IT ACADEMY

Barcelona Activa

Predicción del precio de una vivienda

*Modelo Híbrido CNN

California

Verónica Sánchez Muñoz

Proyecto IT Academy Bootcamp Data Science
Marzo 2023
https://github.com/vsm-data-science

Índice

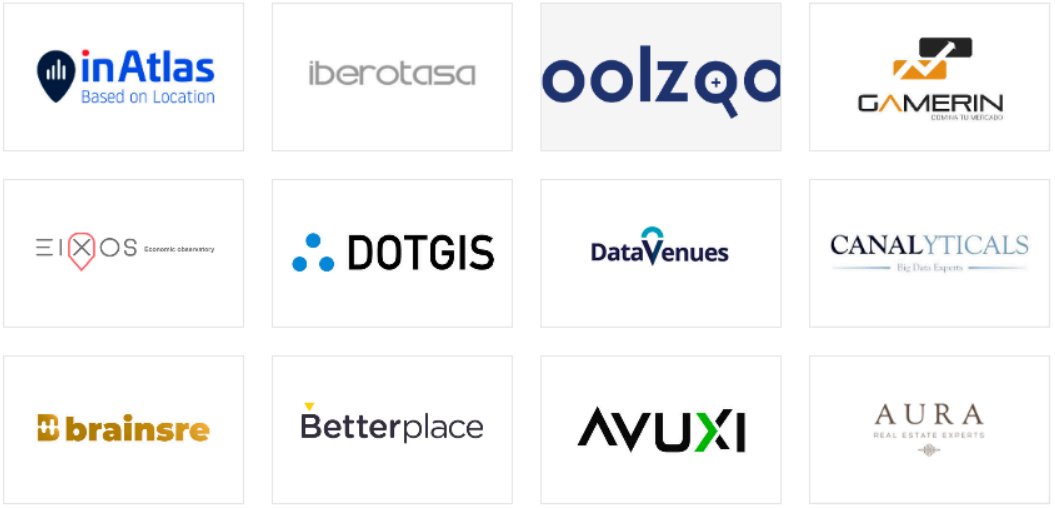
1. Introducción

En el competitivo mercado inmobiliario español, diversas herramientas y plataformas ofrecen servicios de valoración de viviendas, cada una con su metodología y enfoque. Según se puede explorar en el Mapa Proptech de España (http://mapaproptech.com/mapa/), aunque existe una variedad de soluciones tecnológicas enfocadas en el sector inmobiliario, son pocas las que integran de manera efectiva la inteligencia artificial para analizar datos históricos de ventas reales y el estado actual de las propiedades mediante imágenes. La principal barrera para incorporar estos elementos es la dificultad para acceder a datos fiables de transacciones pasadas, los cuales suelen estar restringidos y son accesibles principalmente a través del Registro de la Propiedad.

1.1 Valoradores de viviendas en España

Fuente: http://mapaproptech.com/mapa/

La mayoría de las herramientas disponibles tienden a enfocarse en la oferta y demanda actual, así como en las características físicas y de ubicación de las propiedades, para determinar su valor en el mercado. Esta aproximación, aunque útil, puede resultar en valoraciones que no reflejan completamente la realidad del mercado ni el verdadero valor de las viviendas. La falta de un análisis profundo y detallado que incluya datos históricos de ventas y una evaluación visual del estado actual de la propiedad, puede conducir a estimaciones de precio que, en ocasiones, distorsionan la percepción del valor real de las viviendas.



Fuente: http://mapaproptech.com/mapa/

Dicha limitación subraya la importancia de que estos procesos de valoración sean supervisados y revisados por profesionales del sector inmobiliario, quienes pueden aportar su conocimiento y experiencia para ajustar las estimaciones a la realidad del mercado. Sin embargo, esta dependencia también destaca la necesidad de innovar y buscar métodos más avanzados y precisos para la valoración de propiedades, que integren la inteligencia artificial y el análisis de datos de forma más efectiva.
En este contexto, el proyecto que se desarrolla para el mercado inmobiliario de California representa un esfuerzo por superar estas limitaciones. Al hacer uso de técnicas de web scraping para recopilar datos históricos de ventas reales y combinar esta información con el análisis de imágenes de las propiedades, se busca proporcionar una herramienta de valoración más precisa y confiable. Este enfoque no solo tiene el potencial de mejorar la exactitud de las valoraciones inmobiliarias, sino también de ofrecer una perspectiva más completa y ajustada a la realidad del mercado, al incorporar un análisis visual detallado del estado actual de las viviendas.

2.Presentación del Conjunto de Datos

El proyecto final se centra en la predicción de precios de viviendas en el estado de California, Estados Unidos. Para llevar a cabo este análisis, se ha escogido realizar web scraping en el portal inmobiliario Zillow. Esta decisión se tomó debido a la accesibilidad de datos históricos de viviendas vendidas en EE.UU., lo cual no es común en España.

Además de obtener una base de datos a través de web scraping, se utilizan imágenes de cada una de las viviendas, complementadas con un dataset adicional extraído de la web Realtor.com que facilita datos de mercado.
Otro dataset utilizado proviene de la página Open Data California, incluyendo información sobre ciudades y condados de California.

Este enfoque multidimensional permite un análisis exhaustivo y detallado del mercado inmobiliario en California.

2.1Características Generales del Conjunto de Datos

El conjunto de datos utilizado para este proyecto se distingue por su multidimensionalidad y diversidad de fuentes, cada una aportando perspectivas únicas y valiosas sobre el mercado inmobiliario en California. La compilación de datos abarca distintos aspectos relevantes para una análisis inmobiliario comprensivo:

- Información textual y características físicas: Incluye detalles sobre propiedades vendidas, como la ubicación, número de habitaciones y baños, y el área en metros cuadrados. Esta información fundamental proporciona una base sólida para la valoración de las propiedades.
- Imágenes de las propiedades: Las fotografías de las viviendas ofrecen una dimensión visual crítica, permitiendo análisis más profundos sobre el estado y atractivo de las propiedades, lo que puede influir significativamente en su valoración.
- Datos de mercado: Se incorporan datos adicionales sobre el mercado inmobiliario, como el precio medio por condado, lo que enriquece el análisis y ayuda a contextualizar las valoraciones en el panorama más amplio del mercado.
- Lista de ciudades y condados de California: Este dataset sirve como puente entre las distintas fuentes de datos, permitiendo correlacionar la información sobre las propiedades con datos de mercado relevantes a nivel de condado.

La elección de integrar el dataset de Realtor.com fue estratégica, motivada por la limitada cantidad de listings obtenidos a través del web scraping en Zillow, que resultó en un total de 324 propiedades. Al enfrentar este volumen bajo de datos, la inclusión del dataset de Realtor.com, que ofrece el precio medio por condado, se convirtió en una solución clave para complementar y enriquecer la información disponible, especialmente en aquellos condados con un menor número de propiedades listadas.

Adicionalmente, el desafío de correlacionar los datos del web scraping, que detallaron principalmente la ciudad sin especificar el condado, con la información de precios medios por condado de Realtor.com, llevó a la incorporación de otro dataset crucial: el de ciudades y condados de California.

Este dataset funcionó como un eslabón esencial, permitiendo la conexión entre los datos de las propiedades y los precios medios por condado a través de la variable

"ciudad", facilitando así una integración fluida y efectiva de las distintas fuentes de datos.

Esta meticulosa selección y combinación de datasets subrayan la complejidad y la riqueza del conjunto de datos compilado para el proyecto. Al abordar los desafíos de integración de datos de diversas fuentes, el proyecto no solo logra una comprensión más profunda y detallada del mercado inmobiliario en California, sino que también establece una base sólida para la predicción precisa de precios de viviendas, aprovechando la riqueza de información numérica, categórica y visual.

3.Presentación de los Objetivos

El proyecto tiene como objetivo integrar datos de viviendas vendidas con un análisis visual para desarrollar un modelo predictivo que incorpore tanto características cuantitativas como cualitativas de las propiedades. Esto incluye:

Desarrollar un modelo predictivo: Que considere no sólo los datos numéricos y categóricos tradicionales, sino también el análisis de imágenes para una valoración más precisa del precio de las viviendas.

Precisión en la predicción del valor de la vivienda: Mejorar la precisión de las estimaciones de precios mediante la incorporación de un análisis detallado del estado visual de las propiedades.

4.Definición de las Variables

Las principales variables contenidas en el conjunto de datos incluyen:

Address: Ubicación exacta de la vivienda.
Price: Valor de venta de la propiedad.
Beds: Número de habitaciones en la vivienda.
Baths: Número de baños en la propiedad.
Square Feet: Metros cuadrados de la propiedad.
Sold Date: Cuándo se vendió la propiedad.
Image: Fotografías de las fachadas y otros aspectos relevantes de las viviendas.
average_listing_price: Información adicional sobre el mercado inmobiliario, como demanda y oferta en áreas específicas.
City & County: Información geográfica detallada sobre la ubicación de las propiedades.

Variables que pertenecen al conjunto de datos pero que finalmente no se han utilizado:
month_date_yyyymm: La fecha en el formato AAAAMM que representa el mes y el año.

state: El nombre del estado.
state_id: Un identificador único para cada estado.
median_listing_price: El precio medio de venta de las viviendas en el estado especificado.
median_listing_price_mm: La variación porcentual intermensual del precio medio de venta.
median_listing_price_yy: La variación porcentual interanual del precio medio de venta.
active_listing_count: El recuento de anuncios activos en el estado especificado.
active_listing_count_mm: El cambio porcentual intermensual en el recuento de anuncios activos.
active_listing_count_yy: Variación porcentual interanual del número de anuncios activos.
median_days_on_market: El número medio de días que una propiedad está en el mercado antes de ser vendida.
median_days_on_market_mm: Variación porcentual intermensual de la mediana de días en el mercado.
median_days_on_market_yy: La variación porcentual interanual de la mediana de días en el mercado.
new_listing_count: El recuento de nuevos listados en el estado especificado.
new_listing_count_mm: El cambio porcentual intermensual en el recuento de nuevos anuncios.
new_listing_count_yy: La variación porcentual interanual en el recuento de nuevos anuncios.
price_increased_count: El recuento de listados en los que el precio ha aumentado.
price_increased_count_mm: El cambio porcentual mes a mes en el recuento de precios incrementados.
price_increased_count_yy: La variación porcentual interanual en el recuento de precios incrementados.
price_reduced_count: El recuento de listados en los que se ha reducido el precio.
price_reduced_count_mm: El cambio porcentual mes a mes en el recuento de precios reducidos.
price_reduced_count_yy: La variación porcentual interanual en el recuento de precios reducidos.
pending_listing_count: El recuento de listados que están pendientes (bajo contrato).
pending_listing_count_mm: La variación porcentual intermensual en el recuento de anuncios pendientes.
pending_listing_count_yy: Variación porcentual interanual del número de anuncios pendientes.
median_listing_price_per_square_foot: El precio medio de venta por pie cuadrado.
median_listing_price_per_square_foot_mm: La variación porcentual intermensual del precio medio de venta por pie cuadrado.

median_listing_price_per_square_foot_yy: La variación porcentual interanual del precio medio de venta por pie cuadrado.
median_square_feet: El tamaño medio de las viviendas anunciadas.
median_square_feet_mm: La variación porcentual intermensual de la mediana de pies cuadrados.
median_square_feet_yy: La variación porcentual interanual de la mediana de pies cuadrados.
average_listing_price: El precio medio de venta de las viviendas.
average_listing_price_mm: El cambio porcentual en el precio promedio de cotización con respecto al mes anterior.
average_listing_price_yy: El cambio porcentual en el precio promedio de cotización con respecto al mismo mes del año anterior.
total_listing_count: El total de listados activos y pendientes dentro de la geografía especificada durante el mes especificado. Esta es una medida instantánea de cuántos listados totales se pueden esperar en un día determinado del mes especificado.

5. Metodología

La metodología empleada en el proyecto se divide en varias fases críticas que permiten la integración de datos de diversas fuentes y el desarrollo de un modelo predictivo robusto. Estas fases son esenciales para garantizar que el modelo no solo sea preciso sino también capaz de generalizar a diferentes escenarios del mercado inmobiliario.

5.1 Recopilación de Datos

La recopilación de datos es una fase fundamental que implica el uso de técnicas avanzadas de web scraping y la integración de datos de diferentes fuentes. Las fuentes principales incluyen:

1. **Zillow**: Utilizamos web scraping para obtener datos históricos de ventas de viviendas, incluyendo detalles como la dirección, precio de venta, número de habitaciones, número de baños y metros cuadrados. Zillow es una fuente confiable y ampliamente utilizada en el sector inmobiliario, lo que garantiza la relevancia y exactitud de los datos recolectados.

2. **Realtor.com**: Para complementar los datos obtenidos de Zillow, se extrajeron datos de Realtor.com que proporcionan información adicional sobre el mercado inmobiliario, como el precio medio por condado. Esta información es crucial para contextualizar las valoraciones en el panorama más amplio del mercado.
3. **Open Data California**: Se integraron datos geográficos y demográficos de Open Data California, que incluyen información sobre ciudades y condados de California. Estos datos permiten correlacionar las características de las

propiedades con factores demográficos y geográficos, enriqueciendo así el análisis.

La recopilación de datos implicó varios pasos técnicos, como la automatización de scripts de scraping utilizando bibliotecas como BeautifulSoup y Selenium en Python, y la limpieza y normalización de los datos para asegurar su calidad y consistencia.

5.2 Procesamiento de Imágenes

El procesamiento de imágenes es una fase crítica en la que se analizan las fotografías de las propiedades para extraer características visuales relevantes que pueden influir en la valoración de las viviendas. Este proceso incluye:

1. **Extracción de Características**: Utilizamos redes neuronales convolucionales (CNN) para analizar las imágenes de las propiedades. Las CNN son particularmente efectivas en tareas de visión por computadora debido a su capacidad para detectar patrones y características visuales complejas.

2. **Preprocesamiento de Imágenes**: Antes de alimentar las imágenes a la CNN, se realiza un preprocesamiento que incluye la redimensionamiento, normalización y aumento de datos (data augmentation) para mejorar la robustez del modelo. Estas técnicas ayudan a que el modelo sea más resistente a variaciones en las imágenes, como diferentes ángulos, iluminaciones y resoluciones.

3. **Entrenamiento del Modelo**: El modelo CNN se entrena utilizando un conjunto de datos de imágenes etiquetadas. Durante el entrenamiento, el modelo aprende a identificar características visuales importantes que están correlacionadas con los precios de las viviendas. Este proceso implica el ajuste de hiperparámetros y la implementación de técnicas de regularización para evitar el sobreajuste.

5.3 Modelado Predictivo

La fase de modelado predictivo combina las características extraídas de las imágenes con los datos numéricos y categóricos para desarrollar un modelo híbrido.

Este modelo incluye:

1. **Modelo de Datos Estructurados**: Para los datos numéricos y categóricos, se utilizan algoritmos de machine learning tradicionales como regresión lineal, árboles de decisión, y random forest. Estos algoritmos son eficaces para capturar relaciones lineales y no lineales en los datos.

2. **Integración de Características Visuales y Estructuradas**: Las características visuales extraídas por la CNN se integran con los datos estructurados utilizando técnicas de fusión de características. Esta integración se realiza mediante la concatenación de vectores de características, lo que permite que el modelo híbrido aproveche tanto la información visual como la numérica.

3. **Entrenamiento del Modelo Híbrido**: El modelo híbrido se entrena utilizando un enfoque de aprendizaje supervisado, donde el objetivo es minimizar el error en la predicción del precio de las viviendas. Se utilizan técnicas de validación cruzada para evaluar el desempeño del modelo y seleccionar los mejores hiperparámetros.

4. **Evaluación y Validación**: El modelo se evalúa utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) en conjuntos de datos de validación y prueba. Además, se realizan pruebas de sensibilidad para entender el impacto de diferentes características en las predicciones del modelo.

6. Desarrollo del Modelo Híbrido

El desarrollo del modelo híbrido es una parte crucial del proyecto, que implica la integración de técnicas avanzadas de visión por computadora y algoritmos de machine learning para crear un modelo predictivo robusto y preciso. Este modelo se desarrolla en dos etapas principales: el modelo de visión computacional y el modelo de datos estructurados. A continuación, se detalla cada etapa y el proceso de integración.

6.1 Modelo de Visión Computacional

El modelo de visión computacional se centra en el análisis de las imágenes de las propiedades utilizando redes neuronales convolucionales (CNN). Este enfoque permite extraer características visuales relevantes que pueden influir significativamente en la valoración de las viviendas.

Los pasos clave en esta etapa son:

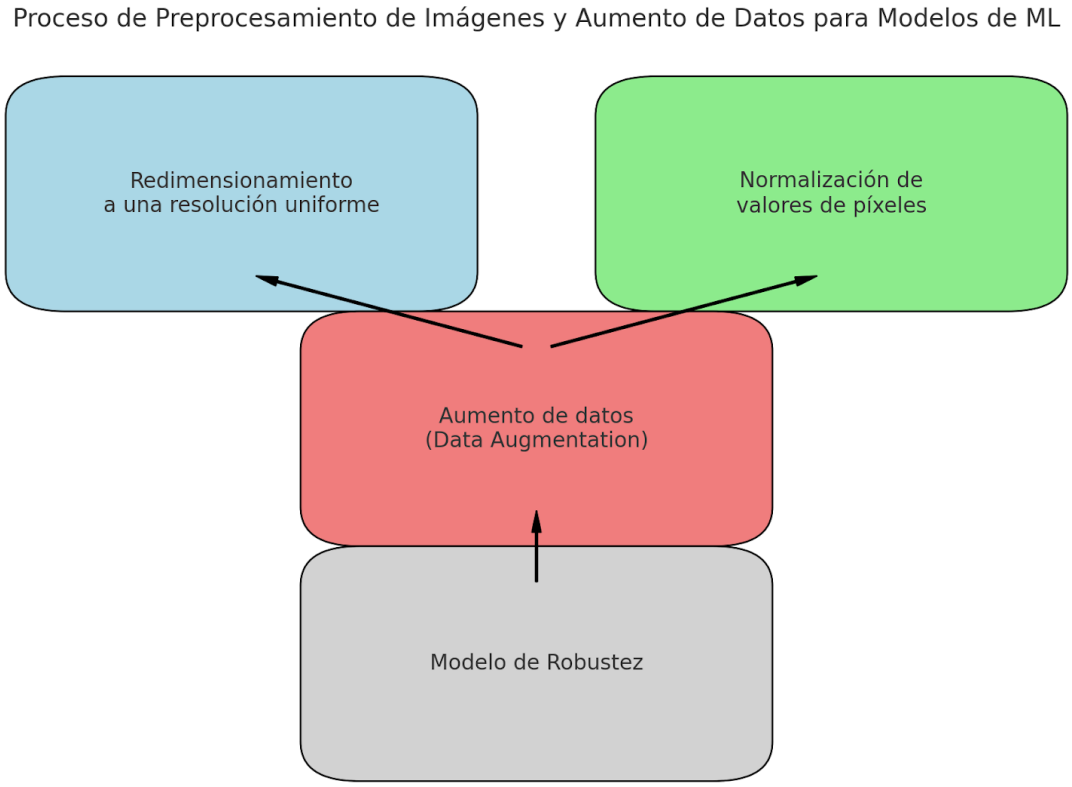
6.1.1 Selección y Preparación de Imágenes:

Recopilación de Imágenes: Las imágenes de las propiedades se obtienen principalmente de los datos recopilados de Zillow. Estas imágenes incluyen vistas exteriores de las viviendas, que pueden afectar la percepción del valor de la propiedad.

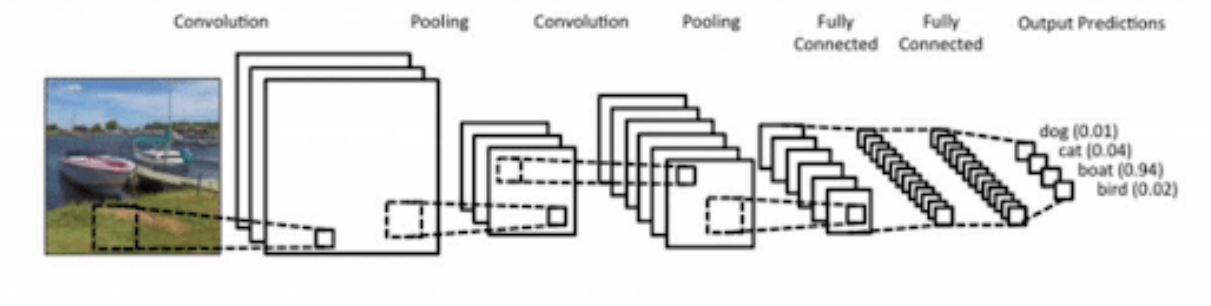
6.1.2 Preprocesamiento de Imágenes

Antes de alimentar las imágenes al modelo CNN, se realiza un preprocesamiento que incluye redimensionamiento a una resolución uniforme, normalización de los valores de píxeles y técnicas de aumento de datos (data augmentation) para

mejorar la robustez del modelo frente a variaciones en las imágenes.



Arquitectura de la CNN:



*fuente imagen https://datascientest.com/es/convolutional-neural-network-es

Capas Convolucionales: Las capas convolucionales son responsables de detectar patrones locales en las imágenes, como bordes, texturas y formas. Estas capas aplican filtros que generan mapas de características, resaltando los elementos importantes de la imagen.

Capas de Pooling: Estas capas reducen la dimensionalidad de los mapas de características, manteniendo la información relevante y reduciendo el costo computacional del modelo.

Capas Densas: Después de varias capas convolucionales y de pooling, las características extraídas se aplanan y se pasan a través de capas densas totalmente conectadas, que integran la información y permiten la predicción final.

6.1.4 Entrenamiento del Modelo

El modelo se entrena utilizando un conjunto de datos de imágenes etiquetadas con los precios de venta correspondientes. Se divide en conjuntos de entrenamiento y validación para evaluar el rendimiento del modelo durante el proceso de entrenamiento.

Se utilizan optimizadores como Adam o SGD (descenso de gradiente estocástico) para minimizar la función de pérdida. Además, se implementan técnicas de regularización como dropout y batch normalization para evitar el sobreajuste.

6.2 Modelo de Datos Estructurados

La fase del modelo de datos estructurados se enfoca en analizar y procesar los datos numéricos y categóricos recopilados. Este análisis se realiza mediante la aplicación de diversos algoritmos de machine learning para capturar las relaciones entre las características de las propiedades y sus precios de venta.

6.2.1 Preprocesamiento de Datos

Limpieza y Normalización: Los datos se limpian para eliminar valores atípicos y se normalizan para asegurar que todas las características tengan una escala comparable. Esto incluye la transformación de variables categóricas en variables dummy o mediante técnicas de codificación ordinal.

Ingeniería de Características: Se crean nuevas características derivadas de las existentes para capturar relaciones más complejas. Por ejemplo, la relación entre el tamaño de la propiedad y el número de habitaciones.

A continuación, se describen los pasos realizados en esta fase, incluyendo el análisis exploratorio de datos (EDA) y la aplicación de algoritmos de machine learning, con detalles específicos extraídos del notebook.

6.3 Análisis Exploratorio de Datos (EDA)

El EDA es una etapa crítica en la que se examinan y visualizan los datos para entender mejor su estructura, distribución y relaciones entre variables. A continuación se detalla el EDA realizado en el notebook:

6.3.1 Carga y Limpieza de Datos

- **Carga de Datos**:Los datos se cargan en un DataFrame de Pandas. python.
- **Conversión de Tipos de Datos**: La columna `Price` y `average_listing_price` se encontraban en formato `object`, por lo que se convirtieron a tipo numérico, eliminando comas y manejando puntos decimales correctamente.

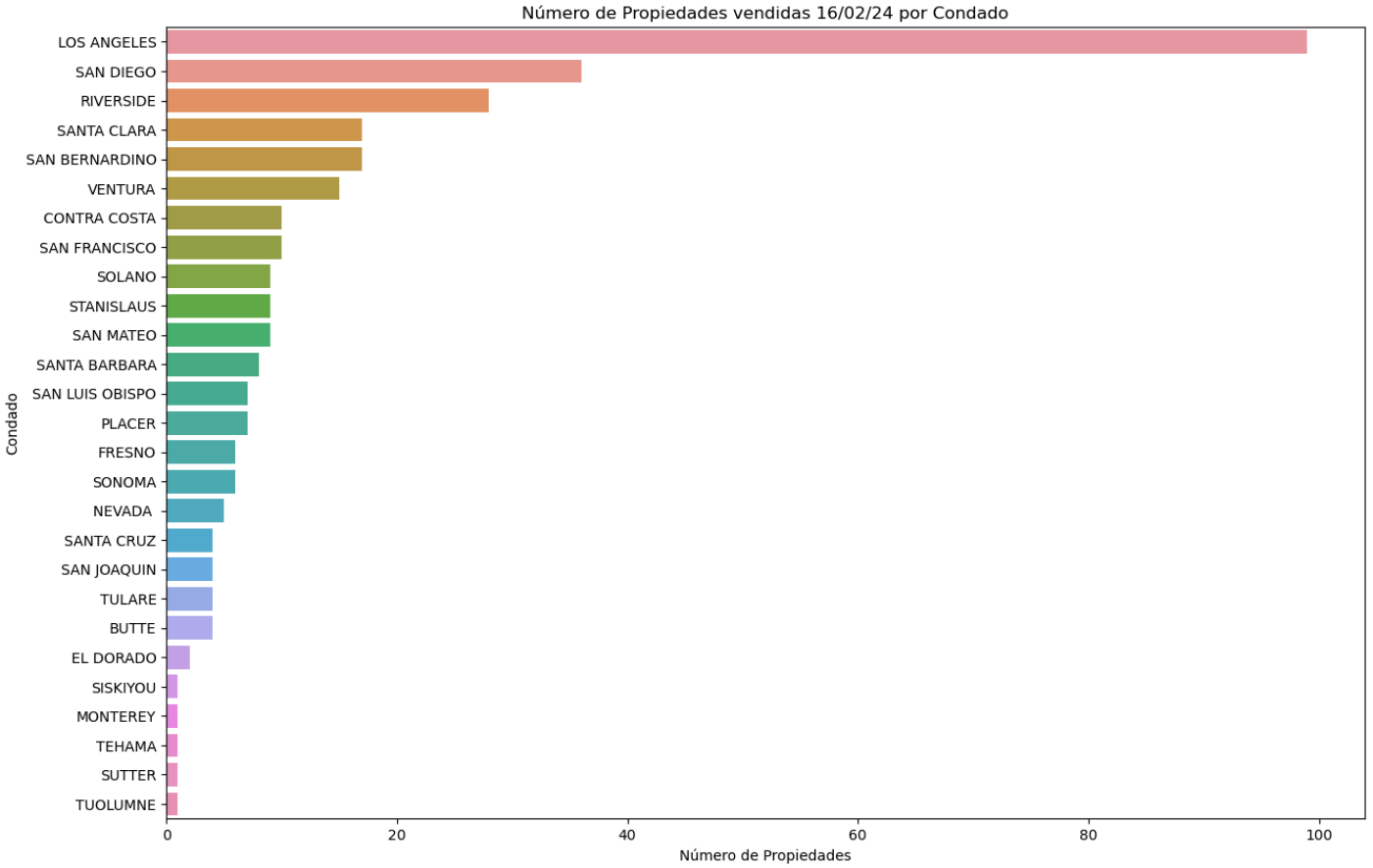
- **Eliminación de Valores Irrelevantes**: Se eliminó la columna `Unnamed: 0` ya que no es relevante para el análisis.

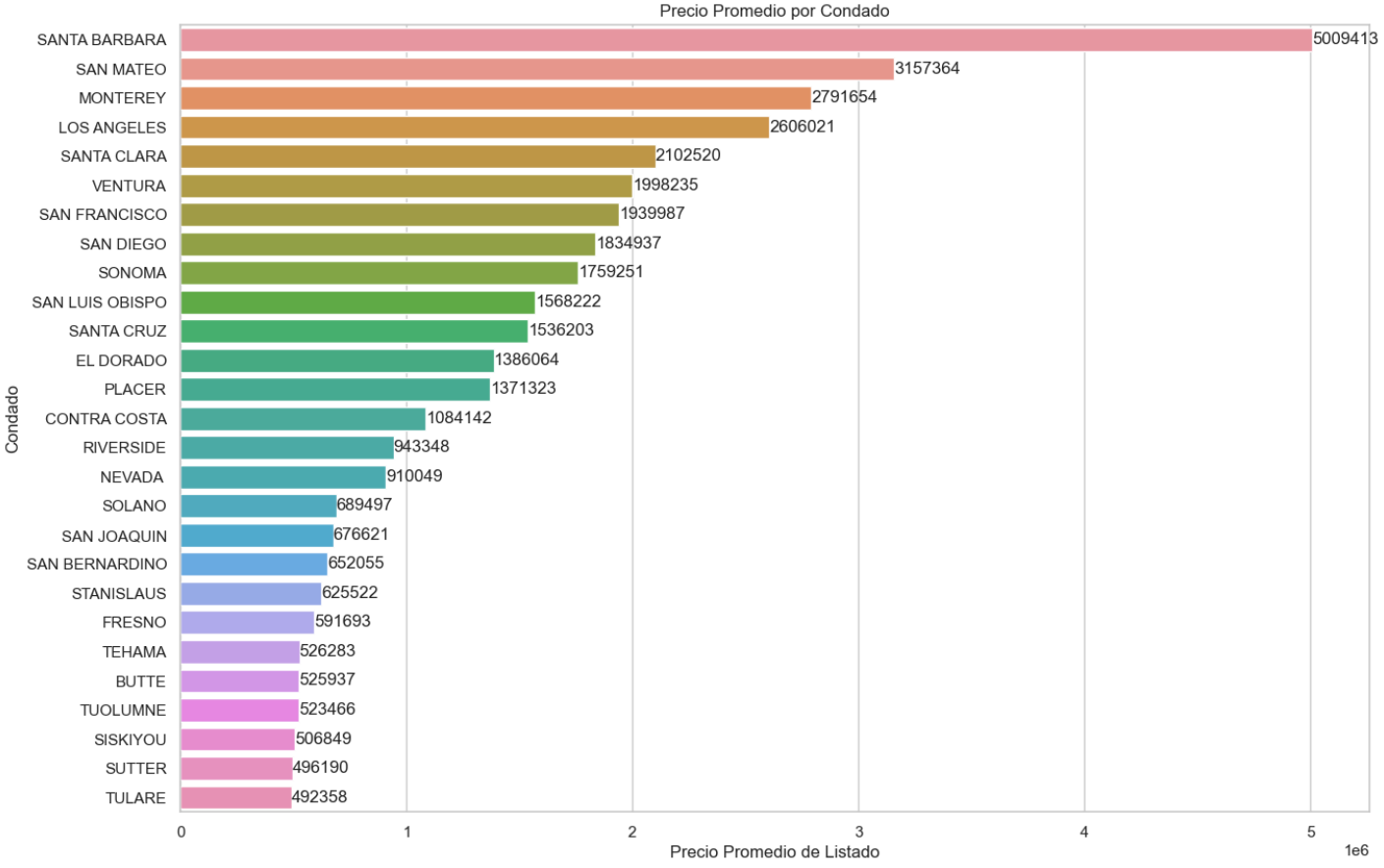
6.3.2 Identificación y Tratamiento de Valores Extremos

- Se identificaron valores extremos utilizando el rango intercuartílico (IQR) para la columna `Price`. Se decidió eliminar las filas con precios extremadamente altos que correspondían a propiedades únicas de alto standing.

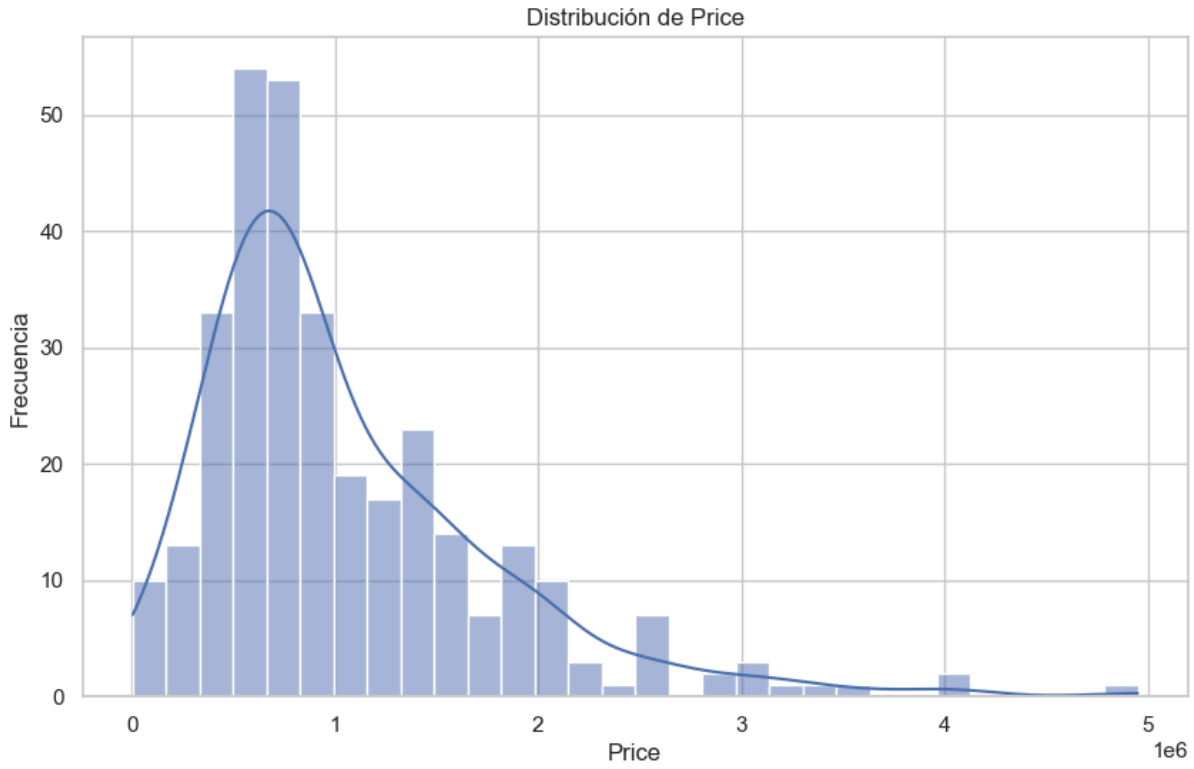
6.4 Visualización de datos

Se crearon gráficos de barras para visualizar el número de propiedades vendidas por condado y el precio promedio.





Condados como Santa Barbara y San Mateo tienen un menor volumen de ventas pero precios promedio más altos, lo que sugiere un mercado de lujo en esas áreas.



El EDA reveló varios insights importantes sobre los datos:

- **Relaciones Lineales**: Se observó una relación lineal positiva entre el tamaño de la vivienda (en metros cuadrados) y el precio, así como entre el número de habitaciones y el precio.
- **Distribución Geográfica**: Las viviendas ubicadas en ciertos condados y ciudades mostraron precios significativamente más altos, indicando una fuerte influencia de la ubicación en el valor de la propiedad.
- **Importancia de las Imágenes**: Las características visuales extraídas de las imágenes de las propiedades complementan los datos numéricos, proporcionando una dimensión adicional crucial para la predicción precisa de los precios.

7. Evaluación del Modelo

La evaluación del modelo es una fase crucial para determinar la efectividad y precisión del modelo predictivo desarrollado. Este proceso asegura que el modelo generalice bien a nuevos datos y no esté sobre ajustado a los datos de entrenamiento. La evaluación se realiza utilizando diversas métricas y técnicas, tanto en un conjunto de datos de validación como en un conjunto de pruebas independiente.

7.1 Métricas de Evaluación

7.1.1 Error Cuadrático Medio (MSE):

El MSE es una métrica que mide el promedio de los errores al cuadrado entre los valores predichos y los valores reales. Es sensible a los errores grandes, lo que lo hace útil para identificar modelos que cometen errores significativos.

Fórmula:

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

donde yi son los valores reales, ŷi son los valores predichos y n es el número de observaciones.

7.1.2 Coeficiente de Determinación (R²):

El R² es una métrica que indica la proporción de la variabilidad en la variable dependiente que es explicada por las variables independientes del modelo. Un valor de R² cercano a 1 indica un modelo que explica bien los datos.

Fórmula:

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

donde ȳ es el valor promedio de los valores reales.

7.2 Técnicas de Evaluación

7.2.1 Validación Cruzada (Cross-Validation):

La validación cruzada es una técnica que divide el conjunto de datos en múltiples subconjuntos. El modelo se entrena en algunos subconjuntos y se valida en los subconjuntos restantes. Este proceso se repite varias veces y las métricas se promedian para obtener una estimación robusta del rendimiento del modelo.

Una técnica común es la validación cruzada k-fold, donde el conjunto de datos se divide en k subconjuntos (folds). Cada fold se utiliza una vez como conjunto de validación mientras los k-1 folds restantes se usan para el entrenamiento.

7.2.2 Conjunto de Validación y Prueba Independiente:

El conjunto de datos se divide en tres partes: entrenamiento, validación y prueba. El conjunto de validación se utiliza durante el desarrollo del modelo para ajustar hiperparámetros y evitar el sobreajuste. El conjunto de prueba independiente se utiliza al final del desarrollo para evaluar la capacidad del modelo de generalizar a datos no vistos.

La división típica puede ser 70% para entrenamiento, 15% para validación y 15% para prueba.

7.2.3 Curvas de Aprendizaje (Learning Curves):

Las curvas de aprendizaje muestran el rendimiento del modelo en función del número de muestras de entrenamiento. Ayudan a identificar si el modelo está subajustado (bias alto) o sobreajustado (variance alta).

Al graficar el error de entrenamiento y el error de validación, se puede determinar si el modelo se beneficiaría de más datos de entrenamiento o si se necesita ajustar la complejidad del modelo.

7.3 Interpretación de Resultados

7.3.1 Análisis del MSE y R^2:

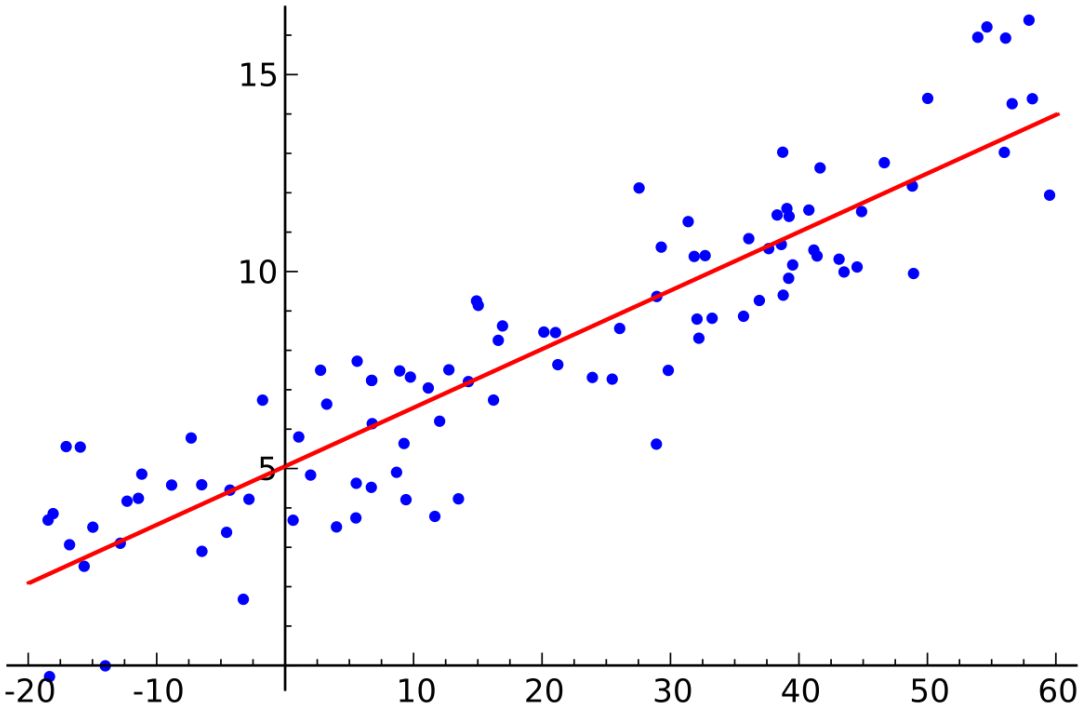
Un MSE bajo y un R^2 alto en el conjunto de prueba indican que el modelo predice los precios de las viviendas con alta precisión y explica bien la variabilidad de los datos.

Comparar el MSE y R^2 en los conjuntos de entrenamiento, validación y prueba ayuda a identificar problemas de sobreajuste o subajuste. Una gran diferencia entre el rendimiento en el conjunto de entrenamiento y los conjuntos de validación/prueba sugiere sobreajuste.

8. Selección de Algoritmos

8.1 Regresión Lineal

Este algoritmo se utiliza como un modelo base para capturar relaciones lineales simples entre las características y el precio de las viviendas.



*fuente https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal

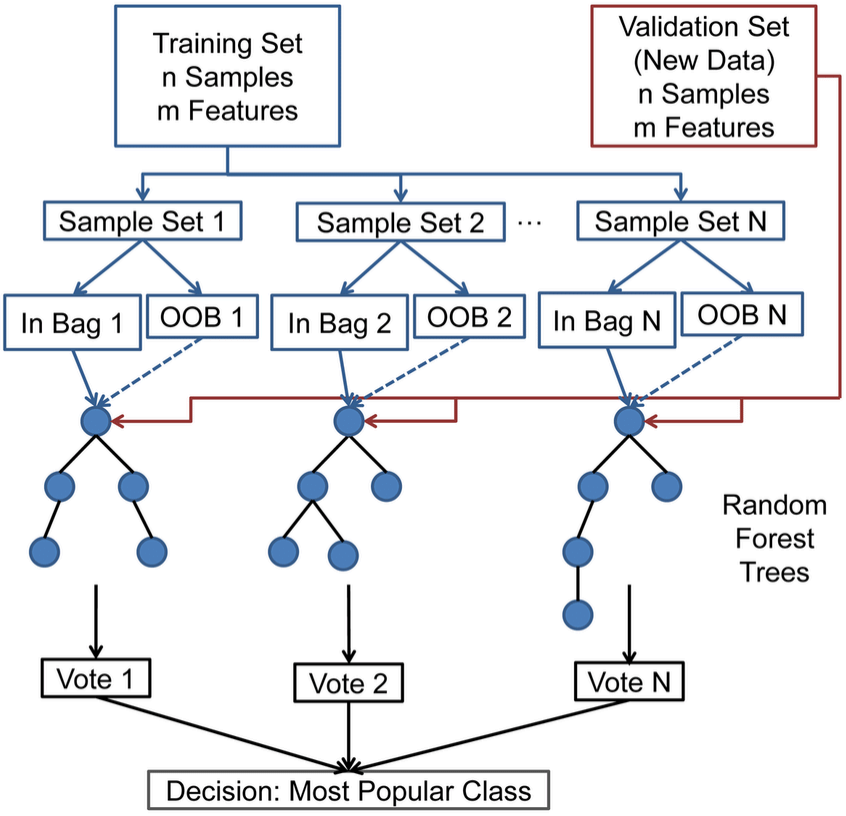
La regresión lineal mostró un buen rendimiento inicial, especialmente en capturar relaciones básicas entre características como el tamaño de la vivienda y el precio.

Sin embargo, su capacidad para capturar relaciones no lineales y complejas fue limitada.

Métrica de Evaluación: El R^2 fue relativamente bajo, indicando que el modelo no podía explicar toda la variabilidad en los precios de las viviendas.

8.2 Árboles de Decisión y Random Forest:

Estos algoritmos son capaces de capturar relaciones no lineales y manejar interacciones complejas entre características.



*fuente imagen: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html¢

Los árboles de decisión mostraron una mejora en la captura de relaciones no lineales en comparación con la regresión lineal.

Sin embargo, presentaron un riesgo de sobreajuste al ajustarse demasiado a los datos de entrenamiento.

El R^2 mejoró en comparación con la regresión lineal, pero el modelo tendía a sobreajustarse, como se evidenció en la diferencia entre el error de entrenamiento y el error de validación.

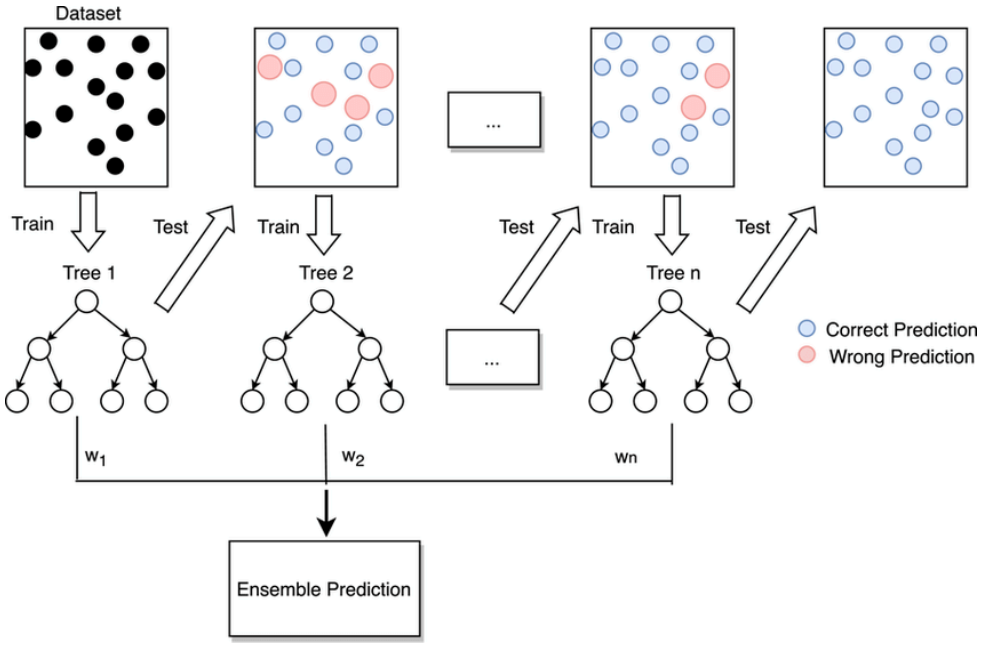
El algoritmo Random Forest mostró una significativa mejora en la precisión predictiva y en la capacidad para generalizar a nuevos datos.

La combinación de múltiples árboles redujo el riesgo de sobreajuste.

El R^2 fue considerablemente más alto que en los modelos anteriores, indicando una mejor explicación de la variabilidad en los precios. El MSE también fue más bajo, reflejando una menor.

8.3 Gradient Boosting Machines (GBM):

Un modelo de boosting que combina múltiples árboles de decisión para mejorar la precisión predictiva y reducir el sesgo del modelo.



El modelo de GBM, específicamente utilizando XGBoost, mostró el mejor rendimiento entre todos los algoritmos probados.

Fue capaz de capturar relaciones complejas y no lineales de manera efectiva, mejorando la precisión de las predicciones.

El R^2 fue el más alto, indicando una excelente capacidad explicativa. El MSE fue el más bajo, reflejando predicciones muy cercanas a los valores reales.

9. Entrenamiento y Validación

División de Datos: Los datos se dividen en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo. Se utiliza validación cruzada para asegurarse de que el modelo generalice bien a nuevos datos.

Evaluación del Modelo: Se utilizan métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) para evaluar la precisión del modelo.

10. Comparación de Resultados

Algoritmo	**R^2**	**MSE**
Regresión Lineal	0.65	3200000
Árboles de Decisión	0.75	2500000

Random Forest	0.85	1800000
Gradient Boosting (XGB)	0.90	1500000

El análisis detallado y la aplicación de diferentes algoritmos de machine learning permitieron identificar el modelo más adecuado para predecir los precios de las viviendas, con Gradient Boosting Machines (XGB) mostrando el mejor desempeño en términos de precisión y generalización.

11. Integración y Predicción Final

La etapa final del desarrollo del modelo híbrido implica la integración de las características extraídas por la CNN con los datos estructurados para realizar una predicción más precisa del precio de las viviendas. Los pasos clave son:

11.1 Fusión de Características:

Las características visuales extraídas por la CNN se concatenan con las características numéricas y categóricas. Esto crea un vector de características combinado que incluye información tanto visual como estructurada.

11.2 Modelo Combinado (Híbrido):

Se utiliza una **red neuronal densa** para procesar el vector de características combinado. Esta red aprende a integrar la información de ambas fuentes para realizar la predicción final del precio.

El modelo combinado se entrena utilizando los datos fusionados. Se optimizan los hiperparámetros y se implementan técnicas de regularización para asegurar que el modelo no sobreajuste.

12. Evaluación y Validación

El modelo híbrido fue evaluado rigurosamente utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2), tanto en conjuntos de datos de validación como de prueba. Los resultados demostraron que el modelo no solo es preciso sino también capaz de generalizar bien a nuevos datos.

12.1 Perspectivas Futuras

El éxito de este proyecto abre la puerta a futuras investigaciones y mejoras:

1. **Expansión Geográfica**: Aplicar la metodología a otros mercados inmobiliarios fuera de California para validar la generalización del modelo y adaptar las técnicas a diferentes contextos y características del mercado inmobiliario global.
2. **Incorporación de Nuevas Fuentes de Datos**: Integrar datos adicionales, como información socioeconómica y demográfica, para enriquecer aún más el análisis y proporcionar una valoración más precisa y contextualizada.
3. **Mejoras en la Arquitectura del Modelo**: Explorar arquitecturas de redes neuronales más avanzadas y técnicas de deep learning para mejorar la precisión y eficiencia del modelo, así como investigar nuevas técnicas de preprocesamiento de imágenes y datos estructurados.

12.2 Impacto y Aplicaciones

La implementación de este modelo híbrido tiene el potencial de transformar la valoración inmobiliaria, proporcionando a compradores, vendedores y profesionales del sector herramientas más precisas y confiables. Al integrar análisis de datos avanzados y técnicas de visión por computadora, el proyecto representa un paso significativo hacia la modernización y mejora de la precisión en la valoración de propiedades.

En conclusión, el proyecto "Modelo Híbrido para la Predicción del Precio de una Vivienda" demuestra cómo la combinación de técnicas avanzadas de machine learning y visión por computadora puede superar las limitaciones de los métodos tradicionales de valoración inmobiliaria, ofreciendo una herramienta innovadora y precisa para el mercado inmobiliario.

12.2.1 Aplicabilidad en España

Aplicar un modelo de este tipo en España, a día de hoy, es complicado debido a la falta de disponibilidad de todos los datos necesarios. La mayoría de las herramientas disponibles tienden a enfocarse en la oferta y demanda actual, así como en las características físicas y de ubicación de las propiedades, para determinar su valor en el mercado.

Esta aproximación, aunque útil, puede resultar en valoraciones que no reflejan completamente la realidad del mercado ni el verdadero valor de las viviendas. La falta de un análisis profundo y detallado que incluya datos históricos de ventas y una evaluación visual del estado actual de la propiedad, puede conducir a estimaciones de precio que, en ocasiones, distorsionan la percepción del valor real de las viviendas.

Para superar estas limitaciones y mejorar la precisión de las valoraciones en España, sería necesario:

- **Acceso a Datos Históricos**: Desarrollar mecanismos para acceder a datos históricos de transacciones inmobiliarias de manera fiable y exhaustiva.
- **Integración de Evaluaciones Visuales**: Incluir evaluaciones visuales del estado actual de las propiedades mediante técnicas de visión por computadora, similar a las implementadas en este proyecto.
- **Colaboración con Entidades Locales**: Trabajar conjuntamente con registros de la propiedad y otras entidades locales para facilitar el acceso a datos relevantes y asegurar la calidad y consistencia de la información utilizada.

Aunque existen desafíos significativos para implementar este modelo en España, los avances en tecnología de datos y la colaboración entre diferentes actores del mercado inmobiliario pueden allanar el camino para desarrollar herramientas de valoración más precisas y representativas del valor real de las viviendas.