¿Cómo realizar la limpieza y análisis de los datos?

PRACTICA 2 – Tipología y ciclo de vida de los datos

Alejandro Alonso Anda y Eduardo Lopez
Máster en Ciencia de Datos

Curso 2025.1

Índice

1 Descripción del dataset

El conjunto de datos utilizado en esta práctica ha sido creado específicamente para este proyecto a partir del *web scraping* de una página web de una inmobiliaria. El objetivo principal del dataset es analizar las características de los inmuebles ofertados y estudiar su relación con el precio de venta, con el fin de construir un modelo que permita estimar dicho precio a partir del resto de variables disponibles.

Este tipo de análisis resulta especialmente relevante en el ámbito inmobiliario, donde el precio de un inmueble depende de múltiples factores como la superficie, la localización o el número de habitaciones. Disponer de modelos explicativos o predictivos en este contexto puede aportar valor tanto a compradores como a vendedores, facilitando la toma de decisiones informadas. Además, al tratarse de datos obtenidos mediante técnicas de *web scraping*, el conjunto de datos presenta problemas habituales de calidad del dato, lo que lo convierte en un caso adecuado para aplicar procesos de integración, limpieza y análisis, tal y como se plantea en el enunciado de la práctica.

El dataset está compuesto por un total de 1869 registros, correspondientes a anuncios individuales de inmuebles, y 9 variables. Estas variables incluyen atributos numéricos, categóricos y campos de texto descriptivo. En la Tabla 1 se resumen las principales variables del conjunto de datos junto con su tipo original y una breve descripción.

Variable	**Tipo original**	**Descripción**
ID	Categórica	Identificador único del anuncio
livingSpaceIcon.svg	Texto	Superficie construida del inmueble
propertyAreaIcon.svg	Texto	Superficie de la parcela
descripcion	Texto	Descripción libre del anuncio
zona	Categórica	Zona geográfica del inmueble
localizacion	Categórica	Localidad o barrio
precio	Texto	Precio anunciado del inmueble
bedroomIcon.svg	Numérica	Número de habitaciones
bathroomIcon.svg	Numérica	Número de baños

Tabla 1: Variables originales del conjunto de datos

Como se observa, varias variables que conceptualmente son numéricas aparecen inicialmente codificadas como texto, especialmente aquellas relacionadas con las superficies y el precio. Asimismo, el conjunto de datos presenta valores faltantes en distintas variables. Estas características justifican la necesidad de aplicar una fase exhaustiva de limpieza y acondicionamiento de los datos antes de proceder a su análisis, que se desarrollará en los apartados siguientes.

2 Integración y selección de los datos

En esta fase se ha llevado a cabo la integración y selección de los datos de interés con el objetivo de preparar el conjunto de datos para su posterior limpieza y análisis. El dataset utilizado procede de una única fuente, concretamente del *web scraping* de una página web inmobiliaria, por lo que no ha sido necesario realizar un proceso de integración de múltiples fuentes externas. No obstante, sí ha sido necesario realizar una integración de tipo semántico para dotar de coherencia y significado a las variables obtenidas durante el proceso de extracción.

En particular, algunas variables presentaban nombres poco descriptivos derivados directamente del scraping de elementos gráficos de la página web, identificados por sufijos como `.svg`. Estos nombres dificultan la interpretación del contenido de las variables y su posterior análisis. Por este motivo, se procedió a renombrar dichas variables utilizando denominaciones claras y semánticamente

significativas en castellano, alineadas con el significado real de los atributos que representan. En concreto, las variables relacionadas con las superficies y las características del inmueble fueron renombradas como `superficie_construida`, `superficie_parcela`, `habitaciones` y `baños`.

Una vez realizado el renombrado, se procedió a la selección de las variables relevantes para el objetivo analítico del proyecto, que consiste en la estimación del precio de los inmuebles a partir de sus características. En este proceso se descartaron aquellas variables que no aportaban información estructurada útil para el análisis, como el identificador único del anuncio o la descripción textual libre, y se conservaron únicamente las variables que describen de forma directa las características físicas del inmueble, su localización y el precio anunciado.

El conjunto de datos resultante tras la fase de integración y selección está compuesto por 1869 registros y 7 variables, que se detallan en la Tabla 2 junto con su tipo de dato actual y una breve descripción.

Variable	**Tipo actual**	**Descripción**
superficie_construida	Texto	Superficie construida del inmueble
superficie_parcela	Texto	Superficie de la parcela
habitaciones	Numérica	Número de habitaciones
baños	Numérica	Número de baños
zona	Categórica	Zona geográfica del inmueble
localizacion	Categórica	Localidad o barrio
precio	Texto	Precio anunciado del inmueble

Tabla 2: Variables seleccionadas tras la fase de integración

El resumen inicial del dataset seleccionado muestra que varias variables que conceptualmente son numéricas continúan estando codificadas como texto y que existen valores faltantes en distintas columnas, especialmente en las variables relacionadas con las superficies y con el número de habitaciones y baños. Estas características ponen de manifiesto la necesidad de aplicar una fase posterior de limpieza y acondicionamiento de los datos, que se aborda en el siguiente apartado.

3 Limpieza de los datos

En este apartado se ha llevado a cabo la limpieza y el acondicionamiento del conjunto de datos seleccionado, con el objetivo de mejorar su calidad y garantizar que los análisis posteriores se realicen sobre una base de datos coherente, homogénea y adecuada desde el punto de vista estadístico. Dado el origen de los datos —obtenidos mediante técnicas de *web scraping*—, esta fase resulta especialmente relevante, ya que es habitual encontrar problemas relacionados con la codificación de variables, valores faltantes y heterogeneidad en las unidades de medida.

3.1 Identificación y gestión de datos faltantes

En primer lugar, se analizó la presencia de valores faltantes en el dataset. Se observó que varias variables presentaban valores nulos, especialmente aquellas relacionadas con las superficies del inmueble y con el número de habitaciones y baños. Sin embargo, un análisis detallado del contexto de los datos permitió concluir que estos valores faltantes no se debían necesariamente a una pérdida de información, sino a situaciones en las que determinadas variables no eran aplicables a ciertos tipos de inmuebles.

Por ejemplo, algunos inmuebles no disponen de parcela, por lo que la variable correspondiente a la superficie de la parcela carece de sentido en dichos casos. De forma similar, la ausencia

de información sobre habitaciones o baños puede indicar que estas características no existen o no son relevantes para el tipo de inmueble anunciado. En este contexto, la imputación de valores mediante medidas de tendencia central introduciría información artificial que no refleja la realidad del inmueble.

Por este motivo, se optó por sustituir los valores faltantes por cero en aquellas variables en las que la ausencia del valor indica explícitamente la inexistencia del atributo. Esta estrategia permite conservar el tamaño del dataset y mantener la coherencia semántica de las variables.

3.2 Conversión de unidades y análisis de las superficies

Un aspecto especialmente relevante del proceso de limpieza fue la gestión de las variables de superficie. Estas variables incluían valores expresados tanto en metros cuadrados como en hectáreas, lo que introducía una fuente clara de incoherencia.

Para resolver este problema, se implementó un proceso de detección y conversión automática de unidades, transformando todos los valores a una unidad común: metros cuadrados. Tras este proceso, se generaron nuevas variables numéricas homogéneas.

La Figura 1 muestra la distribución de la superficie construida una vez homogeneizadas las unidades. Se observa una distribución claramente asimétrica, con presencia de valores extremos coherentes con la heterogeneidad del mercado inmobiliario.

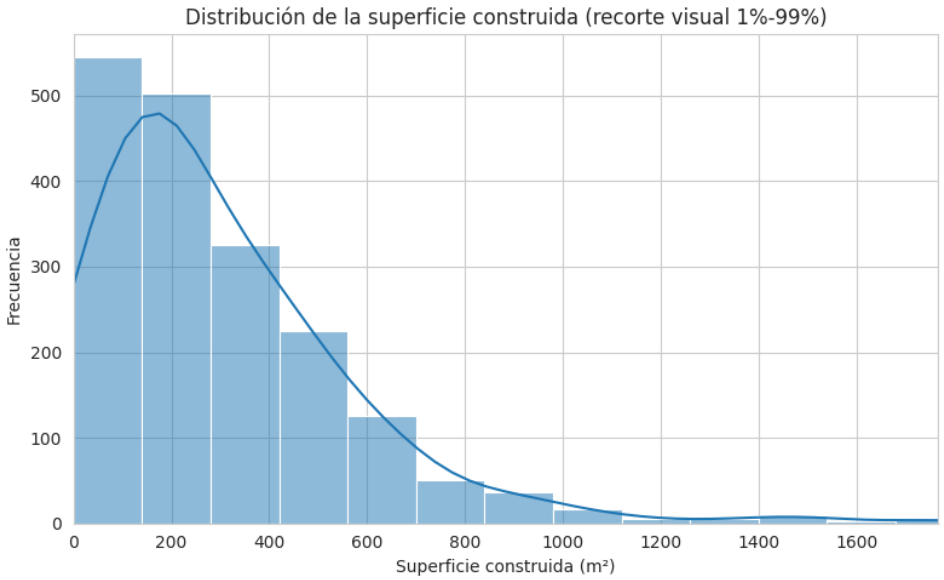

Figura 1: Distribución de la superficie construida tras la conversión de unidades

3.3 Limpieza y análisis de la variable precio

La variable `precio` estaba originalmente codificada como texto e incluía símbolos monetarios y separadores de miles. Se realizó un proceso de limpieza para eliminar estos elementos y convertir el precio a un valor numérico.

La distribución del precio tras la limpieza se muestra en la Figura 2. Al igual que ocurre con las superficies, se observa una distribución asimétrica con valores extremos, característica habitual en datos inmobiliarios.

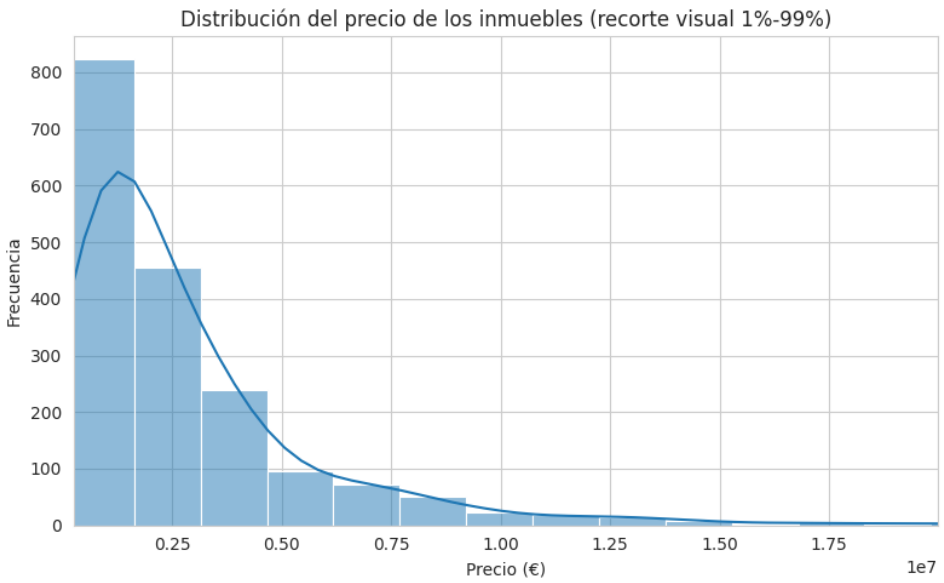

Figura 2: Distribución del precio de los inmuebles

Dado que estos valores extremos son reales y coherentes con la naturaleza del mercado, se decidió conservarlos para el análisis posterior. Con el objetivo de mejorar la legibilidad de las representaciones gráficas, se aplicaron recortes visuales basados en percentiles, sin eliminar ni modificar los datos originales.

4 Análisis de los datos

En este apartado se han aplicado distintas técnicas de análisis de datos con el objetivo de extraer información relevante a partir del conjunto de datos limpio. El análisis se ha estructurado en tres fases: un análisis exploratorio inicial para identificar patrones y relaciones entre variables, la aplicación de modelos de aprendizaje supervisado y no supervisado, y finalmente un contraste de hipótesis para evaluar diferencias significativas entre grupos.

4.1 Análisis exploratorio de los datos

Antes de aplicar técnicas de modelado, se realizó un análisis exploratorio con el objetivo de estudiar la relación entre el precio de los inmuebles y algunas de sus principales características, así como identificar patrones que justificaran los análisis posteriores.

4.1.1 Relación entre precio y superficie

La relación entre el precio del inmueble y la superficie construida se muestra en la Figura 3. Se observa una relación positiva entre ambas variables, lo que indica que, en general, los inmuebles de mayor superficie tienden a presentar precios más elevados. No obstante, también se aprecia una elevada dispersión, especialmente para valores altos de superficie, lo que sugiere la influencia de otros factores no observados en el dataset, como la calidad del inmueble o su estado de conservación.

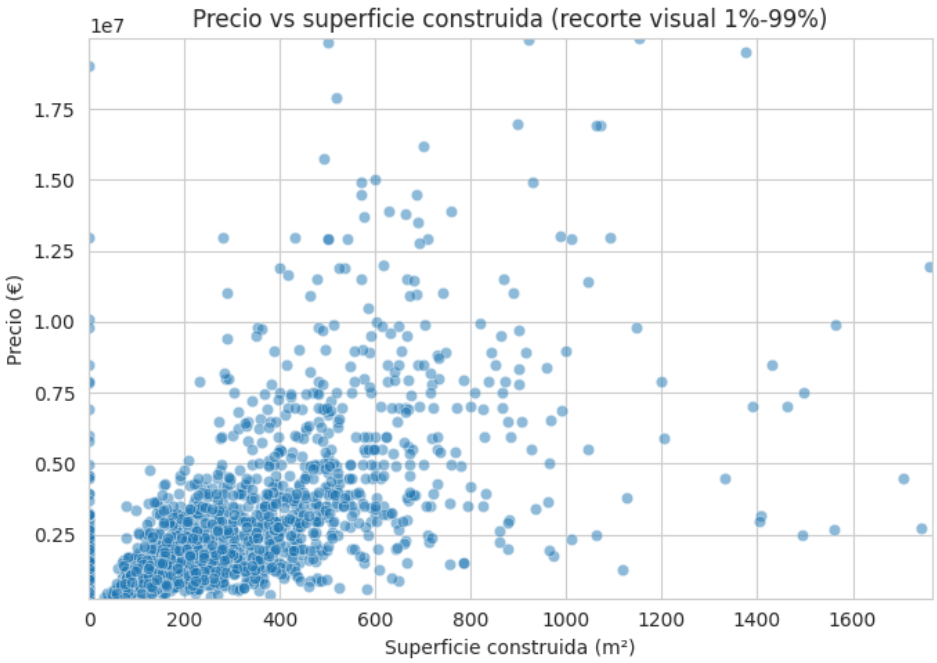

Figura 3: Relación entre el precio y la superficie construida

4.1.2 Precio según número de habitaciones

La Figura 4 muestra la distribución del precio en función del número de habitaciones. Se aprecia un incremento general del precio a medida que aumenta el número de habitaciones, aunque con una variabilidad considerable dentro de cada grupo. Este comportamiento indica que, si bien el número de habitaciones es un factor relevante, no resulta suficiente por sí solo para explicar completamente el precio del inmueble.

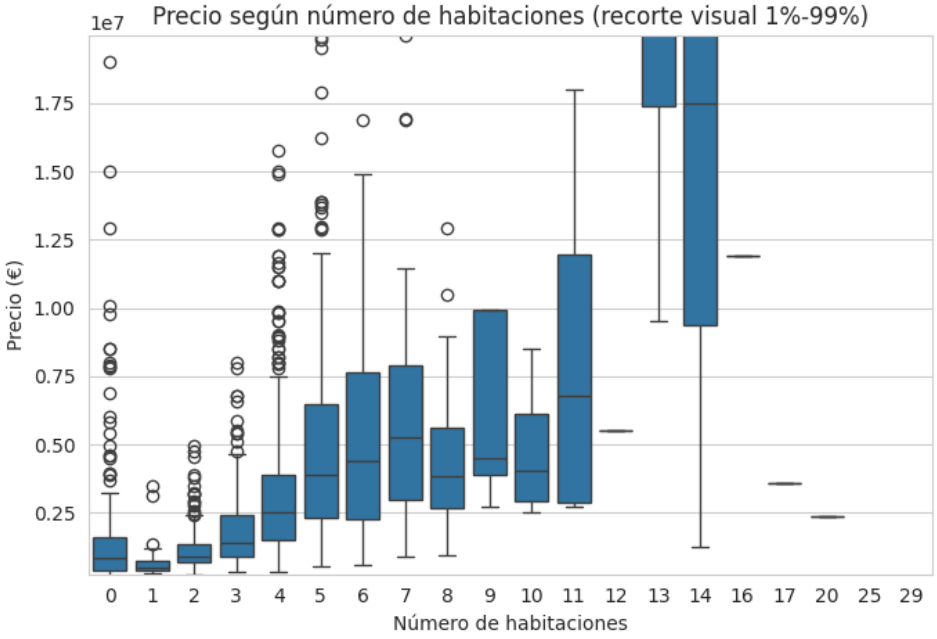

Figura 4: Distribución del precio según el número de habitaciones

4.1.3 Precio según zona geográfica

La distribución del precio por zona geográfica se presenta en la Figura 5. Se observan diferencias claras entre zonas, lo que pone de manifiesto la importancia de la localización en la determinación del precio y justifica el planteamiento posterior de un contraste de hipótesis para evaluar si dichas diferencias son estadísticamente significativas.

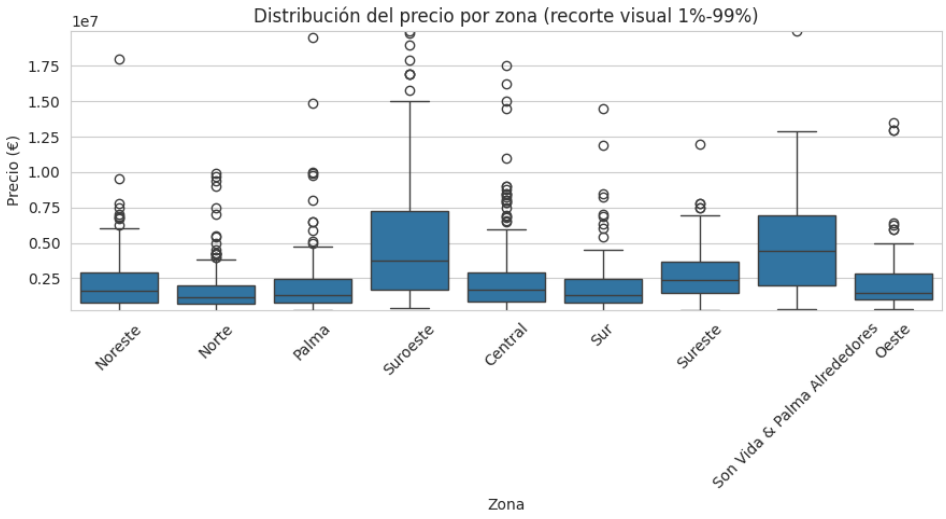

Figura 5: Distribución del precio por zona geográfica

4.2 Modelos supervisados y no supervisados

4.2.1 Modelo supervisado: estimación del precio

Como modelo supervisado se planteó un problema de regresión, tomando como variable objetivo el precio del inmueble y utilizando como variables explicativas la superficie construida, la superficie de la parcela, el número de habitaciones, el número de baños y las variables categóricas de localización (`zona` y `localizacion`).

Dado que el precio es la variable objetivo, se eliminaron previamente los registros con valores nulos en esta variable, quedando un total de 1832 observaciones para el entrenamiento y la evaluación de los modelos. El conjunto de datos resultante se dividió en un 80 % para entrenamiento y un 20 % para test.

Se evaluaron dos modelos de regresión distintos:

- Un modelo de regresión lineal, como aproximación base.
- Un modelo de *Random Forest Regressor*, capaz de capturar relaciones no lineales entre las variables.

El rendimiento de los modelos se evaluó utilizando las métricas MAE (error absoluto medio), RMSE (raíz del error cuadrático medio) y el coeficiente de determinación R^2. Los resultados obtenidos sobre el conjunto de test se resumen en la Tabla 3.

Tabla 3: Resultados de los modelos de regresión

Modelo	**MAE**	**RMSE**	R^2
Random Forest Regressor	1 137 723	2 308 574	0,54
Regresión lineal	1 501 228	2 527 115	0,45

Los resultados muestran que el modelo *Random Forest* obtiene un mejor rendimiento en todas las métricas consideradas, explicando aproximadamente el 54 % de la variabilidad del precio. Esto sugiere que la relación entre las características del inmueble y su precio presenta un comportamiento no lineal, coherente con lo observado en el análisis exploratorio.

4.2.2 Modelo no supervisado: clustering de inmuebles

Como modelo no supervisado se aplicó un algoritmo de *k-means* con el objetivo de identificar grupos de inmuebles con características similares, sin utilizar el precio como variable de entrada. Para ello, se emplearon únicamente variables numéricas relacionadas con las características físicas del inmueble: superficie construida, superficie de la parcela, número de habitaciones y número de baños, previamente estandarizadas.

Para determinar el número óptimo de clusters se utilizó el coeficiente de *silhouette*, evaluando distintos valores de k. El valor que maximiza este coeficiente es $k = 7$, por lo que se seleccionó este número de clusters.

El análisis de los clusters obtenidos muestra la existencia de segmentos claramente diferenciados, que incluyen desde inmuebles de tamaño medio y características estándar hasta grupos más reducidos correspondientes a inmuebles de grandes dimensiones y parcelas muy extensas. Esta segmentación permite identificar tipologías de vivienda diferenciadas dentro del mercado inmobiliario analizado y aporta una visión complementaria al análisis supervisado.

4.3 Contraste de hipótesis

Finalmente, se realizó una prueba de contraste de hipótesis con el objetivo de analizar si existen diferencias significativas en el precio de los inmuebles entre distintas zonas geográficas. Para ello, se seleccionaron las dos zonas con mayor número de observaciones: *Suroeste* y *Central*.

Antes de aplicar el contraste, se verificaron los supuestos necesarios para el uso de pruebas paramétricas. En primer lugar, se evaluó la normalidad de la distribución del precio en ambas zonas mediante el test de Shapiro–Wilk, obteniéndose valores $p < 0{,}05$ en ambos casos, lo que indica que las distribuciones no siguen una distribución normal. Asimismo, el test de Levene para la homogeneidad de varianzas también rechazó la hipótesis de igualdad de varianzas.

Dado que no se cumplen los supuestos de normalidad ni de homocedasticidad, se optó por aplicar una prueba no paramétrica, concretamente el test de Mann–Whitney U, para comparar los precios entre ambas zonas. El resultado del contraste arrojó un valor p prácticamente nulo, lo que indica la existencia de diferencias estadísticamente significativas entre los precios de las dos zonas analizadas.

Desde un punto de vista descriptivo, los inmuebles situados en la zona Suroeste presentan precios claramente superiores a los de la zona Central, lo que refuerza la relevancia de la localización como uno de los factores clave en la determinación del precio de los inmuebles.

5 Conclusiones

En esta práctica se ha desarrollado un proceso completo de análisis de datos partiendo de un conjunto de datos creado específicamente mediante técnicas de *web scraping*. A lo largo del trabajo se han abordado de forma estructurada todas las fases fundamentales de un proyecto de análisis de datos, desde la descripción inicial del dataset hasta la aplicación de modelos de aprendizaje supervisado y no supervisado, así como técnicas de inferencia estadística.

En primer lugar, se ha puesto de manifiesto la importancia de una fase exhaustiva de limpieza y acondicionamiento de los datos. El dataset original presentaba problemas habituales en datos extraídos de fuentes web, como variables numéricas codificadas como texto, unidades heterogéneas, valores faltantes y nombres de variables poco descriptivos. La correcta identificación del significado semántico de estos problemas ha permitido aplicar estrategias de limpieza adecuadas, como la conversión explícita de unidades de superficie a una medida homogénea o el tratamiento diferenciado de valores faltantes cuando estos representaban atributos no aplicables y no información ausente.

En segundo lugar, el análisis supervisado ha permitido comprobar que las características físicas y de localización de los inmuebles explican una parte relevante del precio, aunque no su totalidad. El modelo de *Random Forest* ha mostrado un mejor rendimiento que la regresión lineal, lo que sugiere la existencia de relaciones no lineales entre las variables explicativas y el precio. No obstante, los resultados obtenidos también reflejan la complejidad inherente al mercado inmobiliario y la influencia de factores no observados que no están presentes en el dataset analizado.

Por otro lado, el análisis no supervisado mediante técnicas de *clustering* ha permitido identificar distintos grupos de inmuebles con características diferenciadas, lo que aporta una visión complementaria al análisis supervisado. La segmentación obtenida muestra la coexistencia de tipologías de vivienda claramente distintas dentro del conjunto de datos, desde inmuebles de tamaño medio hasta propiedades con grandes superficies de parcela y características singulares.

Finalmente, el contraste de hipótesis ha puesto de manifiesto la existencia de diferencias estadísticamente significativas en el precio de los inmuebles entre distintas zonas geográficas, reforzando la importancia de la localización como uno de los factores clave en la determinación del precio. La verificación previa de los supuestos de normalidad y homocedasticidad ha permitido seleccionar la prueba estadística adecuada, garantizando la validez de las conclusiones inferenciales.

Como posibles líneas de mejora, cabría ampliar el dataset con nuevas variables relevantes, como el año de construcción, el estado del inmueble o variables socioeconómicas del entorno, así como explorar técnicas de modelado más avanzadas o enfoques de validación más exhaustivos. En cualquier caso, el trabajo realizado demuestra la utilidad de aplicar de forma integrada técnicas de limpieza, análisis exploratorio, modelado y contraste estadístico para extraer conocimiento a partir de datos reales.

6 Resolución del problema

En esta práctica se planteó como objetivo analizar un conjunto de datos inmobiliarios con el fin de comprender qué factores influyen en el precio de los inmuebles y evaluar si los resultados obtenidos permiten responder a la pregunta planteada inicialmente.

Tras el proceso de integración, limpieza y acondicionamiento de los datos, se obtuvo un dataset coherente y homogéneo, adecuado para su análisis. La correcta interpretación del significado de los valores faltantes, así como la conversión explícita de unidades de superficie a una medida común, permitió evitar la introducción de sesgos artificiales y garantizó la validez de los análisis posteriores.

Los resultados del análisis supervisado muestran que es posible explicar una parte relevante del precio de los inmuebles a partir de sus características físicas y de localización. El modelo de *Random Forest* obtuvo un mejor rendimiento que la regresión lineal, lo que indica que la relación entre las variables explicativas y el precio no es estrictamente lineal. No obstante, el valor del coeficiente de determinación obtenido refleja que una parte significativa de la variabilidad del precio no queda explicada por las variables disponibles en el dataset.

Esta explicación parcial de la variabilidad es consistente con la naturaleza del problema analizado. De manera lógica, el precio de un inmueble depende también de factores relacionados con la calidad de la construcción, los acabados, el estado de conservación o las prestaciones del inmueble, variables que no están expuestas ni estructuradas en la página web de origen y, por tanto, no han podido ser incorporadas explícitamente al análisis. La ausencia de estas variables introduce una variabilidad inherente que limita la capacidad explicativa de cualquier modelo construido únicamente a partir de las características observables.

Desde esta perspectiva, los resultados obtenidos no solo son coherentes, sino que aportan una interpretación adicional relevante. La discrepancia entre el precio observado y el precio esperado según el modelo puede interpretarse como una señal indirecta de la calidad del inmueble. En este sentido, el análisis sugiere que es posible plantear, como parte de la solución al problema, la inferencia aproximada de características latentes no observadas a partir del residuo del modelo, abriendo la puerta a futuras líneas de trabajo orientadas a estimar atributos no observados a partir de la información disponible.

Por otro lado, el análisis no supervisado permitió identificar distintos grupos de inmuebles con características claramente diferenciadas. En particular, se observan clusters que corresponden a viviendas residenciales estándar, pero también grupos formados por propiedades con grandes superficies de parcela y escasa o nula superficie construida, que pueden interpretarse como solares o terrenos. Asimismo, se identifican otros clusters asociados a inmuebles con características singulares, donde la relación entre superficie, número de estancias y precio difiere sustancialmente del resto del mercado. Estas diferencias contribuyen a explicar la elevada heterogeneidad del dataset y la presencia de valores extremos en algunas variables.

Finalmente, el contraste de hipótesis confirmó la existencia de diferencias estadísticamente significativas en el precio de los inmuebles entre distintas zonas geográficas. Este resultado permite responder de forma clara a la pregunta planteada, mostrando que la localización es un factor determinante en el precio de los inmuebles y validando empíricamente una de las hipótesis implícitas del análisis.

En conjunto, los resultados obtenidos permiten afirmar que el problema planteado ha sido resuelto de manera satisfactoria. A través de un proceso estructurado de integración, limpieza, análisis y validación de los datos, ha sido posible extraer conocimiento relevante a partir de un conjunto de datos real y responder a la pregunta inicial de forma fundamentada, identificando además nuevas oportunidades de análisis derivadas de las limitaciones del propio dataset.

7 Consideraciones éticas y legales

El uso de técnicas de *web scraping* para la obtención de datos plantea una serie de consideraciones éticas y legales que deben ser tenidas en cuenta a lo largo de todo el ciclo de vida del dato. En este proyecto, la obtención de la información se ha realizado a partir de una página web inmobiliaria de acceso público, sin requerir autenticación, credenciales de usuario ni mecanismos de acceso restringido.

Desde el punto de vista legal, los datos recopilados corresponden exclusivamente a información sobre inmuebles anunciados (características físicas, localización general y precio), y no incluyen

datos de carácter personal que permitan identificar a personas físicas, de acuerdo con la definición establecida en el Reglamento General de Protección de Datos (RGPD). Por tanto, el tratamiento realizado no afecta a datos personales ni vulnera los principios de protección de la privacidad de los usuarios de la plataforma.

Asimismo, durante el proceso de extracción se han respetado los principios de uso responsable de los recursos web, evitando realizar un número elevado de peticiones en un corto intervalo de tiempo que pudiera afectar al funcionamiento del sitio web de origen. La finalidad del scraping ha sido exclusivamente académica, sin ánimo de lucro ni uso comercial de los datos obtenidos.

Desde una perspectiva ética, se ha procurado mantener la integridad y fidelidad de la información recopilada, evitando manipulaciones que pudieran distorsionar el significado original de los datos. Las transformaciones aplicadas durante las fases de limpieza y preparación han tenido como único objetivo mejorar la calidad del dato y facilitar su análisis, sin alterar su contenido semántico.

Por último, en coherencia con los principios de transparencia y reproducibilidad, todo el proceso de obtención, limpieza y análisis de los datos ha sido documentado y el código utilizado se ha puesto a disposición en el repositorio del proyecto. Esto permite evaluar de forma crítica las decisiones adoptadas a lo largo del ciclo de vida del dato y garantiza un uso responsable y éticamente justificado de las técnicas de análisis de datos empleadas.

8 Código

Para la realización de esta práctica se ha optado por desarrollar todo el código en Python, en lugar de utilizar R, a pesar de que este último se menciona como opción preferente en el enunciado. Esta decisión se ha tomado de forma consciente y justificada, atendiendo a distintos factores relacionados con el contexto del proyecto y con el propio proceso de análisis.

En primer lugar, el tamaño del dataset analizado es relativamente reducido, por lo que la velocidad de procesado no constituye un factor determinante en este caso. Ambas herramientas permiten trabajar de forma eficiente con volúmenes de datos de este tamaño, por lo que no existe una ventaja significativa en términos de rendimiento que justifique la elección de un lenguaje frente al otro.

En segundo lugar, Python ofrece una elevada versatilidad para abordar de manera integrada todas las fases del proyecto analítico. A través de librerías como `pandas`, `scikit-learn` y `scipy`, ha sido posible realizar la integración, limpieza, análisis exploratorio, modelado supervisado y no supervisado, así como el contraste de hipótesis, utilizando un único entorno de trabajo coherente. Esta versatilidad resulta especialmente adecuada para proyectos que combinan tratamiento de datos, análisis estadístico y aprendizaje automático.

Por último, la elección de Python responde también a una cuestión de idiosincrasia personal y experiencia previa. El uso de un lenguaje con el que se tiene mayor familiaridad facilita la escritura de código más claro, estructurado y correctamente comentado, lo que redunda en una mayor calidad del trabajo final y en una mejor trazabilidad entre el código desarrollado y los resultados obtenidos.

El código desarrollado se ha organizado de forma modular, separando claramente las distintas fases del proyecto (diagnóstico, integración y selección, limpieza, análisis y representación de resultados). Todo el código utilizado se adjunta en el repositorio Git entregado junto con esta memoria, cumpliendo con los requisitos especificados en el enunciado de la práctica.