HOUSE PRICE PREDICTION

PROYECTO FINAL

INTRODUCCIÓN A LA CIENCIA DE DATOS

Valentina Plada, Gonzalo Nieto, Guillermo Pons y Agustina Soto

Índice

Definición de pregunta

Este proyecto se centra en el desarrollo e implementación de un modelo de regresión para estimar el valor de viviendas utilizando el House Price Regression Dataset (Home Value Insights). A partir de técnicas avanzadas de analítica de datos y aprendizaje automático, el modelo fue entrenado y evaluado para predecir precios de forma precisa y consistente, integrando características clave de las propiedades y de su entorno. El objetivo principal es optimizar la estimación del valor de mercado, reduciendo al mínimo el error de predicción y proporcionando una herramienta confiable para la toma de decisiones. La solución resultante es robusta, escalable y apta para integrarse en procesos reales de evaluación de inmuebles.

Descripción de los datos

Visión general de los datos

El conjunto de datos utilizado para este proyecto (house_price_regression_dataset.csv) es un dataset, diseñado para un problema de regresión supervisada. Contiene 1.046 registros (filas) y 8 variables (columnas).

El objetivo principal es predecir la variable dependiente *House_Price* (precio de la vivienda) utilizando las 7 variables predictoras proporcionadas:

1. *Square_Footage*: Pies cuadrados de la vivienda.
2. *Num_Bedrooms*: Número de habitaciones.
3. *Num_Bathrooms*: Número de baños.
4. *Year_Built*: Año de construcción.
5. *Lot_Size*: Tamaño del lote en acres.
6. *Garage_Size*: Capacidad del garaje (en número de autos).
7. *Neighborhood_Quality*: Calificación de la calidad del barrio (escala 1-10).

Tablas estadísticas

Las variables se dividen en numéricas (continuas o discretas) y categóricas (ordinales) para un mejor análisis.

Tabla de variables numéricas continuas

Esta tabla resume las principales estadísticas descriptivas para las variables continuas y la variable objetivo.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
House_Price	1046	640.310,94	501.130,96	111.626,85	405.343,7	635.825,1	831.929	11.000.000
Square_Footage	1044	2832.54	1289.006	0	1755.5	2888.5	3868.25	9000

Lot_Size	1046	2.81	1.37	0.50	1.72	2.85	3.95	15
Year_Built	1046	1986.47	20.60	1950	1969	1986	2004	2022

Tabla 1. *Estadísticas descriptivas.*

Tabla de variables categóricas y numéricas discretas

Esta tabla muestra los valores únicos para las variables que, aunque almacenadas como números, representan categorías o conteos discretos.

Variable	Tipo	Valores Únicos
Num_Bedrooms	Discreto	**[1, 2, 3, 4, 5, 10, 15]**
Num_Bathrooms	Discreto	**[1, 2, 3, 6]**
Garage_Size	Discreto	**[0, 1, 2, 10]**
Neighborhood_Quality	Ordinal	**[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]**

Tabla 2. *Descripción variables.*

Distribución de los datos

El análisis exploratorio de datos (EDA) revela patrones críticos sobre las relaciones entre las variables.

Distribución de House Price: El histograma muestra una asimetría a la derecha, típica de precios inmobiliarios. La mayoría de los precios se agrupan entre 100.000 y 400.000, pero existen valores extremadamente altos (por ejemplo, 11.000.000), posiblemente outliers.

Distribución de Square_Footage: La distribución es asimétrica a la derecha, la mayoría de las casas están entre 1500 y 4500 pies cuadrados, pero existen algunos valores extremadamente altos arriba de 6000-8000, lo que estira la cola y eleva la media.

Distribución de Year_Built: La distribución es relativamente uniforme desde 1950 hasta 2020 (no hay un año que predomina). La media es 1986, ubicada en el centro del rango.

Distribución Num_Bathrooms: Distribución multimodal y discreta, se concentran valores en 1,2 y 3 baños. Existen algunos casos extremadamente altos (casas muy grandes), 6 baños, que probablemente sean outliers. La media es 1,98, coherente con la concentración en 2 baños.

Distribución Lot_Size: También es right-skewed, con valores que se concentran entre 0,5 y 5, pero hay casos atípicos arriba de 10-14 acres. La media es 2,81, influida por la presencia de outliers.

Distribución Neighborhood_Quality: Variable discreta entre 1 y 10, muy similar a una distribución uniforme.

Matriz de correlación: En general, la matriz de correlación muestra que la variable que más explica el precio de la vivienda es Square_Footage, con una correlación fuerte de 0.65, seguida por Lot_Size, que presenta una relación moderada de 0.35. El número de baños y el año de construcción se relacionan de forma débil con el precio, mientras que la calidad del barrio prácticamente no muestra relación lineal. Además, las correlaciones entre las variables explicativas son muy bajas, lo que indica ausencia de multicolinealidad.

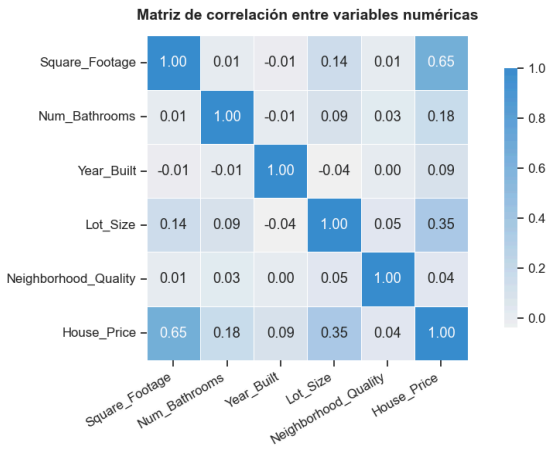

Figura 1. *Matriz de correlación.*

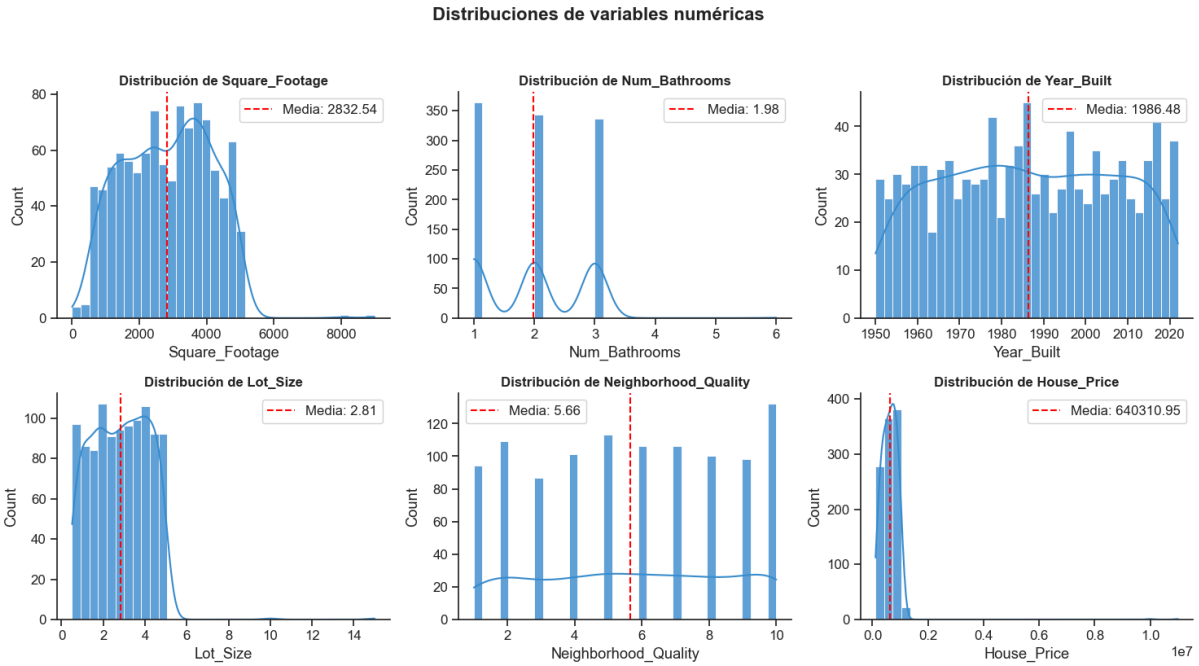

Figura 2. *Distribuciones de variables numéricas.*

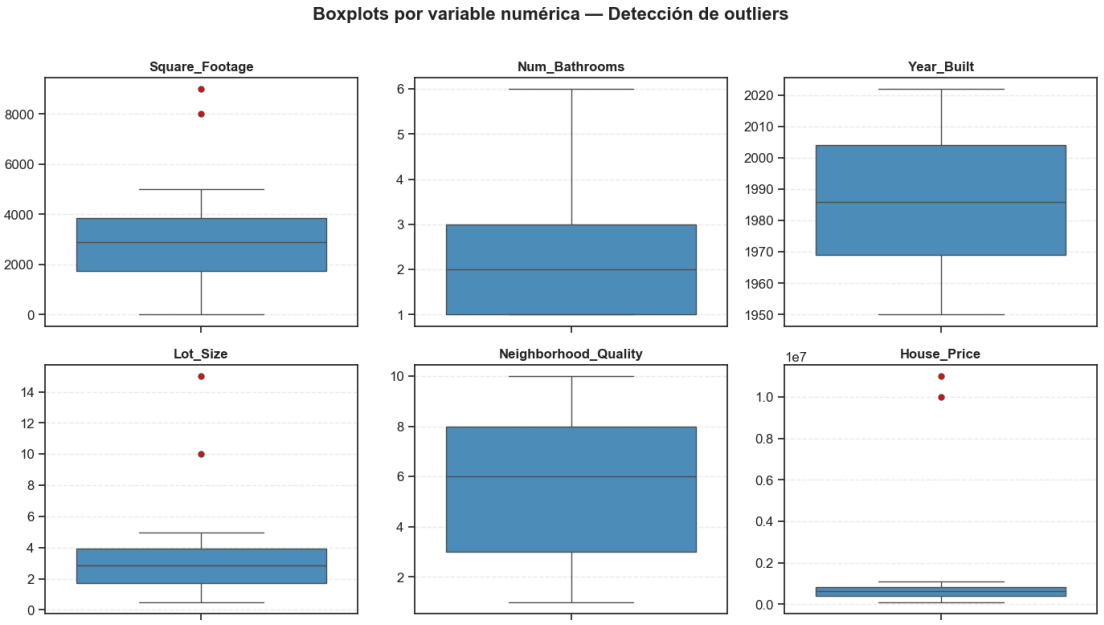

Figura 3. *Boxplots.*

Limpieza de datos

La etapa de limpieza de datos tuvo como objetivo asegurar la calidad, coherencia y trazabilidad del conjunto antes de realizar cualquier análisis descriptivo o inferencial. Para ello, se aplicaron procedimientos estandarizados orientados a: (i) corregir registros redundantes o inconsistentes, (ii) homogeneizar formatos y tipologías de variables, (iii) tratar valores faltantes y atípicos y (iv) documentar cada transformación para garantizar la auditabilidad del trabajo. Estas acciones buscan reducir sesgos, evitar la doble contabilización de evidencia y preservar la validez de los resultados.

Eliminación de duplicados

Por motivos de calidad y validez estadística, se implementó un proceso de deduplicación para identificar y eliminar registros repetidos. La duplicación puede originarse por errores de integración de fuentes, reenvíos del mismo formulario o fallas de captura. Mantener duplicados sobreestima el tamaño muestral efectivo, sesga estadísticas descriptivas, viola el supuesto de independencia entre observaciones y, en contextos de modelado, puede inducir “fugas” de información que inflan artificialmente el desempeño.

Criterio y procedimiento: Se definió “duplicado” como toda coincidencia exacta en las variables clave que identifican de manera unívoca la observación. Bajo esta definición, se detectaron apariciones múltiples y se retuvo únicamente la primera observación, eliminando las restantes. Se efectuaron controles de consistencia para confirmar que los registros removidos no aportaran información diferenciada relevante en campos no clave.

Resultado: Como consecuencia de este proceso, el tamaño del conjunto se redujo de 1.046 a 1.003 observaciones, eliminando 43 registros (≈ 4,11%). Este ajuste no implica pérdida de información única, sino la corrección de redundancias que distorsionaban las métricas. Los criterios que usamos para eliminar duplicados, junto con el registro de los identificadores afectados, quedaron documentados para asegurar su trazabilidad y poder verificarlos más adelante.

Tratamiento de inconsistencias en Square_Footage

En esta etapa se evaluó la calidad de la variable *Square_Footage* (superficie cubierta) para asegurar su coherencia con el resto de los atributos. Al inspeccionar el conjunto, se detectaron cuatro observaciones con *Square_Footage* = 0 que, sin embargo, presentan valores positivos en número de dormitorios, baños y precio de venta, esta combinación es

incompatible con el concepto de superficie habitable y, por tanto, se interpreta como error de carga y no como terrenos sin construcción. Asimismo, se identificaron dos registros con *Square_Footage* faltante (NaN). En consecuencia, se consideró que las cuatro superficies igual a cero constituyen valores inválidos y se trataron como faltantes, sumando seis casos a imputar.

En lugar de eliminar estas observaciones, lo que reduciría innecesariamente el tamaño muestral y podría introducir sesgo, se optó por imputar la superficie mediante el método de K vecinos más cercanos (KNN). Para ello, los ceros se recodificaron como NaN, y la estimación de *Square_Footage* se basó en propiedades con características similares consideradas informativas para la similitud, tales como número de dormitorios y baños, año de construcción, tamaño del lote, tamaño del garaje, calidad del vecindario y precio de la vivienda.

Tratamiento del formato de los datos

Durante la etapa inicial de depuración se detectaron inconsistencias en la forma de registrar ciertas variables. En las columnas *Num_Bedrooms* y *Garage_Size* coexistían valores expresados como números (0, 1, 2, 3, etc.) con sus equivalentes escritos en palabras (zero, one, two, three, four, five). Esta mezcla de formatos impedía tratar dichas variables como estrictamente numéricas y podía distorsionar cualquier análisis descriptivo o modelo posterior.

Adicionalmente, en las variables *Square_Footage* y *House_Price* los valores aparecían como texto, utilizando separadores de miles (,) y puntos decimales, lo cual también dificultaba su conversión directa a formato numérico.

Para solucionar estos problemas se realizó un proceso de estandarización de formatos. En primer lugar, se reemplazaron las palabras zero, one, two, three, four y five por sus respectivos valores numéricos (0, 1, 2, 3, 4 y 5), eliminando también espacios en blanco o caracteres residuales. En segundo lugar, se limpiaron las columnas *Square_Footage* y *House_Price* eliminando los separadores de miles y convirtiendo el contenido de texto a números decimales. De esta forma, todas las variables que representan cantidades quedaron expresadas de manera uniforme y adecuadas para su análisis cuantitativo.

Outliers

Para justificar la eliminación de los outliers en el precio de las casas, es importante dejar claro cómo se detectaron, por qué representan un problema para el análisis y por qué es metodológicamente correcto excluirlos en este contexto.

En primer lugar, la detección de valores atípicos se realizó mediante un boxplot de la variable *House_Price*. Este gráfico mostró que la gran mayoría de las observaciones se concentra en un rango acotado (aproximadamente hasta algo más de un millón de unidades monetarias), mientras que se identificaron dos observaciones con precios extraordinariamente elevados, del orden de los diez millones. Para formalizar esta observación visual se empleó el criterio estadístico basado en el rango intercuartílico (IQR): se calcularon el primer cuartil (Q1), el tercer cuartil (Q3) y el IQR = Q3 − Q1, y se definieron como outliers aquellos valores que exceden el límite superior Q3 + 1.5 * IQR. Bajo este criterio, los dos registros detectados quedaron claramente fuera del rango considerado “normal” para la distribución de precios.

Desde el punto de vista analítico, mantener estos valores extremos en el conjunto de datos puede generar varios problemas. En particular, los outliers tienden a sesgar medidas de tendencia central y dispersión (como la media y la desviación estándar), a aumentar artificialmente la varianza residual y a distorsionar la estimación de los parámetros en modelos de regresión (por ejemplo, haciendo que la pendiente se ajuste de forma desproporcionada para explicar solo un par de observaciones muy extremas). Dado que el objetivo del proyecto es analizar y modelizar el comportamiento “típico” del mercado de viviendas de este dataset, la presencia de un número tan reducido de observaciones extremadamente altas, que no son representativas del resto de la muestra, puede conducir a conclusiones poco robustas y a modelos con menor capacidad de generalización.

Además, estos valores tan elevados pueden responder a dos situaciones: (i) errores de registro o carga de datos (por ejemplo, un precio con ceros de más o una unidad mal especificada), o (ii) propiedades de características excepcionales (casas de lujo que conforman un segmento de mercado distinto del resto de las observaciones). En cualquiera de los dos casos, y especialmente considerando que solo se trata de dos registros frente a más de mil observaciones, su influencia sobre los resultados es desproporcionada en relación con la información que aportan. Por ello, resulta razonable tratarlos como outliers y excluirlos del análisis principal.

En consecuencia, la decisión de eliminar los outliers de *House_Price* se considera metodológicamente correcta porque:

1. Se apoyó en un criterio estadístico estándar y objetivo (regla de IQR).
2. Los valores atípicos representan una proporción mínima de la muestra, pero tienen un impacto potencialmente muy alto sobre las estimaciones.
3. El foco del estudio está en el comportamiento promedio del mercado de viviendas contenido en el dataset, y no en casos extremos o segmentos de lujo.

Entonces, al depurar la base de datos y trabajar con un conjunto de precios más homogéneo y representativo, se mejora la calidad de los análisis descriptivos y se incrementa la estabilidad e interpretabilidad de los modelos predictivos que se construyan sobre estos datos.

Tras la eliminación de los valores atípicos conforme al criterio definido, se generó el nuevo diagrama de caja, que refleja la distribución depurada de la variable y permite una comparación más precisa de su mediana, rango intercuartílico y posibles valores extremos residuales.

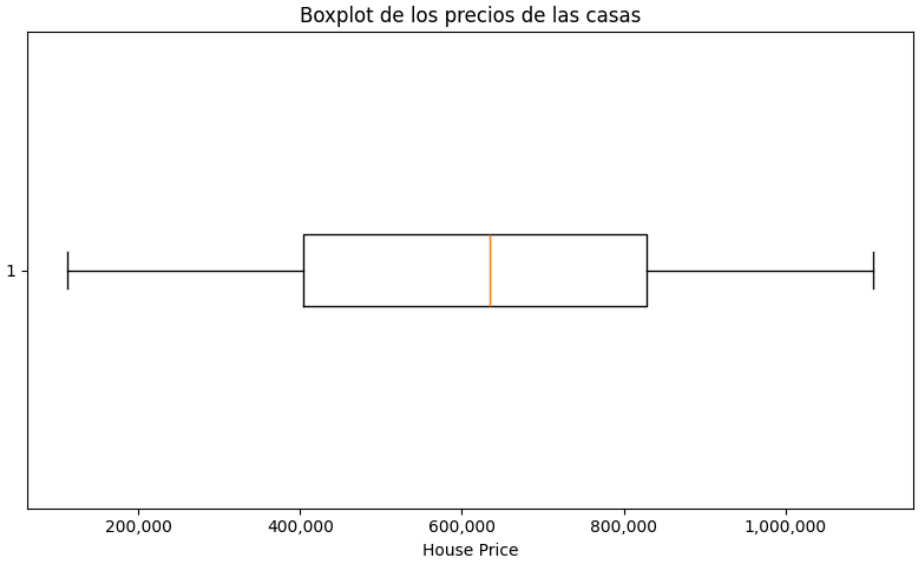

Figura 4. *Boxplot sin outliers.*

Feature Engineering

Estandarización de variables

En la fase de feature engineering se comenzó estandarizando un subconjunto de variables continuas con el objetivo de homogeneizar sus escalas y facilitar el desempeño de modelos sensibles a la magnitud de los predictores. En particular, se trabajó sobre *Square_Footage*, *Lot_Size* y *Neighborhood_Quality*, aplicando un StandardScaler que transforma cada columna a puntuaciones z (media 0 y desvío estándar 1) a partir de la distribución observada. De este modo, se evita que diferencias de unidades o rangos dominen las distancias o los coeficientes durante el ajuste.

Cambios en las variables

Con el fin de obtener una medida más informativa y directamente interpretable, la variable *Year_Built* (año de construcción) se transformó en antigüedad de la propiedad. Esta nueva característica se calculó como la diferencia entre un año de referencia (2025) y el año de construcción (*Antigüedad* = Año_referencia − *Year_Built*). La transformación permite modelar el efecto del paso del tiempo, como el desgaste, obsolescencia o actualizaciones, de manera más lineal y estable que con el año bruto, reduce la colinealidad con otras variables temporales y facilita la comparación entre inmuebles relevados en distintos años.

A su vez, se transformó el precio de las viviendas a su escala logarítmica. El logaritmo natural atenúa la asimetría típica de los precios inmobiliarios, reduce la influencia de valores extremadamente altos, estabiliza la varianza entre rangos de precio y favorece relaciones más próximas a la linealidad con los predictores. Además, la interpretación de los coeficientes se vuelve más intuitiva, ya que los cambios en la variable explicativa se asocian con variaciones porcentuales en el precio. En suma, trabajar con el precio en escala logarítmica aporta robustez estadística, mayor normalidad de los residuales y una lectura económica más clara de los efectos estimados.

Creación de variables

En esta etapa se incorporaron variables derivadas para capturar relaciones estructurales y efectos de interacción que no están explícitos en los predictores originales.

Primero, se construyeron relaciones de tamaño y proporciones. *Build_Ratio* contrasta la superficie cubierta con el tamaño del lote y aproxima el grado de ocupación del terreno, valores altos sugieren mayor densidad edificada, mientras que valores bajos indican mayor espacio libre. *Bedroom_Density* expresa cuántos dormitorios hay por unidad de superficie cubierta y funciona como un indicador de compacidad del layout interior. *Bathroom_Ratio* relaciona baños y dormitorios, aproximando la comodidad del hogar, valores mayores suelen asociarse con segmentos de mayor calidad.

Luego se añadieron interacciones entre variables para modelar efectos conjuntos. *Quality_Garage* combina la calidad del vecindario con el tamaño del garaje, capturando que el aporte de un garaje puede ser mayor en zonas de alto nivel. *Area_Quality* multiplica superficie y calidad del vecindario, permitiendo que el efecto de los metros cuadrados varíe según el entorno. *Size_Bedroom_Interaction* incorpora el hecho de que la utilidad de más superficie puede depender del número de dormitorios (por ejemplo, espacios mejor distribuidos en viviendas más grandes).

Se incluyeron indicadores binarios con interpretación directa. *Has_Garage* identifica propiedades que efectivamente cuentan con garaje, diferenciándolas de las que no lo tienen. *Is_New* marca inmuebles con antigüedad reducida (cinco años o menos), útil para capturar primas asociadas a menor desgaste u obsolescencia.

Como medida de calidad, se creó *Size_per_Bedroom*, que resume la superficie cubierta disponible por dormitorio, valores más altos tienden a reflejar mayor amplitud de los ambientes.

Para robustecer estas nuevas variables, se controlaron divisiones por cero y valores faltantes que pudieran originar infinitos o ausencias, sustituyéndolos por ceros tras la limpieza. Esto garantiza que las nuevas variables sean utilizables en análisis exploratorios y modelos sin fallos de ejecución (aunque, metodológicamente, puede considerarse imputar valores distintos de cero cuando el contexto lo justifique).

En conjunto, estas variables enriquecen la representación del inmueble y del contexto, mejoran la capacidad del modelo para capturar no linealidades y heterogeneidad de efectos y, en general, aportan interpretaciones más finas sobre densidad, confort, interacción entre calidad y tamaño, y atributos clave como la presencia de garaje o la novedad de la propiedad.

Resumen de las variables creadas:

Variable	Descripción breve
Build_Ratio	Proporción entre superficie cubierta y tamaño del lote; aproxima la densidad de construcción.
Bedroom_Density	Dormitorios por unidad de superficie cubierta; indica compacidad del interior.
Bathroom_Ratio	Relación baños/dormitorios; aproxima el nivel de comodidad del hogar.
Quality_Garage	Interacción calidad del vecindario × tamaño de garaje; captura el valor del garaje según el entorno.
Area_Quality	Interacción superficie cubierta × calidad del vecindario; permite que el efecto de los m² varíe por zona.
Size_Bedroom_Interaction	Interacción superficie cubierta × número de dormitorios; refleja cómo la utilidad del espacio depende de los ambientes.
Has_Garage	Indicador binario de presencia de garaje (1 si tiene, 0 si no).
Is_New	Indicador binario de propiedad nueva o reciente (antigüedad ≤ 5 años).
Size_per_Bedroom	Superficie cubierta promedio por dormitorio; proxy de amplitud de ambientes.

Tabla 3. *Variables creadas.*

Feature Selection

La selección de variables busca identificar el subconjunto de predictores con mayor poder explicativo y menor redundancia, para mejorar la generalización del modelo y su interpretabilidad. Se evaluó la relevancia (correlación con el objetivo) y la no colinealidad entre atributos.

Exploración de dependencias lineales y colinealidad

En esta etapa se delimitó el análisis a las variables numéricas del conjunto, excluyendo las de texto o categóricas, para evaluar relaciones lineales de manera consistente. Con ese subconjunto se estimó la matriz de correlaciones de Pearson y se la representó mediante un mapa de calor, lo que permite detectar patrones de asociación y posibles colinealidades entre predictores. Adicionalmente, se calculó la correlación de cada variable con la variable objetivo transformada (*House_Price_log*) y se ordenaron los resultados de mayor a menor, obteniendo un ranking preliminar de relevancia. Esta lectura sirve como filtro inicial para la selección de atributos y para anticipar pares de variables redundantes, con la salvedad de que la correlación capta solo relaciones lineales y no implica causalidad.

Para evaluar la multicolinealidad entre los predictores más asociados con el precio, se seleccionó un subconjunto de variables (*Square_Footage, Size_Bedroom_Interaction, Size_per_Bedroom, Lot_Size* y *House_Age*), se añadió una constante y se calculó el Variance Inflation Factor (VIF) para cada una. El VIF cuantifica cuánto se infla la varianza de un coeficiente por su correlación lineal con el resto: valores cercanos a 1 indican independencia, por encima de 10, colinealidad severa. Los resultados mostraron VIF muy elevados para S*quare_Footage, Size_Bedroom_Interaction* y *Size_per_Bedroom*, lo que evidencia que las dos últimas están fuertemente redundadas por la primera (todas dependen del tamaño interior).

Con base en este diagnóstico, se procedió a depurar el set de predictores: se eliminaron *Size_Bedroom_Interaction* y *Size_per_Bedroom* por su alta multicolinealidad con *Square_Footage*, y se conservaron las demás variables derivadas que aportan información independiente para la predicción del precio (por ejemplo, *Lot_Size* y *House_Age*, cuyos VIF fueron ≈1). Esta decisión reduce la varianza de las estimaciones, mejora la estabilidad numérica del modelado y favorece la interpretabilidad al mantener *Square_Footage* como medida canónica de tamaño, evitando duplicar efectos del mismo constructo.

Modelos Preliminares

Modelo lineal

Se estimó una regresión lineal para explicar *House_Price_log* usando como predictores: *Square_Footage, Lot_Size, Num_Bathrooms, Neighborhood_Quality, Garage_Size, Build_Ratio, Quality_Garage, Area_Quality, Has_Garage, Is_New* y *House_Age*. El conjunto se dividió en entrenamiento (80%) y prueba (20%).

En el conjunto de prueba, el modelo mostró:
MAE = 0.0942, MSE = 0.0172, RMSE = 0.1310 y R^2 = 0.9350.
Dado que el objetivo está en logaritmos, estos errores se interpretan aproximadamente como ≈10 % de error absoluto medio (MAE) y ≈14 % de error cuadrático medio (RMSE) sobre el precio en escala original.

La validación cruzada muestra un MAE entre 50.000 y 59.000 USD, con un promedio de 54.569 USD y una desviación estándar de 3.260 USD. Esto indica que el modelo lineal es estable y consistente entre folds, pero sostiene un error elevado de forma sistemática, lo que sugiere que la relación entre las variables y el precio no está siendo capturada adecuadamente bajo un modelo estrictamente lineal.

Finalmente, el gráfico muestra que los puntos se alinean alrededor de la diagonal de referencia, confirmando visualmente el buen ajuste del modelo y revelando solo desvíos leves en los precios más altos y más bajos.

Entonces, el modelo lineal muestra un desempeño estable pero limitado, logra capturar parte de la variabilidad del precio en escala logarítmica, pero mantiene errores elevados de manera consistente, tal como se evidencia en la validación cruzada. Su comportamiento es coherente y estable entre folds, aunque la precisión alcanzada es moderada y refleja que la relación entre las variables y el precio no es completamente lineal.

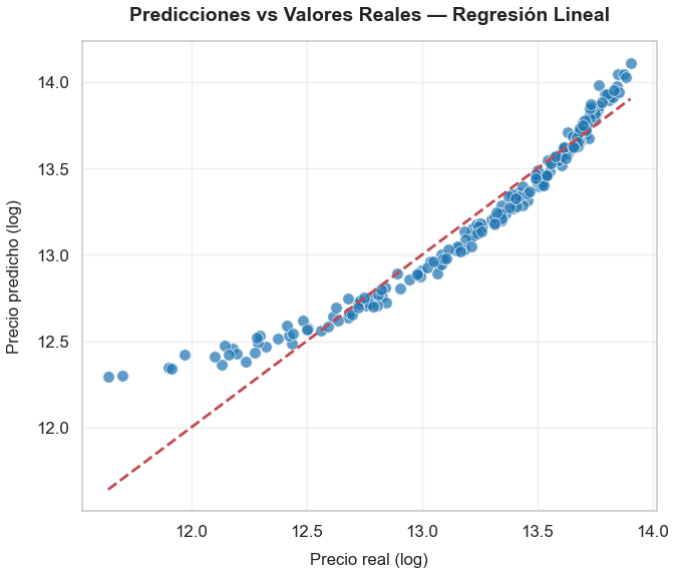

Figura 5. *Modelo lineal, predicciones vs valores reales.*

Random Forest

Se entrenó un RandomForestRegressor con 200 árboles, profundidad sin restricción, random_state=42 y paralelización completa (n_jobs=-1). El modelo se ajustó sobre las mismas variables predictoras usadas en el lineal y se evaluó en el conjunto de prueba (20%).

Obtuvo MAE = 0.0352, RMSE = 0.0514 y R^2 = 0.9900. Dado que el objetivo está en logaritmos, estos errores se traducen en desviaciones porcentuales medianas muy bajas al volver a la escala original del precio, indicando una capacidad predictiva notablemente superior al modelo lineal.

La validación cruzada muestra un MAE (en USD) por fold que se mueve entre 17.685 y 19.736 USD, con un promedio de 18.482 USD y una desviación estándar de solo 820 USD. Estos valores indican que el modelo Random Forest es muy estable y consistente, el error se mantiene bajo en todos los folds y la variación entre particiones es pequeña.

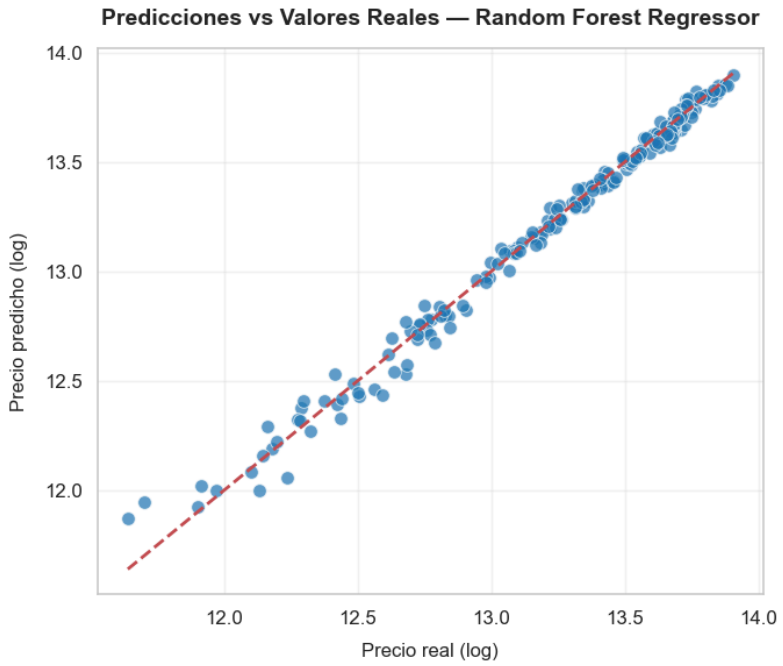

Figura 6. *Modelo Random Forest, predicciones vs valores reales.*

XGBoost

En esta etapa del trabajo se utilizó un modelo de árboles de decisión potenciados (XGBoost) para predecir la variable objetivo. La intuición detrás de este enfoque es combinar muchos árboles simples (cada uno con capacidad limitada) de manera secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por los anteriores. De este modo, el modelo va “aprendiendo” de sus propios desaciertos y ajustando progresivamente sus predicciones, lo que suele traducirse en un alto poder explicativo sin necesidad de estructuras excesivamente complejas en cada árbol individual.

Además, se incorporan mecanismos para controlar el sobreajuste, como limitar la profundidad de los árboles y trabajar con subconjuntos de datos y variables en cada iteración. Intuitivamente, esto obliga al modelo a aprender patrones generales de la relación entre las variables y no solo memorizar casos específicos del conjunto de entrenamiento, favoreciendo así una mejor capacidad de generalización.

Los resultados de la validación cruzada muestran un MAE por fold que varía entre aproximadamente 17.810 y 21.483 USD, con un promedio de 19.343 USD y una desviación estándar de 1.232 USD. Esto indica que el modelo tiene un desempeño estable, aunque con algo más de variabilidad entre folds en comparación con otros modelos, reflejada en una desviación estándar moderada. Aun así, el error se mantiene relativamente bajo en todas las particiones, lo que sugiere que el modelo logra capturar adecuadamente la

relación entre las variables y el precio. En conjunto, la validación cruzada confirma que XGBoost posee una buena capacidad de generalización, aunque con un rendimiento ligeramente menos homogéneo entre folds.

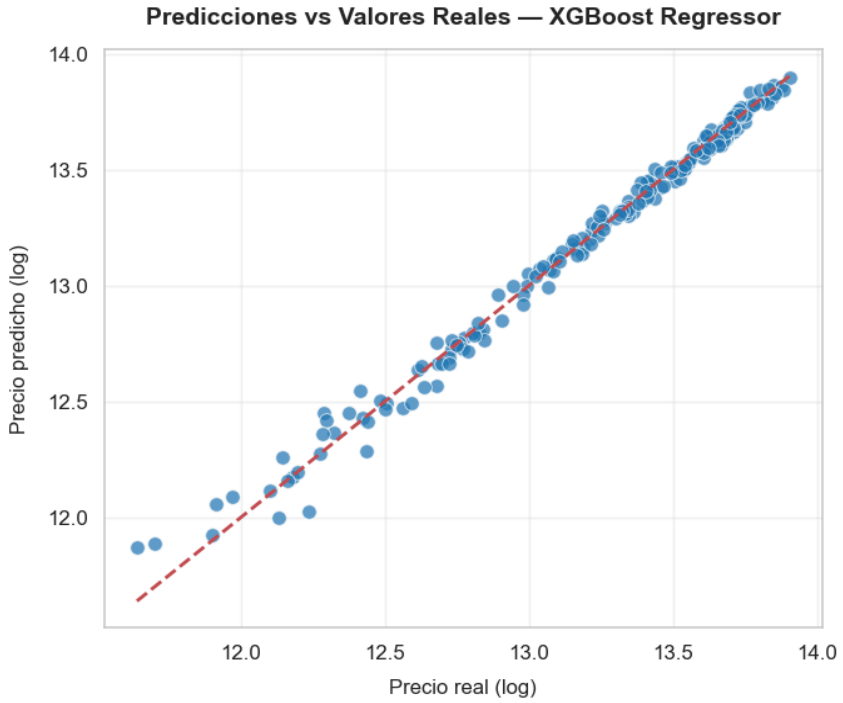

Figura 7. *Modelo XGBoost, predicciones vs valores reales.*

K-Nearest Neighbors (KNN)

Para contrastar con los modelos basados en árboles, también se estimó un modelo de K-Nearest Neighbors (KNN) para regresión con (k = 5) vecinos. La intuición de este método es sencilla, para predecir el valor de la variable objetivo de una nueva observación, el modelo busca en el conjunto de datos aquellas observaciones “más parecidas” (las 5 más cercanas en el espacio de variables explicativas) y asigna como predicción el promedio de sus valores observados. Es decir, asume que un caso se parece en su resultado a los casos que se le parecen en sus características. Se trata de un enfoque puramente “local” y no paramétrico, no impone una forma funcional global, sino que se apoya en la estructura de vecindades de los datos.

La validación cruzada del modelo KNN muestra un MAE (en USD) por fold entre aproximadamente 115.000 y 124.000 USD, con un MAE promedio de 118.456 USD y una desviación estándar cercana a 3.128 USD. Esto indica que el modelo es estable entre folds,

ya que la variación del error es relativamente baja, pero al mismo tiempo evidencia que el error es muy elevado de manera consistente. El hecho de que todos los folds presenten un MAE tan alto sugiere que KNN no logra capturar adecuadamente la estructura del problema y comete errores grandes de forma sistemática.

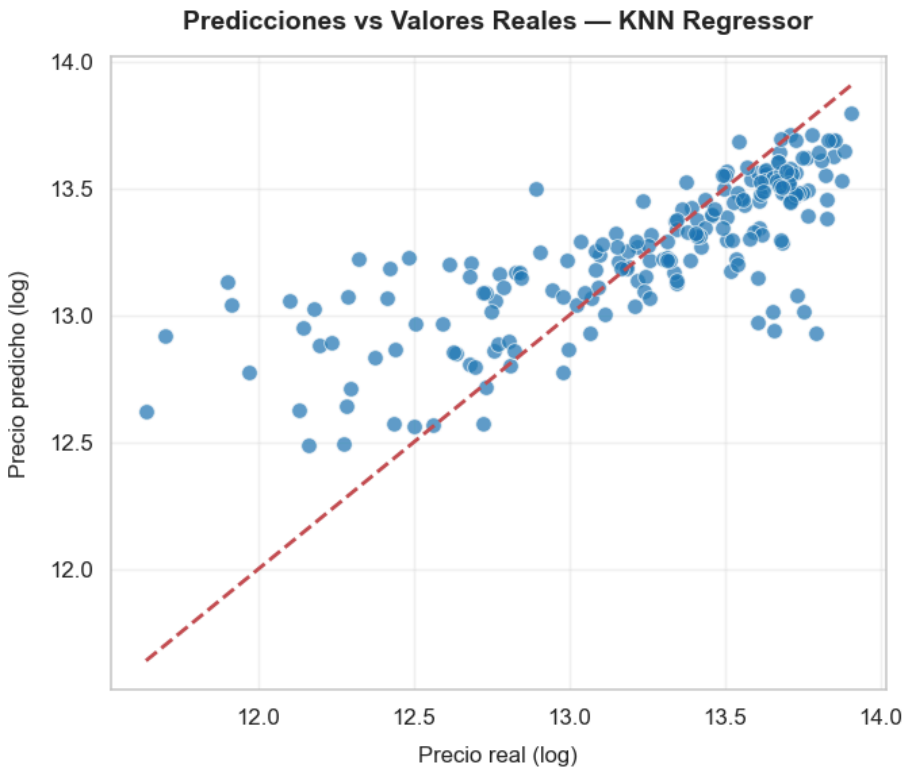

Figura 8. *Modelo KNN, predicciones vs valores reales.*

Implementación y resultados

Resultados

Se calcularon las métricas de todos los modelos en escala logarítmica y en USD para comparar los resultados y ver que modelo lograba predecir mejor la variable precio de la vivienda. A continuación una tabla con las métricas de cada modelo, calculadas en el train y test set.

Métricas en escala logarítmica — Train y Test

	Modelo	MAE_Train	RMSE_Train	R²_Train	R²_Ajustado_Train	MAE_Test	RMSE_Test	R²_Test	R²_Ajustado_Test
0	Lineal	0.093400	0.120200	0.938700	0.937900	0.094200	0.131000	0.935000	0.931200
1	Random Forest	0.014500	0.021200	0.998100	0.998100	0.035200	0.051400	0.990000	0.989400
2	XGBoost	0.010700	0.013700	0.999200	0.999200	0.034600	0.050400	0.990400	0.989800
3	KNN	0.179900	0.253500	0.727500	0.723700	0.238100	0.339800	0.562700	0.537200

Figura 9. *Métricas en escala logarítmica.*

MAE en dólares — Train y Test

	Modelo	MAE_Train_USD	MAE_Test_USD
0	Lineal	53407.000000	50424.000000
1	Random Forest	7115.000000	17678.000000
2	XGBoost	6154.000000	17496.000000
3	KNN	94333.000000	123879.000000

Figura 10. *MAE en dólares.*

Luego de identificar a Random Forest y XGBoost como los modelos con mejor desempeño inicial, se realizó un proceso de ajuste de hiperparámetros para evaluar si era posible mejorar su precisión y capacidad de generalización. Para ello, se entrenaron múltiples iteraciones de cada modelo modificando parámetros clave midiendo en cada caso el MAE en escala logarítmica y el MAE en dólares, tanto en train como en test. Este procedimiento permitió analizar cómo afectan los hiperparámetros al comportamiento del modelo, buscando reducir el error y, al mismo tiempo, disminuir la brecha entre el conjunto de entrenamiento y el de prueba.

Iteración de modelos

Como se mencionó anteriormente, se realizó una iteración de los dos modelos con mejor desempeño, cambiando los hiperparametros e iterando 5 veces. A su vez, se calcularon las métricas en en train y test set.

Aunque la última iteración del Random Forest obtuvo un MAE_Test_USD levemente inferior al del modelo original, la mejora fue mínima aprox 15 USD, una diferencia insignificante dentro del rango de precios del dataset. Además, las demás métricas se mantuvieron prácticamente iguales. Esto indica que el modelo original ya estaba adecuadamente ajustado y que las variaciones introducidas en los hiperparámetros no aportaron una mejora significativa ni en precisión ni en capacidad de generalización.

Random Forest — Iteraciones de Hiperparámetros

	Iteración	n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	MAE_Train_log	MAE_Test_log	MAE_Train_USD	MAE_Test_USD
0	1	200	None	2	1	sqrt	0.020600	0.051600	10679.000000	25218.000000
1	2	400	None	2	1	sqrt	0.020300	0.051900	10450.000000	25380.000000
2	3	300	None	4	2	sqrt	0.029100	0.053900	14943.000000	26272.000000
3	4	600	None	4	2	sqrt	0.029300	0.054000	15020.000000	26530.000000
4	5	300	None	2	1	None	0.014400	0.035200	7100.000000	17621.000000

Figura 11. *Iteraciones modelo Random Forest.*

Las iteraciones realizadas sobre XGBoost no lograron mejorar de manera significativa el desempeño del modelo original. Aunque algunas configuraciones produjeron leves variaciones en el MAE ninguna reducción fue suficientemente relevante como para superar al modelo base. Las diferencias en MAE_Test_USD estuvieron dentro de márgenes muy pequeños, sin aportar mejoras reales en la capacidad predictiva. Además, ciertas configuraciones redujeron demasiado el error en entrenamiento, aumentando el riesgo de sobreajuste sin beneficios en los datos de prueba. En conjunto, los resultados indican que el modelo XGBoost original ya estaba adecuadamente ajustado, y que los cambios explorados en los hiperparámetros no brindaron ventajas adicionales en precisión ni en generalización.

XGBoost — Iteraciones de Hiperparámetros

	Iteración	n_estimators	learning_rate	max_depth	subsample	colsample_bytree	MAE_Train_log	MAE_Test_log	MAE_Train_USD	MAE_Test_USD
0	1	200	0.100000	4	0.900000	0.900000	0.011800	0.034900	6713.000000	17222.000000
1	2	300	0.050000	5	0.900000	0.900000	0.008500	0.034500	4915.000000	16946.000000
2	3	400	0.050000	4	0.800000	0.800000	0.012700	0.034600	7288.000000	17482.000000
3	4	500	0.050000	6	0.900000	0.900000	0.001900	0.033800	1156.000000	16701.000000
4	5	600	0.030000	5	0.800000	0.800000	0.007600	0.036000	4434.000000	17800.000000

Figura 12. *Iteraciones modelo XGBoost.*

Los resultados mostraron que, si bien algunas configuraciones lograron pequeñas variaciones en el error, ninguna de las iteraciones produjo mejoras significativas respecto al rendimiento de los modelos originales. Tanto Random Forest como XGBoost demostraron ser modelos ya bien calibrados en su configuración base, con un comportamiento estable y errores reducidos. Por lo tanto, se concluye que los modelos iniciales representan las mejores versiones obtenidas para este problema, sin que los ajustes adicionales aporten beneficios relevantes en términos predictivos.

Conclusión

El proyecto desarrolló un proceso integral de predicción del precio de viviendas, abarcando desde la limpieza de datos y la ingeniería de variables hasta la comparación y ajuste de diferentes modelos. Tras la depuración del dataset, la eliminación de outliers extremos y la creación de nuevas variables relevantes se entrenaron múltiples modelos con el fin de encontrar un modelo correcto para la predicción de la variable estudiada.

Los resultados mostraron diferencias claras entre los enfoques utilizados. El modelo lineal, si bien estable y altamente interpretable, presentó errores sistemáticamente elevados, lo que evidencia que la relación entre las variables y el precio no es puramente lineal. El KNN, por su parte, mostró dificultades importantes para generalizar, con errores muy altos y desempeño inestable. En contraste, los modelos Random Forest y XGBoost demostraron excelente capacidad predictiva, alcanzando los valores más bajos de MAE y mostrando una brecha reducida entre el train y el test set, lo que indica una sólida capacidad de generalización.

Posteriormente, se realizó un ajuste manual de hiperparámetros sobre los dos mejores modelos. Aunque se exploraron diversas configuraciones, ninguna iteración logró mejorar de manera significativa el rendimiento de los modelos originales, lo que sugiere que su configuración base ya se encontraba cerca del óptimo para este conjunto de datos. En particular, XGBoost destacó por obtener el MAE más bajo en dólares, lo que lo posiciona como el modelo más adecuado para el objetivo de predecir precios en su escala real.

En conclusión, el trabajo permitió identificar que los modelos Random Forest y XGBoost fueron los más eficaces para este tipo de problema, combinando precisión, estabilidad y capacidad de generalización. El proceso completo, desde la preparación de los datos hasta la validación cruzada, fortaleció la robustez del análisis y permitió llegar a un modelo final confiable para la predicción del precio de viviendas.