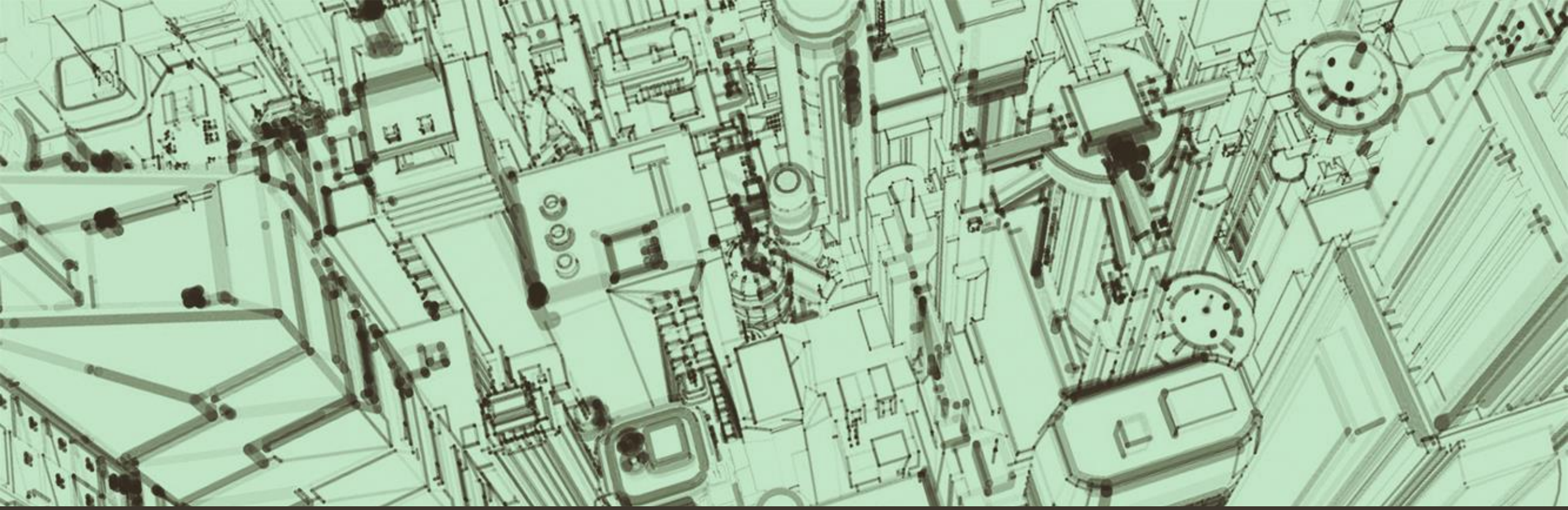


Clasificación y Predicción del Precio de las Propiedades en Bogotá

Natalia Triana Pulido, Sharon Gutierrez Perea y Brandom Alvarez Posada

Introducción y Justificación

Este proyecto analiza datos inmobiliarios de la ciudad de Bogotá con el objetivo de identificar la relación entre la seguridad en las distintas localidades y el valor de las propiedades. Para ello, se integran tres fuentes de datos: propiedades, localidades y tasa de seguridad.

El análisis busca comprender cómo factores sociales, como las tasas de homicidios y hurtos, pueden influir en los precios del mercado inmobiliario, aportando información valiosa para la toma de decisiones en el sector.

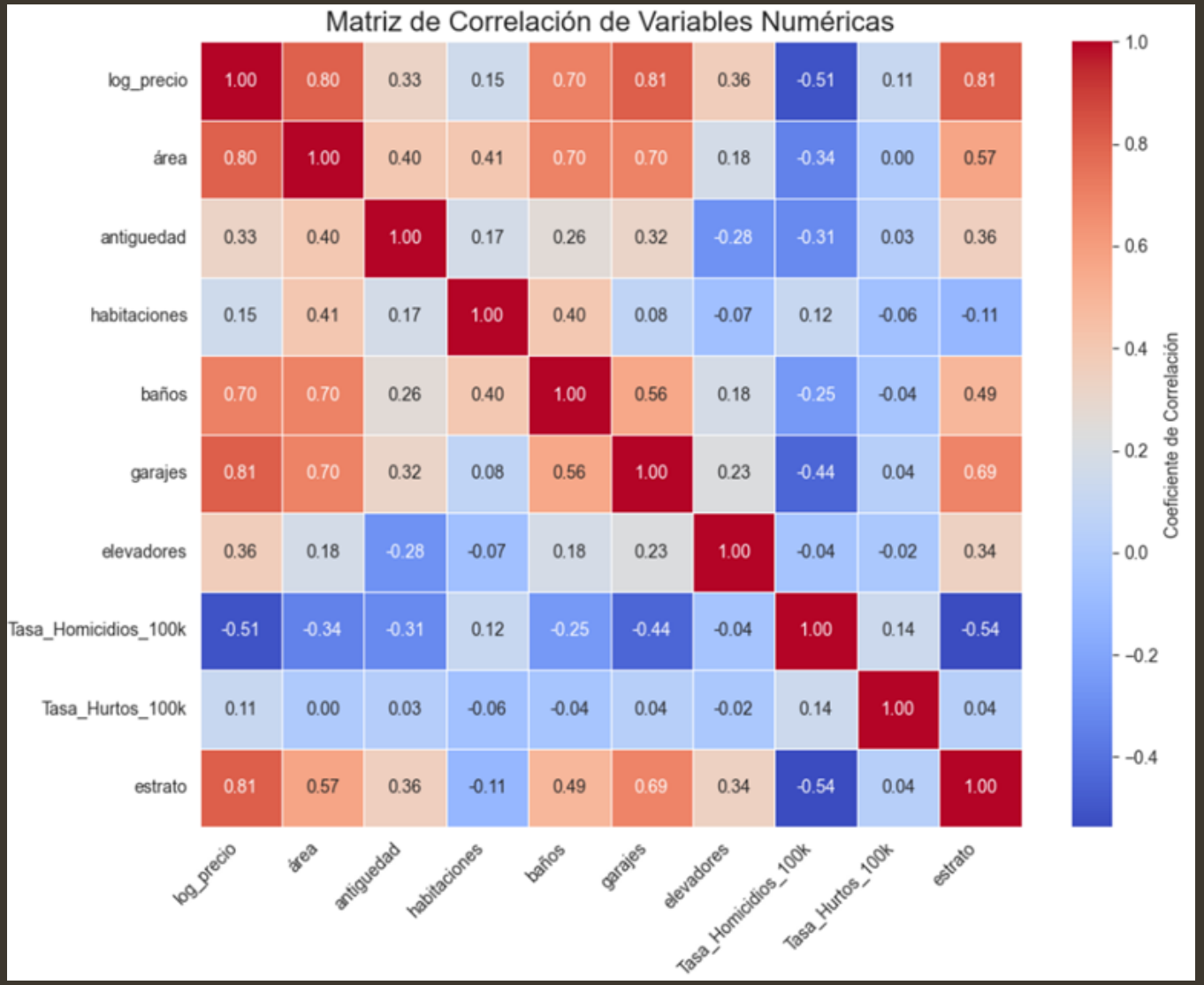
Objetivos del Proyecto

1. Unificar las distintas fuentes de datos en un único conjunto.
2. Limpiar y procesar los datos para garantizar su calidad.
3. Aplicar técnicas de análisis exploratorio para identificar patrones relevantes.
4. Entrenar un modelo de Clustering para clasificar propiedades según su nivel de lujo.
5. Entrenar un modelo de Gradient Boosting (XGBoost) para predecir el precio de las propiedades.
6. Entrenar un modelo de Random Forest Regressor para predecir el precio de las propiedades.

EDA: Hallazgos clave

(Análisis exploratorio de datos)

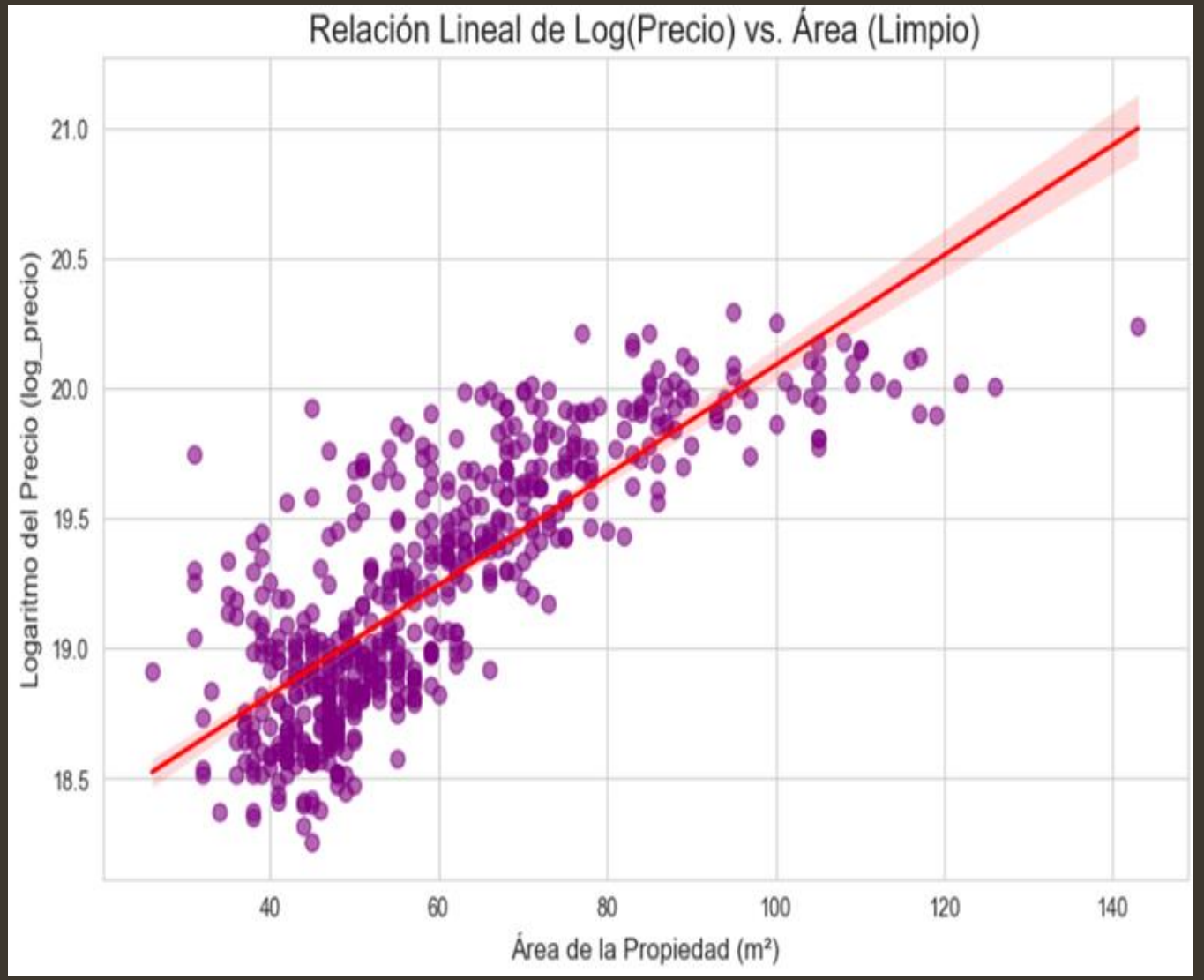
1. Existe una correlación positiva entre el área y el precio de las propiedades.
2. El estrato socioeconómico influye directamente en el valor del inmueble.
3. La cantidad de garajes impacta de forma moderada en el precio final.
4. Las variables área, estrato y garajes se identificaron como las más relevantes para el modelado.



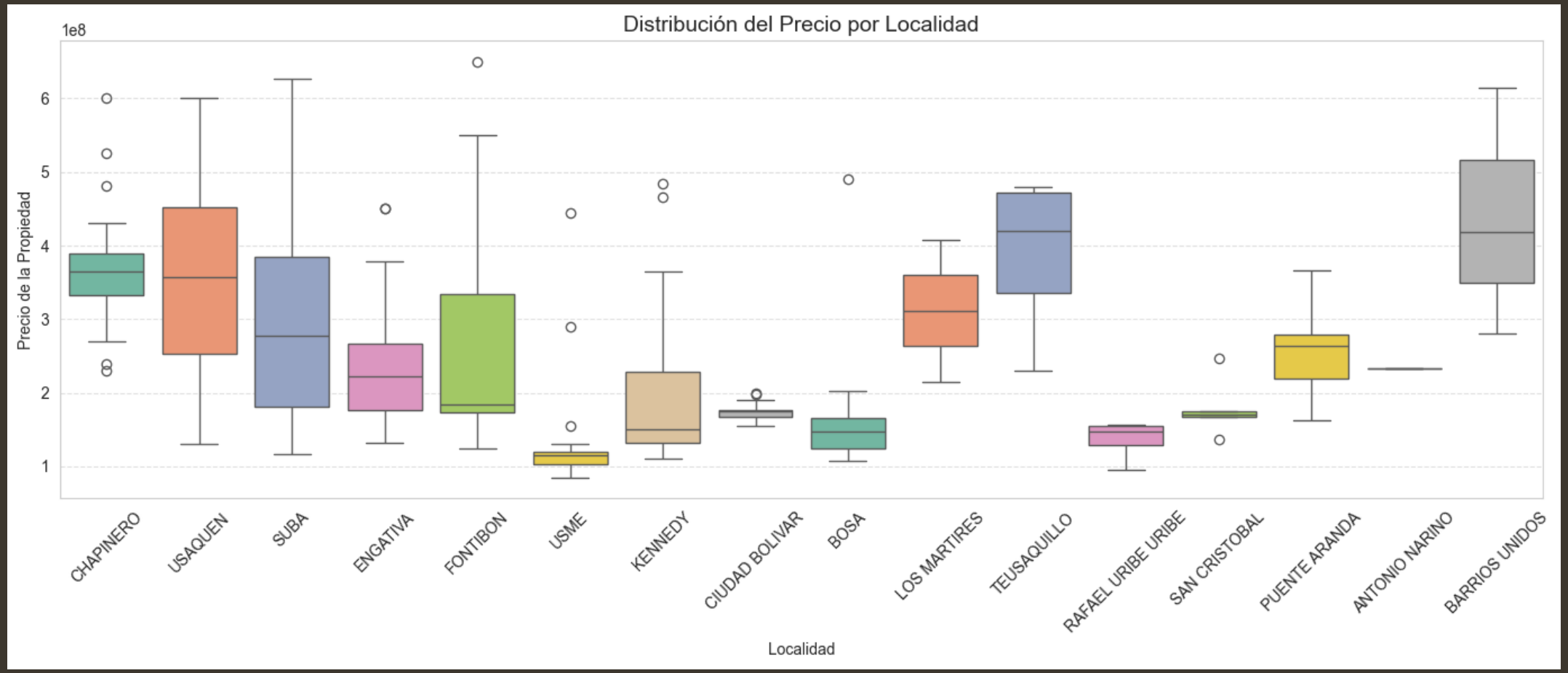
5. Se detectaron Outliers en área y precio, asociados a propiedades de lujo.

- La media disminuyó **$7.8 M**, indica que los outliers eran valores muy altos que inflaban el promedio.
- La dispersión se redujo en **$21.8 M**, mostrando que el conjunto de datos ahora es más homogéneo y menos afectado por valores extremos.

Outliers (Precio)			
Antes		Después	
Media	Media	Desv. Est.	Desv. Est.
$260,780,218.49	$252,950,264.39	$148,218,268.87	$126,397,554.74

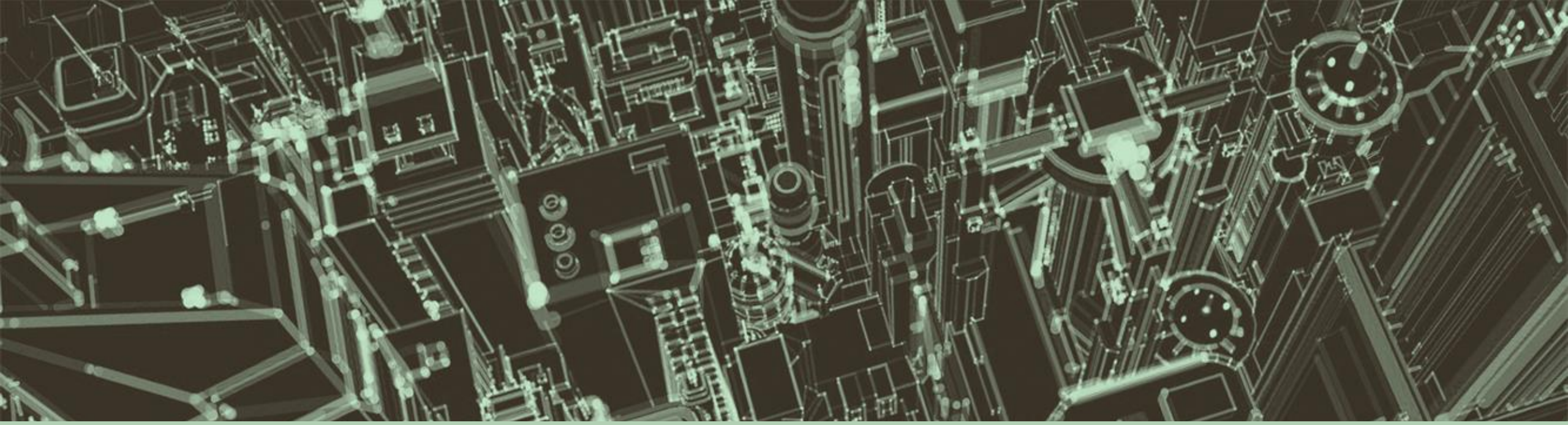


6. Algunas localidades similares presentan diferencias de precio por factores externos como ubicación o seguridad.



Procesamiento de Datos

1. Se identificaron valores nules y dio el respective tratamiento.
2. Se modificaron los nulos de las columnas de gas y remodelado por “No”.
3. Los nulos de la columna de estratos fueron cambiados por el valor correspondiente al barrio.
4. Se añadió el valor de la taza de homicidio y de hurto manualmente a la localidad Antonio Nariño.
5. Las columnas con “Si” y “No” fueron transformadas a valores booleanos.
6. Se tipificaron las columnas: remodelado, deposito, zona_de_lavanderia, gas, parqueadero.
7. No se encontraron duplicados.
8. Se eliminó las columnas: descripción, conjunto, nombre, barrio y dirección.

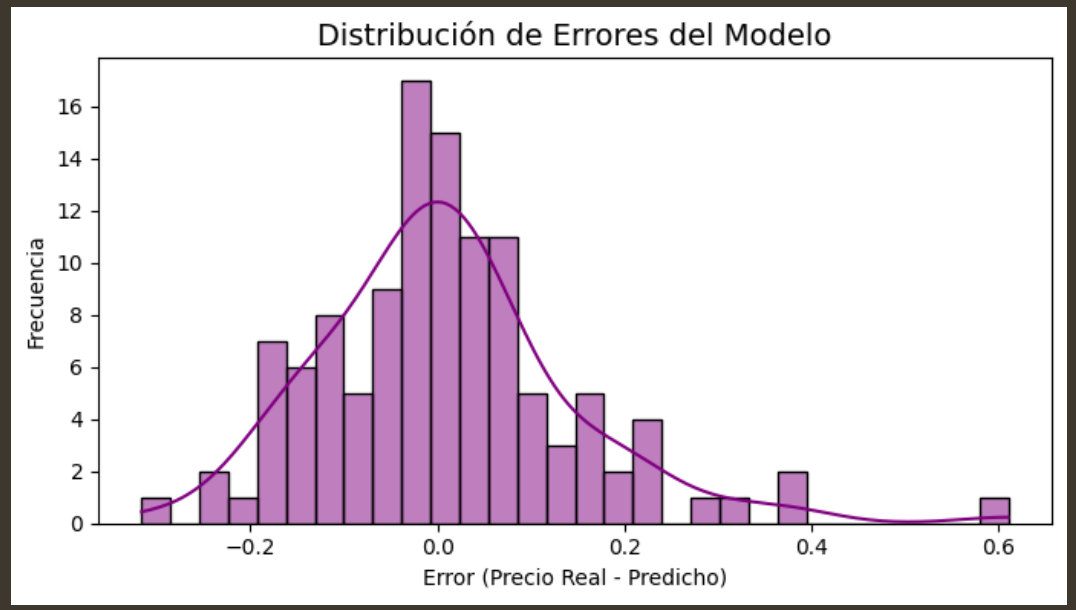


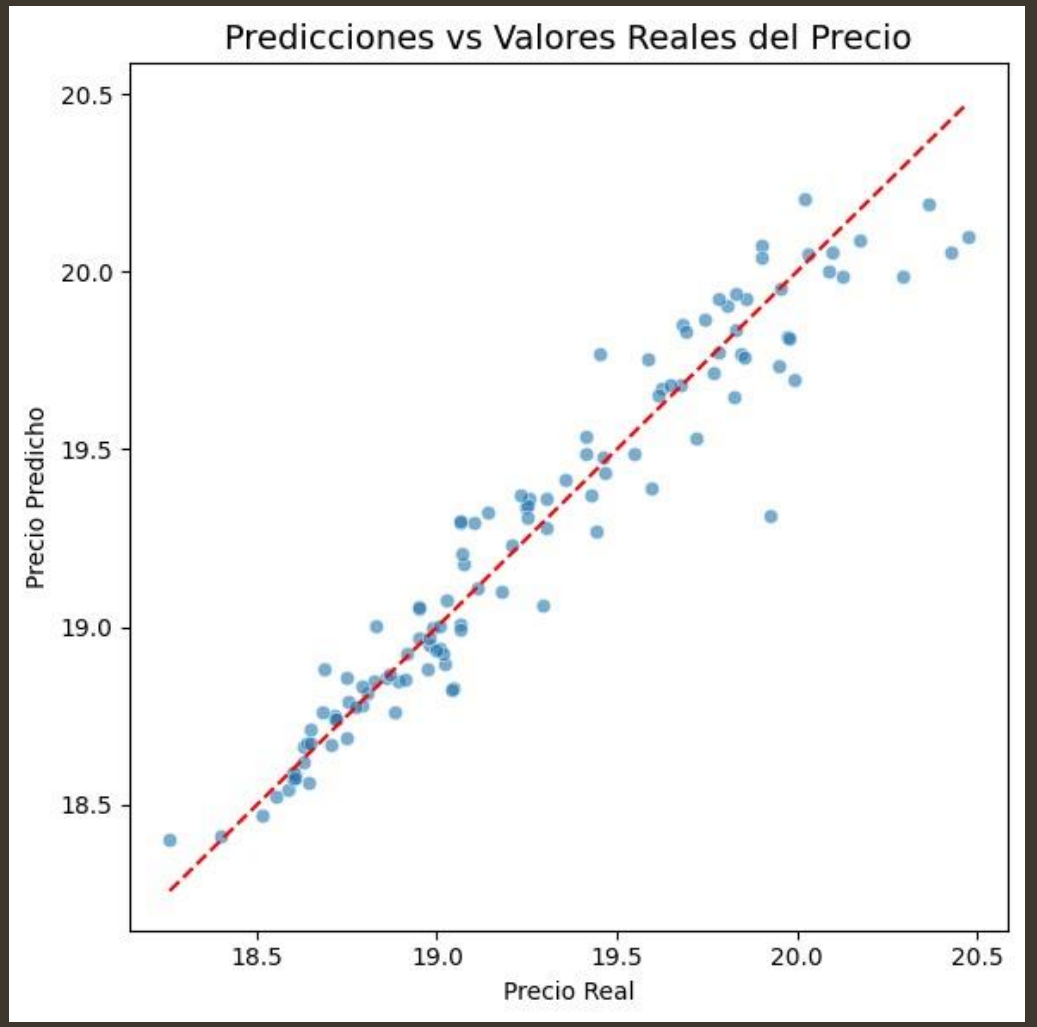
Modelado

1. Random Forest Regressor
2. Gradient Boosting (XGBoost)
3. K-Means Clustering

1. Random Forest Regressor

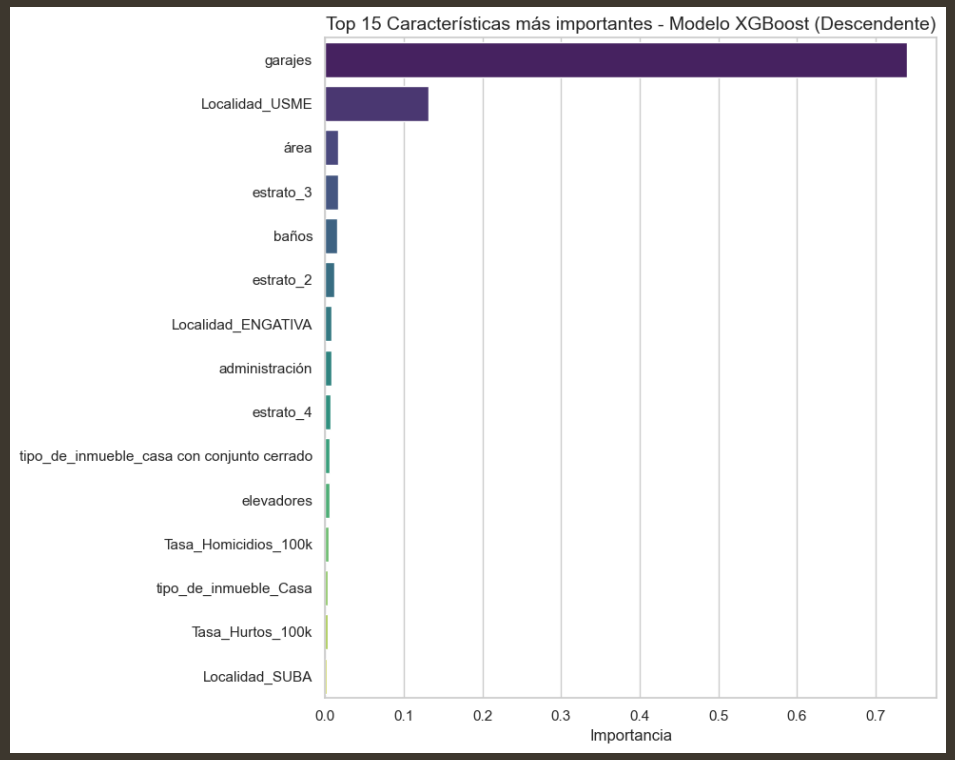
- El modelo alcanzó un R^2 (coeficiente de determinación) de 0.93 y un RMSE (error cuadrático medio) de 0.14, demostrando una alta exactitud en la estimación de precios según las características físicas y de entorno.

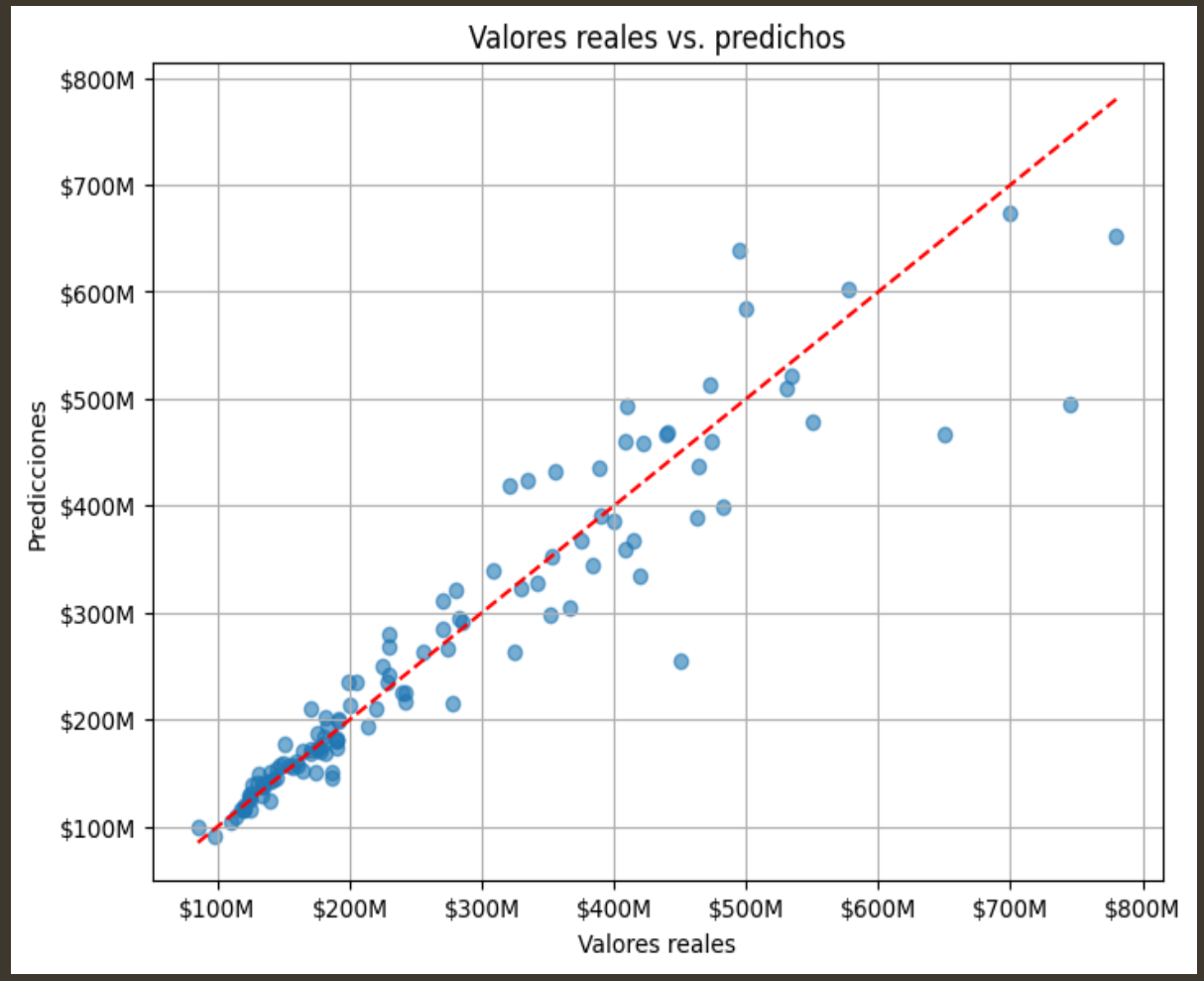




2. Gradient Boosting (XGBoost)

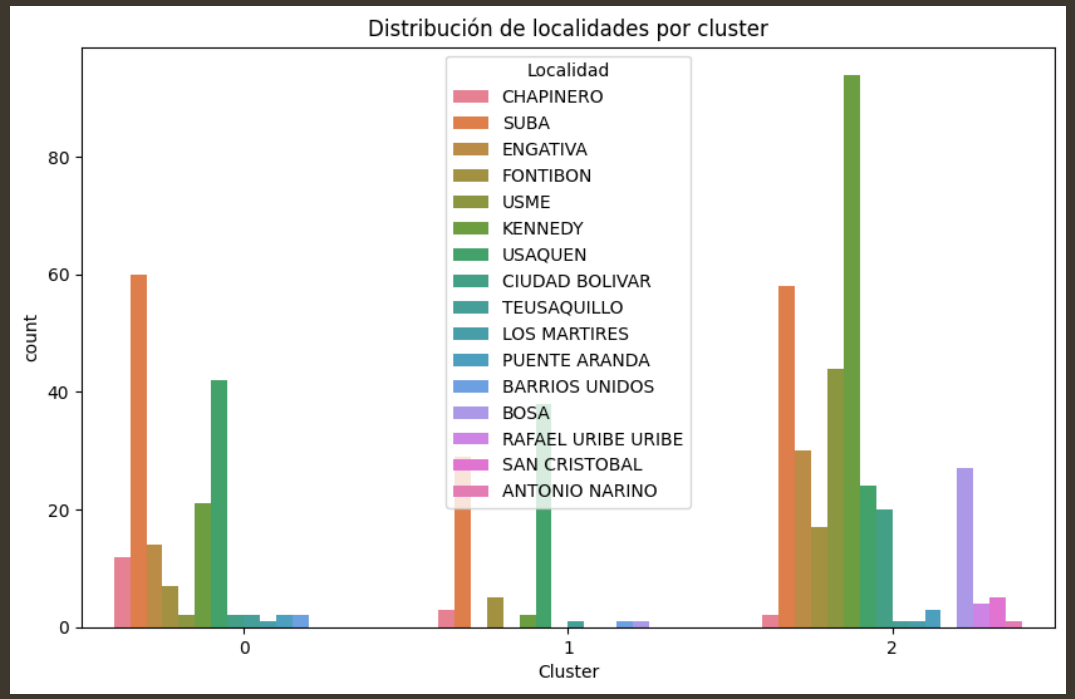
- El modelo alcanzó un R^2 de 0.8905 y un RMSE de 49.7M (9% error) indica un modelo bastante preciso.

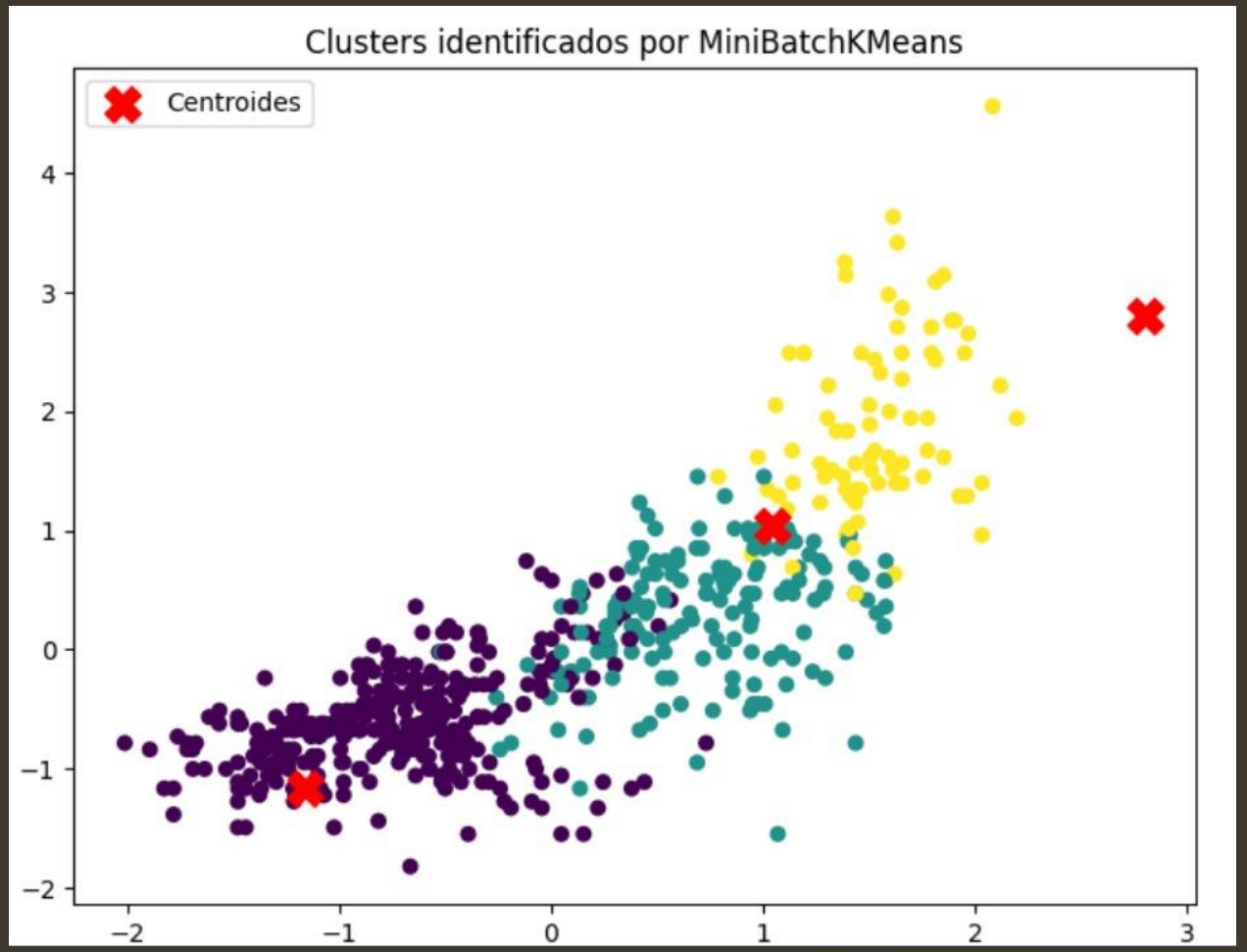




3. K-Means Clustering

- Segmentación efectiva, con buena separación entre grupos (Calinski-Harabasz: 1018.164), baja similitud entre ellos (Davies-Bouldin: 0.816) y cohesión aceptable (Silueta: 0.55).



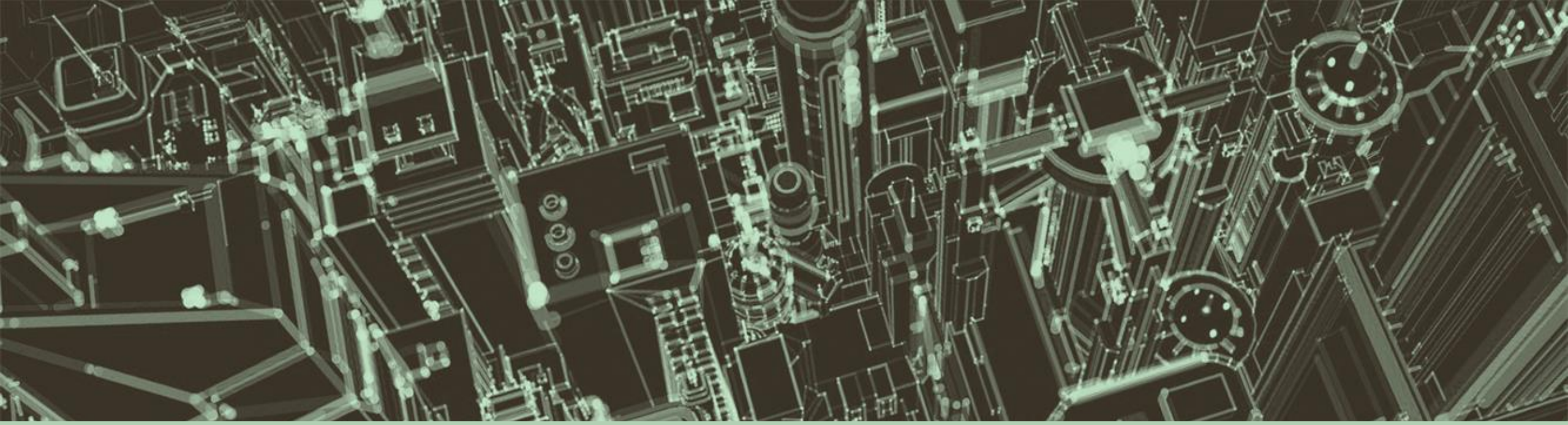


Resultados y Conclusiones

- Los modelos implementados mostraron un desempeño sólido y alineado con los objetivos del proyecto.
- El modelo Random Forest Regressor alcanzó un R^2 de 0.93, demostrando una excelente capacidad predictiva.
- XGBoost ofreció un buen equilibrio entre precisión y eficiencia, siendo útil para datos complejos y heterogéneos.
- K-Means Clustering permitió identificar agrupaciones de propiedades según su nivel de lujo, complementando el análisis predictivo.
- Las variables área, estrato y garajes mostraron un impacto significativo en el valor de las propiedades.
- El proyecto evidenció la viabilidad de aplicar técnicas de machine learning en el análisis inmobiliario, combinando predicción y segmentación para generar información valiosa.

Siguientes Pasos / Futuro del Proyecto

- Incorporar más variables contextuales, como accesibilidad, transporte o cercanía a servicios.
- Optimizar los hiperparámetros de los modelos para mejorar su desempeño.
- Implementar un panel interactivo que permita visualizar las predicciones y clusters en tiempo real.
- Explorar modelos más avanzados, como redes neuronales o ensambles híbridos, para comparar su rendimiento.



¡Gracias!