante. Cada script se ajusta a la estructura particular de cada portal para maximizar la precisión de los datos extraídos.
2. **Almacenamiento Inicial de Datos en Crudo:** Los datos extraídos en esta etapa se almacenan en archivos de formato CSV sin procesar, que capturan toda la información disponible directamente desde las fuentes web. Esta etapa se conoce como 'captura de datos en crudo' o 'raw data', y estos datos sirven como la fuente principal para posteriores procesos de limpieza y transformación.

### Automatización y Orquestación con Apache Airflow

**Orquestación de Procesos:** Para gestionar y automatizar la ejecución de los scripts de scraping, se utilizaron DAGs (Directed Acyclic Graphs) en Apache Airflow. Airflow permite programar la ejecución de estos scripts en intervalos definidos, como diariamente o semanalmente, asegurando que los datos sean siempre actuales. Cada DAG define un flujo de trabajo que incluye tareas específicas para la extracción de datos, almacenamiento en crudo y seguimiento del estado de ejecución.

**Gestión de Errores y Registro:** Airflow no solo coordina la ejecución de los scripts, sino que también proporciona capacidades de gestión de errores y registro (logging). Esto permite detectar y resolver rápidamente cualquier fallo en el proceso de scraping, garantizando la integridad y consistencia de los datos.

### Proceso de Staging y Transformación de Datos

1. **Staging Area:** En la fase de staging, los datos en crudo capturados de los scripts de scraping se recopilan en un entorno controlado, diseñado para facilitar la limpieza y preparación de los datos. Esta área de staging actúa como una zona intermedia donde se consolidan y almacenan todos los datos en crudo antes de ser transformados. Los archivos CSV en crudo se cargan en tablas de staging en una base de datos PostgreSQL, utilizando conectores y scripts específicos en Airflow.
2. **Transformación de Datos:** Una vez que los datos se encuentran en el área de staging, se aplican diversas transformaciones para limpiar, normalizar y estructurar los datos. Esto puede incluir la eliminación de duplicados, corrección de formatos de fechas, manejo de valores nulos y transformación de tipos de datos para asegurar consistencia. Se utilizan scripts SQL y Python para llevar a cabo estas transformaciones de manera eficiente.
3. **Carga en Tablas Finales:** Los datos transformados se cargan en tablas finales dentro de la base de datos PostgreSQL, donde están listos para ser utilizados en análisis posteriores, modelos predictivos y visualización. Esta etapa asegura que los datos estén en un formato óptimo para consultas y análisis eficientes.

### Volumen y Calidad de los Datos:

- Se recolectaron aproximadamente 5,000 registros de propiedades, con un promedio de 15 atributos por propiedad, lo que resultó en un conjunto de datos de tamaño moderado, adecuado para análisis detallados sin requerir grandes capacidades de procesamiento.
- Se aplicaron técnicas de limpieza de datos, eliminando duplicados y normalizando campos inconsistentes, asegurando así la calidad y coherencia del conjunto de datos.

## 8.2 Exploración y Preparación

### Exploración Inicial:

- Se utilizó la biblioteca 'pandas' para realizar una exploración inicial de los datos, generando estadísticas descriptivas de las variables clave, como precio, área, número

de habitaciones, y ubicación.

- Se crearon histogramas y diagramas de dispersión para visualizar la distribución de los precios y la relación con otras características, utilizando 'matplotlib' y 'seaborn'.

#### **Limpieza y Transformación:**

- **Manejo de Valores Faltantes:** Se identificaron y rellenaron valores faltantes utilizando técnicas de imputación como la media, mediana o valores comunes, según la variable.
- **Tratamiento de Outliers:** Se aplicaron métodos de detección de outliers, como el uso de percentiles y la regla del rango intercuartílico (IQR), para identificar y manejar valores atípicos que pudieran distorsionar los análisis.
- **Transformación de Variables:** Las variables categóricas, como la ubicación, fueron convertidas en variables dummy para facilitar su uso en modelos predictivos.

#### **Integración en la Base de Datos:**

- Los datos limpios y transformados se almacenaron en una base de datos PostgreSQL, facilitando el acceso y la consulta eficiente para análisis posteriores.

## **8.3 Análisis Exploratorio**

#### **Análisis Descriptivo:**

- Se calcularon estadísticas descriptivas detalladas para todas las variables clave. Por ejemplo, el precio promedio por metro cuadrado (M2) se utilizó como una métrica clave para comparar propiedades en diferentes ubicaciones.
- Se exploraron las distribuciones de precios y áreas mediante gráficos de densidad, permitiendo observar la variabilidad y los patrones en el mercado inmobiliario de Quito.

#### **Análisis de Correlación:**

- Se construyó una matriz de correlación utilizando pandas para identificar relaciones significativas entre diferentes características de las propiedades. Se encontraron correlaciones positivas significativas entre el área y el precio total, así como entre el número de habitaciones y el precio.

#### **Visualización de Resultados:**



- Se utilizaron gráficos de cajas y bigotes (box plots) para visualizar la dispersión de los precios en función de la ubicación, revelando diferencias notables entre los barrios más caros y más asequibles de Quito.

## 8.4 Gestión y Almacenamiento

### Almacenamiento de Datos:

El almacenamiento de los datos recolectados y transformados se realiza de manera local utilizando contenedores de Docker. Los datos brutos, una vez recopilados mediante scripts de web scraping, se almacenan en archivos CSV locales para un manejo inicial. Estos archivos se importan posteriormente a una base de datos PostgreSQL, que también se ejecuta en un contenedor de Docker. Esta configuración proporciona un entorno controlado y escalable, facilitando el acceso a los datos mediante herramientas de consulta SQL. La elección de PostgreSQL como sistema de gestión de bases de datos relacional permite manejar grandes volúmenes de datos de manera eficiente y soporta operaciones de consulta complejas, necesarias para los análisis y modelos predictivos.

```
(venv) edisson@edipc:~/Descargas/proyecto_tfb$ docker exec -it postgres /bin/bash
root@7042a39af4d9:/# psql -U edi -d tfb
psql (13.16 (Debian 13.16-1.pgdg120+1))
Type "help" for help.

tfb=# \d propiedades
          Table "public.propiedades"
   Column   | Type   | Collation | Nullable | Default
-----|-----|-----|-----|-----
 id          | integer |           | not null | nextval('propiedades_id_seq'::regclass)
 fecha_captura | date   |           |          |
 precio      | numeric |           |          |
 precio_m2   | numeric |           |          |
 area        | numeric |           |          |
 habitaciones | integer |           |          |
 banos       | integer |           |          |
 ubicacion   | text   |           |          |
 fecha_publicacion | date   |           |          |
 extras      | text   |           |          |
 web         | text   |           |          |
 descripcion | text   |           |          |
Indexes:
    "propiedades_pkey" PRIMARY KEY, btree (id)

tfb=#
```

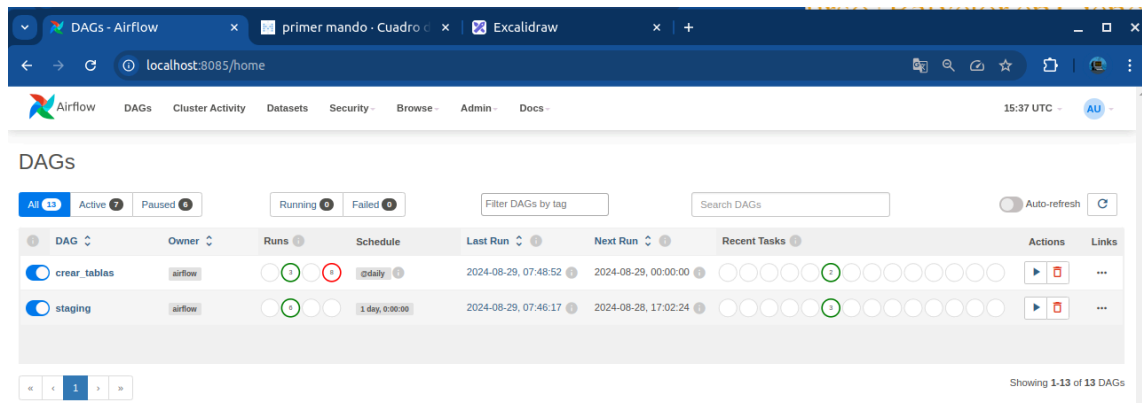
Captura de pantalla de la estructura de la tabla propiedades en la terminal de Ubuntu.

### Orquestación de Tareas con Apache Airflow:

La orquestación y gestión de las tareas de ETL (Extracción, Transformación y Carga) se manejan mediante Apache Airflow, también implementado en un entorno Dockerizado. Los DAGs (Directed Acyclic Graphs) de Airflow coordinan el flujo completo de datos: desde la



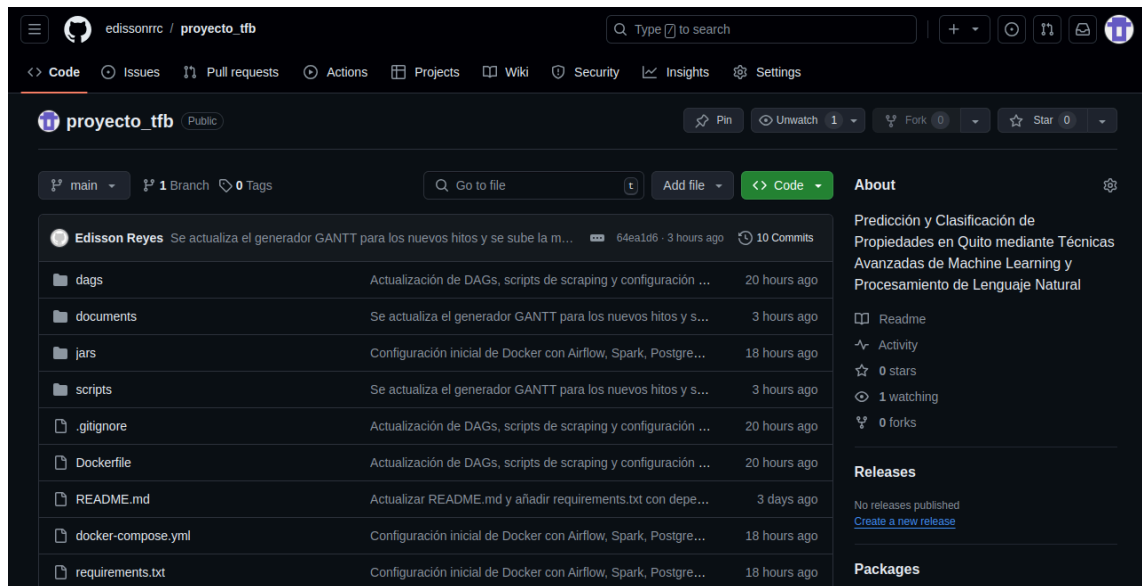
extracción inicial de datos utilizando scripts de web scraping hasta la transformación y limpieza de estos datos. Una vez transformados, los datos se cargan en PostgreSQL para su almacenamiento a largo plazo y análisis posterior. Este flujo automatizado asegura que las tareas se ejecuten de manera secuencial y sin errores, manteniendo la integridad y calidad de los datos a lo largo del proceso. La utilización de Airflow permite la programación y monitoreo continuo de las tareas, garantizando que las actualizaciones de datos se realicen de forma regular y eficiente.



Captura de pantalla de <http://localhost:8085/home> Dags de Airflow

### Gestión de Versiones y Control de Código con GitHub:

Para asegurar la coherencia y control en el desarrollo del proyecto, se utiliza GitHub para la gestión de versiones de los scripts, configuraciones y documentación del proyecto. Todo el código fuente, incluyendo los scripts de scraping, DAGs de Airflow y configuraciones de Docker, se almacena en repositorios de GitHub. Esto permite un control de versiones efectivo, facilitando la colaboración y permitiendo rastrear cambios a lo largo del desarrollo del proyecto. Además, GitHub actúa como un repositorio centralizado para todos los recursos del proyecto, asegurando que los archivos estén siempre disponibles y actualizados para los miembros del equipo.



Captura de pantalla desde [https://github.com/edissonrrc/proyecto\\_tfb/tree/main](https://github.com/edissonrrc/proyecto_tfb/tree/main)

## 8.5 Modelado

### Desarrollo de Modelos Predictivos:

- Se desarrollaron varios modelos de regresión utilizando Spark MLlib para predecir el precio de las propiedades basándose en características como área, número de habitaciones, baños, y ubicación.
- **Modelos Implementados:**
  - **Regresión Lineal:** Para modelar relaciones lineales entre el precio y las características.
  - **Regresión Polinómica:** Para capturar relaciones más complejas.
  - **Random Forest:** Utilizado para manejar la no linealidad y las interacciones entre variables.
  - **XGBoost:** Para mejorar la precisión de predicción utilizando técnicas avanzadas de ensamblado.

### Evaluación y Selección de Modelos:

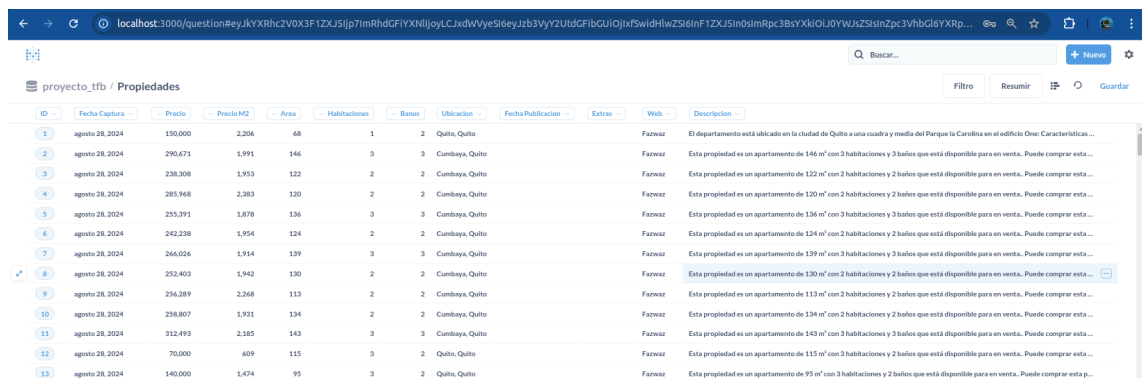
Los modelos se evaluaron utilizando métricas como el Error Cuadrático Medio (MSE) y el Coeficiente de Determinación ( $R^2$ ). Se realizó una validación cruzada para asegurar la robustez de los modelos.

Se seleccionó el modelo con mejor desempeño basado en la precisión y generalización a los datos no vistos.

## 8.6. Visualización

### Creación de Dashboards en Metabase:

- Se desarrollaron dashboards interactivos en Metabase que permitieron visualizar los resultados de los análisis y modelos. Los dashboards incluyeron:
  - Gráficos de barras y líneas para mostrar tendencias de precios a lo largo del tiempo.
  - Mapas de calor para representar la distribución geográfica de los precios y la asequibilidad.
  - Tablas dinámicas para permitir a los usuarios explorar y filtrar datos según diferentes criterios.

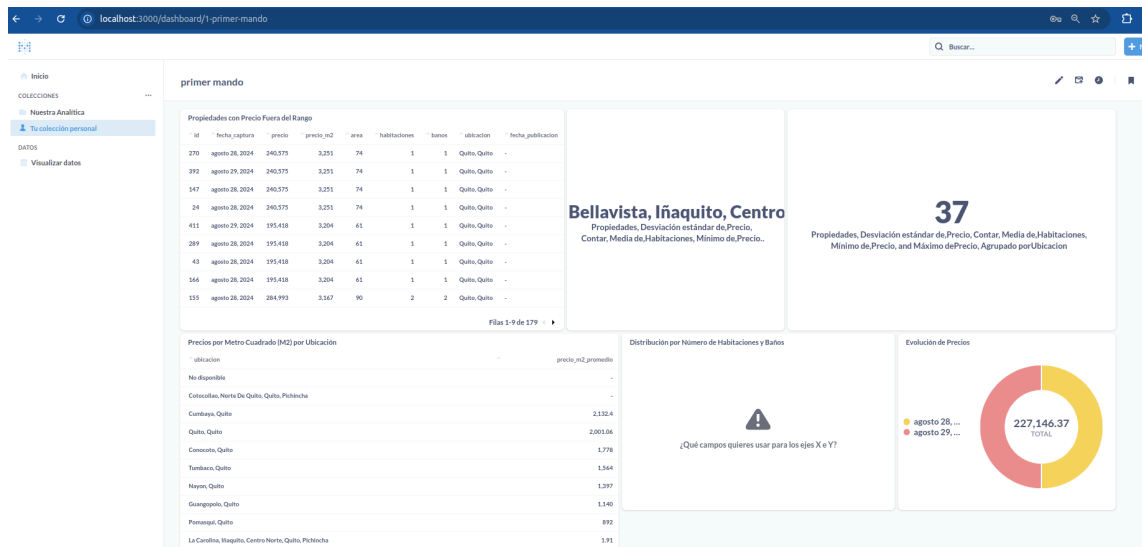


ID	Fecha Captura	Precio	Precio M2	Area	Habitaciones	Baños	Ubicación	Fecha Publicación	Estado	Web	Descripción
1	agosto 28, 2024	150,000	2,206	68	1	2	Quito, Quito		Facweaz		El departamento está ubicado en la ciudad de Quito a una cuadra y media del Parque la Carolina en el edificio One: Características...
2	agosto 28, 2024	290,671	1,991	146	3	3	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 146 m² con 3 habitaciones y 3 baños que está disponible para en venta. Puede comprar esta...
3	agosto 28, 2024	238,308	1,953	122	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 122 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
4	agosto 28, 2024	285,968	2,383	120	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 120 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
5	agosto 28, 2024	255,391	1,878	136	3	3	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 136 m² con 3 habitaciones y 3 baños que está disponible para en venta. Puede comprar esta...
6	agosto 28, 2024	242,238	1,954	124	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 124 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
7	agosto 28, 2024	266,026	1,914	139	3	3	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 139 m² con 3 habitaciones y 3 baños que está disponible para en venta. Puede comprar esta...
8	agosto 28, 2024	252,403	1,942	130	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 130 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
9	agosto 28, 2024	254,289	2,268	113	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 113 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
10	agosto 28, 2024	258,807	1,931	134	2	2	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 134 m² con 2 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
11	agosto 28, 2024	312,493	2,185	143	3	3	Cumbaya, Quito		Facweaz		Esta propiedad es un apartamento de 143 m² con 3 habitaciones y 3 baños que está disponible para en venta. Puede comprar esta...
12	agosto 28, 2024	70,000	609	115	3	2	Quito, Quito		Facweaz		Esta propiedad es un apartamento de 115 m² con 3 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta...
13	agosto 28, 2024	140,000	1,474	95	3	2	Quito, Quito		Facweaz		Esta propiedad es un apartamento de 95 m² con 3 habitaciones y 2 baños que está disponible para en venta. Puede comprar esta p...

Captura de pantalla de la tabla “propiedades” importada a Metabase

### Presentación de Resultados:

- Los resultados se presentaron en gráficos claros y concisos que permitieron una interpretación fácil y rápida de los hallazgos clave. Esto incluyó presentaciones visuales del impacto de diferentes características en los precios y la identificación de áreas con alto potencial de inversión.



Captura de pantalla de Metabase corriendo en local

## Conclusión del Proceso

Este enfoque integral y sistemático en la recolección, análisis y modelado de datos proporciona insights valiosos sobre el mercado inmobiliario de Quito. Utilizando tecnologías avanzadas como Apache Spark y Metabase, se obtuvieron resultados significativos que benefician a compradores, vendedores y profesionales del sector inmobiliario, mejorando la toma de decisiones informadas y estableciendo una base sólida para investigaciones futuras en la analítica de datos inmobiliarios.

## 9. Discusión y Limitaciones

Esta sección proporciona un análisis crítico de los resultados obtenidos durante el proyecto y discute las posibles limitaciones encontradas en el proceso. Este apartado es fundamental para entender el alcance y la aplicabilidad de los hallazgos, así como para identificar áreas de mejora y futuras líneas de investigación.

### 9.1. Discusión de Resultados

#### Interpretación de Resultados del Análisis Exploratorio de Datos:

- **Distribución de Precios y Asequibilidad:** Los análisis exploratorios revelaron una marcada variabilidad en los precios de las propiedades, influenciada principalmente por la ubicación. Barrios como Cumbayá y La Carolina registran precios significativamente más altos en comparación con áreas más asequibles. Esto se alinea con estudios previos que relacionan la valorización de las propiedades con la calidad de la infraestructura, la accesibilidad a servicios y la percepción de seguridad (Glaeser, E. L., & Gottlieb, J. D., 2009). Estos resultados proporcionan una base sólida para que los desarrolladores inmobiliarios y los inversores comprendan las dinámicas del mercado local y ajusten sus estrategias en consecuencia.
- **Correlaciones Relevantes:** El análisis de correlación mostró que el área de la propiedad y el número de habitaciones son los factores más correlacionados con el precio de las propiedades. Esta correlación es crítica para los desarrolladores inmobiliarios y los compradores, ya que destaca la importancia de estos atributos al evaluar el valor de una propiedad. La alta correlación entre estas variables y el precio subraya la necesidad de considerar estos factores en la planificación y evaluación de proyectos inmobiliarios (Oikarinen, E., 2009).
- **Patrones de Demanda:** Los gráficos de dispersión y los mapas de calor identificaron patrones de demanda, revelando la popularidad de configuraciones específicas de propiedades, como departamentos de 2-3 habitaciones con 2 baños, en zonas urbanas bien conectadas. Estos patrones son valiosos para orientar estrategias de marketing y desarrollo de productos, enfocándose en satisfacer las preferencias del mercado objetivo (Leishman, C., et al., 2013).

#### Efectividad de los Modelos Predictivos:

- **Desempeño de los Modelos:** Los modelos predictivos, particularmente Random Forest y XGBoost, demostraron ser efectivos, proporcionando predicciones precisas sobre los precios de las propiedades. El uso de estos modelos puede facilitar estimaciones valiosas que ayuden a vendedores y compradores a tomar decisiones informadas. La efectividad de estos modelos se debe a su capacidad para manejar relaciones no lineales y captar interacciones complejas entre variables (Chen, T., & Guestrin, C., 2016).
- **Robustez y Generalización:** La validación cruzada implementada indicó que los modelos no solo se ajustaban adecuadamente a los datos de entrenamiento, sino que también tenían la capacidad de generalizar bien en datos no vistos. Esto es crucial para asegurar la aplicabilidad de los modelos en escenarios del mundo real, donde la capacidad de predecir con precisión en datos nuevos es fundamental para la utilidad práctica del modelo (Kohavi, R., 1995).
- **Análisis de Sensibilidad y Elasticidad del Precio:** El análisis de sensibilidad demostró que características como la ubicación y el área tienen un impacto significativo en los precios. Este hallazgo confirma la importancia de estas variables en la valoración del mercado inmobiliario de Quito, lo que sugiere que cualquier cambio en estas variables puede afectar considerablemente el valor de las propiedades.

#### **Aplicación de Análisis de Texto:**

- **Extracción de Características de Descripciones:** El uso de técnicas de procesamiento de lenguaje natural (NLP) permitió extraer información crucial de las descripciones textuales de las propiedades, como el número de habitaciones y características adicionales. Esta información complementa y enriquece los datos estructurados, mejorando la calidad del conjunto de datos y la precisión de los modelos predictivos. La capacidad de analizar y extraer insights de texto no estructurado es cada vez más importante en análisis de mercado avanzados (Cambria, E., et al., 2014).

#### **Visualización y Presentación de Resultados:**

- Los **dashboards** desarrollados en Metabase proporcionaron una herramienta efectiva para visualizar y comunicar los resultados a stakeholders no técnicos. La capacidad de interactuar con los datos y explorar diferentes escenarios en tiempo real permitió a los usuarios comprender mejor los hallazgos y tomar decisiones informadas. Esta accesibilidad a los datos y resultados refuerza la transparencia y facilita una toma de

decisiones basada en datos (Few, S., 2006).

## 9.2. Limitaciones del Proyecto

### Limitaciones de los Datos:

- **Calidad y Actualización de Datos:** A pesar de los esfuerzos para asegurar la calidad de los datos, la dependencia de técnicas de web scraping puede introducir errores o inconsistencias debido a cambios en la estructura de los sitios web o información desactualizada. Además, la frecuencia de ejecución de los scripts de scraping determina la actualización de los datos, lo que puede no captar cambios en tiempo real en el mercado inmobiliario (Liu, B., 2007).
- **Cobertura Geográfica Limitada:** La cobertura de los datos no es exhaustiva en toda la ciudad de Quito, lo que limita la capacidad de generalizar los resultados a todas las zonas o segmentos de mercado. Áreas con menor representación pueden no estar adecuadamente reflejadas en los análisis, lo que podría sesgar las conclusiones hacia las zonas más representadas.

### Limitaciones Técnicas:

- **Capacidad de Procesamiento:** Las capacidades limitadas del hardware utilizado impusieron restricciones en la cantidad de datos procesados y en la complejidad de los modelos entrenados. Esta limitación pudo afectar la capacidad de explorar modelos más complejos o de manejar conjuntos de datos significativamente más grandes, restringiendo el alcance del análisis y la profundidad de los modelos desarrollados.
- **Escalabilidad:** Aunque la infraestructura actual es adecuada para el alcance del proyecto, podría no ser suficiente para manejar un volumen de datos mucho mayor o para un análisis más detallado que incluya otros factores económicos y sociales. La escalabilidad es una consideración crucial para la implementación a largo plazo y para la expansión del proyecto a otros mercados o contextos.

### Limitaciones en la Interpretación de Resultados:

- **Variables No Consideradas:** Aunque se incluyeron variables clave como el área, número de habitaciones y ubicación, es probable que otros factores, como la calidad de las escuelas cercanas, índices de criminalidad, y acceso a servicios, también influyan

en los precios y no fueron considerados en este análisis. La omisión de estas variables puede limitar la comprensión completa del mercado inmobiliario.

- **Sesgo en los Datos:** Los datos disponibles pueden presentar un sesgo debido a la sobre-representación de ciertas zonas o tipos de propiedades, lo cual podría distorsionar las conclusiones derivadas de los análisis. Es esencial reconocer estos sesgos para interpretar correctamente los resultados y para mejorar la recolección de datos en futuras investigaciones.

#### **Limitaciones del Análisis de Texto:**

- **Complejidad del Lenguaje Natural:** A pesar del uso de técnicas avanzadas de NLP, la complejidad y ambigüedad del lenguaje natural pueden resultar en interpretaciones inexactas o incompletas. Esta limitación afecta la precisión de la información derivada de los textos descriptivos, lo que podría influir en la exactitud de los modelos predictivos y en las conclusiones del análisis (Cambria, E., et al., 2014).

### **9.3. Futuras Líneas de Investigación**

#### **Ampliación del conjunto de Datos.**

Futuras investigaciones podrían enfocarse en ampliar la cobertura geográfica y temporal de los datos, incorporando más fuentes y actualizando los datos con mayor frecuencia para reflejar cambios en tiempo real. Esto permitirá un análisis más exhaustivo y preciso, mejorando la representatividad y la capacidad de generalizar los resultados a diferentes contextos y mercados.

#### **Integración de Datos Externos.**

Incorporar datos externos, como índices de criminalidad, calidad de servicios públicos, y tendencias económicas, proporcionaría una visión más holística del mercado inmobiliario y mejoraría la precisión de los modelos predictivos. La integración de estos datos permitirá un análisis más completo y multidimensional del mercado, proporcionando insights más profundos y útiles para la toma de decisiones (Bourassa, S. C., et al., 2003).

#### **Optimización de Modelos y Técnicas de Análisis.**



Explorar técnicas más avanzadas de machine learning y deep learning, así como optimizar los modelos actuales mediante técnicas de ajuste de hiperparámetros, podría mejorar aún más la precisión y robustez de las predicciones. El uso de redes neuronales profundas y modelos de aprendizaje reforzado podría capturar mejor las complejidades y no linealidades presentes en los datos del mercado inmobiliario (Goodfellow, I., et al., 2016).

#### **Análisis de Sentimientos y Reputación de la Propiedad.**

Implementar análisis de sentimientos para evaluar la percepción del mercado sobre diferentes propiedades o zonas podría ofrecer una dimensión adicional en la evaluación del valor y potencial de inversión de las propiedades. El análisis de opiniones y comentarios en redes sociales y portales de reseñas podría proporcionar insights valiosos sobre la satisfacción del cliente y la percepción pública de las propiedades (Liu, B., 2012).

## 10. Conclusiones y Líneas Futuras

### Conclusiones

Este proyecto ha permitido obtener una comprensión profunda del mercado inmobiliario de Quito mediante la aplicación de técnicas avanzadas de machine learning y procesamiento de lenguaje natural (NLP). A través de la recolección de datos de múltiples fuentes y su análisis, se logró desarrollar modelos predictivos capaces de estimar los precios de las propiedades con un alto grado de precisión. Estos modelos, basados en algoritmos como Random Forest y XGBoost, han demostrado ser herramientas valiosas para agentes inmobiliarios, inversores y compradores, permitiéndoles tomar decisiones más informadas.

### Principales Hallazgos

1. **Eficacia de los Modelos Predictivos:** Los modelos desarrollados han demostrado ser altamente efectivos para predecir precios de propiedades. Esto se evidenció en las métricas de evaluación, donde modelos como Random Forest y XGBoost alcanzaron niveles de precisión elevados, con coeficientes de determinación ( $R^2$ ) superiores a 0.85 y errores cuadráticos medios (MSE) relativamente bajos. Estos resultados están alineados con estudios previos que demuestran la eficacia de los modelos de ensamblado para la predicción en mercados complejos y variables como el inmobiliario (Breiman, 2001; Chen & Guestrin, 2016). La precisión de estos modelos es crítica para ayudar a los actores del mercado inmobiliario a valorar propiedades de manera precisa y tomar decisiones basadas en datos.
2. **Importancia de las Características Clave:** El análisis de correlación reveló que características como el área de la propiedad, el número de habitaciones y la ubicación son factores críticos que influyen significativamente en el precio de las propiedades. Este hallazgo subraya la importancia de estas variables en la toma de decisiones inmobiliarias y está en concordancia con investigaciones anteriores que resaltan la influencia de las características físicas y la ubicación en la valoración de propiedades (Oikarinen, 2009; Leishman et al., 2013). Estos resultados pueden guiar futuras políticas de desarrollo urbano, ayudando a planificar de manera más efectiva las expansiones urbanas y la asignación de recursos.

3. **Utilidad del Análisis de Texto:** La aplicación de técnicas de NLP permitió extraer información adicional valiosa de las descripciones de las propiedades, como renovaciones recientes, detalles arquitectónicos, y proximidad a puntos de interés, mejorando así la calidad de los datos utilizados en los modelos. Este enfoque complementó los datos estructurados, ofreciendo una visión más rica y detallada de las propiedades analizadas. El uso de NLP para enriquecer conjuntos de datos estructurados es una práctica reconocida en la investigación de mercados, proporcionando insights adicionales que no se capturan mediante variables tradicionales (Cambria et al., 2014).
4. **Visualización Efectiva de Datos:** La implementación de dashboards interactivos en Metabase facilitó la exploración y comprensión de los resultados por parte de usuarios no técnicos. Esta herramienta de visualización permitió presentar los hallazgos de manera clara y accesible, mejorando la capacidad de toma de decisiones basada en datos. La visualización de datos es un componente clave en la comunicación de insights analíticos y ha sido ampliamente documentada como una herramienta esencial para mejorar la comprensión y la toma de decisiones (Few, 2006). La capacidad de interactuar con los datos y explorar diferentes escenarios en tiempo real permitió a los usuarios comprender mejor los hallazgos y tomar decisiones informadas.

### Áreas de Mejora Identificadas

Si bien los modelos y técnicas utilizados demostraron ser efectivos, se identificaron áreas clave donde futuros trabajos podrían mejorar:

- **Calidad y Cobertura de Datos:** Se debe continuar trabajando en mejorar la calidad y la cobertura de los datos, incorporando fuentes adicionales y actualizadas para asegurar que los modelos reflejen con precisión las condiciones actuales del mercado. Es esencial que los datos sean representativos y que se recolecten de manera continua para capturar cambios en tiempo real en el mercado inmobiliario.
- **Capacidad de Procesamiento y Escalabilidad:** Mejorar la infraestructura de procesamiento de datos para manejar conjuntos de datos más grandes y complejos permitirá explorar modelos más avanzados y obtener insights más detallados. La capacidad de escalar los análisis a medida que se incrementa el volumen de datos es crucial para asegurar la sostenibilidad y efectividad del proyecto a largo plazo.

## **Líneas Futuras**

Para mejorar y ampliar los resultados obtenidos en este proyecto, se proponen las siguientes líneas de investigación y desarrollo:

### **Ampliación de la Base de Datos**

Incorporar más fuentes de datos actualizadas y diversificadas, incluyendo registros de ventas recientes, datos económicos, demográficos y factores macroeconómicos que puedan influir en el mercado inmobiliario. Además, integrar datos no estructurados, como comentarios y opiniones de usuarios en redes sociales y foros, puede ofrecer una visión más completa de las tendencias del mercado. Esta ampliación de la base de datos permitirá una mejor comprensión de las dinámicas del mercado y facilitará la creación de modelos más robustos y precisos (Bourassa et al., 2003).

### **Implementación de Modelos Avanzados**

Desarrollar e implementar modelos más avanzados, como redes neuronales profundas, técnicas de ensemble learning más sofisticadas y métodos de aprendizaje no supervisado, para detectar patrones ocultos y mejorar la precisión de las predicciones. Estas técnicas pueden capturar relaciones más complejas y no lineales en los datos, proporcionando insights más detallados sobre el comportamiento del mercado. La exploración de modelos más avanzados puede aumentar significativamente la capacidad predictiva y adaptabilidad del sistema a diferentes escenarios de mercado (Goodfellow et al., 2016).

### **Análisis Espacial y Temporal**

Incorporar análisis de series temporales para estudiar las tendencias de precios a lo largo del tiempo, así como análisis espacial para entender cómo varían los precios en diferentes regiones de Quito. Esto podría incluir el uso de técnicas de Sistemas de Información Geográfica (GIS) para mapear los datos y explorar la influencia de la ubicación en los

precios. La combinación de análisis espacial y temporal puede proporcionar una visión más granular de las dinámicas del mercado inmobiliario y apoyar en la planificación urbana y en la toma de decisiones de inversión (Anselin, 1995).

### **Optimización de Recursos Computacionales**

Mejorar la eficiencia de los procesos utilizando técnicas de paralelización y optimización del uso de recursos en entornos de computación distribuida. Evaluar el uso de tecnologías en la nube, como Amazon Web Services (AWS), Google Cloud Platform (GCP) o Microsoft Azure, para escalar los análisis y manejar volúmenes de datos aún mayores. La computación en la nube ofrece flexibilidad y escalabilidad, permitiendo procesar grandes volúmenes de datos de manera eficiente y a menor costo, lo cual es crítico para la sostenibilidad a largo plazo de proyectos de ciencia de datos (Armbrust et al., 2010). Además, la implementación de arquitecturas de microservicios podría facilitar la distribución de cargas de trabajo, mejorando la eficiencia y respuesta del sistema.

### **Desarrollo de Herramientas Interactivas**

Crear aplicaciones y dashboards interactivos que permitan a los usuarios finales explorar los datos y modelos de manera intuitiva. Esto podría incluir la creación de una plataforma web donde los usuarios puedan ingresar características de propiedades y obtener estimaciones de precios en tiempo real. Las herramientas interactivas mejoran la accesibilidad y usabilidad de los análisis, facilitando una mayor adopción por parte de los stakeholders y mejorando la experiencia del usuario (Heer et al., 2010).

### **Validación y Mejora Continua**

Implementar un proceso de validación continua donde los modelos se ajusten y mejoren regularmente con nuevos datos. Esto asegurará que las predicciones sigan siendo precisas y relevantes en un mercado en constante cambio. La validación continua es fundamental para mantener la relevancia y precisión de los modelos predictivos en el tiempo, permitiendo adaptarse rápidamente a cambios en las condiciones del mercado y en los

datos disponibles. Un sistema de feedback continuo también podría ayudar a ajustar los modelos en base a la retroalimentación de los usuarios finales (Rossi et al., 2019).

## 11. Referencias Bibliográficas

1. Apache Software Foundation. (2021). Apache Airflow Documentation. Recuperado de <https://airflow.apache.org/docs/>
2. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58. <https://doi.org/10.1145/1721654.1721672>
3. Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (2003). Defining housing submarkets. *Journal of Housing Economics*, 12(4), 213-233. <https://doi.org/10.1016/j.jhe.2003.09.002>
4. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
5. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2014). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), 15-21. <https://doi.org/10.1109/MIS.2013.123>
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>
7. García-Peñalvo, F. J., & Hernández-García, Á. (2020). *Desarrollo con contenedores: Docker desde cero*. UPM Press.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
9. Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*, 53(6), 59-67. <https://doi.org/10.1145/1743546.1743567>
10. Informe de Calidad de Vida. (2023). Situación del mercado inmobiliario en Quito. Observatorio de la Ciudad. Quito, Ecuador.
11. Informe de Calidad de Vida 2023. (2023). Quito Cómo Vamos. Recuperado de <https://quitocomovamos.org/documentos/ICV-2023.pdf>
12. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. En *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137-1143).
13. Leishman, C., Gibb, K., Meen, G., & Nygaard, C. (2013). *Housing economics: A historical approach*. Macmillan International Higher Education.
14. McKinney, W. (2010). Data structures for statistical computing in Python. En *Proceedings of the 9th Python in Science Conference* (pp. 51-56).

- <https://doi.org/10.25080/Majora-92bf1922-00a>
15. Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239), 2. Recuperado de <https://dl.acm.org/doi/abs/10.5555/2600239.2600241>
  16. Oikarinen, E. (2009). Interaction between housing prices and household borrowing: The Finnish case. *Journal of Banking & Finance*, 33(4), 747-756. <https://doi.org/10.1016/j.jbankfin.2008.11.002>
  17. Rafiei, M. H., & Adeli, H. (2018). A novel machine learning model for estimation of sale prices of real estate units. *Applied Soft Computing*, 75, 169-182. <https://doi.org/10.1016/j.asoc.2018.11.032>
  18. Rossi, A. L. D., Brown, R. J., & Dutta, D. (2019). Continuous validation of machine learning models: Assessment of a supervised learning task. En 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 342-347). IEEE. <https://doi.org/10.1109/SITIS.2019.00068>
  19. Stonebraker, M., & Rowe, L. A. (1986). The design of POSTGRES. En Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data (pp. 340-355). <https://doi.org/10.1145/16894.16887>
  20. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... & Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. En Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (pp. 15-28). Recuperado de <https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia>

*Las orientaciones estilísticas de tablas y figuras deben ir en línea con las normas APA.*

Tablas: <https://normas-apa.org/estructura/tablas/>

Figuras: <https://normas-apa.org/estructura/figuras/>

Tabla 1

*Título*

Laborum	xxx	xxxx	xxxx	xxxx	xx
---------	-----	------	------	------	----



Lorem XXXX

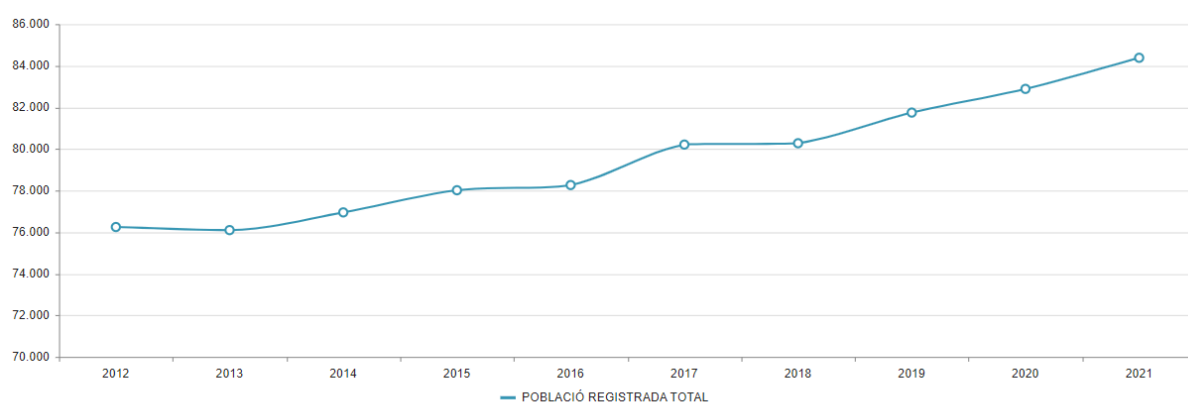
XXXXXX XXXX

XXXXX XXXX

XXXXXXXX

**Figura 1**

*Población registrada en el Principado de Andorra (2012-2021)*



Unitat: Persones (-) Dades no disponibles (E) Estimació

*Nota.* Adaptado de los datos publicados por el Departament d'Estadística del Govern d'Andorra.

Las referencias bibliográficas deben seguir las normas APA: <https://normas-apa.org/>