

DATOS Y FINANZAS

I SEMINARIO SOBRE EL IMPACTO DE LAS TECNOLOGÍAS DE BIG DATA EN LAS FINANZAS

Organizado por los institutos de investigación IADCOM e IGEDECO
Comisión de doctorado subárea FINANZAS
Maestrías de Finanzas y de Gestión y Análisis de Datos

Valuaciones inmobiliarias e Inteligencia Artificial

Diseño de un modelo de valuación de aprendizaje automático
para la Ciudad de Buenos Aires

Por
Mg. Ignacio Sofia

ignaciosofia@gmail.com

INTRODUCCIÓN

Los desarrollos inmobiliarios son:

Inversiones a largo plazo

Con lenta tasa de repago

Grandes costos hundidos

Involucran riesgos sobre:

Demanda

Tierra

Costo de la construcción

PROPTECH



Compañías inmobiliarias que incorporan los avances tecnológicos para optimizar procesos y recopilar datos.

Este trabajo busca plantear la base teórica para realizar un modelo de valuación del metro cuadrado en la Ciudad de Buenos Aires incorporando las técnicas actuales de Análisis de Datos y con una dimensión de Machine Learning; logrando brindar a los participantes del sector no sólo una valuación presente del precio por metro cuadrado, sino también anticipar tendencias en el movimiento de estos valores.

Tendría una doble finalidad:

- Se podría, por un lado, saber el valor de un inmueble en particular usando todos los comparables posibles.
- Pero también resultará útil para analizar tendencias y predecir movimientos de mercado. El modelo podrá “aprender” de la información relevante para autoajustarse de manera cada vez más precisa y anticipar subas o bajas de precio.

TEORÍA DE LAS TASACIONES

En Argentina, el Tribunal de Tasaciones de la Nación fue designado como organismo rector para la tasación de bienes (Ley 21.626).

Tres métodos de valoración de inmuebles:

A

Integración de Costos Parciales

B

Capitalización de Rentas

C

Comparación de valores
venales

BIG DATA Y MACHINE LEARNING EN EL MERCADO INMOBILIARIO

Big Data, más que un concepto, es una disciplina que busca extraer valor de los datos almacenados, formulando predicciones a través de los patrones observados. Trabaja con conjuntos de datos tan grandes y complejos que precisa de aplicaciones informáticas dedicadas para procesar la información adecuadamente.

El aprendizaje automático o **Machine Learning** es una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras, a través de la contrastación con la información disponible, vayan ajustando sus modelos de cálculo, mejorando su desempeño. Es decir, logrando programas de predicción cada vez más precisos.

MACHINE LEARNING vs MODELO HEDÓNICO

El uso de metodologías de aprendizaje automático para estimar los precios de los inmuebles presenta varias ventajas:

No requieren que los datos de entrenamiento estén distribuidos normalmente.

Ahora pueden completarse rápidamente con la potencia y velocidad de las computadoras modernas.

Los errores de estimación pueden reducirse al mínimo gracias a los conjuntos de datos de entrenamiento y de prueba.

ALGORITMOS SUPERVISADOS

- Los algoritmos de aprendizaje supervisados pretenden identificar una función que sea capaz de ofrecer previsiones precisas de la muestra. Un algoritmo de aprendizaje automático buscará una función $F(x)$ que produzca la menor pérdida de predicción esperada, por ejemplo regresión lineal y la máquina de vectores de apoyo.

NO SUPERVISADOS

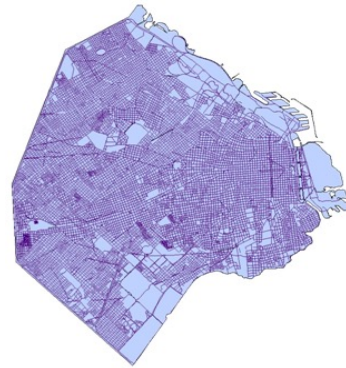
- Los métodos de aprendizaje no supervisado se aplican para descubrir las relaciones entre los elementos de un conjunto de datos sin etiquetas. Debido a esta naturaleza, puede descubrir estructuras ocultas dentro de los datos que no se pueden observar fácilmente. Solucionan problemas de agrupación y asociación, por ejemplo k-means y redes neuronales.

SEMISUPERVISADOS

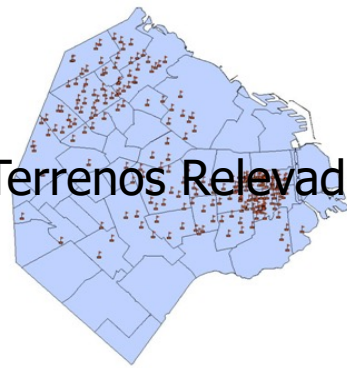
- El aprendizaje semisupervisado se refiere al caso en que hay una gran cantidad de datos de entrada (X) y sólo algunos de los datos están etiquetados (Y). Los científicos de datos pueden adoptar técnicas de aprendizaje supervisado para hacer predicciones sobre los datos no etiquetados, entrenar el modelo con un algoritmo de aprendizaje supervisado y hacer predicciones sobre los datos de prueba.

Mapas georreferenciados de la Ciudad en ArcGis

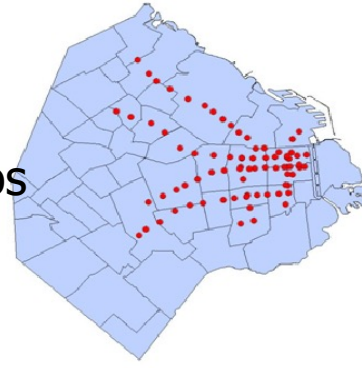
La Ciudad



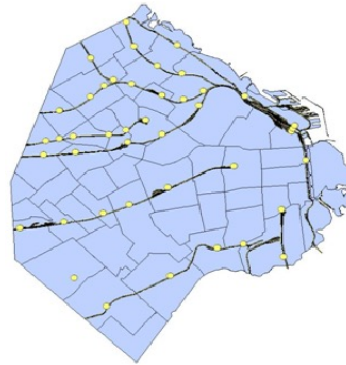
Terrenos Relevados



Estaciones de Subte



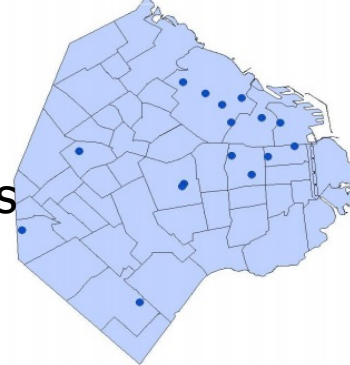
Estaciones de ferrocarril



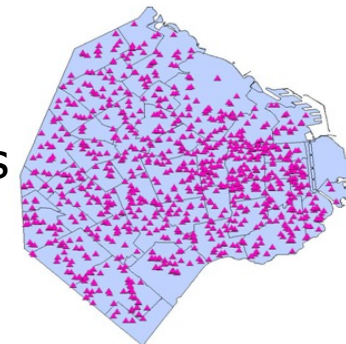
Clínicas y Hospitales



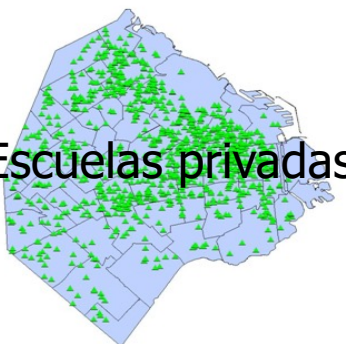
Shoppings



Escuelas públicas



Escuelas privadas



Comisarías

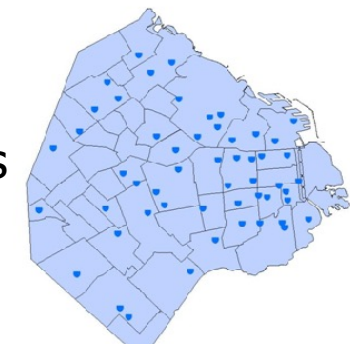


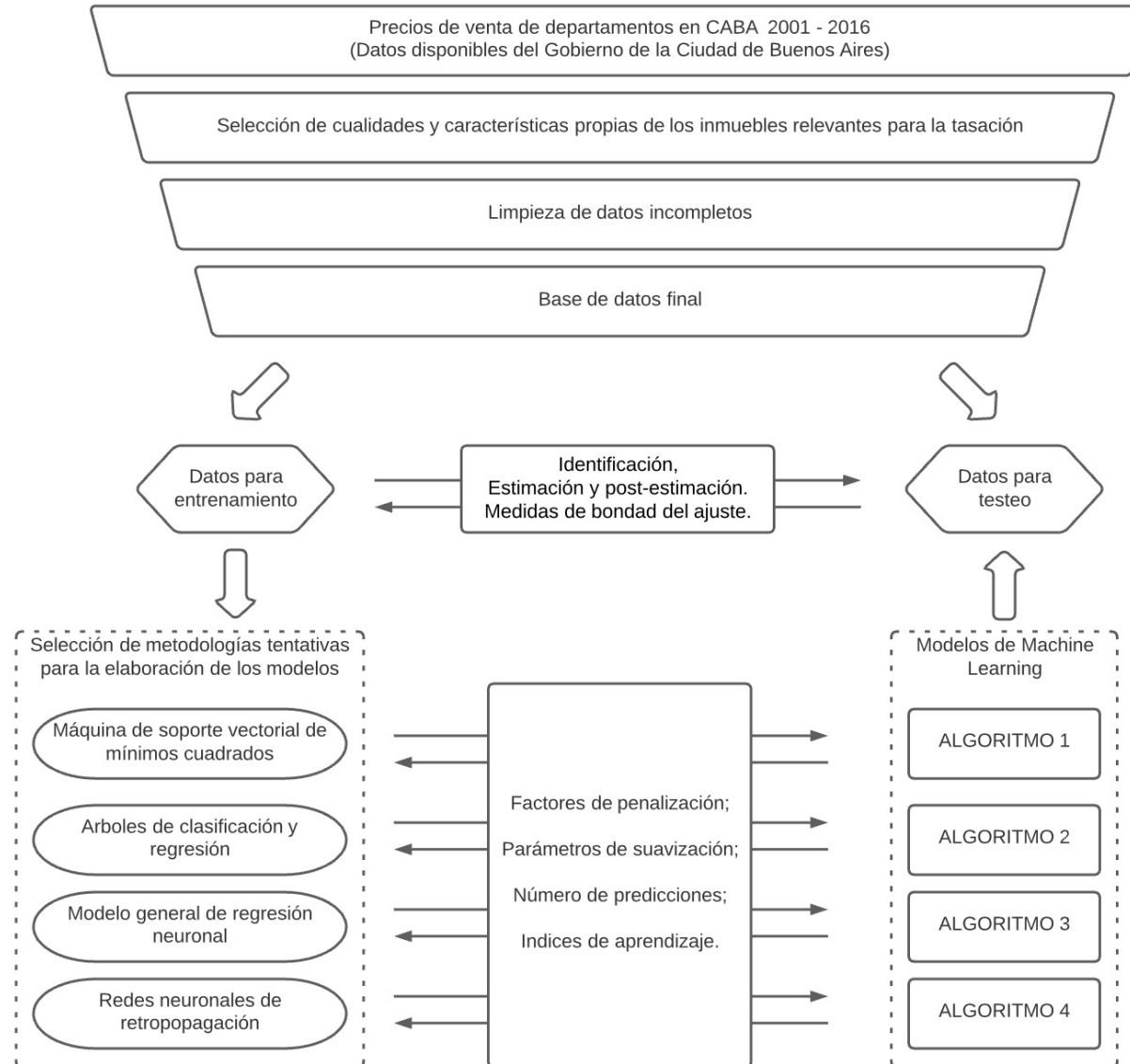
Tabla resumen de las variables incorporadas al modelo

ln inc	SHOP	HOSP	FFCC
Min. :4.927	Min. : 0.000713	Min. : 0.000452	Min. : 0.000520
1st Qu.:5.799	1st Qu.: 0.011637	1st Qu.: 0.006710	1st Qu.: 0.006312
Median :6.006	Median : 0.018556	Median : 0.010285	Median : 0.009913
Mean :6.003	Mean : 0.125617	Mean : 0.117761	Mean : 0.117315
3rd Qu.:6.161	3rd Qu.: 0.025634	3rd Qu.: 0.014661	3rd Qu.: 0.013723
Max. :6.992	Max. :23.660088	Max. :23.743510	Max. :23.750412

SUBTE	ESCPU	ESCPR	POLI
Min. : 0.000972	Min. : 0.000036	Min. : 0.000152	Min. : 0.000295
1st Qu.: 0.003149	1st Qu.: 0.001160	1st Qu.: 0.001097	1st Qu.: 0.003859
Median : 0.005672	Median : 0.001890	Median : 0.001645	Median : 0.006151
Mean : 0.116873	Mean : 0.109147	Mean : 0.109076	Mean : 0.113664
3rd Qu.: 0.012305	3rd Qu.: 0.002781	3rd Qu.: 0.002392	3rd Qu.: 0.008504
Max. :23.752938	Max. :23.733724	Max. :23.742594	Max. :23.743327

z1	z2	z3	z4
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.0000
Mean :0.2511	Mean :0.2511	Mean :0.2511	Mean :0.2466
3rd Qu.:0.5000	3rd Qu.:0.5000	3rd Qu.:0.5000	3rd Qu.:0.0000
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000

DESARROLLO DEL PROGRAMA PREDICTIVO



Resultados del modelo Maquina de Soporte Vectorial de Mínimos Cuadrados

Residuals:

Min	1Q	Median	3Q	Max
-0.51352	-0.08200	0.00466	0.07824	0.54101

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.46494	0.03667	176.278	< 2e-16	***
Z1		-0.86644	0.03038	-28.522	< 2e-16 ***
Z2		-0.53525	0.03145	-17.021	< 2e-16 ***
Z3		-0.37258	0.02905	-12.827	< 2e-16 ***
SHOP	-3.16424	0.70337	-4.499	1.13e-05	***
HOSP	7.24888	1.91944	3.777	0.000207	***
FFCC	-3.76885	2.32764	-1.619	0.106897	
SUBTE	-0.42327	1.14593	-0.369	0.712220	
ESCPU	17.71031	6.47321	2.736	0.006748	**
ESCPR	-19.05327	6.70472	-2.842	0.004924	**
POLI	1.44430	3.02581	0.477	0.633620	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1397 on 212 degrees of freedom

Multiple R-squared: 0.8521, **Adjusted R-squared: 0.8451**

F-statistic: 122.1 on 10 and 212 DF, **p-value: < 2.2e-16**

CONCLUSIONES

El modelo en general es significativo y explica más del 84% de la variación de precios.

Las variables de Zona, tienen relevancia.

La cercanía a hospitales, centros de entretenimiento y educación también influyen en el precio.

En cambio, la cercanía a medios de transporte y a comisarías no tuvieron relevancia estadística.

El aporte más significativo en esta investigación es que abre la puerta a un futuro desarrollo de un modelo de valuación utilizando inteligencia artificial.

En una futura investigación, tomaremos la totalidad de los inmuebles indexados por el Gobierno de la Ciudad. Separaremos el 70 % de los datos para entrenamiento y el 30% para testeo.