ICADE

ANALISIS DE FACTORES QUE AFECTAN AL PRECIO DE LA VIVIENDA MEDIANTE EL USO DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL

Autor: Mª Fernanda Rius Matas

Director: Raúl González Fabre

MADRID | Abril 2024

RESUMEN

El análisis de los factores que inciden en el precio de la vivienda mediante técnicas de inteligencia artificial emerge como un enfoque crucial en el contexto del mercado inmobiliario. La complejidad inherente a este mercado, con sus múltiples variables interrelacionadas, demanda herramientas avanzadas para una comprensión profunda y una toma de decisiones más precisa. Esta investigación se centra en la aplicación de algoritmos de aprendizaje automático y análisis de datos para identificar patrones, tendencias y relaciones ocultas en conjuntos de datos inmobiliarios. Los objetivos incluyen no solo comprender los determinantes del precio de la vivienda, sino también optimizar la toma de decisiones en el mercado inmobiliario. La metodología abarca desde el análisis exploratorio de datos y la revisión de literatura hasta la implementación de modelos predictivos, con el fin de proporcionar herramientas eficientes y precisas para la toma de decisiones estratégicas. Los resultados obtenidos de este estudio no solo informarán sobre los factores más influyentes en los precios de la vivienda, sino que también ofrecerán recomendaciones estratégicas para los actores clave del mercado inmobiliario, contribuyendo así a una comprensión más completa y una gestión más eficaz de este sector fundamental en la economía.

Palabras clave: mercado inmobiliario, precio de la vivienda, machine learning, oferta y demanda, mercado de alquiler, compraventa, evolución del precio.

ABSTRACT

The analysis of factors affecting housing prices through artificial intelligence techniques emerges as a crucial approach in the real estate market context. The inherent complexity of this market, with its multiple interrelated variables, demands advanced tools for deep understanding and more precise decision-making. This research focuses on applying machine learning algorithms and data analysis to identify patterns, trends, and hidden relationships in real estate datasets. Objectives include not only understanding the determinants of housing prices but also optimizing decision-making in the real estate market. The methodology spans from exploratory data analysis and literature review to the implementation of predictive models, aiming to provide efficient and accurate tools for strategic decision-making. The findings from this study will not only inform about the most influential factors in housing prices but also offer strategic recommendations for key players in the real estate market, thus contributing to a comprehensive understanding and more effective management of this fundamental sector in the economy.

Keywords: real estate market, housing price, machine learning, supply and demand, rental market, buying and selling, price evolution

ÍNDICE

INDICE DE FIGURAS

INDICE DE TABLAS

1. Introducción

El mercado inmobiliario, como indican diversos estudios (Leamer, 2007; Mooya, 2016), desempeña un papel crucial en la actividad económica y pieza fundamental del propio ciclo económico. Los precios de la vivienda han sido protagonistas en crisis económicas y financieras pasadas, y su correcta valoración es esencial tanto para compradores y vendedores como para instituciones financieras. Por ello, comprender los factores que influyen en el precio de la vivienda es esencial para múltiples actores en el mercado inmobiliario, desde compradores y vendedores hasta desarrolladores y analistas. Sin embargo, la complejidad y la interdependencia de estos factores plantean desafíos significativos para su análisis y predicción precisos. En este contexto, el uso de técnicas de inteligencia artificial (IA) emerge como una herramienta poderosa para abordar esta complejidad y alcanzar una mayor comprensión sobre el mercado inmobiliario. A través de la aplicación de algoritmos de aprendizaje automático y análisis de datos, se busca identificar patrones, tendencias y relaciones ocultas en conjuntos de datos inmobiliarios, con el fin de proporcionar una comprensión más completa y precisa de los determinantes del precio de la vivienda.

La inteligencia artificial influye en el sector inmobiliario español al automatizar tareas repetitivas, como la búsqueda de propiedades y la atención al cliente a través de chatbots y asistentes virtuales, agiliza el proceso de compraventa de inmuebles. Asimismo, los algoritmos de aprendizaje automático permiten valoraciones más precisas y rápidas de las propiedades, mientras que las herramientas de realidad virtual ofrecen a los compradores una experiencia de visualización mejorada. Sin embargo, todos estos avances conllevan preocupaciones sobre la privacidad de los datos y el impacto en el empleo debido a la automatización (Rivera Carmona, 2023).

En este capítulo se abordarán las razones fundamentales que respaldan la elección del tema de investigación, así como la delineación de los objetivos principales y secundarios del trabajo. Además, se expondrá la metodología que se seguirá para llevar a cabo esta investigación. Este análisis inicial proporcionará una visión clara de la motivación detrás del enfoque elegido, estableciendo el marco conceptual que guiará el desarrollo y la ejecución de este trabajo.

a. Objetivos

El objetivo principal de este trabajo de fin de grado es el estudio de los factores que afectan al precio de la vivienda mediante la aplicación de técnicas de Inteligencia Artificial, para optimizar la toma de decisiones en el mercado inmobiliario.

Para alcanzar este propósito, se plantean objetivos secundarios, que incluyen el análisis de los conceptos clave del mercado inmobiliario y la identificación de eventos pasados que han tenido un impacto en los precios de la vivienda. Además, se llevará a cabo una exploración detallada de datos e información del sector para comprender tendencias y patrones, así como la identificación de las características más influyentes en los precios de las propiedades inmobiliarias. La investigación también implica la implementación de modelos de aprendizaje automático con el fin de predecir el precio de los activos inmobiliarios, seguido de la validación y evaluación de dichos modelos mediante métricas correspondientes. Finalmente, se explorará el impacto potencial de estos hallazgos en las decisiones estratégicas y financieras de las empresas del sector inmobiliario.

Este enfoque integral busca proporcionar herramientas más eficientes y precisas para la toma de decisiones en el mercado inmobiliario al tener una visión más clara sobre el mercado inmobiliario en su conjunto y las variables con más influencia en el precio de la vivienda.

b. Metodología

Con el propósito de alcanzar los objetivos establecidos, se seguirá una metodología estructurada que empezará con el análisis exploratorio de los datos tras la revisión de la literatura, con el objetivo de comprender el contexto del mercado inmobiliario.

A nivel teórico el enfoque del trabajo se centrará en la exploración de fuentes de información tanto primarias como secundarias para respaldar los mensajes clave de la investigación. Se utilizarán herramientas como la biblioteca en línea de la Universidad Pontificia Comillas, Google Académico y diversos recursos como artículos web, informes corporativos y blogs.

Se procederá a la recopilación de los datos utilizados para el análisis, centrándose en

variables clave como país, ubicación, año de construcción, número total de plantas, piso del apartamento, número de habitaciones y baños, área total del apartamento y precio en dólares estadounidenses. Se llevará a cabo también el procesamiento de los datos para garantizar su calidad y coherencia.

Una vez recopilados y preparados los datos, se realizará un análisis exploratorio detallado para identificar patrones, tendencias y posibles relaciones entre variables. Esto proporcionará una base comprensiva antes de la aplicación de técnicas de inteligencia artificial y técnicas de visualización de datos para llevar a cabo un análisis predictivo del precio de la vivienda y sacar conclusiones sobre que variables son más determinantes en el precio.

c. Desarrollo

El desarrollo integral del trabajo sobre el estudio de los factores que afectan al precio de la vivienda, y, por tanto, al mercado inmobiliario se desglosa en cinco fases clave: metodología, marco teórico, análisis de datos y resultados, recomendaciones y conclusiones.

En el capítulo dos del trabajo, el marco teórico, está enfocado en proporcionar una revisión exhaustiva de la literatura existente sobre el mercado inmobiliario, además de definir de manera clara los conceptos clave que sustentan el estudio sobre el precio de la vivienda y los factores o variables que más influyen en esto. La metodología de análisis ofrece una descripción detallada del análisis específico empleado en el estudio. Esto incluye la explicación de las variables analizadas, las técnicas estadísticas utilizadas y los procedimientos aplicados.

Después, a lo largo del tercer capítulo, se desarrolla el análisis de datos y describirán y presentarán los hallazgos derivados de él análisis de datos. El procedimiento detallado del análisis se explica al principio del capítulo tres.

Por último, en los capítulos cinco y seis, de recomendaciones y conclusiones respectivamente, se sintetizan los resultados del estudio con el objetivo de obtener una visión general y estratégica de la situación. Buscando resumir los resultados, proporcionar soluciones estratégicas y prácticas adecuadas para los actores claves del

mercado inmobiliario. Poe último, las conclusiones presentan una visión global de la investigación, resaltando los puntos clave y ofreciendo una reflexión sobre su importancia en el contexto general del mercado inmobiliario.

2. Marco teórico del mercado inmobiliario

a. Factores que inciden sobre el precio y el volumen en el mercado inmobiliario

Según el informe del Observatorio Inmobiliario (BBVA Research, 2023) en los últimos años el mercado de la vivienda en Europa ha experimentado cambios significativos entre los que podemos destacar el aumento de precios. Este aumento de precio ha sido notable mayormente en zonas urbanas y turísticas, impulsado por la demanda, la escasez de oferta y factores como la inversión extranjera. La construcción de nuevas viviendas no ha podido mantenerse al ritmo de la demanda, lo que ha contribuido al encarecimiento y a una mayor competencia entre las propiedades disponibles, algo que se ha visto directamente reflejado en el aumento de precio tanto para compras como arrendamientos.

La inversión extranjera también ha desempeñado un papel importante en el mercado inmobiliario a nivel europeo. Inversores extranjeros, especialmente de fuera de la Unión Europea, han mostrado interés en adquirir propiedades en Europa, tanto para fines turísticos o de inversión, como para residir en ellos.

La pandemia de COVID-19 también ha tenido efectos en el mercado de la vivienda. A grandes rasgos podemos decir que inicialmente debido a la incertidumbre de la pandemia se produjo una disminución de transacciones, sin embargo, postpandemia la demanda de viviendas más espaciosas y con áreas exteriores aumentó, impulsando algunos segmentos de este mercado. (BBVA Research, 2023)

Una de las primeras cosas que debemos analizar al abordar este tema es el escenario internacional complejo del mercado de la vivienda. Las proyecciones económicas del Fondo Monetario Internacional (FMI) para principios de 2022 señalaban una debilitación debido a factores como la pandemia, el aumento de precio del petróleo y

gas, así como problemas persistentes en la oferta. Esto ha desencadenado un aumento generalizado de la inflación, algo de lo que todos somos conscientes hoy en día (Fondo Monetario Internacional, 2022).

Es interesante mencionar cómo la subida de los tipos de interés puede incidir en el mercado de la vivienda. Habitualmente, los aumentos en los tipos de interés llevan consigo un encarecimiento del crédito hipotecario. Esto puede impactar en la demanda de viviendas, provocando una desaceleración del mercado al hacer las hipotecas menos accesibles para algunos compradores. Por otro lado, para aquellos que ya tienen hipotecas, de tasa variable en concreto, un alza de los tipos puede significar cuotas mensuales más altas, lo que puede afectar a la estabilidad financiera.

Cuando analizamos el panorama macroeconómico del sector inmobiliario, identificamos ciertas variables fundamentales que ejercen influencia en su dinámica. Entre estas variables se encuentran el crecimiento económico, la tasa de desempleo, el Producto Interno Bruto (PIB), la inflación y los tipos de interés. Estos elementos son clave para comprender y prever movimientos en el mercado inmobiliario. Además, al evaluar los precios de las propiedades, surge un concepto esencial: el cálculo del precio del suelo. Este se refiere al valor proyectado a partir de las futuras rentas que se esperan en ese lugar específico. Es decir, es una evaluación anticipada basada en las potenciales ganancias derivadas del área en cuestión.

Teniendo todo esto en cuenta, es comprensible que las condiciones de acceso a la vivienda se estén convirtiendo en una de las cuestiones más debatidas tanto en los gobiernos nacionales como en Europa. Tras las crisis financieras, los precios de la vivienda suelen bajar, lo que facilita el acceso a la vivienda. Sin embargo, en las principales ciudades europeas los precios se han recuperado muy rápido y las posibilidades de adquisición o alquiler no han mejorado. Esta situación tiene un impacto negativo en la economía y la sociedad. Este problema de altos precios en el mercado inmobiliario en las ciudades de la Unión Europea es recurrente, los precios están impulsados por la especulación. Además, el futuro del mercado inmobiliario europeo apunta hacia una transformación en la forma en que se perciben y utilizan los activos inmobiliarios. Según los expertos del sector, estamos viendo una transición hacia una gama más diversa de activos de inversión, que van más allá de las oficinas tradicionales y el comercio minorista. Este cambio se ve impulsado por avances tecnológicos y

cambios en los hábitos de consumo, lo que lleva a una difuminación de las fronteras entre los diferentes sectores. Como resultado, los inversores están cada vez más abiertos a utilizar los activos inmobiliarios para diversos fines, desde comercio minorista hasta espacios de trabajo flexibles (PricewaterhouseCoopers, 2023).

Los ciclos demográficos representan otro factor primordial en la configuración del mercado de la vivienda en Europa. Las variaciones en la estructura poblacional, que abarcan desde el envejecimiento demográfico hasta los flujos migratorios y las tasas de natalidad, ejercen una influencia trascendental en la dinámica de la demanda y oferta de viviendas. Estos cambios inciden directamente en la geografía y características de la demanda residencial, generando una necesidad adaptativa en términos de tipos de viviendas, ubicación y servicios asociados. Las migraciones internas y externas, junto con los patrones de densidad poblacional, son determinantes en la reconfiguración de los mercados inmobiliarios, impulsando ajustes en la planificación urbana para adecuarse a estos movimientos demográficos. Este panorama demográfico desempeña un papel crucial al identificar cómo las tendencias poblacionales moldean la demanda residencial y cómo las políticas urbanísticas deben ajustarse para abordar estas transformaciones. (Garcia Vega, 2023)

b. Escenario macroeconómico e impacto

En el transcurso del año 2022, la economía mundial experimentó una marcada desaceleración, representada por una reducción sustancial en la tasa de crecimiento del Producto Interno Bruto (PIB) a nivel global. Según las estimaciones proporcionadas por el Fondo Monetario Internacional (FMI), el crecimiento económico mundial pasó de un sólido 6% en 2021 a un 3,2% proyectado para 2022. Esta disminución en la tasa de crecimiento del PIB tiene implicaciones de gran alcance, particularmente para las economías avanzadas y emergentes que muestran patrones de crecimiento esperados del 1,8% y 3,7%, respectivamente, durante el mismo período. La desaceleración del PIB global ha impactado significativamente a las economías avanzadas, reflejando un cambio marcado en su trayectoria de crecimiento. Este fenómeno se atribuye a una serie de factores, incluida la persistencia de la pandemia de COVID-19, la incertidumbre geopolítica derivada de conflictos como la invasión rusa en Ucrania, y la complejidad de restaurar las políticas económicas a un estado de normalidad tras las perturbaciones

generadas por la pandemia en años previos. (Rodríguez-López, 2022)

A pesar de la desaceleración global, las economías emergentes muestran una mayor resistencia con un crecimiento proyectado del 3,7%. Esta tasa, aunque reducida en comparación con años anteriores, señala un crecimiento relativamente sólido en el contexto actual. Sin embargo, estas economías enfrentan desafíos propios, incluida la volatilidad derivada de los mercados globales, la incertidumbre en las políticas económicas de las principales economías y la presión sobre los precios de los productos básicos, lo que podría afectar su estabilidad económica a corto y mediano plazo. En conjunto, la desaceleración del PIB global representa un desafío significativo para las economías a nivel mundial, requiriendo una gestión cuidadosa de políticas económicas, estrategias de estímulo y medidas regulatorias para mitigar su impacto y fomentar un futuro crecimiento sostenible. (Rodríguez-López, 2022)

En el artículo de “Recent Trends in Real Estate Research” de Breuer y Steininger (2020), señalan que, en la mayoría de los países el sector inmobiliario representa una parte significativa de la economía, medida por su volumen, participación en el PIB y fuerza laboral. En los ejemplos concretos de Estados Unidos y Alemania, el valor total del mercado inmobiliario en 2018 ascendió a $46,4 y $8,3 mil millones respectivamente, lo que demuestra su importancia económica sustancial. También se resalta en el artículo la interconexión del sector inmobiliario con los mercados financieros, especialmente a través de instrumentos como las hipotecas y los valores respaldados por activos. La crisis financiera de 2007-2008, que comenzó en el sector inmobiliario de Estados Unidos y posteriormente se extendió a los mercados financieros globales, ejemplifica esta influencia del sector inmobiliario en la economía en su conjunto. En el contexto especifico del mercado inmobiliario alemán, Breurer y Steininger (2020), también destacan que, como consecuencia del envejecimiento de la población, la urbanización y la tendencia a hogares unipersonales, ha aumentado el número de hogares y que todos estos aspectos han provocado un auge en la demanda de viviendas.

Las perspectivas sobre el PIB para los años siguientes muestran un panorama algo más alentador. En particular, las revisiones al alza del crecimiento del PIB en España para 2024, que pasó del 1,5 % al 2,1 %, y la proyección para 2025, a pesar de una leve

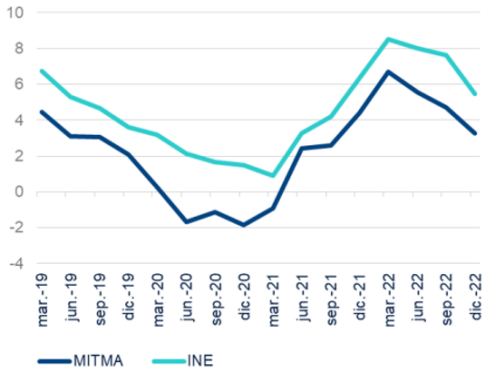
reducción, del 2,0 %, sugieren una recuperación gradual de la actividad económica. Estos datos pueden tener implicaciones importantes para diversos sectores, incluido el mercado de la vivienda, donde un aumento del PIB suele asociarse con un mayor impulso en la demanda de viviendas. Políticas para impulsar la demanda de vivienda o la fijación de precios podrían tener efectos a largo plazo en el sector. Las políticas de impulso a la demanda de vivienda o de fijación de precios podrían tener efectos nocivos a largo plazo en el sector. El Gobierno ha anunciado avales para los jóvenes, a través del Instituto de Crédito Oficial (ICO), con el objetivo de facilitar el acceso a una vivienda, lo que podría influir en la dinámica del mercado inmobiliario en el futuro cercano (BBVA Research, 2024).

c. Evolución de los precios de la vivienda en la Unión Europea y España

Como ya notaron hace unos años Carbó Valverde & Rodríguez Fernández (2018), "la evolución de los precios de la vivienda en la Unión Europea muestra una heterogeneidad considerable". Esta evolución no sigue un patrón regular, tiene aumentos y disminuciones que no son únicamente atribuibles a factores estacionales, sino que también reflejan la naturaleza fluctuante de un mercado cuya dirección a mediano plazo aún no está claramente establecida.

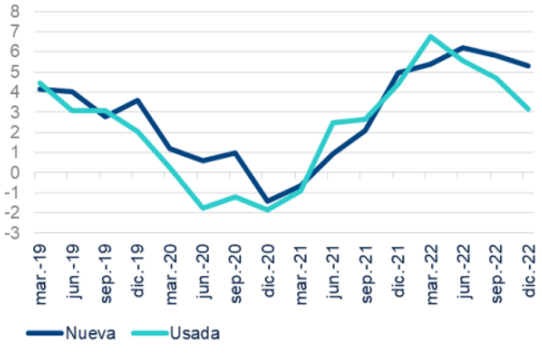
Podemos apreciar en el gráfico proporcionado por BBVA Research, basado en los datos del INE y MITMA, la representación visual de la variación en los precios de la vivienda desde marzo de 2019 hasta diciembre de 2022 en España.

Figura 1. Precio de la vivienda en España 2019-2022



Fuente: BBVA Research, INE y MITMA

Figura 2. Variación interanual precio de la vivienda en España



Fuente: BBVA Research y MITMA

Podemos observar en estos gráficos elaborados por BBVA Research que el aumento del

precio de la vivienda ha sido tanto en vivienda nueva como usada, existiendo una correlación bastante alta entre ambas. Durante el año 2022 los precios de la vivienda en España comenzaron a bajar, para vivienda nueva y usada.

En cuanto a la demanda de viviendas, el panorama del mercado inmobiliario español en 2022 evidencia una desaceleración significativa en las ventas de viviendas desde mediados de año, contrastando con el crecimiento inicial del 6,2% respecto al año anterior. Este descenso se atribuye principalmente al aumento de los tipos de interés y la consecuente moderación del crecimiento económico. Las ventas de propiedades se han situado por debajo de los niveles prepandémicos desde octubre, afectando mayormente a todas las comunidades autónomas, siendo la venta de viviendas de segunda mano y extranjeras las más afectadas. A pesar del incremento del 30.5% en las operaciones realizadas por extranjeros, el crecimiento de las transacciones se ha desacelerado considerablemente. Asimismo, se observa un cambio en los patrones hipotecarios, con una reducción en los préstamos a tipo fijo y un aumento en las amortizaciones como respuesta al aumento de los tipos de interés. (BBVA Research, 2023)

La construcción de nuevas viviendas, así como las ventas de viviendas, se ha visto significativamente afectada. La incertidumbre generada por el Anteproyecto de Ley de Vivienda, sumada al encarecimiento de las materias primas y la crisis derivada del conflicto en Ucrania, contribuyeron al retraso y la ralentización de múltiples proyectos en el sector de la construcción. Si bien las ventas de viviendas lograron superar los niveles de 2019 en la mayoría de las regiones, la situación de los visados para nuevas viviendas no fue igualmente prometedora.

En el contexto del mercado de alquiler en España, se evidencia una preocupante disparidad entre el aumento de los costos de alquiler y el crecimiento de los salarios desde 2015. Esta discrepancia ha generado una situación de vulnerabilidad económica para numerosos hogares, como revela un estudio reciente. Especialmente alarmante es la situación de los jóvenes, quienes enfrentan desafíos adicionales debido a la precariedad laboral y los salarios bajos, lo que ha llevado a que una proporción significativa de sus ingresos se destine al alquiler y gastos asociados. Esta realidad pone de relieve la necesidad de abordar de manera urgente las disparidades en el mercado de vivienda y las consecuencias socioeconómicas que acarrean para la población española (Rodríguez, 2019)

Un aspecto preocupante en cuanto al mercado español es la discrepancia entre la creciente demanda y la oferta de vivienda nueva. A pesar de un aumento notable en la demanda, la construcción no logró satisfacer la necesidad de nuevos hogares. Esta brecha entre la oferta y la demanda ha generado una situación compleja y desafiante para el mercado inmobiliario español. Además, la baja iniciación de viviendas durante los últimos trimestres, en relación con los visados y la formación de nuevos hogares, revela un desajuste en la recuperación del sector. Las estadísticas reflejan que la cantidad de visados de obra nueva se ha mantenido relativamente baja en comparación con la creación de nuevos hogares, señalando una discrepancia en la recuperación del mercado. Los desafíos adicionales, como la escasez de mano de obra cualificada, problemas de inseguridad jurídica, y la lentitud en las tramitaciones de licencias, se suman al panorama. El encarecimiento y la escasez de ciertos materiales, junto con el incremento de los salarios en el sector, sugieren una posible limitación del crecimiento futuro en la construcción de viviendas (Taltavull, P., 2020).

d. Proyecciones y desafíos en el mercado inmobiliario

A partir del análisis mediante un modelo de regresión múltiple enfocado en la evolución de la prime yield en el sector inmobiliario, la empresa Colliers ha realizado proyecciones relevantes. Estas proyecciones indican una fase crítica prevista para el segundo semestre del año 2022, durante la cual se espera que los precios de los activos alcancen su punto más bajo. Específicamente, se prevé que la prime yield del mercado de oficinas se acerque al 4,9% en este período. Este periodo crítico señala el fin de la peor etapa del mercado, con la anticipación de que los tipos de intervención alcancen su punto máximo y potencialmente generen oportunidades de inversión a medio plazo. Para los años 2024 y 2025, prevén una estabilidad relativa en la prime yield, manteniéndose entre el 4,6% y el 4,8%, seguido de una proyección de compresión adicional en 2026, alcanzando un 4,3% (descendiendo 60 puntos básicos desde sus máximos). Lo que implica una continuidad en las condiciones del mercado. Esta estabilidad del prime yield sugiere una tendencia a mantener los precios de los activos dentro de ciertos límites, lo que podría reflejar un equilibrio en la oferta y la demanda en ese período. Las proyecciones del modelo auguran oportunidades de inversión atractivas para aquellos que realicen adquisiciones durante los últimos trimestres de

2023 y el primer semestre de 2024. Se espera que estas adquisiciones ofrezcan rentabilidades considerables debido a la compresión de yields a mediano plazo y al incremento de las rentas durante este período. Estos análisis revelan un panorama cambiante en el mercado inmobiliario español, presentando oportunidades y desafíos cruciales para los inversores y actores del mercado en la toma de decisiones estratégicas a futuro (Colliers, 2023).

En cuanto al sector residencial, de acuerdo con informes de Colliers, durante 2023 ha experimentado una leve disminución en el volumen de transacciones debido al aumento de los tipos de interés y la inflación, lo que ha afectado principalmente a la demanda de viviendas. A pesar de esto, la producción de viviendas nuevas se mantiene estable, especialmente en áreas con alta demanda y escasez de suelo. El mercado de alquiler ha ganado protagonismo, aunque la inversión institucional se ha estancado debido a mayores rendimientos y costos financieros. Por otro lado, el mercado de lujo sigue siendo sólido, con una demanda sostenida por parte de compradores extranjeros. Se espera que, para el próximo año, los desafíos principales incluyan la reducción del coste financiero, la estabilización de los costos de construcción y la escasez de mano de obra cualificada. (Colliers, 2023).

Puede ser esencial ahondar en la importancia de la diversificación de los productos hipotecarios en el mercado actual, específicamente en relación con la accesibilidad para diferentes segmentos demográficos. La innovación en productos financieros, como hipotecas a tasas fijas o variables con condiciones más flexibles, podría representar una solución para mitigar los efectos adversos de los aumentos de los tipos de interés en la demanda de viviendas.

Además, sería relevante considerar estudios que aborden el impacto específico de las restricciones regulatorias en la oferta de viviendas, y cómo estas podrían ser flexibilizadas para fomentar la construcción de nuevos hogares y atenuar la discrepancia entre oferta y demanda en el mercado.

También se puede explorar más a fondo cómo las tendencias tecnológicas están influenciando la comercialización y transacciones inmobiliarias en la era digital. El análisis detallado de la influencia de las plataformas online y las estrategias digitales en la dinámica del mercado podría ofrecer una visión más completa de cómo estas herramientas están afectando las decisiones de compra y alquiler.

Asimismo, un enfoque en cómo los programas gubernamentales o iniciativas locales están abordando la accesibilidad a la vivienda para jóvenes y otros grupos marginados sería crucial. Examinar el impacto de tales políticas en la dinámica del mercado inmobiliario podría proporcionar una perspectiva valiosa sobre las soluciones a largo plazo para la situación actual de precios elevados y la falta de acceso a la vivienda en las principales ciudades europeas. Estas adiciones se centran en áreas específicas que podrían complementar la información existente y ofrecer una perspectiva más amplia y detallada sobre el mercado inmobiliario.

La creciente preocupación por la sostenibilidad y el impacto ambiental ha generado un cambio significativo en el mercado inmobiliario, donde la demanda de edificaciones sostenibles y energéticamente eficientes está ganando terreno (Arrieta, Grajales, & Padilla, 2023). La adopción de prácticas de responsabilidad social empresarial (RSE) en el ámbito inmobiliario ha transformado no solo la manera en que se construyen edificaciones, sino también la relación de las empresas con las comunidades locales y el entorno ambiental. Una de las empresas comentada en este informe, a través de su iniciativa 'Edificio Verde', ha logrado reducir un 30% las emisiones de carbono en sus proyectos de construcción mediante el uso de materiales sostenibles y tecnologías de eficiencia energética. Este enfoque no solo ha mejorado la huella ambiental, sino que también ha generado un ahorro del 20% en costos operativos a largo plazo. Además, iniciativas como 'Comunidades Sostenibles' de la empresa ABC han involucrado a residentes locales en programas de educación ambiental y han destinado el 5% de sus ganancias anuales a proyectos comunitarios, mejorando así la calidad de vida de las áreas circundantes.

Estas prácticas innovadoras son solo un ejemplo de cómo las empresas inmobiliarias están liderando el camino hacia un desarrollo urbano más sostenible y equitativo. Además, las tendencias emergentes en el sector indican una mayor demanda de edificaciones con certificaciones de sostenibilidad, como LEED o BREEAM, lo que refleja la creciente preferencia del mercado por propiedades que promueven un estilo de vida más sostenible. A medida que la RSE continúa evolucionando en el sector inmobiliario, se prevé una mayor colaboración con gobiernos locales para implementar regulaciones más estrictas y fomentar prácticas más responsables. Estas acciones no

solo impulsan la imagen corporativa de las empresas, sino que también influyen significativamente en la calidad de vida de las comunidades y en la preservación del entorno natural (Arrieta, Grajales, & Padilla, 2023).

La industrialización está emergiendo como un modelo prometedor en el sector inmobiliario, liderando el futuro de la vivienda. Con el hormigón como material predominante en estas construcciones, aproximadamente el 30% de los edificios nuevos ya incorpora algún componente industrializado. Este enfoque, respaldado por avances tecnológicos, tiene el potencial de representar entre el 30% y el 40% de las nuevas viviendas para el año 2030, según estimaciones optimistas. La industrialización de viviendas implica la aplicación de tecnología en la fabricación de elementos en serie, lo que resulta en estructuras completas que se ensamblan fuera de fábrica. Este enfoque, que puede ser modular o basado en componentes, ofrece una serie de ventajas, como la sostenibilidad, la reducción de los plazos de construcción y la mejora en la calidad de las viviendas. Los elementos estructurales, las fachadas, las cubiertas, las ventanas, los baños y los paneles de cocina son algunos de los componentes más frecuentemente industrializados en las viviendas (Torío, 2021)

Las ventajas de este modelo de vivienda industrializada incluyen su enfoque en la sostenibilidad, siendo más respetuoso con el medio ambiente que el método de construcción tradicional, además de reducir los plazos de construcción, entre otros beneficios. Sin embargo, sigue habiendo limitaciones para los promotores debido a los costos iniciales, además de la evidente necesidad de continuar la inversión en investigación sobre este tema, en España, el porcentaje de casas industrializadas es inferior al 2%, muy por debajo del resto de Europa (El Economista, 2023).

En este contexto de la vivienda industrializada, empresas como Avintia, AEDAS Homes y ACR están liderando el proceso en España, estableciendo objetivos ambiciosos para la integración de métodos modernos de construcción en sus proyectos. Alianzas estratégicas y el desarrollo de nuevas tecnologías impulsan aún más esta tendencia, con la esperanza de que el 35% de las obras sean industrializadas para 2025 (Torío, 2021). Aunque las viviendas unifamiliares industrializadas son una realidad, el próximo desafío es expandir este enfoque a la construcción en altura.

En conclusión, las proyecciones y desafíos en el mercado inmobiliario revelan un panorama dinámico y en evolución constante, con estimaciones de estabilidad relativa para los próximos años, emergiendo la industrialización y la sostenibilidad como aspectos clave para el futuro del sector inmobiliario, buscando promover una construcción más eficiente y responsable.

3. Análisis

a. Datos utilizados

En este trabajo, se empleará un conjunto de datos compuesto por información detallada sobre propiedades inmobiliarias. Los datos incluyen variables clave como país, ubicación exacta, año de construcción, número total de pisos en el edificio, piso del apartamento, cantidad de habitaciones y baños, área total y habitable de la vivienda, así como el precio en dólares estadounidenses. Se trata de un archivo CSV que después del correcto tratamiento de valores faltantes y coherencia de datos será utilizado para el análisis de factores que afectan al precio de la vivienda.

Esta parte analítica del trabajo sigue el siguiente proceso:

1. Selección de la base de datos para el trabajo, a partir de las disponibles en Kaggle, una plataforma colaborativa de bases de datos, que permite tener una muestra de datos del mercado inmobiliario para la posterior aplicación de técnicas de machine learning.

2. Preprocesamiento de datos, eliminando registros con valores faltantes en los casos en los que no sea posible trabajar con ellos debido a la poca información que presentan o por ser datos claramente erróneos al constar en el archivo con años posteriores al actual. Empezando con la eliminación de las observaciones para las que no hay datos de ‘price_in_USD’, ya que es la variable objetivo de este trabajo.

3. Análisis exploratorio de los datos, con el fin de conocer las variables que se utilizaran en los modelos, identificar conexiones y conocer un poco más sobre los datos disponibles para el estudio.

4. Estandarización y selección de las variables más representativas, o con mayor relevancia para el trabajo.
5. Partición de los datos en training set y test set, dividiendo los datos en conjunto de entrenamiento del 80% y conjunto de prueba de 20%.
6. Entrenamiento de los modelos de aprendizaje supervisado, y cálculo de las métricas necesarias para la evaluación de los modelos.

El conjunto de datos seleccionado ofrece un amplio abanico de variables que proporcionan una visión detallada del mercado inmobiliario. Las diversas características incluidas en este conjunto permiten un análisis de propiedades de alta gama, abarcando aspectos clave como la ubicación geográfica, detalles específicos del inmueble y sus características físicas.

- País: ubicación donde se encuentra la propiedad.
- Ubicación: dirección especifica dentro del país
- Año de construcción del edificio.
- Número total de pisos del edifico: número total de pisos o niveles del edificio
- Piso del apartamento: piso o planta en la que se encuentra el apartamento dentro del edificio.
- Habitaciones del apartamento: número total de habitaciones en el apartamento.
- Baños del apartamento: número de baños en el apartamento
- Área total de la vivienda: superficie total en metros cuadrados.
- Área habitable de la vivienda: superficie habitable del total de metros cuadrados.
- Precio en USD: precio de la propiedad listado en dólares estadounidenses (USD). Se trata de dólares nominales por lo que no se ha llevado a cabo ningún ajuste por inflación.

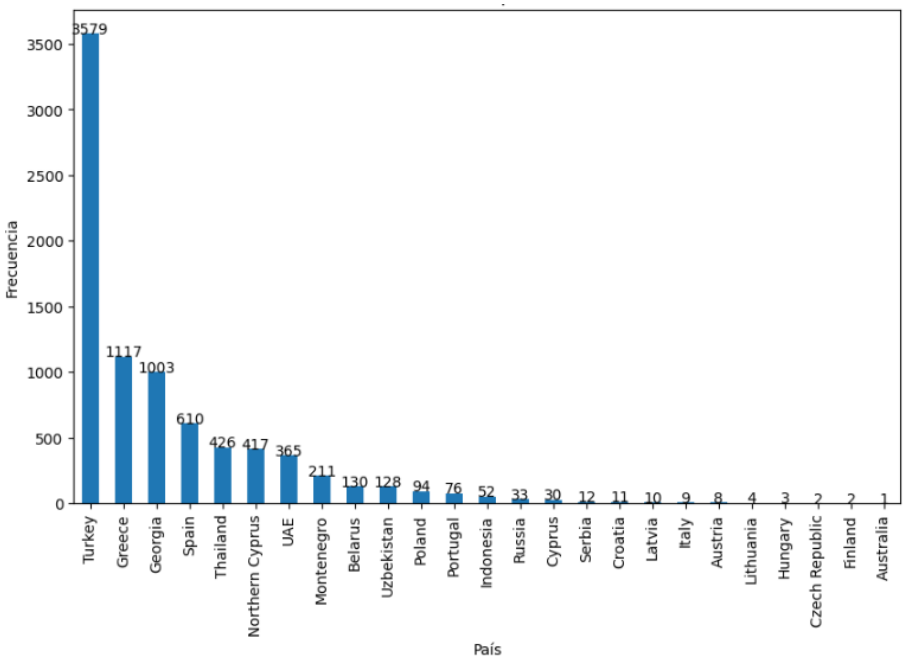
Tras analizar el listado de variables en el conjunto de datos de estudio, podemos afirmar

que los datos contienen información valiosa, desde el país y la ubicación de la propiedad hasta detalles más específicos como el año de construcción del edificio, el número total de pisos, la disposición del apartamento en el edificio, la distribución de habitaciones y baños, así como la superficie total y habitable de la vivienda. Además, la inclusión del precio en dólares estadounidenses permite una evaluación clara del valor de mercado de estas propiedades.

Esto presenta una sólida base para llevar a cabo un análisis profundo del mercado inmobiliario. La diversidad de variables proporciona la oportunidad de realizar análisis comparativos, identificar tendencias, y obtener conclusiones fundamentadas. La amplitud y detalle de esta información ofrecen un panorama sólido para la toma de decisiones informadas y estratégicas en el ámbito inmobiliario.

Dadas las variables disponibles en el conjunto de datos y viendo la distribución por países que presentaban los datos y que se puede apreciar en la figura 3, se consideró la posibilidad de enfocar el estudio en el mercado inmobiliario de Turquía, ya que las características del conjunto de datos parecen estar más alineadas con dicho mercado al ser el país con mayor presencia en los datos.

Figura 3. Frecuencia de cada país en los datos



Fuente: Elaboración propia

Sin embargo, tras probar el mismo análisis realizado en este trabajo solo para las 3.579

observaciones de Turquía no se obtuvieron diferencias significativas en los resultados de los modelos, por lo que se ha optado por trabajar con todas las observaciones sin distinción por país. Es importante tener en cuenta que, analizar todas las observaciones sin distinción y tratarlas como si pertenecieran al mismo país, lo cual se ha hecho al eliminar la variable país para entrenar los modelos, puede llevar a una interpretación sesgada de los resultados. Las diferencias económicas, sociales y políticas entre países pueden tener un impacto significativo en el mercado inmobiliario. Por lo tanto, es importante reconocer estas limitaciones y cómo la suposición de que todas las observaciones pertenecen al mismo país puede no reflejar con precisión la complejidad y la diversidad del mercado inmobiliario real.

b. Preprocesamiento y Análisis Exploratorio de los Datos

El proceso de análisis empieza con la revisión y preparación de los datos. Se ha llevado a cabo un riguroso proceso de preprocesamiento para garantizar la calidad, coherencia y consistencia de los datos. Esto implica la detección de valores atípicos y la normalización o estandarización de datos según sea necesario, así como la creación de nuevas variables relevantes para el estudio.

En cuanto a la detección de valores atípicos, se definieron como tal aquellos que mostraban valores negativos en algunas variables como número de habitaciones, o valores extremadamente grandes en otros casos. Además, se detectaron registros con años de construcción superiores al año actual (2024), los cuales también han sido considerados como valores atípicos y eliminados de la base datos para este trabajo.

En la siguiente tabla podemos ver un resumen estadístico y descriptivo de las variables que conforman el conjunto de datos:

Figura 4. Descripción de las variables presentes en los datos.

	building_construction_year	building_total_floors	apartment_floor	apartment_rooms	apartment_bedrooms	apartment_bathrooms	apartment_total_area	apartment_living_area	price_in_USD
count	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000	8.333000e+03
mean	2018.148326	10.066963	4.880187	2.822033	1.920317	1.484219	98.164887	88.350774	3.296149e+05
std	11.820675	10.870891	7.096037	1.074241	0.952993	0.655416	64.494490	58.441850	6.095362e+05
min	1894.000000	1.000000	-1.000000	1.000000	1.000000	1.000000	22.000000	1.000000	1.390000e+04
25%	2021.000000	4.000000	1.000000	2.000000	1.000000	1.000000	58.000000	53.000000	1.110570e+05
50%	2023.000000	6.000000	2.000000	3.000000	2.000000	1.000000	82.000000	75.000000	1.912680e+05
75%	2023.000000	13.000000	5.000000	3.000000	2.000000	2.000000	120.000000	109.000000	3.529340e+05
max	2024.000000	115.000000	105.000000	10.000000	10.000000	10.000000	1165.000000	1165.000000	1.815170e+07

Fuente: Elaboración propia

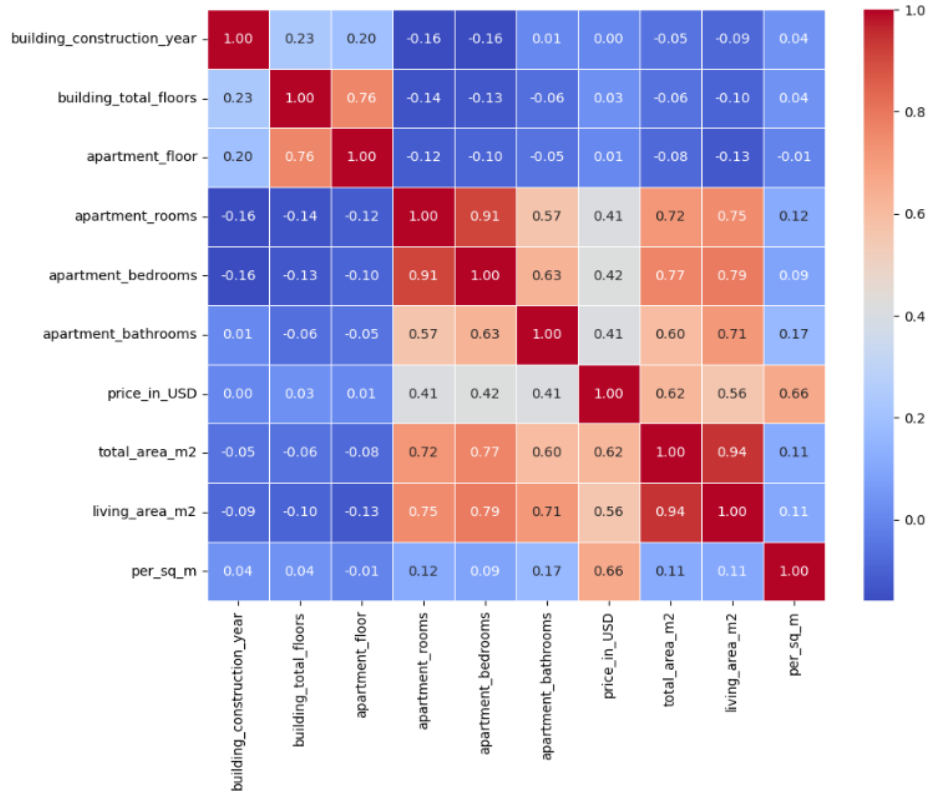
Sobre la estandarización de datos, es un proceso importante en el preprocesamiento de datos para modelos predictivos, especialmente para aquellos que se basan en métodos que asumen una distribución normal de los datos o que son sensibles a la escala de las variables. En el caso de los modelos predictivos utilizados, como la regresión lineal, los árboles de decisión y las redes neuronales, la estandarización puede ayudar a mejorar la convergencia del algoritmo y la interpretación de los coeficientes. La estandarización implica transformar las variables para que tengan una media de cero y una desviación estándar de uno. Esto se logra restando la media de cada variable y dividiendo por su desviación estándar. Esto hace que todas las variables tengan la misma escala y evita que las variables con escalas más grandes dominen el modelo.

Respecto al tratamiento de datos y teniendo en cuenta las variables disponibles en el conjunto de datos cabe destacar que las variables título y ubicación, siendo una descripción de la vivienda y la ubicación exacta del inmueble en el conjunto de datos respectivamente, no aportan información valiosa para los modelos de aprendizaje supervisado que se van a utilizar en el trabajo por lo que esta variable ha sido obviada.

Por otro lado, las variables “apartment_living_area” y “apartment_total_area” contenían las unidades (m^2), por lo que se ha eliminado para poder tratar con esta variable como números y poder incluirlas en los modelos predictivos, así como con el resto de las variables.

Después de este trabajo de preprocesamiento de datos, contamos con un total de 8.333 observaciones y la correlación entre las variables puede apreciarse en la siguiente visualización:

Figura 5. Matriz de Correlación entre las variables del modelo.



Fuente: Elaboración propia.

En general y a la vista de las correlaciones entre variables, las variables relacionadas con el tamaño del apartamento (área total, área habitable, número de dormitorios y baños) y la planta de la vivienda presentan las mayores correlaciones con el precio de la vivienda. Esto sugiere que dichos factores son los que más influyen en el precio de los apartamentos en este conjunto de datos.

Para finalizar con el análisis exploratorio de los datos, se realizó una regresión lineal utilizando el método de los mínimos cuadrados ordianrios (OLS) para investigar la relación entre las variables predictoras y la variable objetivo, que es el precio en dólares de las viviendas. A continuación, se presentan los resultados de la regresión lineal, que incluyen los coeficientes estimados, errores estándar, estadísticas t y p-values para cada variable predictora. Estos resultados proporcionan información sobre la fuerza y la dirección de la relación entre las variables independientes y la variable dependiente.

Figura 6. Factores que determinan el precio de la vivienda

OLS Regression Results

Dep. Variable:	price_in_USD	R-squared:	0.481
Model:	OLS	Adj. R-squared:	0.481
Method:	Least Squares	F-statistic:	882.8
Date:	Sat, 06 Apr 2024	Prob (F-statistic):	0.00
Time:	12:32:56	Log-Likelihood:	-96414.
No. Observations:	6666	AIC:	1.928e+05
Df Residuals:	6658	BIC:	1.929e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.525e+06	1.02e+06	-2.476	0.013	-4.52e+06	-5.26e+05
building_construction_year	1148.7410	505.469	2.273	0.023	157.860	2139.622
building_total_floors	5769.5843	680.462	8.479	0.000	4435.661	7103.508
apartment_floor	939.8863	1003.176	0.937	0.349	-1026.659	2906.432
apartment_rooms	-9.989e+04	7722.265	-12.936	0.000	-1.15e+05	-8.48e+04
apartment_bathrooms	-1.004e+04	1.17e+04	-0.856	0.392	-3.3e+04	1.3e+04
apartment_total_area	3053.2365	295.645	10.327	0.000	2473.677	3632.796
apartment_living_area	5395.8952	313.179	17.229	0.000	4781.964	6009.826

Omnibus:	7937.466	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2597595.311
Skew:	5.916	Prob(JB):	0.00
Kurtosis:	98.981	Cond. No.	3.64e+05

Fuente: Elaboración propia

Al analizar los resultados, es importante observar los valores p (p-value) asociados a cada coeficiente. Un valor p menor que 0.05 indica una significancia estadística, lo que sugiere que el coeficiente estimado es significativamente diferente de cero y que la variable correspondiente tiene un impacto significativo en el precio de la vivienda. En este caso, observamos que las variables 'building_construction_year' y 'building_total_floors' tienen valores p muy bajos, lo que indica una fuerte evidencia en contra de la hipótesis nula de que los coeficientes asociados a estas variables son cero. Por otro lado, las variables 'apartment_floor' y 'apartment_bathrooms' muestran valores p más altos, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula de que los coeficientes asociados a estas variables son cero, lo que significa que podrían no ser estadísticamente significativos para predecir el precio de la vivienda.

4. Desarrollo de Modelos de Predicción para el precio de la vivienda.

En el ámbito del análisis inmobiliario, la capacidad de anticipar el precio de la vivienda desempeña un papel fundamental, así como el conocimiento de las variables con mayor influencia en dicho precio, proporcionando información clave para las partes

interesadas, desde inversores hasta planificadores urbanos.

La implementación de modelos de predicción representa una herramienta esencial para comprender y prever las dinámicas del mercado inmobiliario. En este trabajo, se evaluará la efectividad de los modelos de regresión: el Modelo de Regresión Lineal con regulación Ridge, el Modelo Árboles de Decisión, Bosques Aleatorios, Maquinas de Vectores de Soporte y Redes Neuronales. Estas herramientas, ampliamente reconocidas por su eficacia en distintos contextos, ofrecen enfoques complementarios para abordar la tarea de predecir el precio de la vivienda así ofrecer una visión sobre el rendimiento del modelo según los diferentes algoritmos.

Las métricas que elegidas para evaluar el rendimiento de los modelos son: el coeficiente de determinación $R2$, el coeficiente de correlación de Pearson, el MAPE (Error Porcentual Absoluto Medio) y el MPE (Error Porcentual Medio). El coeficiente de determinación indica la proporción de la varianza en la variable dependiente que es explicada por la variable independiente, mientras que el coeficiente de correlación de Pearson mide la fuerza y dirección de la relación lineal entre las variables. El MAPE calcula el promedio de los errores porcentuales absolutos entre las predicciones y los valores reales, mientras que el MPE calcula el promedio de los errores porcentuales sin considerar su signo. Ambos proporcionan una medida del error porcentual entre las predicciones y los valores reales, con el MAPE ofreciendo una evaluación más intuitiva debido a su consideración de la magnitud del error.

a. Modelos de Regresión Lineal

El Modelo de Regresión Lineal asume una relación lineal entre las variables predictoras y la variable objetivo (precio de la vivienda). Este enfoque clásico proporciona una interpretación sencilla de cómo cada variable afecta al resultado. Sin embargo, su simplicidad puede limitar su capacidad para capturar relaciones no lineales complejas presentes en conjuntos de datos inmobiliarios. El primer acercamiento a la regresión lineal se debe al método de los mínimos cuadrados de Legendre en 1805.

La ecuación de regresión lineal puede expresarse de la siguiente manera:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Donde:

- *y* es la variable dependiente que se trata de predecir, el precio en este caso.
- *X1, X2, ..., Xn* son las variables independientes que se utilizan para predecir y.
- $\beta 0, \beta 1, ..., \beta n$ son los coeficientes que representan la relación entre las variables independientes y la variable dependiente.
- ϵ es el término de error, que representa la diferencia entre el valor predicho y el valor real de *y*.

El primer paso para implementar el modelo de regresión múltiple en el conjunto de datos del trabajo es comenzar por la estandarización de las características o variables, luego se utilizará la regresión lineal para aprender la relación entre estas características y los precios de los apartamentos. Posteriormente, evaluamos el rendimiento del modelo con datos de prueba, utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R2).

Por otro lado, en el enfoque de regresión Ridge, cuyo estimador fue introducido por Hoerl y Kennard en 1970, se seleccionan las variables más relevantes para el análisis. Se utiliza un proceso llamado búsqueda de hiperparámetros para encontrar la mejor configuración para nuestro modelo Ridge, lo que nos permite ajustar el modelo de manera óptima para hacer predicciones más precisas. Luego, entrenamos y evaluamos el modelo Ridge con las características seleccionadas, comparando su desempeño con el modelo de regresión lineal.

Tabla 1. Resultados Modelos de Regresión Lineal y Ridge

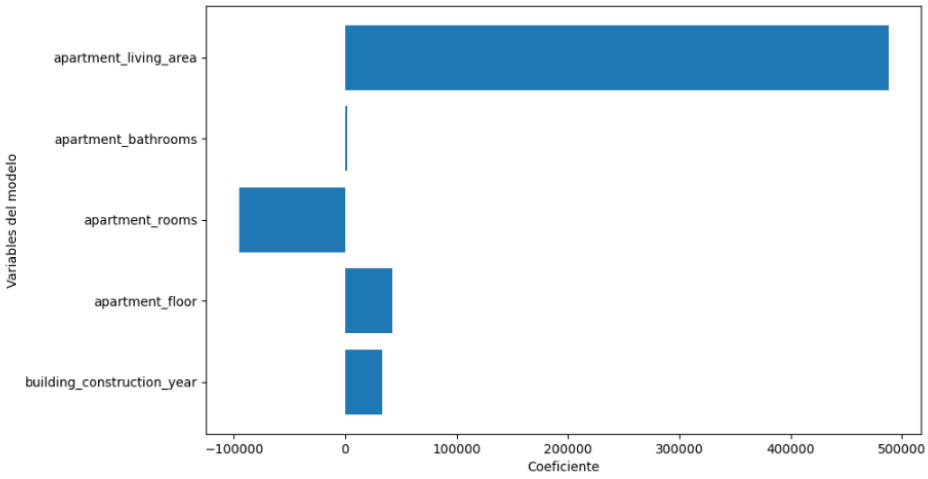
Modelo	Métrica	Valor
Regresión Lineal	R2	0,226
	Coef. de Correlación de Pearson	0,602
	MAPE	109,55%
	MPE	-42,78%
Regresión Ridge	R2	0,212

Fuente: Elaboración propia

Los resultados mostrados en la tabla contienen las métricas más relevantes para evaluar el desempeño de los modelos de regresión. Ambos modelos tienen un coeficiente de determinación (R2) similar, indicando que explican una proporción similar de la variabilidad en los datos. Sin embargo, el modelo de regresión lineal múltiple muestra un coeficiente de correlación de Pearson alto, lo que sugiere una relación lineal fuerte entre las variables. Además, el MAPE y MPE indican que las predicciones del modelo de regresión lineal tienen un error porcentual medio relativamente bajo en comparación con el modelo de regresión Ridge.

En el siguiente gráfico y la posterior tabla, pueden observarse los coeficientes correspondientes a cada variable en el modelo de regresión Ridge, estos coeficientes representan la magnitud y dirección de la influencia que tiene cada variable del modelo en el precio de la vivienda. Si un coeficiente es positivo, significa que un aumento en esa variable está asociado con un aumento en el precio, mientras que si es negativo está asociado con la disminución del pecio.

Figura 7. Coeficientes del modelo de Regresión Ridge



Fuente: Elaboración propia

Figura 8. Valor numérico de los coeficientes con Regresión Ridge

	Coeficiente
building_construction_year	1148.741047
building_total_floors	5769.584321
apartment_floor	939.886270
apartment_rooms	-99893.033849
apartment_bathrooms	-10038.272203
apartment_total_area	3053.236498
apartment_living_area	5395.895220

Fuente: Elaboración propia

Se observa que la variable "apartment_living_area" tiene el coeficiente más grande y positivo, lo que sugiere que el área habitable del apartamento tiene una influencia muy fuerte y positiva en su precio. Por otro lado, la variable "apartment_rooms" tiene un coeficiente negativo, lo que indica que el número de habitaciones está asociado con una disminución en el precio del apartamento. Un aumento de un año de construcción del edificio se asocia con un incremento de aproximadamente 1148,74 dólares en el precio, mientras que cada piso adicional en el edificio aumenta el precio en alrededor de 5769,58 dólares. Sin embargo, cada habitación adicional en el apartamento se relaciona con una disminución de aproximadamente 99893,03 dólares en el precio, al igual que cada baño adicional, que se asocia con una reducción de 10038,27 dólares. Tanto el área total como el área habitable de la vivienda muestran una asociación positiva con el precio, con aumentos de aproximadamente 3053,24 y 5395,90 dólares, respectivamente. Estos resultados sugieren que, aunque características como la antigüedad y el tamaño del edificio tienden a aumentar el precio, el número de habitaciones y baños puede tener un impacto negativo significativo. Sin embargo, es importante recordar que estos

resultados están sujetos a un $R2$ de 0,212, lo que indica que el modelo puede no capturar completamente la variabilidad de los datos.

b. Modelo de Árboles de Decisión

En el contexto de la regresión, los árboles de decisión dividen el espacio de características en regiones rectangulares y asignan una predicción a cada región. El proceso de construcción de un árbol de decisión implica dividir repetidamente el espacio de características en subconjuntos más pequeños y homogéneos, de manera que las predicciones sean lo más precisas posible. Esto se logra seleccionando la característica y el punto de corte que mejor separan las observaciones de entrenamiento. Además, es importante destacar que los árboles de decisión tienen una estructura jerárquica compuesta por un nodo raíz, ramas, nodos internos y nodos hoja, lo que les permite realizar evaluaciones basadas en las características disponibles para formar subconjuntos homogéneos. (IBM, s.f.)

En contraste, el Modelo de Bosques Aleatorios o “Random Forest” es una técnica más flexible y robusta que se beneficia de la combinación de múltiples árboles de decisión. Este modelo puede manejar relaciones no lineales y capturar interacciones complejas entre variables. Además, proporciona una medida de la importancia relativa de cada variable en la predicción. En resumen, los bosques aleatorios son un enfoque avanzado de modelado predictivo que aprovechan la aleatorización y la agregación para proporcionar resultados precisos en la predicción de evento (Rigatti, 2017)

En cuanto a las conclusiones que ofrece cada modelo, el Modelo de Regresión Lineal puede ser eficaz cuando las relaciones son aproximadamente lineales y se prioriza la interpretabilidad. Por otro lado, el Modelo de Bosques Aleatorios es más adecuado para capturar relaciones no lineales y manejar conjuntos de datos complejos, aunque su interpretabilidad puede ser más desafiante. Después de entrenar el modelo con el conjunto de entrenamiento previamente definido, se realizan predicciones sobre el conjunto de datos de prueba y las métricas de evaluación del modelo obtenidas son las siguientes:

Tabla 2. Resultados de Árboles de Decisión

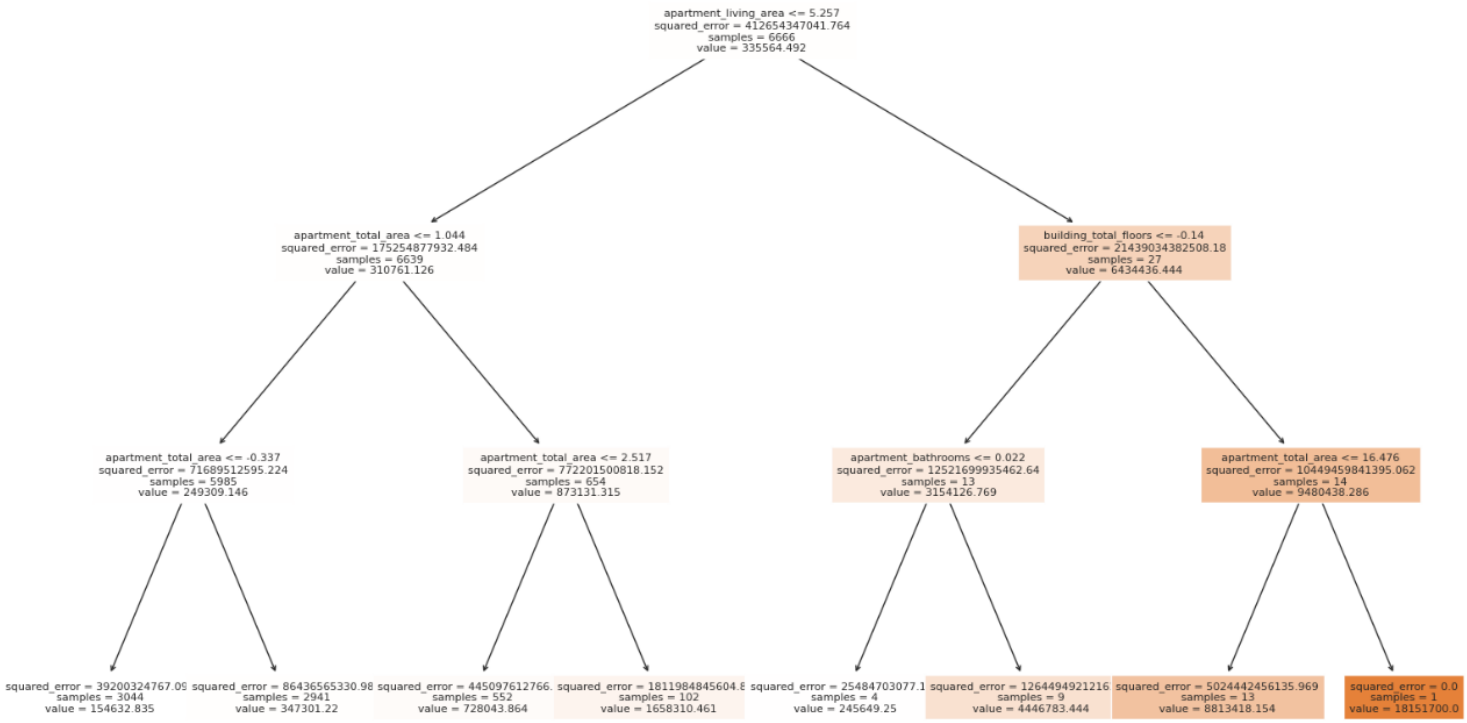
Métrica	Valores
R2	-0,12
Coeficiente de Correlación Pearson	0,51
MAPE	76,21%
MPE	-41,91%

Fuente: Elaboración propia

A la vista de estos resultados y los anteriores sobre los modelos de regresión, en este caso el modelo de árboles de decisión tiene un coeficiente de determinación (R2) negativo y una correlación de Pearson ligeramente menor en comparación con la regresión lineal, muestra una mejora en los errores porcentuales absolutos y medios. Esto indica que el árbol de decisión puede proporcionar una mejor precisión en términos porcentuales en la predicción de los precios de los apartamentos en comparación con la regresión lineal. Sin embargo, aun así, es necesario considerar mejoras adicionales para ambos modelos.

El modelo de árbol de decisión puede ser visualizado para proporcionar una compresión más clara de cómo se toman las decisiones dentro del modelo. Utilizando la función *'plot_tree'* de la biblioteca *'sklearn.tree'* junto con *'matplotlib.pyplot'* para generar una representación gráfica del árbol de decisión entrenado. En la visualización resultante, cada nodo del árbol representa una característica de entrada, cada rama muestra una regla de decisión basada en esa característica, y cada hoja indica la predicción final. Esta representación visual nos permite interpretar fácilmente cómo el modelo clasifica o regresa valores basados en las características de entrada. A continuación, se muestra la visualización del árbol de decisión resultante:

Figura 9. Estructura Árbol de Decisión



Fuente: Elaboración propia

Gracias a esta visualización podemos ver que la raíz del árbol de decisión está asociada con la variable “apartment_living_area” lo que sugiere su importancia como variable para la clasificación de los inmuebles. Las diferentes ramas del árbol van mostrando las diferentes condiciones que pueden tomar las variables. Destaca la importancia del precio por metro cuadrado, el área total y el número de dormitorios como los principales factores influyentes en la clasificación del precio de los apartamentos.

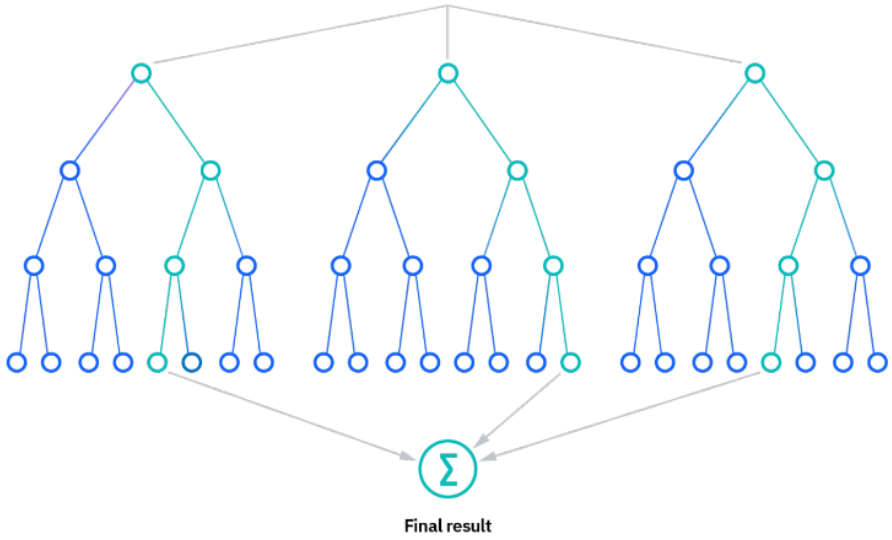
c. Modelo de Bosques Aleatorios

Los bosques aleatorios son un algoritmo fundamental en el campo del aprendizaje automático que se basa en la combinación de múltiples árboles de decisión para producir predicciones precisas y robustas. Este enfoque, desarrollado por Leo Breiman y Adele Cutler en 1996, ofrece varias ventajas clave sobre otros modelos de aprendizaje automático. Este tipo de algoritmo permite abordar el problema de sobreajuste que tienden a tener los algoritmos de Árboles de Decisión. Cada árbol se entrena en una muestra aleatoria del conjunto de datos original, lo que garantiza una baja correlación entre los árboles individuales. Además, en cada división de un árbol, solo se considera

un subconjunto aleatorio de características, lo que añade más diversidad al modelo y reduce el riesgo de sobreajuste.

El funcionamiento de este algoritmo se basa en tres hiperparámetros principales: tamaño de nodo, número de árboles y número de características muestreadas, que se establecen antes del entrenamiento. Consiste en una colección de árboles de decisiones, cada uno formado por una muestra de datos extraída de un conjunto de entrenamiento con sustitución, llamada muestra de programa de arranque. Se agrega aleatoriedad mediante la selección de un subconjunto aleatorio de características en cada árbol, lo que reduce la correlación entre ellos (IBM).

Figura 10. Ejemplo de funcionamiento de los Bosques Aleatorios



Fuente: IBM

Aunque los bosques aleatorios ofrecen muchas ventajas, como la reducción del sobreajuste y la flexibilidad en el manejo de datos, también presentan desafíos, como el procesamiento lento y el consumo de recursos. Sin embargo, con una configuración adecuada y una comprensión completa de sus capacidades y limitaciones, los bosques aleatorios pueden ser una herramienta poderosa para resolver una variedad de problemas de aprendizaje automático en el mundo real (IBM, s.f.). Igual que en los otros modelos expuestos anteriormente, en la siguiente tabla se muestran algunas métricas de evaluación del modelo para obtener una visión completa del rendimiento del modelo de Bosques Aleatorios.

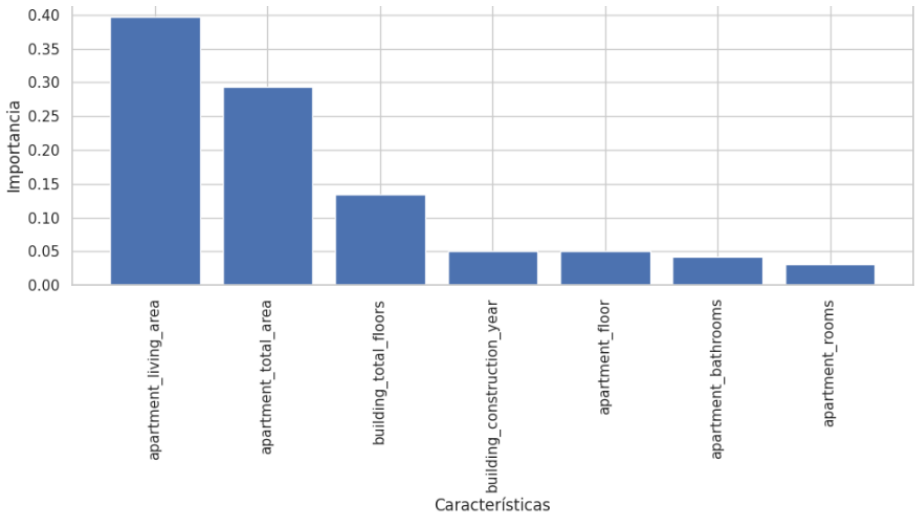
Tabla 3. Resultados del Modelo de Bosques Aleatorios

Métrica	Valores
R2	0,31
Coeficiente de Correlación Pearson	0,65
MAPE	68,54%
MPE	-47,51%

Fuente: Elaboración propia

Los resultados obtenidos con este modelo superan, términos generales, a los discutidos previamente lo que sugiere que los Bosques Aleatorios predicen mejor el precio de la vivienda. Para este modelo de Bosques Aleatorios la importancia de las variables o características en los resultados obtenidos pueden observarse en la siguiente figura, siendo el área de la vivienda (tanto total como habitable) la variable más relevante.

Figura 11. Importancia de las variables en el modelo de Bosques Aleatorios



Fuente: Elaboración propia

d. Modelo de Máquinas de Vectores de Soporte (SVM)

Otro enfoque utilizado en este trabajo es el modelo de Maquinas de Vectores de Soporte (SVM), un algoritmo de aprendizaje supervisado que tiene como objetivo encontrar el hiperplano que mejor separa las clases o características en el espacio de las

características (IBM, 2017)

La implementación de este tipo de modelos en el análisis implica la optimización de parámetros clave como el tipo de kernel. Lo primero es entrenar el modelo, para poder después evaluar su rendimiento. Como base de este algoritmo SVM se encuentra la función de pérdida o pérdida hinge. Esta función mide la distancia entre la predicción del modelo y las etiquetas reales de los datos de entrenamiento. La pérdida hinge penaliza las predicciones incorrectas de manera proporcional a su distancia desde el hiperplano de decisión y se define como:

$$\text{Hinge Loss}(y, f(x)) = \max(0, 1 - y \cdot f(x))$$

Donde:

- *y* es la etiqueta de clase verdadera (1 para la clase positiva, -1 para la clase negativa).
- *f*(*x*) es la predicción del modelo SVM.

Otra fórmula importante para los modelos de SVM es la Formula del Problema de Optimización que puede expresarse como:

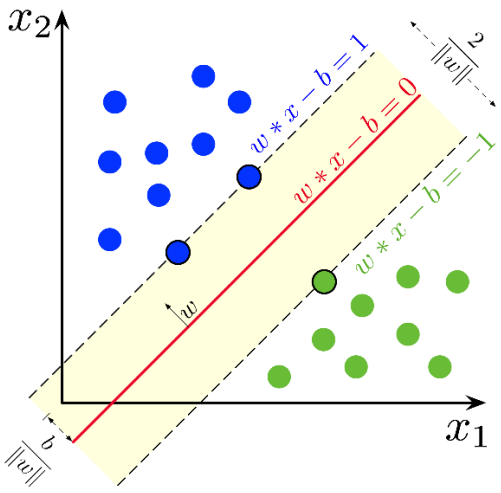
$$\min_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \text{Hinge Loss}(y_i, f(x_i))$$

Donde:

- *w* es el vector de pesos.
- *b* es el término de sesgo.
- *C* es el parámetro de regularización, que controla la compensación entre la maximización del margen y la minimización del error de clasificación.
- *N* es el número de muestras de entrenamiento.

En la siguiente figura puede apreciarse una explicación grafica del hiperplano óptimo de separación en los modelos de SVM:

Figura 12. Ejemplo de hiperplano óptimo en SVM



Fuente: Torres Barrán (2024)

Durante el entrenamiento del modelo, el algoritmo SVM ajusta los parámetros del modelo (pesos w y sesgo b) para minimizar la perdida hinge. Las métricas de evaluación de rendimiento del modelo obtenidas después de entrenar el modelo con los datos son las siguientes:

Tabla 4. Resultados del Modelo SVM

Métrica	Valores
R2	-0,0368
Coeficiente de Correlación Pearson	0,532
MAPE	78,44%
MPE	-36,2%

Fuente: Elaboración propia

Los resultados de este modelo pueden estar bastante alejadas de los valores reales viendo los resultados de las métricas evaluadas, al ser el coeficiente de determinación (R2) negativo, lo que indica que el modelo no se ajusta bien a los datos y es peor que los otros mostrados anteriormente.

e. Redes Neuronales

Las redes neuronales son modelos de computación inspirados en la estructura interconectada de las neuronas en el cerebro humano. Estas redes están compuestas por capas de nodos interconectados, que reciben entradas, realizan cálculos y generan salidas. A través de un proceso de aprendizaje a partir de los datos, las redes neuronales pueden reconocer patrones, clasificar datos y predecir eventos futuros (Atria, s.f.)

Son útiles en modelos de regresión porque pueden descomponer las entradas en capas de abstracción y aprender de ejemplos para reconocer patrones complejos en los datos. Al igual que el cerebro humano, las redes neuronales pueden ajustar automáticamente las conexiones entre sus elementos individuales durante el entrenamiento para llevar a cabo una tarea específica de manera precisa. Esto les permite modelar relaciones complejas entre variables de entrada y salida, lo que resulta especialmente valioso en problemas de regresión donde se busca predecir un valor continuo a partir de múltiples variables de entrada (Mathworks, 2024).

Con el objetivo de evaluar el rendimiento de este modelo, igual que en los casos anteriores, se han calculado las siguientes métricas de evaluación:

Tabla 5. Resultados del Modelo de Redes Neuronales

Métrica	Valor
Coeficiente de determinación (R2)	0,12
Coeficiente de correlación de Pearson	0,618
MPE	-56,42%
MAPE	79,06%

Fuente: Elaboración propia

En este caso, el R2 sugiere que el modelo explica aproximadamente el 13% de la variabilidad en los precios de las propiedades. Al tener un coeficiente de Pearson alto, de 0,62, esto indica que la correlación entre las predicciones del modelo y los valores reales es fuerte y positiva, un MPE negativo indica que las predicciones son, en

promedio, menores que los valores reales, y por último, un MAPE del 80,42% significa que las predicciones, de media, tienen ese error en comparación con los valores reales, lo que es un valore demasiado alto como para poder decir que el algoritmo de redes neuronales predice correctamente.

f. Comparación de modelos

La comparación entre los diferentes modelos utilizados en este trabajo es fundamental para concluir cuál de ellos es el que predice mejor el precio de la vivienda, que es la variable objetivo de este trabajo. Además, al evaluar diferentes modelos, se obtiene una compresión más profunda de como cada algoritmo aborda el problema y cuáles son sus fortalezas y debilidades.

En general, el coeficiente de determinación (R2) es considerado una de las métricas más importantes en la evaluación de modelos de predicción de resultados, al ser una medida que proporciona la variabilidad que es explicada por el modelo en cuestión. UN valor de R2 más cercano a uno indica un mejor ajuste del modelo a los datos y que por tanto explica mejor la variabilidad en los precios de las viviendas. En la siguiente tabla podemos observar el valor de R2 para los distintos modelos:

Figura 13. Comparativa de R2 para los diferentes modelos

Modelo	R2
Regresión Lineal	0,226
Regulación Ridge	0,212
Árboles de Decisión	-0.12
Bosques Aleatorios	0,31
SVM	-0,037
Redes Neuronales	0,12

Fuente: Elaboración propia

Es importante reconocer que cada modelo tiene sus propias características distintivas que ofrecen ventajas y limitaciones específicas en el contexto de la predicción del precio de la vivienda. La Regresión Lineal se destaca por su simplicidad y facilidad de interpretación, lo que la convierte en una opción atractiva para análisis preliminares. La Regulación Ridge, por otro lado, aborda eficazmente el problema del sobreajuste al introducir penalizaciones en los coeficientes, lo que resulta en una mayor estabilidad en la predicción. Los Árboles de Decisión son fáciles de entender y visualizar, lo que facilita la interpretación de los resultados, pero tienden a ser propensos al sobreajuste, especialmente en conjuntos de datos complejos y este sobreajuste es lo que mejoran los modelos de Bosques Aleatorios. Las Máquinas de Vectores de Soporte (SVM) son especialmente eficientes en espacios de alta dimensión y pueden manejar conjuntos de datos con una separación no lineal, aunque su desempeño puede verse afectado por la elección del kernel y la optimización de los parámetros. Por último, las Redes Neuronales tienen la capacidad de capturar patrones complejos y no lineales en los datos, lo que las hace ideales para problemas de alta complejidad, pero su entrenamiento puede ser computacionalmente costoso, requerir grandes cantidades de datos y presentar desafíos en la interpretación de los resultados.

Teniendo en cuenta que la comparación de modelos se está haciendo en base al coeficiente de determinación, el modelo que mejor predice es el de Bosques Aleatorios, con un valor de 0,31. Este resultado sugiere que el modelo tiene mayor capacidad para explicar la variabilidad de los datos, lo cual puede atribuirse a la naturaleza del algoritmo de Bosques Aleatorios, que combina múltiples árboles de decisión y utiliza la aleatoriedad de características para reducir la correlación entre los árboles individuales. Al integrar las predicciones de múltiples árboles, el modelo de Bosques Aleatorios puede capturar una gama más amplia de patrones en los datos y generalizar mejor a nuevos conjuntos de datos, lo que se traduce en un mejor rendimiento predictivo en comparación con los otros modelos considerados.

5. Análisis de los Resultados

La inteligencia artificial (IA) desempeña un papel cada vez más relevante en el mercado inmobiliario, siendo fundamental para socimis, promotoras, y entidades financieras, entre otros actores. Estos algoritmos permiten procesar grandes volúmenes de datos para predecir con mayor precisión el valor de una propiedad, identificar tendencias del mercado, y personalizar recomendaciones para los clientes. Ofrece nuevas oportunidades para la toma de decisiones estratégicas basadas en análisis predictivos, así como para la identificación de oportunidades de inversión y áreas de riesgo en el mercado inmobiliario. Por todo esto, la inteligencia artificial está transformando la forma en que se realizan las valoraciones y la toma de decisiones en el mercado inmobiliario, ofreciendo una mayor precisión y eficiencia en los procesos de negocio (Pierna, 2018).

En el análisis comparativo de modelos de predicción de precios inmobiliarios, se han evaluado varios modelos, desde regresión lineal hasta redes neuronales, con el objetivo de determinar cuál de ellos ofrece la mejor capacidad predictiva. Sin embargo, los resultados obtenidos a lo largo de este trabajo revelan ciertas carencias en el rendimiento de los modelos, lo que plantea interrogantes sobre su efectividad en la predicción de precios de propiedades.

Existen diversas razones por las cuales los modelos pueden no haber alcanzado un nivel de precisión deseado:

I. Limitación de Variables: si bien se incluyen variables importantes en los datos como el año de construcción, el número de pisos o el área de las viviendas, es posible que existan otras variables relevantes que hayan limitado la capacidad predictiva del modelo.

II. Tamaño y calidad del conjunto de datos: un conjunto de datos pequeño puede no capturar toda la variabilidad presente en el objeto de estudio y teniendo en cuenta que los datos originales de Kaggle contaban con 147.537 observaciones pero que tras eliminar las observaciones con valores ‘N/A’ en las variables ‘apartment_total_area’, ‘apartment_bedrooms’, ‘apartment_bathrooms’ y ‘building_construction_year’ se tienen 8.333 observaciones esto ha podido influir en la fiabilidad de los modelos de predicción.

III. Multicolinealidad: La multicolinealidad entre las variables presentes en los datos también puede haber afectado al rendimiento del modelo, esta se da cuando dos o más variables están altamente correlacionadas entre sí, lo que puede dificultar la interpretación de los coeficientes del modelo y conducir a estaciones sesgadas.

IV. Ajuste del modelo: Dado que el RMSE del conjunto de prueba es menor que el del conjunto de entrenamiento y los valores de RMSE son altos en general, es posible que el modelo necesite ajustes adicionales para mejorar su capacidad predictiva. Esto podría incluir la consideración de características adicionales, la exploración de modelos más complejos o la aplicación de técnicas de regularización para evitar el sobreajuste.

En definitiva, todos estos factores han podido influenciar en la capacidad de predicción del precio de la vivienda y que por tanto las estimaciones obtenidas no sean tan buenas como cabría esperar, pero sí que hay una conclusión más clara sobre que variables, de las presentes en el conjunto de datos, afectan más al precio de la vivienda.

6. Conclusiones

La creciente influencia de la inteligencia artificial (IA) en el sector inmobiliario español está redefiniendo la manera en que se realizan las transacciones y se toman decisiones. Aunque la automatización de tareas y procesamiento de grandes volúmenes de datos ofrecen eficiencia y precisión, también plantean desafíos en cuanto a la protección de datos y el futuro del empleo. No obstante, la IA presenta oportunidades estratégicas para los actores del mercado al proporcionar análisis predictivos detallados y personalizar experiencias para los clientes, marcando un cambio significativo en la dinámica del sector (Rivera Carmona, 2023; Pierna, 2018).

En cuanto a los resultados del análisis en este trabajo sobre el sector inmobiliario y las variables más influyentes en la predicción del precio de la vivienda, se revela lo siguiente:

En primer lugar, se identificaron variables fundamentales como el tamaño del apartamento, el año de construcción del edificio y el número de habitaciones y baños como factores determinantes en la predicción del precio de la vivienda. Esto hace

referencia a la importancia de aspectos específicos del inmueble y su entorno en la valoración del mercado, lo cual no se ha tenido especialmente en cuenta en los modelos ya que solo se han utilizado las variables numéricas.
En segundo lugar, entre los modelos de predicción evaluados, se destacó el modelo de Bosques Aleatorios como el más eficaz en la estimación del precio de la vivienda, ofreciendo una mayor capacidad para explicar la variabilidad en los datos.

Dadas las limitaciones identificadas en todos los modelos empleados, como la posibilidad de incorporar variables adicionales relevantes y realizar ajustes para mejorar la precisión de las predicciones, es crucial reconocer la limitación que supone analizar todas las observaciones sin distinción y tratarlas como si pertenecieran al mismo país. Esta aproximación, realizada al eliminar la variable país para entrenar los modelos, podría resultar en una interpretación sesgada de los resultados. Las diferencias económicas, sociales y políticas entre países pueden tener un impacto significativo en el mercado inmobiliario, lo que significa que la suposición de homogeneidad entre las observaciones puede no reflejar con precisión la complejidad y diversidad del mercado real. Por lo tanto, es esencial reconocer estas limitaciones y abordarlas adecuadamente para garantizar una interpretación más precisa y sólida de los resultados obtenidos.

Para futuras líneas de investigación sobre este tema, sería interesante incluir en los datos variables complementarias que puedan enriquecer el análisis del mercado inmobiliario, como indicadores socioeconómicos locales o regionales, índices de calidad de vida, tasas de criminalidad, acceso a servicios públicos, entre otros. Asimismo, se podría considerar la aplicación de metodologías más avanzadas, con el fin de potenciar la capacidad predictiva de los modelos y profundizar en la comprensión de las relaciones subyacentes entre las variables.

7. Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, M.ª Fernanda Rius Matas estudiante de E2 + Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Inversiones en timberland y forestland: la perspectiva de las carteras de inversión" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
2. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
3. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
4. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
5. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.
Fecha: marzo 2024

Firma: _____

8. Bibliografía:

Arrieta, L., Grajales, G. y Padilla, J. (2023). "La responsabilidad social de las empresas con el medio ambiente, caso de estudio Arrendamientos COPABIENES Ltda." (Trabajo de grado). Corporación Universitaria Minuto de Dios, Bogotá – Colombia. [En línea]. Disponible en: https://repository.uniminuto.edu/handle/10656/18487

BBVA Research. (2023). "Observatorio inmobiliario 1S23, Primer semestre 2023." (BBVA Research.) [En línea]. Disponible en: https://www.bbvaresearch.com/wp-content/uploads/2023/04/Real-Estate-EW_1Q23.pdf

BBVA Research. (6 de marzo de 2024). Situación España. Marzo 2024. Recuperado de https://www.bbvaresearch.com/publicaciones/situacion-espana-marzo-2024/#:~:text=Se%20revisa%20al%20alza%20el,hasta%20el%202%2C0%25

Breuer, W., & Steininger, B. I. (2020). Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms. Journal of Business Economics, 90, 963–974. https://doi.org/10.1007/s11573-020-01005-w

Carbó Valverde, S., & Rodríguez Fernández, F. (2018). "El mercado de la vivienda en Europa: viejas costumbres y nuevos desafíos." Cuadernos de Información Económica, 266, 81-92. [En línea]. Disponible en: https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_CIE/266art09.pdf

CBRE. "Informe de tasaciones segundo trimestre de 2023." [En línea]. Disponible en: https://www.cbre.es/insights/reports/como-ha-evolucionado-el-precio-de-la-vivienda-en-espana

Colliers. (2023). "Forecasting Yields." https://www.colliers.com/es-es/research/forecasting-yields

Colliers. (2023). El sector residencial en España en 2023 [Informe interno]. Recuperado de: https://www.colliers.com/es-es/research/informe-residencial-2023

Fondo Monetario Internacional. (2022). Perspectivas de la economía mundial: Afrontar la crisis del costo de vida. Washington, DC: Autor. Recuperado de: https://www.imf.org/es/Publications/WEO/Issues/2022/10/11/world-economic-outlook-october-2022

García Vega, C. I. (2023). "Las ciudades frente a los retos del siglo XXI y espacios públicos para el desarrollo, aproximaciones desde los casos de Morelia y Puebla en México" (Tesis de maestría). Benemérita Universidad Autónoma de Puebla. Recuperado de: https://hdl.handle.net/20.500.12371/18598

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55–67. https://doi.org/10.2307/1267351

Hott, C., Monnin, P. "Fundamental Real Estate Prices: An Empirical Estimation with International Data." J Real Estate Finan Econ, 36, 427–450 (2008). https://doi.org/10.1007/s11146-007-9097-8

IBM. (s.f.). Árboles de decisión. https://www.ibm.com/docs/es/spss-statistics/saas?topic=trees-creating-decision

IBM. (s.f.). Random forest. Recuperado de https://www.ibm.com/es-es/topics/random-forest

Inflation: what does the academic research say? (s. f.). Man Institute | Man Group. [En línea]. Disponible en: https://www.man.com/maninstitute/inflation-what-the-academic-research-says

Leamer, E. E. (2007). Housing is the business cycle. NBER Working Paper Series, No. 13428. Recuperado de http://www.nber.org/papers/w13428

Pavlov, A. D., & Wächter, S. M. (2010). "Subprime lending and real estate prices." Real Estate Economics, 39(1), 1-17. https://doi.org/10.1111/j.1540-6229.2010.00284.x

Pierna, E. (2018, febrero 26). El papel de la inteligencia artificial en la valoración inmobiliaria. LinkedIn. https://es.linkedin.com/pulse/el-papel-de-la-inteligencia-artificial-en-valoraci%C3%B3n-elena-pierna

PricewaterhouseCoopers. (2023). "Tendencias en el mercado inmobiliario en Europa 2023." PwC. [En línea]. Disponible en: https://www.pwc.es/es/real-estate/tendencias-mercado-inmobiliario-europa-2023.html

Rigatti, S. J. (2017). Random Forest. Journal of Insurance Medicine, 47, 31–39. http://meridian.allenpress.com/jim/article-pdf/47/1/31/1736157/insm-47-01-31-39_1.pdf

Rivera Carmona, J. C. (2023). Cómo la inteligencia artificial puede cambiar el sector inmobiliario en España. Fotocasa Blog Profesional. https://blogprofesional.fotocasa.es/como-la-inteligencia-artificial-puede-cambiar-el-sector-inmobiliario-en-espana/

Rodríguez, O. (11 de abril de 2024). Los españoles pagan un 28% más por su alquiler e ingresan un 16% más por su salario desde 2015. El Independiente. https://www.elindependiente.com/

Rodríguez-López, J. (2022). El mercado de vivienda en un escenario internacional complejo. Ciudad Y Territorio Estudios Territoriales, 54(212), 469–482. https://doi.org/10.37230/CyTET.2022.212.11

Rodríguez-López, J. (2023). "Mercado de vivienda: destaca la fortaleza de la demanda." Ciudad y Territorio, Estudios Territoriales, LV (215), 223-240. https://doi.org/10.37230/CyTET.2023.215.13

Samadani, S., & Costa, C. J. (2021). "Forecasting real estate prices in Portugal: A data science approach." 16th Iberian Conference on Information Systems and Technologies. https://doi.org/10.23919/cisti52073.2021.9476447

Taltavull, P. (2020). "Los cuatro retos del mercado inmobiliario para las ciudades." Panorama Social, 32, 107-126. [En línea]. Disponible en: https://www.funcas.es/wp-content/uploads/2021/02/Paloma-Taltavull.pdf

Torío, E. (2021). “La industrialización guía el futuro de la vivienda”. El Economista. https://www.eleconomista.es/vivienda/noticias/11500024/11/21/La-industrializacion-guia-el-futuro-de-la-vivienda.html

9. Anexos

```
from google.colab import drive
drive.mount('/content/drive')
import pandas as pd

# Ruta al archivo XLSX en Google Drive
file_path_xlsx = '/content/drive/My Drive/data_TFG_BA.xlsx'

# Leer el archivo XLSX
df = pd.read_excel(file_path_xlsx)

# Imprim irel DataFrame
print(df.head())
```

```
Mounted at /content/drive
                                        title     country  \
0  1 room studio apartment 22 in Budva, Montenegro  Montenegro
1         1 room apartment 23 in Bangkok, Thailand    Thailand
2         1 room apartment 23 in Pattaya, Thailand    Thailand
3             1 room apartment in Phuket, Thailand    Thailand
4  1 room studio apartment 23 in Budva, Montenegro  Montenegro

                                location  building_construction_year  \
0  Budva, Bar, Bar Municipality, Montenegro                        2016
1                      Bangkok, Thailand                        2023
2    Chon Buri Province, Pattaya, Thailand                        2023
3        Phuket Province, Phuket, Thailand                        2021
4  Budva, Bar, Bar Municipality, Montenegro                        2018

   building_total_floors  apartment_floor  apartment_rooms  \
0                      4              2.0                2
1                      8              2.0                2
2                      8              2.0                2
3                      7              7.0                1
4                      4              4.0                2

   apartment_bedrooms  apartment_bathrooms  apartment_total_area  \
0                   1                    1                    22
1                   1                    1                    23
2                   1                    1                    23
3                   1                    1                    23
4                   1                    1                    23

   apartment_living_area  price_in_USD
0                     20        102702
1                     23         59546
2                     23         50831
3                     23         59000
4                     20         62101
```

```
df.head(1)
```

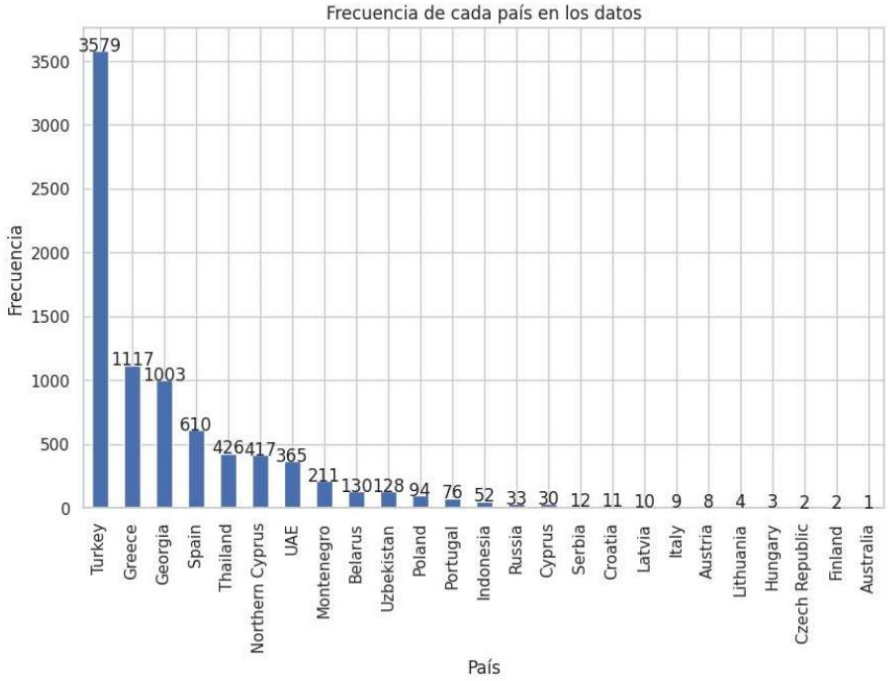
	title	country	location	building_construction_year	building_total_floors	apartment_floor	apartment_rooms	apartment
	1 room studio		Budva, Bar,					

Next steps: View recommended plots

```
df.describe()
```

	building_construction_year	building_total_floors	apartment_floor	apartment_rooms	apartment_bedrooms	apartment_bathrooms
count	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000	8333.000000
mean	2018.148326	10.066963	4.880187	2.822033	1.920317	1.484219
std	11.820675	10.870891	7.096037	1.074241	0.952993	0.655416
min	1894.000000	1.000000	-1.000000	1.000000	1.000000	1.000000
25%	2021.000000	4.000000	1.000000	2.000000	1.000000	1.000000
50%	2023.000000	6.000000	2.000000	3.000000	2.000000	1.000000
75%	2023.000000	13.000000	5.000000	3.000000	2.000000	2.000000
max	2024.000000	115.000000	105.000000	10.000000	10.000000	10.000000

```
import matplotlib.pyplot as plt


plt.figure(figsize=(10, 6))
df['country'].value_counts().plot(kind='bar')
plt.xlabel('País')
plt.ylabel('Frecuencia')
plt.title('Frecuencia de cada país en los datos')

for i, v in enumerate(df['country'].value_counts()):
    plt.text(i, v + 0.1, str(v), ha='center')
plt.show()
```




```
#LIBRERIAS NECESARIAS

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.metrics import mean_squared_error
from sklearn.neural_network import MLPRegressor


from sklearn.preprocessing import StandardScaler
# Dividir los datos en conjunto de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(df[[ 'building_construction_year', 'building_total_floors', 'apartment_floor', 'apa

# Inicializa el objeto StandardScaler
scaler = StandardScaler()

# Ajusta el scaler a datos de entrenamiento y transforma los datos
X_train_scaled = scaler.fit_transform(X_train)

# misma transformación a datos de prueba
X_test_scaled = scaler.transform(X_test)
```

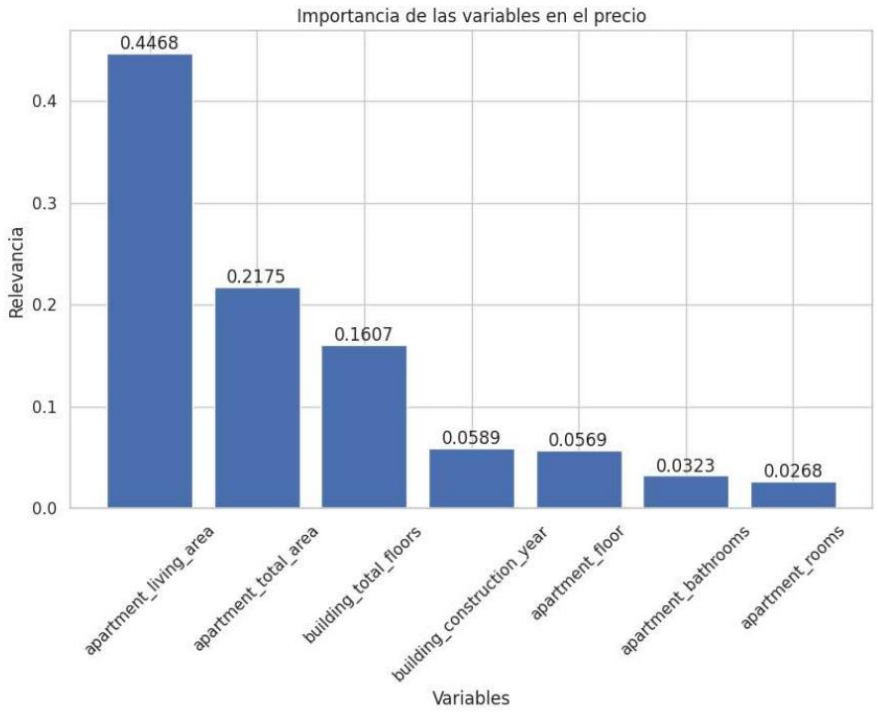
```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# Evaluar el modelo
y_pred_lr = lr_model.predict(X_test)
mse_lr = mean_squared_error(y_test, y_pred_lr)
print("Error cuadrático medio (MSE) para regresión lineal múltiple:", mse_lr)

# Coeficientes del modelo
coefficients = pd.DataFrame(lr_model.coef_, X_test.columns, columns=['Coeficiente'])
print("\nCoeficientes del Modelo:")
print(coefficients)
import statsmodels.api as sm

# Añadir una constante al conjunto de características
X_train_with_const = sm.add_constant(X_train)

# Crear y ajustar el modelo de regresión lineal múltiple
lr_model_sm = sm.OLS(y_train, X_train_with_const).fit()

# Obtener los resultados del modelo
print(lr_model_sm.summary())
```

```
Error cuadrático medio (MSE) para regresión lineal múltiple: 159541138027.37323

Coeficientes del Modelo:
                           Coeficiente
building_construction_year   1148.741047
building_total_floors        5769.584321
apartment_floor               939.886270
apartment_rooms            -99893.033849
apartment_bathrooms        -10038.272203
apartment_total_area         3053.236498
apartment_living_area        5395.895220
                            OLS Regression Results
==============================================================================
Dep. Variable:           price_in_USD   R-squared:                       0.481
Model:                            OLS   Adj. R-squared:                  0.481
Method:                 Least Squares   F-statistic:                     882.8
Date:                Tue, 23 Apr 2024   Prob (F-statistic):               0.00
Time:                        07:28:22   Log-Likelihood:                -96414.
No. Observations:                6666   AIC:                         1.928e+05
Df Residuals:                    6658   BIC:                         1.929e+05
Df Model:                           7
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                       -2.525e+06   1.02e+06     -2.476      0.013   -4.52e+06   -5.26e+05
building_construction_year  1148.7410    505.469      2.273      0.023     157.860    2139.622
building_total_floors       5769.5843    680.462      8.479      0.000    4435.661    7103.508
apartment_floor              939.8863   1003.176      0.937      0.349   -1026.659    2906.432
apartment_rooms            -9.989e+04   7722.265    -12.936      0.000   -1.15e+05   -8.48e+04
apartment_bathrooms        -1.004e+04   1.17e+04     -0.856      0.392    -3.3e+04     1.3e+04
apartment_total_area        3053.2365    295.645     10.327      0.000    2473.677    3632.796
apartment_living_area       5395.8952    313.179     17.229      0.000    4781.964    6009.826
==============================================================================
Omnibus:                     7937.466   Durbin-Watson:                   1.992
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2597595.311
Skew:                           5.916   Prob(JB):                         0.00
Kurtosis:                      98.981   Cond. No.                     3.64e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.64e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
from sklearn.metrics import mean_squared_error
import numpy as np

# Calcular el error del conjunto de entrenamiento
y_pred_train = lr_model.predict(X_train)
mse_train = mean_squared_error(y_train, y_pred_train)
rmse_train = np.sqrt(mse_train)
print("RMSE para conjunto de entrenamiento:", rmse_train)

# Calcular el error del conjunto de prueba
mse_test = mean_squared_error(y_test, y_pred_lr)
rmse_test = np.sqrt(mse_test)
print("RMSE para conjunto de prueba:", rmse_test)
```

```
RMSE para conjunto de entrenamiento: 462618.51519795833
RMSE para conjunto de prueba: 399426.01070457743
```

```
import matplotlib.pyplot as plt

# Entrenar el modelo de árbol de decisión
dt_model = DecisionTreeRegressor()
dt_model.fit(X_train, y_train)

# Obtener la importancia de las características
importances = dt_model.feature_importances_

# Ordenar las características por importancia
indices = importances.argsort()[::-1]
sorted_features = [X_train.columns[i] for i in indices]
sorted_importances = importances[indices]

# Visualizar la importancia de las características
plt.figure(figsize=(10, 6))
bars = plt.bar(range(X_train.shape[1]), sorted_importances, tick_label=sorted_features)
plt.xticks(rotation=45)
plt.xlabel('Variables')
plt.ylabel('Relevancia')
plt.title('Importancia de las variables en el precio')

# Agregar leyenda de valor exacto
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, height, round(height, 4), ha='center', va='bottom')

plt.show()
```




```
from sklearn.model_selection import cross_val_score

scores = cross_val_score(lr_model, X_train, y_train, cv=5, scoring='neg_mean_squared_error')
print("Error cuadrático medio (MSE) para validación cruzada:", -scores.mean())
```

```
Error cuadrático medio (MSE) para validación cruzada: 218076275464.63617
```

Modelo de Regresión Lineal Múltiple

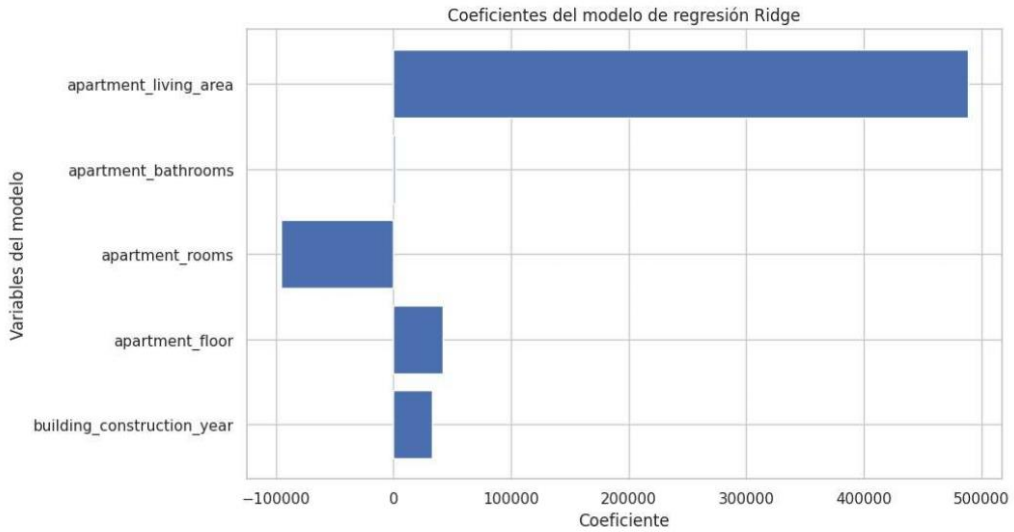
```
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import numpy as np
scaler = StandardScaler()

# Estandarizar las características
X_train_std = scaler.fit_transform(X_train)
X_test_std = scaler.transform(X_test)

# Entrenar el modelo de regresión lineal múltiple
lr_model = LinearRegression()
lr_model.fit(X_train_std, y_train)
y_pred_lr = lr_model.predict(X_test_std)


mse_lr = mean_squared_error(y_test, y_pred_lr)
mae_lr = mean_absolute_error(y_test, y_pred_lr)
rmse_lr = np.sqrt(mse_lr)
r2_lr = r2_score(y_test, y_pred_lr)
pearson_corr_lr = np.corrcoef(y_test, y_pred_lr)[0, 1]
mape_lr = np.mean(np.abs((y_test - y_pred_lr) / y_test)) * 100
mpe_lr = np.mean((y_test - y_pred_lr) / y_test) * 100

print("Métricas para regresión lineal múltiple:")

print("Coeficiente de determinación (R^2):", r2_lr)
print("Coeficiente de correlación de Pearson:", pearson_corr_lr)
print("Error porcentual absoluto medio (MAPE):", mape_lr)
print("Error medio porcentual (MPE):", mpe_lr)
```

```
Métricas para regresión lineal múltiple:
Coeficiente de determinación (R^2): 0.22617894422866902
Coeficiente de correlación de Pearson: 0.6016640282649326
Error porcentual absoluto medio (MAPE): 109.55147355106891
Error medio porcentual (MPE): -42.779484962825705
```

```
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np
# Variables seleccionadas
selected_features = ['building_construction_year', 'apartment_floor', 'apartment_rooms', 'apartment_bathrooms', 'apartment_living_area']

# Regularización con Ridge y búsqueda de hiperparámetros
param_grid = {'ridge__alpha': [0.1, 1, 10]}
ridge_pipeline = Pipeline([('scaler', StandardScaler()), ('ridge', Ridge())])
grid_search = GridSearchCV(ridge_pipeline, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train[selected_features], y_train)

# Mejor modelo encontrado
best_model = grid_search.best_estimator_

# Evaluación del modelo en el conjunto de prueba
y_pred = best_model.predict(X_test[selected_features])
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mejor modelo de regresión Ridge:")
print("Mejor valor de alpha:", grid_search.best_params_['ridge__alpha'])
print("Error cuadrático medio (MSE):", mse)
print("Coeficiente de determinación (R^2):", r2)
```

```
Mejor modelo de regresión Ridge:
Mejor valor de alpha: 10
Error cuadrático medio (MSE): 162385113186.84073
Coeficiente de determinación (R^2): 0.2123848351499877 4
```

```
import matplotlib.pyplot as plt

# Obtener los coeficientes del mejor modelo Ridge
coefficients = best_model.named_steps['ridge'].coef_
feature_names = selected_features

plt.figure(figsize=(10, 6))
plt.barh(feature_names, coefficients)
plt.xlabel('Coeficiente')
plt.ylabel('Variables del modelo')
plt.title('Coeficientes del modelo de regresión Ridge')
plt.show()
```



Regresion RIdge

```
from sklearn.linear_model import Ridge, Lasso

# Entrenar el modelo de regresión Ridge
ridge_model = Ridge(alpha=10.0)
ridge_model.fit(X_train_std, y_train)


# Predicciones en el conjunto de prueba para Ridge
y_pred_ridge = ridge_model.predict(X_test_std)

mse_ridge = mean_squared_error(y_test, y_pred_ridge)
mae_ridge = mean_absolute_error(y_test, y_pred_ridge)
rmse_ridge = np.sqrt(mse_ridge)
r2_ridge = r2_score(y_test, y_pred_ridge)

print("Métricas para regresión Ridge:")
print("Error cuadrático medio (MSE):", mse_ridge)
print("Error absoluto medio (MAE):", mae_ridge)
print("Raíz del error cuadrático medio (RMSE):", rmse_ridge)
print("Coeficiente de determinación (R^2):", r2_ridge)
```

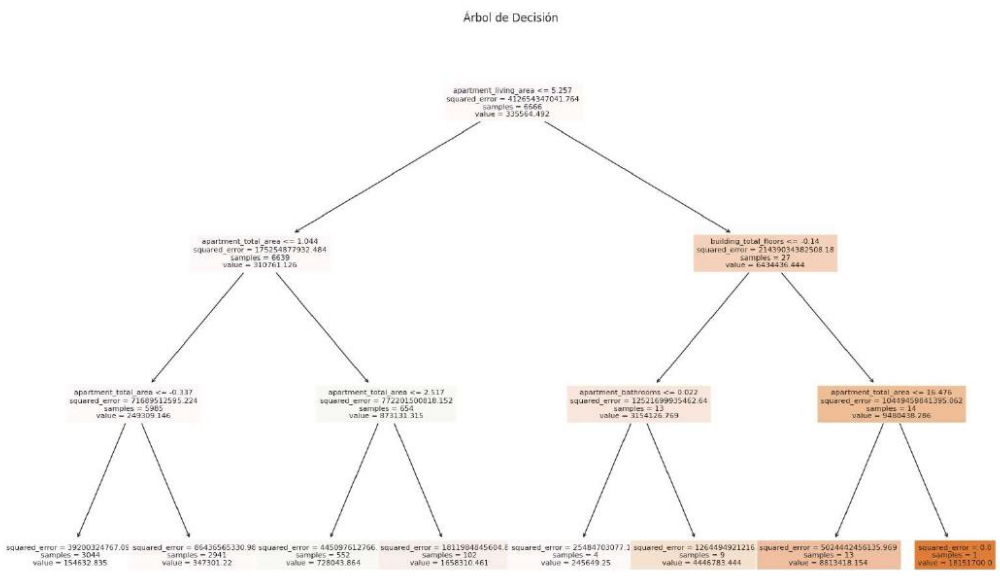
```
Métricas para regresión Ridge:
Error cuadrático medio (MSE): 159354045382.1208
Error absoluto medio (MAE): 220364.26605422952
Raíz del error cuadrático medio (RMSE): 399191.74012261425
Coeficiente de determinación (R^2): 0.227086398130944
```

Decision trees

```
from sklearn.tree import DecisionTreeRegressor

# Entrenar el modelo de árbol de decisión
dt_model = DecisionTreeRegressor()
dt_model.fit(X_train_std, y_train)

# Predicciones en el conjunto de prueba para árbol de decisión
y_pred_dt = dt_model.predict(X_test_std)

# Calcular métricas para árbol de decisión
mse_dt = mean_squared_error(y_test, y_pred_dt)
mae_dt = mean_absolute_error(y_test, y_pred_dt)
rmse_dt = np.sqrt(mse_dt)
r2_dt = r2_score(y_test, y_pred_dt)

import numpy as np
mape_dt = np.mean(np.abs((y_test - y_pred_dt) / y_test)) * 100
mpe_dt = np.mean((y_test - y_pred_dt) / y_test) * 100
pearson_corr_dt = np.corrcoef(y_test, y_pred_dt)[0, 1]

# Imprimir los resultados
print("Métricas adicionales para árbol de decisión:")
print("Coeficiente de determinación (R^2):", r2_dt)
print("Error porcentual absoluto medio (MAPE):", mape_dt)
print("Error medio porcentual (MPE):", mpe_dt)
print("Coeficiente de correlación de Pearson:", pearson_corr_dt)
```

```
Métricas adicionales para árbol de decisión:
Coeficiente de determinación (R^2): -0.11731573450733923
Error porcentual absoluto medio (MAPE): 76.21242732013499
Error medio porcentual (MPE): -41.9123578517138
Coeficiente de correlación de Pearson: 0.5095465957908362
```

```
from sklearn.tree import DecisionTreeRegressor, plot_tree
import matplotlib.pyplot as plt

# Entrenar el modelo de árbol de decisión con una profundidad máxima de 3 para poder visualizarlo mejor
dt_model = DecisionTreeRegressor(max_depth=3)
dt_model.fit(X_train_std, y_train)

plt.figure(figsize=(20,12))
plot_tree(dt_model, feature_names=X_train.columns, filled=True, fontsize=8)
plt.title("Árbol de Decisión ")
plt.show()
```

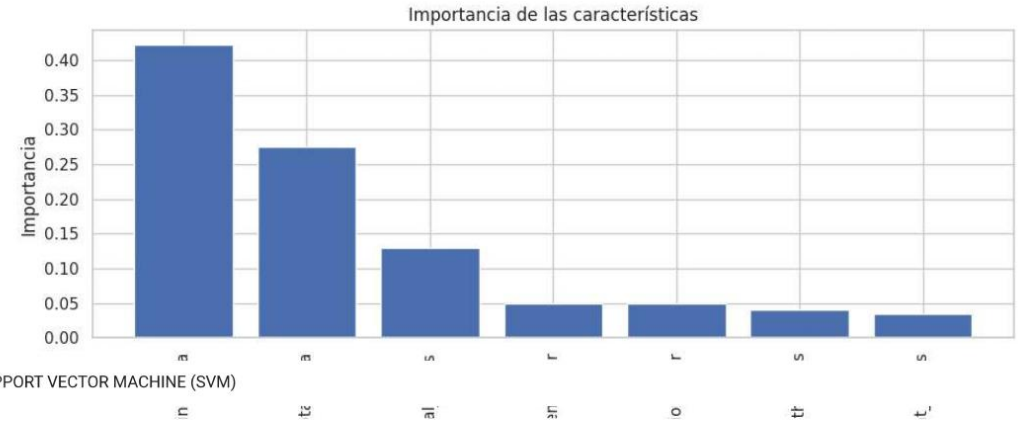


RANDOM FOREST

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
import numpy as np

# Entrenar el modelo de Random Forest
rf_model = RandomForestRegressor()
rf_model.fit(X_train_std, y_train)

# Predicciones en el conjunto de prueba para Random Forest
y_pred_rf = rf_model.predict(X_test_std)

# Calcular métricas para Random Forest
mse_rf = mean_squared_error(y_test, y_pred_rf)
mae_rf = mean_absolute_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mse_rf)
r2_rf = r2_score(y_test, y_pred_rf)

mape_rf = np.mean(np.abs((y_test - y_pred_rf) / y_test)) * 100
mpe_rf = np.mean((y_test - y_pred_rf) / y_test) * 100
pearson_corr_rf = np.corrcoef(y_test, y_pred_rf)[0, 1]

# Imprimir los resultados adicionales
print("Métricas adicionales para Random Forest:")
print("Coeficiente de determinación (R^2):", r2_rf)
print("Error porcentual absoluto medio (MAPE):", mape_rf)
print("Error medio porcentual (MPE):", mpe_rf)
print("Coeficiente de correlación de Pearson:", pearson_corr_rf)
```

```
Métricas adicionales para Random Forest:
Coeficiente de determinación (R^2): 0.30567456652638214
Error porcentual absoluto medio (MAPE): 68.547657586603
Error medio porcentual (MPE): -47.50953846320999
Coeficiente de correlación de Pearson: 0.649798598487092
```

```
import matplotlib.pyplot as plt

# Obtener la importancia de las características
importances = rf_model.feature_importances_
features = X_train.columns

# Ordenar las características por importancia
indices = np.argsort(importances)[::-1]

# Visualizar la importancia de las características
plt.figure(figsize=(10, 6))
plt.title("Importancia de las características")
plt.bar(range(X_train.shape[1]), importances[indices], align="center")
plt.xticks(range(X_train.shape[1]), features[indices], rotation=90)
plt.xlabel("Características")
plt.ylabel("Importancia")
plt.tight_layout()
plt.show()
```



SUPPORT VECTOR MACHINE (SVM)

```
from sklearn.svm import SVR
import numpy as np

# Entrenar el modelo de Máquinas de Vectores de Soporte (SVM)
svm_model = SVR(kernel='linear')  # Especifica el kernel lineal para SVM
svm_model.fit(X_train_std, y_train)

# Realizar predicciones en el conjunto de prueba con el modelo SVM
y_pred_svm = svm_model.predict(X_test_std)

# Calcular métricas de evaluación para SVM
mae_svm = mean_absolute_error(y_test, y_pred_svm)
mse_svm = mean_squared_error(y_test, y_pred_svm)
r2_svm = r2_score(y_test, y_pred_svm)
absolute_errors = np.abs(y_test - y_pred_svm)
mape_svm = np.mean(absolute_errors / y_test) * 100
mpe_svm = np.mean((y_test - y_pred_svm) / y_test) * 100

print("Métricas de evaluación para SVM:")
print("Coeficiente de determinación (R^2):", r2_svm)
print("Métricas adicionales para SVM:")
print("Error porcentual absoluto medio (MAPE):", mape_svm)
print("Error medio porcentual (MPE):", mpe_svm)

from scipy.stats import pearsonr
pearson_coefficient, _ = pearsonr(y_test, y_pred_svm)
print("Coeficiente de correlación de Pearson (r):", pearson_coefficient)
```

```
Métricas de evaluación para SVM:
Coeficiente de determinación (R^2): -0.03676727855269579
```