UNIVERSIDAD CENTRAL	FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS MAESTRÍA EN ANALÍTICA DE DATOS REGISTRO PROPUESTA DE TRABAJO DE GRADO 2024-II

REGISTRO DE TRABAJO DE GRADO

FECHA			2024

DATOS DEL ESTUDIANTE (S)

NOMBRES: Juan Sebastian							APELLIDOS: Blanco Peña
TIPO IDENTIFICACIÓN:	T.I.		C.C.	X	C.E.		NÚMERO: 1022417508
CORREO INSTITUCIONAL: jblancop@ucentral.edu.co							TELÉFONO: 3023469617

NOMBRES: Carlos Santiago							APELLIDOS: Acosta Achury
TIPO IDENTIFICACIÓN:	T.I.		C.C.	X	C.E.		NÚMERO: 1233688342
CORREO INSTITUCIONAL: cacostaa1@ucentral.edu.co							TELÉFONO: 3118968568

NOMBRES: Samuel Stiben							APELLIDOS: Suescun Hernandez
TIPO IDENTIFICACIÓN:	T.I.		C.C.	X	C.E.		NÚMERO: 1022427440
CORREO INSTITUCIONAL:							TELÉFONO: 3208283997

MODALIDAD DE TRABAJO DE GRADO (Seleccione una opción)

	II. Modalidad de profundización: ☐
	a. Trabajo monográfico ☐

Línea de profundización

AVAL DEL DOCENTE DIRECTOR

NOMBRES: LUIS ANDRES CAMPOS MALDONADO	DEPARTAMENTO: INGENIERÍA Y CIENCIAS BÁSICAS
CORREO INSTITUCIONAL: lcamposm@ucentral.edu.co	TELÉFONO-EXT. : +573016627377

COMPONENTES

1. TÍTULO DEL TRABAJO DE GRADO

Sistema Predictivo de Precios de Inmuebles mediante Web Scraping y Machine Learnin

2. INTRODUCCIÓN Y JUSTIFICACIÓN (máximo 1500 palabras)

El mercado inmobiliario es uno de los sectores económicos de mayor dinamismo y relevancia en Colombia, tan solo en el 2024, según el DANE "el sector inmobiliario aportó entre el 8.5% y 8.8% del PIB, lo que equivale a más de 125 billones de pesos". (La Nota Económica, 2024)

Así, tener una aproximación a los costos de viviendas y apartamentos es esencial tanto para consumidores y comerciantes como para entidades financieras. Sin embargo, la información en diversos portales (como Metro Cuadrado, 100 Cuadras y Finca Raíz) complica la adquisición de datos consistentes que faciliten la aplicación de comparaciones fiables.

Frente a esta situación, se quiere crear una solución completa que incluya la extracción automatizada de datos (scraping web) de portales web especializadas, la depuración y tratamiento de la información, y la creación de un modelo de aprendizaje automático (regresión lineal, árboles de decisión o redes neuronales.) enfocado en estimar los precios de propiedades de acuerdo con factores como la zona, cantidad de baños, metros cuadrados, entre otros. Además, se proyecta el desarrollo de una interfaz de usuario o chatbot que facilite la consulta interactiva de la información producida.

La importancia práctica de este instrumento reside en su habilidad para respaldar la toma de decisiones de inversores, instituciones financieras y posibles compradores, proporcionando datos exactos y a tiempo. Adicionalmente, la innovación tecnológica se manifiesta en la fusión de métodos de web scraping, procesamiento de datos y modelos predictivos, lo que facilita la adopción de un enfoque multidisciplinario en proporción con las tendencias presentes en la analítica de datos. Por último, la contribución académica del proyecto se manifiesta en su aporte al saber en el área de la analítica aplicada al sector de bienes raíces, ofreciendo metodologías y hallazgos que podrán ser utilizados en futuros estudios o usos comerciales

3. OBJETIVO GENERAL

Desarrollar un sistema que, a través de la extracción automatizada de datos de portales inmobiliarios, el procesamiento y análisis de la información, y la aplicación de técnicas de machine learning, permita predecir los precios de casas y apartamentos, complementado con una interfaz para la consulta interactiva de resultados.

4. OBJETIVOS ESPECÍFICOS

1. Recolección y tratamiento de datos: Utilizar técnicas de web scraping para obtener datos pertinentes (ubicación, características, precios, etc.) de portales de bienes raíces, después de un proceso de limpieza y normalización de la información.
2. Elaborar y valorar modelos de predicción: Elaborar y capacitar modelos de aprendizaje automatizado que faciliten la predicción del precio de propiedades basándose en las variables recolectadas, comprobando su rendimiento a través de métricas estándar.
3. Proporcionar una interfaz interactiva: Elaborar y poner en marcha una aplicación web o chatbot que habilite a los usuarios para consultar y visualizar las proyecciones, facilitando así la toma de decisiones.

5. ANTECEDENTES Y MARCO TEÓRICO (máximo 3000 palabras)

En años recientes, se ha notado un incremento en el interés por utilizar métodos de análisis de datos y aprendizaje automático para estimar los precios de propiedades. Varias investigaciones a nivel mundial han evidenciado que los modelos predictivos fundamentados en algoritmos como la regresión lineal, árboles de decisión, Random Forest y redes neuronales pueden calcular eficientemente el valor de las propiedades residenciales (Antipov & Pokryshevskaya, 2012; Peterson & Flanagan, 2009). Este interés ha crecido paralelamente al incremento de la disponibilidad de información en internet, donde sitios especializados divulgan datos precisos que son útiles para la evaluación de bienes raíces.

En el contexto latinoamericano, y particularmente en Colombia, se han realizado iniciativas que combinan la extracción de datos mediante web scraping y el análisis predictivo. Por ejemplo, proyectos que han utilizado información extraída de portales inmobiliarios (como Metro Cuadrado, 100 Cuadras y Finca Raíz) han permitido consolidar bases de datos actualizadas para el desarrollo de modelos de predicción (Nunes Ariza & Manrique Piramanrique, 2024). Estos estudios resaltan que, a pesar de la volatilidad del mercado inmobiliario y la complejidad de sus múltiples variables —desde características estructurales del inmueble hasta factores socioeconómicos de la zona—, el uso de modelos de aprendizaje automático mejora significativamente la precisión en la estimación del precio de inmuebles, tanto para venta como para arriendo.

Asimismo, trabajos desarrollados en otros países han validado la aplicación de técnicas de web scraping para recolectar grandes volúmenes de datos no estructurados, que posteriormente son procesados y transformados en variables de entrada para los modelos predictivos (Picardo, 2019; Sharma et al., 2024). La combinación de ambas metodologías, la obtención automatizada de datos y el modelado predictivo ha demostrado ser una alternativa robusta frente a los métodos tradicionales de tasación, los cuales suelen depender en gran medida de la experiencia subjetiva de un tasador y de comparaciones aisladas de mercado.

Además, en ciudades como Bogotá o Medellín, estudios han mostrado que factores como la cercanía a servicios de transporte masivo, la accesibilidad vial o el estrato socioeconómico pueden

influir de manera significativa en la formación del precio de los inmuebles (Chica & Botero, 2023). Estas características refuerzan la necesidad de contar con bases de datos amplias y actualizadas que permitan al modelo capturar la heterogeneidad de los distintos barrios o localidades, superando así la visión tradicional basada en promedios generales.

Es importante destacar que el uso de técnicas de web scraping no solo se limita al ámbito inmobiliario. Diversos estudios en áreas como el comercio electrónico y la monitorización de precios han utilizado esta técnica para extraer información en tiempo real, lo que respalda la viabilidad de su aplicación en el sector de bienes raíces (Caraffa, 2022). Estos antecedentes confirman que el enfoque propuesto para el presente proyecto se encuentra en línea con tendencias actuales y ofrece un potencial real para mejorar la toma de decisiones en el mercado inmobiliario, mediante la disponibilidad de estimaciones más objetivas y basadas en grandes volúmenes de datos.

Marco Teórico

El marco teórico de este proyecto se sustenta en dos ejes principales: las técnicas de extracción de datos (web scraping) y los modelos de aprendizaje automático aplicados a la predicción de precios de inmuebles.

Web Scraping y Recolección de Datos

El web scraping es una técnica que consiste en extraer información de páginas web de manera automatizada, convirtiendo datos no estructurados (generalmente en formato HTML) en datos estructurados que pueden almacenarse en bases de datos o archivos (Kho, 2018). En el contexto inmobiliario, esta técnica permite recolectar información relevante como precios, ubicación, características físicas (número de habitaciones, baños, área en metros cuadrados) y otros atributos publicados en portales especializados. También es factible capturar elementos contextuales, como la disponibilidad de áreas comunes o la cercanía a puntos de interés, si el portal provee dicha información.

Diversas herramientas y librerías en Python —como BeautifulSoup y Selenium— han sido ampliamente utilizadas para implementar procesos de scraping. Estas herramientas facilitan la navegación por la estructura de las páginas web y permiten identificar las etiquetas HTML que contienen la información de interés (Van Rossum, 1989). La automatización de este proceso es fundamental para actualizar constantemente la base de datos, ya que el mercado inmobiliario es altamente dinámico y la información publicada en línea se modifica de forma continua. Además, implementar estrategias de control de acceso (por ejemplo, respetar los Términos de Uso de cada portal y hacer uso de retrasos o rotación de IP) es esencial para evitar bloqueos o acciones en contra del scraping.

Modelos Predictivos y Aprendizaje Automático

El segundo pilar del proyecto es la aplicación de modelos de aprendizaje automático para predecir el precio de los inmuebles. Entre los enfoques más utilizados se encuentra el modelo hedónico de valoración, que postula que el precio de una vivienda puede descomponerse en el valor implícito de sus características observables (Rosen, 1974). Este modelo, basado típicamente en una regresión lineal, ha sido la base de muchos estudios en valoración inmobiliaria, aunque presenta limitaciones al capturar relaciones no lineales o interacciones complejas entre las variables.

Para superar estas limitaciones, se han propuesto alternativas basadas en algoritmos de Machine Learning. Por ejemplo, los árboles de decisión y sus variantes (Random Forest, Gradient Boosting) permiten modelar relaciones complejas y no lineales, ofreciendo una mayor capacidad predictiva y robustez ante variaciones en los datos (Park & Bae, 2015). Asimismo, las redes neuronales han mostrado resultados prometedores en contextos donde las interacciones entre las variables son altamente complejas (Sharma et al., 2024), aunque su interpretabilidad puede ser menor en comparación con modelos basados en árboles.

La selección de un modelo adecuado depende de diversos factores, entre ellos la calidad y cantidad de datos recolectados, la relevancia de las variables seleccionadas y las métricas de evaluación utilizadas. En este sentido, se emplearán indicadores como el error cuadrático medio (RMSE), el coeficiente de determinación (R^2) y el error absoluto medio (MAE) para evaluar y comparar el desempeño de los distintos modelos. Estas métricas permiten cuantificar la precisión del modelo y su capacidad para generalizar sobre nuevos datos, lo que resulta esencial si el objetivo es contar con una herramienta útil para la toma de decisiones. En algunos casos, también puede contemplarse una clasificación de inmuebles en rangos de precio (bajo, medio, alto), lo que implicaría evaluar métricas de clasificación como Accuracy o Recall.

Integración de Datos, Validación y Desarrollo de la Interfaz

Una vez que los datos han sido extraídos y procesados, es necesario integrarlos en un solo repositorio que sirva de base para el modelado. La limpieza de datos, que incluye la eliminación de duplicados, el manejo de valores nulos y la normalización de variables, es una etapa crítica para garantizar la calidad del conjunto de datos y, en consecuencia, el rendimiento de los modelos predictivos (Raschka & Mirjalili, 2017). Adicionalmente, se pueden incorporar variables contextuales —por ejemplo, información georreferenciada y demográfica— para refinar aún más las estimaciones.

Como parte del proceso de validación, es recomendable diseñar experimentos de Cross-Validation (K-Fold) que ayuden a prevenir el sobreajuste y a evaluar la estabilidad del modelo ante diferentes subconjuntos de datos. Asimismo, el uso de un conjunto de prueba totalmente independiente permite medir la capacidad de generalización del modelo en escenarios que simulan datos reales futuros, un aspecto de gran relevancia en un mercado tan dinámico como el inmobiliario.

Por otro lado, el proyecto contempla el desarrollo de una interfaz o chatbot que permita a los usuarios consultar de manera interactiva las predicciones generadas por el modelo. Esta herramienta no solo facilitará el acceso a la información, sino que también contribuirá a la toma de decisiones de inversionistas, entidades financieras y potenciales compradores. El uso de tecnologías web modernas que incluyan visualizaciones claras (por ejemplo, paneles interactivos o gráficos de dispersión) ofrece a los usuarios una experiencia más intuitiva a la hora de analizar y comparar resultados. Con ello se busca que tanto expertos como usuarios sin formación técnica puedan aprovechar la solución.

En resumen , la combinación de web scraping y aprendizaje automático ofrece una solución innovadora y robusta para la predicción de precios de inmuebles. Los antecedentes revisados confirman la eficacia de estas técnicas en distintos contextos, haciendo hincapié en su utilidad para capturar relaciones complejas entre las características de los inmuebles y el precio final. El marco teórico expuesto justifica el uso de modelos predictivos avanzados y estrategias de recolección de datos automatizadas, destacando la importancia de evaluar cuidadosamente la calidad del dataset y de aplicar metodologías de validación que minimicen el riesgo de sobreajuste.

Asimismo, se resalta la relevancia de integrar estas técnicas en una solución amigable para el usuario final, a fin de facilitar la toma de decisiones en el mercado inmobiliario. La posibilidad de contar con una herramienta que permita predecir valores de compra o de arriendo con un margen de error aceptable y que, además, se actualice con la información más reciente de los portales

inmobiliarios, incrementa la transparencia y eficiencia en las negociaciones. De esta forma, el proyecto se alinea con las tendencias globales en analítica de datos, al tiempo que atiende necesidades específicas del mercado colombiano y, en particular, del entorno urbano de Bogotá.

6. FUENTE DE LOS DATOS (Cómo se obtendrán los datos)

Los datos serán obtenidos de sitios web de reconocida relevancia en el sector inmobiliario en Colombia, tales como:

- **Metro Cuadrado**
- **100 Cuadras**
- **Finca Raíz**
- **Properati**

Se utilizarán técnicas de web scraping, implementadas en Python (usando librerías como BeautifulSoup y/o Selenium), para extraer información acerca de propiedades en venta y arriendo.

Cada registro (fila) del dataset representará una propiedad e incluirá información clave como:

Título del anuncio: breve descripción que destaca atributos llamativos del inmueble.

Precio: valor publicado en pesos colombianos (COP), incluyendo posibles indicaciones sobre negociabilidad.

Ubicación: barrio o sector, determinante en la valoración inmobiliaria.

Superficie: área total en metros cuadrados, tanto construidos como privados.

Número de habitaciones y baños: características esenciales para evaluar la propiedad.

Tipo de propiedad: apartamento, casa, penthouse, etc.

Antigüedad del inmueble: nuevo o usado, aspecto influyente en el precio.

Amenidades: existencia de parqueadero, balcón, áreas comunes, piscina o gimnasio.

Descripción del vendedor: texto libre del anuncio útil para análisis de lenguaje natural (NLP).

URL del anuncio: enlace directo al inmueble para validaciones adicionales.

La mayoría de estos datos son de libre acceso; sin embargo, se verificarán los términos de uso de cada sitio para asegurar el cumplimiento de normativas.

7. APLICACIÓN Y/O APORTE ESPECÍFICO AL CAMPO

A continuación, se exponen los principales aportes que este proyecto brinda al campo de la **analítica de datos** y al **sector inmobiliario**, detallados en formato de ítems para mayor claridad:

1. **Optimización de la toma de decisiones**
   - La integración de **web scraping** y **modelos predictivos** permite una estimación más precisa de los precios de inmuebles (compra y arriendo).
   - Inversionistas, compradores particulares y entidades financieras dispondrán de información objetiva que reduce la incertidumbre en la fijación de precios y en la evaluación de oportunidades de inversión.
2. **Impulso a la innovación en el sector inmobiliario**
   - El proyecto emplea **técnicas avanzadas de Machine Learning** (Random Forest, LightGBM, redes neuronales, entre otras) para procesar grandes volúmenes de datos extraídos de portales en línea.
3. **Desarrollo de una interfaz de usuario accesible**
   - Mediante una **aplicación web o chatbot**, se facilita la interacción con el sistema de predicción, permitiendo que usuarios sin formación técnica consulten información de manera ágil.
   - Se prevén funcionalidades como **búsquedas personalizadas**, visualización de métricas (errores de predicción, variables más influyentes) y comparativas de precios por zonas.
4. **Generación de conocimiento y metodologías replicables**
   - El enfoque metodológico (CRISP-DM) y el uso de librerías para **web scraping** y análisis estadístico/ML en Python pueden ser adaptados a otros contextos o geografías con características inmobiliarias similares.
   - Se sientan bases para futuros estudios en analítica de datos aplicada a mercados dinámicos, extendiendo la metodología a escalas mayores o incorporando nuevas variables (índices macroeconómicos, proyecciones de crecimiento urbano, etc.).

8. METODOLOGÍA O ACTIVIDADES ESPECIFICAS

El proyecto se desarrollará a través de un enfoque estructurado en cinco fases, fundamentado en la metodología CRISP-DM, adaptada específicamente al análisis y predicción de precios inmobiliarios mediante técnicas de web scraping y aprendizaje automático.

Fase 1: Comprensión del problema

En el contexto actual, el mercado inmobiliario presenta una dinámica altamente compleja y volátil, caracterizada por la dispersión y heterogeneidad de los precios de inmuebles que se ofertan en diversas plataformas digitales. Esta variabilidad se debe a múltiples factores, tales como la ubicación, las características físicas de las propiedades, las condiciones socioeconómicas del entorno y la temporalidad de las transacciones. Sin embargo, la información disponible en línea se encuentra fragmentada y no siempre es sistematizada, lo que dificulta la obtención de estimaciones precisas y consistentes que apoyen la toma de decisiones en el ámbito de la inversión inmobiliaria.

El presente proyecto se propone abordar esta problemática a través del desarrollo de un sistema integral de análisis predictivo que utilice técnicas avanzadas de web scraping y aprendizaje automático. La finalidad es consolidar grandes volúmenes de datos provenientes de portales inmobiliarios (por ejemplo, Metro Cuadrado, Finca Raíz y 100 Cuadras), sometiéndolos a un riguroso proceso de limpieza y transformación, para finalmente entrenar modelos que estimen con alta precisión el precio de venta y arriendo de inmuebles en Bogotá.

Desde una perspectiva analítica, el primer paso consiste en comprender a fondo el problema: identificar las variables críticas que influyen en la valoración inmobiliaria, evaluar la calidad y disponibilidad de los datos, y definir los requisitos técnicos y funcionales que deberá satisfacer el sistema. Este proceso implica además la revisión de literatura especializada y estudios previos en el ámbito de la predicción de precios de bienes raíces, para fundamentar la elección de métodos y técnicas analíticas. Asimismo, se delimitará el alcance del proyecto, estableciendo las fronteras de análisis tanto en términos geográficos (por ejemplo, centrado en el mercado de Bogotá) como en la naturaleza de los datos (datos de libre acceso versus datos restringidos).

El éxito del proyecto dependerá de la capacidad de integrar y armonizar datos heterogéneos para generar insights accionables que no solo tengan relevancia académica, sino que también aporten valor práctico a inversionistas, entidades financieras y otros actores del sector inmobiliario. Esta fase inicial sienta las bases para la estructuración del sistema, orientando la metodología hacia la obtención de resultados robustos y escalables en un entorno de alta complejidad.

Fase 2: Comprensión de los datos

En esta fase se llevará a cabo una exploración exhaustiva del conjunto de datos recolectado, lo que resulta fundamental para comprender la calidad, el alcance y las limitaciones de la información que se utilizará en el modelado predictivo. Dado que los datos provienen de múltiples portales inmobiliarios mediante técnicas de web scraping, es imperativo evaluar su consistencia y relevancia para el análisis.

Inicialmente, se realizará una inspección minuciosa de los datos extraídos para identificar las variables clave (precio, ubicación, área, número de habitaciones, baños, entre otros) y evaluar la estructura y formato de cada campo. Se aplicarán técnicas de análisis exploratorio de datos (EDA) utilizando herramientas estadísticas y visuales, lo que permitirá detectar patrones, tendencias, valores atípicos y posibles inconsistencias. Este proceso incluye la generación de distribuciones de frecuencias, diagramas de caja (boxplots) y análisis de correlaciones para entender las relaciones interdependientes entre las variables.

Posteriormente, se procederá a la limpieza y transformación de los datos. Este paso abarca la eliminación de duplicados, el tratamiento de valores nulos y la normalización de formatos, asegurando que la información esté en condiciones óptimas para el análisis. Asimismo, se crearán variables derivadas que puedan potenciar el poder predictivo de los modelos, como ratios o indicadores compuestos que integren múltiples dimensiones de los inmuebles.

El conocimiento profundo de los datos obtenido en esta fase permitirá ajustar el proceso de extracción y preparación, garantizando la calidad de la información y reduciendo el ruido que podría afectar la eficiencia de los algoritmos de aprendizaje automático. Este análisis preliminar es esencial para fundamentar la selección de modelos y las estrategias de validación que se aplicarán en fases posteriores del proyecto.

Fase 3: Modelado y análisis

En esta fase se transformarán los datos brutos recolectados en la fase anterior en un conjunto de datos estructurado y apto para el modelado predictivo. Este paso es crucial, ya que la calidad y precisión de los modelos de aprendizaje automático dependen en gran medida del procesamiento y normalización de la información.

Se realizarán las siguientes actividades:

1. **Limpieza de Datos:**

   Se eliminarán registros duplicados y se corregirán errores en la entrada de datos. Se identificarán valores nulos o inconsistentes y se aplicarán técnicas de imputación—por ejemplo, sustitución por la mediana o métodos de imputación multivariante—para minimizar la pérdida de información y garantizar la integridad del dataset.

2. **Transformación y Estandarización:**

   Se convertirá la información a formatos homogéneos, normalizando variables (por ejemplo, fechas, unidades de medida) y codificando variables categóricas mediante técnicas como one-hot encoding o label encoding, permitiendo su procesamiento por algoritmos de machine learning. Además, se evaluará la posibilidad de generar variables derivadas que potencien la capacidad predictiva del modelo.

3. **Integración de Datos de Múltiples Fuentes:**

   Dado que los datos provienen de diversos portales inmobiliarios, se fusionarán en un único repositorio, asegurando la coherencia y compatibilidad entre las distintas fuentes. Este proceso incluirá la conciliación de variables con nomenclaturas diferentes y la resolución de conflictos en la información.

4. **Análisis Exploratorio Post-Preparación:**

   Una vez depurados y transformados, se realizará un análisis exploratorio para confirmar la calidad del dataset. Se generarán estadísticas descriptivas, distribuciones de frecuencia

y análisis de correlaciones, lo que permitirá detectar posibles anomalías residuales y validar la adecuación de los datos para el proceso de modelado.

El flujo de trabajo se implementará utilizando herramientas de procesamiento de datos en Python, como pandas, NumPy y Scikit-learn, asegurando un proceso reproducible, documentado y robusto que sirva de base sólida para la fase de modelado y análisis predictivo.

Fase 4. Modelado de datos

En esta fase se procederá a la construcción y validación de los modelos predictivos utilizando el conjunto de datos ya preparado y depurado. El objetivo es desarrollar algoritmos que permitan estimar con precisión el precio de los inmuebles en Bogotá, a partir de las variables relevantes extraídas de los portales inmobiliarios.

El proceso de modelado se dividirá en las siguientes actividades:

1. **Selección de Algoritmos:**

   Se evaluarán múltiples técnicas de aprendizaje automático, entre las que se incluyen modelos tradicionales como la regresión lineal y los árboles de decisión, así como métodos más complejos como Random Forest, Gradient Boosting y redes neuronales. La elección inicial se fundamentará en la capacidad de cada modelo para capturar relaciones no lineales y manejar la heterogeneidad de las variables.

2. **Entrenamiento y Ajuste de Hiperparámetros:**

   Cada algoritmo se entrenará utilizando una parte del conjunto de datos (por ejemplo, 80% para entrenamiento y 20% para validación) y se aplicarán técnicas de validación cruzada para evitar el sobreajuste. Se realizará un ajuste fino de los hiperparámetros mediante métodos como GridSearchCV, de modo que se optimice el desempeño del modelo en términos de precisión y generalización.

3. **Evaluación del Desempeño:**

   Se utilizarán métricas de evaluación robustas como el Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación (R^2) para comparar el rendimiento de los modelos. Estas métricas permitirán cuantificar la precisión en la estimación de los precios y seleccionar el modelo que logre el mejor balance entre exactitud y robustez frente a datos no vistos.

4. **Interpretación y Análisis de Resultados:**

   Una vez entrenados los modelos, se procederá a analizar la importancia de las variables y la estabilidad de las predicciones. Se utilizarán técnicas de interpretación, tales como análisis de importancia de características y gráficos de residuales, para entender el comportamiento del modelo y validar que los resultados sean coherentes con el conocimiento del mercado inmobiliario.

Esta fase es fundamental, ya que la calidad del modelo predictivo impactará directamente en la capacidad del sistema para ofrecer estimaciones precisas y, en consecuencia, en la toma de decisiones informadas en el ámbito inmobiliario. Los resultados obtenidos en esta etapa servirán de base para la integración final del modelo en la interfaz interactiva, asegurando un producto escalable y de alta aplicabilidad en el campo.

Fase 5: Evaluación y despliegue

En esta fase se validará integralmente el modelo predictivo y la interfaz desarrollada, asegurando que cumplan con los objetivos del proyecto y aporten valor práctico a los usuarios. Se emplearán los siguientes procesos:

1. **Validación del Modelo Predictivo:**

   Se utilizarán técnicas de validación cruzada para confirmar la robustez y estabilidad de los modelos entrenados. Se evaluará el desempeño mediante métricas clave como el RMSE, MAE y el coeficiente de determinación (R^2) en conjuntos de datos de entrenamiento y prueba. Asimismo, se analizarán los residuales y la importancia de las variables para garantizar que las predicciones sean coherentes y precisas.

2. **Pruebas de Usabilidad de la Interfaz:**

   Se diseñará un plan de pruebas orientado a usuarios finales (compradores, inversionistas y entidades financieras) para evaluar la facilidad de uso, la rapidez de respuesta y la claridad en la presentación de resultados de la aplicación web o chatbot. Se emplearán encuestas de satisfacción y estudios de usabilidad que permitan recoger feedback cualitativo y cuantitativo, facilitando ajustes en la experiencia de usuario.

3. **Análisis de Impacto y Retroalimentación:**

   Los resultados de la validación técnica y de las pruebas de usabilidad serán analizados de forma conjunta para identificar áreas de mejora en el modelo y la interfaz. Se documentarán las incidencias detectadas y se realizarán iteraciones de ajuste tanto en los algoritmos de predicción como en la presentación interactiva de la información, asegurando que el producto final sea escalable y se ajuste a las necesidades del mercado.

Este enfoque evaluativo, basado en técnicas estadísticas y en la retroalimentación directa de los usuarios, garantizará la implementación de una solución confiable y de alto rendimiento en el contexto inmobiliario.

9. RECURSOS

El desarrollo del presente trabajo de grado requiere la articulación de recursos tecnológicos, humanos y de infraestructura, indispensables para asegurar la ejecución rigurosa y eficiente del proyecto.

Recursos Tecnológicos

- **Infraestructura computacional personal**: cada uno de los integrantes cuenta con equipos portátiles con capacidad de procesamiento suficiente (mínimo 16 GB RAM, procesadores de al menos 4 núcleos) para el desarrollo de modelos de analítica de datos y procesos de scraping.
- **Lenguajes y entornos de programación**: se emplearán herramientas de código abierto como Python (con librerías como Scikit-learn, Pandas, BeautifulSoup, Flask), R y entornos como Jupyter Notebook y RStudio.
- **Herramientas de análisis y visualización**: se utilizarán Power BI, Tableau Public y Plotly para representar visualmente hallazgos clave que faciliten la interpretación por parte de usuarios no técnicos.
- **Plataformas de colaboración**: Google Drive, GitHub y Trello para control de versiones, planificación ágil y almacenamiento compartido de información.

Recursos Humanos

- El equipo de trabajo está conformado por tres estudiantes de la Maestría en Analítica de Datos, cada uno con formación previa en matemáticas, estadística y programación.
- El acompañamiento metodológico y conceptual estará a cargo del docente director asignado por la Facultad, quien guiará tanto los aspectos técnicos del modelado como los criterios de evaluación del producto final.

Recursos de Datos

- Se contará con acceso a sitios web públicos como Metrocuadrado, Finca Raíz y 100 Cuadras, desde donde se extraerán los datos mediante técnicas de scraping.
- Para el tratamiento de estos datos, se aplicarán técnicas de limpieza, imputación, transformación y enriquecimiento, asegurando la calidad y validez de la información empleada en el modelado.

Conectividad y Licencias

- Todos los integrantes del grupo cuentan con conexión estable a internet.
- No se requiere de licencias pagas adicionales para el desarrollo del proyecto, dado que se utilizarán herramientas de código abierto y de libre acceso.

10. CRONOGRAMA DE ACTIVIDADES

	Semanas de ejecución de cada actividad															
ACTIVIDADES A REALIZAR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Actividad 1																
Actividad 2																
Actividad 3																
Actividad 4																
Actividad 5																
Actividad 6																
Actividad 7																
Actividad 8																

11. PRESUPUESTO (En caso de modalidad Investigación) **Y FUENTES DE FINANCIACIÓN** (En caso de modalidad Profundización)

Cuantificar en dinero los recursos con los que se cuenta y que son requeridos para llevar a feliz término la propuesta. Hacer el presupuesto tiene el propósito académico de que el estudiante realice un ejercicio similar al que debería hacer para calcular el valor de un proyecto que realizaría para un particular.

Nota: El grupo se hace responsable de los costos que pueda implicar el desarrollo del proyecto.

Manejar una estructura de ejemplo del presupuesto..

12. RESULTADOS ESPERADOS

Se espera obtener un sistema integrado que cumpla con los siguientes resultados:

- Una base de datos consolidada y actualizada con información de inmuebles extraída de diversas plataformas.
- Un modelo de machine learning que permita predecir los precios de casas y apartamentos con una precisión aceptable.
- Una aplicación web o chatbot que ofrezca una experiencia interactiva para consultar las predicciones y otros detalles relevantes.
- Documentación técnica y un informe final que resuma las metodologías empleadas, los resultados obtenidos y las oportunidades de mejora para futuros trabajos.

13. BIBLIOGRAFÍA

- Antipov, D., & Pokryshevskaya, E. (2012). Prediction of housing prices using statistical learning. Journal of Real Estate Finance and Economics, 45(3), 327–348.
- Caraffa, G. (2022). Real-time price tracking in e-commerce through web scraping. International Journal of Big Data & Analytics, 7(4), 56–68.
- Chica, M., & Botero, A. (2023). Valoración inmobiliaria en zonas de expansión urbana en Medellín. Revista de Ingeniería y Sociedad, 12(1), 45–58.
- Kho, J. (2018). How to Web Scrape with Python in 4 Minutes. Towards Data Science. Recuperado de https://towardsdatascience.com/how-to-web-scrape-with-python-in-4-minutes-bc49186a8460
- Nunes Ariza, P., & Manrique Piramanrique, J. (2024). Uso de web scraping y aprendizaje automático para la estimación de precios de vivienda en Colombia. Tesis de Maestría, Universidad X.
- Park, J., & Bae, S.-H. (2015). Predicting house prices using machine learning algorithms. Expert Systems with Applications, 42(18), 7020–7028.
- Peterson, D. F., & Flanagan, D. (2009). Housing price prediction using neural networks. Applied Artificial Intelligence, 23(6), 485–500.
- Picardo, O. (2019). Integrating web scraping with machine learning for real estate data analytics. Proceedings of the 12th Latin American Data Science Conference, 48–56.
- Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. Packt Publishing.
- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55.
- Sharma, A., et al. (2024). Comparative analysis of machine learning models for housing price prediction. Journal of Real Estate Research, 46(2), 155–174.

14. FIRMAS

FIRMA DEL ESTUDIANTE:	**FIRMA DEL DOCENTE DIRECTOR O TUTOR**

15. DATOS DE TRÁMITE COMITÉ DE INVESTIGACIÓN DEL PROGRAMA (Espacio para diligenciar por el Comité del Programa)

No. CONSECUTIVO	
No. ACTA	
FECHA	