

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

CARRERA: INGENIERIA EN SISTEMAS DE LA INFORMACIÓN

Trabajo de Titulación como requisito previo para la obtención del título de ingeniero en sistemas de información

Desarrollo de un modelo predictivo basado en machine learning para la estimación de precios del mercado inmobiliario en la ciudad de quito mediante la extracción automatizada de datos web

Autor: Fausto Adrian Chancusig Nacimba
Director: Ing. Francisco Rodríguez Clavijo

QUITO, JUNIO 2026

DEDICATORIA

A mis padres que siempre me han brindado su apoyo sin importar las dificultades y siempre me pusieron como su prioridad.

A mi familia que siempre se dio el tiempo de preguntar cómo iba con mis estudios.

Cada uno es un pilar muy importante en mi vida que espero siempre tener presente.

.

AGRADECIMIENTO

Agradezco en especial a mis padres, quienes todos los días se preocuparon y estuvieron pendientes de mis estudios, siempre apoyándome y velando por mi futuro. Sé el esfuerzo, tiempo y sacrificio que hicieron para que pueda continuar con mi estudio, y quiero que sepan que siempre voy a tener eso presente en mi vida.

Agradezco a mi familia y, en especial, a mis abuelitos, que siempre se dieron el tiempo de preguntar sobre mis estudios, me alentaron y me impulsaron a mejorar y dar lo mejor de mí.

Gracias, mamá y papá, por haberme dado el privilegio de poder estudiar una carrera universitaria. Los amo.

RESUMEN

La presente disertación tiene como propósito el desarrollo de un modelo predictivo basado en algoritmos de Machine Learning para la estimación objetiva de precios del mercado inmobiliario en la ciudad de Quito, utilizando información obtenida mediante técnicas de extracción automatizada de datos web. El mercado inmobiliario quiteño se caracteriza por su dinamismo y por una marcada asimetría de información entre los actores involucrados, lo que dificulta la toma de decisiones económicas fundamentadas.

La metodología empleada se fundamentó en el estándar CRISP-DM y comprendió las siguientes etapas: la recolección automatizada de datos desde los principales portales inmobiliarios ecuatorianos; la limpieza, transformación y estructuración del conjunto de datos; el entrenamiento y comparación de distintos algoritmos de aprendizaje supervisado, entre ellos Regresión Lineal Múltiple, Árboles de Decisión, Bosques Aleatorios y Redes Neuronales; y finalmente, la evaluación del rendimiento de los modelos mediante métricas de error estándar para seleccionar la arquitectura con mayor precisión.

Como resultado se espera obtener una herramienta computacional funcional capaz de estimar el precio de mercado de un inmueble a partir de sus características estructurales y geográficas, contribuyendo así a reducir la asimetría de información del sector y aportando una base tecnológica escalable para futuras aplicaciones en el ámbito de la tecnología aplicada a la propiedad.

Palabras clave: Machine Learning, web scraping, modelo predictivo, mercado inmobiliario, Quito, aprendizaje supervisado, regresión, ciencia de datos.

ABSTRACT

The purpose of this dissertation is the development of a predictive model based on Machine Learning algorithms for the objective estimation of real estate prices in the city of Quito, using information obtained through automated web data extraction techniques. The Quito real estate market is characterized by its dynamism and by a marked information asymmetry among the actors involved, which hinders well-founded economic decision-making.

The methodology was based on the CRISP-DM standard and comprised the following stages: the automated collection of data from the main Ecuadorian real estate portals; the cleaning, transformation, and structuring of the dataset; the training and comparison of different supervised learning algorithms, including Multiple Linear Regression, Decision Trees, Random Forests, and Neural Networks; and finally, the evaluation of model performance through standard error metrics to select the most accurate architecture.

As a result, a functional computational tool was obtained, capable of estimating the market price of a property from its structural and geographic characteristics, thereby helping to reduce the sector's information asymmetry and providing a scalable technological foundation for future applications in the field of property technology.

Keywords: Machine Learning, web scraping, predictive model, real estate market, Quito, supervised learning, regression, data science.

CONTENIDO

DEDICATORIA ..... **II**

AGRADECIMIENTO ..... **III**

RESUMEN ..... **IV**

ABSTRACT ..... **V**

CONTENIDO ..... **VI**

ÍNDICE DE ILLUSTRACIONES ..... **IX**

ÍNDICE DE TABLAS ..... **IX**

CAPÍTULO I: INTRODUCCIÓN **1**

1.1. Planteamiento del problema **1**

1.2. Justificación **2**

1.3. Objetivos **3**

1.3.1. Objetivo general **3**

1.3.2. Objetivos específicos **3**

1.4. Alcance **4**

CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA **5**

2.1. Marco teórico **5**

2.1.1. Antecedentes y estado del arte **5**

2.1.2. Teoría de la valoración inmobiliaria **5**

2.1.3. Web Scraping **6**

2.1.4. Aprendizaje automático supervisado **6**

2.2. Marco conceptual **7**

2.2.1. Machine Learning o Aprendizaje Automático **7**

2.2.2. Web Scraping **7**

2.2.3. Modelo Predictivo **7**

2.2.4. Dataset o Conjunto de Datos **8**

2.2.5. Mercado Inmobiliario **8**

2.2.6. Quito **8**

2.2.7. Inmueble **8**

2.2.8. Plataformas Inmobiliarias 9
2.2.9. Properati 9
2.2.10. Plusvalía 10
..... 11
CAPÍTULO III: DESARROLLO DEL TRABAJO 12
3.1. Metodología de Trabajo 12
3.2. Fase 1: Comprensión del Negocio 14
3.3. Fase 2: Comprensión de los Datos 15
3.3.1. Identificación y selección de fuentes de datos web 15
3.3.2. Diseño e implementación de la extracción automatizada 15
3.4. Fase 3: Preparación de los Datos 21
3.4.1. Integración y selección de variables 22
3.4.2. Estandarización y limpieza de registros 22
3.5. Fase 4: Modelado 23
3.5.1. Selección de algoritmos predictivos 24
3.5.2. Partición del conjunto de datos 24
3.5.3. Entrenamiento y ajuste de hiperparámetros 25
3.6. Fase 5: Evaluación 26
3.6.1. Métricas de evaluación 26
3.6.2. Comparación de resultados 27
..... 28
3.7. Fase 6: Despliegue 28
3.7.1. Serialización del modelo 28
3.7.2. Diseño del flujo de inferencia 29
CAPÍTULO IV: ANÁLISIS DE RESULTADOS 30
4.1. Configuración del entorno de desarrollo 30
4.1.1. Estructura del proyecto 30
4.2. Conjunto de datos final 31
4.3. Resultados del entrenamiento de los modelos 31
4.4. Selección del mejor modelo 32
4.5. Importancia de las variables 33
4.6. Capacidad predictiva del modelo 33
4.7. Estimación de precios con el modelo final 34

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES **36**
5.1. Conclusiones **36**
5.2. Recomendaciones **37**
BIBLIOGRAFÍA **39**

ÍNDICE DE ILUSTRACIONES

Ilustración 1 Visualización de la plataforma web de Properati 10
Ilustración 2 Ilustración 1Visualización de la plataforma web de Plusvalia..... 11
Ilustración 3Página con filtro de Quito y solo propiedades en venta..... 16
Ilustración 4 Precio mediano por metro cuadrado según sector. 18
Ilustración 5 Composición porcentual de tipos de inmueble en los sectores con mayor oferta. 19
Ilustración 6 Relación entre el área típica y el precio por metro cuadrado de cada sector. 20
Ilustración 7 Visualización de los Dataset en crudo. 23
Ilustración 8 Visualización de los Dataset después de todo el proceso de limpieza 23
Ilustración 9 Comparación de los resultados de cada modelo en cada Dataset..... 26
Ilustración 10 Ranking del mejor modelo 28
Ilustración 11 Comparación del R^2 alcanzado por cada modelo en los tres conjuntos de datos 32
Ilustración 12 Importancia de cada variable en las predicciones del modelo Random Forest 33
Ilustración 13 Comparación entre el precio real y el precio estimado por el modelo en cada conjunto de datos.... 34

ÍNDICE DE TABLAS

Tabla 1 Indicadores de sectores representativos del mercado inmobiliario de Quito. 20
Tabla 2 Variables que componen el conjunto de datos final. 31
Tabla 3 Desempeño de los cuatro algoritmos en los tres conjuntos de datos. 32
Tabla 4 Ejemplos de precios estimados por el modelo final. 34

CAPÍTULO I: INTRODUCCIÓN

1.1. Planteamiento del problema

Para abordar esta situación, la presente disertación consiste en el diseño, desarrollo y evaluación de una herramienta tecnológica que automatice la recopilación de datos provenientes de diversas fuentes en línea y aplique técnicas de inteligencia artificial para estimar el valor comercial de las propiedades a partir de sus características.

El aspecto central o problema principal de la investigación es determinar de qué manera un modelo predictivo basado en Machine Learning, alimentado por datos extraídos automáticamente de la web, puede estimar con alta precisión y confiabilidad los precios del mercado inmobiliario en la zona urbana de Quito y sus valles aledaños.

En función de la problemática planteada, se formula la siguiente pregunta principal de investigación:

¿De qué manera un modelo predictivo basado en Machine Learning, alimentado con datos extraídos automáticamente de portales inmobiliarios, puede estimar con alta precisión los precios del mercado de bienes raíces en la zona urbana de Quito?

Asimismo, se desprenden las siguientes preguntas secundarias que orientarán el desarrollo del proyecto:

- ¿Qué técnicas de extracción automatizada de datos web resultan más eficientes y escalables para recolectar información estructurada y no estructurada de los portales inmobiliarios locales?
- ¿Cuál es el proceso óptimo de limpieza y transformación de los datos recolectados para asegurar que el modelo de aprendizaje automático reciba información de calidad?
- ¿Qué algoritmos específicos de Machine Learning presentan el mejor rendimiento y menor mar gen de error al analizar las características de los bienes raíces en este mercado particular?

1.2. Justificación

El mercado inmobiliario en la ciudad de Quito se caracteriza por su constante dinamismo y por la alta dispersión de la información referente a los precios de las propiedades. Actualmente, compradores, vendedores y agentes independientes enfrentan dificultades para establecer u obtener valoraciones objetivas y actualizadas en tiempo real. A menudo, el sector depende de tasaciones empíricas, conocimiento anecdótico o comparaciones manuales que consumen mucho tiempo y están sujetas a sesgos humanos.

Esta investigación es necesaria e importante porque propone una solución tecnológica orientada a democratizar el acceso a la información del valor de los bienes raíces. El principal aporte de este trabajo es proveer una herramienta matemática y computacional objetiva, fundamentada en datos reales y actuales del mercado local, que reduzca la asimetría de información existente y optimice la toma de decisiones económicas tanto de oferentes como de demandantes.

El proyecto es plenamente factible de realizar. La existencia de múltiples portales inmobiliarios de acceso público permite la recolección de grandes volúmenes de datos mediante técnicas de extracción automatizada, web scraping. Adicionalmente, el procesamiento de esta información y el posterior entrenamiento del modelo de Machine Learning se llevarán a cabo utilizando lenguajes de programación y librerías de código abierto ampliamente documentadas y accesibles, como Python, Pandas, Scikit-learn y BeautifulSoup.

En síntesis, la utilidad de esta disertación radica en entregar un modelo predictivo funcional que demuestre la viabilidad de combinar la automatización de datos con la inteligencia artificial. Este modelo sentará una base tecnológica escalable que podría integrarse en el futuro a plataformas del sector PropTech, tecnología aplicada a la propiedad en Quito, beneficiando directamente a la economía local y a los actores del mercado inmobiliario.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un modelo predictivo mediante algoritmos de Machine Learning para estimar los precios del mercado inmobiliario en la ciudad de Quito, utilizando información obtenida a través de técnicas de extracción automatizada de datos web

1.3.2. Objetivos específicos

- Sistematizar los fundamentos teóricos y el estado del arte referentes a los algoritmos de aprendizaje automático y las metodologías de recolección de datos en la web aplicados a la valoración de bienes raíces.
- Construir un flujo de software automatizado que permita extraer, limpiar y estructurar la información de las ofertas inmobiliarias publicadas en los principales portales digitales locales.
- Entrenar y evaluar el rendimiento de diferentes algoritmos predictivos mediante métricas de error estándar, con el fin de seleccionar el modelo que ofrezca la mayor precisión en el cálculo del valor de las propiedades.

1.4. Alcance

El alcance espacial del presente proyecto de titulación se limita a la zona urbana y los valles aledaños de la ciudad de Quito. En el ámbito tecnológico, el estudio se enfoca exclusivamente en la extracción de datos de texto y numéricos desde portales inmobiliarios ecuatorianos de acceso público, excluyendo el procesamiento de imágenes o el análisis de variables macroeconómicas externas. El modelado de datos se delimita al uso de algoritmos de Machine Learning supervisado orientados a tareas de regresión.

El resultado esperado, una vez concluida la investigación, es un modelo predictivo de software completamente funcional. Esta herramienta será capaz de recibir las características estructurales y geográficas de un inmueble tales como ubicación, metros cuadrados para devolver un precio de mercado estimado. Además, el alcance incluye la entrega del documento académico con el análisis comparativo del rendimiento de los diferentes algoritmos evaluados.

CAPÍTULO II: FUNDAMENTACIÓN TEÓRICA

2.1. Marco teórico

2.1.1. Antecedentes y estado del arte

El desarrollo de modelos predictivos para bienes raíces es un tema ampliamente explorado a nivel global, y la evidencia disponible demuestra la eficacia de los algoritmos de inteligencia artificial, en especial de los métodos de ensamble, para alcanzar una alta precisión en la estimación de precios. Un ejemplo representativo es el estudio realizado por Yazdani en 2021, que comparó el desempeño de las redes neuronales, el Random Forest y el método tradicional de regresión hedónica sobre el mercado de vivienda de la ciudad de Boulder, en Estados Unidos. Sus resultados mostraron que el Random Forest y las redes neuronales ofrecen mejores estimaciones que la regresión hedónica, confirmando que la relación entre las características de una vivienda y su precio no es lineal.

Esta tendencia se confirma de manera más amplia en la revisión sistemática elaborada por Gorjian en 2025, que analizó veintitrés investigaciones internacionales sobre predicción de precios de vivienda. En ese trabajo, el Random Forest resultó ser el modelo más utilizado y el de mejor rendimiento, al ser seleccionado en catorce de los veintitrés estudios revisados. La revisión también señala que, si bien los modelos hedónicos siguen siendo valiosos porque ayudan a entender qué variables influyen más en el precio, como la cercanía a los centros urbanos o las características propias del inmueble, la combinación de ambos enfoques mejora al mismo tiempo la precisión y la interpretación de los resultados.

A nivel regional, los estudios aplicados en ciudades latinoamericanas destacan la viabilidad de estos modelos, aunque subrayan el reto técnico que representa el preprocesamiento de la información y el manejo de los valores atípicos, propios de los mercados inmobiliarios de la región, tal como lo advierte Medina Giraldo en su trabajo de 2023.

En el ámbito local existen aproximaciones pioneras en la ciudad de Quito que ya emplean técnicas de Machine Learning y extracción de datos desde portales inmobiliarios, como la desarrollada por Vilca en 2022. El presente proyecto de titulación se diferencia de estos antecedentes al proponer una metodología integral que combina la extracción automatizada de

datos actualizados con la evaluación comparativa de múltiples arquitecturas predictivas, atendiendo de forma específica las particularidades del mercado quiteño.

2.1.2. Teoría de la valoración inmobiliaria

La teoría de la valoración inmobiliaria comprende los enfoques de mercado y los métodos comparativos de precios, fundamentales para entender qué características físicas y geográficas determinan el valor de un inmueble. Entre estos métodos destacan el enfoque comparativo de mercado, el enfoque del costo y el enfoque de capitalización de rentas, cada uno con criterios particulares aplicables al análisis de propiedades en zonas urbanas.

Para efectos del presente trabajo, se adopta principalmente el enfoque comparativo de mercado, ya que es el más compatible con el uso de datos masivos obtenidos de portales en línea y con la aplicación posterior de modelos de regresión basados en variables observables.

2.1.3. Web Scraping

Esta sección abarca los fundamentos de la arquitectura cliente-servidor, los protocolos de comunicación HTTP/HTTPS y las técnicas de análisis estructural de documentos de hipertexto que permiten la recolección masiva y automatizada de datos no estructurados disponibles en la red.

Asimismo, se revisan las consideraciones éticas y legales del web scraping, las buenas prácticas de respeto a los archivos robots.txt y a los términos de uso de los portales, así como las estrategias técnicas para gestionar limitaciones impuestas por los servidores, como el control de tasa de solicitudes y la rotación de identificadores de usuario.

2.1.4. Aprendizaje automático supervisado

En esta sección se exploran los fundamentos matemáticos y estadísticos de los algoritmos de regresión, específicamente modelos como la Regresión Lineal Múltiple, los Árboles de Decisión y Random Forests. Cada uno de estos algoritmos posee fortalezas particulares que serán evaluadas en el contexto del mercado inmobiliario quiteño.

Adicionalmente, se incluyen las teorías sobre las métricas de evaluación de error entre ellas el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el

Coeficiente de Determinación (R^2)— que permiten medir de forma objetiva la precisión y la generalización de los modelos entrenados

2.2. Marco conceptual

2.2.1. Machine Learning o Aprendizaje Automático

El aprendizaje automático es una rama fundamental de la inteligencia artificial y las ciencias de la computación que se centra en el desarrollo de algoritmos capaces de identificar patrones complejos dentro de grandes volúmenes de datos. En lugar de ser programado con instrucciones explícitas y estáticas para cada tarea, el sistema mejora su rendimiento probabilístico y capacidad de resolución a medida que interactúa y se expone a mayor información empírica (Universidad Internacional de Andalucía [UNIA], 2026). Esta disciplina permite automatizar procesos de análisis que simulan el reconocimiento lógico humano, posibilitando clasificaciones y estimaciones de alta precisión en entornos con múltiples variables.

2.2.2. Web Scraping

El web scraping o raspado web es una técnica de recolección sistemática y automatizada de información alojada en páginas de internet. Se ejecuta mediante scripts o herramientas de software especializadas que procesan la estructura subyacente de un sitio web, extrayendo los datos mediante reglas lógicas definidas por el desarrollador para transformarlos en información estructurada (Red de Innovación Turística [Red Intur], 2023). Al aplicarse en investigaciones sectoriales, permite consolidar grandes volúmenes de datos dispersos (como listados de propiedades, precios y características) de manera rápida, optimizando la fase de recolección de información.

2.2.3. Modelo Predictivo

En el ámbito del machine learning, un modelo predictivo es una construcción algorítmica y matemática diseñada para estimar comportamientos, valores o tendencias futuras fundamentándose en el análisis estadístico de la información pasada. Para lograrlo, "el proceso consiste en crear un modelo algorítmico que permite ingresar datos históricos, hacer un entrenamiento según el enfoque que se requiera y, a partir de este, encontrar patrones para predecir alguna información futura" (Universidad de Las Américas [UDLA], 2024, p. 16).

Algoritmos como las regresiones múltiples son ejemplos clásicos de cómo estos modelos infieren variables dependientes basándose en covariables conocidas.

2.2.4. Dataset o Conjunto de Datos

Un dataset constituye la materia prima estructural para el entrenamiento, validación y evaluación de cualquier algoritmo analítico. En la arquitectura de modelos de estimación, su correcta generación, limpieza y formateo resultan críticos; un conjunto con datos sesgados o atípicos (outliers) no tratados comprometerá la fiabilidad de las predicciones. Por esta razón, la literatura especializada lo define como "la piedra angular de todo proyecto de machine learning y deep learning" (Méndez Mangana & Barraquer Kargacin, 2023, p. 177).

2.2.5. Mercado Inmobiliario

El mercado inmobiliario es el entorno económico donde convergen la oferta y la demanda para la transacción (ya sea compra, venta o alquiler) de bienes raíces. Esta dinámica de intercambio no es estática, sino que fluctúa constantemente en función de variables macroeconómicas, tasas de acceso a financiación, políticas urbanísticas y las características geográficas de los activos (Inmogesco, 2024). Su análisis desde la ciencia de datos exige integrar factores tanto de las características físicas de la propiedad como del entorno económico que la rodea.

2.2.6. Quito

Geográfica y demográficamente, el Distrito Metropolitano de Quito, capital del Ecuador, es una urbe emplazada en la hoya de Guayllabamba que se divide administrativamente en parroquias urbanas y rurales. Durante las últimas décadas, la ciudad ha experimentado una importante expansión de su mancha urbana y una fuerte tasa de crecimiento poblacional, extendiendo sus fronteras habitacionales particularmente hacia la ruralidad y los valles periféricos (Quito Cómo Vamos, 2024). Esta dispersión territorial y aumento de densidad convierten a la ciudad en un espacio altamente dinámico para la evaluación de fenómenos de urbanización y la tasación de tierras.

2.2.7. Inmueble

Desde una perspectiva legal y económica, un inmueble es un bien patrimonial tangible

que se caracteriza por su fijación e inmovilidad espacial; es decir, no puede ser trasladado sin que se altere de manera fundamental su estructura, naturaleza o el suelo donde se asienta. Las tierras, edificaciones habitacionales y construcciones comerciales son los componentes principales de esta categoría, funcionando como activos de alto valor de transacción y resguardo de inversión (Jara Flórez, s.f.).

2.2.8. Plataformas Inmobiliarias

Las plataformas inmobiliarias son ecosistemas digitales diseñados para centralizar, gestionar y facilitar la interacción entre la oferta y la demanda de bienes raíces. Funcionan como intermediarios tecnológicos o catálogos virtuales donde propietarios, agentes y constructoras pueden publicar el inventario disponible para venta o alquiler, permitiendo a los usuarios buscar y comparar opciones según sus necesidades. Sin embargo, desde la perspectiva del análisis de datos, estas plataformas trascienden su función comercial y actúan como repositorios masivos de información semiestructurada. Por lo tanto, en el desarrollo de modelos de estimación, estos portales se convierten en la fuente primaria de información empírica, la cual es extraída para construir los conjuntos de datos que alimentarán los algoritmos.

2.2.9. Properati

Properati es un portal inmobiliario de alcance latinoamericano especializado en la comercialización de bienes raíces que se diferencia por su fuerte enfoque analítico y de ciencia de datos. Más allá de funcionar como un catálogo de propiedades, su infraestructura "utiliza datos abiertos y crea herramientas de análisis para brindar información adicional a los futuros compradores, inquilinos o inversores" (Explora Latam, 2021, párr. 3). Al procesar y publicar periódicamente información sobre el valor del metro cuadrado por barrios y zonas, funciona como un valioso referente empírico para contrastar y validar la información recolectada para modelos de estimación.

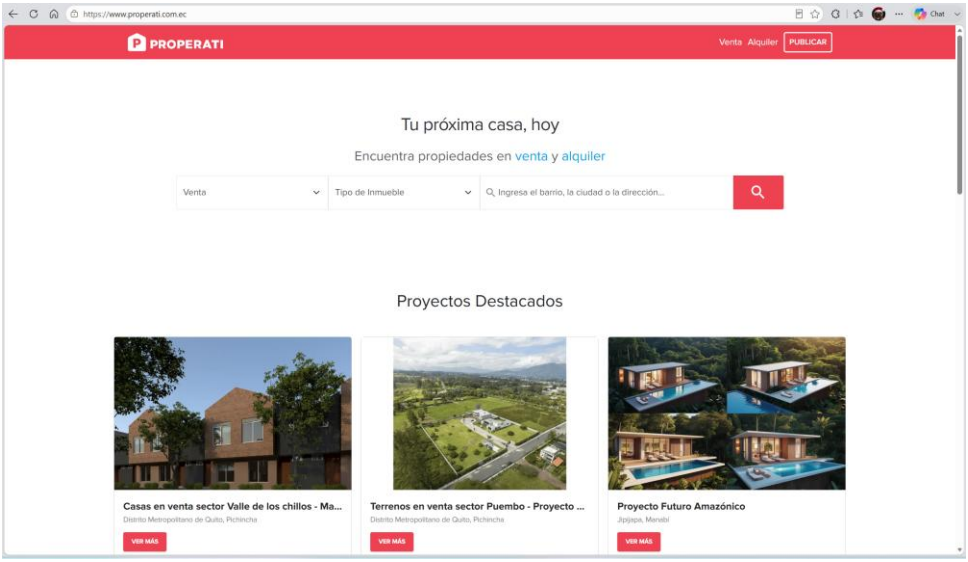


Ilustración 1 Visualización de la plataforma web de Properati

2.2.10. Plusvalía

En la teoría económica, la plusvalía representa el incremento del valor monetario de un bien inmueble a lo largo del tiempo, impulsado habitualmente por factores extrínsecos como la dotación de infraestructura urbana, mejoras en la conectividad o dinámicas de densidad poblacional. Por otro lado, en el entorno tecnológico del Ecuador, Plusvalía se consolida como el portal inmobiliario líder y de mayor penetración en el país. Recientemente integrado a redes operativas internacionales como el holding tecnológico QuintoAndar (Ekos Negocios, 2021), este portal alberga el inventario más grande y dinámico de anuncios de venta y alquiler en el Distrito Metropolitano, convirtiéndose en el espacio web predilecto para el raspado de datos y la estructuración de la oferta real.

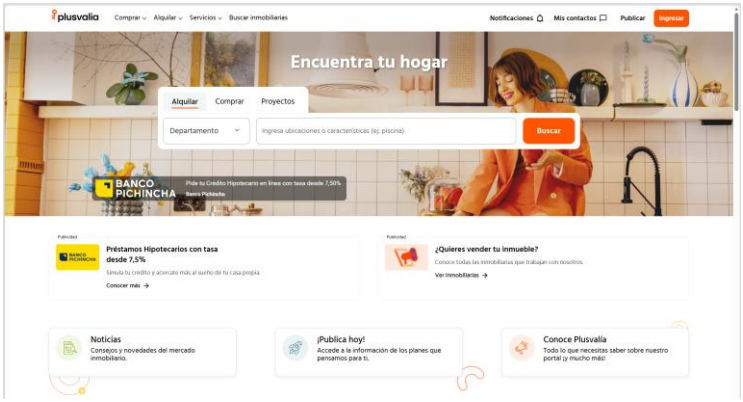


Ilustración 2 Visualización de la plataforma web de Plusvalia

CAPÍTULO III: DESARROLLO DEL TRABAJO

En este capítulo se detalla paso a paso el trabajo técnico y metodológico que se llevó a cabo para implementar el modelo predictivo, respetando las fechas del plan de disertación. A continuación, se muestra la estructura de los temas que se van a abordar.

3.1. Metodología de Trabajo

El presente capítulo detalla el proceso técnico e ingenieril llevado a cabo para el diseño, entrenamiento e implementación del modelo predictivo orientado al mercado inmobiliario de la ciudad de Quito. Para garantizar el rigor científico, la organización sistemática y la reproducibilidad del proyecto, el desarrollo del trabajo se fundamenta en la metodología CRISP-DM.

Se seleccionó CRISP-DM por ser un marco de trabajo de estándar abierto, validado y ampliamente adoptado en la industria para el desarrollo de proyectos de ciencia de datos y Machine Learning. La principal ventaja de este enfoque metodológico radica en su naturaleza estructurada e iterativa, la cual permite alinear de manera directa los objetivos del dominio de estudio en este caso, la dinámica y valoración de los bienes raíces con la implementación técnica de los algoritmos de predicción.

El ciclo de vida del presente proyecto informático se articula en las seis fases secuenciales establecidas por este estándar, las cuales dictan la estructura de los apartados subsiguientes:

1. Comprensión del Negocio: Establecimiento de los objetivos del proyecto y definición de las variables que influyen en la valoración inmobiliaria a nivel local.
2. Comprensión de los Datos: Identificación de los portales web de acceso público, evaluación de los métodos de extracción como web scraping y análisis exploratorio inicial de la información recolectada.
3. Preparación de los Datos: Consolidación de los repositorios de datos, limpieza de valores nulos o atípicos y estandarización de las variables tecnológicas y geográficas para su posterior procesamiento.
4. Modelado: Selección de las técnicas de aprendizaje automático, partición de los conjuntos

de entrenamiento y prueba, y parametrización de los algoritmos predictivos.

5. Evaluación: Medición del rendimiento de los modelos matemáticos construidos mediante métricas de error estandarizadas para determinar el algoritmo de mayor precisión.
6. Despliegue: Serialización del modelo óptimo y establecimiento del flujo de inferencia para la estimación de precios sobre nuevas propiedades.

Para la recolección de datos se evaluaron los principales portales inmobiliarios de acceso público con presencia en el mercado ecuatoriano. Tras dicho análisis se seleccionó el portal Plusvalía (plusvalia.com) como fuente primaria de información, debido a que concentra uno de los mayores volúmenes de oferta de inmuebles en la ciudad de Quito y a que forma parte de un grupo internacional especializado, lo que se traduce en una estructura de publicación homogénea y en un catálogo amplio y actualizado de manera diaria.

La caracterización técnica de la fuente puso de manifiesto dos aspectos que resultaron determinantes para el diseño de la extracción. En primer lugar, las páginas de resultados o listados se generan dinámicamente mediante JavaScript y están protegidas por un sistema anti-bot comercial (DataDome) que despliega desafíos de verificación (CAPTCHA) ante los accesos automatizados, lo que impide su recolección directa. En segundo lugar, se comprobó que las páginas de detalle de cada inmueble, identificadas por rutas del tipo /propiedades/..., sí son accesibles de forma directa e individual.

El examen del archivo robots.txt del portal confirmó esta distinción y orientó la estrategia ética y técnica del proyecto. Dicho archivo autoriza de manera explícita el rastreo de las fichas, declara la ubicación de los mapas de sitio (sitemaps) del portal y restringe el acceso a sus interfaces de programación internas. Esta lectura permitió delimitar el trabajo exclusivamente a recursos cuyo acceso se encuentra permitido por la propia fuente.

De manera complementaria se valoraron fuentes alternativas, entre ellas conjuntos de datos abiertos orientados a la investigación (Properati) y servicios comerciales de extracción. Estas opciones se documentaron como vías de respaldo, priorizando para el presente trabajo la extracción directa de las fichas autorizadas por ser la alternativa más reproducible y de menor costo.

Adicionalmente, se seleccionó como segunda fuente primaria el portal Properati Ecuador (properati.com.ec). A pesar de pertenecer al mismo grupo empresarial que Plusvalía, presenta una infraestructura técnica notoriamente distinta en particular, una arquitectura de renderizado del lado de la cliente basada en componentes web, lo que aporta diversidad metodológica y fortalece la triangulación del conjunto de datos final.

En el proceso de selección se evaluaron otros portales que finalmente fueron descartados con justificación documentada. Mercado Libre Inmuebles Ecuador (inmuebles.mercadolibre.com.ec) se descartó por volumen insuficiente: el portal registra aproximadamente 135 inmuebles activos a nivel nacional en sus categorías residenciales, cifra que arroja una muestra demasiado pequeña al delimitarse a la ciudad de Quito.

3.2. Fase 1: Comprensión del Negocio

El primer paso de la metodología CRISP-DM consiste en comprender los objetivos y requisitos del proyecto desde una perspectiva del dominio de aplicación, transformando este conocimiento en la definición de un problema de minería de datos. En el contexto del presente trabajo, el "negocio" o dominio de estudio corresponde al mercado inmobiliario de la ciudad de Quito.

El objetivo principal desde esta perspectiva es mitigar la asimetría de información existente en el sector para poder fijar el precio en el mercado de Quito. En consecuencia, los actores involucrados compradores, vendedores y agentes independientes a menudo deben depender de tasaciones empíricas, conocimiento anecdótico o comparaciones manuales que consumen tiempo y están sujetas a sesgos humanos. El modelo predictivo se plantea, por tanto, como una herramienta de uso general orientada a democratizar el acceso a valoraciones objetivas, permitiendo, por ejemplo, que un oferente sepa a qué precio publicar una propiedad o que un demandante evalúe si un inmueble se encuentra sobrevalorado respecto a la media del mercado.

Para que el modelo logre capturar la dinámica de este mercado, se determina que el valor comercial de una propiedad no es aleatorio, sino que está condicionado por un conjunto de características tangibles que deben ser parametrizadas. Entre los factores críticos identificados para el entrenamiento del algoritmo se encuentran las características geográficas (ubicación en

la zona urbana o valles) y las características estructurales del inmueble, tales como la superficie en metros cuadrados, el número de habitaciones y la antigüedad.

Finalmente, el criterio de éxito del proyecto se define como la obtención de una herramienta computacional funcional, capaz de procesar estas características físicas y geográficas para devolver un precio de mercado estimado. La validez de esta solución tecnológica se medirá mediante métricas de error estándar, buscando el algoritmo que ofrezca la mayor precisión y confiabilidad para consolidar una base tecnológica aplicable al sector inmobiliaria en Quito.

3.3. Fase 2: Comprensión de los Datos

Esta fase se enfoca en la identificación de las fuentes de información, la recolección inicial de los datos y la evaluación de las herramientas técnicas necesarias para extraerlos de manera efectiva.

3.3.1. Identificación y selección de fuentes de datos web

Para la recolección de datos, se evaluaron los principales portales inmobiliarios de acceso público con presencia en el mercado ecuatoriano. Tras este análisis, se definió trabajar con dos fuentes primarias:

- Plusvalía (plusvalia.com): Se seleccionó como fuente principal porque concentra uno de los mayores volúmenes de oferta en Quito. Al pertenecer a un grupo internacional, mantiene una estructura de publicación homogénea y actualizada.
- Properati Ecuador (properati.com.ec): Se incluyó como segunda fuente primaria fortalecer el conjunto de datos final.

En el proceso de evaluación para elegir otras alternativas que finalmente fueron descartadas. Mercado Libre Inmuebles se descartó por no contar con un volumen suficiente de avisos para la ciudad de Quito.

3.3.2. Diseño e implementación de la extracción automatizada

Durante la fase de validación inicial en la web de plusvalía, se detectó una discrepancia importante entre el número total de resultados reportados por los buscadores de los portales y la

cantidad de registros efectivamente recuperados mediante la estrategia basada en sitemaps. Esta inconsistencia evidenció que el acceso vía sitemaps no lograba capturar la totalidad de la oferta publicada, ya que algunos listados dinámicos o filtrados no se reflejaban en dichos archivos.

Ante esta limitación, se procedió a implementar un enfoque complementario basado en la automatización de un navegador real mediante Selenium. Esta técnica permitió replicar el comportamiento de un usuario humano para acceder a los listados de búsqueda. Para garantizar la estabilidad del proceso y evitar la activación de los mecanismos de bloqueo (CAPTCHA).

- Simulación de comportamiento humano: Se integraron tiempos de espera aleatorios entre las acciones de scroll y la navegación entre páginas.
- Gestión de la carga: Se configuró el navegador para esperar a la renderización completa de los componentes dinámicos de cada página antes de realizar la extracción.

Este ajuste permitió superar las restricciones de acceso y asegurar la recolección íntegra de la muestra, demostrando que la automatización de navegadores, aunque más demandante en términos de recursos computacionales, es un método más robusto para capturar el volumen total de datos en plataformas web.

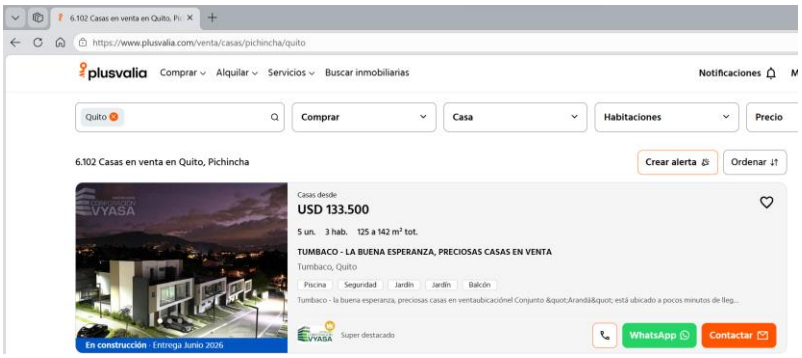


Ilustración 3Página con filtro de Quito y solo propiedades en venta

Se puede observar que la cantidad de propiedades que se puede obtener son un aproximado de 6.101 mientras que al usar el método de sitemaps solo se obtuvieron un total de 2457

3.3.3. Exploración y caracterización del mercado por sector

Una vez integrados y depurados los registros provenientes de ambos portales, se realizó un análisis exploratorio orientado a comprender la estructura del mercado inmobiliario de Quito

a nivel geográfico. El conjunto consolidado reúne 9.644 propiedades distribuidas en numerosos sectores de la ciudad. Con el fin de asegurar la fiabilidad estadística de los indicadores, el análisis se concentró en los sectores que cuentan con al menos 30 registros, criterio que dejó 44 sectores representativos que concentran aproximadamente el 93 % de la oferta total. Como indicador central de comparación se adoptó el precio por metro cuadrado (USD/m²), por tratarse de una medida de plusvalía independiente del tamaño de cada propiedad y, por tanto, comparable entre zonas.

El análisis del precio mediano por metro cuadrado reveló una marcada jerarquía geográfica de la plusvalía. Los sectores de mayor valor por metro son el Valle de Tumbaco (1.136 USD/m²), Centro Norte (1.008), Lumbisí (943), San Juan de Cumbayá (922) y Nayón (893). En el extremo opuesto se ubican Quitumbe (410), Cotocollao (467), Carapungo (479) y el Sur de Quito (498). La mediana global entre sectores se situó en 624 USD/m², valor que sirve como referencia para clasificar cada zona por encima o por debajo del promedio del mercado.

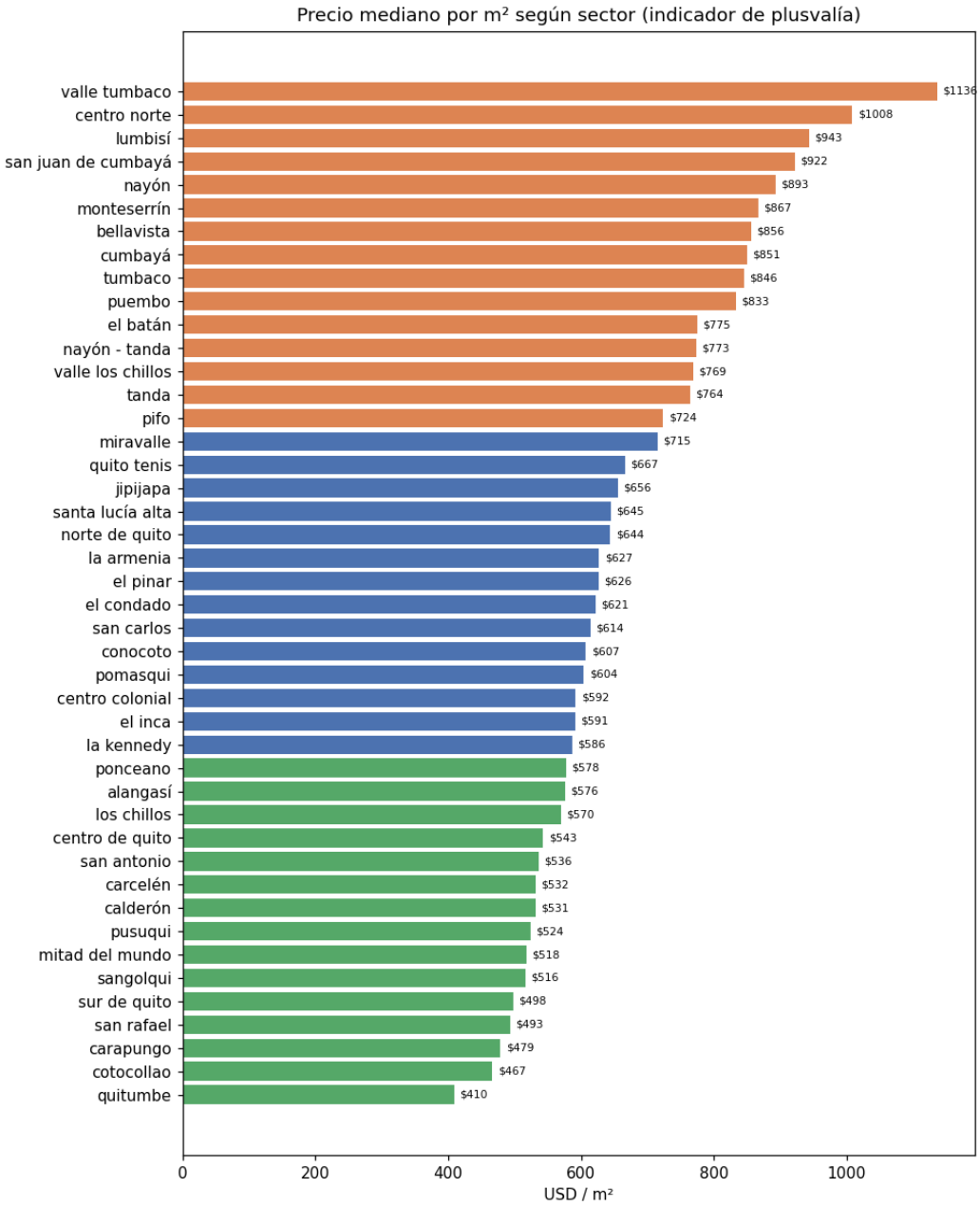


Ilustración 4 Precio mediano por metro cuadrado según sector.

La composición por tipo de inmueble explica buena parte de estas diferencias. Los sectores de los valles orientales (Cumbayá, Tumbaco, Puembo, Nayón) están compuestos casi en su totalidad por casas, mientras que Centro Norte y Norte de Quito se encuentran dominados por departamentos y suites. Esta diferencia de tipología explica que Centro Norte presente un precio por metro cuadrado muy elevado pese a tener áreas reducidas: corresponde a un mercado

de departamentos céntricos, pequeños pero altamente cotizados por unidad de superficie.

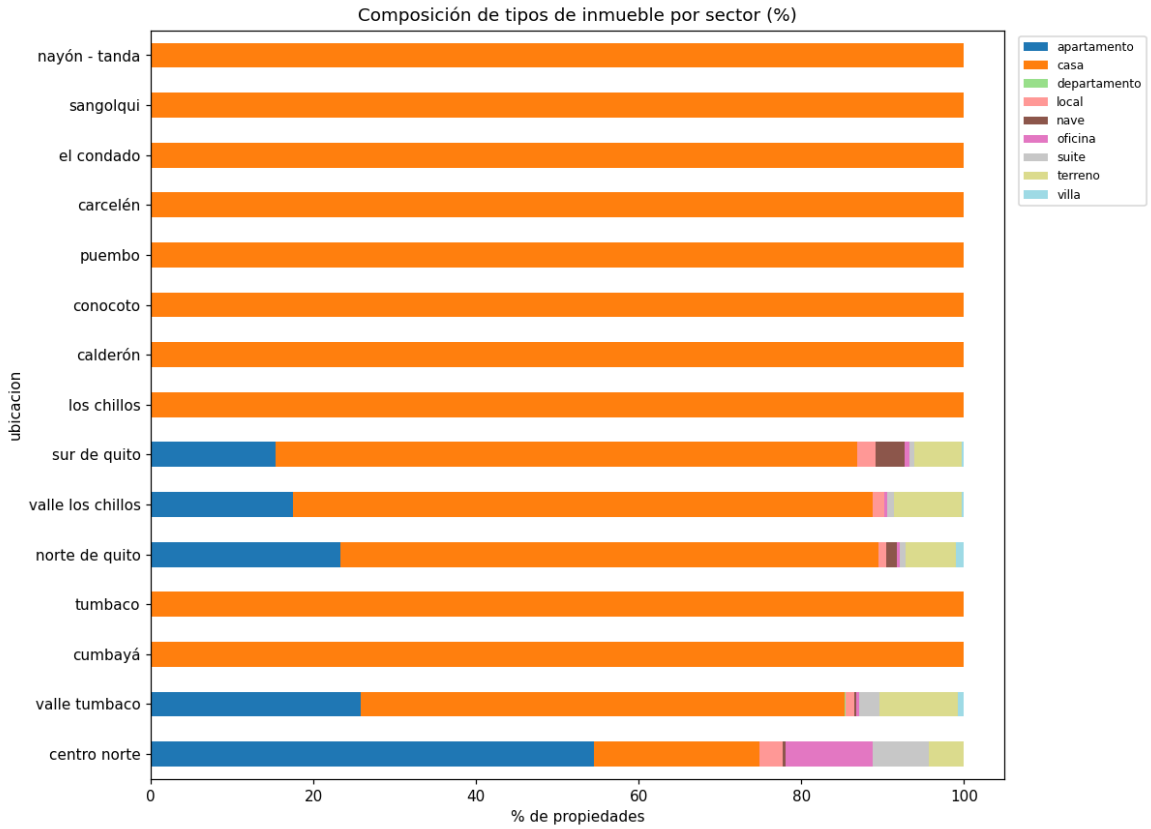


Ilustración 5 Composición porcentual de tipos de inmueble en los sectores con mayor oferta.

Al contrastar el área mediana frente al precio por metro cuadrado se distinguen dos perfiles dominantes de mercado. Por un lado, sectores de área reducida y precio por metro cuadrado elevado, asociados al mercado de departamentos urbanos (Centro Norte). Por otro, sectores de gran superficie (Cumbayá, Monteserrín, Nayón) con plusvalía consolidada en vivienda unifamiliar de mayor tamaño. En la parte inferior del gráfico se agrupan los sectores periféricos de carácter económico.

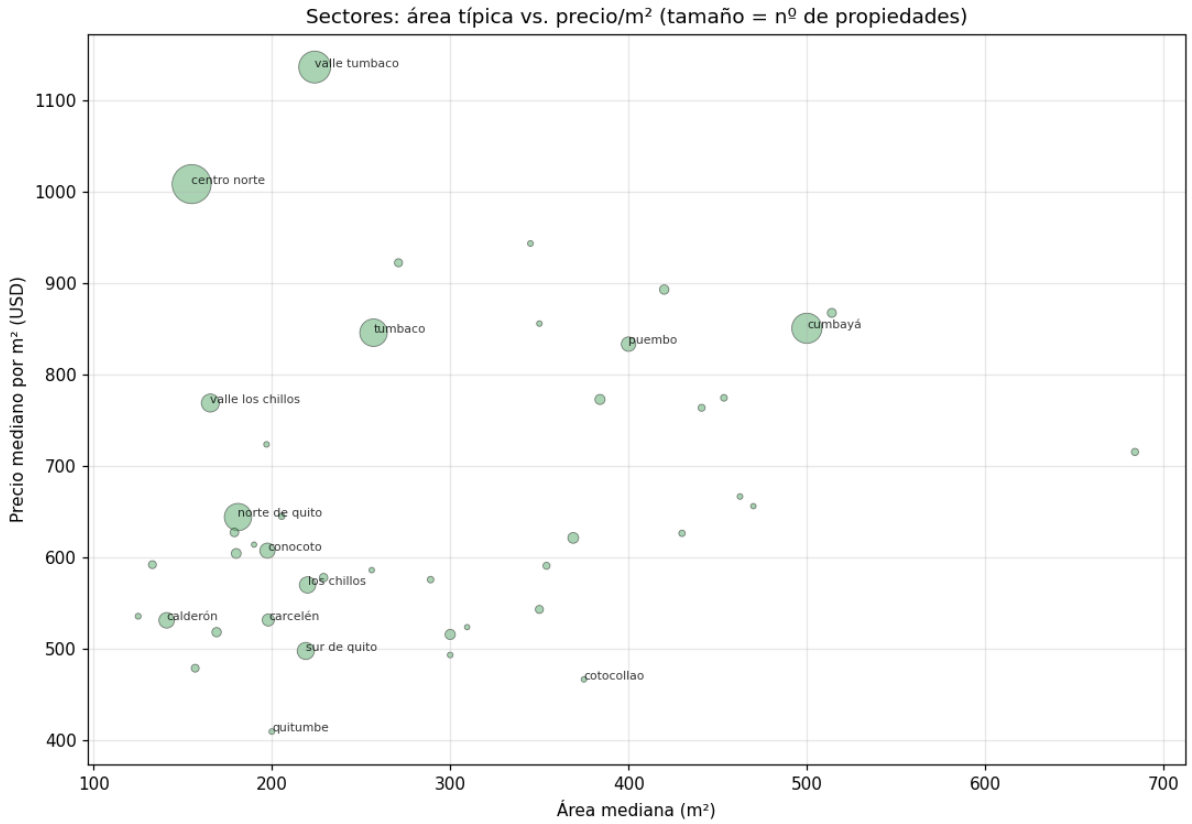


Ilustración 6 Relación entre el área típica y el precio por metro cuadrado de cada sector.

La siguiente tabla resume los indicadores de una selección de sectores representativos, ordenados de mayor a menor plusvalía:

Sector	n	Precio med. (USD)	USD/m²	Área med. (m²)	Tipo dominante
Valle de Tumbaco	1.079	230.000	1.136	224	Casa
Centro Norte	1.616	169.000	1.008	155	Departamento
Nayón	94	349.000	893	420	Casa
Monteserrín	86	487.868	867	514	Casa
Cumbayá	965	425.000	851	500	Casa
Puembo	223	358.000	833	400	Casa
Sur de Quito	313	98.000	498	219	Casa
Carapungo	66	79.000	479	157	Casa
Quitumbe	35	90.000	410	200	Casa

Tabla 1 Indicadores de sectores representativos del mercado inmobiliario de Quito.

A partir de los hallazgos anteriores, los 44 sectores analizados pueden agruparse en cuatro perfiles de mercado:

- **Valles orientales premium**, que incluyen a Cumbayá, Tumbaco, Nayón, Puembo y Monteserrín: casas amplias de 400 a 500 m², con los precios totales más altos y un precio por metro cuadrado superior a 830 USD.
- **Núcleo urbano de departamentos**, concentrado en el Centro Norte y Norte de Quito: predominio de departaamento y suites, áreas pequeñas, pero precio por metro cuadrado muy alto y el mayor volumen de transacciones.
- **Zona intermedia consolidada**, conformada por Conocoto, Los Chillos, Sangolquí, El Condado y Ponceano: precio por metro cuadrado medio, de entre 500 y 700 USD, y buena relación entre tamaño y precio, propio de la clase media.
- **Periferia económica**, ubicada en sectores como Calderón, Carapungo, Quitumbe, Sur de Quito y Pomasqui: los precios por metro cuadrado más bajos, desde 410 hasta 520 USD, con mercados homogéneos y orientados a vivienda accesible.

Esta caracterización confirma que la ubicación constituye un factor estructural en la formación de los precios inmobiliarios de Quito, hallazgo que se ve reflejado posteriormente en la fase de modelado, donde la variable de ubicación resultó ser la segunda de mayor importancia predictiva, únicamente por detrás del área. Asimismo, la detección durante esta exploración de valores atípicos tales como precios y superficies inconsistentes justificó los criterios de filtrado y depuración aplicados en la fase de preparación de los datos.

3.4. Fase 3: Preparación de los Datos

Una vez recolectada la información cruda desde las fuentes web, se procedió a la fase de preparación de los datos. Esta etapa es crítica para garantizar que los algoritmos de aprendizaje automático reciban información estructurada, estandarizada y libre de ruido. Para este proceso se utilizó la librería pandas en el entorno de Python, aplicando un flujo de trabajo secuencial sobre los archivos extraídos.

3.4.1. Integración y selección de variables

El primer paso consistió en la homologación de los conjuntos de datos obtenidos de ambos portales inmobiliarios para permitir su integración en una estructura tabular única. Durante este proceso, se realizó una reducción de dimensionalidad eliminando variables que no aportan valor predictivo al modelo. Las columnas referentes a la dirección web, la metadata de paginación y la estampa temporal fueron descartadas del dataset definitivo.

3.4.2. Estandarización y limpieza de registros

Para homogeneizar la información contenida en las variables restantes, se aplicó una serie de transformaciones a nivel de registro:

- Normalización de texto: Se ejecutó una conversión general a minúsculas en todas las variables de tipo cadena de caracteres. Esto evita la duplicidad de categorías generada por diferencias en el uso de mayúsculas.
- Categorización del tipo de inmueble: Sobre la columna título, se aplicó un filtro de extracción por coincidencia de palabras clave. El algoritmo recorrió cada celda conservando únicamente los términos que definen la categoría arquitectónica tales como casa, departamento, local, suite, terreno, oficina, nave o villa, descartando el resto del texto descriptivo o publicitario introducido por los anunciantes.
- Depuración de la variable objetivo: La columna de precios contenía caracteres alfabéticos y símbolos monetarios incrustados. Se aplicaron reglas de limpieza para extraer de forma estricta los valores numéricos. En los casos donde la celda original carecía de un precio, el sistema asignó automáticamente un valor nulo tipificado como Not a Number, marcándolo como dato faltante para su posterior tratamiento.
- Segmentación geográfica jerárquica: La limpieza de la columna ubicación presentó un desafío particular debido a la heterogeneidad en el registro de los oferentes. Para extraer el macro sector geográfico de forma precisa, se programó una lógica de segmentación que divide la cadena de texto separada por comas y extrae sistemáticamente el tercer elemento leyendo de derecha a izquierda, obteniendo así un nivel de zonificación estandarizado para la ciudad.

- Extracción de variables continuas y discretas: Para las características físicas del inmueble como dormitorios, baños y area, se aplicaron funciones de limpieza que eliminaron los sufijos de texto, conservando exclusivamente los valores numéricos para transformar las columnas a tipos de datos enteros y flotantes.

```
import pandas as pd
import re

# 1. CARGAMOS LOS DATOS
# Aquí simulo los datos para que el notebook funcione, pero en tu caso debes usar:
df = pd.read_csv('/content/dataset_properati.csv')


display(df.head())
```

	url	titulo	precio	ubicacion	dormitorios	banos	area	pagina	fecha_extraccion
0	https://www.properati.com.ec/proyecto/14032-32...	Terrenos en venta sector Puembo - Proyecto La ...	Desde $ 210.000	Puembo, Valle Tumbaco, Quito, Pichincha	NaN	NaN	NaN	1	2026-06-17T10:07:03
1	https://www.properati.com.ec/proyecto/14032-32...	El Tesoro - Casas sector Valle de Los chillos	Desde $ 119.000	Amaguaña, Sur De Quito, Quito, Pichincha	3 dormitorios	3 baños	Desde 127 m²	1	2026-06-17T10:07:03
2	https://www.properati.com.ec/detalle/14032-32-...	Apartamento en Venta en La Carolina	$ 390.000	La Carolina, Iñaquito, Centro Norte, Quito, Pi...	3 dormitorios	4 baños	260 m²	1	2026-06-17T10:07:03
3	https://www.properati.com.ec/detalle/14032-32-...	Apartamento en Venta en La Carolina	$ 125.000	La Carolina, Iñaquito, Centro Norte, Quito, Pi...	2 dormitorios	2 baños	65 m²	1	2026-06-17T10:07:03
4	https://www.properati.com.ec/detalle/14032-32-...	Apartamento en Venta en Iñaquito	$ 186.311	Iñaquito, Centro Norte, Quito, Pichincha	2 dormitorios	3 baños	86 m²	1	2026-06-17T10:07:03

Ilustración 7 Visualización de los Dataset en crudo.

	titulo	precio	ubicacion	dormitorios	banos	area
0	terreno	210.000	valle tumbaco	NaN	NaN	NaN
1	casa	119.000	sur de quito	3.0	3.0	127.0
2	apartamento	390.000	centro norte	3.0	4.0	260.0
3	apartamento	125.000	centro norte	2.0	2.0	65.0
4	apartamento	186.311	centro norte	2.0	3.0	86.0

Ilustración 8 Visualización de los Dataset después de todo el proceso de limpieza

3.5. Fase 4: Modelado

La cuarta fase de la metodología CRISP-DM tiene por objetivo seleccionar y aplicar técnicas de modelado sobre los datos preparados, calibrando sus parámetros para obtener el mejor desempeño predictivo posible. En el presente trabajo, el modelado se realizó sobre tres configuraciones de datos el conjunto Plusvalia el cual está compuesto exclusivamente por casas, el conjunto Properati y un conjunto Unificado resultante de la armonización de ambos, con el fin de predecir el precio de venta de las propiedades del Distrito Metropolitano de Quito.

3.5.1. Selección de algoritmos predictivos

Dado que la variable objetivo, el precio, es continua, el problema se aborda como una tarea de regresión supervisada. Se seleccionaron cuatro algoritmos de naturaleza complementaria, con el propósito de contrastar enfoques de complejidad creciente y justificar empíricamente la elección final:

a) Regresión Lineal. Se incorpora como modelo de referencia. Asume una relación lineal entre las variables predictoras y el precio, y ofrece máxima interpretabilidad: cada coeficiente expresa el efecto marginal de una variable. Su limitación es que no captura relaciones no lineales ni interacciones entre variables, frecuentes en el mercado inmobiliario.

b) Árbol de Decisión. Particiona recursivamente el espacio de características mediante reglas de decisión, lo que le permite modelar relaciones no lineales e interacciones de manera natural y con alta interpretabilidad. Su principal desventaja es la tendencia al sobreajuste cuando crece sin restricciones.

c) Random Forest. Es un método de ensamble que construye numerosos árboles de decisión sobre muestras aleatorias de los datos y de las variables, y promedia sus predicciones. Esta agregación reduce drásticamente la varianza del árbol individual conservando su capacidad para capturar no linealidades, lo que suele traducirse en un desempeño superior y más estable en datos tabulares.

d) Red Neuronal. Modelo conexionista capaz de aproximar funciones altamente no lineales mediante capas de neuronas con funciones de activación. Se incluye para evaluar si un enfoque de aprendizaje profundo aporta mejoras frente a los métodos basados en árboles en un problema tabular de tamaño moderado.

La elección de estos cuatro algoritmos permite cubrir el espectro que va desde un modelo lineal simple e interpretable hasta modelos no lineales de mayor capacidad, sustentando la comparación con criterios tanto teóricos como empíricos.

3.5.2. Partición del conjunto de datos

Para evaluar la capacidad de generalización de los modelos es decir, su desempeño sobre datos no observados durante el entrenamiento el conjunto de datos se dividió mediante la técnica

de hold-out en dos subconjuntos: 80 % para entrenamiento y 20 % para prueba. El subconjunto de prueba se mantuvo completamente aislado del proceso de ajuste y se utilizó únicamente para la evaluación final, evitando así estimaciones optimistas del error. La partición se realizó con una semilla aleatoria fija para garantizar la reproducibilidad del experimento.

Complementariamente, durante el ajuste de hiperparámetros se empleó validación cruzada de K particiones. Esta técnica divide el conjunto de entrenamiento en cinco pliegues; en cada iteración, cuatro pliegues se utilizan para entrenar y el restante para validar, rotando el rol de validación entre todos los pliegues. La métrica reportada es el promedio de las cinco iteraciones, lo que proporciona una estimación más robusta y estable del desempeño, menos dependiente de una partición particular y útil para detectar sobreajuste.

Antes de la partición se aplicó un preprocesamiento encapsulado en un Pipeline, de modo que las transformaciones codificación one-hot de las variables categóricas ubicación y tipo de inmueble, y estandarización de las variables numéricas (área, dormitorios y baños) se ajustaran exclusivamente con los datos de entrenamiento dentro de cada pliegue. Este encadenamiento evita la fuga de información desde el conjunto de prueba hacia el de entrenamiento.

3.5.3. Entrenamiento y ajuste de hiperparámetros

Cada algoritmo se entrenó y optimizó mediante búsqueda exhaustiva en rejilla combinada con la validación cruzada de cinco particiones descrita, utilizando el coeficiente de determinación R^2 como criterio de selección. La configuración interna de cada modelo fue la siguiente:

- Regresión Lineal: no requiere hiperparámetros de ajuste; se entrenó por mínimos cuadrados ordinarios como modelo de referencia.
- Árbol de Decisión: se exploraron la profundidad máxima y el mínimo de muestras por hoja, parámetros que controlan la complejidad del árbol y previenen el sobreajuste.
- Random Forest: se fijó el número de árboles en 200 y se ajustaron la profundidad máxima y el mínimo de muestras por hoja, equilibrando capacidad de ajuste y generalización.

- Red Neuronal - MLP: se evaluaron distintas arquitecturas de capas ocultas y el coeficiente de regularización L2. Se empleó un máximo de 600 iteraciones con parada temprana para detener el entrenamiento cuando la validación dejaba de mejorar, evitando el sobreajuste.

Dataset	Modelo	R^2 (test)	R^2 (val. cruzada)	RMSE (USD)	MAE (USD)
Plusvalia	Regresión Lineal	0.687	0.671	151 181	92 033
Plusvalia	Árbol de Decisión	0.720	0.696	143 058	81 238
Plusvalia	**Random Forest**	**0.808**	**0.777**	**118 232**	**61 016**
Plusvalia	MLP	0.698	0.680	148 364	87 392
Properati	Regresión Lineal	0.590	0.585	104 876	68 386
Properati	Árbol de Decisión	0.658	0.618	95 715	60 939
Properati	**Random Forest**	**0.693**	**0.638**	**90 800**	**57 400**
Properati	MLP	0.590	0.582	104 897	68 414
Unificado	Regresión Lineal	0.673	0.662	133 658	78 342
Unificado	Árbol de Decisión	0.688	0.690	130 683	73 506
Unificado	**Random Forest**	**0.783**	**0.759**	**108 810**	**58 929**
Unificado	MLP	0.686	0.662	131 110	75 268

Ilustración 9 Comparación de los resultados de cada modelo en cada Dataset

3.6. Fase 5: Evaluación

El objetivo de la quinta fase de CRISP-DM es evaluar de forma objetiva la calidad de los modelos construidos. Aquí se comprueba si realmente cumplen con la meta principal del proyecto, orientada a predecir el precio de las propiedades con la mayor exactitud posible. Para llevar a cabo esta validación, se evaluaron todos los modelos frente al conjunto de prueba. Este subconjunto corresponde al 20 % de los datos que fue reservado exclusivamente para este fin y que no se utilizó durante el entrenamiento, lo que permite obtener una estimación imparcial de cómo se comportarán los modelos ante nueva información.

3.6.1. Métricas de evaluación

Al tratarse de un problema de regresión, el desempeño se midió mediante tres indicadores complementarios. Sea n el número de observaciones de prueba, y_i el precio real, $\hat{y}_i$ el precio predicho por el modelo y $\bar{y}$ el promedio de los precios reales.

Error Absoluto Medio (MAE), es el promedio de las diferencias absolutas entre el valor

real y el predicho. Se expresa en las mismas unidades que la variable objetivo (dólares), por lo que su interpretación es directa —representa el error promedio en dólares— y otorga el mismo peso a todos los errores:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Raíz del Error Cuadrático Medio (RMSE): raíz del promedio de los errores al cuadrado. También en USD, pero penaliza con más fuerza los errores grandes, por lo que es sensible a valores atípicos.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Coeficiente de Determinación (R^2): proporción de la varianza del precio explicada por el modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

En conjunto, R^2 mide la bondad de ajuste relativa, mientras que MAE y RMSE expresan el error en dólares: MAE da el error típico y RMSE alerta sobre errores grandes.

3.6.2. Comparación de resultados

Se presenta una tabla en donde se pueden observar de mejor manera los resultados obtenidos, se presenta en orden del que fue más efectivo al que menos.

Dataset	Ranking	Modelo	R2_test	RMSE_USD	MAE_USD
Plusvalia	**1**	**Random Forest**	**0.808**	**118,232**	**61,016**
Plusvalia	2	Arbol Decision	0.720	143,058	81,238
Plusvalia	3	MLP	0.698	148,364	87,392
Plusvalia	4	Reg. Lineal	0.687	151,181	92,033
Properati	**1**	**Random Forest**	**0.693**	**90,800**	**57,400**
Properati	2	Arbol Decision	0.658	95,715	60,939
Properati	3	Reg. Lineal	0.590	104,876	68,386
Properati	4	MLP	0.590	104,897	68,414
Unificado	**1**	**Random Forest**	**0.783**	**108,810**	**58,929**
Unificado	2	Arbol Decision	0.688	130,683	73,506
Unificado	3	MLP	0.685	131,110	75,268
Unificado	4	Reg. Lineal	0.673	133,658	78,342

Ilustración 10 Ranking del mejor modelo

3.7. Fase 6: Despliegue

La fase de despliegue constituye el cierre del pipeline automatizado: su objetivo es poner el modelo entrenado a disposición de los usuarios finales, de manera que el conocimiento generado pueda utilizarse de forma operativa para estimar el precio de nuevas propiedades. En este proyecto, el despliegue se materializó en dos componentes: la serialización del modelo ganador y el diseño de un flujo de inferencia.

3.7.1. Serialización del modelo

Una vez seleccionado el algoritmo más preciso que dado el caso es Random Forest, este se entrenó por última vez utilizando la totalidad del conjunto de datos disponible, con el fin de aprovechar toda la información para la versión de producción. El modelo resultante se serializó, es decir, se almacenó como un objeto binario en un archivo de disco mediante la librería joblib, especialmente eficiente para objetos de scikit-learn que contienen grandes arreglos numéricos. Alternativamente, la librería estándar pickle de Python permite realizar la misma operación.

La serialización es un paso fundamental porque evita tener que reentrenar el modelo desde cero cada vez que se necesita una predicción un proceso que consume tiempo y recursos de cómputo. En su lugar, el modelo se entrena una sola vez y, a partir de ese momento, es suficiente con cargar el archivo en memoria para utilizarlo de inmediato. Un aspecto clave del

diseño es que no se guardó únicamente el algoritmo, sino el pipeline completo, que incluye las etapas de preprocesamiento junto con el modelo entrenado. De este modo se garantiza que, en producción, los datos de entrada reciban exactamente las mismas transformaciones aplicadas durante el entrenamiento, eliminando inconsistencias.

3.7.2. Diseño del flujo de inferencia

El flujo de inferencia define como un usuario o un sistema externo puede obtener una predicción de precio de forma instantánea a partir de las características de una vivienda. Este flujo se implementó en un script de Python independiente y consta de cuatro etapas:

1. Carga del modelo Al iniciarse, el script carga en memoria el modelo serializado mediante joblib.load(). Esta operación se realiza una única vez, de modo que las predicciones posteriores son prácticamente instantáneas.
2. Ingreso de datos. El usuario proporciona las características de la propiedad —de forma obligatoria el metraje (m^2) y la ubicación (sector de la ciudad), y de forma opcional el número de dormitorios, baños y el tipo de inmueble. La entrada puede realizarse por argumentos de línea de comandos o mediante un modo interactivo que solicita los datos por teclado. Cuando un parámetro opcional se omite, el script asigna automáticamente el valor por defecto definido en los metadatos.
3. Construcción del registro. Los datos ingresados se organizan en una estructura tabular con exactamente las mismas columnas que se emplearon durante el entrenamiento. El pipeline serializado se encarga internamente de aplicar el preprocesamiento.
4. Predicción y respuesta. El modelo procesa el registro y devuelve el precio estimado en dólares, que se presenta al usuario de manera inmediata.

CAPÍTULO IV: ANÁLISIS DE RESULTADOS

En este capítulo se presentan y analizan los resultados obtenidos después de aplicar todo el proceso descrito en el capítulo anterior. Primero se detalla el entorno con el que se trabajó y la forma en que se organizó el proyecto. Luego se describe el conjunto de datos final y se muestran los resultados de los modelos entrenados, comparando su desempeño para identificar cuál fue el más preciso. Finalmente, se analiza qué tan bien predice el modelo elegido y se muestran algunos ejemplos reales de estimación de precios.

4.1. Configuración del entorno de desarrollo

Todo el desarrollo se realizó en el lenguaje Python, apoyándose en varias librerías especializadas según la tarea. Para el manejo y la limpieza de los datos se emplearon pandas y numpy. El entrenamiento y la evaluación de los modelos se hicieron con scikit-learn, que es una de las herramientas más usadas en aprendizaje automático. La extracción de la información desde los portales inmobiliarios se apoyó en Selenium, que permite automatizar un navegador real. Para guardar el modelo final ya entrenado se utilizó joblib, y las gráficas se elaboraron con matplotlib.

El trabajo se llevó a cabo dentro de cuadernos de Jupyter, un entorno que permite combinar el código, sus resultados y los comentarios en un mismo documento, lo cual facilita seguir el proceso paso a paso. El proyecto se ejecutó en un computador personal de gama media, cuya capacidad fue suficiente para el volumen de datos manejado, sin necesidad de recurrir a servidores especializados.

4.1.1. Estructura del proyecto

El proyecto se organizó en componentes separados, cada uno con una función clara dentro del flujo de trabajo, de modo que fuera fácil de mantener y de entender:

- Un script de extracción encargado de recolectar la información desde los portales web.
- Los archivos de datos en crudo, tal como se obtuvieron de las páginas, sin ningún tratamiento.
- Los archivos de datos ya limpios y estandarizados, listos para el análisis.
- El cuaderno de modelado, donde se entrenan y comparan los algoritmos.
- El modelo final guardado en un archivo, para poder reutilizarlo sin volver a entrenarlo.
- Un script de predicción que toma los datos de una vivienda nueva y devuelve su precio estimado.

4.2. Conjunto de datos final

Después de aplicar la integración y la limpieza descritas en el capítulo anterior, se obtuvo un conjunto de datos consolidado con 9.644 propiedades. Este dataset reúne la información de los dos portales en una sola estructura, con seis variables que describen cada inmueble. La siguiente tabla resume el significado de cada una:

Variable	Descripción
Tipo de inmueble	Categoría a la que pertenece la propiedad, como casa o departamento.
Precio	Valor de venta publicado, expresado en dólares estadounidenses.
Ubicación	Sector de la ciudad donde se encuentra el inmueble.
Dormitorios	Número de dormitorios de la propiedad.
Baños	Número de baños de la propiedad.
Área	Superficie del inmueble en metros cuadrados.

Tabla 2 Variables que componen el conjunto de datos final.

La variable que se quiere predecir es el precio, mientras que las demás funcionan como características que ayudan al modelo a estimarlo. Cabe señalar que los precios corresponden a valores de oferta publicados por los anunciantes, expresados en dólares, que es la moneda oficial del Ecuador.

4.3. Resultados del entrenamiento de los modelos

Se entrenaron cuatro algoritmos diferentes sobre tres conjuntos de datos: el de Plusvalía, el de Properati y el unificado que combina ambos. Para medir su desempeño se usaron tres indicadores. El R^2 muestra qué porcentaje del comportamiento de los precios logra explicar el modelo, donde un valor cercano a uno es mejor. El RMSE y el MAE miden el error de las predicciones en dólares, por lo que mientras más bajos, mejor. Los resultados se resumen en la siguiente tabla, donde se resalta el mejor modelo de cada conjunto:

Conjunto	Modelo	R^2	RMSE (USD)	MAE (USD)
Plusvalia	Regresión Lineal	0,687	151.181	92.033
Plusvalia	Árbol de Decisión	0,720	143.058	81.238
Plusvalia	**Random Forest**	**0,808**	**118.232**	**61.016**
Plusvalia	Red Neuronal	0,698	148.364	87.392
Properati	Regresión Lineal	0,590	104.876	68.386

Properati	Árbol de Decisión	0,658	95.715	60.939
Properati	**Random Forest**	**0,693**	**90.800**	**57.400**
Properati	Red Neuronal	0,590	104.897	68.414
Unificado	Regresión Lineal	0,673	133.658	78.342
Unificado	Árbol de Decisión	0,688	130.683	73.506
Unificado	**Random Forest**	**0,783**	**108.810**	**58.929**
Unificado	Red Neuronal	0,686	131.110	75.268

Tabla 3 Desempeño de los cuatro algoritmos en los tres conjuntos de datos.

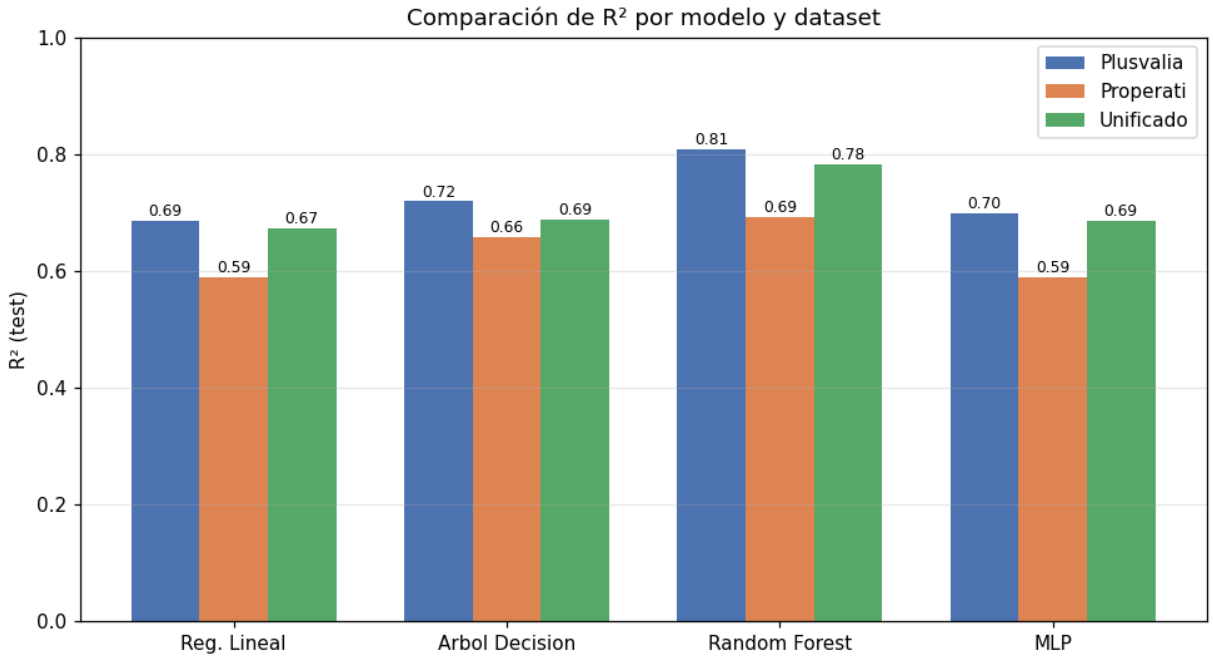


Ilustración 11 Comparación del R² alcanzado por cada modelo en los tres conjuntos de datos

Tanto en la tabla como en la gráfica se observa con claridad que el Random Forest fue el modelo con mejor rendimiento en los tres casos. Logró el mayor R² y, al mismo tiempo, los errores más bajos. La regresión lineal y la red neuronal obtuvieron resultados parecidos entre sí y bastante por debajo, mientras que el árbol de decisión quedó en una posición intermedia.

4.4. Selección del mejor modelo

A partir de estos resultados se seleccionó el Random Forest como modelo definitivo del proyecto. Su mejor desempeño se explica porque combina muchos árboles de decisión y promedia sus respuestas, lo que le permite captar relaciones complejas entre el tamaño, la ubicación y el precio sin equivocarse tanto como un solo árbol. El mejor resultado se dio en el

conjunto de Plusvalía, formado únicamente por casas, donde el modelo alcanzó a explicar alrededor del ochenta por ciento del comportamiento de los precios. Esto tiene sentido, ya que es un conjunto más homogéneo y con mayor cantidad de registros.

4.5. Importancia de las variables

Una ventaja del Random Forest es que permite conocer cuánto influyó cada variable en sus predicciones. Como muestra la siguiente gráfica, el área de la propiedad fue, de lejos, el factor más determinante, seguido por la ubicación. El número de baños, de dormitorios y el tipo de inmueble tuvieron un peso mucho menor.

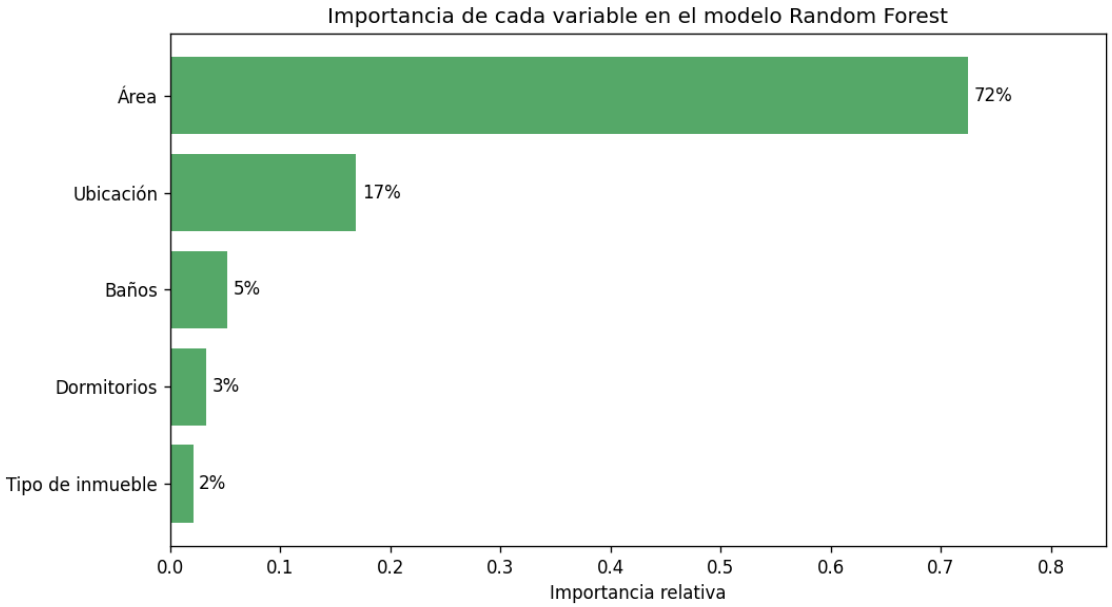


Ilustración 12 Importancia de cada variable en las predicciones del modelo Random Forest

Este resultado coincide con la lógica del mercado inmobiliario: el tamaño de una vivienda es lo que más define su valor, pero la zona en la que se encuentra marca una diferencia importante. Que la ubicación aparezca en segundo lugar confirma lo que ya se había observado en el análisis por sector, donde los precios cambiaban notablemente de un barrio a otro.

4.6. Capacidad predictiva del modelo

Para evaluar qué tan acertadas son las estimaciones, se compararon los precios reales con los precios que predijo el modelo sobre propiedades que no había visto antes. En la siguiente gráfica, cada punto es una vivienda; mientras más cerca esté de la línea diagonal, más exacta fue la predicción.

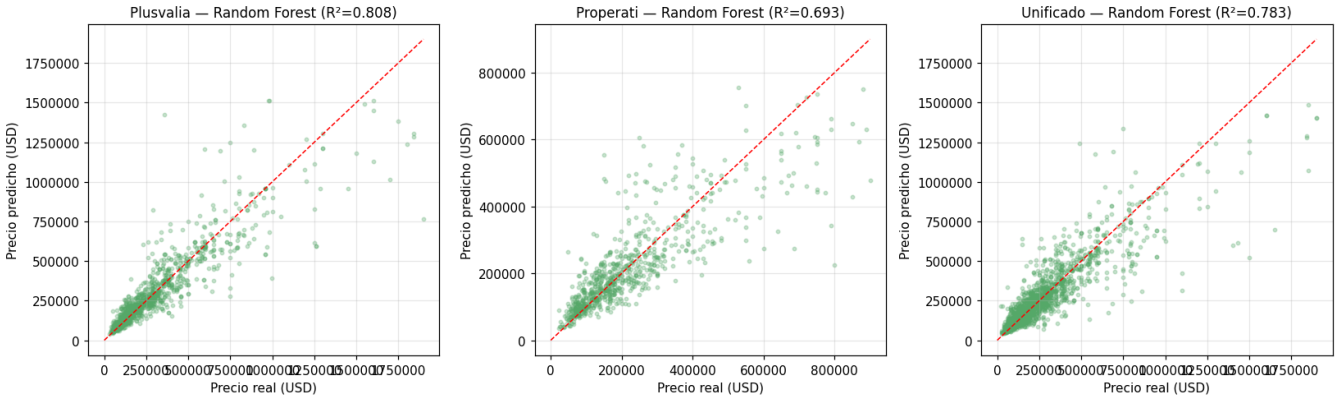


Ilustración 13 Comparación entre el precio real y el precio estimado por el modelo en cada conjunto de datos.

Se aprecia que la mayoría de los puntos se agrupa alrededor de la línea, lo que indica que el modelo acierta bastante bien en el rango de precios más común. Las diferencias más grandes aparecen en las propiedades de precios muy altos, que son menos frecuentes y, por lo tanto, más difíciles de estimar. En promedio, el error del modelo se mantiene dentro de un margen razonable para una herramienta de estimación inicial, que sirve como una referencia confiable más que como una tasación exacta.

4.7. Estimación de precios con el modelo final

Una vez entrenado y guardado, el modelo quedó listo para usarse de forma inmediata. Basta con ingresar las características de una vivienda, principalmente su metraje y su ubicación, para obtener al instante un precio estimado, sin necesidad de repetir todo el proceso de entrenamiento. La siguiente tabla muestra algunos ejemplos de estimaciones generadas por el modelo:

Tipo	Área	Ubicación	Precio estimado
Casa	180 m²	Cumbayá	$ 109.423
Casa	120 m²	Tumbaco	$ 119.779
Departamento	90 m²	Centro Norte	$ 131.436

Tabla 4 Ejemplos de precios estimados por el modelo final.

Estos resultados demuestran que el modelo responde de manera coherente con el mercado: asigna precios más altos a las casas amplias de los valles y reconoce el alto valor por metro de los departamentos en zonas céntricas. Con esto, el objetivo de contar con una

herramienta funcional de estimación de precios queda cumplido.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

En este capítulo se recogen las conclusiones a las que se llegó después de desarrollar el proyecto y se comparan con los objetivos que se plantearon al inicio. La idea es mostrar de forma clara qué se consiguió, qué dificultades aparecieron en el camino y qué aprendizajes deja el trabajo. Al final se incluyen algunas recomendaciones pensadas para quienes quieran retomar o mejorar esta propuesta más adelante.

5.1. Conclusiones

Se cumplió el objetivo general del proyecto, ya que se logró construir un modelo capaz de estimar el precio de una vivienda en Quito a partir de información real tomada de internet. De los cuatro algoritmos que se probaron, el Random Forest fue el que mejor respondió. En el caso de las casas consiguió explicar alrededor del ochenta por ciento del comportamiento de los precios, lo que representa un resultado sólido para una primera herramienta de estimación en el mercado local.

En cuanto al primer objetivo específico, se revisaron y organizaron los fundamentos teóricos sobre aprendizaje automático y sobre las técnicas de recolección de datos en la web aplicadas a la valoración de bienes raíces. Esta base teórica fue la que permitió tomar decisiones con criterio a lo largo de todo el desarrollo, desde la elección de los algoritmos hasta la forma de medir su rendimiento.

El segundo objetivo específico también se cumplió. Se construyó un flujo automatizado que extrae, limpia y organiza la información publicada en los portales Plusvalía y Properati. El mayor reto en esta parte fue la inconsistencia de los datos, ya que cada portal publica los precios, las ubicaciones y las características de manera distinta. Para resolverlo se aplicó un proceso de limpieza y homologación que dejó toda la información en una estructura única y lista para ser analizada.

Respecto al tercer objetivo específico, se entrenaron y compararon cuatro algoritmos usando métricas de error estándar. El Random Forest fue el más preciso en todos los conjuntos de datos analizados, por encima de la regresión lineal, el árbol de decisión y la red neuronal. Por esta razón fue el modelo seleccionado como solución final. Los resultados confirmaron algo que ya señala la literatura sobre el tema, y es que para este tipo de datos los modelos basados en

árboles suelen rendir mejor que los modelos más simples o que las redes neuronales.

El análisis de los datos dejó además un hallazgo interesante sobre el mercado de Quito. El tamaño de la propiedad es el factor que más influye en su precio, pero la ubicación ocupa el segundo lugar y marca diferencias muy claras entre sectores. Zonas como el Valle de Tumbaco, Cumbayá o Nayón concentran los precios más altos por metro cuadrado, mientras que sectores del sur y de la periferia norte resultan considerablemente más accesibles. Esto demuestra que el modelo no solo predice un número, sino que aprende la misma lógica geográfica que conoce cualquier persona familiarizada con la ciudad.

Por último, el proyecto se cerró con una herramienta funcional. El modelo entrenado se guardó en un archivo reutilizable y se creó un programa sencillo que, al ingresar el metraje y la ubicación de una casa, devuelve un precio estimado de inmediato y sin necesidad de volver a entrenar nada. Con esto se alcanza el resultado esperado dentro del alcance del trabajo, que era contar con un modelo predictivo completamente operativo.

5.2. Recomendaciones

Se recomienda incorporar nuevas variables que también influyen en el precio de una vivienda y que en este trabajo no se consideraron, como la antigüedad del inmueble, el número de parqueaderos o su estado de conservación. Incluir estos datos ayudaría a que las estimaciones sean todavía más exactas.

Conviene actualizar la información de manera periódica y volver a entrenar el modelo cada cierto tiempo. El mercado inmobiliario cambia constantemente, por lo que un modelo entrenado con datos antiguos puede perder precisión si no se mantiene al día.

Sería útil unificar los nombres de los sectores, ya que cada portal nombra de forma distinta zonas que en realidad son la misma o que están muy cerca. Una estandarización geográfica más fina permitiría obtener un mapa de precios más limpio y comparaciones más justas entre barrios.

Como siguiente paso natural, se recomienda llevar la herramienta a una interfaz web sencilla para que cualquier usuario pueda consultarla desde su navegador, sin necesidad de conocimientos técnicos. Esto acercaría el proyecto a compradores, vendedores y agentes que son, al final, los principales beneficiarios.

Finalmente, se sugiere validar las estimaciones del modelo con precios reales de cierre y con la opinión de tasadores profesionales. Los datos utilizados corresponden a precios de oferta, que suelen ser algo más altos que el valor final de venta, por lo que esta comparación ayudaría a calibrar mejor la herramienta y a darle mayor confianza de cara a un uso real.

BIBLIOGRAFÍA

Bodero Poveda, E. M., & Guerrero Morejón, K. A. (2023). Aplicación del método ARIMA para el pronóstico del avalúo de bienes raíces. Perspectivas Sociales y Administrativas, 1(1), 26-43.

Gorjian, M. (2025). Integrating machine learning and hedonic regression for housing price prediction: A systematic international review of model performance and interpretability (MPRA Paper No. 125676). Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/125676/

Inmogesco. (2024). ¿Qué es el mercado inmobiliario?. Recuperado de https://inmogesco.com/blog/que-es-el-mercado-inmobiliario/

Jara Flórez, H. (s.f.). Clasificación de bienes muebles e inmuebles. Dialnet. Recuperado de https://dialnet.unirioja.es

Medina Giraldo, A. T. (2023). Predicción de los precios de vivienda en la ciudad de Medellín y el Área Metropolitana [Tesis de maestría, Universidad Internacional de La Rioja]. Re-UNIR.

Méndez Mangana, C., & Barraquer Kargacin, A. (2023). Introducción a la creación de un conjunto de datos para machine learning. Annals d'Oftalmologia, 31(4), 177-184.

Quito Cómo Vamos. (2024). Informe de Calidad de Vida 2024. Recuperado de https://quitocomovamos.org

Red de Innovación Turística. (2023). Web Scraping: Iniciación. Recuperado de https://red-intur.org

Universidad de Las Américas [UDLA]. (2024). Creación de un modelo predictivo basado en inteligencia artificial [Tesis de maestría, Universidad de Las Américas]. Repositorio Institucional UDLA.

Universidad Internacional de Andalucía [UNIA]. (2026). Machine Learning: concepto, beneficios y aplicaciones. Recuperado de https://www.unia.es/vida-universitaria/blog/machine-learning-que-es

Vilca, W. (2022). Análisis automático para el establecimiento de un método basado en el uso de técnicas de Machine Learning para predecir el valor de un inmueble en la ciudad de Quito.