DeepKnowledge

DATA CONSULTING

↘PRODUCTO DE DATOS

PRECIO INMOBILIARIO EN LA CDMX

INTEGRANTES:
MARTINEZ BAUTISTA JENNIFER MAGALI
PANIAGUA SUAREZ ALAN ABRAHAM
ROJAS LAGUNAS KEVIN ANTONIO
TAPIA RODRÍGUEZ JAMEL IRAIS

Seminario de Estadística II
(Ciencia de Datos) 2026-01

ÍNDICE

DeepKnowledge
DATA CONSULTING

PRODUCTO FINAL

ANÁLISIS SOBRE EL PRECIO INMOBILIARIO DE LA CDMX.

Introducción

El mercado inmobiliario de la Ciudad de México es uno de los más dinámicos y complejos de América Latina. La determinación del precio de una propiedad no solo depende de sus características físicas, como la superficie construida o el número de recámaras, sino también de una intrincada red de factores geográficos, económicos y sociales que varían significativamente entre alcaldías. Elementos como la accesibilidad, el desarrollo urbano, la densidad poblacional y la plusvalía histórica influyen directamente en la valuación de los inmuebles y generan, en consecuencia, un entorno altamente heterogéneo y difícil de modelar.

Problema por resolver

Tradicionalmente, los procesos de valuación han dependido de métodos manuales, criterios subjetivos o comparativos informales, lo que introduce incertidumbre, sesgos y riesgos financieros tanto para compradores como para empresas del sector. Ante este panorama, surge la necesidad de desarrollar herramientas basadas en Ciencia de Datos que permitan automatizar, sistematizar y aumentar la precisión de la estimación del valor de un inmueble, reduciendo la dependencia de juicios humanos y aprovechando la riqueza de datos disponibles.

Objetivo del proyecto

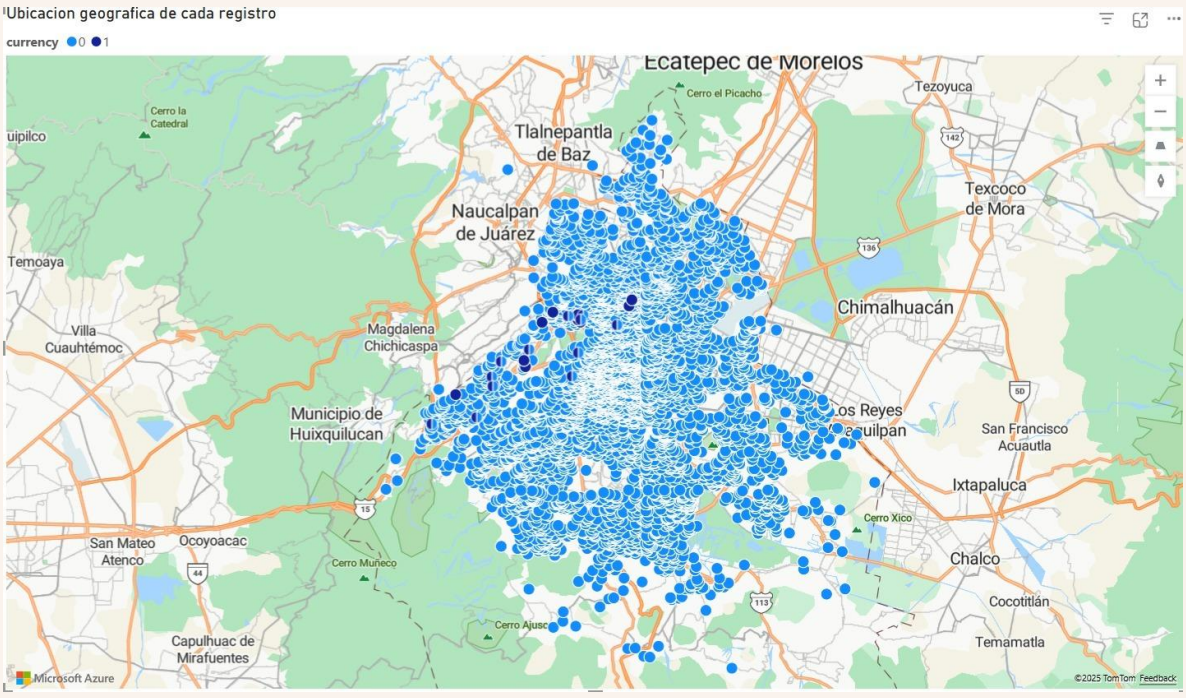
El presente proyecto tiene como objetivo desarrollar un **Producto de Datos capaz de mitigar la incertidumbre en la valuación inmobiliaria**, mediante la aplicación de técnicas avanzadas de Ciencia de Datos. Utilizando el conjunto de datos *“Housing Prices in CDMX”* obtenido de Kaggle, se llevó a cabo un análisis integral que abarca desde la validación de la Calidad de Datos hasta la implementación de modelos predictivos y técnicas de segmentación.
Este producto permite comprender de forma profunda los factores que influyen en el precio de los inmuebles en la Ciudad de México, así como generar modelos y visualizaciones que faciliten su interpretación y futura toma de decisiones.

1. Calidad de Datos

El análisis inició con una revisión exhaustiva del conjunto original de más de 18,000 registros, evaluando cinco dimensiones fundamentales: duplicidad, completitud, consistencia, conformidad e integridad.

- ✓ **Duplicidad:** Se identificaron y eliminaron registros duplicados, reduciendo la base a aproximadamente **15,281 observaciones**. Esto fue fundamental para evitar sesgos por publicaciones repetidas.
- ✓ **Completitud:** Se depuraron registros con valores faltantes en atributos críticos (precio, superficie, geolocalización), garantizando cero valores nulos en variables relevantes.
- ✓ **Consistencia:** Se verificó la relación entre price, price_aprox_local_currency y price_aprox_usd. Para evitar multicolinealidad y fuga de información, se decidió conservar únicamente **price_aprox_local_currency** como variable objetivo.
- ✓ **Integridad Geoespacial:** Se filtraron propiedades ubicadas estrictamente dentro de los límites reales de la CDMX.
- ✓ **Conformidad (Outliers):** Se aplicaron métodos de detección de valores atípicos utilizando **percentiles 10–90** en superficie y precio, dado que los datos inmobiliarios presentan colas pesadas.

Resultado: Se obtuvo un conjunto depurado final de **12,775 registros** (CleanHousingDataCDMX.csv).



2. Análisis Exploratorio e Ingeniería de Datos

Con la base de datos depurada, se desarrolló un análisis exploratorio detallado para comprender la distribución de las variables fundamentales y detectar patrones iniciales. Se identificó que el mercado analizado está dominado por departamentos (74.14%), seguido por casas (25%), mientras que locales comerciales (0.79%) y PH representan menos del 1%. Esta composición confirma el carácter predominantemente vertical del desarrollo habitacional en la CDMX.

Las gráficas de dispersión entre superficie y precio revelaron una tendencia creciente: a mayor superficie construida, mayor precio, aunque con variabilidad significativa producto de factores de ubicación y calidad constructiva. Asimismo, se observó que las propiedades ofrecidas en dólares se concentran en zonas de alto valor económico.

Posteriormente, se llevaron a cabo transformaciones clave en la etapa de ingeniería de datos:

Principales tratamientos realizados

- **Eliminación de columnas redundantes** relacionadas con el precio.
- **Codificación de variables categóricas** mediante *One Hot Encoding* para tipo de propiedad y alcaldía.
- **Conversión de moneda a un indicador binario**.
- **Filtrado geográfico** para conservar únicamente inmuebles dentro de la CDMX.
- **Eliminación de outliers** mediante percentiles 10–90.
- **Escalado de variables numéricas** cuando fue necesario para modelos sensibles (SVR, KNN, RNA).
- **Construcción de la variable categórica price_cat** mediante cuartiles para clasificación.

Estas transformaciones garantizaron un conjunto de datos consistente, estadísticamente robusto y adecuado para la fase de modelado.

3. Modelado

Para evaluar el comportamiento del precio de los inmuebles, se realizaron tanto modelos de regresión (precio continuo) como modelos de clasificación (rango de precio).

La división del conjunto de datos se realizó mediante una proporción **70% para entrenamiento y 30% para prueba**, manteniendo esta configuración constante en todos los modelos para asegurar comparabilidad.

Antes del entrenamiento, todas las transformaciones necesarias (escalado, codificación, selección de variables) se ajustaron únicamente sobre el conjunto de entrenamiento, evitando así cualquier tipo de fuga de información.

3.1 Modelos de Regresión

Se buscaron modelos capaces de predecir el precio aproximado en moneda local. Se probaron los siguientes enfoques:

- **Regresión Lineal Múltiple:**
  - R^2 Train: ≈ 0.4852 y R^2 Test: ≈ 0.5257
  - Diagnóstico: Desempeño moderado. Captura tendencias generales, pero no logra explicar la complejidad y variabilidad específica del mercado inmobiliario.
- **Regresión Polinomial:**
  - R^2 Train: ≈ 0.6173 y R^2 Test: Negativo (-1188.10)
  - Diagnóstico: **Sobreajuste extremo.** La explosión de dimensionalidad resultó incompatible con los datos, generando un modelo que no generaliza.
- **SVR (Kernel RBF):**
  - R^2 Train: ≈ -0.0291 y R^2 Test: ≈ -0.03
  - Diagnóstico: Deficiente. A pesar del escalado, no logró modelar las relaciones complejas ante la gran dispersión de los datos.
- **Árbol de Decisión Regressor:**
  - R^2 Train: ≈ 0.653 y R^2 Test: ≈ 0.632.
  - Diagnóstico: **Mejor Modelo.** Logró capturar interacciones no lineales clave entre la superficie, la zona y el tipo de propiedad, manteniendo una buena capacidad de generalización.

3.2 Modelos de Clasificación (cuartiles de precio)

Se buscó clasificar los inmuebles en 4 rangos de precio (cuartiles). Los resultados de accuracy en el conjunto de prueba fueron:

- **Decision Tree Classifier:**
  - Accuracy (Test): 0.63
  - Observaciones: Desempeño aceptable e interpretable, aunque sensible a pequeños cambios en los datos de entrenamiento.
- **KNN (K-Nearest Neighbors):**
  - Accuracy (Test): ≈ 0.25
  - Observaciones: Muy bajo rendimiento. Se vio afectado severamente por la "maldición de la dimensionalidad".
- **Red Neuronal (RNA):**
  - Accuracy (Test): 0.57 – 0.58
  - Observaciones: Capturó cierta complejidad tras 300 épocas de entrenamiento, pero no logró superar el desempeño de los modelos basados en árboles.
- **Random Forest Classifier:**
  - Accuracy (Test): **0.66** (Train: 0.78)
  - Observaciones: **Mejor Clasificador.** Mostró ser robusto y eficiente, reduciendo la varianza inherente de los árboles simples y manejando mejor la no linealidad de los datos.

3.4. Resultados de Aprendizaje No Supervisado

Se aplicaron técnicas para segmentar el mercado sin etiquetas predefinidas, con el fin de encontrar patrones ocultos.

1. **K-Means:**
   - Utilizando variables de superficie y precio escalados, se determinó un **K óptimo de 6** mediante el método del codo.
   - Los clusters resultantes fueron coherentes, logrando agrupar las propiedades en zonas y rangos de valor lógicos con la realidad del mercado.

2. **PCA (Análisis de Componentes Principales):**
    - Las primeras 3 componentes explicaron el **32.7% de la varianza**.
    - Esta técnica resultó útil para la visualización global de la estructura de los datos, pero insuficiente como modelo predictivo único debido a la complejidad de la información.
3. **DBSCAN:**
    - No logró detectar clusters relevantes.
    - Debido a la altísima densidad de inmuebles y la continuidad urbana de la CDMX, el algoritmo formó un solo cluster masivo, resultando inadecuado para segmentar vecindarios específicos.

4. Mejor Modelo

El **Árbol de Decisión Regressor** se posicionó como el mejor modelo para la predicción del precio en moneda local debido a su capacidad para capturar relaciones no lineales entre las variables. Esta propiedad es importante en mercados complejos donde variables como superficie, tipo de propiedad y alcaldía o delegación interactúan entre sí de manera no lineal.

Además, maneja adecuadamente variables categóricas codificadas mediante One Hot Encoding, como las 16 alcaldías o los distintos tipos de propiedad. Mientras otros modelos sufren por la dimensionalidad y las combinaciones posibles, el árbol puede utilizar estas variables para construir divisiones relevantes que reflejan diferencias sustanciales entre zonas de alta y baja plusvalía, así como, su tolerancia al ruido y su capacidad para minimizar el impacto de valores atípicos. Con un **R^2 de prueba cercano a 0.63**, confirmaron que logra un equilibrio adecuado entre capacidad explicativa y generalización, es decir, qué tanto el modelo logra explicar la variación del precio usando las variables y que tan bien funciona con nuevos datos.

Por otra parte, el **Random Forest Classifier** fue el mejor modelo para clasificar inmuebles por cuartiles de precio porque combina muchos árboles de decisión entrenados sobre distintas muestras del dataset y con diferentes subconjuntos de variables. Esto reduce la varianza, hace el modelo más estable y evita que se sobreajuste a patrones locales o al ruido del mercado inmobiliario.

Además, este modelo maneja muy bien información heterogénea como variables categóricas codificadas, superficies, precios y características complejas de la CDMX. Su capacidad de capturar interacciones entre superficie, zona y tipo de propiedad, sumada a su resistencia a outliers y su tolerancia a ruido estructural, permitió obtener un **accuracy de prueba cercano a 0.66**, el mejor entre todos los

modelos evaluados. También, ofrece interpretabilidad mediante la asignación de importancias a las variables, lo que permite identificar cuáles factores influyen más en la clasificación del rango de precio, siendo la superficie construida, la alcaldía y el tipo de propiedad los más relevantes.

Conclusiones Generales

El estudio permitió comprender en profundidad el comportamiento del mercado inmobiliario de la CDMX y desarrollar un Producto de Datos capaz de evaluar y modelar precios de forma automatizada y basada en evidencia.

Los principales hallazgos fueron:

- La superficie construida y la ubicación son los determinantes más importantes del precio.
- Los modelos lineales son insuficientes para capturar la complejidad del mercado.
- El modelo más adecuado para **predicción de precios** es el **Árbol de Decisión Regressor**.
- Para **clasificación de inmuebles por rangos de precio**, el mejor modelo es el **Random Forest Classifier**.
- Para segmentación, **K-means** ofrece agrupamientos coherentes con la realidad del mercado.

En conjunto, la base de datos representa una fotografía robusta del mercado de vivienda en la Ciudad de México y resulta adecuada para tareas de regresión, predicción de precios, clasificación por tipo de inmueble o zona, así como análisis geoespacial.

ANEXO:

CUADROS COMPARATIVOS DE LOS MODELOS

A continuación se muestran dos cuadros comparativos de los modelos usados, uno de aprendizaje supervisado y otro de no supervisado.

MODELOS

Modelo	Transformaciones y tratamientos aplicados	Métricas (Train / Test)	Resultados y comportamiento del modelo	Por qué NO es recomendable (si aplica)	Conclusión / Mejor uso
Regresión Lineal Múltiple	- One Hot Encoding de tipo de propiedad y alcaldía. - Eliminación de variables con fuga de información (`price_per_m2`). - División 70/30 train-test.	- Train $R^2 \approx$ **0.48** - Test $R^2 \approx$ **0.52**	Modelo estable, sin overfitting. Explica parte del precio, pero no capta la complejidad del mercado.	Captura solo relaciones lineales; el precio depende de muchas interacciones no lineales (ubicación, superficie, mercado).	Útil como modelo base, pero insuficiente para una predicción precisa.
Regresión Polinomial (grado 2)	- One Hot Encoding. - Expansión polinómica (`Polynomial Features`). - Mismo 70/30 train-test. - Regresión lineal sobre los polinomios.	- Train R^2 muy alto. - Test R^2 **negativo extremo** (≈ –1188).	Sobreajuste severo; el modelo aprende ruido y memoriza el train.	Explotación del número de variables implica explosión de dimensionalidad. No generaliza.	**No recomendable** en datasets inmobiliarios; demasiado sensible al ruido.
SVR	- Escalado obligatorio con StandardScaler. - One Hot Encoding previo. - División 70/30.	- Train $R^2 \approx$ **–0.03** - Test $R^2 \approx$ **–0.0291**	Modelo incapaz de capturar la relación entre superficie, precio y ubicación. Tiende a predecir valores cercanos al promedio.	SVR colapsa con muchos dummies (alta dimensionalidad) y alta dispersión. No tolera datasets grandes.	**No recomendable**. Su rendimiento es muy pobre para este tipo de datos.
Decision Tree Regressor	- One Hot Encoding. - Eliminación de outliers. - División 70/30. - Hiperparámetros: `max_depth=6`.	- Train $R^2 \approx$ **0.65** - Test $R^2 \approx$ **0.63**	Captura no linealidad, interacciones y zonas específicas. Generaliza bien.	Puede sobreajustarse si se aumenta la profundidad, pero el modelo usado está controlado.	**Mejor modelo para regresión** según esta práctica.
Decision Tree Classifier	- One Hot Encoding. - Estratificación por cuartiles. - División 70/30.	- Train acc ≈ **0.73** - Test acc ≈ **0.63**	Buen desempeño para clasificar por rangos.	Menos robusto a variaciones en los datos que un ensamble.	Útil, pero existe una mejor alternativa.
KNN Clasificador	- Escalado con StandardScaler. - One Hot Encoding. - n_neighbors=20.	- Train acc ≈ **0.25** - Test acc ≈ **0.2499**	Predice principalmente la clase mayoritaria. No distingue bien los patrones.	Sufre por la alta dimensionalidad (dummies), dispersión y ruido.	**Modelo descartado**, desempeño pobre.
RNA (Red Neuronal Artificial)	- Escalado Min-Max. - Capa oculta de 64 neuronas + dropout. - 300 épocas, EarlyStopping.	- Test acc ≈ **0.57–0.58**	Capta algunas no linealidades, pero no supera a Random Forest.	Requiere más tuning y más variables (amenidades) para destacar.	Aceptable, no óptimo.
Random Forest Classifier	- One Hot Encoding. - Sin necesidad de escalado. - División 70/30.	- Train acc ≈ **0.766** - Test acc ≈ **0.655**	Mayor estabilidad, menos varianza, captura interacciones complejas.	Requiere más recursos computacionales, pero dentro de límites aceptables.	**Mejor modelo para clasificación** de rangos de precio.

MODELOS DE APRENDIZAJE NO SUPERVISADO

Modelo	Transformaciones aplicadas	Resultados obtenidos	Ventajas observadas	Limitaciones para inmuebles CDMX	Conclusión
K-means	- Escalado (StandardScaler). - Variables usadas: superficie cubierta y precio.	- K óptimo ≈ **6** (método del codo). - Clusters bien formados por rangos de precio-superficie.	Agrupa propiedades similares; útil para segmentación.	Sensible a outliers (aunque ya fueron eliminados). No incorpora coordenadas complejas.	**Adecuado** para segmentación básica del mercado.
PCA	- Escalado global. - Aplicado sobre todas las variables.	- PC1, PC2 y PC3 explican **~32.7%** de la varianza.	Útil para visualización 3D y estructura global.	No sintetiza bien un dataset tan heterogéneo. No sirve para predicción.	Herramienta auxiliar, no modelo predictivo.
DBSCAN	- Escalado de coordenadas. - eps=0.3, min_samples=20.	- Detecta **solo 1 cluster masivo**. - Muy pocos puntos como ruido.	Bueno para datos con zonas claramente separadas.	La CDMX es muy densa, entonces DBSCAN no puede detectar fronteras naturales.	**No recomendable** para inmuebles en CDMX.

DeepKnowledge

DATA CONSULTING