Universidad Andrés Bello®

PREDICCIÓN DE PRECIOS DE VIVIENDAS MEDIANTE MACHINE LEARNING Y ANÁLISIS GEOESPACIAL AVANZADO

PORTAFOLIO DE PROYECTOS
ENTREGA FINAL

ALONSO DÍAZ STARI
PROFESOR GUIA: PABLO SCHWARZENBERG

UNIVERSIDAD ANDRES BELLO
FACULTAD DE INGENIERÍA
INGENIERÍA CIVIL INFORMÁTICA
SANTIAGO, CHILE
27 - 06 - 2025

CONTEXTO Y OBJETIVO GENERAL DEL PROYECTO

PROBLEMA

Estimar el precio de propiedades es una tarea compleja debido a la alta cantidad de factores involucrados, especialmente la dependencia espacial y contextual no abordada adecuadamente por los modelos tradicionales.

OBJETIVO

Desarrollar un modelo predictivo robusto para estimar precios de departamentos en Santiago, integrando atributos intrínsecos (vía web scraping) y contexto geoespacial mediante grafos de Puntos de Interés (POIs) y embeddings vectoriales por categoría.

MOTIVACIÓN

El mercado inmobiliario es clave para la economía chilena y la estabilidad financiera de hogares e instituciones. Un modelo preciso mejora decisiones de compra, venta y gestión de riesgos, mitigando errores de valoración con potenciales consecuencias económicas graves.

OBTENCIÓN Y PREPARACIÓN DE LOS DATOS

RECOLECCIÓN DE PUBLICACIONES

Se desarrolló un web scraper en Python para extraer URLs de propiedades en venta desde un portal inmobiliario, enfocando posteriormente el análisis en la Región Metropolitana de Santiago.

EXTRACCIÓN DE ATRIBUTOS

Usando las URLs, mediante otro web scraper se obtuvo información detallada de cada propiedad (dormitorios, baños, orientación, ubicación, etc), consolidada en un dataset en formato CSV.

INCORPORACIÓN DE PUNTOS DE INTERÉS (POI) GEORREFERENCIADOS

Se integró información contextual mediante:

- Mapbox API: Extracción de servicios generales cercanos (educación, salud, seguridad, etc.) por coordenadas.
- Transporte Público (IDEOCUC): Se incorporaron paraderos de buses y estaciones de Metro, estandarizando sus coordenadas con PyProj (WGS 84).

LIMPIEZA DE POIS

Se filtraron POIs irrelevantes, se imputaron categorías faltantes y se eliminaron columnas innecesarias, dejando los datos listos para ser usados en el modelo y cargados a una base de grafos.

Datos del Mundo Real

Problema Central: Datos crudos con alta prevalencia de:

01 Inconsistencias y errores de tipeo (ingreso manual de usuarios son validación de los campos)

02 Gran cantidad de datos faltantes

03 Formatos no estandarizados

Consecuencia: La necesidad de un procesamiento y limpieza meticulosa y exhaustiva, a menudo con revisión manual.

GESTIÓN DE DATOS FALTANTES

"La ausencia de datos es información" (Sperrin et al.).

ESTRATEGIAS APLICADAS

- **Eliminación de Columnas:** Para datos excesivamente faltantes o irrelevantes (ej., *gastos_comunes, fecha_publicado*).
- **Imputación por Mediana + Flag Indicadora:** Para atributos clave (*antiguedad, banos, dormitorios*), preservando datos y señalando la ausencia original.
- **Imputación por Cero (0):** Para atributos donde la ausencia significa "no tiene" (*terraza, estacionamiento, bodegas*).
- **Estandarización y One-Hot Encoding:** Para categóricas (*orientacion, tipo_depto*), manejando variaciones y la ausencia explícitamente.

LIMPIEZA Y OUTLIERS EN ATRIBUTOS NUMÉRICOS CLAVE

COLUMNAS:

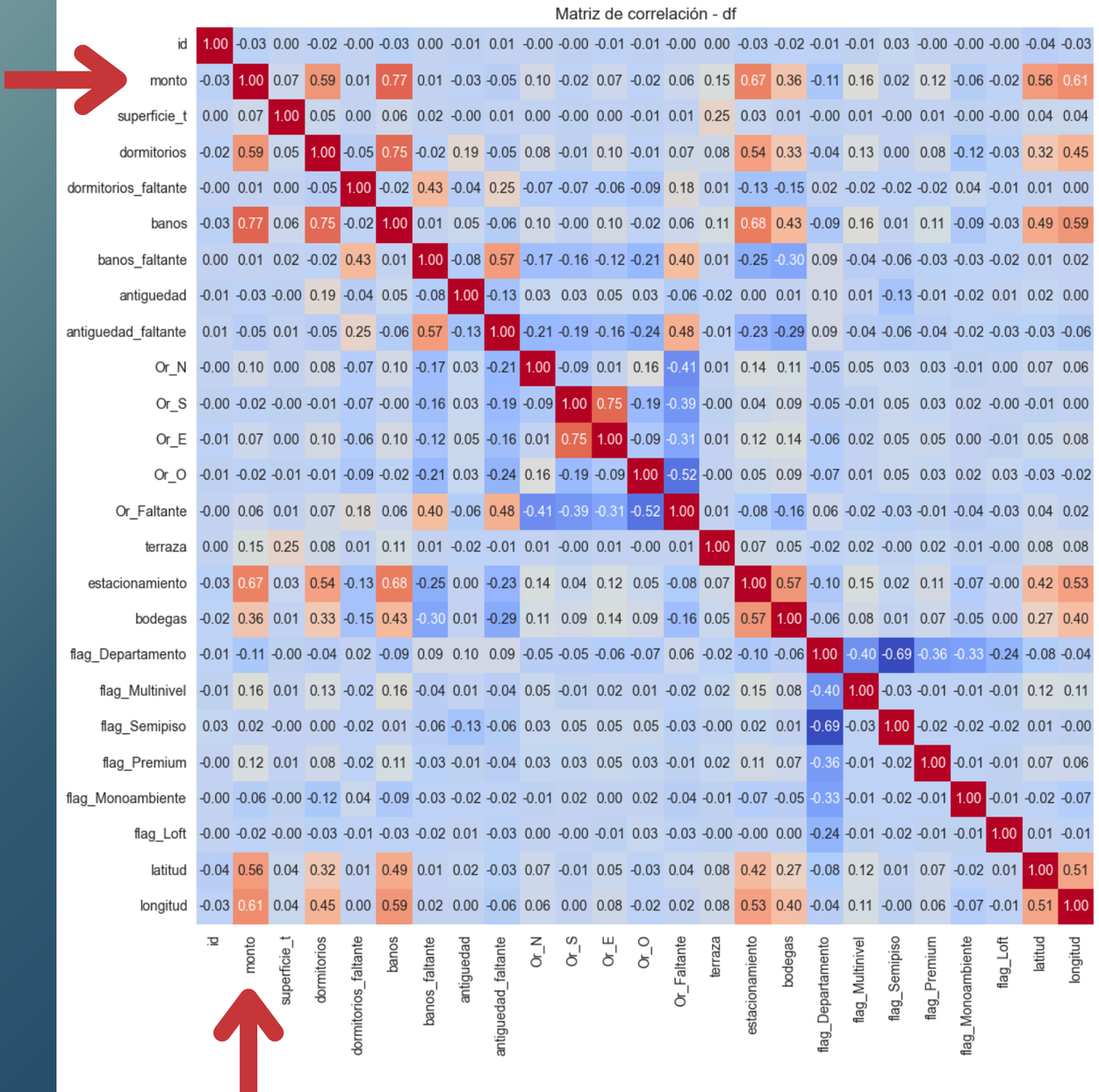
- superficie_t

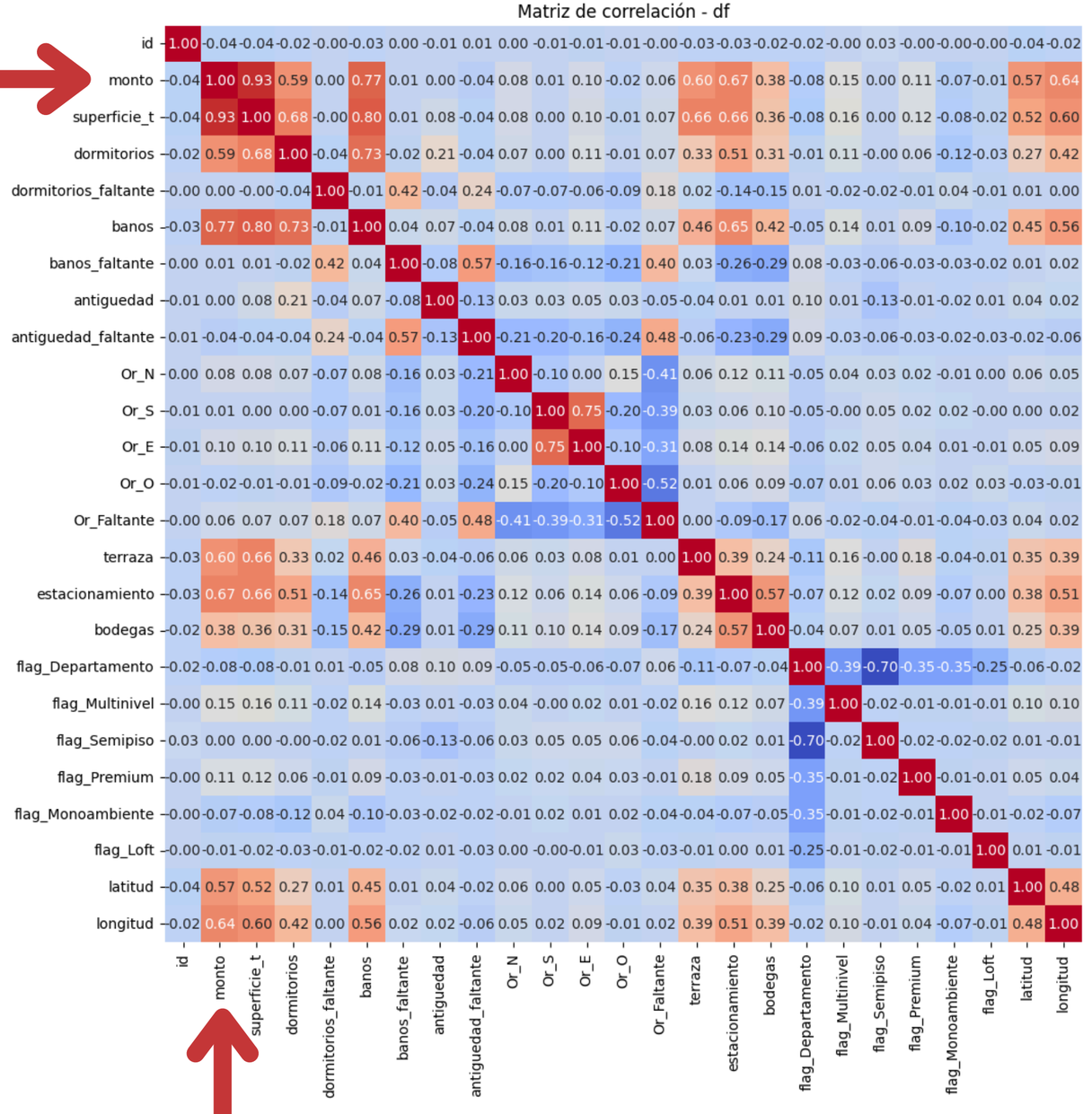
- Eliminación de outliers extremos (>400m^2, <15m^2) por errores de ingreso o no representatividad de un departamento (ej. edificios completos).
- Justificación: Evitar sesgos y mejorar la predicción en el segmento de mercado principal.

- terraza

- Eliminación de outliers (>500m^2).
- Filtrado contextual: >200m^2 solo en comunas de lujo (Lo Barnechea, Vitacura, Las Condes, Peñalolén).
- Justificación: Coherencia con el mercado inmobiliario y eliminación de errores.

EFECTOS DE LA LIMPIEZA





EFECTOS DE LA LIMPIEZA

ANTES

- Los datos brutos presentaban correlaciones débiles o incluso ilógicas, lo que indicaba una alta presencia de ruido y errores. Era muy difícil discernir relaciones significativas entre las características de las propiedades y su precio.

DESPUES

- Tras la limpieza, imputación de datos y eliminación de outliers, las correlaciones se volvieron lógicas, fuertes y coherentes con las expectativas del mercado inmobiliario. La calidad de los datos es fundamental para un modelado efectivo.

OBSERVACIÓN ADICIONAL

- Aumento de colinealidad entre superficie_t, dormitorios, banos.
- Patrones en las _faltante flags (ej., banos_faltante y antiguedad_faltante = 0.57, sugiriendo patrones de ingreso).

DATASETS

Departamentos

id	25215	non-null	int64
monto	25215	non-null	int64
superficie_t	25215	non-null	float64
dormitorios	25215	non-null	int64
dormitorios_faltante	25215	non-null	int64
banos	25215	non-null	int64
banos_faltante	25215	non-null	int64
antiguedad	25215	non-null	int64
antiguedad_faltante	25215	non-null	int64
Or_N	25215	non-null	int64
Or_S	25215	non-null	int64
Or_E	25215	non-null	int64
Or_O	25215	non-null	int64
Or_Faltante	25215	non-null	int64
terraza	25215	non-null	float64
estacionamiento	25215	non-null	int64
bodegas	25215	non-null	int64
flag_Departamento	25215	non-null	int64
flag_Multinivel	25215	non-null	int64
flag_Semipiso	25215	non-null	int64
flag_Premium	25215	non-null	int64
flag_Monoambiente	25215	non-null	int64
flag_Loft	25215	non-null	int64
latitud	25215	non-null	float64
longitud	25215	non-null	float64

POIs

longitude	39052	non-null	float64
latitude	39052	non-null	float64
type	39052	non-null	object
filterrank	39052	non-null	int64
category_en	34315	non-null	object
sizerank	39052	non-null	int64
maki	39052	non-null	object
class	39052	non-null	object
name	33257	non-null	object

Estaciones de metro

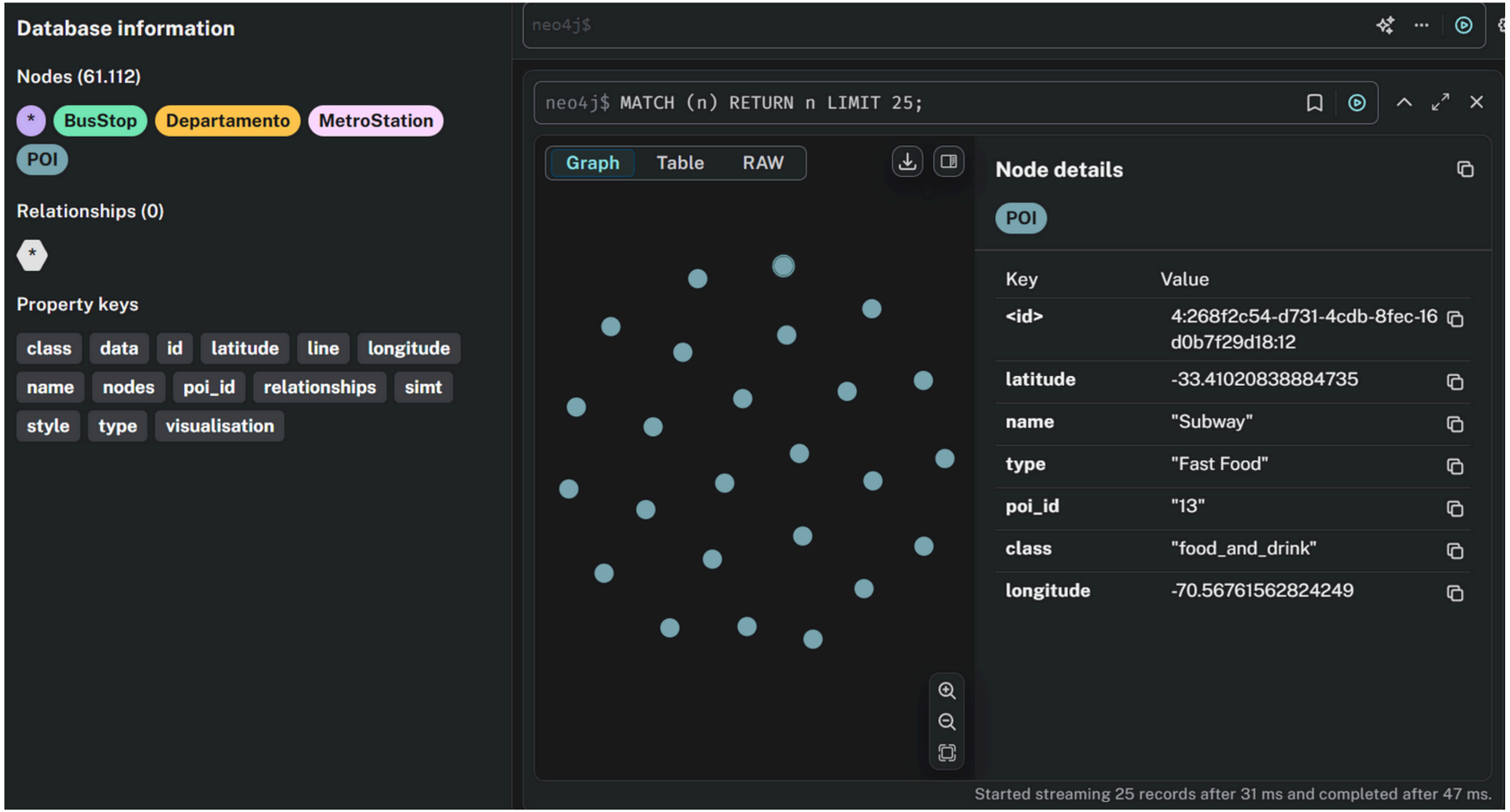
id	117	non-null	int64
nombre	117	non-null	object
tipo	117	non-null	object
linea	117	non-null	object
longitud	117	non-null	float64
latitud	117	non-null	float64

Paraderos

id	12635	non-null	int64
simt	11774	non-null	object
longitud	12635	non-null	float64
latitud	12635	non-null	float64

CARGA EN NEO4J AURA

61.122
NODOS
CARGADOS



AVANCES LOGRADOS Y PRÓXIMOS PASOS

ESTADO ACTUAL

- Datasets de Propiedades y POIs preprocesados y limpios.
- Nodos cargados en Neo4j Aura
- Datos listos para la conexión de los nodos del grafo

PRÓXIMOS PASOS CRUCIALES

- Construcción de Grafo: Representación de propiedades y POIs en Neo4j.
- Generación de Graph Embeddings.
- Desarrollo y Evaluación de Modelos Predictivos: Construcción de modelos de regresión para predecir precios.
- Iteración y Optimización.

GRACIAS