PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Predicción del Valor del Suelo en el Gran Santiago Mediante la Correlación de Características de Uso de Suelo

Cristián Canales - Tomás Sepúlveda - Sebastián Pérez

Introducción

En el siguiente trabajo, estudiaremos las correlaciones entre el modo en que las comunas del Gran Santiago usan el suelo, para predecir el valor del metro cuadrado del suelo. Empleamos distintas técnicas de aprendizaje supervisado y no supervisado para el estudio y cuantificación de la relación entre los distintos usos de suelo existentes. Además, la diferenciación o clasificación posible entre las diferentes comunas propias del estudio.

Nuestra base de datos es privada (no publicada), construida mediante datos públicos. La base de datos se genera a través del cruce de la zonificación EOD, proveniente de la encuesta origen y destino realizada por la Subsecretaría de Transportes (SECTRA) y la información relacionada con predios y líneas de construcción por uso de suelo del Servicio de Impuestos Internos (SII).

La base de datos contiene las 32 comunas que componen el Gran Santiago, divididas en sus respectivas zonas de estudio. Cada comuna muestra el total de metros cuadrados construidos de Comercio, Educación, Residencia, Industria, Servicios y Otros Usos.

Además, cuenta con el total de hogares diferenciados por grupos socioeconómicos. Si bien, la información presente no está actualizada, permite abordar las dinámicas de usos de suelo con datos reales. De este modo, facilita resolver el problema del presente trabajo.

Objetivos

Objetivo general: Cuantificar y modelar la relación entre diferentes usos de suelo en el Gran Santiago para predecir el valor del metro cuadrado, utilizando diferentes técnicas de aprendizaje supervisado.

Objetivos específicos:

- Explorar y limpiar la Base de Datos para su correcta implementación.
- Observar la distribución de los diferentes usos de suelo de las comunas.
- Medir la fuerza de las relaciones entre los distintos usos.
- Mostrar características, patrones y dinámicas comunes.

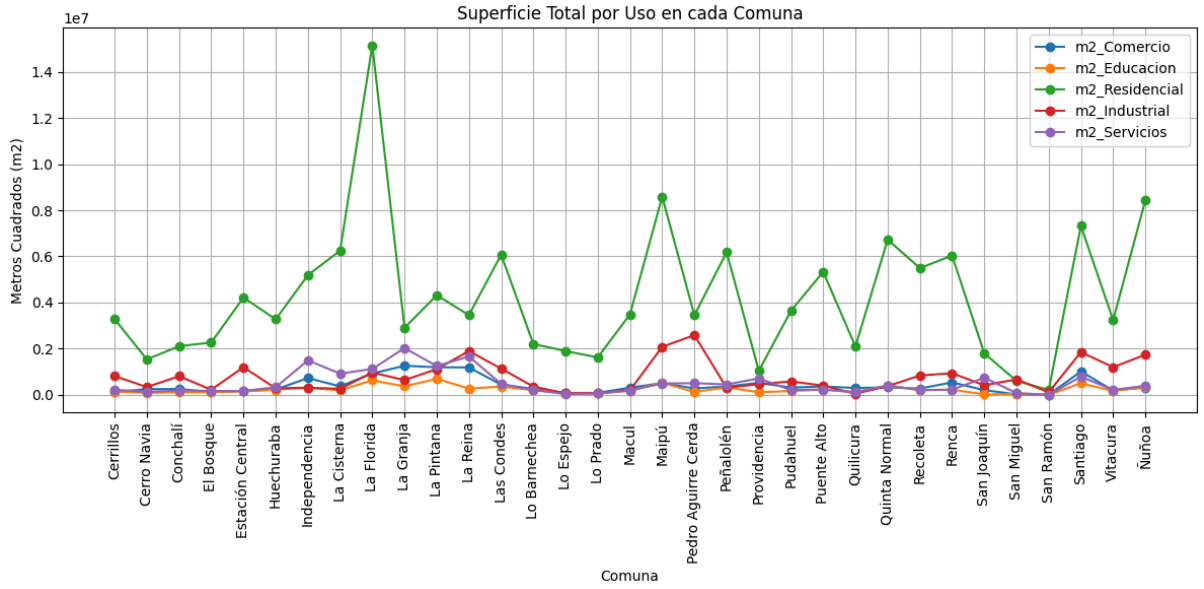
Resultados

1. Análisis exploratorio

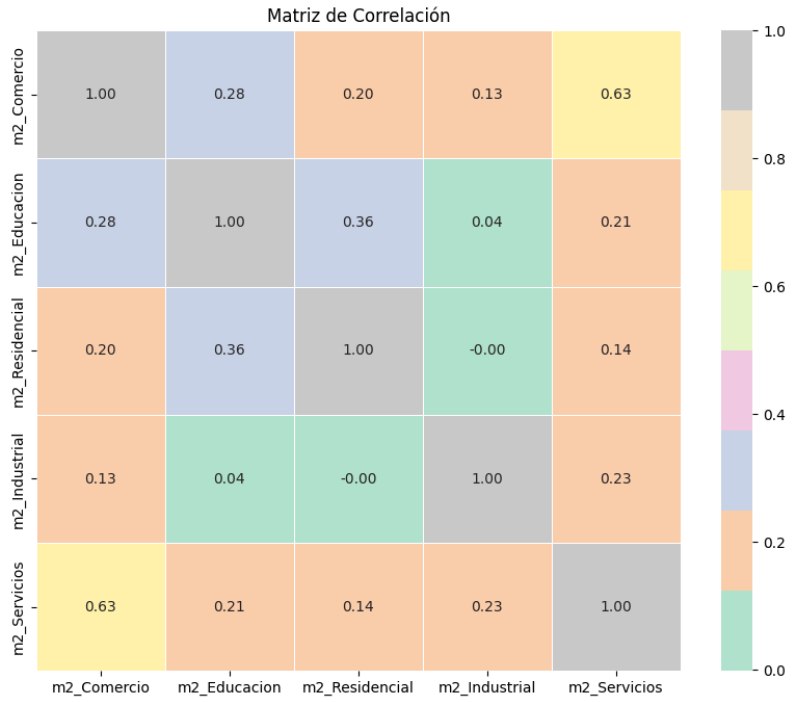
La Siguiente tabla resume la información contenida en el dataset

Columnas	Elementos No Nulos	Tipo de Dato
Zona_EOD	695	int64
m2_Comercio	695	float64
m2_Educacion	695	float64
m2_Residencial	695	float64
m2_Industrial	695	float64
m2_Servicios	695	float64
m2_Otros	695	float64
GSE Alto	695	float64
GSE Medio Alto	695	float64
GSE Medio	695	float64
GSE Medio bajo	695	float64
GSE Bajo	695	float64
Ed. Básica y Media	695	float64
Ed. Superior	695	float64
UF/m2	243	float64
AREA	695	float64
Comuna	695	object
Perimeter	695	float64

De este modo, generamos el siguiente gráfico que compara los metros cuadrados destinados a Comercio, Educación, Residencias, Industrias y Servicios en las 32 comunas del Gran Santiago.



Observamos que los metros cuadrados residenciales dominan de forma homogénea en todas las comunas, constituyendo el principal uso de suelo. En este sentido, **destaca La Florida**, comuna que duplica a sus pares en el uso de suelo residencial. Otro aspecto significativo, es el fenómeno que se observa en La Granja, Pedro Aguirre Cerda, Providencia, San Miguel y San Ramón, **donde coexisten los usos industriales y residenciales de forma evidente.** Esto será analizado en detalle al agrupar las comunas según sus factores determinantes (Punto 1.3: Análisis de Cluster).



Las columnas con mayor correlación **positiva** son: 'm2_Comercio' y 'm2_Servicios', con un Coeficiente de Pearson de 0.6347. Por su parte, las columnas con mayor correlación **negativa** son: 'm2_Residencial' y 'm2_Industrial' con un Coeficiente de Pearson de: -0.0012.

2. Técnicas de aprendizaje Supervisado

Dado el objetivo de medir la fuerza de las relaciones entre los distintos usos y dar la base para poder predecir cambios urbanos según las dinámicas existentes entre usos, nos proponemos ejecutar un modelo de regresión lineal múltiple para cada tipo de uso de suelo. **Buscamos lograr la selección del mejor modelo posible utilizando cada uno de los 6 usos de suelo diferentes como variables dependientes (residencial, por ejemplo) y asignando el resto de los otros usos como variables explicativas.** Por lo tanto, se plantea la idea de encontrar 6 modelos de regresión lineal múltiple, uno que explique cada uno de los distintos usos de suelo, para luego utilizar sus respectivos coeficientes como medida cuantificadora de la fuerza y dirección en la relación de usos.

En el desarrollo de la Ingeniería de Solución tendremos tantos modelos como usos de suelos existan, ya que se buscará la explicación de cada uno de ellos en función de los otros.

$$Comercial = f(Educación, Residencial, Industrial, Servicios, Otros)$$
$$Educación = f(Comercial, Residencial, Industrial, Servicios, Otros)$$
$$Residencial = f(Comercial, Educación, Industrial, Servicios, Otros)$$
$$Industrial = f(Comercial, Educación, Residencial, Servicios, Otros)$$
$$Servicios = f(Comercial, Educación, Residencial, Industrial, Otros)$$

$$Otros = f(Comercial, Educación, Residencial, Industrial, Servicios)$$

El peso de las variables independientes será considerado como **el coeficiente normalizado (beta)** de la regresión lineal múltiple, ya que estos coeficientes determinan la importancia de cada variable independiente dentro del modelo de regresión. Por lo tanto, **el coeficiente con mayor valor será el que tenga el más alto poder de explicación de la variable dependiente.**
Además, estos coeficientes están liberados de sus unidades y varianzas, por lo que son directamente comparables.
Es importante mencionar que **se utiliza el método "backward stepwise"** el cual considera una interacción para incorporar en una primera instancia todas las variables, para luego ir sacando y comparando las diferencias en el rendimiento del modelo mediante diferentes métricas, para la cual se selecciona al R2 ajustado como la más indicada para la elección del modelo.

Los resultados obtenidos son los siguientes:

Usos de Suelo	Comercio	Educación	Residencial	Industrial	Servicios	Otros	R2 Ajustado	Constante
Comercio	x	0,26	0,01	-0,03	0,33	0,06	**0,43**	**6280,55**
Educación	0,09	x	0,03	x	0,01	x	**0,17**	**3628,20**
Residencial	0,62	4,07	x	-0,16	x	0,27	**0,14**	**144413,67**
Industrial	-0,20	0,18	-0,02	x	0,21	1,36	**0,57**	**9773,17**
Servicios	1,11	0,15	x	0,09	x	x	**0,42**	**-2959,11**
Otros	0,10	-0,13	0,01	0,40	x	x	**0,56**	**3235,62**

La x representa las variables no utilizadas por el modelo para explicar de mejor manera la variable dependiente.

Como se mencionó antes, nos enfocamos en encontrar la combinación de variables independientes que entregará el r2 ajustado más alto, para luego, dada esa combinación de variables obtener los coeficientes beta normalizados que es la magnitud y dirección de la relación entre cada uso de suelo y todo el resto de los usos.

Una primera mirada nos muestra que las "x" en los coeficientes determinan que esa variable no está integrada a la explicación de la variable objetivo. Además, se menciona la diagonal de la tabla con puras "x" ya que estadísticamente no tendría incluir una variable dependiente como su propio predictor.

Un coeficiente positivo como el que muestra comercio con educación (0.26) implica que un aumento en los metros cuadrados de educación también se asocia con un aumento en los metros cuadrados comerciales, mientras que comercio con industria con un coeficiente negativo (-0.03) podría estar indicando que ambos usos de suelo compiten por el espacio, o algún conflicto o incompatibilidad.

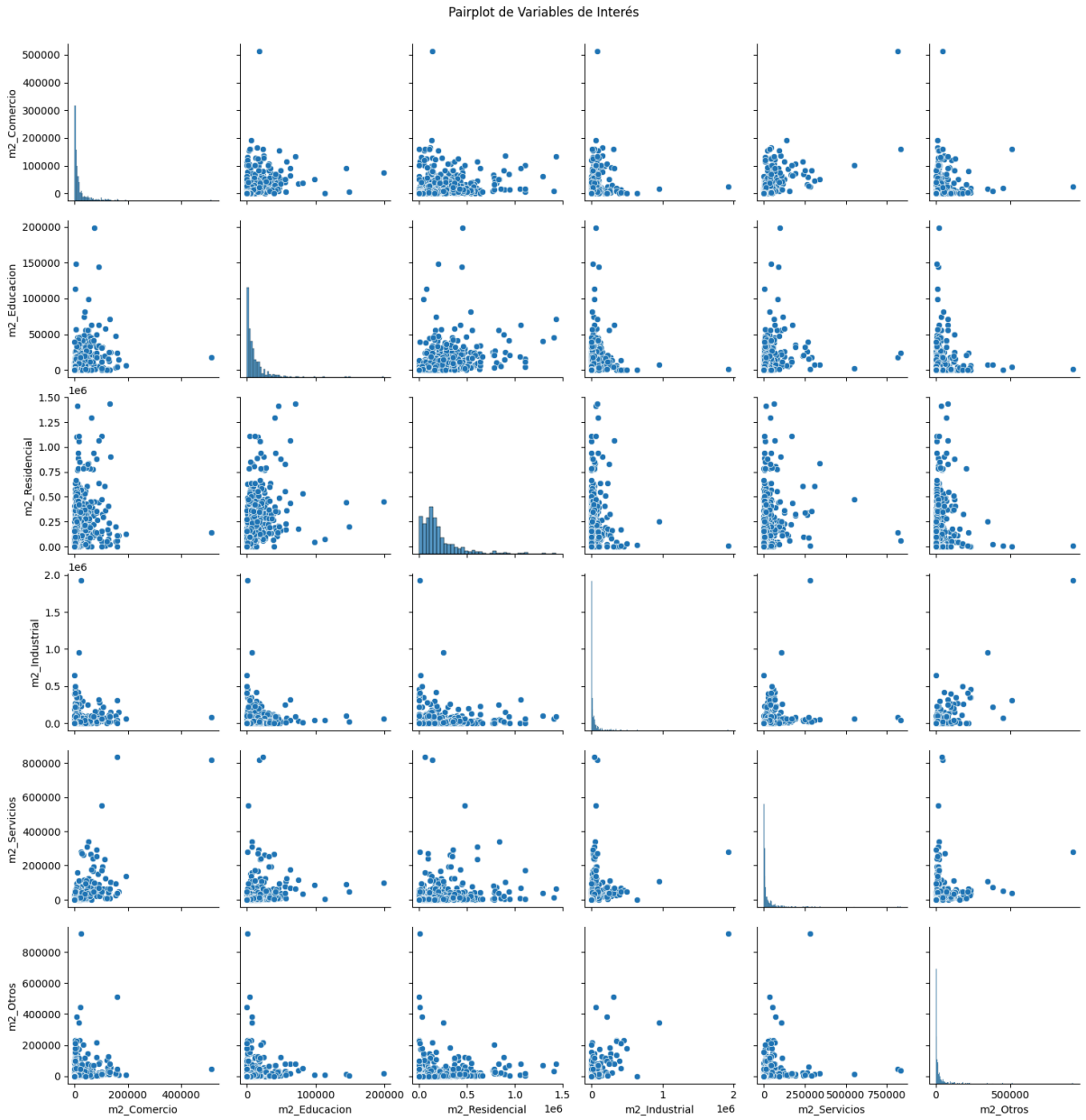
Las relaciones más fuertes que podemos encontrar son servicios con comercio (coeficiente 1.11), los cuales suelen estar presentes en las mismas áreas de alta actividad económica. Otro uso fuertemente relacionado es el residencial con comercio (coeficiente 0.62), pensando en zonas residenciales mixtas, y a su vez, muy fuertemente relacionada con el uso educación (coeficiente 4.07), que es la más fuerte de las relaciones y usos que son extremadamente compatibles. Por último, cabe destacar la coherencia en los resultados al obtener un coeficiente negativo en el modelo residencial cuando se considera el uso industrial (-0.16) demostrando una incompatibilidad urbana.

Las constantes de los modelos indican condiciones fijas o contribuciones independientes de las otras variables al uso de suelo. Resalta fuertemente la constante alta en el uso residencial, dado que es el uso con más metros cuadrados construidos, y posiblemente con más de alguna zona casi exclusivamente residencial.

Finalmente, los valores r2 ajustados obtenidos en cada modelo muestran que tanta variabilidad de los datos están explicados en los modelos. **Los desempeños en general no son muy buenos, y esto se debería a factores adicionales que no se incluyen en los modelos.** Si este fuese el caso se tiene que considerar el ajuste y desarrollo de otros modelos que capturen relaciones más complejas y no lineales.

Es por esto por lo que se considera la regresión lineal múltiple como un buen punto de partida, y que cumple para cuantificar la relación entre usos, pero se deben explorar otros modelos más complejos para lograr una aplicación y mejora en predicción.

Por otro lado, es interesante ver cómo se relacionan las variables de estudio entre sí, debido a que de esta manera podemos visualizar fácilmente si dichas relaciones tienen un comportamiento lineal o más complejo. En base a eso, podemos decidir si aplicar regresión lineal o algoritmos de aprendizaje supervisado más robustos que están diseñados para trabajar con relaciones no lineales como Random Forest, SVR, etc. Es por este motivo que se realizó el siguiente gráfico.



Relaciones entre las variables del estudio

Como podemos apreciar, **el comportamiento entre las variables no parece lineal**. Es por ese motivo que se procedió a realizar el mismo análisis anterior, pero aplicando otros modelos más adecuados según los datos de estudio tales como random forest y árboles de decisión.

1. Random Forest:

Usos de Suelo	Comercio	Educacion	Residencial	Industrial	Servicios	Otros	R2 Ajustado	MSE	RMSE
Comercio	x	x	0.11	0.12	0.64	0.13	0.91	101015622.4351	10050.6528
Educacion	x	x	0.50	x	0.50	x	0.90	26176861.7131	5116.3328
Residencial	0.11	0.38	x	0.17	0.12	0.22	0.90	3855650135.7332	62093.8816
Industrial	x	0.09	x	x	0.25	0.66	0.92	821911598.9870	28669.0007
Servicios	0.58	0.12	0.10	0.09	x	0.10	0.90	390819233.1288	19769.1485
Otros	0.13	0.05	0.12	0.60	0.10	x	0.91	275558391.4584	16599.9515

Tabla 2: Resultados Random Forest

2. Árboles de decisión:

Usos de Suelo	Comercio	Educacion	Residencial	Industrial	Servicios	Otros	R2 Ajustado	MSE	RMSE
Comercio	x	0.10	x	x	0.87	0.03	0.63	432070554.5723	20786.3069
Educacion	0.25	x	0.43	0.32	x	x	0.66	91734152.0933	9577.7947
Residencial	x	0.56	x	0.10	0.09	0.26	0.52	19243696813.3847	138721.6523
Industrial	0.01	0.11	x	x	0.12	0.77	0.84	1669054818.8186	40854.0673
Servicios	x	0.42	0.35	0.19	x	0.04	0.66	1339336196.3928	36596.9424
Otros	x	x	x	0.96	0.04	x	0.80	626276104.6788	25025.5091

Resultados Árboles de decisión

Como se puede apreciar en las imágenes, en base a las mejores combinaciones para cada caso, se obtuvieron valores muy altos para el R2 en la mayoría de los escenarios. Este resultado sugiere que el modelo ajusta muy bien los datos de entrenamiento, lo cual es positivo. Un valor alto valor de R2 indica que el modelo explica una gran parte de la variabilidad en los datos. Sin embargo, es importante notar que un R2 muy bueno no siempre garantiza un buen rendimiento en datos nuevos o no vistos, especialmente cuando se observan señales de sobreajuste.

Por otro lado, los valores obtenidos para el MSE (Mean Squared Error) y RMSE (Root Mean Squared Error) en todos los casos son excesivamente altos. Esto nos indica que **aunque el modelo parece ajustarse bien a los datos de entrenamiento, la magnitud de los errores es considerablemente alta.** Esto es un claro indicativo de sobreajuste.

El hecho de que los valores del MSE y RMSE sean altos refuerza esta idea de sobreajuste, ya que estos indicadores reflejan el error promedio del modelo. Idealmente, para un modelo que generaliza bien, tanto el MSE como el RMSE deberían ser lo más bajos posibles, lo que indicaría que los errores de predicción son pequeños. En resumen, aunque el R2 alto puede parecer muy bueno, los valores de MSE y RMSE nos están indicando que el modelo no tiene una buena capacidad predictiva y puede fallar al aplicarse en un entorno real.

Estudio del valor de suelo

Una vez que hemos cuantificado y modelado la relación entre diferentes usos de suelo en el Gran Santiago, avanzamos a predecir el valor del metro cuadrado. Tal como consignamos en la primera parte del documento, esta variable presenta una abundante cantidad de valores nulos (243). Sin embargo, el valor de los datos existentes nos obliga a emplearlos para determinar si es posible modelar y predecir satisfactoriamente el precio del metro cuadrado.

1. Regresión Lineal

En primera instancia se realizó una regresión lineal en donde se obtuvo lo siguiente:

R^2	MSE	RMSE
0.38	157.14	12.54

Métricas de evaluación para la regresión lineal

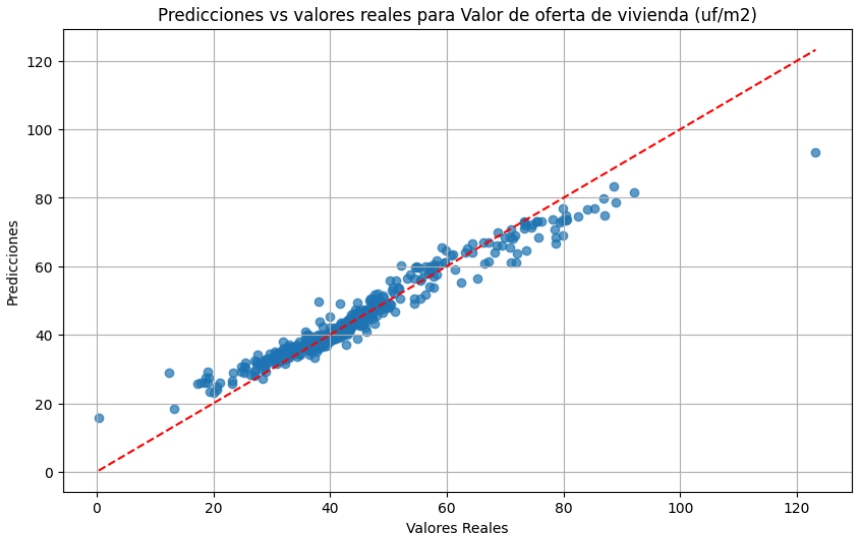
2. Random Forest

Como segundo modelo trabajamos nuevamente con Random forest, para este escenario se obtuvieron los siguientes resultados:

R^2	MSE	RMSE
0.58	107.34	10.36

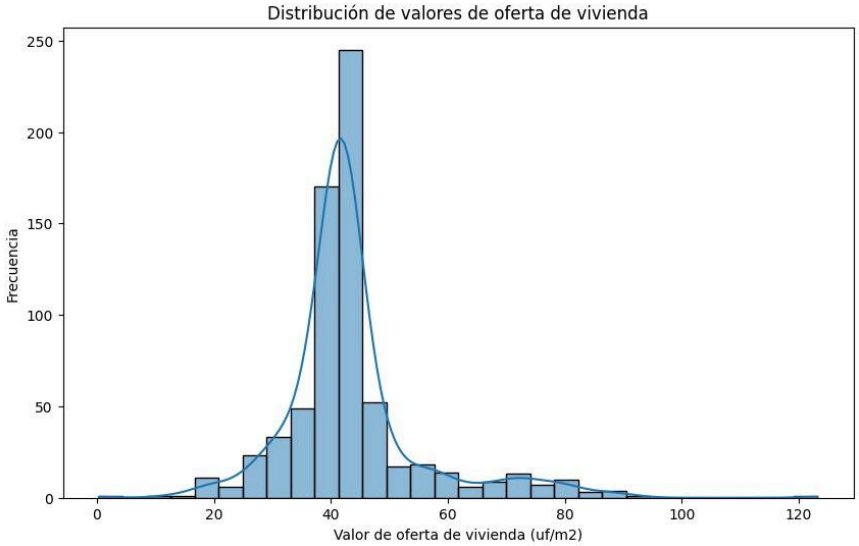
Métricas de evaluación para random forest

Como podemos apreciar, Random Forest tuvo un mejor desempeño en comparación a la regresión lineal, podemos notar un pequeño sobreajuste, pero nada considerablemente extraño. También graficamos los valores predecidos por nuestro modelo y los valores reales para la variable del valor de vivienda, donde se obtuvo lo siguiente.



Gráfica de datos reales y predecidos por el modelo.

Podemos apreciar fácilmente como nuestro modelo fue capaz de predecir de muy buena forma los valores para la variable valor de la vivienda. Podemos notar una pequeña desviación con respecto a los valores reales, lo que podemos explicar por el pequeño sobreajuste de nuestro modelo.

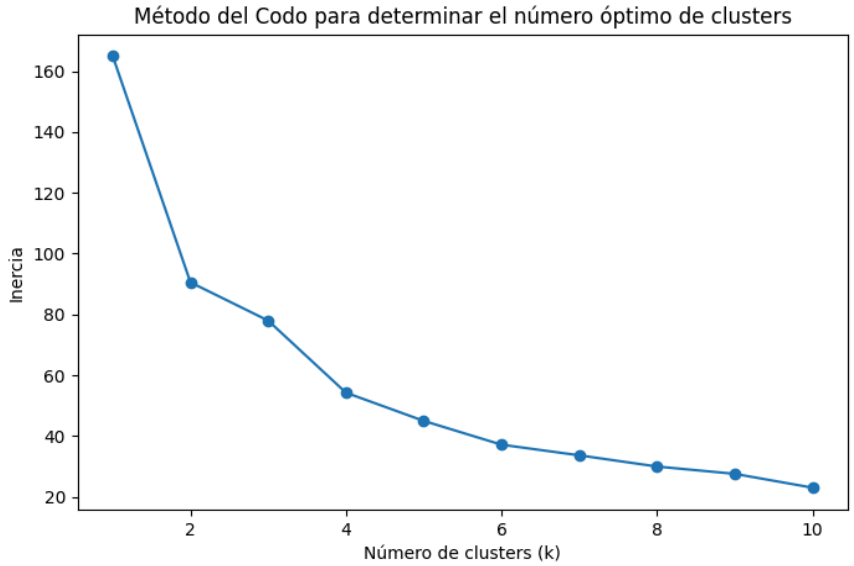


Distribución del precio del metro cuadrado.

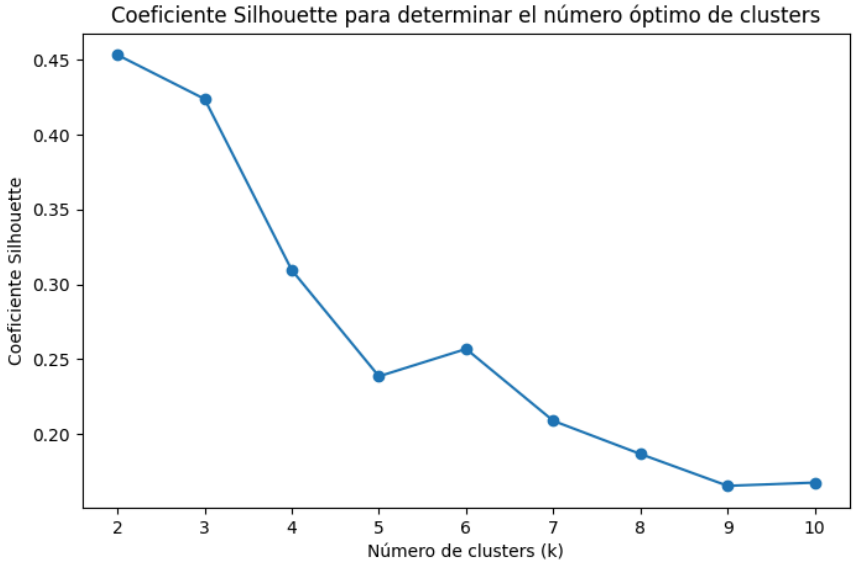
3. Análisis de Cluster

Realizada la predicción de precios, estamos listos para agrupar las comunas. Efectuamos el **Método de Codo**, que permite determinar el número óptimo de clusters. Para tal efecto, iteramos sobre diferentes valores de k (número de clusters) y calculamos la inercia o suma de los errores cuadráticos para cada uno de sus

valores. El momento de disminución de inercia, es el punto óptimo. Lo que es evidente en el valor 4 de nuestro ejercicio.

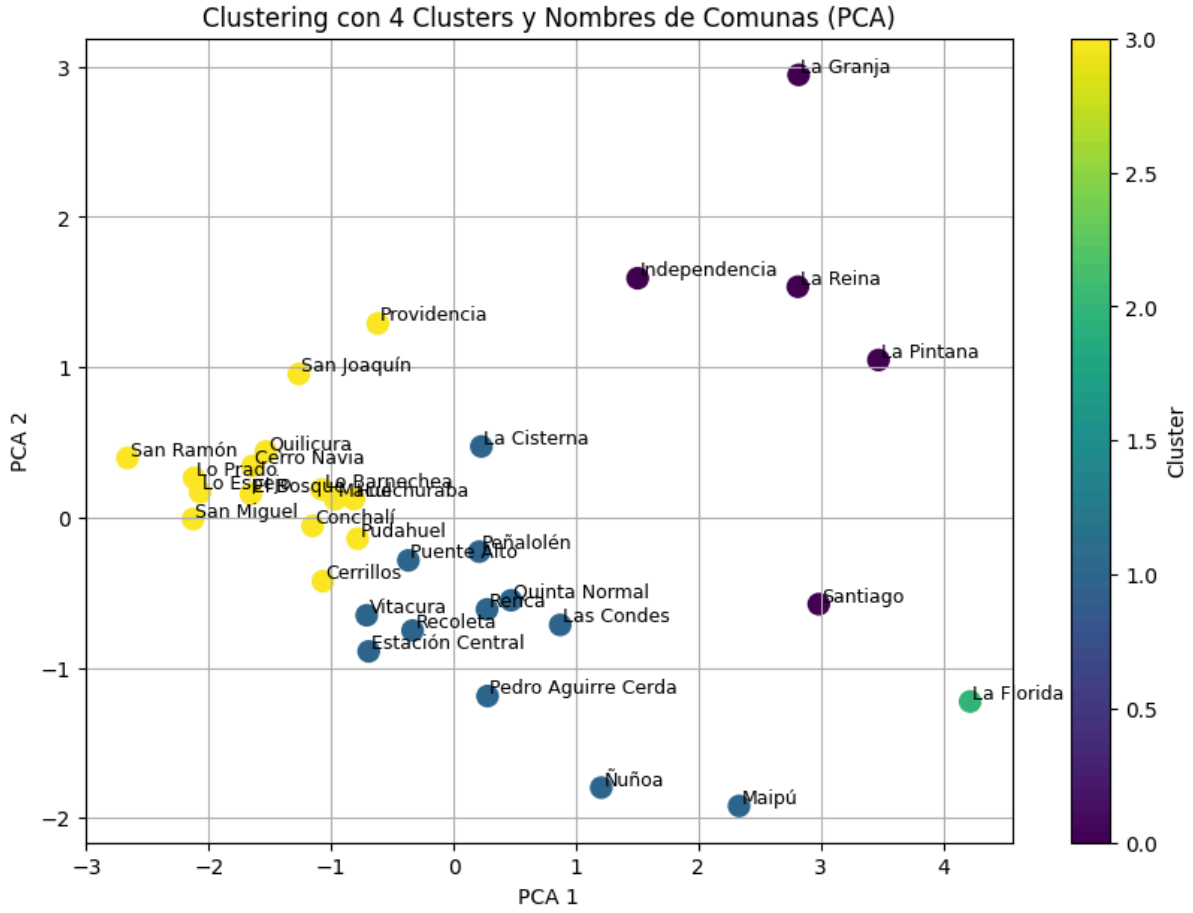


Luego, ejecutamos el **Método Silhouette,** que mide la bondad del agrupamiento dentro del cluster asignado, en comparación con otros clusters. De este modo, obtenemos una métrica de cohesión y separación. En la gráfica correspondiente, observamos un punto óptimo en 5 clusters.



Finalmente, efectuamos la agrupación de las comunas en cuatro (4) clusters mediante el método K-Means y Reducción de Dimensionalidad. Se observan comunas similares en proporción de viviendas, industrias, educación y servicios, así como niveles socioeconómicos de sus residentes.

En específico, **existe similaridad de comunas intuitivamente disímiles** como Providencia y Lo Barnechea con Quilicura, San Ramón, Pudahuel, Cerrillos y otras. Del mismo modo, La Reina comparte grupo con La Granja, Independencia y La Pintana. Y Las Condes y Vitacura con Puente Alto y Ñuñoa. Destaca La Florida muy distinta a las otras comunas.



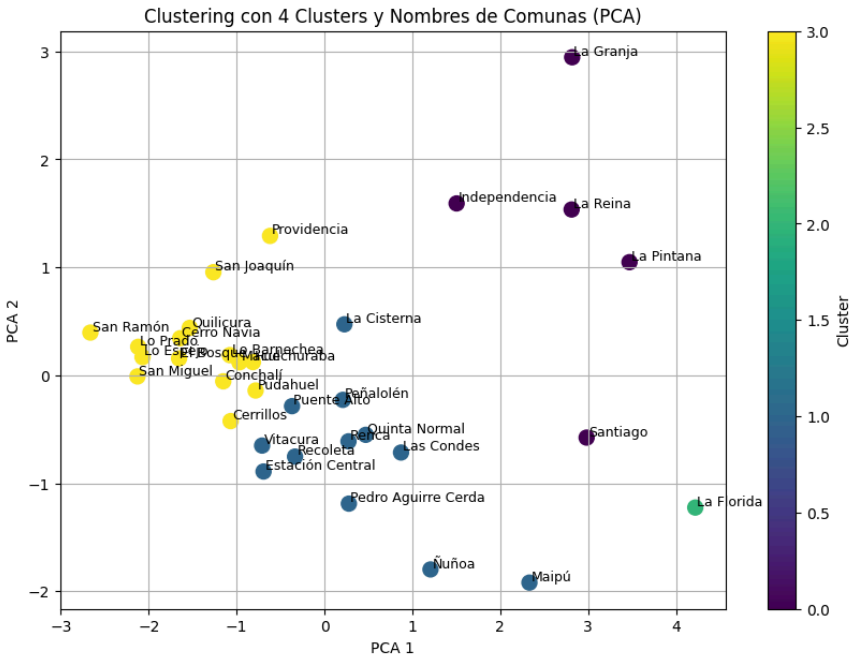
Así, caracterizamos a los grupos de la siguiente forma:

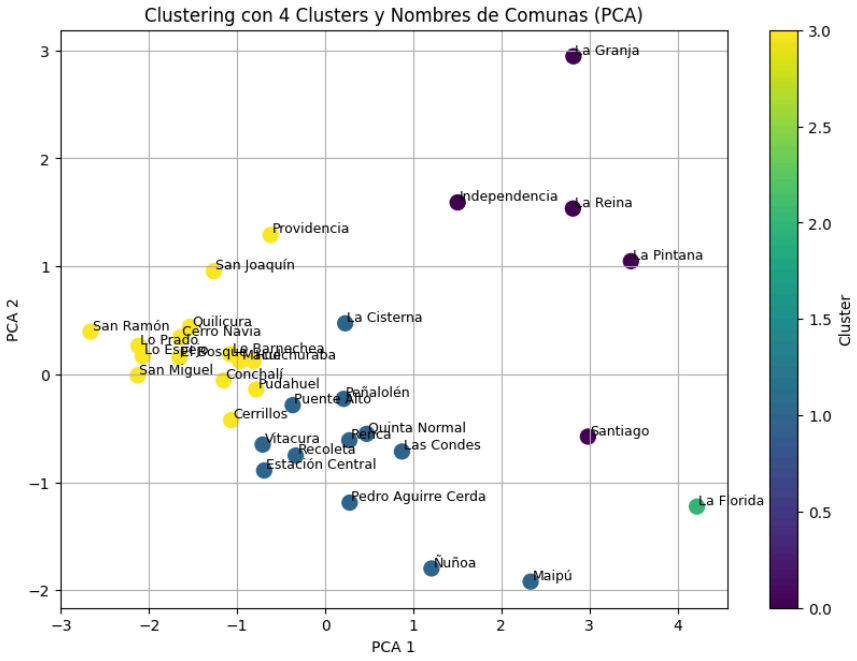
Grupo 1 (amarillo): Grupo homogeneo

Grupo 2: (morado): Grupo comercio

Grupo 3 (azul): Clase media Servicios

Grupo 4 (verde): Residencial





Conclusión

El análisis demuestra una correlación significativa entre ciertas características del uso de suelo en las comunas del Gran Santiago y el valor del suelo. El modelo supervisado implementado, utilizando tanto regresión lineal múltiple como técnicas más avanzadas como Random Forest, permitió predecir el valor del suelo con un alto nivel de precisión.

Aunque se observó un sobreajuste en algunos casos, los resultados sugieren que estas características del uso de suelo son determinantes clave en la valorización del suelo urbano. Adicionalmente, los análisis de agrupamiento por clúster identificaron patrones comunes entre comunas, revelando similitudes inesperadas que pueden influir en la planificación urbana futura.

En resumen, este estudio ofrece una herramienta valiosa para la planificación urbana y la toma de decisiones en políticas de desarrollo territorial, permitiendo predecir de manera precisa el valor del suelo a partir de características del uso de suelo y dinámicas urbanas.

Bibliografía

Servicio de Impuestos Internos. (2015). Información de predios y líneas de construcción por uso de suelo. SII. https://www.sii.cl

Subsecretaría de Transportes. (2015). Encuesta Origen y Destino. SECTRA. https://www.sectra.gob.cl