

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“DETERMINANTES DEL COMPORTAMIENTO DE
PRECIOS DE PROPIEDADES EN CONURBACIONES DE
LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y
BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE
AUTOMÁTICO.”

JOSÉ MIGUEL SANSANA PARRA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Mauricio Olivares Faúndez
Profesor Correferente: José Luis Martí Lara

Agosto - 2023

DEDICATORIA

Para toda persona que, en algún período de la vida, haya desconocido su propio resplandor al verse inmersa en un oscuro monólogo interno. Siempre existirá un atisbo de luz que guiará la salida.

AGRADECIMIENTOS

Sería contraintuitivo no comenzar este apartado reconociendo a los principales influyentes en la persona que soy hoy: Mis padres, Soledad y José Miguel, con los cuales pude interiorizar valores clave como lo son el amor, la bondad, la responsabilidad y disciplina. Valores que me han permitido forjar tanto mi carácter como el camino que he seguido hasta este punto.

Agradezco también a mis hermanos Tomás y Estela. Al primero por estar siempre conmigo en muchas experiencias de vida y a la segunda porque, a pesar de la distancia, logro sentir todo el cariño que nos tenemos. A mi tío Denis, ya que sin él no hubiese podido tener una cómoda estadía en Santiago y cuyo conocimiento en el área de Inteligencia Artificial me ha iluminado en varias ocasiones, incluyendo este escrito. A mis abuelas Cecilia y Zunilda, y a mis abuelos Sergio y José Miguel por siempre haberme apoyado de manera incondicional, orientándome a darle énfasis a mis estudios y a terminar mi memoria a tiempo. A mis tías Marcia y Tita, con quienes he creado bonitos recuerdos y experiencias. A mi tío Guillermo, con aquellas memorables conversaciones frente al pino de las empanadas. Agradezco también a Jhoan, por enseñarme a ser alguien auténtico.

Pero no todo es la sangre, puesto que en mi período escolar y universitario he conocido personas maravillosas que han contribuido a mi desarrollo en todos los ámbitos. Aquí menciono a Bianca, quien me ha demostrado que la vida puede ser bella y colorida, apoyándome en momentos críticos, y a su familia, la cual se ha encargado de hacerme sentir parte de la misma con un cálido trato. A Matías, Pablo, Seba y Agus, por tantas vivencias y conversaciones, ya sea por Discord o en persona. Por supuesto, elogiar a los chicos que conocí durante mi estadía en la USM, en particular a Álvaro, Iván, Max, Joaquín, Cristóbal, Diego, Nico, Peter y Barber, con los cuales viví momentos de supervivencia universitaria durísimos, pero también momentos de compañerismo inigualables. También agradezco a la Cote por brindarme apoyo, más aún en duros momentos de la pandemia.

Agradezco profundamente al profesor Mauricio Olivares por confiar en mi trabajo y en lo que podía ofrecer.

Los apartados previos pueden pensarse como un desesperado intento de englobar a todo aquel que alguna vez tuvo una influencia en mí, sin embargo, es una clara muestra de que aquellos mencionados realmente han tenido un impacto significativo en mi vida, y es este impacto el que me ha traído al lugar donde estoy, recordando cada experiencia que me formó y me permitió llegar al final de una dura carrera universitaria.

Quizás una buena medida del impacto de estas experiencias pueda ser la sonrisa del autor al terminar esta sección.

RESUMEN

La geografía chilena hace que este sea un país muy diverso en variados aspectos, haciéndose presente en la cultura de diversas comunas y conurbaciones. Es por ello que se desea indagar si esta diversidad afecta los atributos que más determinan el precio de venta de un inmueble dependiendo del lugar geográfico de la propiedad. Considerando lo anterior, en este estudio se comparan tres modelos de aprendizaje automático: Regresión lineal, árboles de decisión y GAM mediante técnicas de inferencia y métricas de desempeño. La metodología se centra en la recopilación de datos mediante *web scraping* en Goplacit.com, análisis exploratorio, pre-procesamiento de datos, ajuste de modelos y análisis de inferencia y métricas en base a estos. Los resultados dictan que el árbol de decisión es el más consistente respecto a la inferencia realizada, puesto que muestra los mismos cinco atributos (m^2 útiles, m^2 totales, latitud, longitud y cantidad de baños) como los más influyentes independiente de la conurbación, sumado a un desempeño superior a los otros modelos en cuanto a predicción. A su vez, todos los modelos tienden a bajar su desempeño al reducir la cantidad de datos. GAM baja el desempeño más drásticamente que otros modelos y no provee determinantes concisos. Por último, regresión lineal logra un intermedio de desempeño junto a consistencia, ya que dentro de los seis atributos más importantes entre conurbaciones, permanece un patrón consistente.

Palabras Clave— Aprendizaje automático; *Web scraping*; Interpretabilidad; Precio de viviendas; conurbaciones chilenas

ABSTRACT

Abstract—Chilean geography makes this a very diverse country in various aspects, becoming present in the culture of various communes and conurbations. That is why it is desired to investigate whether this diversity affects the attributes that most determine the sale price of a property depending on the geographical location of it. Considering the above, in this study three machine learning models are compared: Linear Regression, Decision Trees and GAM using inference techniques and performance metrics. The methodology focuses on data collection through *web scraping* on Goplacit.com, exploratory analysis, data pre-processing, model fitting and inference analysis and metrics based on these. The results dictate that the decision tree is the most consistent with respect to the inference made, since it shows the same five attributes (*useful* m^2 , *total* m^2 , latitude, longitude and number of bathrooms) as the most influential independent of the conurbation, added to a better

performance respect to the other models compared. In turn, all models tend to lower their performance by reducing the amount of data. GAM lowers performance more drastically than other models and does not provide concise determinants. Finally, linear regression achieves an intermediate performance along with consistency, since within the six most important attributes between conurbations, a consistent pattern remains. Corresponde a la traducción al idioma inglés del Resumen anterior. Sujeto a la misma regla de extensión del Resumen.

Keywords—Machine Learning; Web scraping; Intrepretability; Property prices; Chilean Conurbations

GLOSARIO

CLP: Moneda utilizada en Chile.

CSV: Del inglés *Comma separated value* corresponde a un tipo de archivo similar a Excel, pero con datos separados por comas en vez de celdas.

DF: Del inglés *DataFrame*, representando una estructura matricial capaz de contener datos en Python.

GAM: Del inglés *Generalized Additive Models*. Familia de modelos de aprendizaje automático.

GC: Conurbación en región de Bío Bío, conocida como Gran Concepción.

GS: Conurbación en región metropolitana, conocida como Gran Santiago.

GV: Conurbación en región de Valparaíso, conocida como Gran Valparaíso.

MAE: Del inglés *Mean Absolute Error*, métrica de desempeño de modelos de regresión.

ML: Del inglés *Machine Learning* o aprendizaje automático.

MSE: Del inglés *Mean Squared Error*, métrica de desempeño de modelos de regresión.

PDP: Del inglés *Partial Dependency Plots*, gráficos que relacionan la influencia de un atributo con la oscilación de una variable objetivo.

R2: Del inglés *R Squared*, métrica de desempeño de modelos de regresión.

RMSE: Del inglés *Root Mean Squared Error*, métrica de desempeño de modelos de regresión.

Scraper: Programa que ejecuta web scraping.

UF: Unidad de fomento de la producción (indicador monetario chileno).

WS: *Web scraping* o raspado web.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
GLOSARIO	VI
ÍNDICE DE FIGURAS	IX
ÍNDICE DE TABLAS	X
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	2
1.1 Contexto y Motivación General	2
1.2 ¿Cómo determinar el precio de un inmueble en Chile y comparar los factores que inciden en este?	2
1.3 Extracción de datos	4
1.4 Objetivos	5
1.4.1 Objetivo general:	5
1.4.2 Objetivos específicos:	5
CAPÍTULO 2: MARCO CONCEPTUAL	6
2.1 <i>Conurbaciones en Chile</i>	6
2.2 Mercado inmobiliario Chileno actual	7
2.3 Modelo hedónico de precios	7
2.4 Goplaceit.com	7
2.5 Aprendizaje automático (ML)	8
2.6 Problemas de regresión y clasificación	8
2.7 Interpretabilidad en ML	9
2.7.1 Definición de interpretabilidad	9
2.7.2 Enfoques <i>model-specific</i> y <i>model-agnostic</i>	9
2.7.3 Enfoques locales y globales	10
2.7.4 Modelos altamente interpretables	11
2.8 Inferencia y predicción en ML	12
2.9 Modelos de ML seleccionados	13
2.9.1 Regresión lineal	13
2.9.2 <i>Generalized additive models</i>	15
2.9.3 Árboles de decisión	16
2.10 Cálculo de "Importancia de atributos" en ML	18
2.10.1 "Importancia de atributos" en regresión lineal	19
2.10.2 "Importancia de atributos" en GAM	19
2.10.3 "Importancia de atributos" en árboles de decisión (CART, Random Forest y XGBoost)	20

2.11	Métricas de evaluación de desempeño de modelos de regresión	21
2.11.1	R^2	21
2.11.2	Mean Squared Error y Root Mean Squared Error	21
2.11.3	Mean Absolute Error	22
2.12	Cross-Validation	22
2.13	Python	23
2.14	Web scraping	23
2.14.1	Librerías de web scraping	23
2.15	Legalidad del web scraping en Chile	24
2.15.1	Acceso a la información de Goplaceit.com	24
2.16	Jupyter Notebook	25
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN		26
3.1	Metodología propuesta	26
3.2	Fase 1: Extracción de datos de propiedades mediante web scraping	27
3.2.1	Consideraciones al extraer datos del Gran Concepción	31
3.2.2	Consideraciones al realizar webscraping de manera prudente	31
3.3	Fase 2: Extracción de datos de entorno de propiedades manualmente	32
3.4	Fase 3: Limpieza y estructuración de datos obtenidos	33
3.4.1	Limpieza y estructuración de datos en Gran Santiago	34
3.4.2	Limpieza y estructuración de datos en Gran Valparaíso	40
3.4.3	Limpieza y estructuración de datos en Gran Concepción	45
3.4.4	Resumen al finalizar la limpieza de datos	48
3.5	Fase 4: Implementación y entrenamiento de modelos	49
CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN		51
4.1	Análisis, inferencia y predicción de modelos	53
4.1.1	Enfoque en GS	53
4.1.2	Enfoque en GV	57
4.1.3	Enfoque en GC	61
4.1.4	Discusión sobre multicolinealidad	64
4.2	Análisis y comparación regional de mismos modelos	65
4.2.1	Enfoque en regresión lineal	66
4.2.2	Enfoque en árbol de decisión	66
4.2.3	Enfoque en GAM	67
CAPÍTULO 5: CONCLUSIONES		69
5.1	Conclusiones Generales	69
5.2	Trabajos futuros	71
ANEXOS		72
REFERENCIAS BIBLIOGRÁFICAS		78

ÍNDICE DE FIGURAS

1	Ejemplos de acercamientos a interpretabilidad de predicción en modelos de regresión.	11
2	Selección de modelos de ML dadas las necesidades del estudio.	12
3	Comparación de GAM con regresiones lineales.	16
4	Partición de espacio de características.	17
5	Árbol de decisión orientado a la regresión.	18
6	Ejemplo de PDP.	19
7	Cross-Validation.	23
8	Metodología propuesta.	26
9	Página principal goplaceit.com/cl.	27
10	Inicio de sesión en Goplaceit.com/cl.	28
11	Sección Goplaceit.com/cl/mapa.	28
12	``Mis Zonas'' en Goplaceit.com/cl/mapa.	29
13	Propiedades dentro de límites en Goplaceit.com/cl/mapa.	29
14	Ejemplo de Información intrínseca del inmueble.	30
15	Ejemplo de Información del entorno de un inmueble.	30
16	Longitud y latitud de la propiedad.	31
17	Distribución en GS de atributos internos.	34
18	Distribución en GS de atributos externos.	35
19	Distribución en GS de delincuencia y densidad.	36
20	Remoción de valores atípicos en GS.	37
21	Transformación logarítmica.	38
22	Remoción de nuevos valores atípicos en GS.	38
23	Balance de tipos de vivienda en GS.	39
24	Balance de comunas en GS.	40
25	Distribución de atributos internos en GV.	41
26	Distribución de atributos externos en GV.	41
27	tasas de delincuencia y densidad en GV.	42
28	Distribución de atributos internos al procesarse en GV.	43
29	Distribución de atributos externos al procesarse en GV.	43
30	Distribución de comunas en GV.	44
31	Desbalance de propiedades en GV.	44
32	Distribución atributos internos en GC.	45
33	Distribución atributos externos en GC.	46
34	Distribución atributos externos en GC.	46
35	Distribución final atributos internos GC.	47
36	Balance de comunas en GC.	47
37	Balance de propiedades en GC.	48
38	Matriz de correlación en GS.	51
39	Matriz de correlación en GV.	52
40	Matriz de correlación en GC.	52
41	Importancia (en porcentaje) de regresión lineal en GS.	53

42	Importancia de árbol de decisión en GS.	55
43	Importancia de GAM en GS.	56
44	Importancia (en porcentaje) de regresión lineal en GV.	58
45	Importancia (en porcentaje) de árbol de decisión en GV.	59
46	Importancia (en porcentaje) de GAM en GV.	60
47	Importancia (en porcentaje) de regresión lineal en GC.	61
48	Importancia (en porcentaje) de árbol de decisión en GC.	62
49	Importancia (en porcentaje) de GAM en GC.	63
50	Coeficientes ajustados en regresión lineal para GS.	65
51	Fragmento de código que muestra la rotación de ``user-agent" tras cada nueva página con imágenes.	72
52	Imagen a modo de ejemplo respecto al almacenamiento de densidad y tasa de delincuencia por comuna.	73
53	Distancia a Paradero y Bomberos imprecisos.	73
54	Ejemplo de <i>DF</i> completo a nivel de conurbación de región metropolitana. . .	73
55	Datos atípicos en metros cuadrados útiles para el GS.	74
56	Evidencia sesgo en GV.	74
57	Correlación en comunas de GS.	75

ÍNDICE DE TABLAS

1	Comunas que forman parte del Gran Santiago.	6
2	Comunas que forman parte del Gran Valparaíso y Gran Concepción.	7
3	Resumen de eliminación de datos por conurbación.	49
4	Suma de porcentajes por tipo de atributo en GS.	56
5	Métricas de modelos en GS.	56
6	Suma de porcentajes por tipo de atributo en GV.	60
7	Métricas de modelos en GV.	60
8	Suma de porcentajes por tipo de atributo en GC.	63
9	Métricas de modelos en GC.	64
10	Suma de porcentajes por tipo de atributo en regresión lineal.	66
11	Métricas de modelos basadas en regresión lineal.	66
12	Suma de porcentajes por tipo de atributo en árbol de decisión.	67
13	Métricas de modelos basadas en árbol de decisión.	67
14	Suma de porcentajes por tipo de atributo en GAM.	67
15	Métricas de modelos basadas en GAM.	68
16	Resumen WebScraping en GS.	72
17	Resumen WebScraping en GV.	72
18	Resumen WebScraping en GC.	72
19	Tabla comparativa de alcances en investigaciones	76
19	Tabla comparativa de alcances en investigaciones (<i>Continuación</i>). Fuente: Elaboración propia	77

INTRODUCCIÓN

Chile es un país con un vasto territorio, el cual posee diversas comunas y conurbaciones que componen puntos neurálgicos a lo largo del mismo. Dada la geografía comentada, muchas conurbaciones tienen características completamente distintas a otras, ya sea por su ubicación geográfica, el clima que poseen, la cultura de sus habitantes, entre otras. Estas diferencias permiten que afloren diversos cuestionamientos respecto a si ciertos factores que inciden de cierta manera en una u otra conurbación, lo harán de magnitud similar en otras. Es uno de estos cuestionamientos el que tomará profunda atención para la investigación, el cuál alude a los determinantes que influyen en el precio de venta de inmuebles, y su formulación es la siguiente: ¿Se aprecian factores distintos al momento de determinar el precio de un inmueble (ya sea casa o departamento) en una u otra conurbación de Chile?, ¿En algunas conurbaciones ciertos atributos influyen más intensamente que otros? Para resolver esta interrogante, se debe recorrer una variedad de herramientas que permitan identificar cierta “importancia” de atributos con respecto a otros.

Para lo anterior, es necesaria una herramienta de análisis de precios de inmuebles, de cuya estructura se pueda extraer cierta “importancia” de ciertos atributos por sobre otros. El modelo hedónico [Rosen, 1974] de precios propone la determinación de precios de inmuebles mediante la ponderación de atributos internos (habitaciones, metros cuadrados, baños, etc.) y atributos externos a la propiedad (distancia a servicios, densidad comunal, ubicación geográfica, etc.). Este modelo de precios puede ser incorporado a modelos de aprendizaje automático, y dependiendo del grado de interpretabilidad de estos, se puede extraer de su ajuste una ponderación de los atributos, permitiendo su comparación en términos de inferencia y desempeño.

En consecuencia, se proponen tres modelos de aprendizaje automático: Regresión lineal, árboles de decisión y *generalized additive models*, con los cuáles se aplicaran técnicas de inferencia para determinar la importancia de los atributos que inciden en el precio de una propiedad, sumado a una comparación del desempeño predictivo de los modelos a modo de validación. Se propone además que la extracción de los datos sea mediante *web scraping*, ya que no se dispone de una base de datos previa.

Se revisará en primer lugar un marco teórico que otorgará los fundamentos de la investigación, desde modelos de *ML* interpretables, su formulación matemática, modo de interpretar sus determinantes y métricas de desempeño, hasta el repaso de maneras éticas de realizar *web scraping* en Chile. Luego, se presentará la metodología, que consiste de cuatro fases: Extracción de datos mediante *web scraping*, extracción manual de datos, limpieza de datos y ajuste de modelos de *ML*. Posteriormente, se exponen y discuten los resultados obtenidos para finalmente concluir respecto a ellos.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

1.1. Contexto y Motivación General

Chile es un país con un vasto territorio, y en este, posee diversas comunas y conurbaciones que componen puntos neurálgicos a lo largo del mismo. Dada esta geografía, muchas conurbaciones tienen características completamente distintas a otras, ya sea por la ubicación geográfica de estas, el clima que poseen, la cultura de sus habitantes, entre otras. Estas diferencias permiten que surjan diversos cuestionamientos respecto a si ciertos factores que inciden de cierta manera en una u otra conurbación, lo harán de magnitud similar en otras. Es uno de estos cuestionamientos el que tomará profunda atención para la investigación, el cuál alude a los determinantes que influyen en el precio de inmuebles y su formulación es la siguiente: ¿Se aprecian factores distintos al momento de determinar el precio de un inmueble (ya sea casa o departamento) en una u otra conurbación de Chile? ¿En algunas conurbaciones ciertos atributos influyen más intensamente que otros? Para resolver esta interrogante, se debe recorrer una variedad de herramientas que permitan identificar cierta “importancia” de atributos con respecto a otros.

1.2. ¿Cómo determinar el precio de un inmueble en Chile y comparar los factores que inciden en este?

Si se va a la raíz del precio del inmueble, existen ciertos métodos que relacionan diversas variables para conducir a un precio final en Chile. Uno de los principales métodos para calcular el avalúo de un bien raíz es la tasación, existiendo muchos tipos, pero siendo pertinente al estudio la tasación comercial urbana mencionada en el manual de tasación [Nacionales, 2007], otorgando cierto peso a variables internas y externas del inmueble. Sin embargo, este tipo de tasación tiene dos defectos para la investigación. La primera razón se refiere a que no es transparente respecto a la manera en que se determinan los precios, por ende, no se puede analizar ni experimentar respecto a factores que inciden en los mismos. La segunda razón, alude a que se consideran variables complejas de medir (aptitudes del suelo, densidad de construcción, materiales, etc.).

Otro método para el cálculo del avalúo de un bien raíz (si bien no tan preciso como el anterior) es utilizar la familia de herramientas *online* (como [tasando.cl, 2022]). Este tipo de herramientas suelen basarse en comparar las características que se ingresan de determinada propiedad con otras similares en un rango cercano, sin embargo, tampoco consigue entregar detalles claros respecto a una comparativa de atributos.

Teniendo en cuenta la problemática anterior, se propone el modelo matemático hedónico

[Rosen, 1974] de precios de viviendas. Este modelo considera variables internas y externas de un inmueble, donde cada variable tiene un peso dentro de la ecuación. Ha sido una herramienta útil durante mucho tiempo para analizar cómo influyen los atributos que componen el precio de un inmueble, generalmente mediante una regresión lineal, ya que proporciona transparencia respecto a qué atributos influyen más y cuáles influyen menos en la predicción. Además, con el auge del aprendizaje automático a partir de la década de 1990, han surgido nuevas alternativas enfocadas en la misma función de encontrar factores que inciden en el precio de un inmueble, utilizando la lógica de atributos hedónicos e incluyéndolos en los procesos internos de los modelos. Con ello, se encuentra una herramienta útil para abordar esta investigación. Sin embargo, se han realizado pocos estudios comparativos entre modelos de aprendizaje automático enfocados en la importancia que ciertos factores tienen sobre otros, en lugar de centrarse principalmente en el rendimiento del modelo.

Continuando con la idea anterior, un motivo importante para fomentar el descubrimiento de la importancia de ciertas variables en la determinación de los precios de los inmuebles es la escasa investigación académica en Chile sobre este tema. La mayoría de los estudios se centran en una comuna en particular [Rosales, 2021], en la comparación de modelos construidos para mejorar la precisión en la predicción de los precios de los inmuebles, sin buscar la inferencia como objetivo principal [Masías *et al.*, 2016] [Bozanic, 2020], o incluso como un paso intermedio para la consolidación de un servicio de predicción de precios de los inmuebles [Reyes Feris, 2022]. Un estudio que aborda la inferencia es [Chardon Schirm, 2019]; sin embargo, se enfoca principalmente en la predicción dentro de Santiago de Chile y utiliza modelos que no ofrecen una interpretación intrínseca. Esto hace que el espectro de investigación en Chile se centre más en la comparación del rendimiento de los modelos que en investigar los factores de relación entre las variables que contribuyen al precio de un inmueble.

A nivel internacional, se observa una proporción similar, donde se abordan principalmente temas de comparación de modelos en términos de precisión, en lugar de un enfoque inferencial [Forys, 2022], [Lahmiri *et al.*, 2023]. También se utiliza el *web scraping* para extraer propiedades de sitios web, pero con propósitos diferentes al aprendizaje automático [Bricongne *et al.*, 2023]. Un enfoque más relevante para esta investigación sería el propuesto por [Rico-Juan y Taltavull de La Paz, 2021], donde se sugiere una comparación de modelos de aprendizaje automático interpretables utilizando herramientas hedónicas. Se destaca su comparación en términos de robustez. Sin embargo, aunque se acerca mucho al objetivo de esta investigación, sigue siendo una comparación basada en una sola comuna y la dicotomía se establece entre el aprendizaje automático y los métodos hedónicos, no entre modelos de aprendizaje automático en sí.

Dicho lo anterior, el problema se resume en: la ausencia de transparencia respecto a cómo se llega a los métodos para determinar el precio de un inmueble en Chile y si, dependiendo del lugar geográfico en que se encuentre la propiedad, esta métrica puede variar, dándole relevancia a atributos que en otros sectores no destacan. A su vez, existe ausencia de comparativa de modelos respecto a este fin. Por último, hay escasa investigación a nivel nacional

e internacional respecto al tema.

1.3. Extracción de datos

Cabe destacar que los atributos a utilizar para la investigación se limitan a los encontrados en el sitio web goplaceit.com, y a su vez, corresponden al enfoque hedónico, donde se encuentran:

- Tipo de propiedad (String)
- Comuna (String)
- Número de habitaciones (Int)
- Número de baños (Int)
- M^2 útiles (float)
- M^2 totales (float)
- Distancia a hospital (float)
- Distancia a Bomberos (float)
- Distancia a Supermercado (float)
- Distancia a Centro comercial (float)
- Distancia a Farmacia (float)
- Distancia a Bencinera (float)
- Densidad de población (de la comuna) (float)
- Tasa de delincuencia en trimestre anterior (en la comuna) (float)
- Precio en UF (float) (Objetivo)

Estos datos serán extraídos mediante técnicas de *web scraping*

Una vez reconocido el problema a afrontar y las aristas del mismo, se definen una serie de objetivos que guiarán la investigación.

1.4. Objetivos

1.4.1. Objetivo general:

- Aplicar técnicas de inferencia sobre modelos de *ML* para encontrar los atributos más relevantes respecto a la predicción de precios de venta de viviendas en las tres conurbaciones mas pobladas de Chile.

1.4.2. Objetivos específicos:

- Contrastar la inferencia realizada a modelos de aprendizaje automático en cada conurbación tratada.
- Analizar y comparar la inferencia realizada en base a los modelos de aprendizaje automático entre las distintas regiones abordadas.
- Visualizar los atributos más importantes por región y modelo en base a la inferencia realizada.
- Comparar modelos de aprendizaje automático utilizados en términos de precisión de predicción.

CAPÍTULO 2

MARCO CONCEPTUAL

A continuación, se llevará a cabo la definición y precisión de todos los conceptos necesarios para el correcto entendimiento de la investigación, desde conceptos demográficos sobre el área geográfica hasta conceptos teóricos relacionados con el aprendizaje automático o *Machine Learning* (ML en adelante) y planteamientos matemáticos de los modelos de ML a utilizar. Se hace hincapié en las herramientas y tecnologías utilizadas.

2.1. Conurbaciones en Chile

Una conurbación es un conjunto de poblaciones o comunas que, dado el crecimiento demográfico y de infraestructura, han conseguido “fusionarse”, de manera que los límites de una comuna terminan donde los límites de otra comienzan. En Chile, este fenómeno se puede observar a lo largo de todo el país y en diversas regiones. Las cinco conurbaciones más pobladas de Chile corresponden al Gran Santiago, Gran Concepción, Gran Valparaíso, Conurbación La Serena y Coquimbo y área metropolitana de Antofagasta. Para la investigación en curso se escogen las tres conurbaciones mas pobladas del país, decisión principalmente motivada por la disponibilidad de datos en dichas conurbaciones, delimitando a su vez los alcances de la investigación. Las comunas de las conurbaciones seleccionadas pueden verse a continuación (tablas 1 y 2):

Comunas Gran Santiago				
Cerrillos	La Granja	Ñuñoa	Quinta Normal	Vitacura
Cerro Navia	La Pintana	Padre Hurtado	Recoleta	
Conchalí	La Reina	Pedro Aguirre Cerda	Renca	
El Bosque	Las Condes	Peñalolén	San Bernardo	
Estación Central	Lo Barnechea	Pirque	San Joaquín	
Huechuraba	Lo Espejo	Providencia	San José de Maipo	
Independencia	Lo Prado	Pudahuel	San Miguel	
La Cisterna	Macul	Puente Alto	San Ramón	
La Florida	Maipú	Quilicura	Santiago	

Tabla 1: Comunas que forman parte del Gran Santiago.
Fuente: Elaboración propia.

Comunas Gran Valparaíso	Comunas Gran Concepción	
Concón	Chiguayante	Lota
Quilpué	Concepción	Penco
Valparaíso	Coronel	San Pedro de la Paz
Villa Alemana	Hualpén	Talcahuano
Viña del Mar	Hualqui	Tomé

Tabla 2: Comunas que forman parte del Gran Valparaíso y Gran Concepción.
Fuente: Elaboración propia.

2.2. Mercado inmobiliario Chileno actual

El estado del mercado inmobiliario en Chile se encuentra en un punto crítico, puesto que en 2022 los retiros de fondos AFP y ayudas fiscales propiciaron un aumento en la inflación, sumado a tasas mas restrictivas, aumento de precios de materiales y escasez de mano de obra, hicieron que el precio de venta de inmuebles estuviese muy por sobre lo esperado. Aún así, se espera que este año 2023 exista un repunte. En cuanto a la oferta nacional proporcionada [de la Construcción, 2022], suele existir una proporción de aproximadamente 85 % de oferta de departamentos y 15 % de oferta en casas.

2.3. Modelo hedónico de precios

Es una forma de asignarle un valor a un bien raíz, basado en que este último está constituido por diversos atributos. Con ellos, es posible establecer que su precio de mercado se puede explicar como un agregado de los precios individuales de cada atributo. El modelo permite identificar la importancia relativa de cada atributo en el valor asignado por el mercado a un bien raíz, mediante lo cual es posible determinar cómo cambiará dicho valor al variar la cantidad y calidad en que se encuentra presente cada uno de estos atributos y, consecuentemente, predecir precios, [Quesada, 2015].

Usualmente, este modelo suele representarse de manera lineal, logarítmica o cuadrática.

2.4. Goplaceit.com

Goplaceit.com es un sitio web internacional que posee dominio en Chile con Goplaceit.com/cl. Este permite la publicación de propiedades, ya sean departamentos o casas, para su arriendo o venta. A su vez, es posible buscar propiedades que ya han sido publicadas y acceder a diversas características de las mismas, como su precio, m^2 construidos, habi-

taciones e incluso características muy detalladas del entorno en que se encuentra, como distancia a estaciones de bomberos, a tiendas comerciales, a centros de salud, a bencineras, entre otros, todo en la medida de metros. Aquella última consideración es la más importante al escoger este sitio como aquel donde se extraerá la información de propiedades, ya que ofrece una gran cantidad de información en un formato estandarizado, tanto de variables internas como del entorno. Si bien existen sitios web como TocToc.com o portalinmobiliario.com que poseen la mayor cantidad de información inmobiliaria de Chile, su información por propiedad no se encuentra del todo estandarizada, principalmente limitada porque se analiza un radio fijo alrededor de la propiedad, usualmente entre 2 y 3 km. Y en caso de no encontrar servicios en ese rango, simplemente no se muestran, llevando a una pérdida de información.

2.5. Aprendizaje automático (ML)

Corresponde a un subconjunto de la inteligencia artificial, donde un sistema basado en un algoritmo aprenda patrones de los datos sin estar explícitamente programado para ello. Un modelo de *ML* corresponde a un archivo o código en algún lenguaje de programación que contiene un algoritmo de *ML*.

2.6. Problemas de regresión y clasificación

Un concepto clave a comprender antes de hablar de los modelos de *ML* como tal, es saber si un modelo sirve para regresión, para clasificación o para ambas tareas.

Los algoritmos de regresión intentan estimar una función intermediaria desde valores de entrada que conduzca a valores de salida numérica o continua. Esto indica que los problemas de regresión se orientaran a valores de salida como cantidades, tamaños, medidas, etc.

Por otro lado, los algoritmos de clasificación intentan estimar la función intermediaria entre las variables de entrada y variables de salida categóricas o discretas. Esto indica que problemas de clasificación se orientan a valores de salida como: “Gato”, “Mayor que X cantidad”, “Sí”, “No”.

Dado que el problema presentado radica en que la variable de salida es el precio de un inmueble, este es considerado como una cantidad continua, y por ende, un problema de regresión.

2.7. Interpretabilidad en ML

2.7.1. Definición de interpretabilidad

En primera instancia no existe una formalización matemática que sea capaz de condensar el concepto de interpretabilidad. Christoph Molnar [Molnar, 2020] corrobora esto mencionando que, al no haber una formalización matemática como tal, se guía por acercamientos vinculados a la definición de la palabra desde diversas perspectivas. Considerando una definición de diccionario, la palabra “interpretar” es entendida como “Explicar acciones, dichos o sucesos que pueden ser entendidos de diferentes modos” [academia española de la lengua, 2021]. Una definición de interpretabilidad alejada de toda formalidad matemática es la que menciona Tim Miller [Miller, 2017], pudiendo entenderse como la capacidad humana de entender el por qué de una decisión. Una definición asociada al contexto de ML reza lo siguiente “La interpretabilidad es el grado en que un humano puede predecir el resultado de un modelo de manera consistente” [Kim *et al.*, 2016]. Por último, se considera a la interpretabilidad como “Una noción específica del dominio, por lo cual no puede existir una definición que sea acorde a todos los propósitos” [Rudin, 2019].

La interpretabilidad bajo una definición sujeta al contexto del ML puede establecerse como la capacidad en que un humano es capaz de entender y explicar el procedimiento o lógica que utiliza un modelo para producir una predicción respecto a un conjunto de datos pre-determinado. Es decir, un modelo cuya estructura de decisión sea transparente para una persona. Sumado a lo anterior, la interpretabilidad en el contexto de ML es definida no como una máxima afirmativa o negativa, sino como una escala con diversos rangos, teniendo en un extremo una alta interpretabilidad y al otro extremo, una interpretabilidad muy baja o nula. Con ello, “un modelo es mas interpretable que otro si las decisiones que este toma para realizar el ajuste de los datos son mas sencillas de comprender para un ser humano que las decisiones del otro modelo.”[Molnar, 2020].

2.7.2. Enfoques *model-specific* y *model-agnostic*

Habiendo estructurado una definición acorde al contexto para la interpretabilidad, es necesario categorizarla y delimitarla en el marco de trabajo que conllevará esta investigación. La primera categorización alude a los enfoques *model-specific*, y *model-agnostic*, los cuales son definidos en base a los usos de estos. En el caso del enfoque *model-specific* se utiliza para interpretar las decisiones que toma un modelo de ML directamente de la estructura del mismo (por ejemplo: el análisis de las decisiones sobre la estructura en un árbol de decisión), mientras que el enfoque *model-agnostic* considera el modelo de ML como una caja negra, y se idea un modelo intermediario que sea capaz de aproximar los *outputs* que tenga el modelo complejo, de manera de interpretar este nuevo modelo, [Stiglic *et al.*, 2020].

La interpretación del enfoque *model-specific* está limitado a modelos en particular, y para

cada uno existe una manera de extraer explicaciones de su lógica examinando sus parámetros internos. [Du *et al.*, 2019]. “Métodos *model-agnostic* pueden ser usados en cualquier modelo de ML y se aplican usualmente *post hoc* (luego del entrenamiento del modelo con los datos) [Molnar, 2020] donde los parámetros internos del modelo no se inspeccionan, ya que el modelo se trata como una caja negra”, [Du *et al.*, 2019].

Las ventajas que ofrece un enfoque *model-specific* radican en que permiten obtener un entendimiento mas profundo de la decisión otorgada por un modelo al conocer como funciona el modelo con sus parámetros internamente. Su desventaja radica en que para modelos más complejos, es difícil utilizar este método considerando que la estructura del modelo es de por sí compleja y las decisiones que toma están ocultas por extensos cálculos internos.

Las ventajas que ofrece un enfoque *model-agnostic* es que son mucho mas flexibles que los de enfoque *model-specific*, puesto que no dependen de un modelo en particular para utilizarse. La desventaja principal es que el enfoque *model-agnostic* realiza una aproximación de características en base al input y output entregado, por lo que puede que resulte en simplificar de sobremanera aquello que se quería buscar en un inicio,[Carrillo *et al.*, 2021].

En base a los dos métodos propuestos, se utilizará el enfoque *model-specific* para conceptos de inferencia sobre la estructura de los modelos, sin embargo, también se dará uso a enfoques *model-agnostic* para analizar los modelos desde un punto de vista de precisión y predicción.

2.7.3. Enfoques locales y globales

Este tipo de enfoques, se diferencian en la manera que tienen respecto a su enfoque de explicabilidad. Estos métodos pueden centrarse en explicar una predicción en particular, conocido como interpretabilidad local [Molnar, 2020], o ir más allá, pudiendo comprender totalmente el comportamiento del modelo y sus decisiones, conocido como interpretabilidad global [Stiglic *et al.*, 2020]. “Las interpretaciones locales de los modelos pueden realizarse diseñando arquitecturas justificadas de modelo en el cual se explique explícitamente de que manera una decisión fue realizada”[Stiglic *et al.*, 2020]. Por otro lado, los modelos reconocidos como globalmente interpretables, contienen una transparencia respecto a su estructura de manera que se puede saber que ocurre en su interior en un ámbito mas abstracto [Du *et al.*, 2019]. Para explicar la salida global que pueda tener un modelo, se necesita primero entrenar dicho modelo, conocer de antemano como funciona el algoritmo, que lo caracteriza y por ultimo conocer la estructura de los datos (pudiendo ser lineal, o tener una forma polinómica) [Lipton, 2016].

Dicho lo anterior, el enfoque a utilizar para esta investigación será global, ya que el objetivo es analizar los determinantes de precios de inmuebles en general (dentro de las regiones establecidas), no analizar cada predicción individualmente.

2.7.4. Modelos altamente interpretables

Una vez entendida la definición de interpretabilidad en el contexto de *ML*, además de los métodos en los que esta se categoriza, es menester poder aplicar dicha definición para seleccionar modelos acordes a la investigación.

En primera instancia, se considera que los algoritmos que son mas simples en su estructura son usualmente mas interpretables que modelos mas complejos, sin embargo no se puede realizar una generalización, y cada caso debería ser analizado por separado [Schmitt, 2022]. Yendo en particular a un tipo de algoritmo, [Hayunga, 2020] asevera que “un árbol de regresión es bastante fácil de entender, sin perder la capacidad de poder identificar y representar un patrón complejo”, destacando además la capacidad de este tipo de algoritmos de capturar relaciones no lineales sin perder el entendimiento de la estructura y lógica que conllevaron a tomar una decisión. [Stiglic *et al.*, 2020] propone la siguiente ilustración (véase Figura 1), la cuál relaciona los tipos de algoritmos dependiendo de la categoría de interpretabilidad que poseen (categorías que fueron detalladas en subsecciones anteriores).

	Global	Local
Model-specific	- Decision trees (depends on depth and number of terminal nodes; Hastie, Tibshirani, & Friedman, 2009; Stiglic, Kocbek, Pernek, & Kokol, 2012), - Linear and logistic regression models (Harrell Jr, 2015), - Generalized linear models (GLM) and generalized additive models (GAM; Hastie <i>et al.</i>, 2009), - Naive Bayes classifier (Kononenko, 1993), - GNNExplainer (Ying <i>et al.</i>, 2019)	- Set of rules (for specific individual; Visweswaran, Ferreira, Ribeiro, Oliveira, & Cooper, 2015), - Decision trees (by tree-decomposition; Visweswaran <i>et al.</i>, 2015), - Visual analytics-based approaches (interactive visualization techniques for interpretation focusing on individual prediction), - k-Nearest neighbors (k-NN; depends on the number of important features, retrieving k-nearest neighbors for interpretation; Yuwono <i>et al.</i>, 2015), - GNNExplainer (Ying <i>et al.</i>, 2019)
Model-agnostic	- Different variants of model compression/knowledge distillation/global surrogate models (Elshawi, Al-Mallah, & Sakr, 2019), - Partial Dependence Plots (PDP; Elshawi, Al-Mallah, & Sakr, 2019), - Individual Conditional Expectation (ICE) plots (Elshawi, Al-Mallah, & Sakr, 2019), - Black Box Explanations through Transparent Approximations (BETA; Lakkaraju, Kamar, Caruana, & Leskovec, 2017) - Model understanding through subspace explanations (MUSE; Lakkaraju <i>et al.</i>, 2019).	- Local interpretable model-agnostic explanations (LIME; Ribeiro <i>et al.</i>, 2016), - Shapley additive explanations (SHAP; Lundberg & Lee, 2017), - Anchors (Ribeiro <i>et al.</i>, 2018), - Attention map visualization, - Model understanding through subspace explanations (MUSE; Lakkaraju <i>et al.</i>, 2019).

Figura 1: Ejemplos de acercamientos a interpretabilidad de predicción en modelos de regresión.

Fuente: [Stiglic *et al.*, 2020].

Un punto importante a considerar en la la imagen (véase figura 1), es que detalla claramente modelos a considerar en caso de utilizar las categorías de *model-specific* y método global de interpretabilidad, dentro de los cuales considera (para regresión) a los árboles de decisión (dependiendo de su profundidad y el número de nodos), la regresión lineal, modelos lineales generalizados y modelos aditivos generalizados. El resto de los que se mencionan en la categoría mencionada son modelos acordes para las tareas de clasificación.

Por último, [Molnar, 2020] menciona en su libro que una manera bastante sencilla de alcanzar un nivel alto de interpretabilidad es utilizar modelos adecuados para ello, y presenta la regresión lineal, logística, modelos lineales generalizados, modelos aditivos generalizados, árboles de decisión, *RuleFit* principalmente (al menos para el caso de que se busque interpretabilidad basada en el paradigma *model-specific*).

Considerando las fuentes analizadas hasta este punto, se decide optar por la utilización de tres modelos, los cuales se enfocan en el enfoque hacia problemas de regresión, los cuales son: Regresión lineal, Métodos aditivos generalizados (que son la generalización de modelos de orden lineal o superiores) y árboles de decisión. En mayor detalle, se escogen por las siguientes razones: Se realiza un primer filtro en torno a modelos que sean útiles para problemas de regresión, luego se comprime dicho conjunto al considerar cuales de esos modelos responden a modelos altamente interpretables. Sumado a ello, se consideran modelos que responden a los enfoques de *model specific* y globales. Dicho proceso se ilustra a continuación (figura 2).

Selección de modelos de ML para el estudio

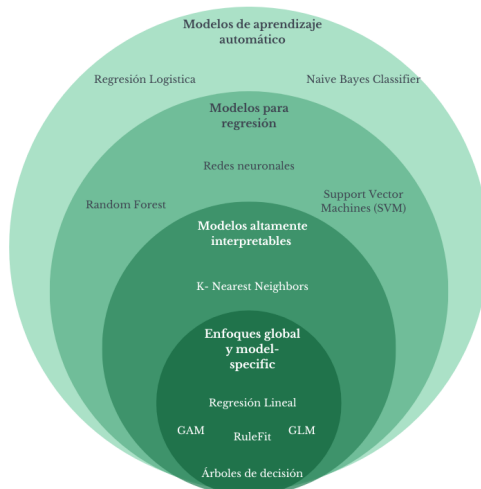


Figura 2: Selección de modelos de ML dadas las necesidades del estudio.

Fuente: Elaboración propia.

Con dichas restricciones los modelos a escoger se reducen a cinco: Regresión lineal, *generalized linear models*, *generalized additive models*, árboles de decisión, *RuleFit*. Tanto regresión lineal, *generalized linear models* y *RuleFit* se utilizan para detectar relaciones lineales en los datos, por lo que solo se escoge la regresión lineal por ser el método clásico al momento de catalogar relaciones lineales. Luego, los árboles de decisión no se comportan bien del todo al observar relaciones completamente lineales, por lo cual pueden aportar una excelente perspectiva que contraste con la regresión lineal. Por último, *generalized additive models* comprenderá un punto medio entre ambos casos, al ser mucho mas flexible que los otros dos modelos mencionados anteriormente, aunque a costa de pérdida de cierta interpretabilidad.

2.8. Inferencia y predicción en ML

En primer lugar, [Kumar, 2021] menciona que un modelo es usado para predecir el valor de respuesta de una variable basado en variables predictivas. Por otro lado, se habla de

inferencia cuando el modelo se utiliza para entender la relación existente entre las variables predictivas y variables de respuesta.

La predicción se centra en el valor como tal de la respuesta que produce la entrada de un modelo de *ML*. Inferencia en cambio va más allá. Esta se hace las preguntas ¿Qué elementos de salida se relacionan con qué elementos de entrada?, ¿Cómo cambia la magnitud de la salida respecto a algún cambio en las variables de entrada?

Ya que se definió anteriormente la interpretabilidad, esta puede vincularse al concepto de inferencia en este punto, ya que un modelo altamente interpretable permitirá realizar inferencia y contestar sus preguntas mas comunes, ya que se conocerá el proceso que utiliza el modelo para categorizar los datos.

Aún así, no son necesariamente excluyentes. Es posible obtener un desempeño alto con un enfoque en la predicción y aún así poder efectuar inferencia de manera correcta.

Dicho lo anterior, el enfoque de esta investigación se centrará en encontrar un buen desempeño predictivo, y al mismo tiempo realizar inferencia sobre los modelos para entender por qué toman determinadas decisiones.

2.9. Modelos de *ML* seleccionados

A continuación se presenta un acercamiento teórico a los modelos de *ML* escogidos:

2.9.1. Regresión lineal

Como primer punto, un análisis de regresión consiste en un método estadístico que investiga relaciones entre una variable y muchas otras. A su vez, un modelo estadístico alude a una descripción sencilla de cierto proceso [Yan y Su, 2003]. Dicho esto, existen tres tipos de regresión, la primera denominada regresión lineal simple (vincula la relación entre dos variables), la segunda es regresión lineal múltiple, donde una variable dependiente se puede asociar a muchas variables independientes y el tercer tipo es una regresión no lineal. En esta investigación se dará énfasis al segundo tipo de regresión.

Regresión lineal múltiple corresponde a un modelo de regresión lineal en el cuál el modelo posee una variable dependiente, la cuál se relaciona a mas de una variable independiente.[Yan y Su, 2003] Matemáticamente, la expresión general de este tipo de regresión es la siguiente (ecuación 1):

$$y_i = \beta_{i0} + \beta_{i1}x_{i1} + \dots + \beta_{ip}x_{ip} + \epsilon_i \quad (1)$$

Donde “y” corresponde a la variable dependiente, $\beta_0, \beta_1, \dots, \beta_p$ son denominados “coeficientes de regresión”, $x_1, x_2, \dots, x_p$ las variables dependientes y ϵ corresponde al error aleatorio. Finalmente, el subíndice i representa el número de observación, mientras que p alude al número total de características a observar.

Este tipo de regresión puede ser representada de manera matricial, de manera de obtener propiedades valiosas para su funcionamiento. Reescribiendo la ecuación 1 de manera matricial, se obtiene (ecuación 2):

$$y = X\beta + \epsilon \quad (2)$$

Donde:

$$X = \begin{pmatrix} x_{11} & x_{12} \dots & x_{1p} \\ x_{21} & x_{22} \dots & x_{2p} \\ \dots & \dots & \dots \\ x_{i1} & x_{i2} \dots & x_{ip} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_i \end{pmatrix} \quad (3)$$

Un punto clave a mencionar es respecto a la estimación de mínimos cuadrados, ya que a partir de aquello es posible realizar un ajuste lineal de los datos para el modelo de regresión. Mínimos cuadrados consiste en minimizar la suma de los cuadrados de los residuos. Estos residuos pueden representarse de la siguiente manera:

Para regresión lineal múltiple sin forma matricial, se despeja la ecuación 1 (nótese que se introduce la notación r_i aludiendo al residuo (ecuación 4):

$$y_i - (\beta_{i0} + \beta_{i1}x_{i1} + \dots + \beta_{ip}x_{ip}) = e_i \quad (4)$$

Sin embargo, esta ecuación considera el valor teórico de y_i , y se requiere el valor de la misma con la predicción realizada para poder contrastarla con los valores observados de los parámetros. Es por esto que la ecuación 4 se modifica desde teórica a práctica de manera que queda la siguiente expresión (ecuación 5):

$$y_i^* - (\beta_{i0} + \beta_{i1}x_{i1} + \dots + \beta_{ip}x_{ip}) = b_i \quad (5)$$

Donde y_i^* corresponde a la función predicha para los parámetros y variables independientes, mientras que ϵ se intercambia por la notación b para denotar el cambio entre teórico y predicho.

Dicho aquello, la función de mínimos cuadrados queda definida por (ecuación 6):

$$\sum_{n=0}^i b_i^2 \quad (6)$$

y en forma matricial (ecuación 7):

$$b = \operatorname{argmin}_{\beta} [(y - X\beta)'(y - X\beta)] \quad (7)$$

Donde el superíndice “'” corresponde a la función luego de la predicción.

Por último, destacar ciertas ventajas y desventajas teóricas respecto a este modelo en particular. En primer lugar, una ventaja es la estructura sencilla de la regresión lineal, la cuál la hace absolutamente transparente en como se realizan las predicciones. Además, considerar que es un modelo bastante utilizado y por ende ampliamente implementado en diversos lenguajes con librerías actualizadas. En cuanto a las desventajas, la mas importante es que el modelo solo es capaz de modelar relaciones lineales, a pesar de que la estructura pueda ser distinta. Esto también influye en que la predicción puede que no sea muy precisa.

Como comentario final, este modelo representará el enfoque hedónico mas utilizado en el análisis de precios de propiedades en base a determinados atributos.

2.9.2. *Generalized additive models*

Es un tipo de modelo lineal generalizado que busca aprender relaciones no lineales entre los datos. La linealidad es bastante difícil de obtener, principalmente porque mientras mas aumenta el número de atributos (o variables independientes) que se desean observar, mas cabe la posibilidad de que la relación entre ellas no sea lineal, [Molnar, 2020].

En cuanto a su formalidad matemática, este tipo de modelo relaja la restricción lineal de que la relación entre la variable dependiente con las independientes debe ser una suma simple con pesos (los parámetros que acompañan a cada variable independiente). En lugar de ello, se asume que la relación que existe puede ser modelada como una suma de ciertas funciones que representan cada característica, viéndose de la siguiente forma (ecuación 8):

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_p(x_p) \quad (8)$$

Las funciones f_i pueden ser funciones con forma paramétrica previamente especificada (por ejemplo, algún tipo de polinomio) o incluso pueden llegar a a especificarse de manera no paramétrica o semi paramétrica.

¿Cómo se realizan estos aprendizajes de funciones no lineales? El método original de ajuste se basaba en el uso de “suavizantes no paramétricos” (aludiendo a la suavidad en la representación de una función) mediante un algoritmo de *backfitting* [Hastie T, 1990]. Este funciona en base a una suavización iterativa sobre los residuos de las funciones (como los vistos en la sección de regresión lineal). Otro método utilizado, es en base al concepto denomi-

nado “splines”. Este concepto alude a funciones que pueden ser combinadas de manera de aproximar funciones arbitrarias.

A continuación, se destaca una diferencia respecto al ajuste que tiene sobre los datos un modelo basado en GAM, comparado a modelos de regresión lineal, donde se observa que hay un ajuste mucho mas fino respecto a los datos (figura 3):

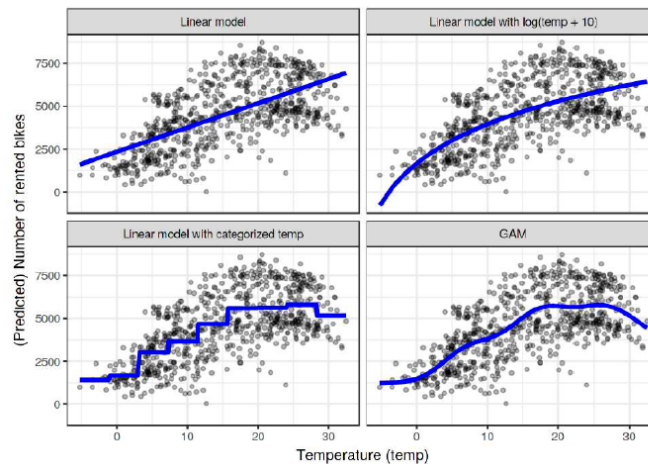


Figura 3: Comparación de GAM con regresiones lineales.

Fuente: [Molnar, 2020].

Hablando respecto a ventajas de este tipo de modelo es posible señalar la gran flexibilidad y adaptabilidad a los datos.

Respecto a las desventajas, este modelo es menos interpretable que uno lineal, por ende se debe intentar no complejizarlo en demasía para obtener resultados mas interpretables.

2.9.3. Árboles de decisión

Los árboles de decisión son un tipo de algoritmo que se encarga de realizar una predicción desde un espacio de entrada a un espacio de salida, $h : X \rightarrow y$, buscando predecir una instancia de x a medida que se viaja desde un nodo raíz hasta un nodo hoja. Considerando cada uno de los nodos inmersos en el camino desde la raíz hasta la hoja, el hijo sucesor se elige en base a una partición respecto al espacio de entrada de la variable independiente (x) en cuestión. El criterio de partición se decide en base tanto a características de x como a reglas definidas con anterioridad [Shalev-Shwartz Shai, 2014].

Un ejemplo de la partición del espacio de características de las variables se observa a continuación (figura 4), donde X_1 y X_2 son las variables independientes, R_1, R_2, R_3, R_4, R_5 son las regiones en las cuales se dividió el espacio y t_1, t_2, t_3, t_4 son las particiones en particular.

Los árboles de decisión pueden dividirse en dos: árboles de clasificación y árboles de deci-

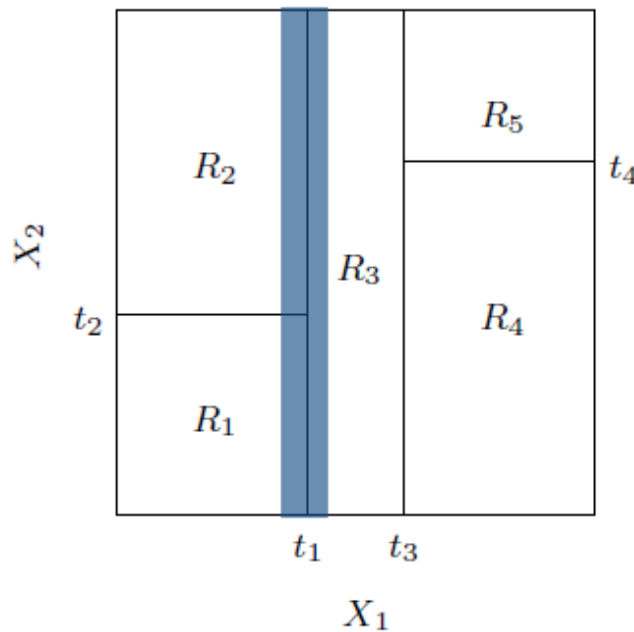


Figura 4: Partición de espacio de características.
Fuente: [Ñanculef R., 2022].

sión. Ambos se diferencian en las salidas que pueden ofrecer, siendo para el primer caso una clase o categoría asociada a los datos, mientras que el segundo caso corresponde a que la salida esté en el espacio de los números reales [Bozanic, 2020].

Un ejemplo de árbol de decisión enfocado en la regresión puede verse a continuación (figura 5), en la cuál se puede observar cómo la estructura del árbol va segmentándose a medida que se cumplen ciertas condiciones de las variables (x_1 y x_2), donde cada nodo que no es una hoja corresponde a una decisión, mientras que las hojas del árbol representan valores continuos finales. Cabe destacar que para esta investigación se utilizará la definición binaria de árbol de decisión, que corresponde a que un nodo solo puede dividirse en dos ramificaciones.

Este problema, de manera idéntica a la regresión lineal, se esmera en reducir la suma del error cuadrático respecto a todas las subdivisiones del árbol, siendo su expresión matemática (ecuación 9):

$$\sum_{j=1}^J \sum_{i \in R_j} (Y_i - Y_{R_j}^*)^2 \quad (9)$$

Donde i corresponde a una observación en particular y j representa la región de las variables independientes, para finalmente Y_i representar la función esperada teóricamente y $Y_{R_j}^*$ el promedio de la predicción obtenida para cada ejemplo en un nodo.

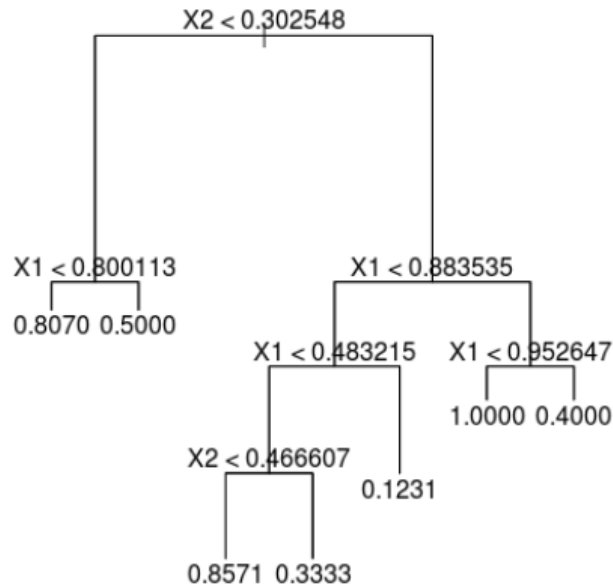


Figura 5: Árbol de decisión orientado a la regresión.
Fuente: [Trajan, 2020].

Existen diversos enfoques para los árboles de decisión, en particular se utilizará el enfoque CART [Ñanculef R., 2022] el cuál es un algoritmo de construcción del árbol que aborda tanto problemas de regresión como de clasificación. Es importante este punto ya que para la presente investigación se utilizará un enfoque de regresión.

Ventajas de los árboles de decisión son diversas, principalmente las vinculadas a que la estructura del árbol está directamente realizada para capturar interacciones entre características de los datos. Además, su visualización hace natural el proceso de obtención de las variables mas relevantes para un problema.

Respecto a las desventajas, un árbol de decisión tendrá problemas trabajando con relaciones completamente lineales. Sumado a ello, el árbol no es muy suave a la hora de representar los datos.

2.10. Cálculo de “Importancia de atributos” en ML

Descritos los modelos de ML a utilizar, es menester analizar de qué manera se puede medir la importancia de un atributo y que tanto influye en la variable objetivo. Cada modelo propuesto anteriormente posee una forma de calcular o medir dicho impacto, los cuales son presentados a continuación.

2.10.1. “Importancia de atributos” en regresión lineal

Refiriéndose a la regresión lineal, la “importancia de atributos” consiste en la magnitud de cada uno de los coeficientes (valor absoluto de los mismos) donde mientras mayor sea, mas importante será [Anand, 2020]. Esto se muestra a continuación (ecuación 10):

$$F_i = |\beta_i| \quad (10)$$

Donde F_i representa la importancia del atributo e “i” itera por cada uno de los coeficientes de regresión, calculando de esta manera cada una de las importancias del atributo.

2.10.2. “Importancia de atributos” en GAM

Al tener una estructura similar, se pensaría que GAM y regresión lineal comparten una forma igualmente similar de calcular importancia para sus atributos. Lamentablemente, GAM no posee una fórmula que permita dicho cálculo de importancia, ni menos examinar coeficientes para su potencial interpretación. Sin embargo, puede utilizarse un método denominado *PDP (Partial Dependency Plots)* [Molnar, 2020] que, para cada uno de los atributos involucrados, se realiza un gráfico cuyo eje vertical corresponde al cambio en el promedio de la variable objetivo debido al eje horizontal, el cual corresponde al cambio del valor de dicha variable. Viendo el gráfico en su totalidad, este indica la función que realiza el modelo GAM para cada uno de los atributos. Sumado a ello, se establece como medida de importancia el valor absoluto de la resta entre los extremos del intervalo del promedio de los valores de la variable objetivo. Un ejemplo se puede observar a continuación:

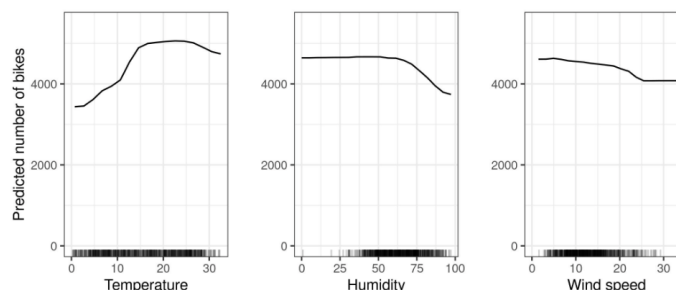


Figura 6: Ejemplo de PDP.

Fuente: [Molnar, 2020].

En la imagen(figura 6) se observan tres *PDP* de ejemplo, donde el eje vertical corresponde a la predicción del número de bicicletas observadas en una localidad, el eje horizontal en cada gráfico representa el cambio en un atributo, en este caso siendo Temperatura, humedad y velocidad del viento. De allí, una medida de importancia es calcular que tanto varió la variable objetivo analizando el rango de esta. Con ello en mente, es posible establecer que

la temperatura es “más importante” que la velocidad del viento en determinar el número de bicicletas circundantes, ya que la variable objetivo varió más con este atributo.

La desventaja de este enfoque, es que no es una manera intrínseca de GAM para obtener interpretabilidad, sumado a que para establecer *PDPs* es necesario dejar todos los atributos restantes que no están siendo analizados en el gráfico constantes, eliminando el factor de incidencia de otras variables en el resultado.

2.10.3. “Importancia de atributos” en árboles de decisión (CART, Random Forest y XG-Boost)

La importancia del atributo para el caso de árboles de decisión en regresión se nutre de un concepto llamado “impureza de un nodo” [Subramanian, 2021] el cual cuantifica el grado en que ciertas muestras pertenecen a diferentes clases.

Matemáticamente se utiliza *Mean Squared Error* (Error cuadrático medio) para medir la impureza en regresión de la forma explicada anteriormente en 9.

Teniendo dicha impureza en mente, se plantea la siguiente formulación (ecuación 11) basada en [Subramanian, 2021] :

$$ni_j^* = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right} \quad (11)$$

la cuál se encarga de calcular la importancia para un nodo, donde

- ni_j es la importancia del nodo j
- w_j peso de muestras que alcanzan el nodo j
- C_j El valor de impureza del nodo j
- $left(j)$ el nodo hijo de la partición izquierda del nodo j
- $right(j)$ el nodo hijo de la partición derecha del nodo j

Finalmente, la importancia de todas las variables se calcula como (ecuación 12):

$$fi_i = \frac{\sum j : jparticionesenatributoi * ni_j}{\sum k : todoslosnodos * ni_k} \quad (12)$$

2.11. Métricas de evaluación de desempeño de modelos de regresión

Una métrica de evaluación de modelos es una medida numérica utilizada para evaluar el rendimiento y la calidad de un modelo de *ML*. Existen diversos tipos de métricas basado en el enfoque que se le da al problema, sin embargo a continuación se exhiben las mas utilizadas [Hasan Parvez, 2023] y las que se utilizaran en esta investigación.

2.11.1. R^2

“ R^2 , también conocido como coeficiente de determinación, es un valor entre 0 y 1 que mide que tan bien el modelo de regresión se ajusta a los datos. R^2 puede ser interpretado como el porcentaje de variación en la variable dependiente del modelo. Mientras mas cerca R^2 se encuentre de 1 (o 100 %), de mejor manera el modelo será prediciendo la variable dependiente.”, [D’Auria, 2021].

El cálculo de esta métrica se reduce a lo siguiente (ecuación 13):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

donde y_i corresponde a la instancia “i” del valor real de la variable objetivo. $\hat{y}_i$ corresponde al valor predicho en la instancia “i” de la variable objetivo. $\bar{y}$ corresponde a la media de los valores predichos.

2.11.2. *Mean Squared Error y Root Mean Squared Error*

“*Mean squared error (MSE)* mide la cantidad de error en modelos estadísticos. Este evalúa la diferencia promedio entre los valores observados y predichos. Cuando un modelo no tiene error, la métrica equivale a 0. A medida que el error del modelo incrementa, el valor de esta métrica incrementa.”[Frost, 2023].

El cálculo de *MSE* se reduce a la siguiente forma (ecuación 14):

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (14)$$

Esta métrica se utiliza para penalizar la aparición de valores atípicos, ya que castiga las diferencias mas notorias. Aún así, existe preferencia por la métrica *Root Mean squared error (RMSE)* la cuál aumenta un poco la interpretabilidad de RMSE incluso penalizando mas los valores atípicos.

La ecuación de RMSE quedaría de la siguiente forma (ecuación 15):

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (15)$$

2.11.3. Mean Absolute Error

“Mean Absolute Error (MAE) es una medida del tamaño promedio de errores en una colección de predicciones, sin considerar su dirección en el proceso. Es medida como la diferencia absoluta promedio entre los valores predichos y los valores reales.” [Deepchecks, 2023] Su ecuación (ecuación 16) corresponde a la siguiente:

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (16)$$

Esta medida suele ser muy interpretable dada la estructura de su cálculo.

2.12. Cross-Validation

“Cross-Validation o CV [Alhamid, 2020] es un procedimiento que provee la habilidad de estimar el desempeño del modelo en datos que aún no se han visto durante el entrenamiento.

Este tiene dos pasos principales: dividir los datos en subconjuntos y rotar el entrenamiento y testing entre ellos. La técnica de división de conjuntos tiene las siguientes propiedades:

- Cada subconjunto tiene aproximadamente el mismo tamaño
- Los datos pueden ser seleccionados aleatoriamente en cada subconjunto
- Todos los subconjuntos son usados en el conjunto de entrenamiento excepto uno, que se utilizará con propósito de pruebas. Este subconjunto de pruebas será rotado hasta que todos los subconjuntos hayan sido una vez conjunto de pruebas (solo una vez).
- Es recomendable que cada ejemplo único sea contenido en solo un subconjunto.”

A continuación se muestra una figura que ejemplifica el procedimiento de CV (figura 7):

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Fold-6	Fold-7	Fold-8	Fold-9	Fold-10
Step-1	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test
Step-2	Train	Train	Train	Train	Train	Train	Train	Train	Test	Train
Step-3	Train	Train	Train	Train	Train	Train	Train	Test	Train	Train
Step-4	Train	Train	Train	Train	Train	Train	Test	Train	Train	Train
Step-5	Train	Train	Train	Train	Train	Test	Train	Train	Train	Train
Step-6	Train	Train	Train	Train	Test	Train	Train	Train	Train	Train
Step-7	Train	Train	Train	Test	Train	Train	Train	Train	Train	Train
Step-8	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
Step-9	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
Step-10	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train

Figura 7: Cross-Validation.

Fuente: [Alhamid, 2020].

2.13. Python

Python es un lenguaje de programación interpretado de alto nivel. Gracias a que es un lenguaje interpretado otorga facilidad para realizar *debugging*, propiciando la productividad y ahorrando tiempo al momento de tener errores.

Este lenguaje es caracterizado principalmente por la facilidad de su sintaxis y la gran cantidad de librerías que dispone para procesamiento de datos, análisis de datos, *ML*, *web scraping*, *Big Data*, entre otras. Considerando que la presente investigación requiere del uso varios de los tópicos anteriormente mencionados, Python se convierte en un lenguaje de programación esencial dadas las necesidades latentes.

2.14. Web scraping

Web scraping se refiere a la recolección de datos a través de cualquier medio que no sea un ser humano extrayendo información manualmente mediante la interacción con un sitio web o un programa interactuando con una API. Esto se puede conseguir mediante el desarrollo de un programa que ejecuta consultas constantes sobre un servidor web, solicita la información y la analiza gramaticalmente para extraer lo necesitado.

Web scraping también engloba diversas técnicas de programación y distintas librerías dependiendo del lenguaje de programación que se utilice para ello. Dentro de las técnicas mas usadas se utiliza análisis de datos, análisis gramatical de lenguaje natural y seguridad de la información [Mitchell, 2018].

2.14.1. Librerías de web scraping

Se presentan una variedad de librerías de *web scraping*, que sirven para diversos propósitos dependiendo de las necesidades requeridas en el lenguaje de programación Python [Reyes Feris, 2022].

- *Beautiful Soup*: Librería que permite desarrollar *web scrapers* en base a la estructura HTML y CSS y sus respectivas etiquetas para localizar la información deseada.
- *Scrappy*: Similar a *beautiful soup* en su modo de operar, sin embargo permite funciones mas avanzadas.
- *Selenium*: Permite realizar *web scraping* más avanzado al considerar estructura Javascript y permitiendo una gran variedad de automatizaciones en sitios web.

2.15. Legalidad del web scraping en Chile

Ya que el *web scraping* consiste en la extracción de información desde una página web de manera automatizada, se deben establecer los límites legales que abarca su uso.

Existe principalmente una ley que abarca de estos temas, siendo Ley N°21.459, la cual reza “Establece normas sobre delitos informáticos, deroga la ley N° 19.223 y modifica otros cuerpos legales con el objeto de adecuarlos al convenio de Budapest”[del congreso nacional de Chile, 2022b] en el cuál corresponde mencionar al artículo 2:

“Artículo 2°.- Acceso ilícito. El que, sin autorización o excediendo la autorización que posea y superando barreras técnicas o medidas tecnológicas de seguridad, acceda a un sistema informático será castigado con la pena de presidio menor en su grado mínimo o multa de once a veinte unidades tributarias mensuales.

Si el acceso fuere realizado con el ánimo de apoderarse o usar la información contenida en el sistema informático, se aplicará la pena de presidio menor en sus grados mínimo a medio. Igual pena se aplicará a quien divulgue la información a la cual se accedió de manera ilícita, si no fuese obtenida por éste.

En caso de ser una misma persona la que hubiere obtenido y divulgado la información, se aplicará la pena de presidio menor en sus grados medio a máximo.”

2.15.1. Acceso a la información de Goplacit.com

Una vez entendido el artículo 2 de la ley N°21.459, cabe investigar que considera Goplacit.com como ilícito dentro de su búsqueda de información, donde se obtiene lo siguiente:

Fragmento de condición número 8 en Términos y condiciones de Goplacit.com ,[Goplacit.com, 2022].

“Los Usuarios tienen derecho a revisar toda la información publicada en el Sitio, pudiendo

descargarla, copiarla o imprimirla solo para fines personales, estando estrictamente prohibida su utilización para fines comerciales.”

Y por último, hay una distinción entre usuario y usuario registrado. Un usuario simple es considerado una visita:

“Visita: Usuario que accede a una o más páginas de información publicada en el Sitio, no exigiéndose su identificación para dicho acceso.”[Goplacelt.com, 2022].

Por ende, al realizar web scraping no se está obteniendo información personal de cada persona que haya publicado, simplemente se busca la información pública y con fines personales, ya que no se dispondrá la fuente de los datos, lo cuál no está penalizado.

2.16. *Jupyter Notebook*

Es un entorno virtual basado en navegadores de la web para crear documentos homónimos. Estos documentos están en formato JSON y están compuestos por un sistema de versiones, de manera de que poder volver a fases anteriores del desarrollo en caso de presentar algún inconveniente en versiones posteriores. También, contiene una lista de celdas, donde cada una posee entradas y salidas, permitiendo ejecutar código o texto en formato *Markdown* de manera independiente.

Jupyter notebook utiliza ciertos núcleos de computación, bajo los cuales permite programación en diversos lenguajes, donde los mas comunes son *Python* y *R*.

Hay dos ventajas principales al trabajar en este tipo de entornos, la primera alude al orden y legibilidad del código al poder ejecutarlo en celdas por separado, incluso pudiendo escribir texto entre celdas para otorgarle un estilo profesional. El segundo punto refiere a la facilidad de realizar *debug* sobre el código, ya que al estar mas ordenado bajo secciones, es posible reconocer mucho mas rápido las causas de un error.

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

3.1. Metodología propuesta

El procedimiento a plantear tiene como fin poder encontrar los atributos (ya sea intrínsecos o externos) que mas influyen en el precio de un inmueble (casa y departamento) mediante ML, comparando dichos modelos para obtener el resultado mas preciso posible. Para lo anterior es necesaria la elaboración de un plan a seguir, de manera de asegurar que cada parte de este obedezca al fin planteado. La siguiente figura (figura 8) condensa los puntos clave de la metodología, subdivididos en fases. Es relevante anunciar que la metodología propuesta se llevará a cabo en archivos Jupyter Notebook en el lenguaje Python.

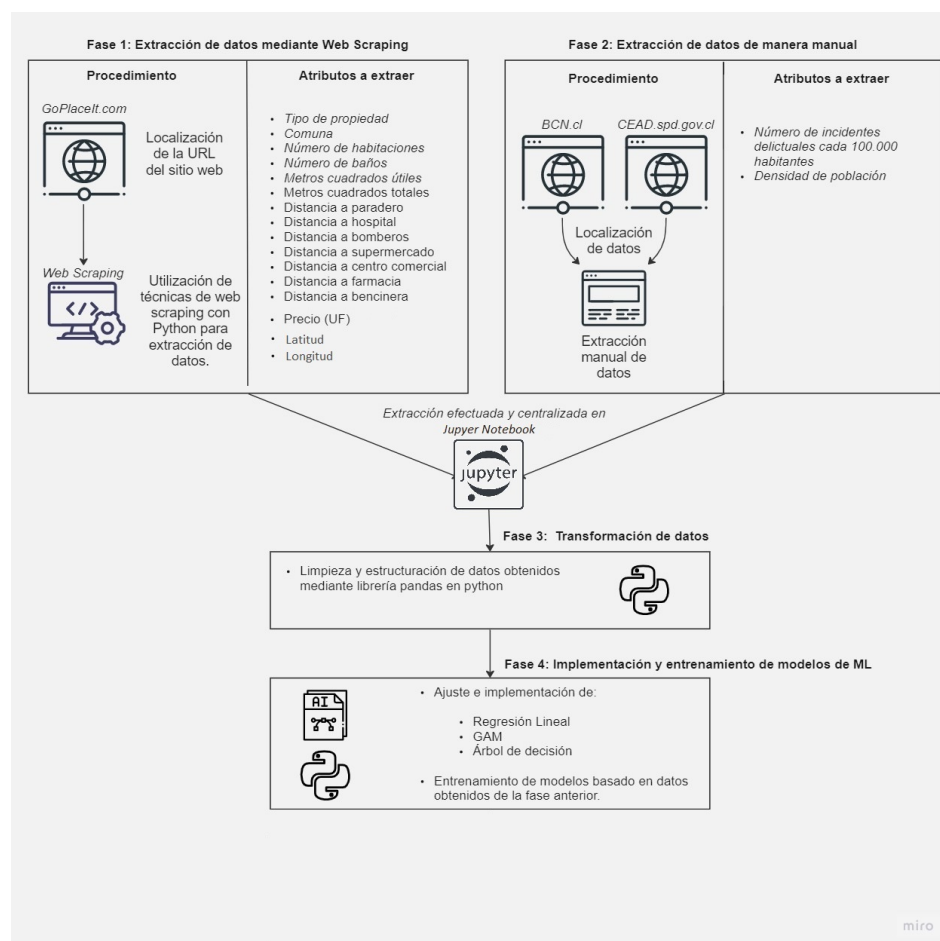


Figura 8: Metodología propuesta.
Fuente: Elaboración propia.

3.2. Fase 1: Extracción de datos de propiedades mediante web scraping

La primera fase de la metodología propuesta (véase figura 8) alude a la extracción de datos mediante *web scraping*. Para ello, se define manualmente una ruta que el *scraper* deberá seguir a cabalidad para conseguir la extracción de los datos. Se agrega además una subsección donde se describen ciertas técnicas utilizadas para que el *web scraping* se realice de manera segura, sin sobrecargar demasiado el servidor donde se hospeda el sitio web objetivo y dando a entender que el comportamiento es similar al de un humano. Sumado a lo anterior, esta primera fase de *web scraping* se realizará utilizando la librería *Selenium* de Python, puesto que es una librería que se utiliza para trabajar con sitios web cuya estructura es dinámica y basada en javascript (tienen botones que presionar, elementos que cargan después de cierto tiempo, etc.). Dicho lo anterior, el *scraper* primero comienza aleatoriamente en una de cuatro páginas posibles (entre youtube.com, emol.com, biobiochile.cl y google.cl) y desde allí, espera una cantidad aleatoria entre uno y dos segundos antes de redirigirse a la página principal de goplaceit.com/cl. Una vez en esta nueva página, se presiona el botón de la sección de “Regístrate o Ingresa” marcada en la imagen en un rectángulo de color rojo (figura 9).



Figura 9: Página principal goplaceit.com/cl.
Fuente: Elaboración propia.

Luego de haber realizado la tarea anterior, el *scraper* escribe (en intervalos pausados de entre dos y cuatro segundos) las credenciales asociadas a una cuenta personal en la sección de ingreso (figura 10) para acceder a los registros almacenados allí. Las credenciales que el *scraper* ingresa corresponden al correo registrado para el ingreso a la cuenta, sumado a la contraseña de la misma. Posterior a ello, el *scraper* presiona el botón de “iniciar sesión”.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

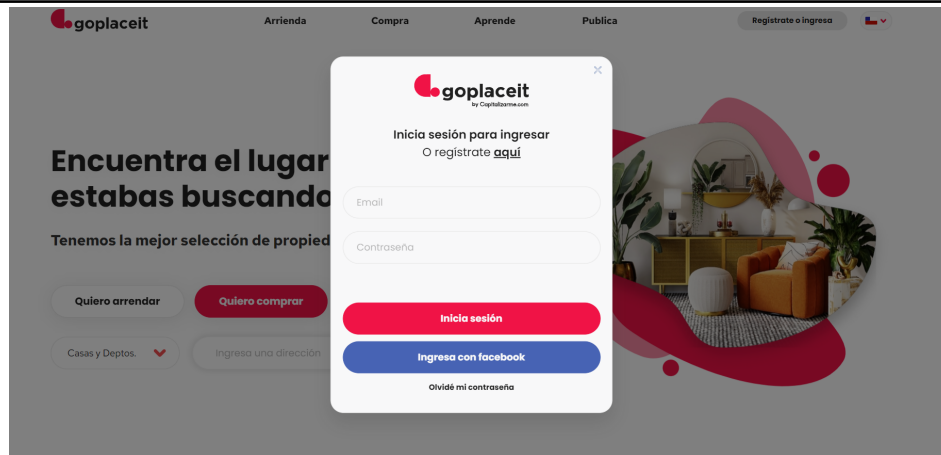


Figura 10: Inicio de sesión en Goplaceit.com/cl.
Fuente: Elaboración propia.

Una vez ingresadas las credenciales, el *scraper* cambia su URL a www.goplaceit.com/cl/mapa (figura 11).

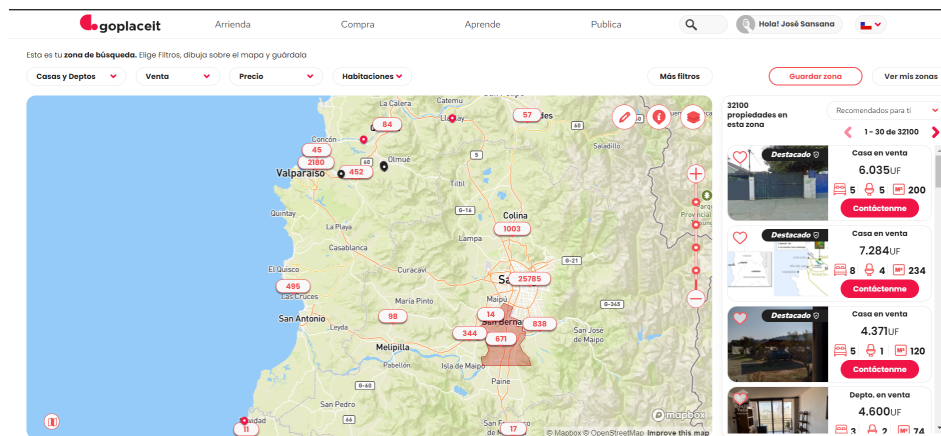


Figura 11: Sección Goplaceit.com/cl/mapa.
Fuente: Elaboración propia.

En esta nueva pestaña, se debe seleccionar “mis zonas”(figura 12) luego de entre uno y cuatro segundos aleatorios. En esta sección es crucial que se haya realizado el paso anterior relacionado con el ingreso de las credenciales, puesto que de no hacerlo, la sección “mis zonas” no mostrará nada y obligará a entrar con las credenciales posteriormente. Cabe destacar que las zonas que aparecen en “mis zonas” corresponden a las comunas definidas en las conurbaciones de Gran Santiago, Gran Valparaíso y Gran Concepción. Esta delimitación de las zonas se realizó con anterioridad, ingresando a la cuenta respectiva y seleccionando el área a extraer (las comunas en particular en este caso).

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

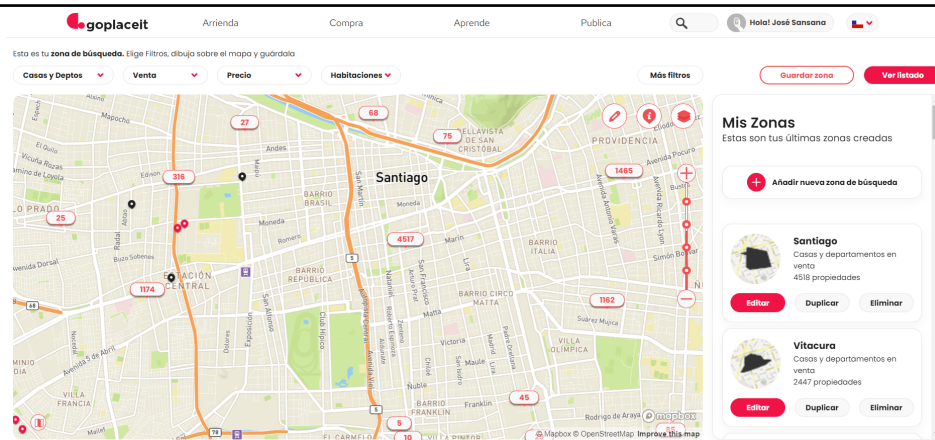


Figura 12: “Mis Zonas” en Goplacelit.com/cl/mapa.

Fuente: Elaboración propia.

En el menú de la derecha, el *scraper* presiona una de las comunas y el sitio web considerará automáticamente las propiedades que se encuentran dentro del área demarcada. Realizado esto, se presiona “ver listado” para ver un listado de todas las propiedades que se encuentran dentro de los límites de la zona. Se ve un ejemplo (figura 13) de la zona de Santiago Centro, con las correspondientes propiedades inmersas en los límites.

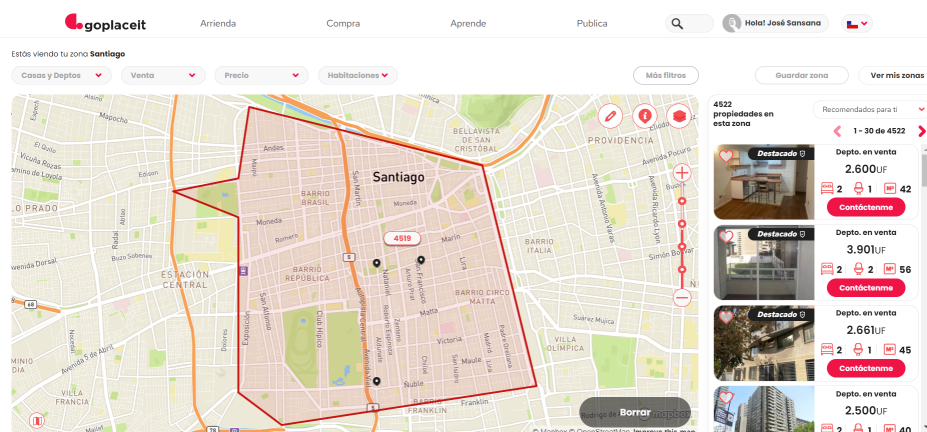


Figura 13: Propiedades dentro de límites en Goplacelit.com/cl/mapa.

Fuente: Elaboración propia.

Habiendo ya descrito el procedimiento en común para cada una de las comunas consideradas, lo que resta es extraer la información de cada una de las propiedades que se encuentran en la zona demarcada. Para ello es necesario presionar cada una de las imágenes de las distintas propiedades, bajo lo cual se abrirá una pestaña adicional que contendrá la información intrínseca y externa del inmueble (véase figuras 14 y 15). Cabe destacar que la cantidad de estacionamientos se extrae tal cual aparece en la página mientras sea un valor numérico. En

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

caso de que la página no ofrezca información mediante la frase “Sin información”, la cantidad de estacionamientos se reemplaza con un “0”.

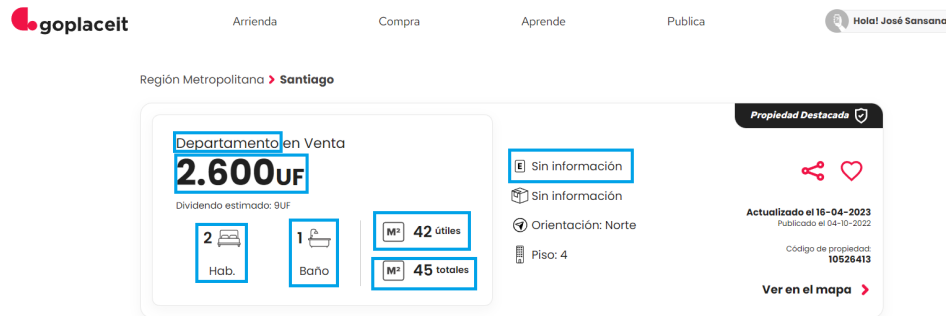


Figura 14: Ejemplo de Información intrínseca del inmueble.
Fuente: Elaboración propia.

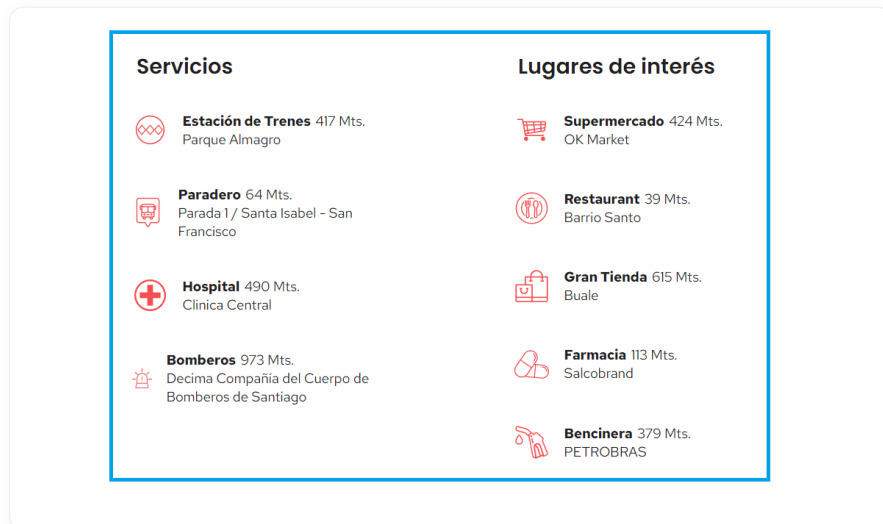


Figura 15: Ejemplo de Información del entorno de un inmueble.
Fuente: Elaboración propia.

También se puede ver (figura 14) un texto que reza “Ver en el mapa”. Este texto es otro componente importante dentro de la estructura del sitio web, puesto que, al presionarlo, se cambia de pestaña a una nueva, la cuál contiene la longitud y latitud de la propiedad dentro del URL (figura 16). Estas longitudes y latitudes son almacenadas debido a su utilidad potencial para la investigación.

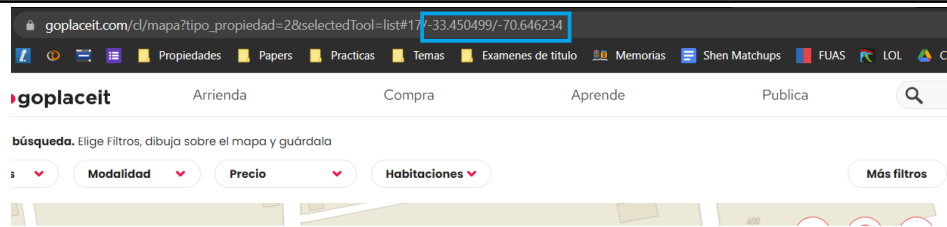


Figura 16: Longitud y latitud de la propiedad.

Fuente: Elaboración propia.

Al momento de haber extraído los datos intrínsecos y externos (con longitud y latitud incluidas) estas se almacenan en una lista y posteriormente en un *dataframe*. Este proceso de extracción se repite con cada una de las imágenes presentes en la sección previamente comentada.

Aún así queda un paso extra, pues al momento en que el *webscraper* llega al final de la página (ha recorrido todas las imágenes mostradas), debe presionar el botón que simula una flecha de color rojo, este puede apreciarse (figura 13) en la parte media derecha. Presionando este botón, el *webscraper* esperará un máximo de 30 segundos y podrá acceder a las propiedades siguientes. Este proceso se repite hasta que ya no se pueda presionar mas el botón rojo. De esta forma se recorrerán todas las propiedades que estén disponibles para una zona geográfica previamente definida.

Al realizar este procedimiento la cantidad de veces necesaria hasta extraer todos los datos de las propiedades, se convierte el *dataframe* de la comuna en un documento csv, almacenando todos los datos relacionados a esa comuna en particular y pudiendo así trabajarse en el entorno *Jupyter Notebook*. Los resultados de cada conurbación pueden verse en las tablas de anexos (figuras reftab:WSGS para GS, 17 para GV y 18 para GC).

3.2.1. Consideraciones al extraer datos del Gran Concepción

Se debe destacar el hecho de que en el Gran Concepción no se pudo obtener información de cada una de las comunas presentes en este, limitando la extracción solo a 6 de estas, siendo: Chiguayante, Concepción, Coronel, Hualpén, San Pedro De La Paz y Talcahuano, ya que Tomé, Lota, Penco, Hualqui no disponían de los datos necesarios (menos de tres datos por comuna y de poca calidad).

3.2.2. Consideraciones al realizar *webscraping* de manera prudente

Al haber completado la primera fase de la metodología (figura 8), es de suma importancia resaltar la prudencia con la que se realizó *web scraping*. La principal motivación para esto

fue evitar sobrecargar los servidores de goplaceit.com a su vez de evitar ser bloqueados por el servidor por una excesiva extracción de datos en un corto período de tiempo.

La primera medida fue comenzar el *web scraping* a partir de un sitio web aleatorio, los cuales son: youtube.com, emol.com, biobiochile.cl y google.cl. Esto como medida cautelar ya que un servidor avanzado puede detectar si se ingresa directamente a la página en la que se encuentra o si se viene redirigido de otro sitio web. Este último emula un comportamiento “mas humano”, y por tanto, no activa alertas de precaución.

Una segunda medida fue, habiendo ya ingresado a goplaceit.com/cl desde el sitio web pre-definido, esperar un tiempo aleatorio entre dos y cuatros segundos en las acciones que se estaban realizando en la página, las cuales son: presionar botón de login, ingresar credenciales, redirigir a página de goplaceit.com/cl/mapa y elegir la zona de la cual se extrajo la información (con todos los botones involucrados).

Una tercera medida consistió en el uso de distintos *user agents*. Cabe destacar, que cada vez que se presiona el botón de flecha roja para que aparezcan nuevas propiedades, se cambia aleatoriamente el componente *user-agent* (figura 51) (que actúa como cliente para los servidores), de manera que no sea específicamente solo un cliente el que solicita la información constantemente. Esto se realiza como medida cautelar para prevenir bloqueos de IP desde el servidor. Otro punto es que, al cambiar de página, se esperaba un tiempo aleatorio entre dos y cuatro segundos.

Por último, la cuarta medida alude a definir tiempos aleatorios entre cero y dos segundos (y en algunos casos entre uno y dos segundos) al momento de realizar las solicitudes al servidor basadas en la estructura HTML de la página. Esto se realizó obteniendo la información interna y externa del inmueble.

3.3. Fase 2: Extracción de datos de entorno de propiedades manualmente

Esta corresponde a la segunda fase mencionada en la metodología (figura 8) y alude a la extracción de datos manuales, mas específicamente a la tasa de delincuencia cada cien mil habitantes y a la densidad de habitantes, ambas por comuna.

La tasa de delincuencia cada cien mil habitantes es obtenida mediante el portal de ayuda a víctimas del ministerio del interior de Chile [de apoyo a víctimas, 2022]. En este, se selecciona la tasa cada cien mil habitantes de la comuna que deseo, considerando todos los delitos registrados en la página y en el último trimestre 2022.

La densidad de habitantes se obtiene a la proyección de población 2023 realizada por la Biblioteca del congreso nacional [del congreso nacional de Chile, 2022a], donde a cada comuna le asignan una proyección de población 2023 basada en el censo 2017.

Toda la información obtenida anteriormente es conglomerada en un archivo de Microsoft

Excel, teniendo un código para cada comuna y sus respectivos índices de delincuencia y densidad poblacional en columnas aledañas. Se muestra un ejemplo (figura 52 de esta extracción).

3.4. Fase 3: Limpieza y estructuración de datos obtenidos

A continuación se describe un procedimiento de limpieza y estructuración aplicado a cada una de las conurbaciones en general, para luego describir procedimientos específicos para cada una.

Teniendo los archivos CSV en un entorno local con los datos de cada una de las comunas involucradas, el siguiente paso alude a la fase 3 de transformación de datos de la metodología propuesta (figura 8), la cuál comienza con la carga de datos de los archivos anteriormente mencionados en el entorno de *Jupyter Notebook*. Para esta sección se usa primordialmente la librería *pandas* de *Python*, ya que permite el manejo de *dataframes* (referidos a partir de ahora como “DF”) para el manejo de datos, sumado a las librerías *numpy* para cálculos, *matplotlib* y *seaborn* para visualización y *base64* junto a *FileLink* y *HTML* para la transformación de *dataframe* a documento CSV.

El primer objetivo, es poder estandarizar los datos de una conurbación dentro de un solo archivo, por ende, se comienza con la carga de los archivos CSV. Se carga cada una de las comunas involucradas en la conurbación como un *DF* y luego se unen mediante la concatenación de filas de un *DF* tras de otro, juntando así todas las comunas en un solo formato.

El segundo objetivo, consiste en eliminar columnas que traigan problemas previos. Estos problemas fueron evidenciados al momento de realizar la extracción de datos. La columna “Paradero” fue considerada imprecisa dado que que en algunas comunas (usualmente de la región metropolitana) otorgaba información precisa de la cercanía a paraderos, sin embargo en comunas no tan céntricas (como Concón, San Pedro de la Paz, Chiguayante, entre otras) solía tener problemas detectando paraderos mediante el mapa de la aplicación, otorgando valores mas altos de los que realmente eran. Mismo caso ocurría con la distancia a la compañía de bomberos mas cercana, ya que aparecen en el mapa las compañías mas grandes, mas no todas las compañías. En cuanto a los restaurantes pasaba el caso opuesto, ya que no existía una estandarización de lo que significaba restaurante (mostraba picadas, locales de comida rápida, Restaurantes finos, entre otros). La siguiente imagen (figura 53) denota un ejemplo de una distancia a paradero y estación de bomberos distorsionada por el mapa en Concón. (Concón tiene su propia compañía de bomberos, la cuál está a menos de 3000 metros de la propiedad mostrada en la imagen. Por otro lado, existen diversos paraderos de transporte público, pero no son considerados).

Habiendo extraído dichos atributos, el paso siguiente corresponde a unir el *DF* conformado con todas las comunas con las estadísticas de densidad y delictuales como las mostradas (figura 52). Para ello, se carga independientemente el archivo excel mencionado y se une

con el *DF* principal mediante las comunas, es decir, ya que tanto el *DF* principal como el *DF* de densidad y delincuencia poseen una columna que refiere a las comunas. Estas comunas son la llave de conexión entre ambos *DF*. Una vez realizado dicho proceso, el *DF* principal queda de la siguiente forma (figura 54).

Cabe destacar que aquel *DF* solo muestra una cantidad parcial de atributos, ya que sería demasiado extenso mostrarlo en su completitud. Aún así, es posible notar que las últimas dos columnas corresponden a la densidad de la comuna y la tasa de delincuencia en el último trimestre de la misma. También es posible notar que estas tasas varían comparando las primeras con las últimas, debido al cambio de comuna.

Teniendo el *DF* completo respecto a los datos, la siguiente misión corresponde al análisis y limpieza de los datos, también conocido como pre-procesamiento, para que su calidad beneficie a los modelos de *ML* que se utilizarán mas adelante.

3.4.1. Limpieza y estructuración de datos en Gran Santiago

La limpieza comienza con un estudio de las distribuciones de los atributos, así como la detección de valores atípicos (comúnmente conocidos como *outliers*), los cuales se definen como datos que no van unidos a la estructura general de la distribución.

Las distribuciones iniciales del GS tanto para los atributos internos como externos se pueden visualizar a continuación (figuras 17 y 18) sumado a la densidad y tasa de delincuencia cada cien mil habitantes (figura 19).

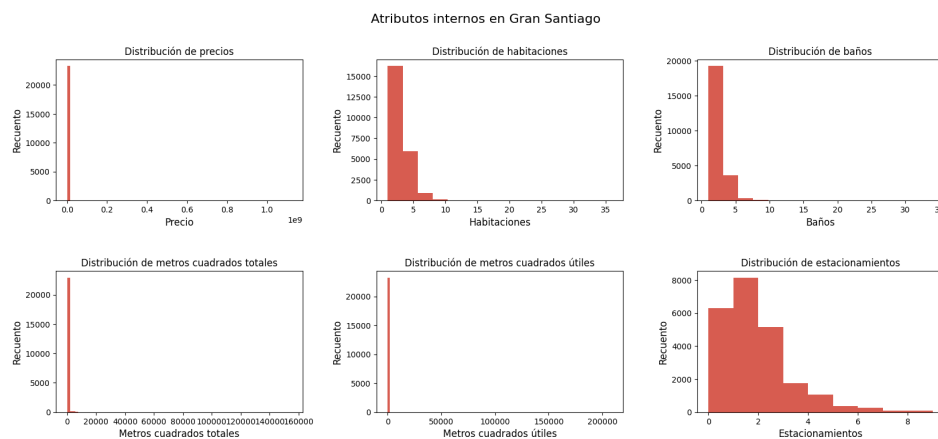


Figura 17: Distribución en GS de atributos internos.

Fuente: Elaboración propia.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

Atributos externos en Gran Santiago

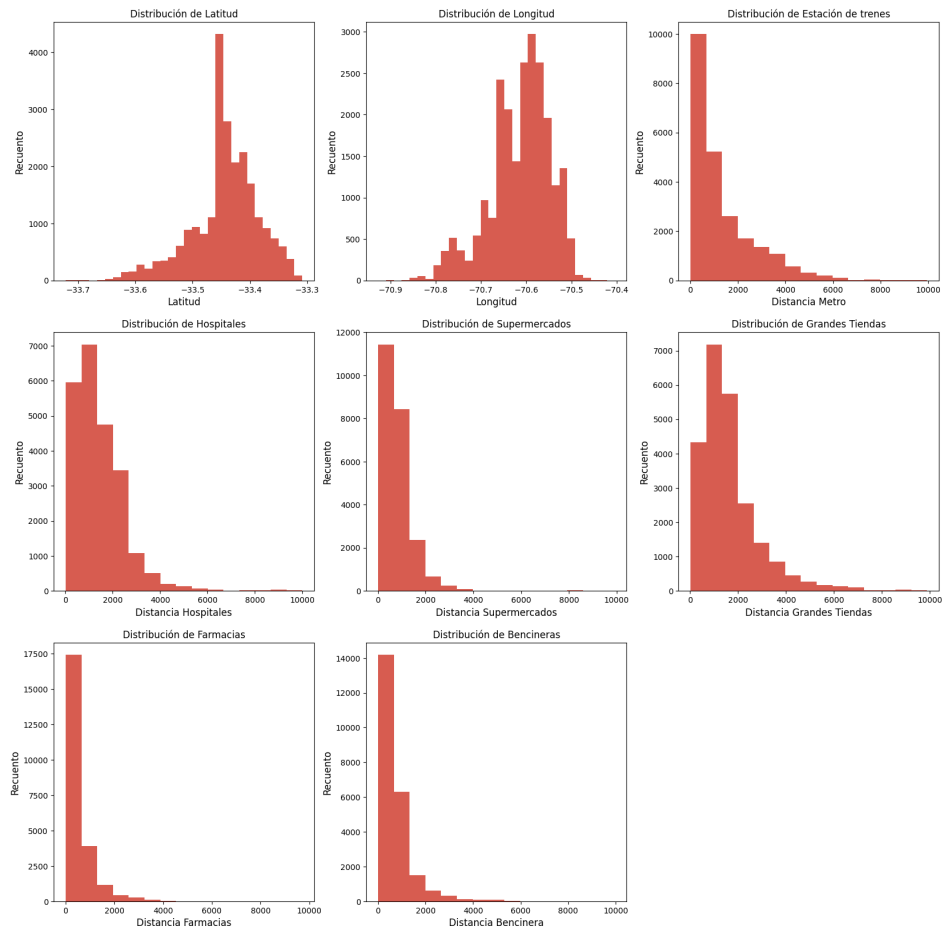


Figura 18: Distribución en GS de atributos externos.
Fuente: Elaboración propia.

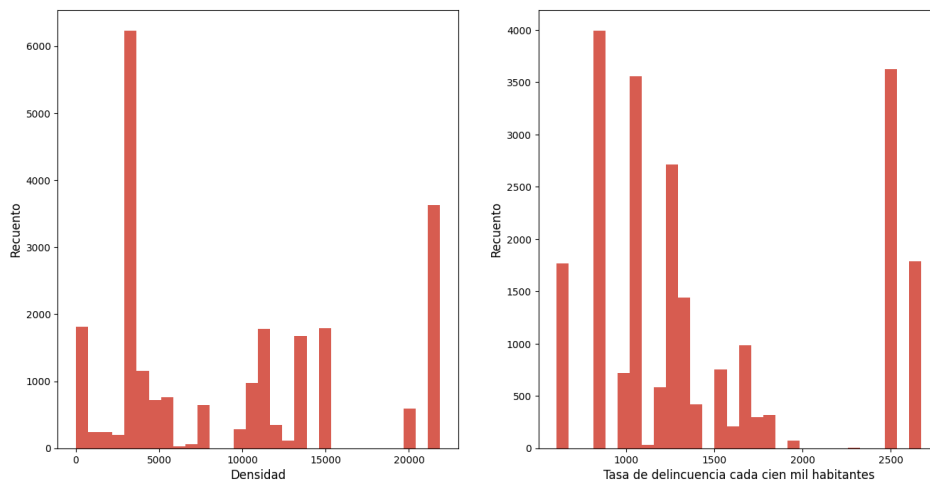


Figura 19: Distribución en GS de delincuencia y densidad.

Fuente: Elaboración propia.

Analizando la primera figura (figura 17), se destacan dos aspectos principales. El primero alude a que la escala de los atributos (a excepción de la cantidad de estacionamientos) parece no ser acorde a la representación de los datos, ya que no se puede ver con claridad una distribución de probabilidad, sino una simple barra. Este tipo de comportamiento es obra de valores atípicos, ya que la mayoría de los valores se concentra al principio de la escala, pero una cantidad ínfima es representada por valores mucho mas grandes, suponiendo un incremento en la escala de valores. Respecto al segundo aspecto, este refiere al sesgo positivo de las distribuciones observadas. Este sesgo destaca por una concentración de los datos mayoritariamente hacia la izquierda de la distribución, y cuyos valores extremos se cargan hacia la derecha. Este tipo de sesgo puede llevar a problemas en análisis de regresión, por lo cuál debe ser manejado.

Sumado a la visualizaciones de las distribuciones, se imprimen los datos extremos de cada uno de los atributos internos para poder observar directamente cuales son los valores atípicos. Un ejemplo de esto se observa a continuación (figura 55).

Gracias a las gráficas de distribución, la visualización de los valores extremos y el análisis de la estructura de los datos, se descubre lo siguiente:

- Todo precio que está sobre 1.000.000 corresponde a un error de la persona que publicó el aviso, ya que el precio debía subirse en UF y se encuentra en CLP, por lo que se realiza la transformación requerida a UF.
- Si el precio corresponde a 1 o si el precio oscila entre 100.000 y 1.000.000 significa que ocurrió un error al ingresarlo o simplemente se trata de propiedades exhuberantes

(con mas de 20 habitaciones y miles de m^2) que cuentan como valor atípico, por lo cuál se descartan dichas filas del *DF*.

- Las propiedades con un número de habitaciones y/o baños mayores a 18 son descartadas por corresponder a valores atípicos.
- Si los m^2 útiles superan los 40.000 o no superan los 11, entonces se descartan por valores atípicos. Mismo caso para los m^2 totales superando los 25.000.
- En algunos casos simplemente los usuarios que publicaban la propiedad ingresaban al revés las cantidades de m^2 útiles con m^2 totales, así que en caso de que estos últimos fuesen menores a los útiles, se reemplazaba el uno por el otro.
- Eliminar precios menores a 6.3 (en escala logarítmica).

Al aplicar estas consideraciones, las distribuciones de los datos internos se ven de manera mucho más uniforme (figura 20).

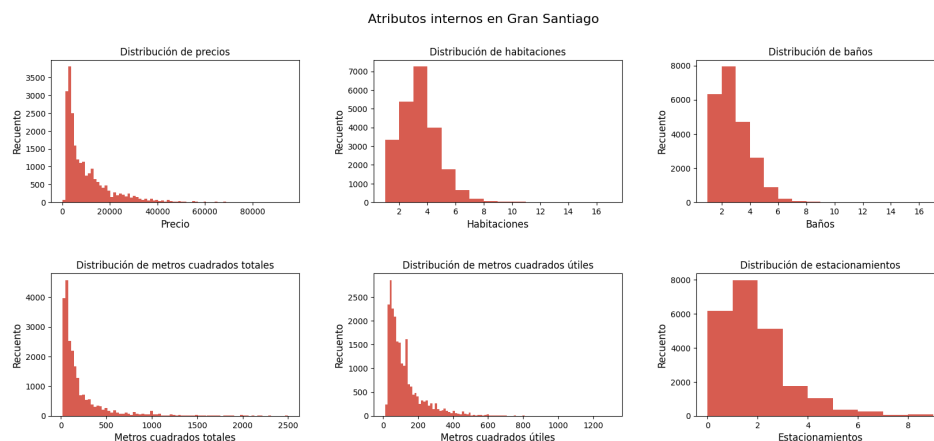


Figura 20: Remoción de valores atípicos en GS.
Fuente: Elaboración propia.

Aún así, aún queda por abordar el sesgo positivo que presentan los atributos. Es por ello que se opta por la utilización de la transformación de las variables a escala logarítmica (es decir, aplicar un logaritmo en base 10 a la escala de los atributos y sumarle 1). Al aplicar esto, se percibe una mejora en la distribución de precios, m^2 útiles y m^2 totales, mas no se verifica una mejora respecto a baños, habitaciones y estacionamientos. Es por ello que se opta dejar estos tres últimos atributos inalterados. Las distribuciones lucen de la siguiente forma tras las transformaciones realizadas (figura 21).

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

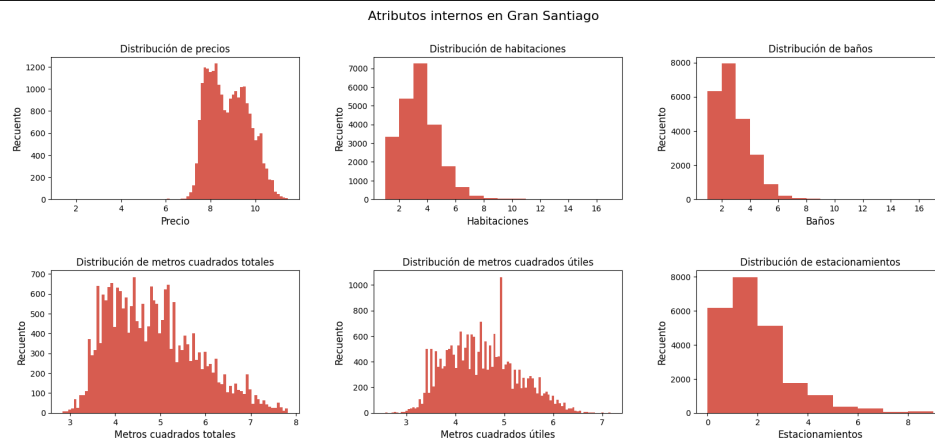


Figura 21: Transformación logarítmica.
Fuente: Elaboración propia.

Al ver esta distribución, se hace evidente una anomalía de valores inferiores a 6.3 en los precios. Haciendo dicha remoción las distribuciones quedan de la siguiente forma:

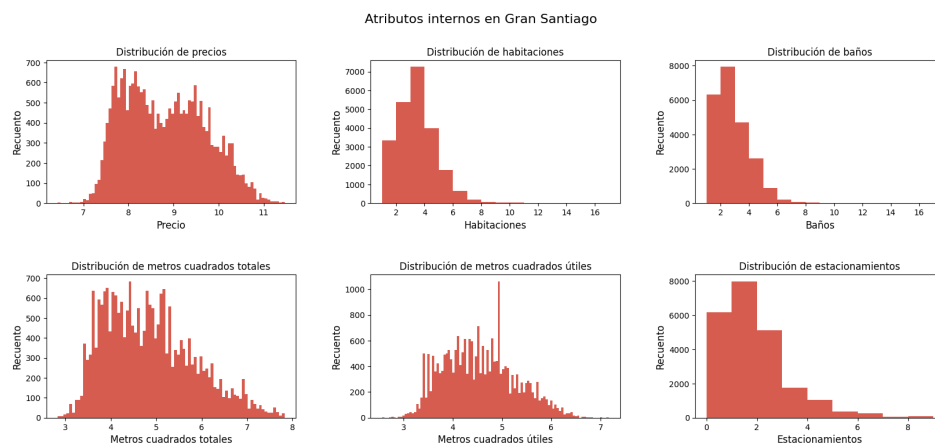


Figura 22: Remoción de nuevos valores atípicos en GS.
Fuente: Elaboración propia.

En cuanto a los atributos externos, permanecen iguales (figura 18). Latitud y longitud no pueden ser transformados a escala logarítmica dada su escala de valores negativos. En cuanto al resto de los atributos externos, se opta por mantenerlos como están, ya que al utilizar escala logarítmica ocurre un sesgo negativo.

Respecto a la densidad y la tasa de delincuencia, al no tener muchos datos debido a que solo existe dicho dato por comuna y no por sector, se opta por dejar tal cual está.

Luego de realizada la limpieza de los datos del GS, deben removerse los datos duplicados, ya que hay veces en que el usuario de goplacit que publica la propiedad la publica dos veces

por error o simplemente son propiedades que cumplen la misma condición (por ejemplo, se venden tres departamentos iguales en un mismo edificio). Los datos duplicados pueden desembocar en problemas en los modelos de *ML*, ya que pueden inducir sesgos. Esto se realiza mediante la función *drop_duplicates* incluida en la librería *pandas*.

Teniendo los duplicados removidos, solo queda un paso para finalizar esta sección, la cuál es transformar las variables categóricas a numéricas. Para ello se opta por transformar las comunas a formato *One Hot Encoding* (cada comuna es una nueva columna de atributos en el *DF*, y se le asigna 0 en la celda que no aparezca y un 1 en caso de aparecer) y las viviendas a formato *Label Encoding* (se categoriza departamentos como 0 y casas como 1 en la misma columna).

A pesar de haber considerado tanto la comuna en la que se encuentra la propiedad como el tipo de vivienda, solo se utilizará este último, ya que las comunas harían que el análisis posterior entre modelos y comunas difiera, principalmente porque cada conurbación tiene un número distinto de comunas, complejizando el análisis (se muestra claramente en la matriz de correlación de la figura 57) ya que todas las comunas tienen un impacto muy bajo en determinar el precio, mas no así la longitud y la latitud) y en segundo lugar un desbalance de filas por comuna. Es por esto que se decide utilizar solo la latitud y longitud para la representación geográfica de la vivienda. En cuanto al balance de estas métricas, se pueden ver a continuación (figuras 23 y 24).

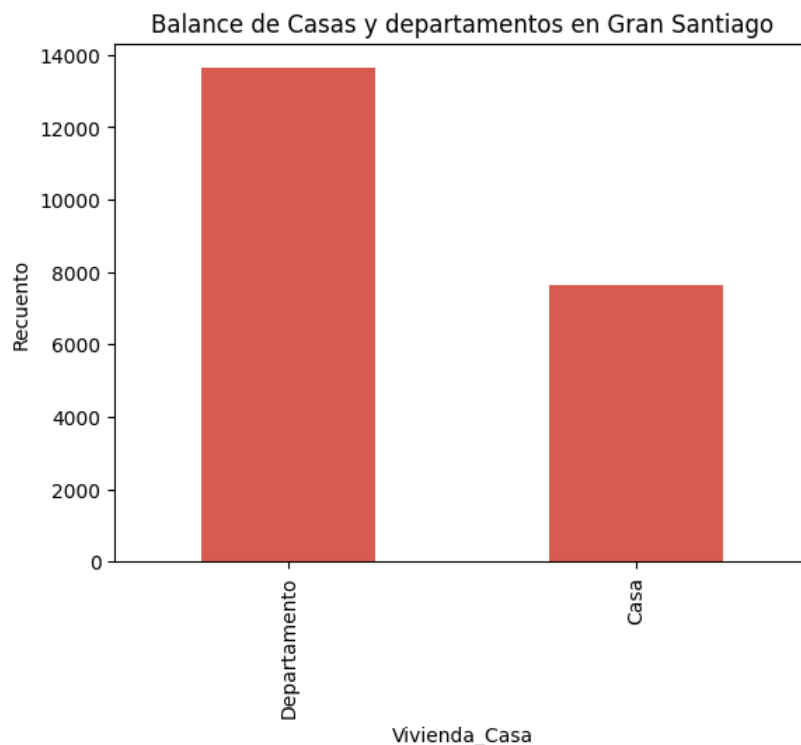


Figura 23: Balance de tipos de vivienda en GS.
Fuente: Elaboración propia.

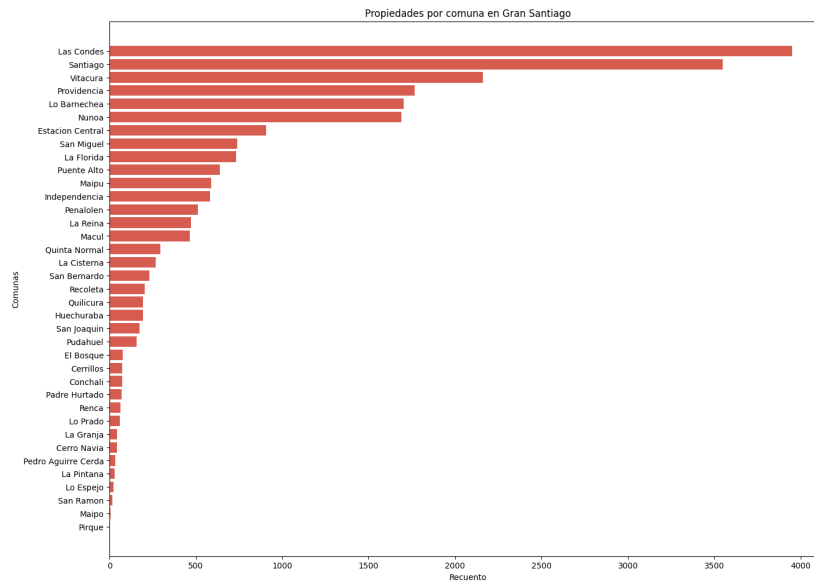


Figura 24: Balance de comunas en GS.
Fuente: Elaboración propia.

Finalmente, se transforma el *DF* a un archivo CSV con toda la información de GS lista para ser cargada.

3.4.2. Limpieza y estructuración de datos en Gran Valparaíso

Los pasos a seguir para la limpieza de datos en GV, son esencialmente los mismos que los utilizados en GS, por lo que esta descripción se enfocará en presentar las distribuciones iniciales y finales de los atributos internos y externos, además de las consideraciones utilizadas para el tratamiento de los datos. Cabe destacar que el balance de comunas y tipos de vivienda es similar a GS. En caso de tener alguna duda sobre como se visualizan determinados ejemplos durante el proceso, por favor revisar la sección de Limpieza y estructura de datos en Gran Santiago.

En primera instancia se observan distribuciones (figuras 25 y 26) con un alto nivel de valores atípicos y con un sesgo positivo. A su vez, se imprimen los valores extremos para tener una mejor noción de como manejarlos. Cabe destacar que la latitud en esta conurbación se comporta de distinta manera a las otras dos, puesto que hay un número considerable de propiedades de Concón (comuna ubicada a una latitud mayor que el Valparaíso y Viña del Mar), por lo que es posible ver una distribución casi bimodal. Sumado a lo anterior, se observa la distribución de tasas de delincuencia y densidad (figura 27).

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

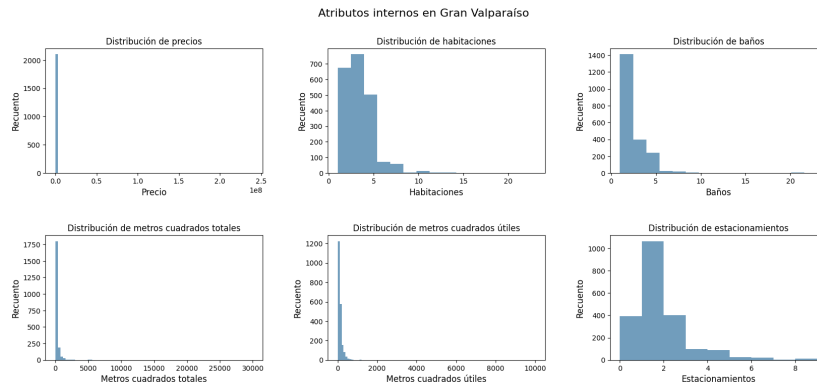


Figura 25: Distribución de atributos internos en GV.
Fuente: Elaboración propia.

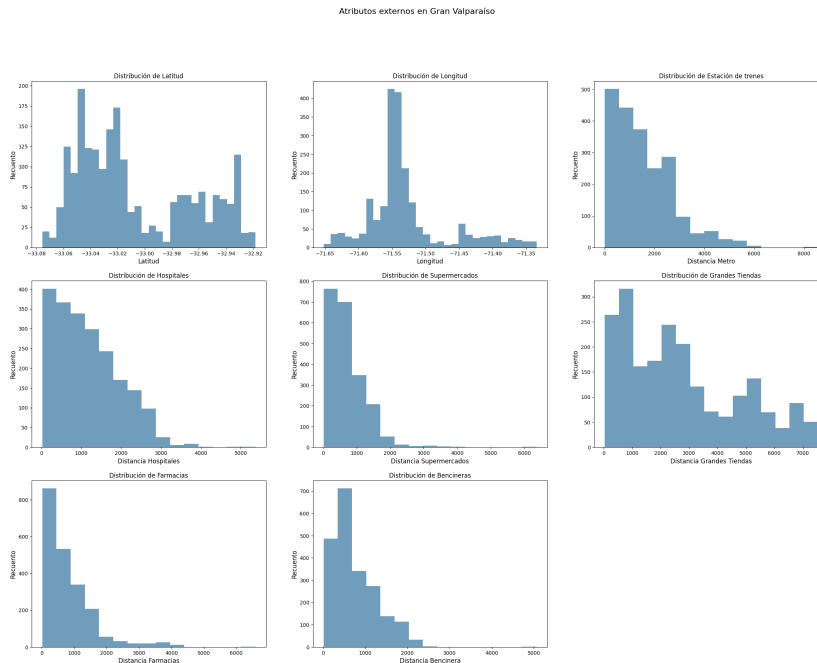


Figura 26: Distribución de atributos externos en GV.
Fuente: Elaboración propia.

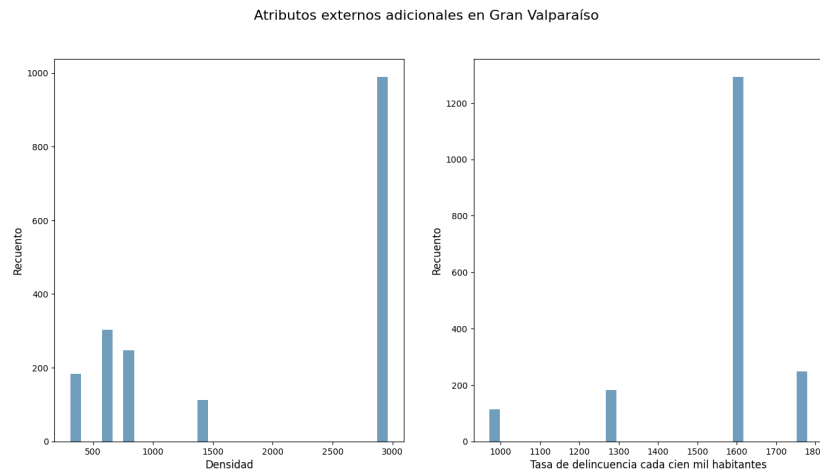


Figura 27: tasas de delincuencia y densidad en GV.
Fuente: Elaboración propia.

Al analizar las distribuciones y los valores extremos, se decanta por tomar las siguientes medidas:

- Solo existe un precio que se transforma desde CLP a UF, siendo 240.000.000 CLP.
- Si el precio es menor a 80 o mayor a 100.000 se descarta por ser valores atípicos por errores de publicación del precio.
- En caso de que la propiedad contenga mas de 15 habitaciones o baños, se descarta de inmediato por considerarse valor atípico.
- Si la propiedad tiene menos de 11 m^2 útiles o totales es descartada por considerarse valor atípico. Mismo caso si los m^2 útiles superan los 1.000 metros y los m^2 totales superan los 1.500 metros.
- Se suma 8.000 metros a la distancia a estación de trenes de las propiedades de Concón por considerar erróneamente una estación antigua de ferrocarriles como estación de metro. La distancia desde esa antigua estación a la estación mas cercana es de 8.000 metros.

Al aplicar estas consideraciones, es posible ver distribuciones mucho mas uniformes, sin embargo poseen el mismo problema de sesgo positivo que en el caso de GS, por lo que se escoge la transformación logarítmica a las variables de Precio, m^2 útiles y m^2 totales. El resto de variables al aplicar la transformación desarrollan un sesgo negativo o una leve mejora no considerable (figura 56).

Con el proceso anterior las distribuciones finales se ven mucho mas uniformes (figuras 28 y 29).

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

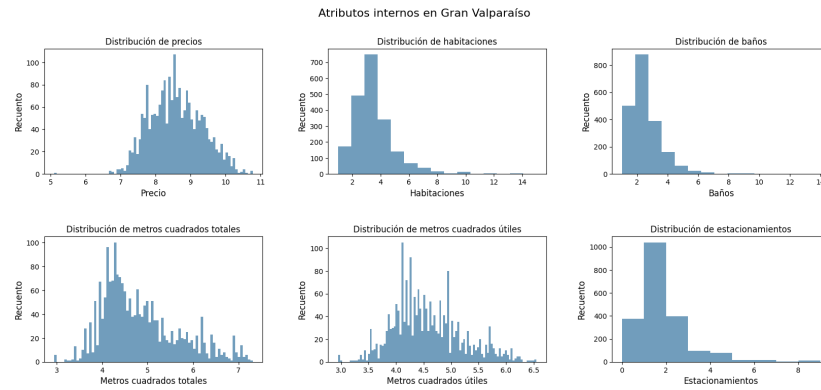


Figura 28: Distribución de atributos internos al procesarse en GV.
Fuente: Elaboración propia.

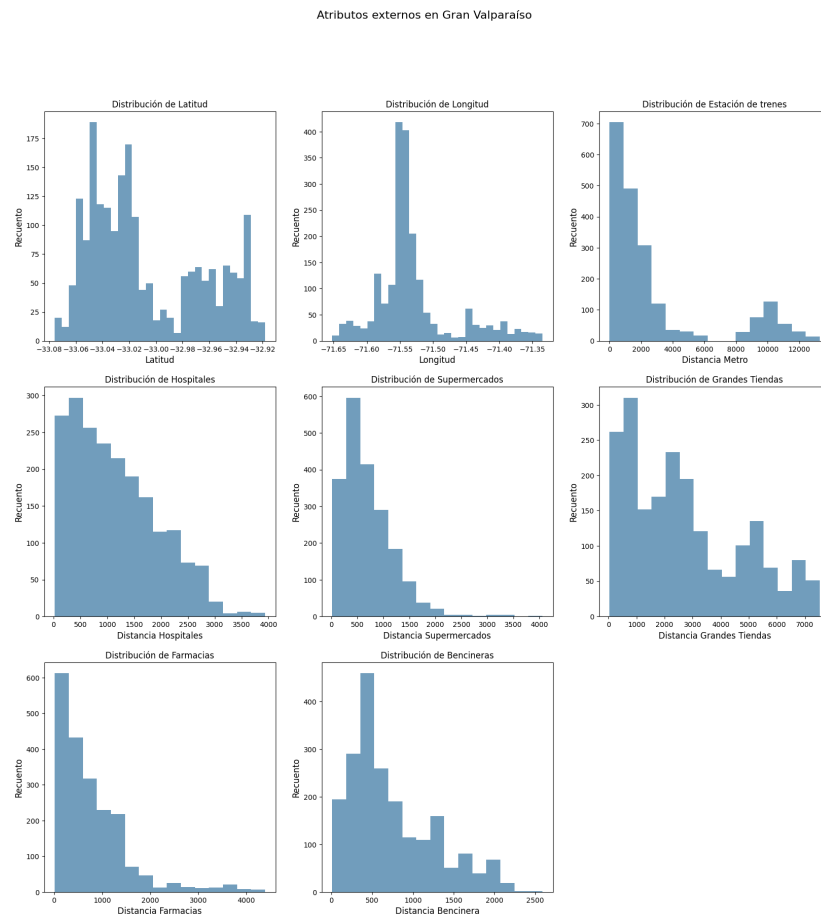


Figura 29: Distribución de atributos externos al procesarse en GV.
Fuente: Elaboración propia.

Se aplica el mismo procedimiento anterior respecto a remover las filas duplicadas y transformar atributos categoricos con *One Hot Encoding* y *Label Encoding*, dejando finalmente solo

la variable “Vivienda_Casa” que establece 1 en caso de ser casa y 0 en caso de departamento. Esto debido a propósitos de uniformidad entre conurbaciones. Se observa un desbalance de ambos casos (figuras 30 y 31).

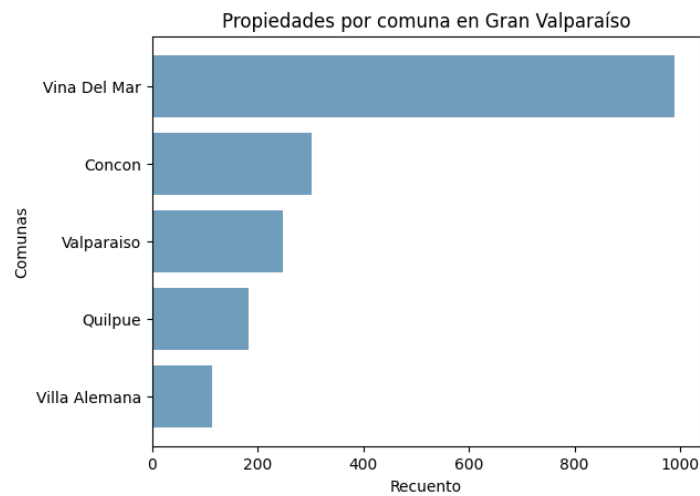


Figura 30: Distribución de comunas en GV.
Fuente: Elaboración propia.

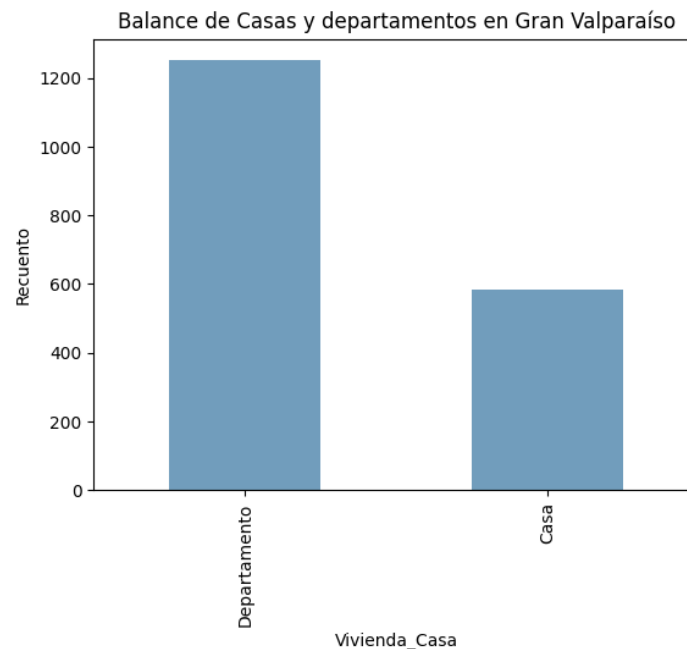


Figura 31: Desbalance de propiedades en GV.
Fuente: Elaboración propia.

También se opta por remover las comunas y representar la información geográfica de la propiedad solamente con la latitud y longitud por los motivos descritos en la limpieza de GS

aunque se deja la variable binaria que indica el tipo de vivienda.

Con el proceso de limpieza realizado, se transforma el *DF* en un archivo CSV que será utilizado para la fase siguiente.

3.4.3. Limpieza y estructuración de datos en Gran Concepción

Continuación del procedimiento de limpieza de datos, esta vez con la conurbación de GC (a excepción de las comunas denotadas en la fase de extracción de datos mediante *web-scraping*).

Observando las distribuciones iniciales (figuras 32 y 33) se nota una evidente similitud respecto a los casos de las conurbaciones anteriores. También se realiza el proceso de observar los valores extremos. Además se observa también la distribución de tasas de delincuencia y densidad (figura 34).

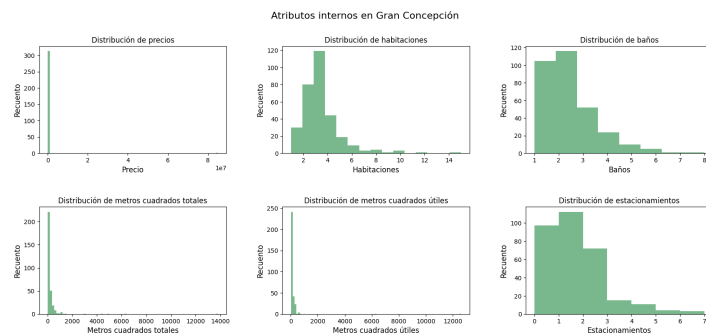


Figura 32: Distribución atributos internos en GC.

Fuente: Elaboración propia.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

Atributos externos en Gran Concepción

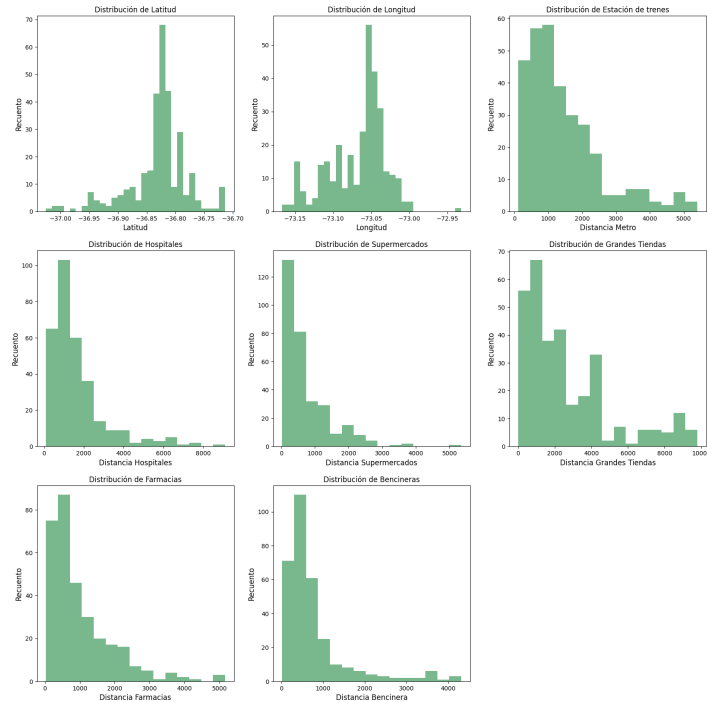


Figura 33: Distribución atributos externos en GC.

Fuente: Elaboración propia.

Dada la cantidad de datos para esta conurbación, las consideraciones a seguir son menores a las anteriores. Con ello en mente es posible determinar lo siguiente:

Atributos externos adicionales en Gran Concepción

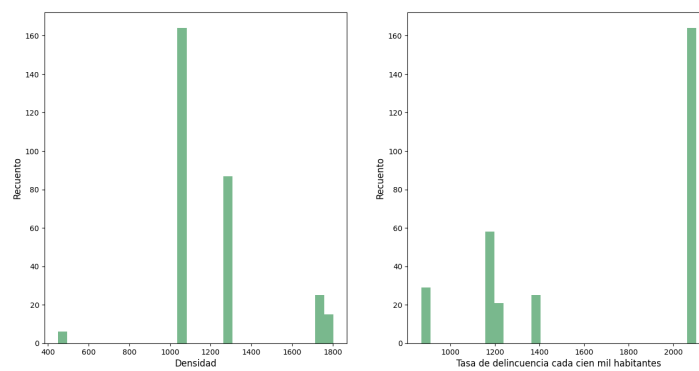


Figura 34: Distribución atributos externos en GC.

Fuente: Elaboración propia.

Dada la cantidad de datos para esta conurbación, las consideraciones a seguir son menores a las anteriores. Con ello en mente es posible determinar lo siguiente:

- Si el precio es mayor a 25.000 UF se descarta del *DF*.
- Si los m^2 útiles de una fila superan los 1.500 y los m^2 totales superan los 4.000, entonces se descarta dicha fila del *DF*

Siguiendo estas consideraciones, a su vez que se aplica el procedimiento de remover duplicados y transformación de variables categoricas mediante *One Hot Encoding* y *Label Encoding*, se obtiene la siguiente distribución final de atributos internos(figura 35) y balances (figuras 37 y 36).

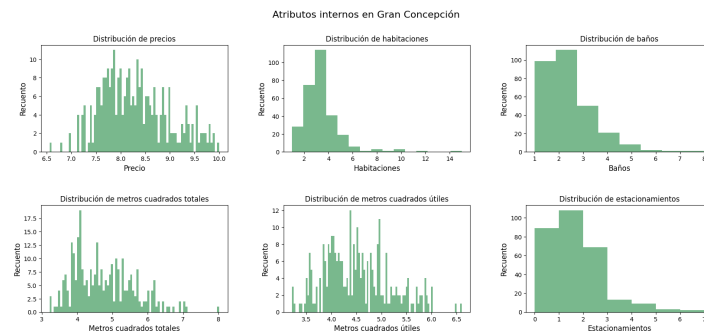


Figura 35: Distribución final atributos internos GC.
Fuente: Elaboración propia.

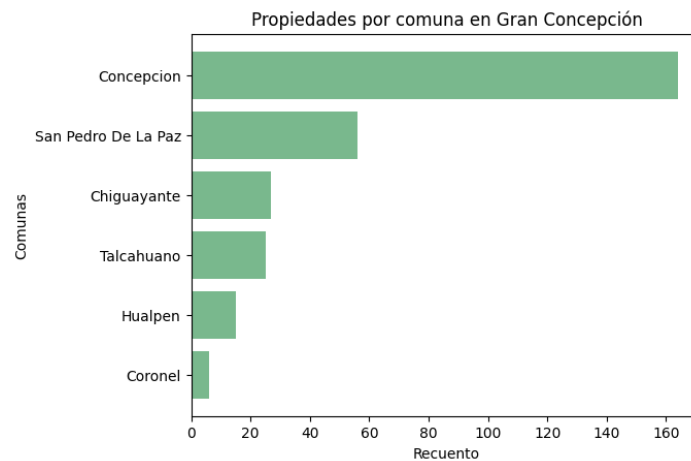


Figura 36: Balance de comunas en GC.
Fuente: Elaboración propia.

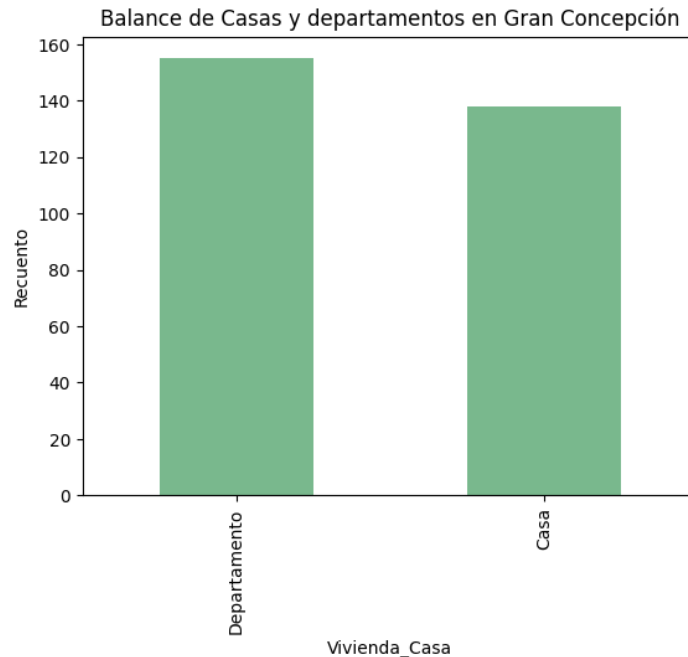


Figura 37: Balance de propiedades en GC.
Fuente: Elaboración propia.

También se opta por remover las comunas y representar la información geográfica de la propiedad solamente con la latitud y longitud por los motivos descritos en la limpieza de GS. La variable de tipo de vivienda permanece.

Por último, se descarga un archivo CSV en base al *DF* de la conurbación para ser cargado en la fase siguiente.

3.4.4. Resumen al finalizar la limpieza de datos

Una vez terminada la limpieza de datos en cada conurbación, se presenta a continuación una tabla (véase tabla 3) a modo de resumen que condensa la cantidad de datos que se eliminaron del conjunto de datos por no aportar al mismo y que cantidad de datos queda finalmente en cada conurbación respecto a todos los datos disponibles en el sitio web de Goplacit.com.

Tabla 3: Resumen de eliminación de datos por conurbación.

Fuente: Elaboración propia.

Resumen eliminación de datos por conurbación			
Conurbación	Datos perdidos	Datos finales	Porcentaje final respecto a Goplacit.com
Gran Santiago	2.006	21.254	76 %
Gran Valparaíso	257	1.836	71 %
Gran Concepción	21	293	71 %

3.5. Fase 4: Implementación y entrenamiento de modelos

Como se estableció previamente, los modelos de ML a utilizar corresponden a regresión lineal, GAM (en específico su variante lineal) y árbol de decisión (CART). Su implementación es a partir de librerías disponibles en Python, utilizando SKlearn y PyGam para los modelos y principalmente numpy, pandas, matplotlib y seaborn para el procesamiento y visualización de los datos. Por supuesto, estas son solo las librerías mas utilizadas, ya que se utilizó una variedad mas amplia.

Se realiza la carga de archivos CSV por cada conurbación y se separan las *features* (columnas) de la variable objetivo (precio del inmueble). Con ello en mente, se diseña una función para cada modelo de ML. Cada función se encarga de escalar las *features* de manera de estandarizar un rango de valores para todas las variables, de lo contrario el modelo podría sesgarse a variables que varían mucho en su propia escala pero no en sintonía con el resto. Luego, se realiza el proceso de *Cross-Validation* descrito en el marco teórico. Se utiliza una división en 5 *folds*, es decir, el conjunto de los datos de cada conurbación se divide en 5, con lo cual 4 (80 % de los datos) se utilizan para entrenamiento y 1 (20 % de los datos) para pruebas, rotando el subconjunto de pruebas en cada iteración. En cada una de estas iteraciones (dependiendo de la función) se ajustan los datos a uno de los tres modelos base y se calculan las métricas de R^2 , RMSE y MAE, de las cuales se guardará un promedio. A su vez, se calcula la *Feature Importance* con los métodos descritos en el marco teórico: Para el caso de la regresión lineal, se extrae un promedio de los coeficientes de regresión tras cada iteración. Respecto al árbol de decisión, se realiza el cálculo de impureza de cada variable tras cada iteración y se almacena un promedio de dicho cálculo. Respecto a GAM lineal, se realiza un *PDP* promedio de cada iteración, del cuál se extrae que *features* hicieron que el precio del inmueble variara mas.

Con lo anterior se cumplen los pasos mas importantes de la fase de ajuste y entrenamiento, escalar para estandarizar sus rangos, utilizar cross validation y extraer promedios tanto del desempeño como de la importancia de atributos que permiten la realización de inferencia.

Habiendo ya cumplido los objetivos de esta sección, simplemente resta visualizar dichos elementos de manera que puedan entenderse de mejor forma los resultados. Esta visualización

se detallará en la sección siguiente.

CAPÍTULO 4

VALIDACIÓN DE LA SOLUCIÓN

Antes de comenzar directamente con el análisis comparativo de modelos, se considera importante añadir las matrices de correlación de cada conurbación (GS en figura 38, GV en figura 39 y GC en figura 40, ya que serán de utilidad al intentar comprender el grado de correlación entre cada atributo con otros.

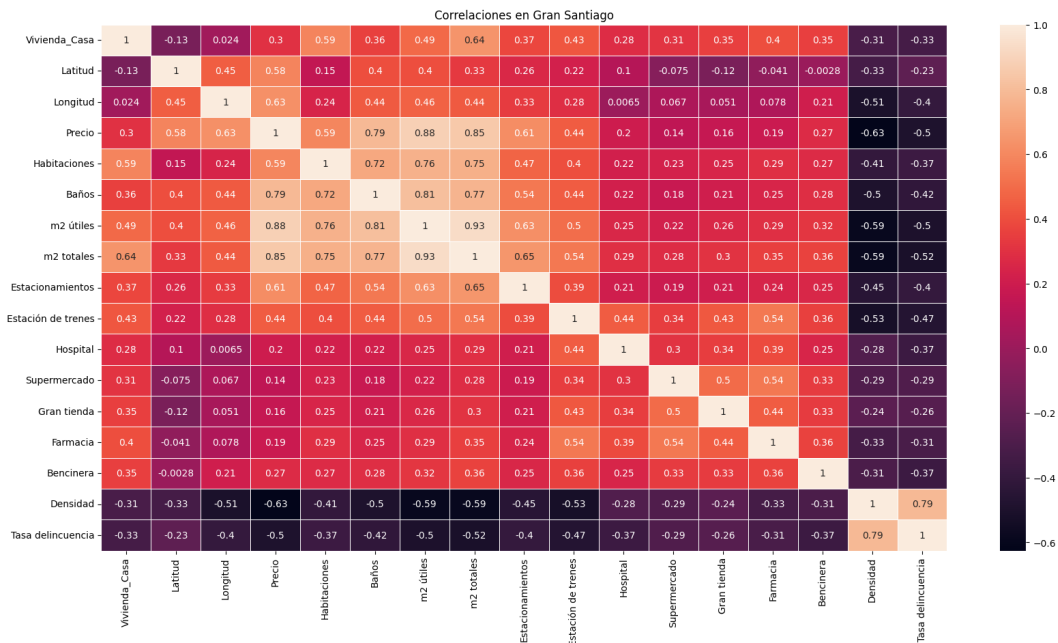


Figura 38: Matriz de correlación en GS.
Fuente: Elaboración propia.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

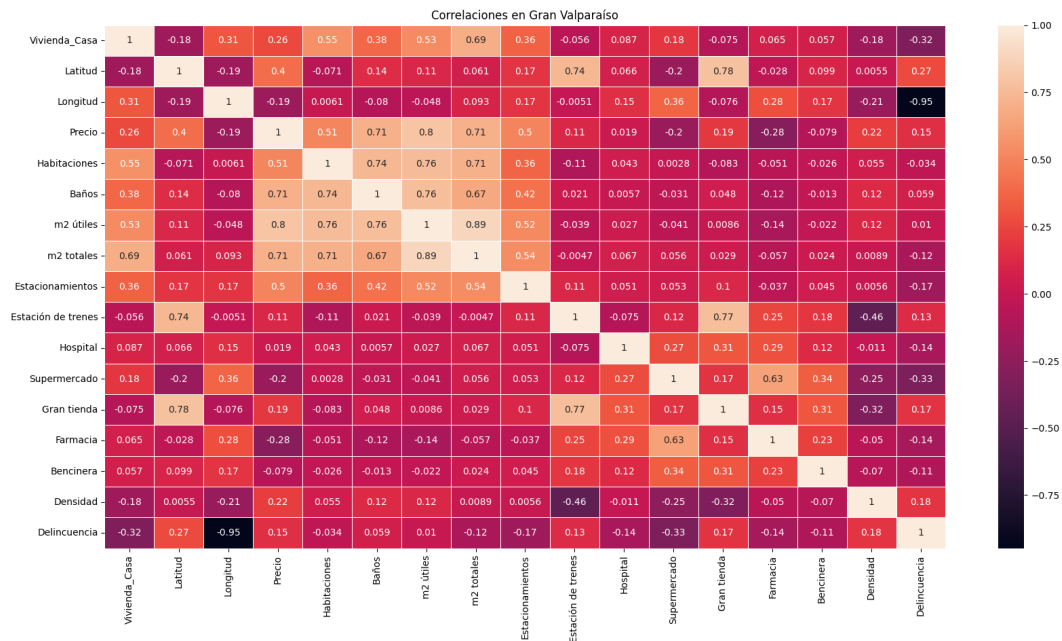


Figura 39: Matriz de correlación en GV.
Fuente: Elaboración propia.

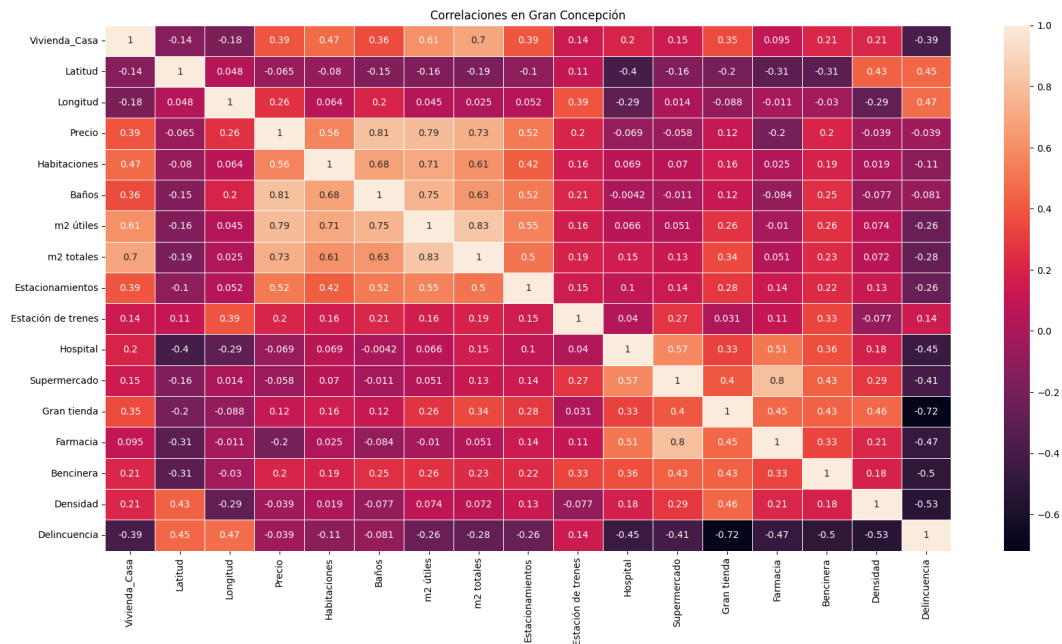


Figura 40: Matriz de correlación en GC.
Fuente: Elaboración propia.

4.1. Análisis, inferencia y predicción de modelos

Esta subsección tiene por objetivo realizar un análisis sobre los resultados obtenidos, tanto por la importancia de atributos calculada para cada modelo, como las métricas de desempeño observadas en cada modelo. Esto se realiza a nivel de conurbación.

Debe recordar el lector que, al aplicar la técnica de *Cross Validation* a cada uno de los modelos, se obtiene un promedio tanto de métricas de desempeño como de los “pesos” asociados a la importancia otorgada por los modelos.

4.1.1. Enfoque en GS

Habiendo realizado el entrenamiento de los modelos para GS, se expone a continuación un promedio de importancia de las variables de cada uno de los modelos tratados. En primer lugar, se detalla la importancia obtenida en el modelo de regresión lineal (véase figura 41).

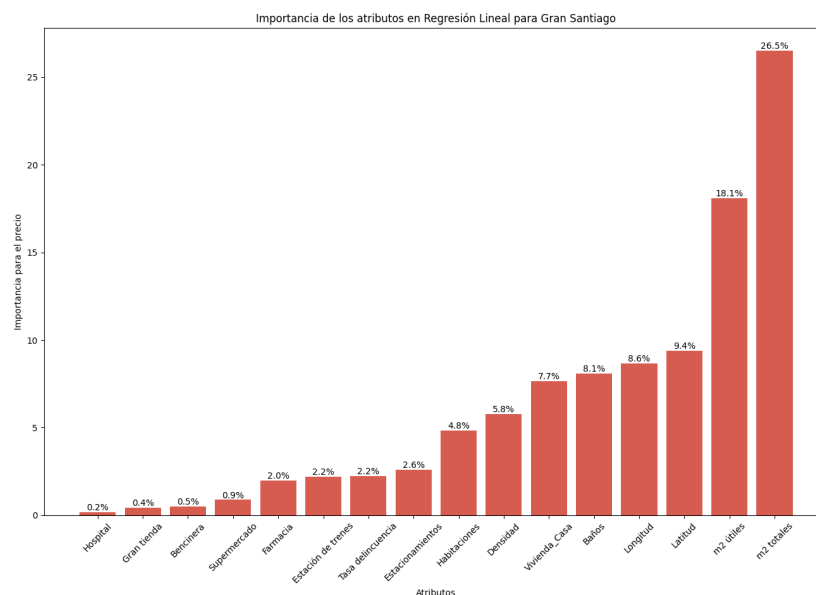


Figura 41: Importancia (en porcentaje) de regresión lineal en GS.
Fuente: Elaboración propia.

Un primer punto a considerar para el análisis de la importancia en regresión lineal es que esta se obtiene mediante los coeficientes de regresión presentes en la ecuación de regresión lineal múltiple (véase ecuación 1). Es importante destacar que los coeficientes son una medida de “peso” asociada a cada uno de los atributos, donde mientras mayor sea el valor absoluto de este, mas importante para el modelo puede considerarse la variable.

Con ello es posible observar que con un modelo lineal de representación de los datos, se

prioriza bastante los m^2 útiles y totales del terreno (44.6 % entre ambos), los cuales cuentan dentro de atributos internos del inmueble (60.1 % en su conjunto), seguido de atributos de la posición geográfica (latitud y longitud con un 18 % en su conjunto), los cuales suman 32.2 % y finalmente el tipo de vivienda con 7.7 % de peso.

También puede observarse que las 5 variables menos influyentes fueron distancias a lugares del entorno, como Hospital, Gran Tienda, Bencinera, Supermercado y Farmacia, solo esta última teniendo un porcentaje de importancia superior al 1 %.

Otro punto importante a destacar es que la interpretación de los coeficientes es solo válida para este conjunto de atributos y las interacciones que estos tienen entre sí. Es decir, puede ser que al eliminar una variable del modelo (sobre todo alguna con alta importancia), puede que exista un comportamiento muy distinto al observado. De todas formas, luego de revisar las regresiones lineal por cada conurbación, se hará hincapié en observaciones generales respecto a este aspecto.

Luego de las observaciones a la regresión lineal, es momento de analizar el comportamiento del árbol de decisión. Es importante mencionar previamente que el árbol de decisión no capta de buena manera la naturaleza lineal de los datos dada su estructura de nodos sin ser capaz de ser representada en una ecuación.

Con ello en mente, la primera observación a destacar (véase figura 42) es el alto impacto de la variable m^2 útiles, la cuál por si sola obtiene un porcentaje de peso del 69.1 %. De por sí constituye una diferencia con la regresión lineal, ya que la mayor importancia de esta última aludía a los m^2 totales, con un 26.5 % de peso.

Una similitud entre el enfoque del árbol respecto a la regresión lineal es que sus 5 variables mas importantes son las mismas, siendo m^2 útiles (69.1 %), longitud (8.9 %), latitud (5.8 %), baños (5.2 %) y m^2 totales (4.7 %), sin embargo disponen de distintos pesos.

Cabe destacar que a excepción de la tasa de delincuencia, longitud y latitud, todas las variables de entorno son inferiores al 1 %.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

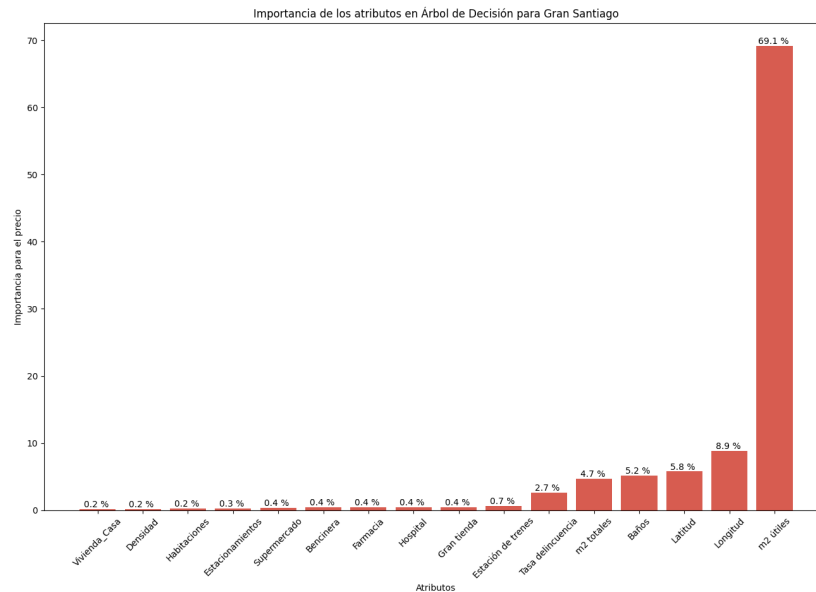


Figura 42: Importancia de árbol de decisión en GS.
Fuente: Elaboración propia.

Al sumar los pesos de los atributos internos se llega a un 79.5 %, otorgando un 20.2 % a los atributos externos y un 0.3 % al tipo de vivienda.

Por último, respecto a GAM (figura 43) que ofrece flexibilidad entre representaciones lineales y no lineales mediante funciones, mantiene un firme 25.9 % para m^2 totales, siendo la variable mas importante en este modelo, y que, por lo bajo, duplica cada variable que le antecede.

De manera similar a los modelos predecesores, mantiene en el tope a m^2 totales (25.9 %), m^2 útiles (9.7 %), latitud (9.3 %) y longitud (7.5 %) en el tope, excepto por la densidad poblacional (7.3 %), que desplaza a baños (6.3 %) al sexto lugar de importancia.

A pesar de darle mucha importancia a la variable de m^2 útiles, este modelo da bastante mas peso a las variables del entorno, llegando en su conjunto a 48.6 %, siendo la mas alta. Las únicas variables que son significativamente menos importantes son el tipo de vivienda y la distancia a farmacia, ambas inferiores a 1.5 %.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

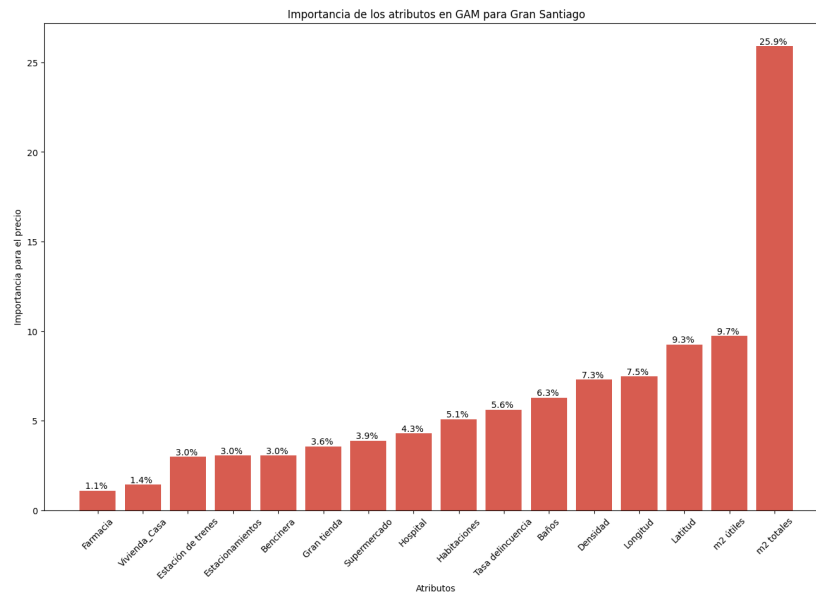


Figura 43: Importancia de GAM en GS.
Fuente: Elaboración propia.

A modo comparativo y de resumen, se muestra una tabla (tabla 4 que suma el porcentaje de atributos internos, externos y el tipo de vivienda en cada uno de los modelos tratados.

Tabla 4: Suma de porcentajes por tipo de atributo en GS.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo en GS			
Modelo	Internos	Externos	Vivienda
Regresión lineal	60.1 %	32.2 %	7.7 %
Árbol de decisión	79.5 %	20.2 %	0.3 %
GAM	50 %	48.6 %	1.4 %

Seguido a la importancia de atributos, se expone una tabla comparativa (véase tabla 5) que denota el desempeño de cada modelo en cuanto a las métricas establecidas:

Tabla 5: Métricas de modelos en GS.
Fuente: Elaboración propia.

Métricas obtenidas en GS			
Modelo	R^2	RMSE	MAE
Regresión lineal	0.906	0.277	0.208
Árbol de decisión	0.972	0.149	0.048
GAM	0.946	0.210	0.151

Es posible observar que el árbol de decisión es aquel que tiene mejor desempeño en todas las métricas, seguido por GAM y finalmente la regresión lineal. Aún así, el hecho de que todos los modelos capten un R^2 superior a 0.9 significa que todos logran representar de manera correcta los datos. Si bien no existe un rango de R^2 predefinido que denote un desempeño excelente, en muchas áreas se tolera aceptable un R^2 superior a 0.7.

A modo de complementar el análisis, se considera observar la matriz de correlación para GS (figura 38) y observar que comportamientos tienen los atributos de m^2 útiles, m^2 totales, latitud y longitud, ya que . Tanto m^2 útiles y totales tienen un nivel de correlación positiva alta (considerando un rango alto de correlación entre 0.7 y 1) respecto al precio (0.88 y 0.85 respectivamente). Esto indicaría que mientras más grande es una propiedad, el precio de la misma tenderá a subir casi linealmente con dicho crecimiento. Mismo comportamiento se observa en la latitud y longitud pero con una correlación positiva mas baja (0.58 y 0.63). Lo cuál indicaría que existe cierta tendencia a que al acercarse al sector oriente de la capital indicaría un aumento de precios, y al mismo tiempo subir hacia el norte.

4.1.2. Enfoque en GV

De manera similar a lo realizado en GS, se observa el porcentaje de importancia en cada uno de los modelos seleccionados.

En primer lugar, regresión lineal (figura 44) entrega que sus primeras 5 variables en nivel de importancia corresponden a m^2 útiles (19.8 %), latitud (13.2 %), m^2 totales (12.5 %), baños (10.9 %) para luego bajar a distancia a farmacia (7.8 %). Esta oportunidad entra un atributo de distancia a las 5 variables mas importantes. Otro punto importante a destacar es la participación mas baja del atributo Longitud (4.6 %). A su vez, las variables que menos influyeron fueron: Distancia a estación de trenes (0.1 %), tasa de delincuencia(1.7 %) y distancia a bencinera(2.3 %), siendo estos atributos externos.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

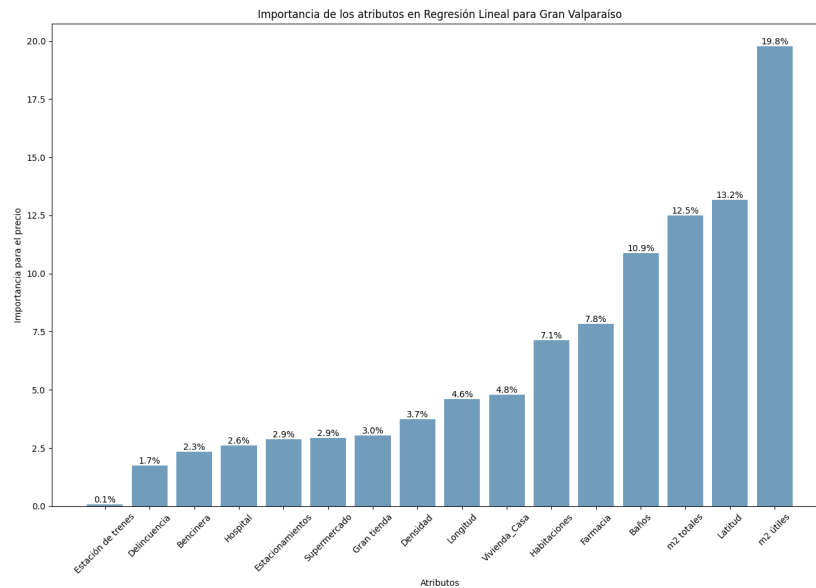


Figura 44: Importancia (en porcentaje) de regresión lineal en GV.
Fuente: Elaboración propia.

Considerando los atributos en su conjunto como internos, externos y el tipo de vivienda, los porcentajes corresponden a 53.2 %, 42 %, 4.8 % respectivamente.

Respecto al árbol de decisión (figura 45), se percibe un comportamiento similar al observado en GS, donde las variables mas importantes son los m^2 útiles (59 %), baños (10.7 %), latitud (9.9 %), m^2 totales (5.2 %), longitud (4.1 %). Nuevamente los m^2 útiles impactan de manera estrepitosa las decisiones tomadas por el árbol.

Mas allá de Latitud y Longitud, el resto de atributos externos no impacta significativamente el resultado.

En total, se consigue un peso del 76.3 % para atributos internos, un 23.5 % para atributos externos, un 0.2 % para el tipo de vivienda.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

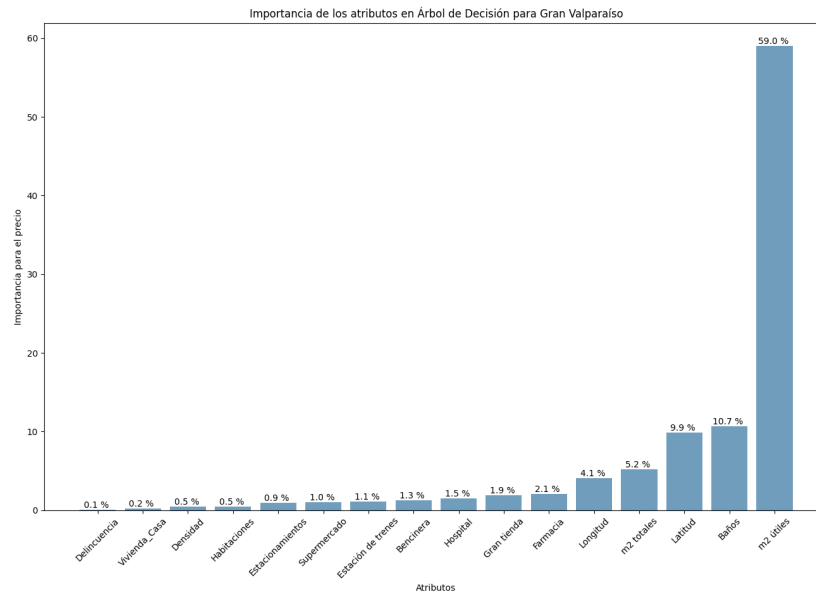


Figura 45: Importancia (en porcentaje) de árbol de decisión en GV.
Fuente: Elaboración propia.

En cuanto a GAM (figura 46), comparte similitud con regresión lineal considerando que el tercer atributo mas importante corresponde a una distancia (en este caso a bencinera) con un 8.7 % de peso. También se puede ver un cambio respecto a la regresión lineal y al árbol de decisión, considerando que esta vez se encuentra entre las primeras 5 posiciones la longitud y no la latitud. El resto de atributos en las primeras 5 posiciones son atributos que ya se veían en el árbol de decisión: m^2 totales (13.7 %), m^2 útiles (11.6 %), y baños (8.4 %).

A su vez, es el tipo de vivienda la variable menos importante para este caso (0.7 %).

Se descubre también un importante suceso, por primera vez se observa una predominancia de peso en el conjunto de atributos externos (57.2 %) respecto a los internos (42.1 %).

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

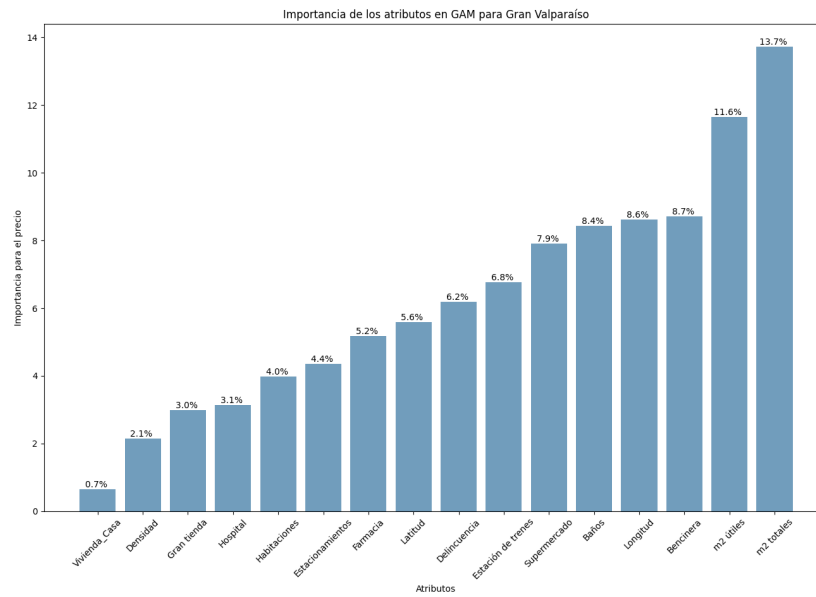


Figura 46: Importancia (en porcentaje) de GAM en GV.
Fuente: Elaboración propia.

Tabla 6: Suma de porcentajes por tipo de atributo en GV.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo en GV			
Modelo	Internos	Externos	Vivienda
Regresión lineal	53.2 %	42.2 %	4.8 %
Árbol de decisión	76.3 %	23.5 %	0.2 %
GAM	42.1 %	57.2 %	0.7 %

Tabla 7: Métricas de modelos en GV.
Fuente: Elaboración propia.

Métricas obtenidas en GV			
Modelo	R^2	RMSE	MAE
Regresión lineal	0.814	0.312	0.229
Árbol de decisión	0.920	0.202	0.063
GAM	0.844	0.285	0.205

Nuevamente se percibe superioridad (tabla 7) del árbol de decisión por sobre el resto, y se mantiene el mismo orden. La diferencia es que se redujeron significativamente los coeficientes de R^2 (aumentando a su vez RMSE y MAE), pudiendo ser debido tanto a la menor cantidad de datos como a una relación no percibida por los modelos.

Atributos que suelen repetirse en todos los modelos dentro de los 5 mas importantes son m^2 útiles, m^2 totales, latitud, baños. Sin embargo, se considera añadir la distancia a Farmacia y longitud. Si se analiza la correlación (figura 39) de estos atributos respecto al precio, puede verse que m^2 útiles y totales siguen manteniendo una correlación positiva fuerte con el precio, sin embargo latitud y longitud tienen una correlación mas débil, viendose una leve tendencia del precio a aumentar cuando la latitud aumenta (la propiedad se mueve hacia Concón) y el precio tiende a aumentar al acercarse la longitud hacia la costa (longitud disminuye) Desde Villa Alemana hasta Viña del Mar. Sumado a lo anterior, la distancia a Farmacia tiene una correlación levemente negativa (-0.28), aludiendo a que esta distancia tiende a aumentar cuando el precio de la vivienda es mas alto.

4.1.3. Enfoque en GC

Se comienza analizando la regresión lineal (figura 47) donde tres variables lideran la importancia, siendo el número de baños (12.4 %), m^2 totales (12.1 %) e índice de delincuencia (11.7 %). Aún así se observa comportamiento similar a otras considerando que a excepción de distancia a farmacia (8.4 %), el resto de distancias se encuentran de 3 % de importancia hacia abajo.

Es importante destacar que la distancia a Farmacia está altamente correlacionada (0.8) con la distancia a supermercado, por lo que es muy probable que la mayoría de las veces una farmacia esté dentro del supermercado.

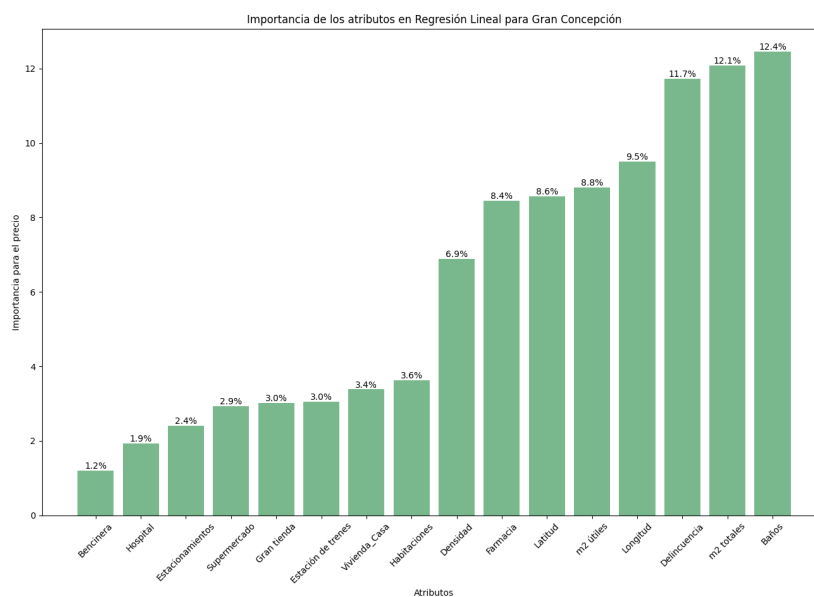


Figura 47: Importancia (en porcentaje) de regresión lineal en GC.
Fuente: Elaboración propia.

Respecto al árbol de decisión (figura 48) se observan dos variables liderando todo el peso, siendo m^2 útiles (37.8 %) y baños (37.1 %) seguido de lejos por m^2 totales (7.9 %). También se encuentran en cuarto y quinto lugar la longitud(4.5 %) y latitud(2.2 %).

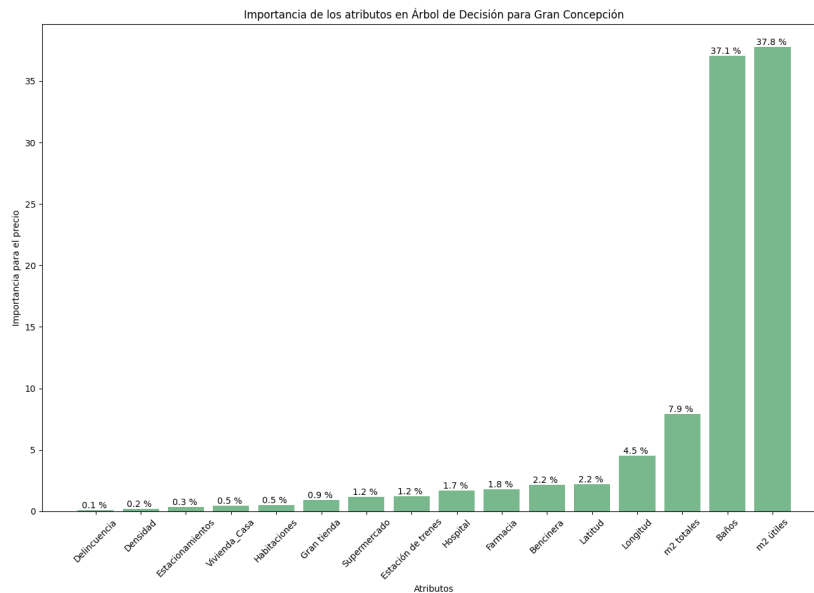


Figura 48: Importancia (en porcentaje) de árbol de decisión en GC.

Fuente: Elaboración propia.

Los resultados mas sorprendentes los entrega GAM (49), cuya variable mas importante corresponde a “Vivienda_Casa” (9 %), seguido de densidad (9 %) y m^2 totales (8.6 %). Aún así, se puede ver una escala uniforme de importancias, disminuyendo gradualmente. Este modelo es el que ha conseguido un mayor balance de importancia entre atributos. Por otro lado, estacionamientos, distancia a Gran Tienda y distancia a Estación de trenes está muy bajo del resto, con 3.5 % como máximo entre ellos.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

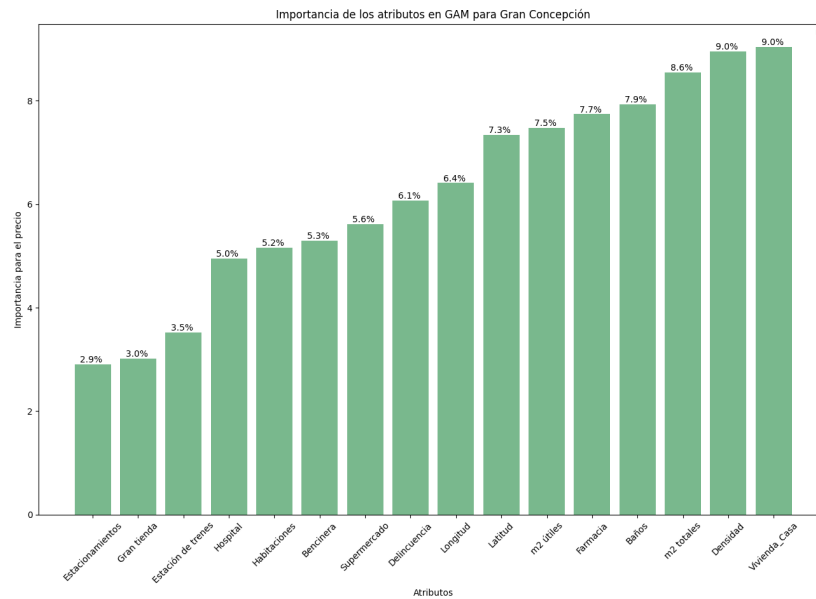


Figura 49: Importancia (en porcentaje) de GAM en GC.
Fuente: Elaboración propia.

Se puede observar (tabla 8) que se mantiene la tendencia de GAM a situarle un mayor peso al conjunto de variables de entorno mas que a variables internas, aunque además regresión lineal entra a esta tendencia, al ser el atributo de delincuencia muy fuerte en importancia. El árbol de decisión, por su parte, en vez de darle mas importancia a los atributos externos, hace exactamente lo contrario, cargando la importancia a atributos internos.

Tabla 8: Suma de porcentajes por tipo de atributo en GC.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo en GC			
Modelo	Internos	Externos	Vivienda
Regresión lineal	39.3 %	57.3 %	3.4 %
Árbol de decisión	83.6 %	15.9 %	0.5 %
GAM	32.1 %	58.9 %	9 %

Al momento de llegar a las métricas (tabla 9) se observa que el árbol de decisión es aquel que mejor desempeño muestra en cada una de las métricas, llegando casi al umbral de 0.93 de R^2 . Los modelos que mas tienden a priorizar variables de entorno por sobre variables internas suelen tener un peor desempeño.

Tabla 9: Métricas de modelos en GC.

Fuente: Elaboración propia.

Métricas obtenidas en GC			
Modelo	R^2	RMSE	MAE
Regresión lineal	0.808	0.299	0.228
Árbol de decisión	0.928	0.181	0.063
GAM	0.771	0.320	0.247

Se puede observar también que la regresión lineal logra obtener métricas de desempeño incluso mejores que GAM.

Un detalle interesante observado en la matriz de correlación es el hecho de que mientras mas distancia exista desde la propiedad a un centro comercial, la tasa de delincuencia tiende a disminuir drásticamente, pudiendo significar que la delincuencia es muy grave en zonas más urbanas.

4.1.4. Discusión sobre multicolinealidad

Al interpretar la regresión lineal es menester considerar que un coeficiente negativo indica que esa variable posee una correlación inversa al incremento de la variable objetivo (en este caso el precio del inmueble) y en caso de ser positivo, indica una correlación positiva entre ambas variables. Dicho esto, es claro percibir que hay coeficientes cuya correlación en el modelo de regresión lineal no se condicen a la correlación observada en la matriz de correlación de las conurbaciones (figuras 38, 39 y 40). Esto es un claro ejemplo de multicolinealidad, lo cuál ocurre cuando hay atributos predictores que estan muy correlacionados entre si (usualmente con un coeficiente mayor a 0.7). Con ello, es claro observar en la matriz de correlación para el GS que en general los atributos internos de un inmueble (a excepción de los estacionamientos) tiene una correlación positiva bastante alta (sobre 0.7 en todos los casos y llegando hasta 0.93). Si se observa como ejemplo la regresión lineal ajustada para GS (figura 50) se nota que vivienda_casa y habitaciones tienen un coeficiente negativo, pero observando la matriz de correlación se observa que su correlación respecto al precio es positiva.

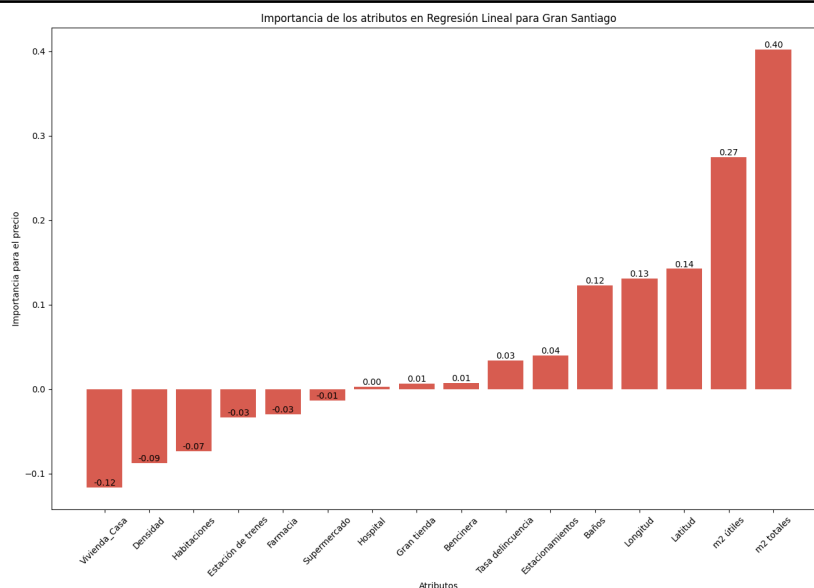


Figura 50: Coeficientes ajustados en regresión lineal para GS.

Fuente: Elaboración propia.

[Frost, 2016] Menciona que el principal problema que puede causar la multicolinealidad es la dificultad de poder interpretar los coeficientes entregados por una regresión lineal (se torna difícil explicar cada impacto de las variables por sí solas), dado que algunas variables, al verse influidas por otras, pueden cambiar su signo de correlación (Aún así, en esta investigación se considera el valor absoluto como la importancia) e incidencia en la variable objetivo.

Al eliminar variables del modelo intentando buscar una mejor interpretación de los coeficientes para la regresión lineal se obtenía una baja de las métricas de desempeño, sumado a que aún faltaba multicolinealidad por eliminar, ya que al eliminar una capa, parecía yacer otra debajo.

Es por esta premisa que el análisis realizado debe considerarse específicamente para este conjunto de atributos y las interacciones entre ellos. Además, se optó por priorizar el desempeño de los modelos, ya que si bien se podía obtener una mejor interpretabilidad con menor desempeño, realmente no ayuda el hecho de que un desempeño sea pobre y la inferencia sea “mas clara”.

4.2. Análisis y comparación regional de mismos modelos

Ya que anteriormente se detallaron los resultados de los modelos para una conurbación en específico sin un enfoque interregional, se analizará este último comportamiento a continuación.

4.2.1. Enfoque en regresión lineal

Respecto al comportamiento general de la regresión lineal respecto a los atributos considerados (tabla 10) A medida que se opta por dar mas importancia a los atributos externos se reduce al mismo tiempo el desempeño (tabla 11) del modelo y viceversa.

Aún así, a pesar de la baja de desempeño que pueda tener la regresión lineal respecto al número de datos vistos, mantiene cierta consistencia en las variables que considera importancia, ya que dentro de las primeras 6 siempre se encuentran m^2 útiles, m^2 totales, baños, alguna variable de distancia, latitud y longitud.

Tabla 10: Suma de porcentajes por tipo de atributo en regresión lineal.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo (regresión lineal)			
Conurbación	Internos	Externos	Vivienda
GS	60.1 %	32.2 %	7.7 %
GV	53.2 %	42.2 %	4.8 %
GC	39.3 %	57.3 %	3.4 %

Tabla 11: Métricas de modelos basadas en regresión lineal.
Fuente: Elaboración propia.

Métricas obtenidas en regresión lineal			
Conurbación	R^2	RMSE	MAE
GS	0.906	0.277	0.208
GV	0.814	0.312	0.229
GC	0.808	0.299	0.228

4.2.2. Enfoque en árbol de decisión

Los árboles de decisión fueron aquellos que otorgaron mejores métricas de desempeño (tabla 13) a lo largo de toda la experimentación (siempre con un R^2 sobre 0.92). Sumado a esto, tienden a ser consistentes con los atributos mas importantes (m^2 útiles y totales, latitud, longitud y baños), dando mucha mas importancia a los atributos internos en general. Aún así, resulta importante destacar que a pesar de que la conurbación de GC tuviera menos datos disponibles, el ajuste fuese levemente mejor en esta conurbación que en GV, siendo además esta última conurbación la cuál mas peso se le dió a las variables del entorno (tabla 12).

Tabla 12: Suma de porcentajes por tipo de atributo en árbol de decisión.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo (Árbol de decisión)			
Conurbación	Internos	Externos	Vivienda
GS	79.5 %	20.2 %	0.3 %
GV	76.3 %	23.5 %	0.2 %
GC	83.6 %	15.9 %	0.5 %

Tabla 13: Métricas de modelos basadas en árbol de decisión.
Fuente: Elaboración propia.

Métricas obtenidas en árbol de decisión			
Conurbación	R^2	RMSE	MAE
GS	0.972	0.149	0.048
GV	0.920	0.202	0.063
GC	0.928	0.181	0.063

4.2.3. Enfoque en GAM

Respecto a GAM, esta intentó balancear los atributos internos con los externos en GS (tabla 14), sin embargo, a medida que se disminuía el número de datos, el desempeño (tabla 15) tiende a bajar de manera mas drástica que otros modelos en todas las métricas.

Puede notarse también que a medida que mas peso le otorgaba a m^2 totales y útiles, el desempeño del modelo parecía mejorar, sin embargo, en el caso del GC no había una variable que realmente fuera mucho mas importante que otra, sino que existía una degradación paulatina.

Tabla 14: Suma de porcentajes por tipo de atributo en GAM.
Fuente: Elaboración propia.

Suma de porcentajes por tipo de atributo - GAM			
Conurbación	Internos	Externos	Vivienda
GS	50 %	48.6 %	1.4 %
GV	42.1 %	57.2 %	0.7 %
GC	32.1 %	58.9 %	9 %

Tabla 15: Métricas de modelos basadas en GAM.

Fuente: Elaboración propia.

Métricas obtenidas en árbol de decisión			
Conurbación	R^2	RMSE	MAE
GS	0.946	0.210	0.151
GV	0.844	0.285	0.205
GC	0.771	0.320	0.247

CAPÍTULO 5

CONCLUSIONES

5.1. Conclusiones Generales

Este escrito surge como una necesidad de hacer transparentes dos aspectos principales respecto a los precios de venta de inmuebles en las conurbaciones más pobladas de Chile. El primero es tener un acercamiento empírico respecto a los determinantes principales del precio de estos inmuebles basado en el modelo hedónico. El segundo alude a la transparencia intrínseca de ciertos modelos de *ML* clásicos que, a diferencia de técnicas modernas y de un rendimiento superlativo (como las redes neuronales artificiales), se puede indagar fácilmente en su estructura para ir más allá de un objetivo predictivo y permitir responder preguntas mucho más profundas, adaptando las mismas al contexto en el que se está desarrollando la solución (Salud, Financiero, Industria, etc.) y acercándose a cómo los modelos de *ML* toman decisiones.

Dicho lo anterior, se detalla la obtención de resultados en favor de los objetivos específicos planteados al inicio de la investigación: En primer lugar, se propuso el contraste de inferencia de modelos por cada una de las conurbaciones vistas en la investigación. Respecto a este tópico, en GS existe consistencia entre los tres modelos respecto a las variables más importantes, siendo principalmente m^2 útiles y totales, latitud, longitud (no necesariamente en orden). Se destaca que el árbol de decisión considera m^2 útiles con un 69.1 % de importancia, estando este atributo muy por sobre los demás (siguiéndolo latitud con 8.9 % de importancia). Se destaca también la diferencia de enfoques de los modelos, donde el árbol de decisión considera m^2 útiles como la variable más importante con creces, mas no así la regresión lineal y GAM, las cuales consideran a los m^2 totales como la variable más importante. Respecto a las variables externas, ninguna distancia a algún elemento (Hospital, metro, supermercado) fue determinante. En el caso de GV se observa en todos los modelos un mayor balance respecto a la importancia de los atributos (los porcentajes de los atributos más importantes bajan su magnitud), no obstante el árbol de decisión mantiene las mismas variables dentro de las cinco más importantes. Se puede apreciar, que en los modelos de regresión lineal y GAM hay una tendencia de variables de entorno más importantes como la distancia a farmacia en caso de la primera y distancia a bencinera respecto a la segunda. Respecto a GC, se nota un balance mucho mayor entre las variables más importantes en los modelos de regresión lineal y GAM (variables no sobresalen mucho más respecto a otras). El árbol de decisión permanece con las mismas variables más importantes, sin embargo distribuye el podio entre m^2 útiles y cantidad de baños. Se destaca que en el caso de la regresión lineal y GAM el peso otorgado en importancia a las variables externas suele ser mayor mientras menos datos se tengan, lo cual no ocurre con el árbol de decisión.

Dentro de lo plasmado en el párrafo anterior, se perciben atisbos de otro de los objetivos, el cuál alude a la comparativa regional de inferencia de un mismo modelo. En primer lugar, la

regresión lineal otorga más importancia a los atributos externos a medida que se utiliza en una conurbación que posee menos datos, sin embargo, logra mantener una consistencia de atributos importantes, oscilando entre m^2 útiles, m^2 totales, latitud, longitud, alguna variable de distancia, y baños. Respecto al árbol de decisión, este se comporta de manera muy similar en todas las conurbaciones, variando ligeramente el peso que le asigna al conjunto de atributos internos y externos. Aún así, no tiene el mismo comportamiento que los modelos contendientes. Por último GAM, presenta inconsistencia respecto a lo que considera importante en cada conurbación, por ejemplo en GC la variable más importante es el tipo de vivienda, una variable que fuera de esa conurbación prácticamente no tiene importancia, sumado a que mientras menos datos se tienen, es el modelo que menos se decanta por una variable en específico, sino que trata de establecer un balance entre todas.

Otro objetivo alude a la comparación de los modelos en torno a su desempeño. Primeramente el árbol de decisión tiene un desempeño superior respecto a sus rivales en cada una de las conurbaciones revisadas, teniendo su R^2 mas bajo con un valor de 0.92. GAM supera a la regresión lineal, solo en GS y GV, sin embargo, es la que mas rápido decae en precisión en todas las métricas (bajando un R^2 desde 0.94, pasando por 0.84 y terminando en 0.77). La regresión lineal si bien no tiene un desempeño sobresaliente, si logra mantener una caída de desempeño mucho más suave que GAM, pasando desde 0.9, por 0.81 y 0.808 en GS, GV, y GC respectivamente.

Por último, pero no por ello menos importante, la visualización de la importancia de las variables (concepto ligado profundamente a la inferencia) se aprecia en todas las figuras de la sección de Validación de la solución. Esta visualización permitió una interpretación mucho más sencilla del comportamiento de los atributos al ser ajustados al modelo.

Con base en lo anterior mencionado, queda en evidencia que en cuanto a desempeño, el árbol de decisión supera con creces a sus contendientes, manteniendo a su vez consistencia respecto a los atributos más importantes, siendo m^2 útiles, m^2 totales, latitud, longitud y cantidad de baños, tan solo variando levemente sus pesos a lo largo de todas las conurbaciones. GAM por su parte es aquel que más disminuye su desempeño mientras menos datos tenga, y aquel en el cuál más fluctúa la importancia de las variables del modelo, priorizando en exceso variables externas. Aún así, dentro de las mejores variables, suelen estar los m^2 útiles y totales, alguna variable geográfica (ya sea latitud, longitud o ambas) y al menos una variable externa. La regresión lineal, por su parte, es un punto intermedio, si bien su consistencia respecto a los atributos más importantes fluctúa más que en el árbol de decisión, sigue manteniendo más consistencia que GAM, sumado a que sus métricas de desempeño no cae estrepitosamente. Con ello, y sintetizando lo obtenido por todos los modelos, en GS el conjunto de atributos más importantes radica en ambas medidas de m^2 (estando muy por encima del resto), latitud, longitud y baños, sumado a la variable de densidad en algunos casos. Respecto a GV, se mantiene una estructura similar, pero dando cierto énfasis a variables de entorno de distancia, como distancia a Farmacia o Bencinera. Por último, en GC los Baños juegan un papel importante junto a los m^2 , sin embargo, el análisis se entorpece al tener una cantidad de datos menor, dándole importancia a la tasa de delincuencia del sector o al tipo

de vivienda.

Una de las motivaciones claves para elaborar esta investigación queda plasmada en la tabla comparativa de investigaciones (tabla 19) en la cuál, se hacen procedimientos similares, más no se intenta abordar un enfoque comparativo regional en Chile, tanto de modelos de *ML* como de la interpretación que se puede extraer de los propios modelos. Aún así, tales enfoques se centran en tópicos que en esta investigación no se tratan (más modelos de *ML*, inferencia con otros atributos, modelos que abarcan la temporalidad dentro de su lógica, etc), por lo cuál, el lector queda invitado a revisarlos.

5.2. Trabajos futuros

Dentro de este tema, es importante destacar varias consideraciones para resaltar en futuras investigaciones. La primera de ellas, se refiere a la falta de temporalidad de los datos. Al no tener en cuenta un periodo de tiempo más amplio (solo abril de 2023), es posible que se hayan omitido ciertos aspectos relevantes del entorno económico. También es necesario considerar que, en esta investigación, al utilizar el modelo hedónico, se excluyen variables de mercado (como la inflación) que pueden tener un impacto significativo en el precio de una propiedad, especialmente si se realiza un análisis temporal de estos precios.

Es crucial considerar la búsqueda de un conjunto de datos más representativo para todas las regiones, incluida la Región Metropolitana. Una de las observaciones más críticas fue que, a medida que se disponía de menos datos, los modelos tendían a disminuir su desempeño. Esto podría lograrse trabajando con sitios web que proporcionen bases de datos de transacciones e información completa de las propiedades, o utilizando técnicas de *web scraping* en diferentes sitios web que capturen de manera exhaustiva el entorno de las viviendas. Además, esto permitiría una mejor estandarización de la inferencia que se puede obtener de los modelos, fortaleciéndolos. Asimismo, considerar más atributos que los presentados en la investigación y que puedan ser replicados en cada una de las conurbaciones podría proporcionar un espectro más amplio para determinar el precio de una propiedad. Por ejemplo, incluir una medida más precisa de la distancia al transporte público, distancia a colegios, bodegas, entre otros. También sería interesante explorar métodos de *imputación* de datos, que consisten en no descartar inmediatamente un atributo debido a la falta de ciertos valores, sino estimarlos en función de los demás atributos y ejemplos disponibles en el mismo conjunto de datos.

ANEXOS

```
#Ciclo que itera a través de todas las imágenes contenidas en la lista
for i in range(1,31):
    #Elemento de imagen guardado para ser clickeado posteriormente
    image = WebdriverWait(driver, 5).until(EC.element_to_be_clickable((By.XPATH, '//*[@id="gpi-property-list-containe

#Rotación de user agents
user_agent = random.choice(user_agents)
options.add_argument(f'user-agent={user_agent}')
```

Figura 51: Fragmento de código que muestra la rotación de “user-agent” tras cada nueva página con imágenes.

Fuente: Elaboración propia.

Tabla 16: Resumen WebScraping en GS.

Fuente: Elaboración propia.

Resumen WebScraping en GS	
Tiempo total	57:03:00
Datos en sitio web	28.155
Datos Extraídos	23.275
Porcentaje Extraído	83 %

Tabla 17: Resumen WebScraping en GV.

Fuente: Elaboración propia.

Resumen WebScraping en GV	
Tiempo total	05:04:11
Datos en sitio web	2.594
Datos Extraídos	2.093
Porcentaje Extraído	81 %

Tabla 18: Resumen WebScraping en GC.

Fuente: Elaboración propia.

Resumen WebScraping en GC	
Tiempo total	00:53:42
Datos en sitio web	412
Datos Extraídos	314
Porcentaje Extraído	76 %

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

A	B	C	D	E	F	G
Comuna	Tiempo scrapi	Registros	Totales	%	Densidad	Delincuencia
Villa Alemana	1177	130	156	83%	1436	971,9
Quilpue	1964	204	231	88%	311	1274,8
Concon	3124	334	454	74%	603	1607,3
Vina Del Mar	9078	1105	1346	82%	2962	1593,3
Valparaiso	2926	320	407	79%	810	1779,1

Figura 52: Imagen a modo de ejemplo respecto al almacenamiento de densidad y tasa de delincuencia por comuna.

Fuente: Elaboración propia.



Paradero 6766 Mts.
Parada 1 Achupallas



Hospital 1152 Mts.
Cruz Roja



Bomberos 3093 Mts.
8a Compañía Bomberos Viña del Mar

Figura 53: Distancia a Paradero y Bomberos imprecisos.

Fuente: Elaboración propia.

años	m2 útiles	m2 totales	Estacionamientos	Estación de trenes	Hospital	Supermercado	Gran tienda	Farmacia	Bencinera	Densidad	Tasa delincuencia
22	22	0	...	546	3641	372	3035	1539	349	4236.0	1972.000267
75	128	0	...	690	3462	172	3215	1725	129	4236.0	1972.000267
88	250	0	...	1021	2775	2339	2826	1523	764	4236.0	1972.000267
120	250	3	...	807	2771	2126	2657	1469	707	4236.0	1972.000267
92	244	2	...	1460	1487	1573	1524	216	590	4236.0	1972.000267
...	...	...	...	...	...	...	...	...	...	...	...
143	182	2	...	1502	1509	348	1587	365	257	3456.2	1038.391824
160	160	1	...	3043	3016	1722	2428	652	1293	3456.2	1038.391824
400	1063	0	...	3803	3998	796	2928	581	588	3456.2	1038.391824
140	297	2	...	1946	2362	691	1163	325	511	3456.2	1038.391824
145	160	2	...	1823	2047	410	682	42	481	3456.2	1038.391824

Figura 54: Ejemplo de DF completo a nivel de conurbación de región metropolitana.

Fuente: Elaboración propia.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

```
m2 útiles:
15380      -5
15381      -5
12751       1
12753       1
2567        1
...
7792      38401
14238     110000
7379      120000
13655     155000
5330      209000
Name: m2 útiles, Length: 23275, dtype: int64
```

Figura 55: Datos atípicos en metros cuadrados útiles para el GS.
Fuente: Elaboración propia.

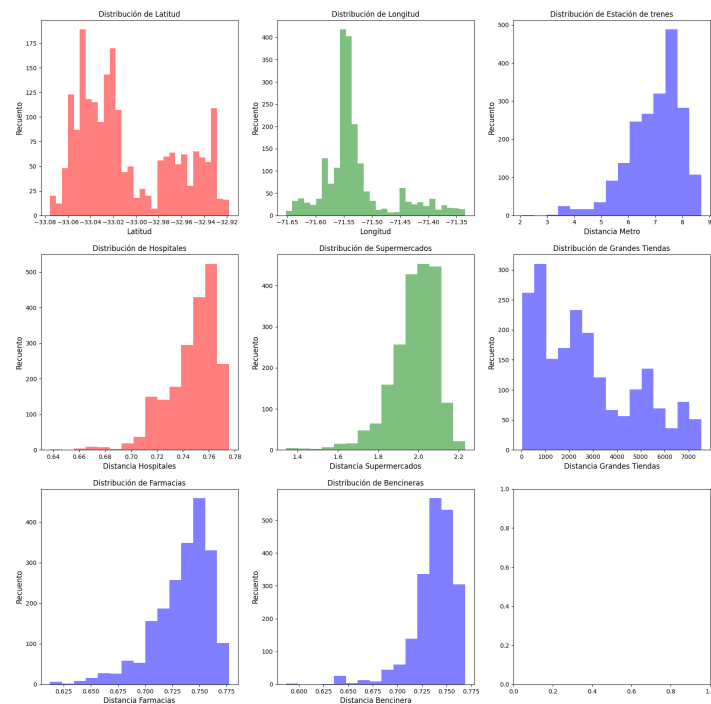


Figura 56: Evidencia sesgo en GV.
Fuente: Elaboración propia.

DETERMINANTES DEL COMPORTAMIENTO DE PRECIOS DE PROPIEDADES EN CONURBACIONES DE LAS REGIONES DE VALPARAÍSO, METROPOLITANA Y BÍO-BÍO: INFERENCIA POR MEDIO DE APRENDIZAJE AUTOMÁTICO.

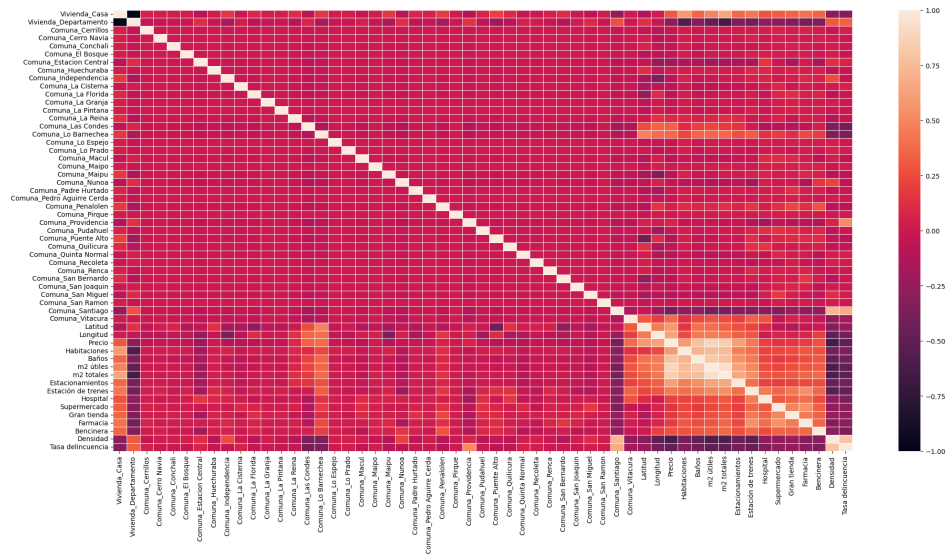


Figura 57: Correlación en comunas de GS.

Fuente: Elaboración propia.

Tabla 19: Tabla comparativa de alcances en investigaciones
Fuente: Elaboración propia.

Autor(es) y año	Predicción de precio	Interpretabilidad de modelos	Comparación de inferencia	Precisión de predicción	interpretación de coeficientes
[Rosales, 2021]	X			X	X
[Masías et al., 2016]	X			X	
[Bozanic, 2020]	X			X	
[Reyes Feris, 2022]	X			X	
[Chardon Schirm, 2019]	X	X	X	X	
[Forys, 2022]	X			X	
[Lahmiri et al., 2023]	X			X	
[Bricongne et al., 2023]					
[Rico-Juan y Taltavull de La Paz, 2021]	X	X	X	X	X
José Sansana, 2023	X	X	X	X	X

Tabla 19: Tabla comparativa de alcances en investigaciones (Continuación). Fuente: Elaboración propia

Autor(es) y año	Comuna en particular	Comuna chilena	Enfoque y comparativa regional chilena	Web scraping para datos
[Rosales, 2021]	X	X		
[Masías et al., 2016]		X		
[Bozanic, 2020]		X		
[Reyes Feris, 2022]		X		X
[Chardon Schirm, 2019]	X			
[Forys, 2022]	X			
[Lahmiri et al., 2023]	X			
[Bricongne et al., 2023]				X
[Rico-Juan y Taltavull de La Paz, 2021]	X			
José Sansana, 2023		X	X	X

REFERENCIAS BIBLIOGRÁFICAS

- [academia española de la lengua, 2021] academia española de la lengua, R. (2021). Real academia española: Diccionario de la lengua española. <https://dle.rae.es>. Accedido: Diciembre 1, 2022.
- [Alhamid, 2020] Alhamid, M. (2020). What is cross-validation? <https://towardsdatascience.com/what-is-cross-validation-60c01f9d9e75>. Accedido: Junio 5, 2023.
- [Anand, 2020] Anand, A. (2020). Feature importance explained. <https://medium.com/analytics-vidhya/feature-importance-explained-bfc8d874bcf>. Accedido: Junio 3, 2023.
- [Bozanic, 2020] Bozanic (2020). Sistema de predicción de precios-venta de inmuebles en el mercado del sector inmobiliario de la región metropolitana de la república de Chile con el uso de algoritmos de machine learning. *Universidad de Chile*.
- [Bricongne et al., 2023] Bricongne, J.-C., Meunier, B., y Pouget, S. (2023). Web-scraping housing prices in real-time: The covid-19 crisis in the uk. *Journal of Housing Economics*, 59:101906. COVID-19's Impacts on Housing Markets.
- [Carrillo et al., 2021] Carrillo, A., Cantú, L. F., y Noriega, A. (2021). Individual explanations in machine learning models: A survey for practitioners. *arXiv preprint arXiv:2104.04144*.
- [Chardon Schirm, 2019] Chardon Schirm, I. (2019). Predicción de rentas en Santiago de Chile utilizando algoritmos de aprendizaje automático. *Universidad de Chile*.
- [D'Auria, 2021] D'Auria, E. (2021). Looking at r-squared. <https://medium.com/@erika.auria/looking-at-r-squared-721252709098>. Accedido: Junio 5, 2023.
- [de apoyo a víctimas, 2022] de apoyo a víctimas, P. (2022). Estadísticas delictuales Chile. <http://cead.spd.gov.cl/estadisticas-delictuales/>. Accedido: Abril 19, 2023.
- [de la Construcción, 2022] de la Construcción, C. C. (2022). Oferta de mercado inmobiliario en Chile. <https://cchc.cl/centro-de-informacion/indicadores/mercado-inmobiliario-oferta-nacional>. Accedido: Abril 19, 2023.
- [Deepchecks, 2023] Deepchecks (2023). Mean absolute error (mae) - deepchecks. <https://deepchecks.com/glossary/mean-absolute-error/>. Accedido: Junio 5, 2023.
- [del congreso nacional de Chile, 2022a] del congreso nacional de Chile, B. (2022a). Biblioteca del congreso nacional de Chile. <https://www.bcn.cl>. Accedido: Abril 19, 2023.
- [del congreso nacional de Chile, 2022b] del congreso nacional de Chile, B. (2022b). Ley 21459. <https://www.bcn.cl/leychile/navegar?idNorma=1177743>. Accedido: Marzo 24, 2023.

- [Du *et al.*, 2019] Du, M., Liu, N., y Hu, X. (2019). Techniques for interpretable machine learning. Department of Computer Science and Engineering, Texas A&M University.
- [Forys, 2022] Forys, I. (2022). Machine learning in house price analysis: regression models versus neural networks. *Procedia Computer Science*, 207:435–445. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022.
- [Frost, 2016] Frost, J. (2016). Multicollinearity in regression analysis. <https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>. Accedido: Junio 10, 2023.
- [Frost, 2023] Frost, J. (2023). Mean squared error (mse) in regression. <https://statisticsbyjim.com/regression/mean-squared-error-mse/>. Accedido: Junio 5, 2023.
- [Goplacelt.com, 2022] Goplacelt.com (2022). Términos y condiciones. <https://www.goplacelt.com/cl/terminos-y-condiciones>. Accedido: Marzo 24, 2023.
- [Hasan Parvez, 2023] Hasan Parvez (2023). Evaluating regression models performance. <https://www.aionlinecourse.com/tutorial/machine-learning/evaluating-regression-models-performance>. Accedido: Junio, 2023.
- [Hastie T, 1990] Hastie T, T. R. (1990). *Generalized Additive Models*. Monographs of Statistics and Applied Probability 43.
- [Hayunga, 2020] Hayunga, P. . (2020). Examining the information content of residuals from hedonic and spatial models using trees and forests. *Journal of Real Estate Finance and Economics*, 60(1-2):170–180.
- [Kim *et al.*, 2016] Kim, B., Khanna, R., y Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.
- [Kumar, 2021] Kumar, A. (2021). Machine learning: Inference and prediction difference. <https://vitalflux.com/machine-learning-inference-prediction-difference/>. Accedido: Diciembre 6, 2022.
- [Lahmiri *et al.*, 2023] Lahmiri, S., Bekiros, S., y Avdoulas, C. (2023). A comparative assessment of machine learning methods for predicting housing prices using bayesian optimization. *Decision Analytics Journal*, 6:100166.
- [Lipton, 2016] Lipton, Z. C. (2016). The mythos of model interpretability. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016).
- [Masías *et al.*, 2016] Masías, Valle, C., Crespo, Vargas, y Laengle (2016). Property valuation using machine learning algorithms: A study in a metropolitan-area of chile. *In Selection at the AMSE Conferences* (p. 97).

- [Miller, 2017] Miller, T. (2017). *Explanation in artificial intelligence: Insights from the social sciences*. School of Computing and Information Systems University of Melbourne, Melbourne, Australia.
- [Mitchell, 2018] Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. O'Reilly Media, Inc.
- [Molnar, 2020] Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.
- [Nacionales, 2007] Nacionales, B. (2007). *Manual de tasación*. Bienes Nacionales.
- [Quesada, 2015] Quesada, J. R. (2015). Uso del método de precios hedónicos, para estimar variación en precios de viviendas producto de nuevas líneas de metro. *Universidad de Chile*.
- [Reyes Feris, 2022] Reyes Feris, A. (2022). Desarrollo de un servicio de predicción de precios de venta y arriendo para inmuebles en Chile, Colombia y México. *Universidad de Chile*.
- [Rico-Juan y Taltavull de La Paz, 2021] Rico-Juan, J. R. y Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? an analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, 171:114590.
- [Rosales, 2021] Rosales, H. (2021). Real estate pricing models using machine learning schemes. *International Conference on Industrial Engineering and Operations Management*.
- [Rosen, 1974] Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34-55.
- [Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206-215.
- [Schmitt, 2022] Schmitt, M. (2022). Interpretable machine learning: Why and how to make your machine learning models interpretable. <https://www.datarevenue.com/en-blog/interpretable-machine-learning>. Accedido: Diciembre 2, 2022.
- [Shalev-Shwartz Shai, 2014] Shalev-Shwartz Shai, B.-D. S. (2014). Understanding machine learning (from theory to algorithms | | decision trees). *Cambridge University Press*.
- [Stiglic et al., 2020] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., y Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379.
- [Subramanian, 2021] Subramanian, D. (2021). The mathematics of decision trees, random forest, and feature importance in scikit-learn and spark. Towards Data Science. Accedido: Junio, 2023.

[tasando.cl, 2022] tasando.cl (2022). Tasaciones gratis y en segundos. <https://tasando.cl>. Accedido: Diciembre 5, 2022.

[Trajan, 2020] Trajan (2020). Decision trees regression and splits. <https://stats.stackexchange.com/questions/490185/decision-tree-regression-and-splits>. Accedido: Diciembre, 2022.

[Yan y Su, 2003] Yan, X. y Su, X. G. (2003). Linear regression analysis. *Theory and Computing*.

[Ñanculef R., 2022] Ñanculef R. (2022). Árboles y otras hierbas. Accedido: Diciembre, 2022.