

**Análisis Multidimensional del Uso del Suelo mediante la aplicación de técnicas de Machine Learning para las comunas del Gran Santiago.**

Cristián Canales

Tomás Sepúlveda

Sebastián Pérez

**Introducción**

En el siguiente trabajo se abordará un problema asociado con el análisis y dinámicas de los distintos usos de suelos urbanos en el Gran Santiago, teniendo como objetivo el uso de distintas técnicas de aprendizaje supervisado y no supervisado para el estudio y cuantificación de la relación entre los distintos usos de suelo existentes, y la diferenciación o clasificación posible entre las diferentes comunas propias del estudio.

Para ello se utiliza una base de datos que se caracteriza por ser privada (no publicada), pero realizada a través de datos públicos. Corresponde al cruce de la zonificación EOD, proveniente de la encuesta origen y destino realizada por la Subsecretaria de Transportes (SECTRA) para el desarrollo de estudios de oferta y demanda en transporte y dinámicas de movilidad urbana, con la información relacionada con predios y líneas de construcción por uso de suelo del Servicio de Impuestos Internos (SII). Este cruce permite contar con las 32 comunas que componen el Gran Santiago, divididas en sus respectivas zonas de estudio, las cuales cuentan con el total de metros cuadrados construidos de Comercio, Educación, Residencia, Industria, Servicios y Otros usos que conforman cada zona. Además, la base de datos cuenta con el total de hogares diferenciados por grupos socioeconómicos propias del estudio. Si bien, la información presente no esta actualizada a la fecha, si permite abordar las dinámicas de usos de suelo con datos reales, y las problemáticas planteadas en el siguiente trabajo.

**Objetivos**

**Objetivo general:** Cuantificar y modelar la relación entre diferentes usos de suelo en el Gran Santiago, utilizando diferentes técnicas de aprendizaje supervisado. Además del agrupamiento de las comunas en estudios utilizando técnicas de aprendizaje no supervisado.

**Objetivo general:**

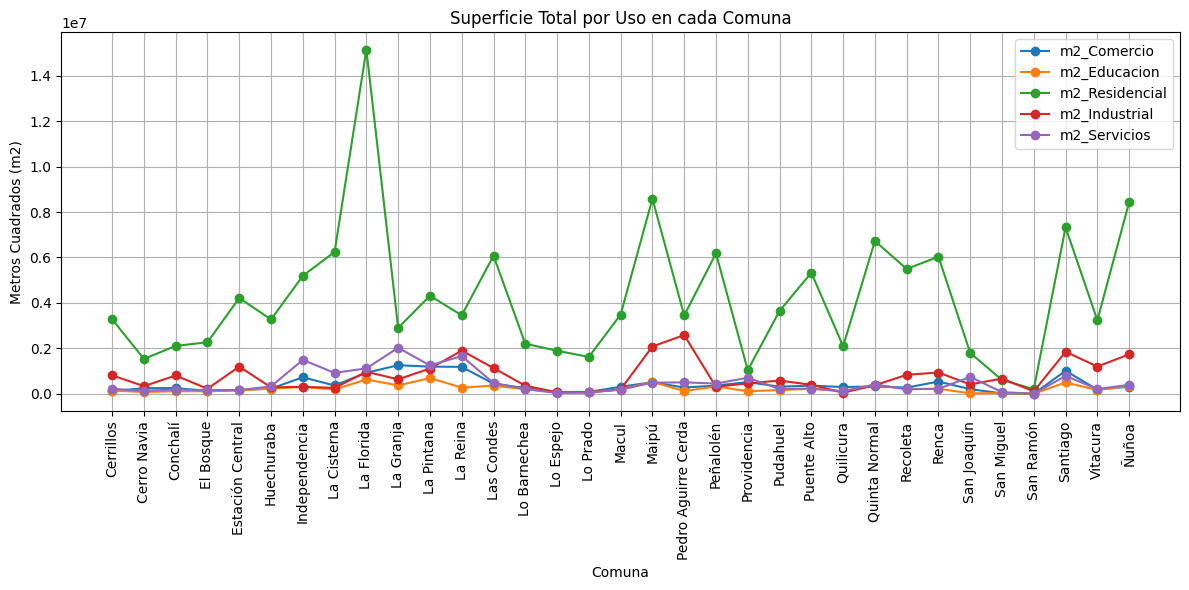
* Preparar y arreglar la base de datos para poder usarla en este estudio con un formato adecuado.
* Realizar un análisis exploratorio de los datos y observar la distribución de los diferentes usos de suelo de las comunas del Gran Santiago.
* Ajustar y comparar diferentes modelos de aprendizaje supervisado para 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.
* Ajustar y comparar diferentes modelos de agrupamiento o cauterización sobre las comunas del Gran Santiago para evidenciar características, patrones y dinámicas comunes, en función de sus proporciones de uso de suelo y cantidad de hogares por grupos socioeconómicos.

**Resultados**

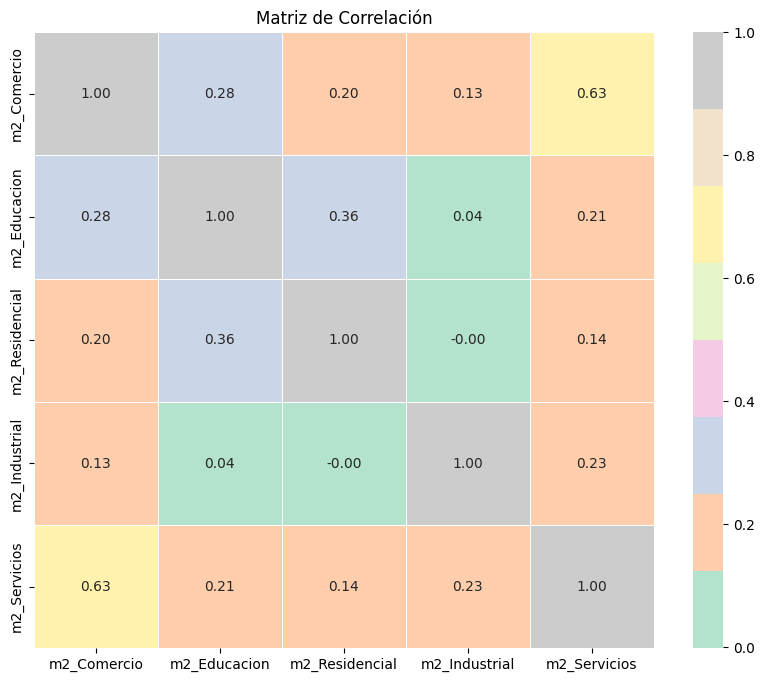
1. Análisis exploratorio

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 |



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).



De este modo, 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.

1. Técnicas de aprendizaje Supervisado

Dado el objetivo de cuantificar las relaciones entre los distintos usos de suelo se propone y aplica distintas técnicas de aprendizaje supervisado. En un comienzo se parte con el ajuste de un modelo de regresión lineal múltiple para cada tipo de uso de suelo. La idea es 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.

Por lo tanto, 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.

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 |

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 (1.11), los cuales suelen estar presentes en las mismas áreas de alta actividad económica. Otro uso fuertemente relacionado es el residencial con comercio (0.62), pensando en zonas residenciales mixtas, y a su vez, muy fuertemente relacionada con el uso educación (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 podría estar dando porque haya factores adicionales que no se estén incluyendo en los modelos, y es importante tener esta consideración para, por ejemplo, realizar una proyección. 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.



Gráfico 1: Relaciones entre las variables de estudio.

Como podemos apreciar, el comportamiento entre las variables no parece ser 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.

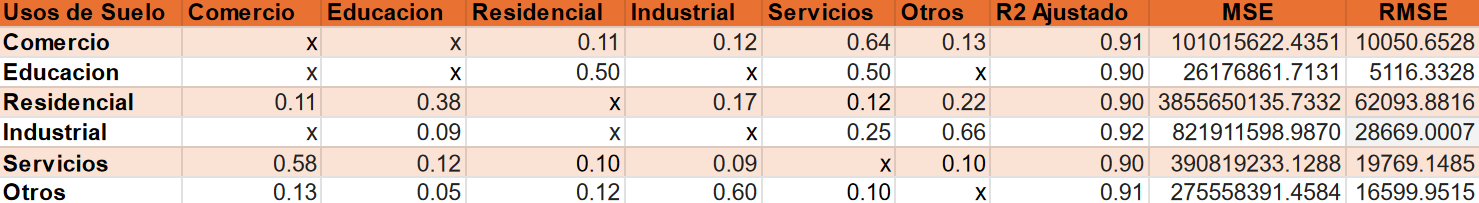
1. **Random Forest:**

Tabla 2: Resultados Random Forest

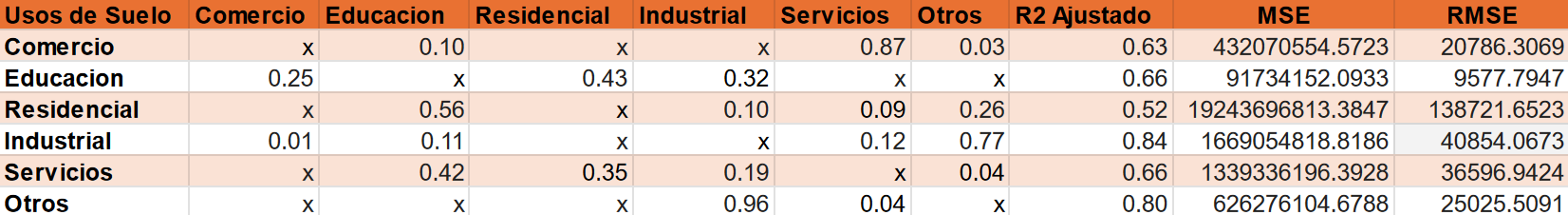
1. **Árboles de decisión:**

Tabla 3: 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.

Finalmente, consideramos que era interesante poder predecir el costo de valor de la vivienda. Tal como se comentó en la primera parte del documento, esta variable presentaba demasiados valores nulos(243). De esta manera utilizamos todas las variables de m2 y de GSE para predecir el valor de viviendas faltantes.

1. **Regresión Lineal**

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

Tabla 4: Métricas de evaluación para la regresión lineal

1. **Random Forest**

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

Tabla 5: 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.

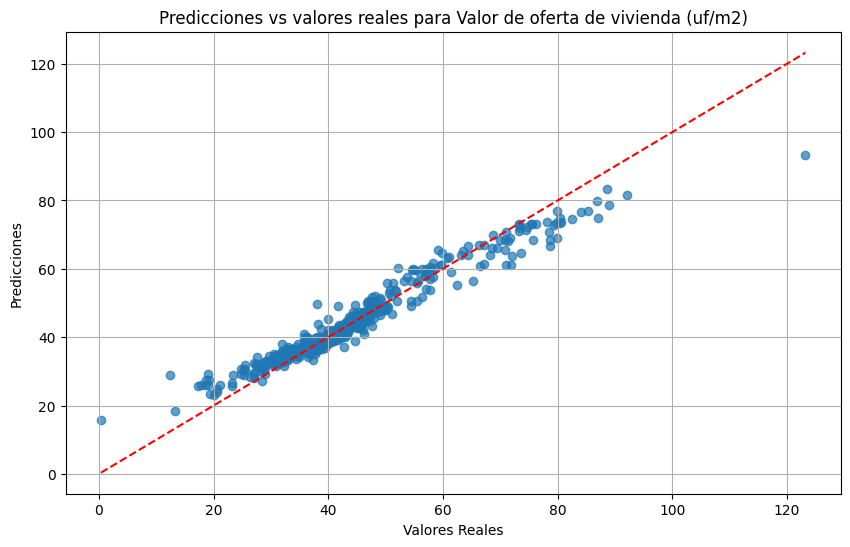


Gráfico 1: Gráfica de datos reales y predecidos por el modelo.

Podemos apreciar fácilmente como nuestro modelo es 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 cual se puede explicar por el pequeño sobreajuste de nuestro modelo.

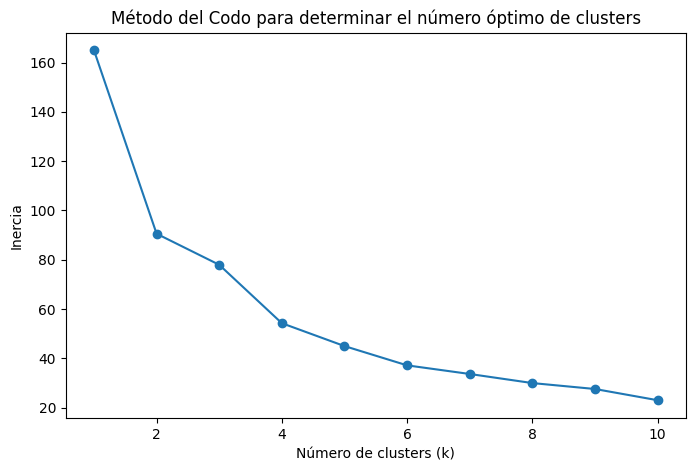
b. Aplicar apropiadamente técnicas del aprendizaje supervisado y seleccionar el mejor modelo posible.

Aplicar correctamente las técnicas de Aprendizaje Supervisado; separación train-test, elección apropiada de indicadores comparativos y técnicas según el tipo de dato de la variable respuesta. En caso de ser necesario aplicar técnicas de balanceo de clases o bien obtienen la probabilidad de corte óptima bajo algún criterio correctamente establecido

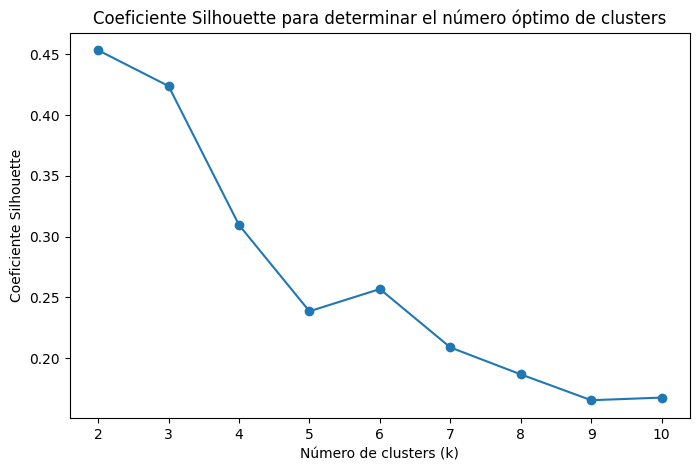
Seleccionan el mejor modelo de aprendizaje supervisado de acuerdo con indicadores claros y correctamente utilizado

1. Análisis de Cluster

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.

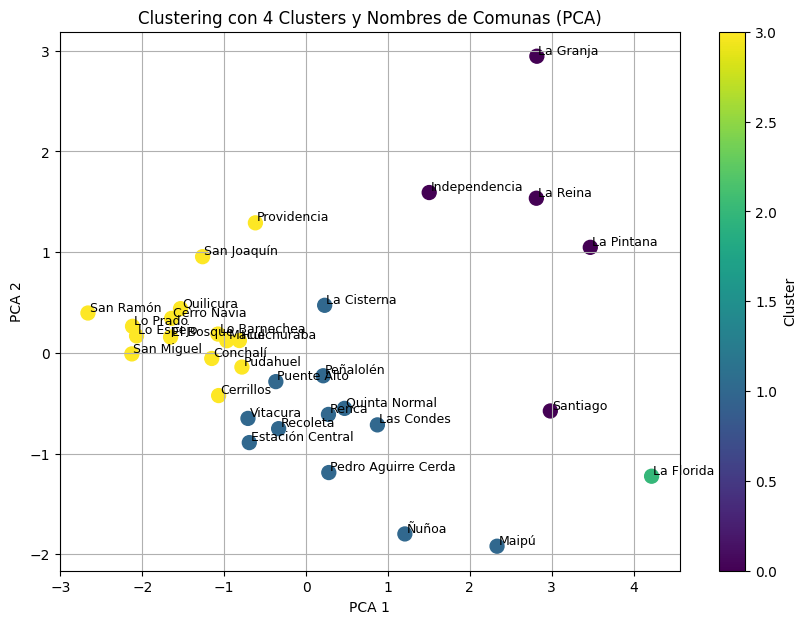


Luego, ejecutamos el **Método Silhouette,** que mide la bondad del agrupamiento dentro del cluster asignado, en comparación con otros clusters. De este modo, obtenemos una métrica de cohesión y separación. En la gráfica correspondiente, observamos un punto óptimo en 5 clusters.



**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.



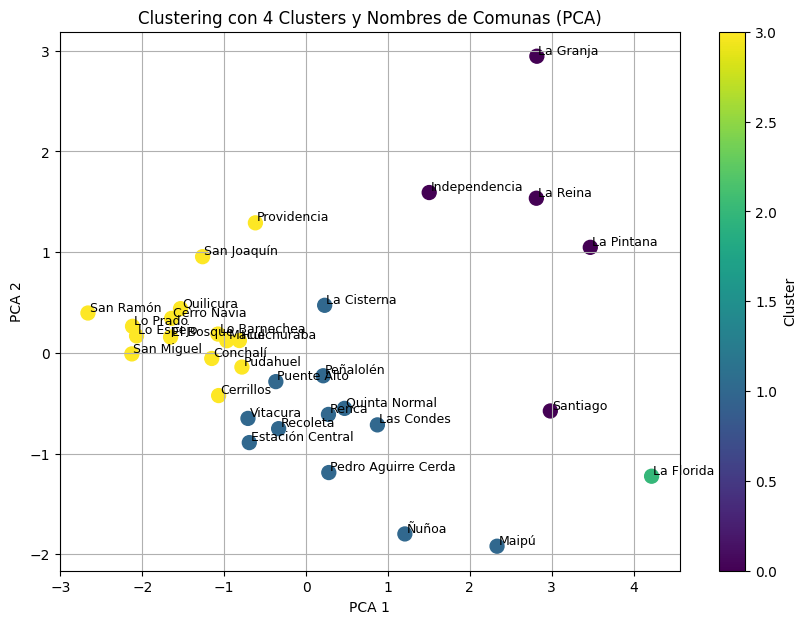
Así, caracterizamos a los grupos de la siguiente forma:

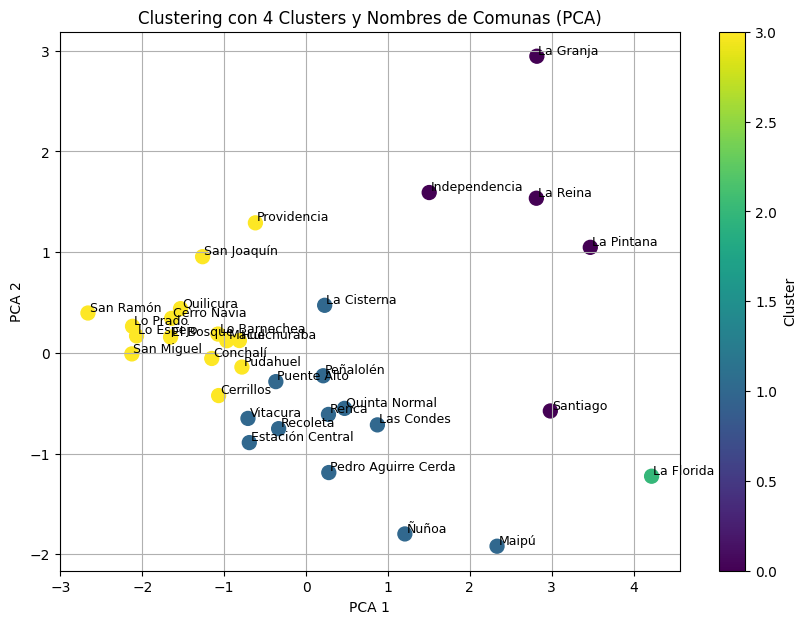
Grupo 1 (amarillo): Grupo homogeneo

Grupo 2: (morado): Grupo comercio

Grupo 3: Clase media Servicios

Grupo 4: Residencial





**Conclusión**

Relacionada con los resultados y objetivos planteados inicialmente.

Las conclusiones se relacionan con el contexto del problem

**Anexos**

Es opcional y contempla cualquier información adicional que consideren relevante.

**Bibliografía**