UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

SISTEMA DE PREDICCIÓN DE PRECIOS-VENTA DE INMUEBLES EN EL MERCADO DEL SECTOR INMOBILARIO DE LA REGIÓN METROPOLITANA DE LA REPÚBLICA DE CHILE CON EL USO DE ALGORITMOS DE MACHINE LEARNING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MIRKO SLOVAN BOZANIC LEAL

PROFESOR GUÍA:
CHARLES THRAVES CORTÉS-MONROY

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
MARCEL GOIĆ FIGUEROA

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: MIRKO SLOVAN BOZANIC LEAL
FECHA: 2020
PROF. GUÍA: CHARLES THRAVES CORTÉS-MONROY

SISTEMA DE PREDICCIÓN DE PRECIOS-VENTA DE INMUEBLES EN EL MERCADO DEL SECTOR INMOBILARIO DE LA REGIÓN METROPOLITANA DE LA REPÚBLICA DE CHILE CON EL USO DE ALGORITMOS DE MACHINE LEARNING

El presente trabajo tiene por objetivo minimizar el error de pronóstico en la predicción de precios-venta en cada uno de los tipos de inmueble y operaciones inmobiliarias existentes en el mercado. Se seleccionan - o mantienen - los mejores algoritmos, con la finalidad de que la empresa pueda mejorar la predicción de los precio-venta asociados a los inmuebles solicitados por sus clientes. Vendiendo un producto más confiable, que facilite y logre clientes fidedignos con el tiempo.
Se cuantificará el beneficio financiero que esta investigación le otorga a la empresa.

Se trabaja con datos de la Región Metropolitana fluctuantes entre marzo del 2019 a mayo del 2019. Éstos poseen características observables que serán de utilidad a la hora de predecir con los algoritmos de predicción enunciados en la presente memoria.
Cabe señalar que fueron extraídos de la página web: *www.portalinmobiliario.com*.

Se realiza la estimación del precio de venta de los bienes inmobiliarios pertenecientes a las siguientes categorías: (1) arriendo-comercial, (2) arriendo-casa, (3) arriendo-departamento, (4) arriendo-oficina, (5) venta-comercial, (6) venta-casa, (7) venta-departamento, (8) venta-oficina. Para ello se utilizarán algoritmos de Machine Learning con métodos de regresión con solo una variable dependiente, normalizada con logaritmo natural. Se testearán 7 modelos de regresión: Linear Regression, Decission Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, Artificial Neural Networking, Kernel Smoothing Regression.

Se ha concluido que Random Forest es el algoritmo más eficiente - en comparación al que posee la empresa - a la hora de predecir los precio-venta analizados. Particularmente por la disminución del MAE en los siguientes escenarios: (1) arriendo de oficinas (MAE: -0,0097), (2) departamentos (MAE: -0,0034) y (3) venta oficinas (MAE: -0,0261) dentro de la Región Metropolitana de la República Chile.
En paralelo, el modelo de la empresa - Generalized Boosted Regression - es un mejor modelo predictivo en el mercado de venta de casas y departamentos.
Considerando los resultados de lo anteriormente expuesto; se concluye un beneficio estadístico-total de $0,039$ puntos MAE para la ganancia. Lo anterior conlleva a una ganancia monetaria-semestral de $CLP 9.750.000.

Se aconseja a la empresa inmobiliaria centrar su análisis estadístico en la predicción segmentada de sus inmuebles (según sus características), en vez de buscar inmuebles similares en la vecindad de un elemento interesante para empezar a predecir en base a ellos.
Finalmente se solicita tener en consideración la optimización paramétrica, pues con ella convergerán a mejores resultados a medida que se minimiza el error asociado.

Para todos aquellos que somos primera generación. Que con el esfuerzo, sangre, sudor y lágrima de nuestros apoderados; estamos hoy leyendo esto. ¡Que siempre nos mueva el amor!

Agradecimientos

En primer lugar a mi querida familia, sin excepción alguna. En particular a mi padre Mirko Bozanic Arellano; quien hasta 5to básico me acompañó en mis estudios de matemáticas. Y a mi madre María Eliana Leal Herrera, que desde 8vo básico me acompañó a las olimpiadas de matemáticas. Sin duda alguna; sin vuestra sangre, sudor, lágrima, y apoyo; yo no estaría hoy escribiendo estas líneas.

Mi paso por la universidad cambió radicalmente mi pensamiento y forma de mirar las cosas. Es por ello que hago mención a todas las personas conocidas en esta larga; memorable; y agradable trayectoria de la cual estaré eternamente agradecido. De ellas puedo destacar a Rubén Rubilar Carillanca, amigo y hermano que, a pesar de haber tenido el primer encuentro en la Escuela de Verano - 2011, la amistad y hermandad persistieron. Tanto en especialidad como en Plan Común. Muchas gracias RubRub por ese apañe y soporte constante ¡sos un crack, que se sepa! A mis primero amigos de universidad; Daniel Bravo Liberona, Bastián Pardo Vergara y Álvaro Jara Moreno quienes sin pedir nada a cambio, me ayudaron siempre que pudieron. Una verdadera y linda amistad. Espero que este trío giles pueda entender la wea de una buena vez por todas.

El tiempo avanza, y uno - sin saberlo - continua encontrándose con personas maravillosas. Ya en especialidad, no puedo dejar de lado la mención de Carl(it)os Barría Arriagada, Ronald Leblebici Garo, Tomás Soto Jara y Macarena Andrade Muñoz.

Todas las personas son únicas, pero más aún ellas en donde destacan - entre un sin fin de cosas maravillosas - la linda voluntad y bondad que poseen. Una bondad admirable, de la cual confieso aprendí mucho; y que será recordada como parte significativa en la enseñanza de vida; dentro de esta gran aventura.

No todo es deber de carrera, o relativas a lo académico-administración universitaria. También estarán en el recuerdo los memorables momento de distensión. Es por ello que hago un particular nombramiento a Tomas Maturana Marchant. Amigo de reflexiones y meditaciones. Un verdadero "Vignolo del siglo XXI"; genuino y pacífico. Elementos clave que fueron complemento significativo en mi aprendizaje y entendimiento de vida. Un fuerte abrazo, y un cafecito a distancia, hermano mío.

Quiero nombrar a una de mis tantas familias. A mi querido equipo de Feria-Fundación Proyecto Reinserción, y a mi querido Departamento de Ingeniería Civil Industrial de la Universidad de Chile. Que en sus hermanos, amigos, compañeros, funcionarios y docentes; fueron para mí personas con quienes siempre se podía contar; hablar y compartir. Muchas, muchas gracias a cada uno. En especial a la Directiva Feria por la Reinserción Social 2016, y el cuarto piso de la Torre Industrial (desde el 2016 al 2020).

A mis maestros, Sr. Richard Weber Haas; Sr. Charles Thraves Cortés-Monroy y Sr. Marcel Goić Figueroa. Por la resiliencia, autocrítica, autoaprendizaje y entendimiento de ML.

Tabla de Contenido

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Antecedentes Generales

1.1.1. Características de la Empresa

Se trabajo en una empresa ubicada en el sector empresarial de la comuna de Las Condes con rubro FONDOS Y SOCIEDADES DE INVERSIÓN Y ENTIDADES FINANCIERAS SIMILARES perteneciente al sector industrial de servicios financieros y servicios profesionales. La empresa posee 2 misiones, y una visión que se presentan a continuación,

1. **Misión Asesoría de Inversiones:** Encontrar las mejores oportunidades de inversión en el mundo inmobiliario [23].

2. **Misión Asesoría de Compra y Venta de Inmuebles:** Facilitar el proceso de venta de un inmueble entregando asesoría integral para todas las etapas de la venta y compra, haciéndonos parte de éste desde la intención de compra o venta hasta la finalización de ésta [23].

3. **Visión:** Ser la mejor empresa tasadora del mercado que utilice algoritmos de Machine Learning en el territorio nacional.

En lo que respecta a la organización de la empresa se deben señalar – en primer lugar – los cargos adscritos en la empresa:

1. **Gerencia de Finanzas:** Levantar financiamiento para las operaciones iniciales de la empresa. Realizar funciones de contabilidad, recursos humanos y control de gestión.

2. **Gerencia de Tecnológica (TI):** Diseñar, mantener y actualizar la infraestructura tecnológica de la empresa, esto incluye el algoritmo optimizador y la página web.

3. **Gerencia de Ventas/Atención al Cliente:** Su principal objetivo es atraer clientes interesados en invertir en el mundo inmobiliario, además de asesorar, informar y

relacionarse con los clientes. No presenta una gerencia general, dado que todos los trabajadores existentes son accionistas o propietarios de la organización, por lo tanto, solo existe una organización según funciones no jerárquicas.

Se detalla al lector que en la constitución empresarial se explicita que los 3 socios fundadores, contando a los gerentes, son representantes legales.
Nótese que no existe un gerente general que comande los lineamientos de la empresa. Esto último porque la empresa está en sus inicios, y los cargos presentan horizontalidad entre quienes componen el organigrama. Este se presenta a continuación,

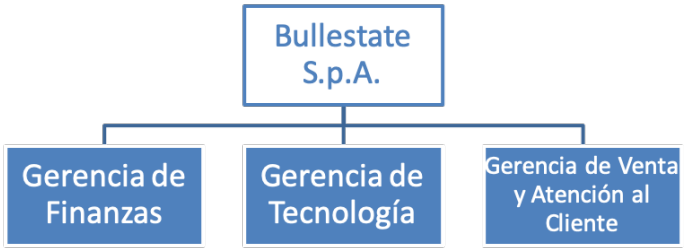

Figura 1.1: Organigrama de la Empresa - Fuente: Elaboración Propia

La empresa desarrolló un algoritmo que permite la obtención de datos de manera directa de la página web: www.portalinmobiliario.com; sobre venta y arriendo de propiedades. Permitiendo así, la comparación de precios entre inmuebles según sus características respectivas.
La empresa entrega de dos servicios:

1. **Asesoría de inversión en propiedades:** se logra una estimación en tiempo real de un valor de mercado predicho para el inmueble de interés. Con esto se da la posibilidad de adelantar la decisión - teniendo nociones de precio venta - al cliente ex ante al dictamen de un precio de tasación. Permitiendo adelantar - o al menos tener nociones - sobre la decisión de transacción para maximizar su utilidad individual. El resultado posee en su defecto, un período corto de tiempo dada el timelapse que tiene la empresa para actualizar nuevamente los datos.

2. **Asesoría para la venta y compra de inmuebles:** tasando propiedades, aconsejando sobre el valor que debería tomar una propiedad a vender acorde al contexto y características que tiene, tomando en cuenta la rapidez de venta y la rentabilidad que espera el cliente. Además, la empresa entrega el servicio de gestión de venta, equivalente al entregado por corredoras inmobiliarias.

Este último en lo que respecta al último servicio nombrado se pueden enunciar 2 unidades que lo componen. Estos se pueden caracterizar de la siguiente manera,

1. **Asesoría de compra de bien inmobiliario:** El servicio inicia a partir del contacto personal de los clientes con la empresa; e incluye el siguiente listado de acciones,

   (a) Levantamiento de requerimientos y características del bien que se desea comprar. Adicionalmente es necesario de antemano tener un objetivo de transacción. Es

decir, si se desea: (1) comprar para vender o; (2) para arrendar o; (3) para vivir.

(b) Búsqueda automatizada en distintos portales web que cumplan con las características impuestas por el cliente.

(c) Mediante algoritmos estadísticos de Machine Learning, se identifican aquellos inmuebles que cumplen de mejor forma los objetivos del cliente.
Actualmente está definido a través de un indicador de rentabilidad.

(d) Se genera un reporte personalizado con las opciones de compra posibles para el cliente.

(e) Se pide al cliente elegir las propiedades de su preferencia.

(f) Se programa visita a propiedades que el cliente selecciona, y se realiza inspección junto al cliente.

(g) En caso de compra, se realiza tramitación legal a través de empresa externa.

(h) Una vez finalizada la compra, se cobra comisión de un 1 % del precio de compra del inmueble.

2. **Asesoría de venta de bien inmobiliario:**

(a) Se realiza una predicción de regresión gratuita de la propiedad. Actualmente se realiza a través de correo electrónico.

(b) En caso de existir interés de venta, se coordina una cita para establecer características de la venta y firma del contrato ("de agente") con la empresa.

(c) Se publica anuncios de venta en los principales portales inmobiliarios.

(d) Se entrega el servicio de gestión de visitas, es decir, se muestra la propiedad a los interesados.

(e) Desarrollo de propuesta comercial con el fin de detallar puntos clave de - y para - la transacción. Ajustada a sus requerimientos y/o necesidades dentro de la operación inmobiliaria. Ello considerando la actualidad del mercado inmobiliario en el país.

(f) Tramitación de compradores, abogados y notaría para finalizar venta.

(g) Cobro de comisión por servicio de un 1 % respecto del valor de venta.

(h) Al finalizar el contrato se le ofrece al cliente la posibilidad de obtener algún servicio después de la transacción. En donde se le facilita la opción de realizar inversiones dentro del sector inmobiliario.

En lo que respecta a los posibles clientes, se puede decir que son todos aquellos que tengan la necesidad de invertir-comprar-vender un inmueble dentro del territorio nacional. Buscando una predicción que muestre el precio de venta con el menor error posible según la información

utilizada por la empresa; para poder tener una referencia lo más clara posible sobre el precio de venta del inmueble observado. Buscando saber un valor que les permita decidir correctamente para la maximización de su respectiva utilidad.

En cuanto a la dimensionalidad de la actividad realizada por la empresa, se pueden nombrar lo siguiente.

1. **En el área de corretaje:** Disponen mensualmente de 10 clientes promedio. Equivalente a 10 operaciones al mes.
   Teniendo una evolución creciente a lo largo del tiempo.

2. **En el área de inversiones:** A diferencia del corretaje, la empresa posee 5 clientes promedio activos en esta área. Cada uno con 3 asesorías. Esto crece en el tiempo.

La ventaja competitiva de la empresa está en la diferencia que existe dentro de los participantes de estos 2 mundos. Por un lado se encuentran los tasadores tradicionales que, dado un conjunto de criterios y características observables, otorgan un valor definido al inmueble. Este valor es el que los bancos - en caso de participar de la hipoteca - escogen como referencia monetaria para tomar sus decisiones.
En cambio la empresa opera prediciendo precios de mercado dado un historial de información para fijar un precio de mercado. He ahí la diferencia, en que el primero es un precio de tasación, mientras que el otro es un precio de mercado. Siendo comúnmente mayor el valor de tasación que el de mercado.

1.1.2. Marco Institucional

El sector de la empresa se divide en partes múltiples a consecuencia de sus asesorías. Estas divisiones son: (1) intermediación financiera y actividades inmobiliarias, y (2) empresariales y de alquiler. En la primera división existen, a nivel nacional, 10.477 empresas; y para el segundo caso existen 5.809 empresas [12].
La empresa cuantifica la ganancia a obtener, producto de la sustitución de sus algoritmos, mediante el uso de la métrica de error MAE. La idea central es posible resumirla en la siguiente frase: *"Considerando la métrica MAE como único criterio cuantificador para calcular la ganancia obtenida por el uso de un nuevo producto, se tendrá - y elegirá - al modelo que posea el valor más bajo en la comparación del modelo de la empresa y el modelo candidato-postulado como mejor alternativa en la sustitución".*

Si se mantiene el modelo de la empresa, no habrá ganancia generada con los modelos postulados. No obstante, si se presenta el caso contrario - en donde el MAE de los modelos postulados es menor al de la empresa en las respectivas categorías - se tendrá una ganancia monetaria plausible. En particular, un porcentaje de la ganancia promedio resultante de la mejor respuesta a la optimización de su utilidad individual. Si el modelo postulado tiene un MAE con menor valor al modelo de la empresa, se debe tomar la diferencia positiva de ambos.

Con un MAE menor al modelo original la empresa, se obtiene una ganancia. Pero ¿cómo

esto se traduce en un monto cuantificable? La respuesta está en la rentabilidad. Históricamente los clientes obtienen una rentabilidad promedio del 20 %. Es decir, considerando un ejemplo ilustrativo, se asumirá que un cliente desea vender su casa. La predicción de venta arroja $100.000.000, y el valor de tasación $120.000.000. Se tiene entonces que el oferente - vendedor del inmueble y respectivo cliente - percibirá una rentabilidad bruta de $20.000.000. Traduciéndose en un +20 % favor (detallado anteriormente). De ese 20 % ya aplicado, la empresa se lleva un cuarto de ese monto. Es decir, el 25 % de la rentabilidad obtenida por el cliente.
Al 20 % del monto del cliente a favor obtenido por el cliente, se debe añadir la diferencia positiva (de ganancia monetaria) detallada con anterioridad. Pues, la predicción va en beneficio del mismo. Sigue que la empresa extrae un cuarto de lo ganado por el cliente. Es decir, el 25 % del (20 % + [diferencia positiva]).

Los actores directos son los clientes, portales inmobiliarios, abogados y corredoras. Entre los actores indirectos se puede destacar al Servicio Nacional del Consumidor (SERNAC). En tanto, las relaciones que la empresa posee con las entidades anteriormente mencionadas son,

1. **Clientes:** La empresa conecta con estos a la hora de ofrecerles un servicio de asesoría para realizar una predicción del precio de venta del (los) inmueble (s) solicitado en cuestión. Por parte de los clientes, estos conectan al momento de aceptar el producto ofrecido por la empresa. Como consecuencia se tiene que la empresa puede fidelizar bajo estas acciones a la persona en cuestión. Si la predicción posee un buen rendimiento, entonces es posible que el cliente vuelva o la empresa sea recomendada.

2. **Portales Inmobiliarios:** La empresa se vincula con los portales inmobiliarios a través de la información. La información se materializa en bases de datos que contienen la información necesaria para predecir los precios de ventas dentro de un modelo predictivo. En paralelo, la empresa está constantemente actualizando sus bases de información. Garantizando al eventual cliente, un análisis lo más fidedigno posible, según el rango temporal analizado.

3. **Abogados:** Encargados de realizar la gestión de compra-venta de las propiedades en caso de que la oferta de la empresa sea aceptada por el consumidor demandante. Actúan como entes moderadores e intermediarios dentro de la operación entre ambos entes. Los estándares utilizados son los exigidos para ser considerados como empresa "tasadora". Entregando de esta manera, servicios fidedignos y normados en la relación existente entre el oferente y aceptante.

4. **Corredoras:** Esta relación surge producto de que, en su gran mayoría, son estas entidades las responsables de gestionar los recursos inmobiliarios. Transformándose en una competencia directa para el área de venta de la empresa.
   Esto último estimula el mercado, creando más - y mejores - precios que minimicen la brecha entre precio mercado y precio de tasación.

5. **Servicio Nacional del Consumidor:** Organismo encargado de velar por el cumplimiento de los derechos del consumidor. En este contexto, aplica como el responsable de observarlos en las consultas de predicción de precios de mercado; como en la transacción de inmuebles.

Los consumidores están protegidos ante cualquier irregularidad legal que exista en el proceso de adquisición finiquitado entre oferente y demandante.

En cuanto a regulaciones se trata, no hay mayores restricciones dentro de la creación de predicciones. No así la protección al consumidor. Para el presente caso se encuentra la ley nº 19.496 en materia de protección al consumidor. La cual indica, entre cosas:

1. El derecho a contar con información veraz y oportuna sobre los bienes y servicios ofrecidos, su precio, condiciones de contratación, etc., a la garantía de los productos, a retractarse de una compra y a respetar lo establecido en los contratos [13].

2. La no discriminación arbitraria por parte de empresas proveedoras de bienes o servicios [13].

3. La regulación de la publicidad engañosa y la eliminación de la llamada "letra chica.en los contratos [13].

Estas son unas (de las tantas) restricciones que posee una empresa tasadora dentro del mundo inmobiliario para no cometer irregularidades con su contraparte.

En lo que respecta a tendencias de mercado, pueden ser nombrados los índices IMACEM e IMACOM como marcadores de tendencia. Muestra de esta puede ser observada en la siguiente sección en la figura 1.3. Si estos suben es señal que los precios de las viviendas también lo están haciendo. Es una implicancia directa, puesto que es un indicador. No se entra necesariamente en el conflicto de correlación no implica causalidad.

1.2. Justificación del Tema

1.2.1. Información del Área de la Empresa

Las funciones principales de la empresa son: (1) prestar asesoría en la compra-venta de inmuebles, y (2) la asesoría de inversión inmobiliaria. Es una start-up que está en sus inicios (menos de 3 años de vida) que posee 4 trabajadores en total. Cada una de las 3 áreas, posee a no más de 2 trabajadores. Es más, únicamente un área (de venta) posee este par de recursos humanos. Las definiciones de estas áreas se presentan a continuación.

- **Área de Tecnología:** Responsable de entregar la propuesta de valor de la empresa. Siendo esta manifestada principalmente en el algoritmo encargado de escanear los portales inmobiliarios en sus páginas web. Creando un producto útil para el análisis de rentabilidad en la empresa.
  A su vez, es la encargada de la plataforma web realizada para informar sobre lo ofrecido por la empresa.
- **Área de Venta:** Responsable de: (1) la coordinación de reuniones con los clientes; (2) entrega de los informes de rentabilidad creados por el área de tecnología; y (3) realización de asesorías de compra-venta de propiedades. Siendo finalmente la representante

en los acuerdos emanados entre ambas partes (oferente y demandante).

- **Área de Finanzas:** Responsable de toda lo contable relacionado a la empresa manteniendo una buena salud financiera en la empresa.

Los posibles clientes son todos aquellos que tengan la necesidad de comprar-vender-invertir en algún inmueble dentro del territorio nacional. Estos buscan conocer el precio de venta más correcto posible, con el fin de poder decidir - y discernir - sobre el precio ofertado en la eventual transacción.

El solicitante de la presente memoria a solicitado un modelo predictivo que obtenga mejores resultados - con el menor error de ajuste posible - al que ya utilizan en la empresa. A su vez, otros puntos a evaluar son: (1) el proceso de filtración previo de datos en el análisis, y (2) el procedimiento de evaluación métrica de las predicciones. Lo anterior producto de la incertidumbre que ésta posee en su proceso de análisis; junto con saber si sus recomendaciones empresariales son certeras, y cómo estas se pueden mejorar si es posible. Todo con el fin de brindar un mejor servicio a sus clientes mediante una predicción de precio de venta lo más cercana posible al precio real de transacción.
En el escenario del cliente se observa que el problema está en la necesidad de poseer una referencia que le permita maximizar su utilidad personal en el proceso de transacción. Más detalladamente, que el precio de venta[1] esté lo más cercano posible al precio de transacción. Disminuyendo la brecha, y permitiendo así una idea más clara entre proceder (o no) dentro de la operación.

El método de la empresa en cuestión está en recoger el elemento a predecir y filtrar por distancias en primera instancia. Se buscan al menos 10 vecinos que estén a máximo 50 metros de distancia con las mismas características tipo-operación más las características observables de: (1) dormitorios, (2) baños y (3) estacionamientos. Sino se encuentran inmuebles, se repite con 100 metros buscando 10 inmuebles como mínimo. Si falla, nuevamente se repite el proceso con 500 metros.
En segunda instancia se mantienen los criterios de características observables, pero se flexibilizan los estacionamientos; permitiendo que los vecinos tengan uno más o uno menos. Si fracasa, se admite cualquier cantidad de estacionamientos. En tercera instancia se repite la esencia de la segunda, solo que esta vez con los dormitorios. En cuarta instancia con los baños; y así sucesivamente.
Teóricamente este modelo además de entrar en sesgo de selección dentro de un rango muy acotado, también tiende al sub-ajuste de predicción por la escasa cantidad de vecinos analizados alrededor. En la presente memoria se evaluarán 8 sub-conjuntos formalizados que poseen una cantidad de entre 1.333 datos como cota mínima (inmuebles comerciales en venta) y 62.915 como cota máxima (departamentos en venta). Una evidente diferencia con la restricción de cota mínima impuesta por la empresa de al menos 10 inmuebles.
La empresa además de utilizar un único algoritmo - Generalized Boosted Regression - no alterna otros modelos en sus análisis que, si bien acota espacios de manera preliminar según sus propios criterios, no es justificable ni suficiente para utilizar un único modelo. Se proponen soluciones para corregir ello, en la sección 1.2.3. de Posibles Alternativas de Solución.
Las causas del problema están en la construcción y ejecución de la metodología en la empresa.

[1]Se hace hincapié en que esta es la palabra que comúnmente se usa en el mundo inmobiliario. Por tanto, debiera tenerse en cuenta también la operación "arriendo" como aquella contenida en ese término.

Más detalladamente en la etapa de evaluación, en donde regresionan la data para obtener los resultados.
Por tanto, la consecuencia del problema está en obtener resultados con alta probabilidad de sesgo y sub-ajustamiento. Destacando más aún el pobre poder de predicción que puede tener el procedimiento empresarial por no considerar aristas clave como: (1) métricas de error, (2) uso monótono de algoritmo y (3) parámetros óptimos. Se concluye, y se hace hincapié en tener en cuenta por parte de la empresa, que se debe tener un cambio en su metodología. Siendo la etapa de evaluación la más importante, por lejos.

Teniendo en consideración la parte inicial de la presente sub-sección, el área beneficiada - y directamente impactada - con los resultados de esta memoria, es el área de tecnología de la empresa. Obteniendo - eventualmente - una mejor metodología. Más robusta y más detallada. Conteniendo los elementos esenciales de todo análisis de Machine Learning en Data Science.

1.2.2. Identificación del Problema u Oportunidad y su Relevancia

La economía política impone supuestos sociales sobre la sociedad para el modelamiento de comportamientos individuales (y grupales) que ayuden al entendimiento de este conjunto de análisis. Una rama es la competencias imperfectas [21], que se definirá más adelante ex post la explicación de algunos conceptos clave para el entendimiento de esta. Dichos conceptos se presentan a continuación.
• **Mercado Competitivo [19]:** Aquel en el que existen muchos compradores y vendedores. Encargados de intercambiar productos "idénticos". Dada la interacción existente entre ambos participantes, es que estos se clasifican como precio-aceptante[2].

Esta competencia puede ser: monopólica u oligopólica; en donde los oferentes serían los "vendedores", y los clientes de la empresa los "compradores" dispuestos a tranzar bienes no diferenciables. Cuando se detalla la característica de "no diferenciable" se hace referencia a 2 (o más) inmuebles que comparten similares características observables dentro del mismo conjunto de análisis. Un ejemplo sería 2 departamentos en arriendo con igual cantidad de baños, dormitorios y metros construidos.
Teniendo en consideración que existen muchos vendedores (oferentes) en el mercado, transando bienes no diferenciables, es que se está en presencia de una competencia monopólica.

• **Competencia Monopólica [19]:** Estructura de mercado en la que muchas empresas venden productos similares pero no idénticos.

Finalmente para saber si es una competencia imperfecta, se debe verificar el no cumplimiento de la mayoría de los supuestos de mercados perfectos. Estos últimos son [30],

1. Información perfecta.

2. Todos los agentes económicos son racionales.

[2]Compradores y vendedores que aceptan el precio de mercado. Manteniendo este precio, los compradores pueden comprar todo lo que deseen; y los vendedores pueden vender todo lo que deseen

3. Hay tantos compradores y productores que ninguno tiene influencia sobre los precios.

4. Los bienes son homogéneos.

5. Los costos marginales se igualan a los ingresos marginales y los precios.

Por tanto este mercado posee un comportamiento dual. Es un mercado competitivo pues hay ofertas que ofrecen inmuebles no diferenciables dentro de la misma clasificación[3] tipo-operación en donde el cliente puede cotizar y elegir la mejor respuesta que maximice su propia utilidad. En paralelo se observa también una competencia monopólica, pues hay muchos oferentes que ofrecen distintos inmuebles. En síntesis resulta ser una combinación de ambos conceptos. Por un lado se observan los match entre vendedor-comprador en el mercado competitivo, y en el otro lado las características de los inmuebles que se transan en el mercado. En particular dentro de las clasificaciones de los inmuebles.

La oportunidad a abordar será la de maximizar las utilidades individuales del comprador y/o vendedor a la hora de decidir qué hacer con el inmueble en cuestión. Esto se ve en manifiesto de manera directa por parte del "comprador"; pero también puede ser aplicado al "vendedor" si es que este solicita un servicio a la empresa. Esto último para tener una mejor referencia del precio mercado, acercándose a él. Entre más lejos esté del precio mercado aumentando la brecha, más baja será la probabilidad de concretar una transacción con un posible interesado por una propiedad. Maximizando así, la utilidad individual del cliente de la empresa.
La empresa presta asesoría financiera – basada en predicciones – para saber cuál es la jugada óptima por parte de quienes estén interesados en: comprar, vender o invertir. De ello hay que tener en cuenta que los costos de vivienda tienen un impacto significativo en individuos, familias, empresas y gobiernos. Recientemente, compañías en línea como Zillow (en EE. UU.) han desarrollado sistemas de propiedades que proporcionan estimaciones automatizadas de los precios de la vivienda sin la necesidad de evaluadores profesionales. Sin embargo, la comprensión de lo que impulsa el valor de las casas es muy limitada [11]. Cualquiera sea el modelo, siempre habrá variables exógenas significativas que falten para poder predecir con exactitud y a la perfección. Las causas subyacentes pueden ser muchas, pero dentro de éstas está la variabilidad que producen los índices económicos como el IMACEC y el IPC que afectan de manera significativa la variabilidad de los precios inmobiliarios. Muestra causal de ello es observable en las Figura 1.2 del informe de CLAPES UC [1],

Las figuras 1.2 y 1.3 evidencian un crecimiento sistemático de los indicadores. Para la figura 1.2 estos indican el crecimiento en el índice de precios (eje Y) de los inmueble casa y departamento (eje X). En paralelo, para la figura 1.3 se muestra un crecimiento conjunto de múltiples indicadores; los cuales son (1) el índice de precios de departamentos y casas referenciados en la figura 1.2; y (2) el IMACEM (Índice Mensual de Actividad Económica) e IMACOM (Índice Mensual de Actividad de la Construcción). En el período de análisis de entre enero 2008 y septiembre 2017. Como elementos del eje X.
El eje Y se ubican los puntos asociados al índice en cuestión.

[3]Este criterio es una combinación de ambos. Operación si es arriendo o compra, y tipo si es un inmueble comercial-casa-departamento-oficina. En total 8 clasificaciones a analizar dentro del capítulo 4.

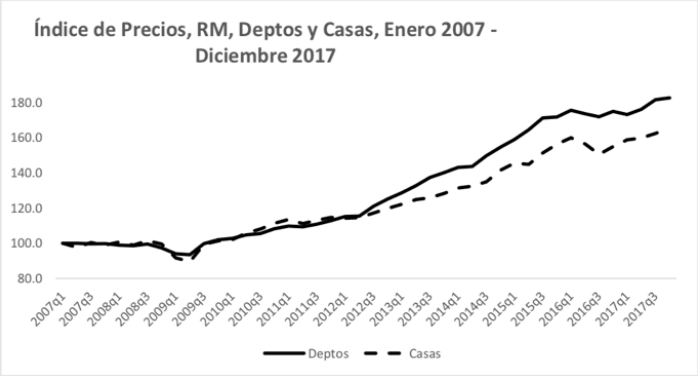

Figura 1.2: Índice de precios para la RM, departamentos y casas, enero 2007 - diciembre 2017 - Fuente: [1]

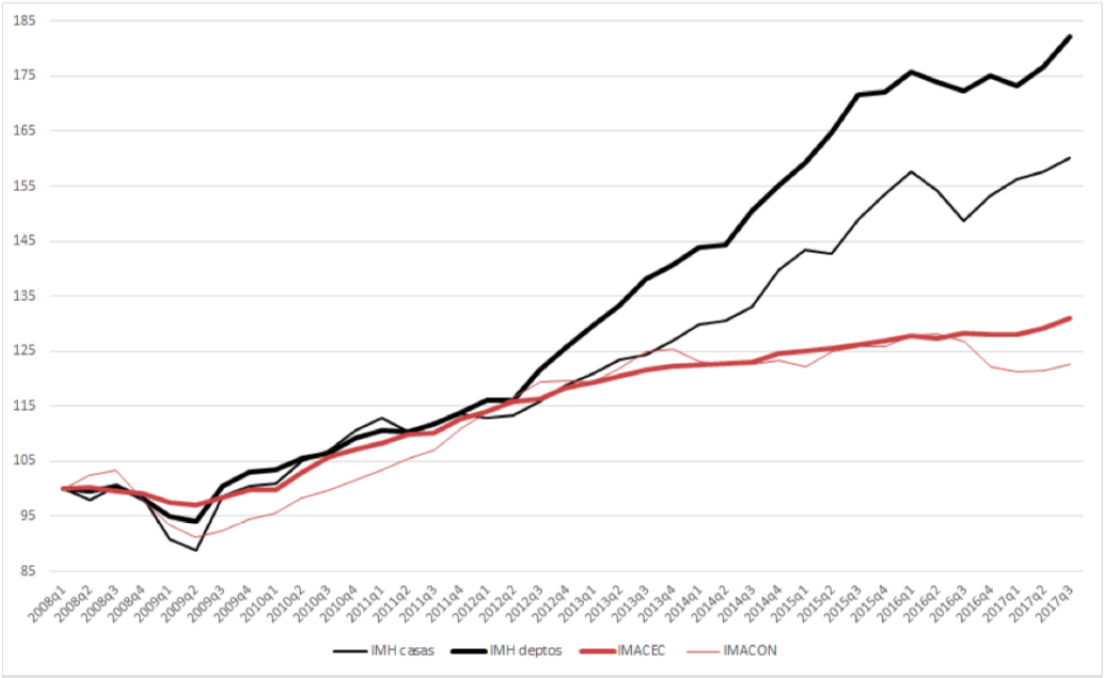

Figura 1.3: Comparación, para la RM, entre (i) el índice de precios del departamento y casas (líneas negras); y (ii) el IMACEM e IMACOM (líneas rojas), para el período enero 2008 - septiembre 2017 - Fuente: [1]

La evolución de las tendencias muestra - entre otras cosas - que las predicciones de precios de mercado para los departamentos y casas aumenta con el correr de los meses. En detalle, considerándolo en términos reales[4], las casas aumentaron su valor en un 65 %. Para el caso de los departamentos se tuvo un alza más considerable, del orden del 82 %. Todo esto ocurrido en la Región Metropolitana en el período anteriormente detallado.
Se observa además que los índices de precio siguen el mismo comportamiento que el IMACEC y el IMACON. Manifestando una correlación clara (positiva) entre los valores. No obstante ello, se debe ser cauto en tener presente que correlación no implica causalidad.

[4]Ajustado por inflación por estar en UF en primera instancia.

Considerando las alineaciones de tendencias, se postula la existencia de variables exógenas que afectan evidentemente el valor del precio del inmueble. Sin importar si existe doble causalidad en las variables, esto se puede solucionar con la incorporación - bien justificada - de una variable instrumental que también sea exógena. Demostrando de esta manera, que existen una (o más) variables significativas que influyan dentro de la fórmula a desarrollarse para la predicción. Pueden postularse, en este escenario, variables como: cercanía de parques, estaciones de metro; acceso a transporte público en general; cercanías a lugares de ocio como malls y cines; entre otros elementos exógenos.
Las variaciones de precio, por tanto, son causadas por diferentes factores a lo largo del tiempo. Creando de esta manera la necesidad, por parte de los interesados, en estar pendientes a cualquier cambio que vaya en su beneficio. Maximizando su utilidad. Sea para vender, comprar o invertir; la empresa ofrece al consumidor la posibilidad de tener una predicción que ofrezca un precio mercado aproximado del inmueble.

La oportunidad vista en el trabajo de la presente memoria está en mejorar el nivel de error arrojado como producto del uso del algoritmo de la empresa, Generalized Boosted Regression, en la predicción del precio de venta del inmuebles de interés; habiendo usado previamente la metodología de la empresa.

Esta metodología será modificada a una nueva que será discutida, y evaluada, entre los nuevos algoritmos postulados y el algoritmo de la empresa en cuestión. Más específicamente en el capítulo 4, subsección 4.1.5 de Quinta Etapa. Se postularán nuevas fórmulas y conjuntos de información que al usar, rescatarán - de manera más óptima - la información. Respecto a este último punto en particular, se reporta que la empresa utiliza las variables geográficas como filtros secuenciados con criterios similares para encontrar vecinos cercanos al elemento de interés; creando clusters. La empresa hace caso omiso de toda información geográfica - como variable independiente - en la fórmula de predicción. Esto tentativamente crearía sesgo de selección. Dado que todo se acota a vecindades cercanas dentro de un radio arbitrario que usa como centro, el elemento de interés. Nada certifica que la distancia elegida por la empresa sea la óptima. De hecho es posible que se estén omitiendo inmuebles relevantes a la hora de analizar el problema.
Por último se destaca que la empresa tampoco posee criterios de lo que es tener una *baja* o *alta* cantidad de datos. En la mayoría de los casos, dada la gran cantidad de filtros previos a la evaluación de regresión, terminan por utilizar una baja cantidad. Incurriendo en modelos sub-ajustados.
La presente memoria evita lo anterior con el uso de 8 sub-conjuntos de información segmentados en 5 partisiones iteradas entre ellas como conjuntos de entrenamiento y testeo, con ayuda del algoritmo K-fold cross-validation. Clasificando los inmuebles según sus características de tipo-operación. Esto se detallará más adelante.

En Machine Learning existe una vasta cantidad de modelos predictivos. En la presente memoria se usarán: (1) Linear Regressions, (2) Decision Tree, (3) Random Forest, (4) Support Vector Machine (SVM), (5) K-Nearest Neighbor, (6) Artificial Neural Networking y (7) Kernel Smoothing Regression. Todos ellos en su fase de regresión en vez de la de clasificación por la naturaleza del problema de predecir valores de precio.
La evaluación de los modelos estará criteriada por las siguientes métricas de error: R^2, MAPE y RMSE.

Dicho ello, el efecto final que tendrá la memoria está en el tratamiento y análisis de datos de la empresa. Los cuales podrán mejorar con la elección correcta del modelo predictivo según los resultados de la métrica de error, con un tratamiento de datos riguroso realizado previamente.

1.2.3. Posibles Alternativas de Solución

Las componentes clave que sirven para vislumbrar una oportunidad de mejora en la presente memoria son, entre otras, las siguientes:

1. La empresa no incorpora variables geográficas, posiblemente significativas. Estas resultan ser unas de las tantas variables categóricas signficativas; transformables a numéricas - vía variable binaria - que no están consideradas dentro de la regresión utilizada por la empresa. Ésta las utiliza en primera instancia, a modo de filtro y por separado. Usándolas como variables cluster, que sirven para crear vecindades entorno al elemento de interés a predecir.

2. La empresa no ajusta de manera adecuada sus modelos para realizar predicciones. Esto pues no optimiza los parámetros a utilizar dentro de la estimación, y solo se queda con los componentes por default dentro del comando del algoritmo a utilizar.

La causa en sí es relevante pues la empresa posee 192.179 datos de información con antecedentes de distintas compras/arriendos de distintas clases de inmuebles. Implicando en una gran cantidad de información que debe ser tratada de la mejor forma posible para rescatar la información más valiosa. Tanto para el cliente, como para la empresa misma en cuestión. Lo anterior obliga a crear un modelo predictivo que utilice de forma muy precisa las componentes esenciales para crear una predicción de los precios de venta solicitados por el cliente. Pudiendo, de esta manera, prestar una asesoría más fidedigna a la ya existente por la empresa. Es por ello que la causa es relevante. Para poder saber, con cierto grado de certeza, los precios de venta de las propiedades involucradas.

Analizando todos los puntos, en la etapa de evaluación se proponen las siguientes puntos a aplicar:

- **Evitar sesgo de selección y sub-ajuste:** Tal y como se detalló con la información de cotas en los sub-conjuntos de información, se crearon 8 sub-bases de la información original. Estos 8 elementos difieren en la caracterización de tipo y operación de cada inmueble. Al contar con una cantidad técnicamente aceptable (> 1333) se abre la posibilidad de analizar un conjunto técnicamente heterogéneo. Abriendo la posibilidad de calibrar de manera aceptable los algoritmos a utilizar dentro de cada análisis.
- **Cambiar ruta de análisis:** Siguiendo el hilo anterior, una forma de evitar el sesgo de selección de la ruta usada en el proceso de evaluación de la empresa - que radica en agrandar el radio desde el inmueble a predecir - se concreta yendo desde lo más general hasta lo más particular manteniendo los conjuntos de análisis. Es decir dadas las bases, se realizará en cada una de ellas una inspección que partirá con los resultados del promedio de las respectivas métricas de error; y finalizará con los mejores modelos

en cada una de las bases según los criterios de métricas de error.

- **K fold cross-validation:** La empresa realiza una única partisión en los conjuntos de análisis. Es decir, una sub-base de prueba, y otra sub-base de test. Esto puede entrar en conflicto pues al tener únicamente 2 bases, se corre el riesgo que el peso de los datos esté más cargada en una de ellas. Se postula como solución la realización de un K Cross-Validation que partisione la sub-base en otros K sub-conjuntos. Con estos últimos se tomará uno como testeo, y los demás como entrenamiento. Al ser K sub-conjuntos, se tendrán K iteraciones de este último paso; por tanto se deberá proceder con el promedio respectivo del *cross* en cada una de las métricas a utilizar.

- **Parámetro óptimo:** Para sorpresa de la investigación, la empresa no busca el parámetro óptimo dentro de un vector que contenga valores como posibles candidatos, lo que levanta nuevamente dudas respecto a la eficiencia del proceso utilizado. Para la presente memoria se postula utilizar, en cada uno de los modelos de regresión que necesiten parámetro, un vector de valores correspondientes a un parámetro único. En el caso de Support Vector Machine, dada la relevancia de todos sus parámetros, se procederá con 2 optimizaciones aplicando ceteris paribus.

- **Métricas de error:** En vista de que la empresa utiliza un único algoritmo, no se ve en la necesidad de evaluar el rendimiento del mismo con respecto otros. Para el uso de distintos algoritmos se es necesaria la incorporación de criterios métricos que permitan distinguir cuál es mejor que el otro; así como también, a las K partisiones creadas en el K fold cross-validation.
  Se postula el uso de las siguientes métricas: R^2, RMSE y MAPE. Observar que R^2 busca un máximo, mientras que RMSE y MAPE buscan un mínimo. La justificación de la incorporación de estas métricas es que una (o dos) certificarán a la otra. En particular R^2 buscando un máximo, debiera coincidir con el resultado de las otras métricas buscando un mínimo.

- **Evitar elección monótona de algoritmos:** Al disponer de distintas etapas y caracterizaciones de los elementos a utilizar para predecir; resulta extraño que se utilice un único algoritmo en todos los contextos. Estos no necesariamente son iguales en comportamiento. Es más, se puede conjeturar que no guardan relación alguna. Por tanto se presta para preguntarse ¿bastará la utilización de un solo algoritmo? ¿será lo más eficiente en todos los escenarios? Es por ello que se propone la utilización de los siguientes algoritmos: Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Kernel Smoothing Estimator, Support Vector Machine y Artificial Neural Networking.

- **Cuantificador de beneficio monetario de implementación:** Si el MAE perteneciente a algún modelo postulado por esta memoria fuese menor al obtenido por parte del modelo de la empresa - en la respectiva categoría de datos - el modelo postulado será mejor. Ergo, será seleccionado como el mejor modelo posible dentro de dicha categoría en cuestión. Trayendo como consecuencia, la sustitución del modelo de la empresa.
  En caso contrario, si el MAE obtenido por el modelo de la empresa resulta ser menor al del modelo postulado; se mantendrá el correspondiente a la empresa. Siendo este el mejor de la respectiva sub-base.
  Dentro de las 8 sub-bases de análisis, se tendrán eventualmente 8 diferencias positivas. Se designará por el símbolo Δ_+ a dicho factor, el cual siempre será positivo por defini-

ción[5]. Dicho esto, la fórmula para calcular la ganancia G semestral neta[6], obtenida con los resultados de la presente memoria, estará dada por

$$G = 1.000.000.000 \times 0,25 \times \left[0,2 + \sum_{I} \Delta_{+}\right]$$

[5]Si es negativo, se mantiene el modelo de la empresa. Sigue que no hay ganancia adicional por cambio de modelo.

[6]La empresa ganó aproximadamente $1.000.000.000 el último semestre.

1.2.4. Propuesta de Valor

En la actualidad las tasaciones son realizadas por un tasador de manera física. Este va un inmueble, y en base a ciertos criterios[7] basados en sus características observables, se valoriza según estas componentes. En la presente memoria se trabajará con datos pertenecientes a la página web www.portalinmobiliario.com en donde cualquier usuario con cuenta asociada puede ofertar un inmueble con ciertas características (véase las figuras A.2 y A.3 del Apéndice A). Si se observa, en la primera parte está la opción "Valor". Este monto hace referencia al precio del cual está dispuesto ofertar el usuario para el inmueble en cuestión.
Para el trabajo de datos, se dispone únicamente de dichas valorizaciones. No están capturadas las variaciones que esta sufre al momento de entrar en acuerdo con el demandante.

Considerando un set de oferentes, y asumiendo que cada uno de ellos oferta un solo inmueble, se tendrá registro de una cantidad igual de inmuebles con su respectivo *precio mercado*. En el presente trabajo se analizarán ofertas iniciadas en marzo del 2019, y eliminadas de la página web en mayo 2019 como mes máximo de eliminada la publicación del oferente. Por tanto se considerarán inmuebles con precios de mercado como conjunto de prueba, para predecir precios de mercado que se encuentren en la base de testeo. Las características de los inmuebles pertenecientes a este último conjunto serán evaluadas en base a inmuebles con una similar caracterización dentro del conjunto de entrenamiento, permitiendo predecir los precios de mercado de dichos inmuebles.
Finalmente se postula como propuesta de valor, el tener predicciones más precisas que aquellas que sean creadas por tasaciones físicas. Precios de mercado que evolucionan concorde a un mercado competitivo de inmuebles que, en caso de tener similares características, son reconocidos como bienes no diferenciables. Evitando valorizaciones subjetivas basadas en criterios estandarizados.

Por construcción, todo individuo racional busca maximizar su utilidad individual minimizando sus costos. En paralelo, dentro de cada adquisición inmobiliaria existe la incertidumbre sobre si *se estará (o no) pagando el precio correcto.* Con la predicción del precio mercado de inmueble el cliente podrá saber de antemano si es conveniente (o no) invertir, comprar o vender dadas las circunstacias.
Un oferente puede ofrecer un valor para ejecutar la transacción. Pero si el cliente sabe de antemano el precio de mercado, este podrá observar qué tan alejado está el precio de oferta del precio mercado. Abriendo el paso a 2 escenarios: (1) comprar barato (precio mercado $>$ precio de venta), para vender caro. Si es que lo desea vender. O bien, (2) arrendar a un precio razonable. Evitando volatilidades monetarias producto del sesgo de información entre el cliente y el oferente.

Finalmente en este punto, se evidencia otra componente de la propuesta de valor. La construcción de un modelo predictivo más preciso al ya usado por la empresa. Para acercarse más al precio de mercado existente, dando más certeza al cliente para que este pueda tener mejor información a la hora de decidir qué hacer en la transacción. Esto crea nuevamente dos escenarios:
• **Para el cliente:** Se podrá dar una mejor referencia del valor de mercado. Dándole una cota

[7]Como la tasación fiscal, metros construídos, tipo de vivienda, etc.

referencial que pueda usar para saber si es conveniente, o no, transaccionar en la operación.
• **Para la empresa:** El poseer precisiones más certeras da permiso a la empresa para posicionarse aún más como marca de predicción de precios de mercado. Aumentando la fidelización de los cotizantes nuevos y antiguos.

1.3. Pregunta de Investigación y Objetivos

1.3.1. Objetivo General

Minimizar los errores de ajuste en cada una de las predicciones de precios-venta pertenecientes a las distintas propiedades de interés dentro de las diferentes clasificaciones tipo-operación del mercado inmobiliario. Todo con el fin de aumentar las ganancias de la empresa.

1.3.2. Objetivos Específicos

1. Corrección de anomalías valóricas existentes en todas las variables existentes en la base de datos. Corrigiendo valores de signo, valores imposibles en el contexto de variable, entre otros.

2. Elegir, y utilizar, los atributos relevantes que pertenezcan a la base; que sirvan para la predicción de los elementos a analizar.

3. Construir un set de modelos predictivos que minimicen el error de ajuste en cada una de las bases a analizar en cada uno de ellos.

4. Se cuantificará monetariamente cuánto es lo ganado por la empresa ante una eventual mejora de los modelos propuestos en la presente memoria.

1.3.3. Resultados Esperados

En cuanto a los productos concretos que se obtendrán en el proyecto se pueden enunciar,

- **Código y método de análisis mejorados:** La idea final está en construir un código que haga correr todos los algoritmos anteriormente nombrados para hacer una evaluación de ellos. Con esto, el código tendrá un set de información donde se deberá elegir el menor (valor) error posible en cada uno de los distintos escenarios posibles. Cuando se habla de escenarios se hace referencia a los distintos contextos que existen dentro de la base (total) de información utilizada por la empresa. En este caso particular serán inmuebles del tipo: comercial, casa, departamento y oficina. Cada uno de ellos en operaciones de arriendo y venta.
- **Rendimiento óptimo y final:** Dada la elección de los algoritmos, y posterior predicción en cada uno de ellos; se mostrarán cuáles deberán ser los mejores algoritmos (con

sus parámetros óptimos) en cada una de las divisiones que se realicen a la base original (expuestas a finales de la explicación del punto anterior) según los criterios métricos de error elegidos (R^2, RMSE, MAPE, MAE).

1.3.4. Alcance del Proyecto

En lo correspondiente a alcances, se presentará a continuación una serie de limitaciones ocurridas dentro de la presente memoria. Estas serán de carácter técnico y teórico dentro de la creación y aplicación del producto realizado en cuestión.
La programación del código necesario para el cumplimiento de lo solicitado por la empresa; estará realizado en el lenguaje "R". Del cual se presentaron las siguientes limitaciones:

- **Información acotada de tiempo:** La información utilizada para predecir los valores dentro de la base de datos, tienen un horizonte temporal iniciado en marzo del 2019 y culminado en mayo de ese mismo año. Teniendo en consideración que son 3 meses de rango, se hace sospechosa la capacidad de utilizar los resultados de la presente memoria como indicadores de decisión por parte de la empresa. Esto pues no captura - en opinión del autor - los elementos necesarios como para discernir cuál modelo (y cuál no) se debe utilizar en un futuro cercano. Destacando que la presente investigación será publicada dentro del lapso mayo-junio del 2020.

Finalmente se reporta al lector que, al tener los códigos en lenguaje "R", la implementación le corresponderá únicamente a la empresa. Esto por el contexto de la misma. Ésta posee una API que recopila la información directamente de www.portalinmobiliario.com. Recibida la información comienzan a utilizar su algoritmo diseñado en Python para obtener resultados. Impidiendo en este paso, la compatibilidad e imposibilidad de implementación.

1.3.5. Metodología de Investigación

La metodología a seguir se basará en la metodología de CRISP-DM la cual se enfoca principalmente al análisis profundo del conocimiento y entendimiento del problema a resolver. Esta es una metodología actualmente muy utilizada en proyectos de aplicación de tecnología que calza con las necesidades del proyecto. Consiste principalmente en un conjunto de tareas que van de lo general a lo específico.
El ciclo de vida de un proyecto realizado con esta metodología consiste en seis etapas. La secuencia de etapas no es rígida, y algunas de ellas son bidireccionales, por lo que se permite revisar parcial o totalmente las etapas anteriores como se muestra en la Figura 1.4 [6]. Estas etapas están definidas como sigue,

1. **Business Understanding:** Entender los objetivos del proyecto y requerimientos desde una perspectiva empresarial. Posteriormente se convierte este conocimiento en un problema de data mining.

2. **Data Understanding:** Se inicia con una colección de data y procede con actividades

Figura 1.4: Etapas del modelo CRISP-DM - Fuente: [6]

en orden para familiarizarse con la data. Identificando problemas con la calidad de los datos, descubrir primeros insights o detectar subconjuntos que permitan realizar hipótesis sobre la información oculta.

3. **Data Preparation:** Etapa que cubre todas las actividades para construir el dataset final. Contiene la construcción de tablas, selección de atributos tanto como la transformación y limpieza de la data.

4. **Modeling:** Aquí varias técnicas de modelamiento son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos.

5. **Evaluation:** En esta etapa se debe construir un modelo (o modelos) que parezcan tener una alta calidad desde una perspectiva de data analysis.

6. **Deployment:** Dependiendo de los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan complejo como implementar un proceso de minería de datos repetible en toda la empresa. En muchos casos, es el cliente, no el analista de datos, quien realiza los pasos de implementación.

Para el cumplimiento de la primera etapa, *Business Understanding* - como ya se dijo - se debe comprender los objetivos del proyecto. Con las etapas de *Data Understanding* y *Data Preparation* se procederá como se muestra a continuación,

- Colectar los datos.
- Hacer una limpieza de estos, identificando problemas de calidad y su completitud (cantidad de campos de información vs cantidad de NULL en cada campo).
- Hacer un análisis de las componentes principales para encontrar factores que expliquen el tiempo de duración de los vendedores.
- Descubrir si existen subconjuntos interesantes para formar hipótesis en cuanto a la información oculta en los datos. Puede que algunos datos se muevan de forma conjunta, por ejemplo.

En la etapa de *Modeling* se seleccionarán y aplicarán las técnicas de modelado que sean

pertinentes al problema. Finalmente, para las etapas de evaluación y análisis de resultados, se revisará el proceso de desarrollo y se validará comparando con data de otros períodos. Los modelos pensados para el ejercicio de solución del problema planteado son,

- **Regresiones Lineales**
- **Árbol de Decisión**
- **Random Forest**
- **Support Vector Machine**
- **Kernel Density Estimators**
- **Redes Neuronales**

Sigue que las herramientas que evaluarán la calidad de las predicciones – dado un set de prueba para cada uno de los modelos – son: (1) R-Squared (R^2), (2) Mean Absolute Percentage Error (MAPE), (3) Root Mean Square Error (RMSE). Para el caso de R^2 es conveniente que el valor sea lo mayor posible; mientras entre menor sea el valor de los demás, más cerca está el modelo de ser el candidato correcto para la solucionar el problema. Los modelos quedan definidos de la siguiente manera:

- **R-Squared R^2:** Métrica encargada de calcular el porcentaje de la varianza explicada. Su fórmula está dada por,

$$R^2 = \frac{\left[\sum (y - \bar{y})(\hat{y} - \bar{y})\right]^2}{\sum \left[y - \bar{y}\right]^2 \sum \left[\hat{y} - \bar{y}\right]^2}$$

- **Mean Absolute Percentage Error (MAPE):** También conocida como Mean Absolute Percentage Deviation (MAPD) es una medida de la precisión de predicción en un método de pronóstico. Por lo general, expresa la precisión como un porcentaje, y se define por la fórmula,

$$MAPE \equiv \frac{100\,\%}{N} \sum_{\mathrm{i}=1}^{N} \left| \frac{y_{\mathrm{i}} - \hat{y}_{\mathrm{i}}}{y_{\mathrm{i}}} \right|$$

- **Root Mean Square Error (RMSE):** El RMSE de valores predichos $\hat{y}_{\mathrm{i}}$ para un total de N valores predichos con sus respectivos valores reales y_{i} generará un total de N datos dados por,

$$RMSE \equiv \sqrt{\frac{\sum_{\mathrm{i}=1}^{N} (\hat{y}_{\mathrm{i}} - y_{\mathrm{i}})^2)}{N}}$$

- **Mean Absolute Error (MAE):** Al igual que los componentes anteriores, sean valores ya predichos $\hat{y}_{\mathrm{i}}$ para un total de N valores predichos con sus respectivos valores y_{i}, se tendrá un MAE dado por,

$$MAE \equiv \frac{\sum_{\mathrm{i}=1}^{N} |\hat{y}_{\mathrm{i}} - y_{\mathrm{i}}|}{N}$$

El escenario de elección ideal en la comparación de 2 modelos, es cuando existe un solo ganador indiscutido. Es decir, en donde 2 ó 3 métricas son reportadas a favor de un contendiente. Pero ¿qué pasaría si el ganador no obtiene la aprobación de las 3 (ó 2) métricas; o hay empates técnicos? Es ahí donde se debe seleccionar cuál métrica es *más relevante* que la otra [14]. Dicho esto, el orden será: $R^2 > RMSE > MAPE$.

• **Primera Prioridad R^2:** Por definición matemática, esta métrica actúa libre de escalas. Es decir, sin importar lo grande que sea el valor, siempre se encontrará acotado entre $(-\infty, 1]$. El valor 1 indica que el modelo de predicción ajusta de forma perfecta al modelo planteado por los valores de testeo, mientras que - entre más negativo - peor es la calidad de la predicción. Este será la primera prioridad por ser libre de escalas. Facilitando la comprensión de la discriminación homogénea entre los modelos postulados.

• **Segunda Prioridad (MAPE):** Es comúnmente usada como función de pérdida de estimación en los problemas de regresión. Su definición permite una interpretación escalada de las predicciones resultantes, gracias a la normalización creada por el factor N. Si el resultado es $0\,\%$ es una predicción perfecta. Mientras que si es cercana a 1, la predicción no está muy bien realizada. Observar que, por definición, solo asegura valores positivos como resultado.

• **Tercera Prioridad (RMSE):** Es la raíz cuadrada del MSE. Al compararlos se identifica que ambos lo hacen de la misma forma en términos de minimización: $MSE(a) > MSE(b) \iff RMSE(a) > RMSE(b)$. Por tanto la única diferencia subyace en la velocidad de flujo en los gradientes de cada uno[8]

$$\frac{\partial}{\partial \hat{y}_{\mathrm{i}}} RMSE = \frac{1}{2\sqrt{MSE}} \frac{\partial MSE}{\partial \hat{y}_{\mathrm{i}}}$$

[8]Se usará RMSE por la incorporación de la raíz cuadrada. Aminorando el espacio de los cambios entre observaciones.

Capítulo 2

Marco Teórico

A continuación, se muestran los métodos a utilizar dentro de la presente memoria para poder segmentar - de principio - y estimar los precios-venta de los inmuebles. Las variables observables y a utilizar dentro de cada modelo serán seleccionadas conforme este posea parámetros, o no. Si no posee parámetros, tendrá un set de variables independientes distinta a aquellos que si lo tendrán.

2.1. Criterio: Uso de Método Freedman-Diaconis

A modo de anticipo, se debe tener en cuenta del cómo una "pendiente de una recta" puede ser considerada la "derivada de una función continua". En una recta: $y = mx + c$ se tiene que c es el valor intercepto que corta el eje Y y m la pendiente. La pendiente, en este caso particular, posee la siguiente formulación,

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

Los dos puntos del numerador y denominador son valores conocidos del conjunto real $\in \mathbb{R}$. Se crea una sola línea recta entre 2 puntos del espacio. Ahora, ¿qué pasa cuando la diferencia entre los elementos de los pares (y_1, y_2) y (x_1, x_2) disminuye a un valor infinitesimalmente pequeño $\Delta \to 0$? Se termina transformando en la derivada de una función continua. Dada por

$$\frac{\Delta y}{\Delta x} \xrightarrow[\Delta \to 0]{} \frac{\partial y}{\partial x}$$

Manifestando así, el principio de la suma de Riemann dentro de una función continua perfectamente diferenciable entre dos valores dentro de su dominio.
La idea central del uso de Freedman-Diaconis como método de aproximación radica justamente en la suma de Riemann, y como esta converge al área bajo la curva de una función continua y diferenciable. Usando el Teorema Fundamental del Cálculo parte 1 [10]

Teorema 2.1 *Si f es una función continua en un intervalo $I \subseteq \mathbb{R}$ y $a \in I$, entonces la función G definida por:*

$$G(x) = \int_a^x f$$

es derivable en $\text{int}(I)$ *y además* $G' = f$ *en* $\text{int}(I)$.

Implicando que para toda función real-continua y diferenciable, habrá una primitiva que la defina.
Esto es de suma ventaja pues establece de base que, con ayuda de intervalos de longitud infinitesimal a través de la suma de Riemann, se puede converger al área bajo la curva de la función. Una muestra gráfica de la convergencia de la suma de Riemann se presenta con el siguiente set de imágenes,

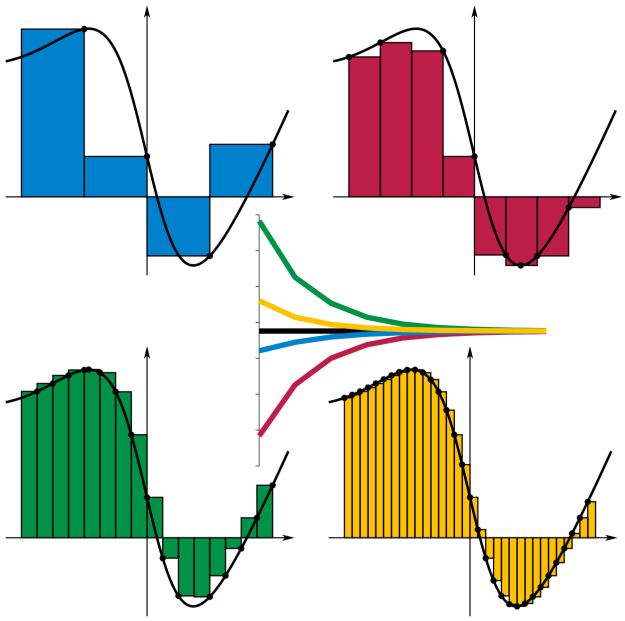

Figura 2.1: Suma de Riemann a Área Bajo la Curva - Fuente: www.wikipedia.org

Finalmente para un set de intervalos con longitud infinitesimalmente, se podrá dibujar el área bajo la curva con dicho conjunto.
Ahora el problema radica en la cantidad *óptima* de intervalos que sirvan para dibujar la curva. Esto pues, en principio, como se tiene un conjunto de datos conocidos - y utilizados para predecir - no se podrá tener una curva empírica de la cual se pueda obtener una función para trazar la suma de Riemann. Es ahí donde entra el método de Freedman-Diaconis, diseñado para minimizar la diferencia entre el área bajo la distribución de probabilidad empírica y el área bajo la distribución de probabilidad teórica. Teniendo así las condiciones expuestas con anterioridad.

Freedman-Diaconis va más allá de la suma de Riemann pues arregla las falla de este, en donde sus cálculos quedan anulados. Tal cual esto se explica en el Corolario (ecuación 2.24) dentro del ensayo de Freedman-Diaconis [17], en donde confirma la existencia de la integral para funciones suaves tipo L_1.
Un ejemplo demostrativo de ello sería pensar en un triángulo (en $\mathbb{R}^2$), pirámide (en $\mathbb{R}^3$),

etc-expandible a dimensiones finitas; en donde la derivada, y por consecuencia la integral, se indefinen en la intersección no continua de los hiperplanos. O coloquialmente dicho, la "punta" del polígono.
Se podría pensar el triángulo como una suma de dos rectas con pendientes de sentido distinto. En ese caso, la suma de Riemann es factible con la suma de ambas funciones; pero nuevamente esto se cae en un punto no diferenciable. Una singularidad.
De ahí la gracia de Freedman-Diaconis, pues no tiene problemas en estos casos. Siendo útil para la creación de una distribución empírica.

El ancho óptimo del intervalo está dado por,

$$\Delta = 2\frac{IQR(x)}{\sqrt[3]{N}}$$

El $IQR(x)$, interquartile range, se define como la diferencia entre los percentiles 75 y 25, o entre los cuartiles superiores y los inferiores $IQR(x) = Q_3 - Q_1$. En otras palabras, el $IQR(x)$ es el primer cuartil sustraído del tercer cuartil; esto se puede observar en la figura 2.2.

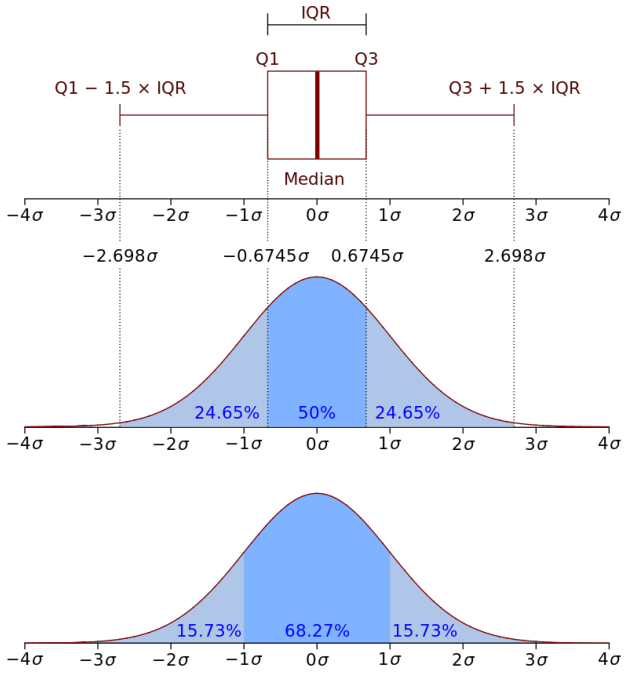

Figura 2.2: Interquartile Range - Fuente: www.wikipedia.org

Esto permite ajustar la cantidad óptima de intervalos que definirán la forma de la curva formada sin necesidad de entrar en una integración. Mas, esto sirve para datos empíricos, como la data observable, que ayudan a la estimación vía entrenamiento y calibración.
Hay una semejanza, en la presente memoria, con la implementación del método Freedman-Diaconis y el algoritmo Kernel Smoothing Regression. Esto radica pues el algoritmo, al igual que el método, busca optimizar *bandwidth* para la predicción al momento de ajustarse. Como es un Kernel Density Estimator, difiere en la calidad/rapidez de estimación según la definición del kernel a utilizar. Una aproximación de lo realizado por el algoritmo Kernel Smoothing Regression se observa en la figura 2.3.

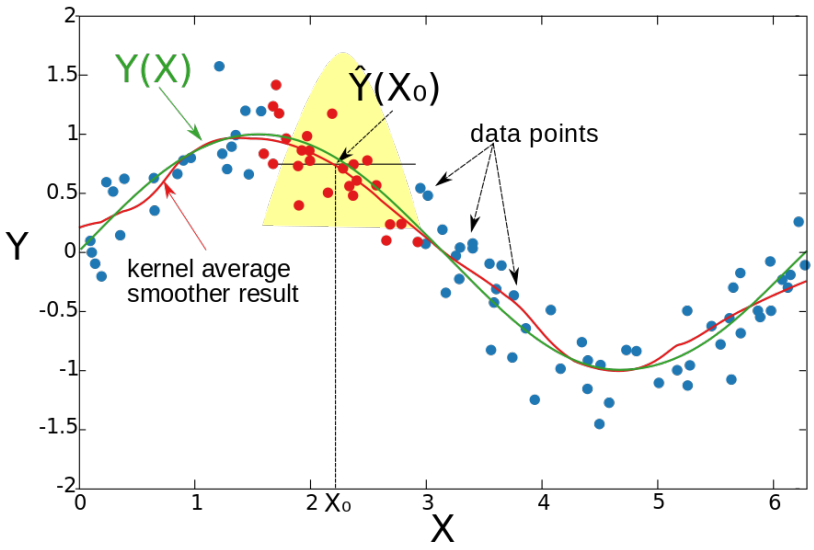

Figura 2.3: Funcionamiento Kernel Smoothing Regression - Fuente: www.wikipedia.org

2.2. Modelamiento

2.2.1. Modelo de la Empresa = Generalized Boosted Regression

El problema que resuelve este algoritmo está en, dado un conjunto de prueba $\{y_{\mathrm{i}}, X_{\mathrm{i}}\}_1^N$ con (Y, X)-valores conocidos, se debe buscar una función de aproximación (estimación) $\hat{F}(X)$, de la función $F^*(X)$ (función real) que mapea X hacia Y, minimizando el valor esperado de alguna función de pérdida especificada $\mathfrak{L}(y, F(X))$ sobre distribución conjunta de todos los (Y, X)-valores En términos comparativos, Mínimos Cuadrados Ordinarios posee las mismas componentes para la resolución del problema: $\vec{y}$: función real, $\hat{y}$: función estimación, $\hat{\varepsilon}$: componente de función de pérdida (SSE: $\sum_{\mathrm{i}\in I} \vec{\varepsilon}^2$) a optimizar (minimizar) sobre el conjunto de entrenamiento compuesto de (Y, X)-valores.

Un procedimiento común es restringir $F(X)$ para ser un miembro de una clase parametrizada de funciones $F(X; P)$ en donde $P = \{P_1, P_2, ...\}$ es un set finito de parámetros cuyos valores conjuntos identifican a los miembros individuales de la clase. Este algoritmo se enfoca en algoritmos de la forma,

$$F(X; \{\beta_m, a_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(X; a_m)$$

Con una función genérica $h(X; a)$, la cual suele ser una función parametrizada simple de variables de entrada X, caracterizada por parámetros $a = \{a_1, a_2, ...\}$. Los términos individuales difieren en los valores a_m elegidos para estos parámetros. Para este escenario particular, aquí resultan de interés las funciones $h(X; a_m)$ de un pequeño árbol de regresión. Como los producidos por CART. Para un árbol de regresión, los parámetros a_m son las variables de división; las ubicaciones divididas, y los medios terminales del nodo de los árboles individuales [18].

2.2.2. Regresión Lineal Multivariada

Modelo encargado de minimizar la suma de errores cuadráticos considerando como base, los supuestos de Gauss-Markov. Estos son,

- **Linealidad:** Dada una función $Y = f(\beta X)$ lineal, esta lo es en base al parámetro β. En particular, $Y_{\mathrm{i}} = \beta_0 + \sum_{j=1}^{N} \beta_j X_{\mathrm{i}j} + \varepsilon_{\mathrm{i}}$ $\forall \mathrm{i} \in I$. En donde i hace referencia a la observación, y j hace referencia a la característica.
- **Muestreo aleatorio $\{Y_{\mathrm{i}}, X_{\mathrm{i}j}\}$:** Indica que la característica observable de la observación Y_{i} (es decir $X_{\mathrm{i}j}$) es independiente de otra característica observable de la misma observación.
- **Esperanza de los factores no observables igual 0 $E(\varepsilon_{\mathrm{i}}) = 0$:** En promedio, para cada observación, el efecto de los factores no observables ε_{i} $\forall \mathrm{i} \in I$ no afecta en el modelo.
- **Correcta identificación:** La matriz X posee rango completo. Es decir, la inversa de la matriz existe. Por tanto, cada columna de la matriz aporta información relevante.

Mientras que cada fila es linealmente independiente una de otra. De modo que existe $[X'X]^{-1}$.

- **Homocedasticidad** $V(\varepsilon_{\mathrm{i}}) = \sigma^2$**:** La varianza de los errores es la misma para todos los factores no observables, es decir, no hay relación entre un factor y otro dentro de la suposición.

Teniendo en consideración los supuestos anteriores, la ecuación matricial lineal de MCO está dada por,

$$\vec{y} = X\beta + \vec{\varepsilon} \Longleftrightarrow y_{\mathrm{i}} = \sum_{j=1}^{J} \beta_j X_{\mathrm{i}j} + \varepsilon_{\mathrm{i}} \quad \forall \mathrm{i} \in I \equiv \{1, ..., N\}$$

Tomando en cuenta la forma matricial del problema $y = X\beta$ basta desarrollar como sigue,

$$y = X\beta \underbrace{\Longrightarrow}_{X'\cdot} X'y = X'X\beta \underbrace{\Longrightarrow}_{[X'X]^{-1}\cdot} [X'X]^{-1}X'y = [X'X]^{-1}X'X\beta \underbrace{\Longrightarrow}_{\exists \mathbb{I}} [X'X]^{-1}X'y = \hat{\beta}$$

De lo anterior basta tomar $\hat{\beta}$, pues $\hat{y} = X\hat{\beta}$. Quedando definida su respectiva predicción.

El objetivo de este modelo está en comparar la efectividad de los modelos de Machine Learning versus el modelo perteneciente a la empresa (que utiliza regresiones lineales), como primera evaluación.

2.2.3. Árboles de Decisión

Los árboles binarios o árboles de decisión son un tipo de algoritmo de aprendizaje supervisado en la cual existe una variable objetivo predefinida. Este tipo de algoritmos divide el espacio de los predictores (las variables independientes) en regiones distintas y no sobrepuestas [24]. En la Figura 2.4 se puede ver un ejemplo de árbol binario el cual posee 6 regiones separadas.

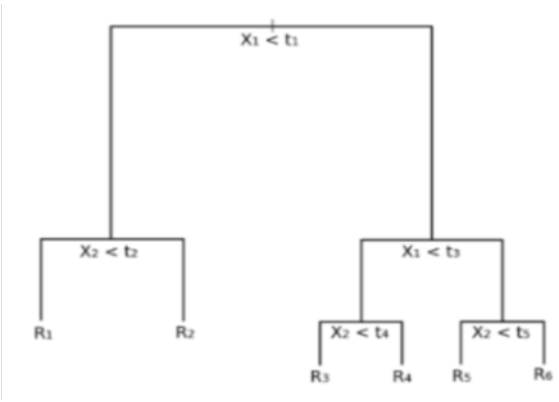

Figura 2.4: Ejemplo de árbol binario con 6 regiones separadas - Fuente: [24]

Este problema, al igual que la regresión lineal, busca reducir la suma de error cuadrático. Es decir mín $\sum_{j=1}^{J} \sum_{\mathrm{i} \in R_{\mathrm{i}}} (Y_{\mathrm{i}} - \hat{Y}_{R_j})^2$ en donde para cada región j y observación i se minimiza el

SSE. Dentro de los árboles de decisión existen 2 tipos: (1) árboles de clasificación y (2) árboles de regresión. La primera esta enfocada en realizar una salida (outcome) que termina por ser una clase perteneciente a la base datos. Por ejemplo, la clasificación de una deuda como riesgo bancario. Finalmente la segunda alternativa, en donde la salida puede ser considerada una variable con números reales. En donde un ejemplo aplicado podría ser la predicción de una población.

Matemáticamente el presente informe se enfocará en un marco basado en un árbol particular llamado árboles de clasificación y regresión, o CART [4]. En la Figura 2.5 se muestra una partición binaria recursiva del conjunto de entrada, junto con la estructura de árbol correspondiente. En este ejemplo, el primer paso divide todo el espacio de entrada en dos regiones según si $x_1 \leq \theta_1$ o $x_1 > \theta_1$ donde θ_1 es un parámetro - o características observable significativa-numérica que sirve de criterio - del modelo. Esto crea dos subregiones, cada una de las cuales se puede subdividir de forma independiente.

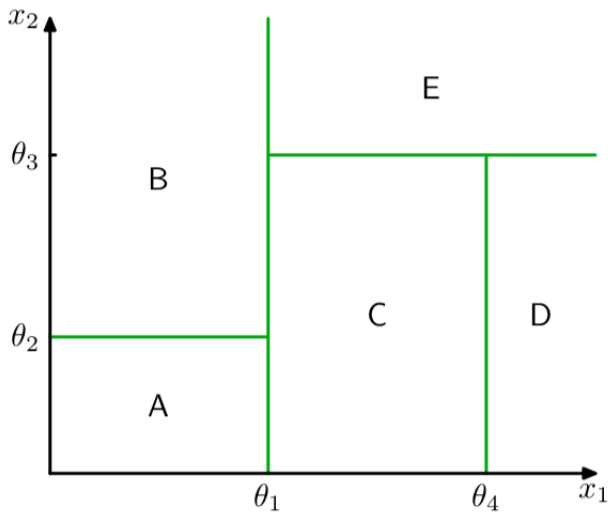

Figura 2.5: Ilustración de un input en 2 dimensiones - Fuente: [2]

Por ejemplo la región $x_1 \leq \theta_1$ se subdivide según $x_2 \leq \theta_2$ o $x_2 > \theta_2$ se da lugar a las regiones designadas como A y B. Dentro de cada región, hay un modelo separado para predecir la variable objetivo. Por ejemplo, en la regresión podríamos simplemente predecir una constante sobre cada región, o en la clasificación podríamos asignar cada región a una clase específica.

Árboles de Regresión

Considérese primero un problema de regresión en el que el objetivo es predecir una sola variable objetivo t a partir de un vector D-dimensional $\mathrm{d} = (x_1, ..., x_D)^T$ de valores de entrada (inputs). La data de entrenamiento consiste en vectores de entrada $\{\mathrm{d}_1, \mathrm{d}_2, ..., \mathrm{d}_N\}$ junto con las correspondientes etiquetas $\{t_1, ..., t_N\}$. En una partición del espacio de entrada se minimiza la función SSE, entonces el valor óptimo de la variable predictiva dentro de cualquier región dada viene dado por el promedio de los valores de t_n para esos puntos de datos que cayeron en esa región. Ahora considérese cómo determinar la estructura de un árbol. Incluso, para un número fijo de nodos en el árbol, el problema de determinar la estructura óptima (incluida

la elección de la variables de entradas para cada división, así como también los umbrales correspondientes) para minimizar el SSE generalmente no es computacionalmente factible, debido a la gran cantidad de posibles soluciones. En cambio, una optimización codiciosa generalmente se puede realizar comenzando con un solo nodo raíz (correspondiente al espacio de entrada completo) y luego haciendo crecer el árbol agregando de a un nodo a la vez. En cada paso habrá un cierto número de regiones candidatas dentro del espacio de entrada que se pueden dividir, lo que corresponde a la adición de un par de nodos hoja (no raíz) al árbol existente. La optimización conjunta en la elección de la región a dividir, la elección de la variable de entrada y el umbral, se puede hacer de manera eficiente mediante una búsqueda exhaustiva que tenga en cuenta que, para una elección dada de la variable dividida y el umbral, la elección óptima de la variable predictiva viene dada por promedio local de los datos, como se señaló anteriormente. Esto se repite para todas las opciones posibles de variable a dividir, y se conserva la que da el menor valor SSE. Dada la estrategia de hacer crecer el árbol, nace la duda de cuándo dejar de agregar nodos. Un enfoque simple sería detenerse cuando la reducción en el error residual cae por debajo de algún umbral. Sin embargo, se encuentra que ninguna de las divisiones disponibles produce una reducción significativa en el error y, sin embargo, después de varias divisiones más, se encuentra una reducción sustancial de errores. Por esta razón, es una práctica común hacer crecer un árbol grande, utilizando un criterio de detención basado en el número de puntos de datos asociados con los nodos de hoja, y luego podar el árbol resultante. La poda se basa en un criterio que equilibra el error residual con una medida de la complejidad del modelo. Si denotamos el árbol de inicio para la poda por T_0 donde se define $T \subset T_0$ coo un sub-árbol de T_0 el cual se puede obtener podando nodos de T_0. Supóngase que los nodos hoja están idexados por $\tau = 1, ..., |T|$, con el nodo hoja τ representando a la region $\mathcal{R}_\tau$ de un espacio de entrada que posee N_τ "data-points" y $|T|$ denotando el total de nodos hoja. La optimización ideal para la region $\mathcal{R}_\tau$ está dada por,

$$y_\tau = \frac{1}{N_\tau} \sum_{x_n \in \mathcal{R}_\tau} t_n$$

Y la correspondiente contribución de la suma cuadrada de residuos está dada por,

$$Q_\tau(T) = \frac{1}{N_\tau} \sum_{x_n \in \mathcal{R}_\tau} \{t_n - y_\tau\}^2$$

El criterio de poda viene dado por,

$$C(T) = \sum_{\tau=1}^{|T|} Q_\tau(T) + \lambda |T|$$

La regularización del parámetro λ determina el trade-off entre el SSE y la complejidad del modelo medido por el número $|T|$ de nodos hoja, y donde este valor es decidido por el método de cross-validation [2].

2.2.4. Random Forest

Así como todos los modelos descritos anteriormente, un árbol binario también tiene problemas de sesgo y de varianza. Esto se intenta disminuir con la metodología de Random

Forest en la cual se utilizan varios árboles de decisión tratados con la técnica de bagging para reducir la varianza de las predicciones. Esta técnica lo que hace es generar subconjuntos dentro del set de entrenamiento para que la correlación de las variables, si es que existe, no afecte en los resultados, reduciendo la varianza. En la Figura 2.6 se ilustra el proceso bagging,

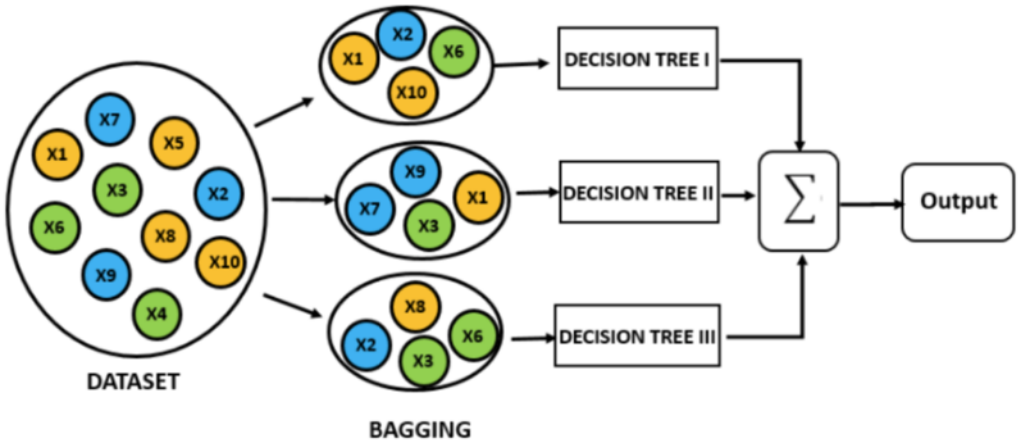

Figura 2.6: Ilustración del proceso bagging - Fuente: [24]

Random Forest para Regresiones

Una forma de reducir la varianza del modelo es construir una gran cantidad de sub-modelos independientes, con sus propios sub-conjuntos de información, para después sacar el promedio. Como solo se dispone de un porcentaje fijo de datos de entrenamiento, no se puede pensar en tener más datos para analizar más, y de mejor forma. Para reparar ello - dentro de las regresiones - Random Forest aplica una técnica llamada *bagging*. Técnica consistente en crear B sub-conjuntos de entrenamiento dentro de, el ya existente, conjunto de entrenamiento. Para que, en cada sub-muestra b, se entrene su respectivo árbol de decisión. Obteniendo un resultado $\hat{f}^{\star b}(x)$. Con todo calculado se procede a calcular el promedio de los B árboles de decisión calculados. Esto queda representado a través de la siguiente fórmula,

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{\star b}(x)$$

Se señala que los requisitos para Random Forests son: (1) la baja correlación entre residuos, y (2) árboles de bajo error. Random Forests disminuye el error promedio de los árboles empleados por el factor $\bar{\rho}$. La aleatoriedad empleada debe apuntar a una baja correlación de residuos [3].

2.2.5. Support Vector Machine

Una máquina de vectores de soporte, SVM por su acrónimo en inglés, es un clasificador definido por un hiperplano separador. En otras palabras, dados los datos de entrenamiento, el algoritmo genera un hiperplano óptimo que categoriza nuevos ejemplos. Se entenderá por hiperplano al plano generado en N dimensiones. En un espacio unidimensional como una recta, un hiperplano es un punto (divide una línea en dos líneas). En un espacio bidimensional como el plano XY, un hiperplano es una recta (divide el plano en dos mitades). En un espacio tridimensional, un hiperplano es un plano corriente (divide el espacio en dos mitades). Este modelo posee parámetros para su ajusta, dentro de los cuales se pueden nombrar: (1) Kernel, (2) Regularización, (3) Gamma y (4) Margen; entre tantos otros [25].

Definición 2.2 ***(Algunos) parámetros de Support Vector Machine****:*
• ***Kernel:*** *Parámetro encargado de definir la separación producto del hiperplano separador de donde se genera el SVM. Estas separaciones pueden ser lineales o no lineales. Caracterización creada como consecuencia de la elección del respectivo kernel. Estos pueden ser cuatro: (1) lineal, (2) radial, (3) polinomial y (4) sigmoidal.*
• ***Regularización:*** *Parámetro que define qué tan estricta debe ser la evasión del SVM sobre los errores de clasificación de los datos de entrenamiento. Un ejemplo de esto se observa en las Figuras 2.7 y 2.8 donde se aprecian dos modelos, uno con alta regularización y otro con baja.*
• ***Gamma:*** *Parámetro que define el entorno de acción dentro de un conjunto de entrenamiento. Un alto gamma es aquel que considera solo los puntos cercanos, y un gamma de bajo valor es aquel que considera - además - los puntos lejanos. Implicando la consideración de outliers dentro del conjunto.*
• ***Margen:*** *Parámetro que define la separación - a través de un hiperplano - entre los subconjuntos de datos. Si el margen es óptimo, entonces la separación será equidistante. En caso contrario, será tendenciado a una de las tantas opciones. Esto se puede en las Figuras 2.9 y 2.10.*

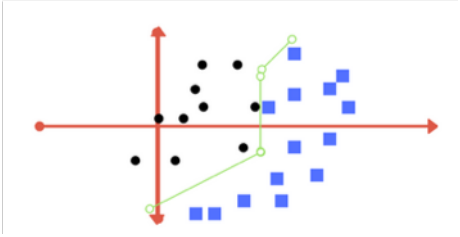

Figura 2.7: Modelo con alto nivel de regularización - Fuente: [25]

Cabe señalar que al ser componentes que dan la posibilidad de una separación lineal y no lineal, necesariamente habrá una disposición no-lineal de datos que no debe ser desarrollada con una configuración lineal. Esto solo se decide exante la ejecución del algoritmo. Por tanto, se debe tener precaución con la elección incorrecta de kernel dentro de cualquier análisis.

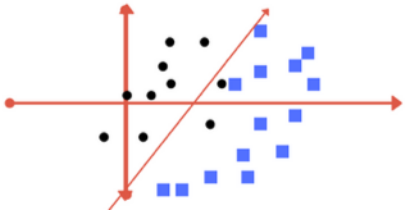

Figura 2.8: Modelo con bajo nivel de regularización - Fuente: [25]

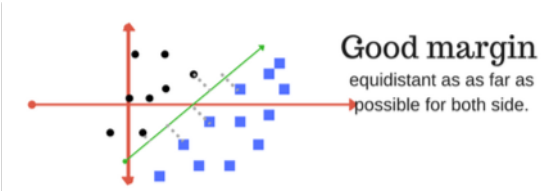

Figura 2.9: Modelo con buen margen - Fuente: [25]

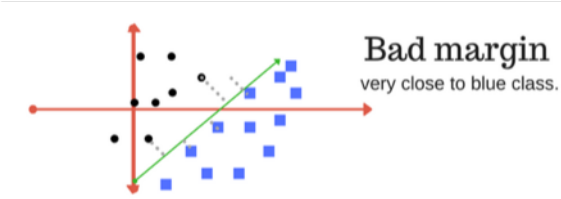

Figura 2.10: Modelo con mal margen - Fuente: [25]

Support Vector Machine para Regresiones

Ahora se extenderá la concepción de SVM al problema de regresión lineal. En una regresión lineal simple, se minimiza una función regularizada dada por,

$$\frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2 + \frac{\lambda}{2}\|w\|^2$$

Para obtener soluciones dispersas, la función de error cuadrático se reemplaza por una función de error ε-insensible [29] que da error cero si la diferencia absoluta entre la predicción $y(x)$ y el objetivo t es menor que ε, donde $\varepsilon > 0$. Un ejemplo simple de una función de error ε-insensible es la propuesta por Christopher M. Bishop [2] la cual posee un costo lineal asociado a los errores fuera de la región insensible. El ejemplo está dado por,

$$\mathbb{E}_\varepsilon(y(x) - t) = \begin{cases} 0, & \text{si } |y(x) - t| < \varepsilon; \\ |y(x) - t| - \varepsilon & \sim \end{cases}$$

Y es ilustrado en la siguiente Figura,
Por tanto hay que minimizar la función de error regularizado dada por,

$$C\sum_{n=1}^{N}\mathbb{E}_\varepsilon(y(x_n) - t_n) + \frac{1}{2}\|w\|^2$$

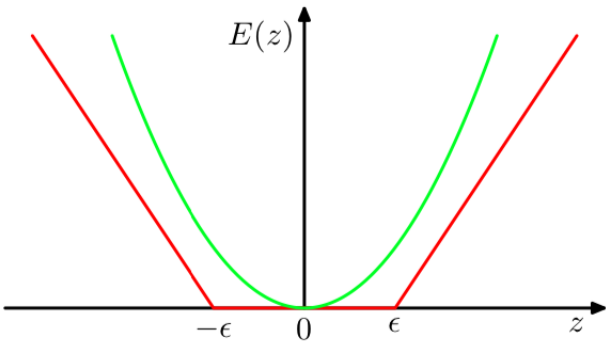

Figura 2.11: Función de error ε-insensible - Fuente: [2]

Para cada "data-point" x_n se necesitan 2 variables de holgura $\xi_n \geq 0$ y $\hat{\xi}_n \geq 0$, en donde $\xi_n > 0$ corresponde a un punto en donde $t_n > y(x_n) + \varepsilon$, y $\hat{\xi}_n > 0$ corresponde a un punto en donde $t_n < y(x_n) - \varepsilon$ como se ilustra en la siguiente Figura,

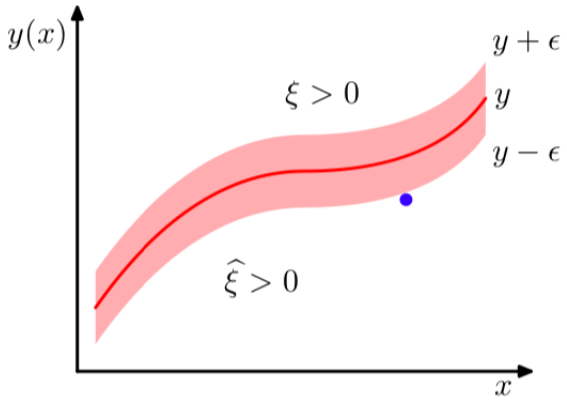

Figura 2.12: Ilustración de regresión SVM - Fuente: [2]

En la Figura se muestra la curva de regresión junto con el "tubo" insensible. También se muestran ejemplos de las variables de holgura ξ y $\hat{\xi}$. Los puntos por encima del tubo ε-insensible tienen $\xi > 0$ y $\hat{\xi} = 0$. Puntos por debajo del tubo ε-insensible tienen $\xi = 0$ y $\hat{\xi} > 0$. Finalmente los puntos adentro del tubo ε-insensible tienen $\xi = \hat{\xi} = 0$.
Expuesto lo anterior, la función error regularizada para SVM de regresiones puede ser escrita como,

$$C \sum_{n=1}^{N} (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|w\|^2$$

El cual debe ser minimizado bajo las siguientes restricciones: $\xi_n \geq 0$ y $\hat{\xi}_n \geq 0$ como ocurrió en las 2 ecuaciones anteriores.

De este modo, se demuestra que hay como máximo νN "data-points" que caen fuera del tubo insensible, mientras que hay al menos νN "data-points" que son "support vectors" y, por lo tanto, se encuentran en el tubo o fuera de él. El uso de SVM para resolver un problema de regresión se ilustra en el siguiente ejemplo usando un conjunto de datos sinusoidales en la Figura 2.13. Aquí los parámetros ν y C se han elegido a mano. En la práctica, sus valores

normalmente se determinarían mediante cross-validation [2].

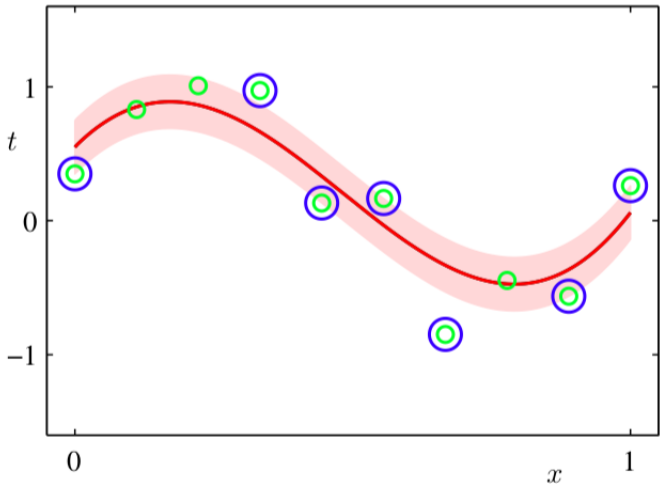

Figura 2.13: Ilustración de ν-SVM de regresión - Fuente: [2]

La Figura 2.13 presenta una regresión aplicada a datos sinusoidales artificiales utilizando kernels gaussianos. La curva de regresión pronosticada se muestra mediante la línea roja, y el tubo insensible corresponde a la región sombreada. Además, los puntos de datos se muestran en verde, y aquellos con vectores de soporte se indican con círculos azules.

2.2.6. K-Nearest Neighbor

Algoritmo que se encarga de clasificar, en base al parámetro k, al (a los) vecino (s) más cercanos según su posición. El parámetro k hace referencia a la cantidad de vecinos cercanos que se busca clasificar o regresionar.
Se supondrá un plano compuesto por los siguientes elementos,

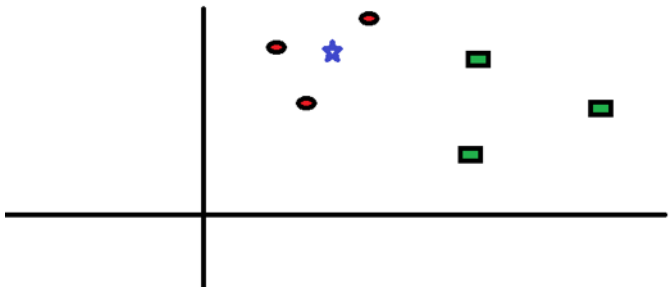

Figura 2.14: Gráfico de caso simple KNN - Fuente: [28]

En donde se llamará a los círculo rojos como (CR) y a los cuadrados verdes como (CV). Se destaca que estamos en búsqueda de la clasificación de la estrella azul para saber a qué conjunto pertenece (CR o CV). Como queremos clasificar a la estrella azul, por el esquema, podemos calcular con $K = 3$ a qué clasificación pertenece. Dicho lo anterior, ahora la estrella azul es el centro del cálculo para descubrir los vecinos más cercanos. El radio del círculo

formado, con la estrella como centro, posee un largo lo suficientemente extenso para abarcar 3 vecinos lo más cercano posibles.

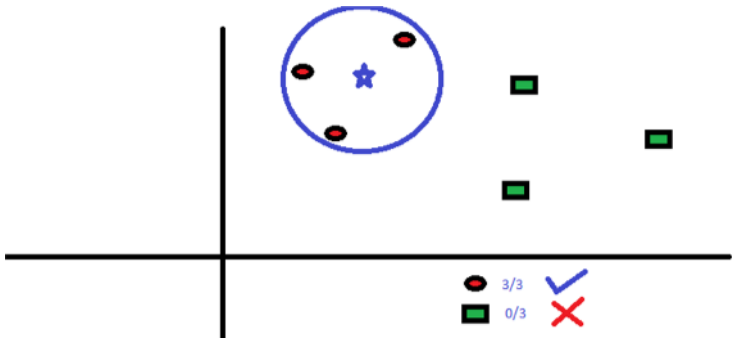

Figura 2.15: Gráfico de clasificación estrella azul - Fuente: [28]

Se observa que los vecinos más cercanos son de clase CR. De modo que la estrella azul – que antes no tenía clasificación - ahora es CR. Es decir, círculo rojo. En esta memoria se usará KNN de regresión. Para predecir números en base a la distancia de los vecinos más cercanos. Una implementación simple de la regresión KNN es calcular el promedio del objetivo numérico de los K vecinos más cercanos. Otro enfoque utiliza un promedio ponderado de la distancia inversa de los K vecinos más cercanos. La regresión KNN utiliza las mismas funciones de distancia que la clasificación KNN [27].

Dichas funciones de distancia - entre otras tantas - son,

- **Distancia Euclidiana:** $\sqrt{\sum_{\mathrm{i}=1}^{k}(x_{\mathrm{i}}-y_{\mathrm{i}})^2}$
- **Distancia Manhattan:** $\sum_{\mathrm{i}=1}^{k}|x_{\mathrm{i}}-y_{\mathrm{i}}|$
- **Distancia Minkowski:** $\left(\sum_{\mathrm{i}=1}^{k}(|x_{\mathrm{i}}-y_{\mathrm{i}}|)^q\right)^{1/q}$

Las tres medidas de distancia anteriores solo son válidas para variables continuas. En el caso de las variables categóricas, debe utilizar la distancia de Hamming, que es una medida del número de casos en que los símbolos correspondientes son diferentes en dos cadenas de igual longitud [27]. Matemáticamente esto es un estimador de densidad de kernel lo cual se define de la siguiente forma,

Kernel Density Estimators

Primero se debe comenzar observando los estimadores de densidad de kernel. Se supondrá que las observaciones son dibujadas desde una probabilidad de densidad desconocida[1] $p(x)$ en algún espacio D-dimensional. Se desea estimar el valor de $p(x)$. Asimismo se considerará una pequeña región $\mathcal{R}$ que contiene a x. La masa de probabilidad asociada con esta región

[1]Esta idea es el fundamento de lo desarrollado en el capítulo 2. Particularmente la subsecciones: 3.1.1. - 3.1.2. - 3.1.3. - 3.1.4. Buscando densidades vía histogramas. Estas últimas logran ser bien aproximadas con los intervalos de Freedman-Diaconis. Todo este desarrollo posee una estructura análoga a lo realizado por el algoritmo Kernel Smoothing Regression; explicado en la presente subsubsección.

está dada por,

$$P = \int_{\mathcal{R}} p(x)\partial x$$

Ahora se procede en suponer que se ha recopilado un conjunto de datos que comprende N observaciones extraídas de $p(x)$. Como cada punto de datos tiene una probabilidad P de caer dentro de $\mathcal{R}$, el número total K de puntos que se encuentran dentro de $\mathcal{R}$ distribuirán de acuerdo a una distribución binomial,

$$Bin(K|N,P) = \frac{N!}{K!(N-K)!}P^K(1-P)^{1-K}$$

De modo que al recordar una distribución binomial genérica,

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

Y teniendo en consideración que,

$$\mathbb{E}(m) \equiv \sum_{m=0}^{N} mBin(m|N,\mu) = N\mu$$

$$\mathbb{V}(m) \equiv \sum_{m=0}^{N}(m-\mathbb{E}(m))^2 Bin(m|N,\mu) = N\mu(1-\mu)$$

Se observa que la fracción media de puntos que caen dentro de la región es, $\mathbb{E}[K/N] = P$. De manera similar utilizando varianza para el mismo factor se obtiene que $\mathbb{V}[K/N] = P(1-P)/N$. Para $N \to \infty$, esta distribución tendrá un peak alrededor de la media. Por tanto,

$$K \simeq NP$$

Sin embargo, si también se asume que la región $\mathcal{R}$ es suficientemente pequeña como para que la probabilidad de densidad $p(x)$ sea aproximadamente constante en la región, entonces se tiene,

$$P \simeq p(x)V$$

Donde V es el volumen de $\mathcal{R}$. Combinando ambas ecuaciones anteriores se obtiene que,

$$p(x) = \frac{K}{NV}$$

Téngase en cuenta que la validez de la ecuación anterior depende de dos supuestos contradictorios: (1) que la región $\mathcal{R}$ sea lo suficientemente pequeña como para que la densidad sea aproximadamente constante sobre la región. Y, sin embargo, lo suficientemente grande (en relación con el valor de esa densidad) que el número K de los puntos que caen dentro de la región son suficientes para que la distribución binomial tenga un peak alrededor de la media (en $N \to \infty$). Se puede explorar el resultado anterior de dos maneras diferentes: (1) O se puede corregir K y determinar el valor de V a partir de los datos, lo que da lugar a la técnica de K-Nearest Neighbor que se discutirá en breve, (2) o se puede arreglar V y determinar K a partir de los datos, dando lugar al enfoque del kernel. Se puede demostrar que tanto el estimador de densidad K-Nearest Neighbor como el estimador de densidad del

kernel convergen con la densidad de probabilidad en el límite $N \to \infty$ siempre que V se contraiga adecuadamente con N, y K crezca con N [15].

Ahora se considerará una pequeña región $\mathcal{R}$ el cual será un hipercubo centrado en el punto x en el que se desea determinar la densidad de probabilidad. Para contar el número K de puntos que se encuentran dentro de esta región, es conveniente definir la siguiente función,

$$k(u) = \begin{cases} 1 & |u_{\mathrm{i}}| \leq 1/2 \ \ \mathrm{i} = 1, ..., D \\ 0 & \sim \end{cases}$$

que representa un cubo unitario centrado en el origen. La función $k(u)$ es un ejemplo de una función del núcleo (kernel). Más adelante se verán las restricciones que debe cumplir el kernel $k(u)$, de modo que existen más funciones posibles además del hipercubo. A partir de la ecuación anterior se tendrá que la cantidad $k(\frac{x-x_n}{h})$ será 1 sí el punto de datos x_n se encuentra dentro de un cubo de lado h centrado en x, y 0 en caso contrario. Por lo tanto, el número total de puntos de datos que se encuentran dentro de este cubo será,

$$K = \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right)$$

Sustituyendo en la ecuación $p(x) = \frac{K}{NV}$ da como resultado la ecuación propuesta por Loftsgaarden y Quesenberry [20],

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{x - x_n}{h}\right)$$

En donde se ha usado $V = h^D$ para el volumen de un hipercubo de tamaño h en D dimensiones. Finalmente - para la ecuación anterior - se puede escoger cualquier otra función kernel $k(u)$ siempre y cuando cumpla con las condiciones,

$$k(u) \geq 0$$

$$\int k(u) \partial u = 1$$

El modelo de clase de densidad dado por la ecuación superior se llama estimador de densidad del kernel, o estimador Parzen.
Tiene un gran mérito en que no haya cómputo previo involucrado en la fase de 'entrenamiento". Porque esto simplemente requiere el almacenamiento del conjunto de entrenamiento. Sin embargo, esto es también una de sus grandes debilidades porque el costo computacional de evaluar la densidad crece linealmente con el tamaño del conjunto de datos. Dando justificación a la advertencia señalada en la desventaja de estos modelos.

2.2.7. Redes Neuronales

Este algoritmo tiene su nombre debido a la similitud que existe (en su funcionamiento) con el cerebro humano. Se compone de neuronas (nodos) de entrada, neuronas (nodos) ocultas,

y neuronas (nodos) de salida. Todas ellas conectadas con "redes" dentro del algoritmo. Este método funciona como caja negra. Estos nodos están conectados entre sí, y la fuerza de sus conexiones entre sí se asigna un valor basado en su fuerza: inhibición (máxima es -1.0) o excitación (máxima es +1.0). Si el valor de la conexión es alto, entonces indica que hay una conexión fuerte. Dentro del diseño de cada nodo, se incluye una función de transferencia. Los nodos de entrada reciben información, en la forma que puede expresarse numéricamente. La información se presenta como valores de activación, donde a cada nodo se le asigna un número, cuanto mayor sea el número, mayor será la activación. Esta información se pasa a través de la red. En función de las fortalezas de conexión (pesos), inhibición o excitación, y las funciones de transferencia, el valor de activación se pasa de nodo a nodo. Cada uno de los nodos suma los valores de activación que recibe; luego modifica el valor basado en su función de transferencia. La activación fluye a través de la red, a través de capas ocultas, hasta que alcanza los nodos de salida. Los nodos de salida luego reflejan la entrada de manera significativa hacia el mundo exterior. La diferencia entre el valor predicho y el valor real (error) se propagará hacia atrás al repartirlos a los pesos de cada nodo según la cantidad de este error del cual es responsable el nodo [26].

Redes Neuronales para Regresiones

Los modelos de regresión y clasificación vistos en esta memoria son por lo general combinaciones lineales de funciones básicas no lineales fijas $\phi_j(x)$ y toman la forma,

$$y(x, w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right)$$

Donde $f(\cdot)$ es la identidad en el caso de la regresión. El objetivo central de redes neuronales es extender esto haciendo que las funciones básicas $\phi_j(x)$ dependan de los parámetros y luego permitir que estos parámetros se ajusten, junto con los coeficientes $\{w_j\}$, durante el entrenamiento. Esto conduce al modelo básico de redes neuronales, que se puede describir a partir de una serie de transformaciones funcionales. Primero se construyen M combinaciones lineales de las variables de entrada $x_1, ..., x_D$ en la forma,

$$a_j = \sum_{\mathrm{i}=1}^{D} w_{j\mathrm{i}}^{(1)} x_{\mathrm{i}} + w_{j0}^{(1)}$$

En donde $j \in \{1, ..., M\}$ y el superíndice (1) indica que los correspondientes parámetros son los de primera capa de la red [22]. Asimismo los parámetros $w_{j\mathrm{i}}^{(1)}$ indican los pesos y los parámetros $w_{j0}^{(1)}$ indican los sesgos. Finalmente, las activaciones de la unidad de salida se transforman utilizando una función de activación apropiada para proporcionar un conjunto de salidas de red y_k. La elección de la función de activación está determinada por la naturaleza de los datos y la distribución supuesta de las variables objetivo. Por lo tanto, para problemas de regresión estándar, la función de activación es la identidad, de modo que

$$y_k = a_k \Longrightarrow y_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} = \sum_{j=1}^{M} w_{kj}^{(2)} h(a_j) + w_{k0}^{(2)}$$

En donde las cantidades a_j se conocen como activaciones. Cada uno de ellos se transforma utilizando una función de activación no lineal diferenciable $h(\cdot)$.

2.3. Análisis Crítico

A continuación se presentará al lector una discusión respecto a las trampas existentes en los modelos descritos. En particular, teniendo en cuenta las desventajas de los modelos entorno a los puntos de (1) gasto computacional, (2) tamaño necesario de bases a utilizar, (3) sensibilidad de los datos.
Sobre gasto computacional se debe tener suma precaución en los algoritmos de "caja negra"; aquellos de alta complejidad; y/o con alta cantidad de parámetros. Esto pues la idea del producto a entregar dentro de esta memoria, es que sea eficiente - entre otras cosas - en los tiempos. Asimismo se puede nombrar la memoria computacional. Pues al tener procesados muchos datos - en una lapso muy largo de tiempo - los computadores *no preparados*[2] no soportan la información resuelta de la predicción, y se terminan *congelando*. Considerando este punto es que se debería tener especial atención en los algoritmos: Artificial Neural Networking, Kernel Smoothing Regression y Support Vector Machine.
En lo que respecta al segundo punto se debe tener en cuenta que la presente memoria utiliza una base compuesta de 192.179 datos. En donde además, cada una de estas observaciones será analizada cada una con 7 características para Linear Regression; y 9 para el resto de los algoritmos. Evidenciando que se tendrá un análisis de una base grande en cantidad de datos. Si es por este punto entonces se debe tener precaución con: Kernel Smoothing Regression, K-Nearest Neighbor y Artificial Neural Networking.
En el último punto de sensibilidad, se debe tener en cuenta qué tan sensibles son los algoritmos a la incorporación de características irrelevante o repetidas dentro de la base de datos. Lo irrelevante se explica sobre el contexto de las predicciones. Si hablamos de precios de inmuebles de nada sirve el precio de las manzanas. No obstante ello puede existir una correlación que permita definir una relación espuria. Claramente esto último es una caricatura pues omite el hecho de que el analista haga la filtración previa de la base de datos. Todo este ejemplo tiene por intención levantar el punto de una relación - posiblemente - espuria.
En vista de las caracterizaciones anteriores, los algoritmos de los cuales se debe tener precaución son los de tipo Kernel Estimator. Por tanto, en detalle, estos son K-Nearest Neighbor y Kernel Smoothing Regression.

Se concluye que hay que tener especial atención con los algoritmos tipo Kernel Estimator, Artificial Neural Networking y Support Vector Machine.

[2]Cuando se utiliza esta caracterización se hace referencia a aquellos que poseen procesadores no tan rápidos (inferior o igual a Intel i5), y una RAM pequeña (igual o menor a 4 GB).

Capítulo 3

Descripción Metodológico

Los datos que serán utilizados en la presente memoria provienen principalmente de una data histórica en la venta y arriendo de diversos inmuebles. Son en total 192.179 datos creados a partir de comienzos de marzo de 2019 a mayo de 2019. Los datos se presentan de manera discreta considerando los días como períodos de análisis. Estos son,

1. **id:** Identificación de la vivienda dentro del sistema del proveedor de la base de datos www.portalinmobiliario.com
2. **id2:** Identificación de la vivienda dadas sus características.
3. **fechapublicacion:** Fecha en que la vivienda fue publicada para su compra/arriendo.
4. **fechascrap:** Fecha cuando la publicación fue retirada del portal.
5. **region:** Región en donde se encuentra el inmueble.
6. **direccion:** Dirección de la vivienda (posee múltiples formatos de entrada).
7. **operacion:** Clasificación de vivienda según: arriendo y venta.
8. **tipo:** Clasificación del inmueble según: “casa”, “comercial”, “departamento”, “estacionamiento”, “oficina” y “sitio”.
9. **precio:** Precio de la vivienda.
10. **dormitorios:** Número de dormitorios.
11. **baños:** Número de baños.
12. **metrosmin:** Metros mínimos construidos en la vivienda.
13. **metrosmax:** Metros máximos pertenecientes al terreno de la vivienda.
14. **estacionamiento:** Número de estacionamientos.

15. **lat:** Ubicación geográfica de latitud.

16. **lon:** Ubicación geográfica de longitud.

Estos datos han sido compilados a partir de una API unida al sitio web *portalinmobiliario.com* perteneciente a una empresa del rubro "FONDOS Y SOCIEDADES DE INVERSIÓN Y ENTIDADES FINANCIERAS SIMILARES" encargada de desarrollar tasaciones en el mercado a partir de un algoritmo de Machine Learning. Dicha empresa será omitida por razones de confidencialidad. Dentro de las variables que efectivamente se van a utilizar dentro de la base se encuentran: (1) datos de localización geográfica, (2) precio de inmuebles, (3) características observables de los inmuebles como lo son: (3.1) precio, (3.2) dormitorios, (3.3) baños, (3.4) metros mínimos-máximos, (3.5) estacionamientos, (3.6) comuna de localización, (3.7) latitud de localización y (3.8) longitud de localización; (4) operación de la vivienda[1] y (5) tipo de vivienda[2] Se realizará entonces un análisis de precio por cada tipo/operación de inmueble para ver cuál debiese ser el valor predicho más ajustado al valor real que sin entrar en sobreajustamiento o overfitting. Para ver qué datos característicos del inmueble son significativos; junto a el valor de otros factores no considerados dentro de la base de datos que puedan ser agregados como variables explicativas.

3.1. Análisis de Datos

A continuación se mostrará un análisis de algunas características significativas pertenecientes a cada uno de los inmuebles. En detalla, dentro de la presente memoria, se trabajará como variable dependiente: `ln(PRECIO)`. La aplicación de logaritmo natural `ln()` a la variable precio, se justifica para disminiur la dispersión existente en los datos de dicha característica. Facilitando una comprensión de diferencias mínimas, dentro de un continuo [16].
En tanto, como variables independientes[3]: `PRECIOS; DORMITORIOS;BAÑOS; METROS MÍNIMOS; METROS MÁXIMOS; ESTACIONAMIENTO y COMUNA`. Utilizando las variables `TIPO`, `OPERACIÓN` e `INTERVALOS DE FREEDMAN-DIACONIS`[4] para agrupar los inmuebles. Como la empresa desea predecir el precio, este pasa a ser el objetivo de la presente sección. Es decir, todos los análisis correspondientes a esta parte de la memoria serán utilizados con ese fin; utilizando las características relevantes y necesarias para formalizar, y orientar, sobre los precios de los inmuebles.
La base contiene la información de inmuebles desde marzo del 2019 a mayo del 2019. En detalle se usará el sub-conjunto perteneciente a inmuebles de la Región Metropolitana, 192.179 datos en total que representan el 81,5 % de toda la base. El otro 18,5 % se reparte en las demás regiones.

Se desea observar la validez de la premisa: *"A conjuntos, más acotados de inmuebles con similares características (clustering); los inmuebles convergen a un valor único, característico*

[1]Encargada de clasificar inmuebles: (1) comprados o (2) vendidos.

[2]Encargada de clasificar inmuebles: (1) comerciales, (2) casa, (3) departamento y (4) oficina.

[3]Linear Regression es el único modelo que no dispondrá de las variables de `LATITUD` y `LONGITUD`.

[4]Intervalos generados con cotas inferiores y superiores óptimas. Resultantes de la partisión creada por el algoritmo de Freedman-Diaconis [17].

de ellos.". Esta frase es la justificación de la metodología utilizada por la empresa. Creando clusters que tienen como centroide al inmueble que se desea predecir (más detalle véase la sub-sección 1.2.1). Esto se manifiesta matemáticamente en reducir la varianza dentro de cada uno de los clusters formados, dado un promedio muestral cualquiera.
La idea de esta sección está en - a través de los resultados - formalizar una idea que permita aceptar (o rechazar) la la eficiencia del algoritmo utilizado por la empresa. Viendo qué tan bueno es tomar elementos con características idénticas; y cercanos entre ellos[5].
Finalmente el valores promedio monetarios a utilizarse, y pertenecientes a los elementos de las segmentaciones, estará en Unidades de Fomento (UF). Usando la valorización registrada el 31 de diciembre del 2019, con un monto de CLP$28.310.

3.1.1. Análisis de Precio

A continuación se presenta un análisis del precio de cada uno de los inmuebles existentes en la base de datos pertenecientes a la Región Metropolitana. Primero, se entenderá de forma generalizada, cómo se comportan los precios sin distinción de tipo u operación. Después se crearán tablas que segmenten dichos resultados según los criterios anteriormente expuestos.

En la base de datos existe un gran problema de dispersión. El mínimo valor de la variable `precio`, en el conjunto de todos los inmuebles de la Región Metropolitana de la República de Chile, es 0,00007059769. Mientras que el máximo está dado por 610.804.129. En vista de ello, no se hace fácil dimensionar la dispersión existente en este conjunto. Más, se hace incompatible la comparación entre los valores por la gran magnitud de la varianza poblacional, y la distancia entre el valor mínimo y máximo dentro del conjunto.
Considerando lo anterior, se hace imposible construir un histograma[6] adecuado que permita visualizar de manera aceptable la distribución de los valores pertenecientes a la variable `precio`.
Gráficamente esto es visualizable en la construcción misma de distintos histogramas. Como los que se presentarán en el siguiente apartado[7].

En la figura 3.1 se hace evidente que a pesar de la concentración de valores al inicio del gráfico, no es posible diferenciar - tanto visual como técnica - los distintos valores que subyacen en los inmuebles. Esto ya, previamente justificado, por la existencia de significativas diferencias entre los valores `precio`; destacado aún más en sus extremos.
La normalización de los elementos vía logaritmo natural se expresa visualmente en la siguiente gráfica.

[5]Este último punto no se analizará métricamente esta sección dado que los elementos no se pueden manipular en esa categoría con los elementos observados.

[6]Como nota al lector, recordar que los histogramas poseen intervalos internos que permiten identificar - vía frecuencia acumulada - cuántos valores (del elemento de interés) hay dentro del rango correspondiente.

[7]Tómese la siguiente precaución respecto a la relación de las barras y la densidad dentro del gráfico: La densidad de probabilidad posee un área bajo la curva igual a 1. De ahí se justifican los valores correspondientes al eje Y. En cambio las barras hacen alusión a la frecuencia acumulada de cada uno de los valores en el histograma, sin representación alguna dentro del eje Y.

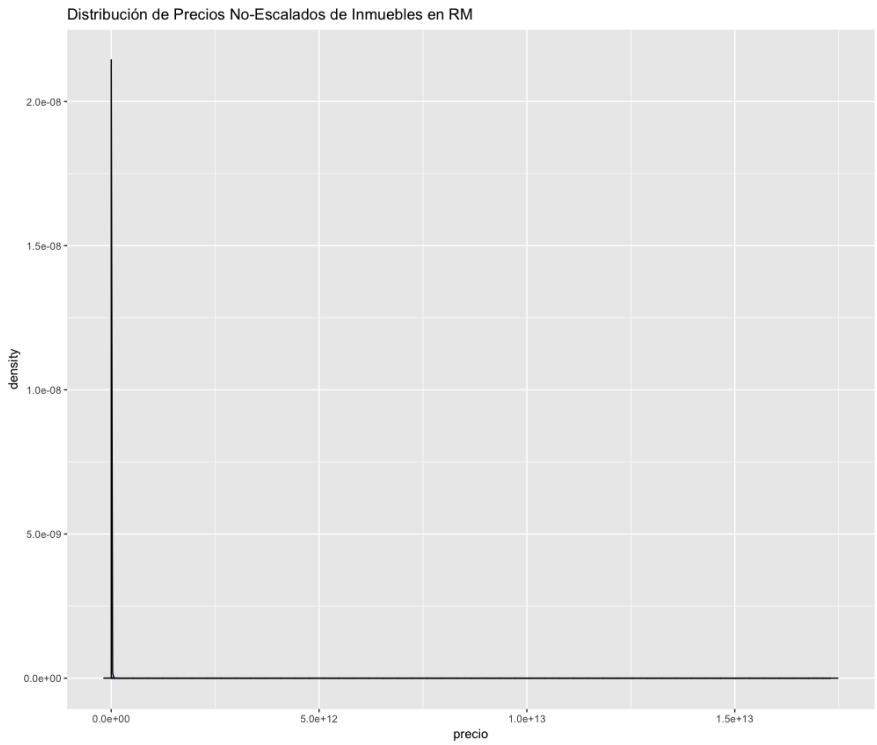

Figura 3.1: Distribución de Precios No-Escalados de Inmuebles en la Región Metropolitana - Fuente: Elaboración Propia

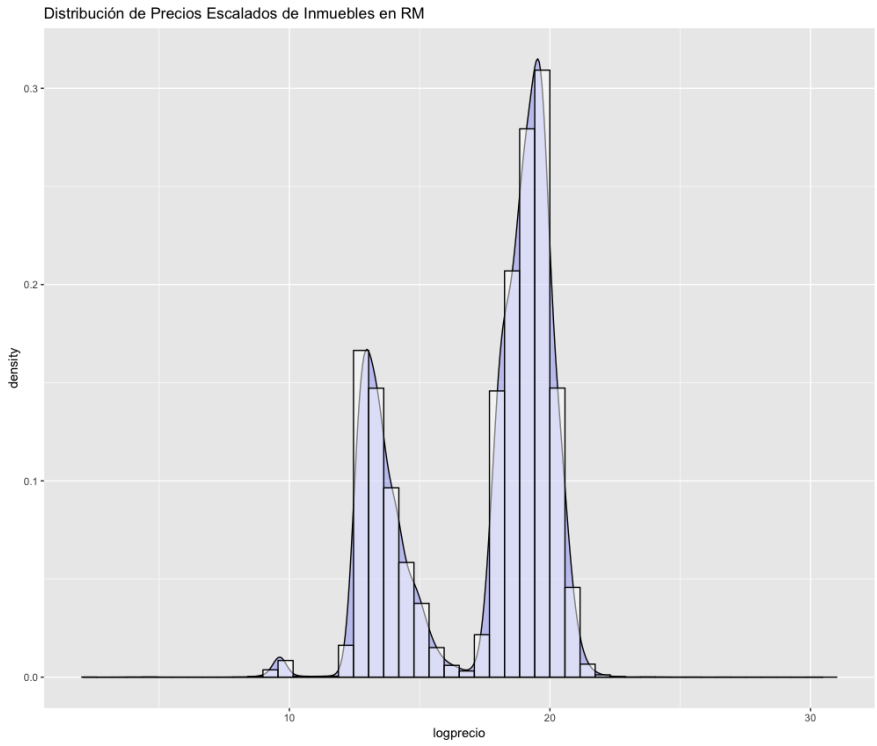

Figura 3.2: Distribución de Precios Escalados de Inmuebles en la Región Metropolitana - Fuente: Elaboración Propia

En el caso de la figura 3.2 se observa una visualización clara de las frecuencias acumuladas. Justificando de esta manera el uso de logaritmo natural para una normalización de valores. Resulta interesante destacar la existencia de 3 campanas; las cuales crecen a medida que se avanza en el eje X.

Al estar todo contenido dentro del gráfico, no hay distinción alguna de lo que se contiene en él. En particular, no hay forma alguna de identificar el origen y composición de las 3 campanas anteriormente señaladas. Es por esto que se plantea al lector, la creación de distintas diferenciaciones. Las cuales podrán ir de manera conjunta en el tiempo, con el fin de modelar de mejor forma las gráficas. Más detalladamente en la convergencia de las barras a las curvas de los respectivos gráficos de densidad.
Lo anterior se justifica en que un mayor número de aciertos en las frecuencias acumuladas - con los respectivos intervalos - hará que el conjunto de barras converja - lo más posible - al área bajo la curva. Dando una estimación predictiva cada vez mejor[8] de la curva a estimar del gráfico formado por el conjunto de los precios-venta de los inmuebles. Esta es la idea central de Kernel Smoothing Regression, un algoritmo que se utilizará en la presente memoria, y que tiene por objetivo optimizar el tamaño de los *bandwidths* para aproximar - de la manera más aceptable posible - el conjunto de bandas a la curva de la función. Un método para encontrar la cantidad óptima - y por consecuencia el ancho óptimo - de los intervalos internos del eje X en un histograma, es usando la partisión de Freedman-Diaconis [17].

3.1.2. Análisis por Operación de Inmueble

En esta oportunidad se analizará el comportamiento de los datos segmentados según su respectiva operación. Es decir, separados entre: arriendo y venta.
La Tabla 3.1 manifiesta dicha segmentación,

	Operación	Tamaño resp. total	Tamaño como subconjunto	Promedio (UF)	Desv. Estándar (UF)
1	Arriendo	34,8 %	66.936	$3{,}25 \cdot 10^2$	$2{,}315 \cdot 10^4$
2	Venta	65,2 %	125.243	$2{,}2 \cdot 10^4$	$2{,}149 \cdot 10^6$

Tabla 3.1: Tabla resultado de inmuebles clasificados según su operación - Fuente: Elaboración Propia

Se evidencia que hay una predominancia de inmuebles en venta. En particular 125.243, mientras que hay 66.936 inmuebles en arriendo. Siendo la cantidad de venta, el doble de la cantidad de arriendos.

Previo a su análisis, primero se debe hacer hincapié en la relación existente entre el promedio y la varianza para tener el permiso a concluir un par de elementos. Algunos escenarios considerando dichas métricas muéstrales pueden ser,

1. **Promedio conocido - Alta varianza:** Sin importar el valor del promedio, si el conjunto posee una alta varianza quiere decir que los datos están más disperso. Más sesgo en la muestra. Por tanto los valores al interior del conjunto son notoriamente distintos dado un alto valor de varianza.

[8] Si el conjunto de barras dibuja en su totalidad (o lo más posible) el área bajo la curva del gráfico.

2. **Promedio conocido - Baja varianza:** Aquí el escenario se divide en 2. Si el promedio es bajo (resp. alto) entonces en presencia de una baja varianza, entonces los datos estarán concentrados más cerca de 0 (resp. más alejado de 0)

Mirando la Tabla 3.1 se tiene que, dado un promedio dado para ambos conjuntos, la varianza destaca por ser excesivamente alta en ambos casos. Cumpliendo con lo señalado en el punto 1 de la enumeración anterior. En arriendo una desviación estándar del orden de $2{,}315 \cdot 10^4$, mientras que en compra tiene una magnitud de $2{,}14910^6$.
Si se tiene en cuenta esto; y se comparan los datos resultantes con los valores reportados por las estadísticas habitacionales (de precios) del Observatorio Urbano del Ministerio de Vivienda y Urbanismo [5], se tiene que en la Región Metropolitana el precio de arriendo promedio es del orden de $226.669 (8,006676 UF). Esto considerando viviendas tipo *hogar*: casas y departamentos.
Creando una diferencia de valores (promedio) igual a 316,9933 UF entre los datos de la empresa menos los datos del informe del Observatorio Urbano del Ministerio de Vivienda y Urbanismo.
Las diferencias son radicalmente altas para ser aceptadas a priori. Por tanto se concluye que una única segmentación por operación no es suficiente para agrupar los datos. Tanto por su valor promedio registrado, como por la varianza obtenida. Esto se puede observar en la siguiente gráfica.

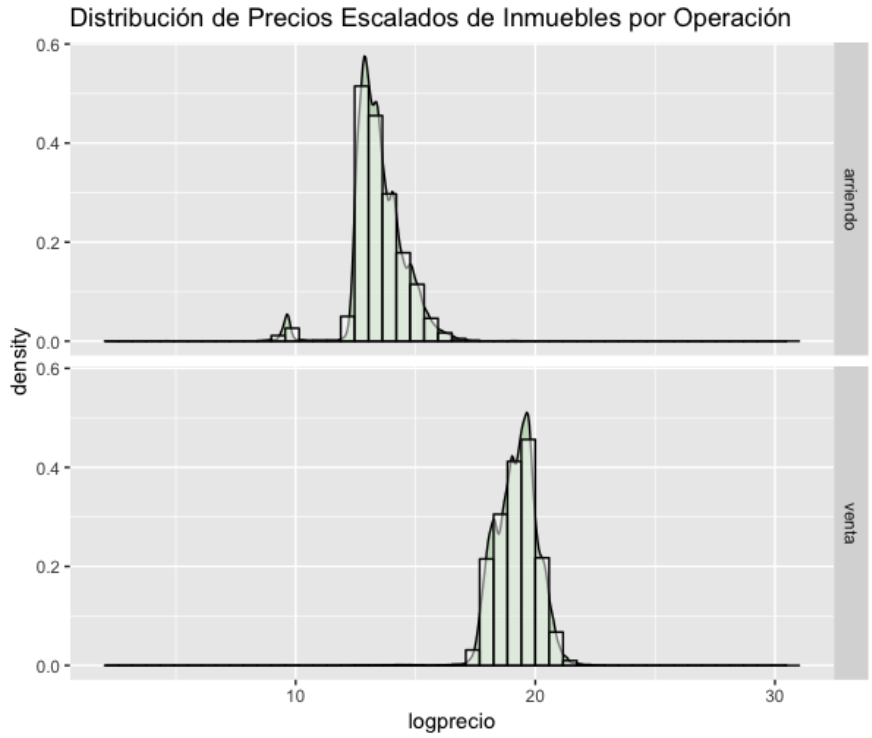

Figura 3.3: Distribución de Precios Escalados de Inmuebles por Operación en la Región Metropolitana - Fuente: Elaboración Propia

Se recuerda al lector que se busca disminuir la magnitud de la varianza para certificar la herramienta usada por la empresa. Clusterizando cada vez más, hasta tener conjuntos con inmuebles de similares características, que permitan tener una varianza poblacional aceptable

para compararlos con el inmueble de interés.

3.1.3. Análisis por Tipo de Inmueble

En esta oportunidad se toma la variable tipo para clasificar subconjuntos de análisis. Aquí las posibles opciones las posibles opciones de clasificación son: (1) casa, (2) comercial, (3) departamento, (4) estacionamiento, (5) oficina o (6) sitio.
Al igual que la sección anterior, aquí se clasificará con la misión de querer visualizar las distintas varianzas muéstrales dentro de cada uno de los sub-conjuntos para diagnosticar si efectivamente bajó la varianza, o no.

	Tipo	Tamaño resp. total	Tamaño como sub-conjunto	Promedio (UF)	Desv. Estándar (UF)
1	Casa	34,4 %	66.155	$1{,}89 \cdot 10^4$	$1{,}594 \cdot 10^6$
2	Comercial	1,67 %	3.218	$1{,}98 \cdot 10^5$	$1{,}077 \cdot 10^7$
3	Departamento	55,9 %	107.336	$7{,}6 \cdot 10^3$	$5{,}899 \cdot 10^5$
4	Estacionamiento	0,0208 %	40	$3{,}08 \cdot 10^3$	$7{,}8740 \cdot 10^3$
5	Oficina	7,76 %	14.920	$3{,}81 \cdot 10^3$	$9{,}8944 \cdot 10^4$
6	Sitio	0,265 %	510	$2{,}87 \cdot 10^4$	$6{,}6633 \cdot 10^4$

Tabla 3.2: Tabla resultado de inmuebles clasificados según su tipo

Observando la Tabla 3.2, se tiene que la mayor concentración de datos se centra en los departamentos. Mientras que los estacionamientos y los sitios poseen una cantidad técnicamente insignificante. Esto último señala que no será conveniente trabajar con dichos tipos de inmuebles, por la posibilidad de entrar en sub-ajuste dentro de la calibración y posterior predicción de modelos. Por tanto, se omitirá desde ahora en adelante trabajar con esos tipos de inmuebles.

Se intentará en la próxima sub-sección continuar con una clasificación; pero esta vez con un criterio conjunto que una estas 2 primeras sub-secciones. Es decir, manteniendo de forma simultánea la clasificación: tipo-operación. La representación gráfica de las distribuciones en los 6 casos anteriores, se observa en la figura 3.4.
La lectura de los histogramas en este caso resulta interesante. Observar que algunos tipos de inmueble poseen 3 montes, mientras que otros poseen 2. Implicando en la existencia de 2 a 3 sub-grupos de interés dentro de dichas categorías.
En cuanto a la idea expuesta anteriormente en la convergencia del número de barras a la curva del gráfico de densidad, se observa cómo esto se cumple - de manera cercana - en el caso de los inmuebles tipo casa. No así en los estacionamientos en donde claramente no hay semejanza. Esto, por tanto, refuerza la idea de implementar intervalos de Freedman-Diaconis.
Finalmente en cuanto a los valores de precio, se confiesa que es extraña la existencia de 3 montes en algunos casos. Se podría pensar que los 2 montes hacen referencia a la operación arriendo-venta, pero pierde sentido la existencia de un tercero. En paralelo, se observa cómo las campanas mayores (en el eje X) convergen por el lado izquierdo a ln(20). Si bien hay una diferencia de magnitud significativa entre los valores traducidos a los originales, se evidencia la existencia de un intervalo relevante dado por 20 ± 2 en el eje X.

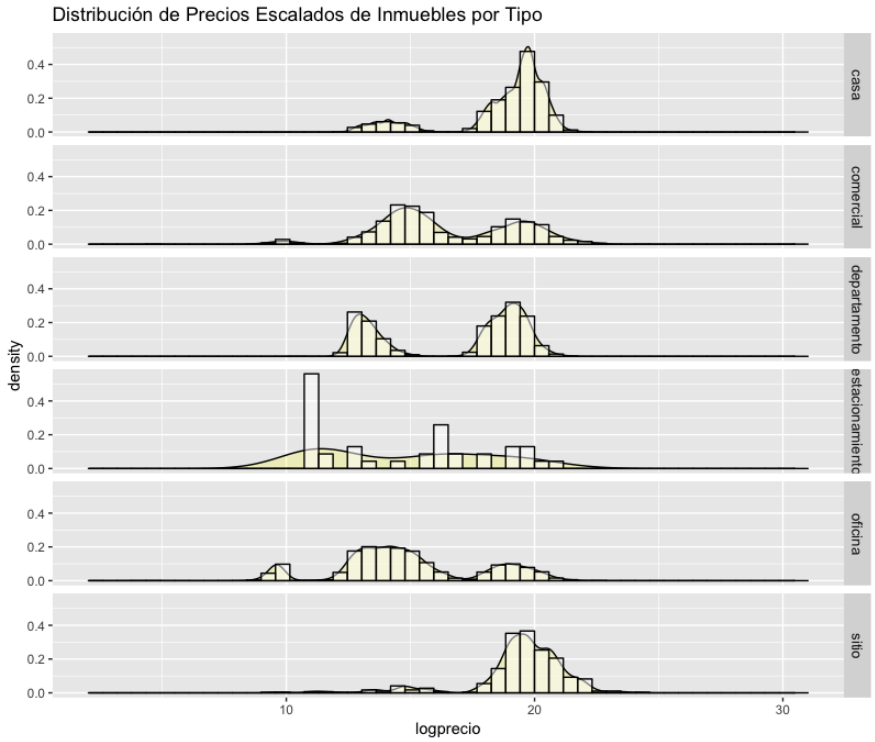

Figura 3.4: Distribución de Precios Escalados de Inmuebles por Tipo en la Región Metropolitana - Fuente: Elaboración Propia

3.1.4. Análisis Combinado Tipo-Operación de Inmueble

Se explicará previamente al lector el cómo se generó - computacionalmente - esta división de elementos dentro de la base de datos.
Teniendo en consideración que cada uno de los inmuebles posee las características tipo y operación, es que se crearon por cada una de estas características, variables dummies asociadas a cada una de ellas. De tal modo que se mantendrá si cumple con la clasificación asociada a la característica. Un ejemplo de ello puede ser: tipo (característica) - casa (clasificación) o operación (característica) - arriendo (clasificación). Lo anterior resumido en la siguiente dummy representativa,

$$y_{\text{i}} = \begin{cases} 0 & \text{No se mantiene observación i.} \\ 1 & \text{Se mantiene observación i.} \end{cases}$$

Gráficamente los inmuebles clasificados de manera conjunta en la combinación tipo-operación, distribuyen de la forma mostrada en la figura 3.5.

A pesar de la incorporación de este criterio conjunto sigue sin ser suficiente para que el conjunto de barras converja al total del área bajo la curva. En este análisis combinado de tipo (4) y operación (2), se tuvieron: $4 \times 2 = 8$ sub-conjuntos de análisis. Al ya estar clasificados los inmuebles con este criterio, ya no hay más formas de "clasificarlos nuevamente" en algún otro criterio distinto.
Para entender mejor, el lector puede observar la primera columna (desde izquierda a derecha)

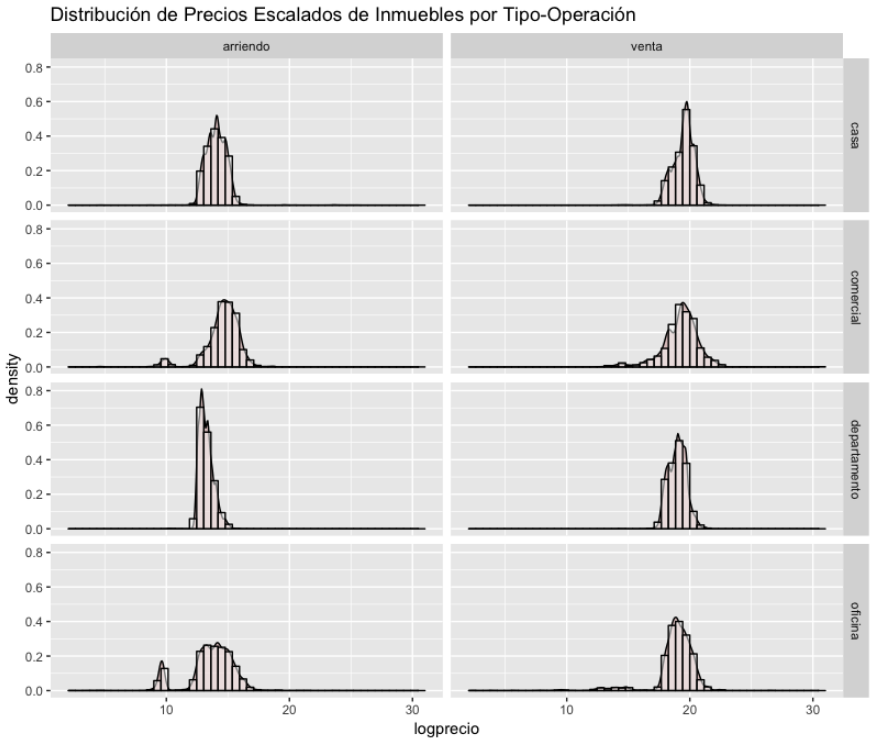

Figura 3.5: Distribución de Precios Escalados de Inmuebles por Tipo-Operación en la Región Metropolitana - Fuente: Elaboración Propia

en cada una de las tablas realizadas en las sub-secciones 3.1.2 y 3.1.3. Ese era el criterio respectivo de cada una de ellas. Una variable fija por cada clasificación. Pero esta vez los inmuebles poseen su clasificación que agota 2 posibilidades (tipo-operación). Al obedecer las variables dummies respectivas, nace la pregunta ¿existirá algún otro criterio conveniente para acotar aún más las varianzas muestrales dentro de estos nuevos conjuntos, permitiendo capturar toda el área bajo la curva para poder estimar cualquier punto dentro de una eventual predicción?
La respuesta se encuentra en la regla planteada por Freedman-Diaconis. La cual consiste en encontrar las cotas mínimas y máximas de los intervalos formalizados al interior del eje X en un histograma. Con el fin de tener intervalos parcializados con un ancho ideal que permita crear un *bandwidth* óptimo [17].

Finalmente teniendo en cuenta los elementos importantes a tener presente con el uso del método de Freedam-Diaconis, y los resultados mostrados en la tabla 3.3, se puede reportar lo siguiente

1. **Espaciado:** A un menor espaciado, mayor será la cantidad de intervalos; y por consecuencia una estimación más ajustada a la curva generada por la función. En esta característica los inmuebles con espaciados mínimos (dentro de todos los existentes) son: (1) departamentos en venta, (2) casas en venta y (3) departamentos en arriendo. Coincidentemente son aquellos sub-conjuntos con mayor concentración de datos, como se verá más adelante. Esta relación es esperable pues, a mayor cantidad de información en un conjunto, mayor es la probabilidad que posea un alto grado heterogeneidad. De

Intervalos de Freedman-Diaconis					
Operación	Tipo	Espaciado	Cota Mínima	Cota Máxima	# Intervalos Internos
Venta	Comercial	0,32	11,1	30,19	60
	Casa	0,06	8,14	30,07	376
	Departamento	0,06	7.91	29,22	392
	Oficina	0,17	2.4	26,43	146
Arriendo	Comercial	0,22	4.61	18,44	64
	Casa	0,11	4,71	25,41	184
	Departamento	0,05	6,31	24,22	393
	Oficina	0,17	2,08	24,91	132

Tabla 3.3: Tabla informativa de características de intervalos Freedman-Diaconis - Fuente: Elaboración Propia

ahí entonces que necesita de más intervalos. Pues necesita capturar los elementos semejanzas en la partisión. Téngase en cuenta que a mayo dispersión de datos, más costará crear una curva. Por tanto, para tener un ajuste óptimo, se necesitarán más intervalos. Sigue que a menor espaciado, mayor cantidad de datos en el conjunto.

2. **Cotas Mínimas:** Estas son muy diversas. Se observa una diferencia significativa entre los valores correspondientes a cada clasificación. Mostrando que el arriendo de inmuebles de oficina son lo más accesible. En contraste de la venta en inmuebles comerciales, que es parte como lo más caro del conjunto. Las cotas mínimas no influyen directamente en la cantidad de intervalos.

3. **Cotas Máximas:** Ídem anterior. Destacan los máximos en inmuebles: (1) casa, (2) departamento y (3) comerciales. Todo el conjunto de venta. Diferenciándose significativamente con el conjunto de arriendo. Se reitera que estas cotas no influyen de manera individual, ni directa, en la cantidad de intervalos.

4. # **Intervalos Internos:** Este es el elemento más importante a rescatar. Una baja cantidad de intervalos se traduce en que los elementos no están lo suficientemente dispersos como para crear otro intervalo adicional. Entre menos intervalos, menos entropía entre los elementos internos del intervalo. En caso contrario, una alta cantidad de los mismos implica que hay que diferenciarse aún más. Sujeto a la suavización de la curva que generan las barras (de frecuencia acumulada) en cada clasificación tipo-operación. Es por ello que es interesante observar los inmuebles comerciales en venta. A pesar de presagiar que son los inmuebles más caros (dadas sus cotas), parece ser que sus valores son muy similares entre ellos. Permitiendo ser agrupados de manera más fácil dentro de un intervalo. No así los departamentos en venta o arriendo que, a pesar de ser parte de los inmuebles con mayor cantidad de elementos, poseen la mayor cantidad de intervalos.

Finalizando la interpretación gráfica, se observa cómo Freedam-Diaconis es útil - y necesario - en los casos que poseen la mayor cantidad de intervalos. Estos inmuebles son: (1) departamentos en arriendo, (2) departamentos en venta, (3) casas en venta. Pues al observar la figura 3.5 se evidencia que las barras creadas no son suficientes para cumplir con el objetivo de con-

vergencia al área bajo la curva. Más, en opinión del autor, son las curvas más complicadas. De ahí se explicaría la necesidad de más intervalos en esos inmuebles en particular[9].

Como se ha separado en un set de intervalos por cada clasificación, se procederá en capturar los resultados métricos de toda la clasificación. Teniendo en cuenta que: *"El promedio del total de los componentes, es el total de los promedios de los componentes".* [10] De este modo se calculará la varianza interna de cada intervalo, y se promediará como un todo; considerando los demás.
Los resultados de cada segmentación se muestran en la Tabla 3.4,

	Operación-Tipo	Tamaño resp. total	Tamaño como sub-conjunto	Promedio (UF)	Desv. Estándar (UF)
1	Arriendo-Casa	4.85 %	9.303	$1{,}16 \cdot 10^3$	$5{,}1962 \cdot 10^4$
2	Arriendo-Comercial	0.984 %	1.885	$1{,}43 \cdot 10^2$	$2{,}3622 \cdot 10^2$
3	Arriendo-Departamento	23.2 %	44.421	$1{,}26 \cdot 10^2$	$9{,}4710 \cdot 10^3$
4	Arriendo-Oficina	5.88 %	11.275	$4{,}52 \cdot 10^2$	$2{,}4474 \cdot 10^4$
5	Venta-Casa	29.7 %	56.852	$2{,}18 \cdot 10^4$	$1{,}7176 \cdot 10^6$
6	Venta-Comercial	0.696 %	1.333	$4{,}77 \cdot 10^5$	$1{,}6733 \cdot 10^7$
7	Venta-Departamento	32.8 %	62.915	$1{,}29 \cdot 10^4$	$7{,}7071 \cdot 10^5$
8	Venta-Oficina	1.90 %	3.645	$1{,}42 \cdot 10^4$	$1{,}9519 \cdot 10^5$

Tabla 3.4: Tabla informativa de resultados en criterio: operación-tipo de inmueble - Fuente: Elaboración Propia

Como se puede observar en la Tabla 3.4, conforme avanzan los criterios, la varianza muestral del promedio de los precios de mercado va disminuyendo a medida que aumentan (y combinan) los criterios a utilizar. La empresa, además de incorporar esta dupla de clasificación, añade características físicas tales como: `BAÑOS, DORMITORIOS, etc...`. En caso de mantener la estructura del presente análisis, este se vería forzado en añadir más y más características para observar si la veracidad de la frase de interés es correcta, o no. Efectivamente la varianza disminuye. Pero no lo suficiente - ni en todos los casos - como para argumentar que es conveniente seguir acotando conjuntos. Esto se observa más detalladamente en los niveles exponenciales en cada uno de los valores muestrales de varianza. En la sección 3.1.3. los factores fluctúan entre 7 y 12, teniendo una moda de 9; mientras, en la sección 3.1.4. los factores fluctúan entre 4 y 14, y no poseen moda. Como se ve, la varianza aumenta en un caso particular de un cluster.
Algebráicamente[11] esto es condición suficiente para demostrar que el incremento de criterios, para predecir características complementarias a las ya utilizadas dentro de un conjunto acotado, no es un argumento válido computacionalmente. Sigue que la metodología de la empresa no es óptima-ideal, y debe cambiarse para obtener mejores predicciones.

[9] Recuérdese que el gráfico señalado es de precios-venta. Y las barras de la figura también corresponden al mismo. Por tanto la predicción con características observables debe ser (mucho) más compleja

[10] **Ejemplo ilustrativo:** Se tiene un conjunto {2,2,3,3,5,5} → prom1(2,3,5) = 3,33...3 | prom2(2,3,5) = 3,33...3 | prom3(prom1,prom2) = 3,33...3 | promTOT(2,2,3,3,5,5) = 3,33...3. Sigue que prom3 = promTOT.

[11] Basta que un elemento en el conjunto de demostración (dominio) no cumpla con la condición en la función (imagen) para argumentar que es infactible. En este caso, mantener la hipótesis que la varianza - para todos los casos - debe bajar con el avanzar de la suma de criterios.

3.1.5. Evolución Temporal de Precios y Cantidad de Inmuebles Vendidos

A continuación se analizará el comportamiento temporal de los precios de venta promedio de los inmuebles; junto con las cantidades ofertadas a lo largo del horizonte de evaluación.

Se analizarán 80 días en total. En los siguientes contextos: (1) evolución diaria del promedio muestral de precios de venta en los 80 días, (2) evolución diaria de la cantidad muestral ofertada en los 80 días, (3) evolución en día de mes de la cantidad muestral ofertada en los 80 días, (4) evolución en día de mes del promedio muestral de precios de venta en los 80 días. Con el objetivo de poder observar si es factible mantener la estructura temporal de la base de datos. Esto refiere a la característica de atemporalidad. Si el tiempo transcurrido fuese relevante, entonces los datos más cercanos al inmueble a predecir, debieran tener un peso mucho mayor. En contraste, aquellos que estén más alejados, deberán tener un peso menor o hasta nulo. Pero aquí remite la pregunta: en caso de ser significativa la posición temporal de los datos, ¿cuál (y cómo) debiera ser la distribución del peso en cada una de ellas?. Entre los algoritmos a utilizar, no todos admiten al peso como input. Solo algunos como Artificial Neural Networking y Kernel Smoothing Estimator. Esto resultaría en un gran problema en caso de que la posición temporal sea significativa, pues habría que - eventualmente - inventar una distribución de peso para el set de información y forzar su implementación, aunque no se pueda dentro del comando de "R". También se debe señalar la existencia implícita de shocks aleatorios derivados de los posibles acontecimientos económico-sociales que ocurren de manera imprevista. No sabemos lo que ocurrirá mañana, por tanto también habría que modelar un shock aleatorio en caso de ser un modelo temporal.
Resumiendo, se informa al lector que se evaluarán los diferentes contextos señalados para diagnosticar si se debe considerar una base de datos temporal o atemporal. Esto se realizará evaluando si los datos pertenecientes a la base de datos son suficientemente representativos a la realidad; a través de una comparación de los datos de la base de la empresa, con los datos de la base del Instituto Nacional de Estadística [8] [9], y después se discutirá sobre si debe (o no) ser una base temporal o atemporal.

Análisis de Evolución Temporal en Base Completa

Recordando el primer paso descrito con anterioridad, se procede en analizar el gráfico de evolución diaria del promedio muestral de precios de venta en los 80 días analizados. De esto se debe extraer que la tendencia de los datos debiese ser la misma que la reportada por el Instituto Nacional de Estadística [8] [9]. Dando de esta manera el permiso a utilizar la base de datos de la empresa, sin sesgo significativo que afecte los futuros valores de predicción.

Observando la Figura 3.6,
se observan 5 outliers. En particular 2 de dichos valores que pueden ser categorizados como outliers mínimos. Correspondientes a los días 02 de marzo y 26 de marzo del presente año. Mientras que, por otra parte, hay 3 outliers máximos ubicados en los días 18 de abril, 28 de abril, y 08 de mayo. Se observa en los datos históricos del Instituto Nacional de Estadística que, el mes de marzo 2019 presentó una variación porcentual interanual de un 19,6 %, como

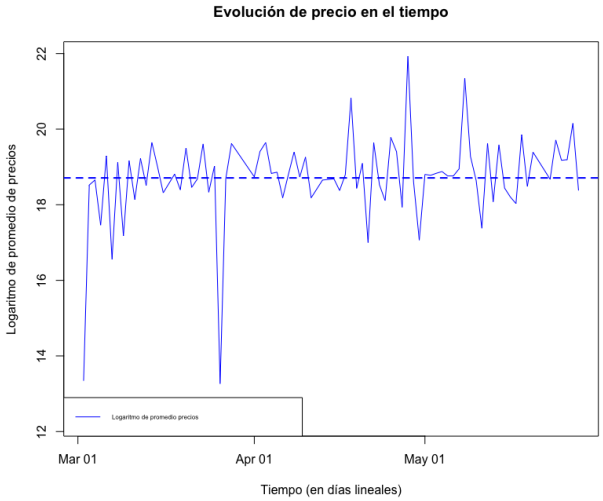

Figura 3.6: Evolución (en días lineales) de `ln(promedios de precios)` en la base completa - Fuente: Elaboración Propia

consecuencia de una mayor cantidad de proyectos en etapa de entrega perteneciente a algunas empresas de proyectos habitacionales. Por su parte, algunas empresas de arriendo también presentaron resultados positivos, a causa de una mayor cantidad de metros cuadrados arrendados.
Lo anterior es consistente con los peaks reportados en marzo (ver Figura 3.6). En donde se cumple empíricamente que a *"menor precio, mayor demanda"* (dentro del set de outliers mínimos). En paralelo, para el caso de outliers máximos se observa que en el mes de abril del 2019, el Índice de Ventas de Actividades Inmobiliarias disminuye. Atribuible posiblemente - entre otras cosas - al aumento de precios disparados y mostrados en los outliers máximos. La evolución del Índice de Ventas de Actividades Inmobiliarias Base Promedio observable desde enero 2017 a julio 2019 se muestra en la Figura A.1.
Ahondando más en lo anterior, se puede observar la evolución temporal de la cantidad de inmuebles operados. El gráfico de evolución diaria de la cantidad muestral ofertada en los 80 días se muestra en la Figura 3.7,

En esta ocasión destacan 2 outliers máximos (correspondientes al 24 y 27 de mayo del 2019) y 3 outliers mínimos (perteneciente al 13 y 15 de abril; y 4 de mayo del 2019).
Se observa que los outliers mínimos del análisis de precio de mercado (02 de marzo y 26 de marzo) no coinciden con algún outlier del análisis de cantidad de inmuebles operados. En cambio los outliers máximos del análisis de precio (18 de abril, 28 de abril y 08 de mayo) coinciden relativamente con los outliers mínimos del análisis de cantidad (13 y 15 de abril; y 04 de mayo). Resaltando una correlación negativa que certifica *"a mayor precio, menor demanda" en arriendo/compra de bienes inmobiliarios.*

Ahora se procede en analizar los inmuebles operados, y sus respectivos precios de venta, en los meses informados; dentro de una ventana de tiempo de 30 días. Esto, para ver sus respectivas tendencias; y analizar la correlación existente entre los valores reportados en el eje Y. Se verá la correlación gracias al uso del coeficiente de correlación de Pearson. Este coeficiente es una medida lineal entre dos variables aleatorias cuantitativas. A diferencia

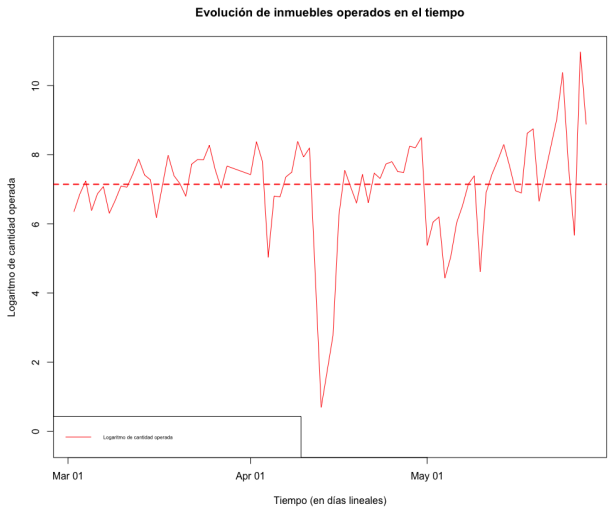

Figura 3.7: Evolución diaria de `ln(cantidad operada de inmuebles)` en la base completa - Fuente: Elaboración Propia

de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables. La única restricción es que las variables a comparar (cantidad operada - cantidad operada y precio en venta - precio en venta) deben ser cuantitativas y continuas. Como se cumple ese requerimiento, se procederá en su uso. La correlación de Pearson está definida como,

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{c\mathbb{V}(X;Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

La hipótesis nula de este comando - en el programa computacional "R" - es: $\rho = 0$, mientras que la hipótesis alternativa es $\rho \neq 0$. Por tanto si el p-valor es cercano a 0, se rechazará la hipótesis nula en favor de la hipótesis alternativa.

La gráfica de la cantidad de inmuebles operados se muestra a continuación,
Utilizando la correlación de Pearson se reporta lo siguiente,

- **MARZO - ABRIL**: $\rho = -0,173845$ y p-valor $= 0,3957$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y abril.
- **ABRIL - MAYO:** $\rho = 0,3994231$ y p-valor $= 0,04322$. Dado el valor de p-valor, se rechaza la hipótesis nula. Sigue que si hay correlación entre abril y mayo.
- **MARZO - MAYO:** $\rho = 0,1659105$ y p-valor $= 0,1659105$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y mayo.

Dados los resultados se esperaría que existiese correlación entre los meses, para poder elegir la atemporalidad. Si todos los meses se comportan igual, entonces no hay necesidad de estar pendiente de estos de manera particular. Sino más bien, serían confiables de manera general por comportarse de manera similar.

Continuando con los precios de venta, se representa la evolución de estos - en el mismo lapso de tiempo de 30 días - en la figura 3.9.
Si se aplica la misma metodología de la correlación se obtienen los siguientes valores,

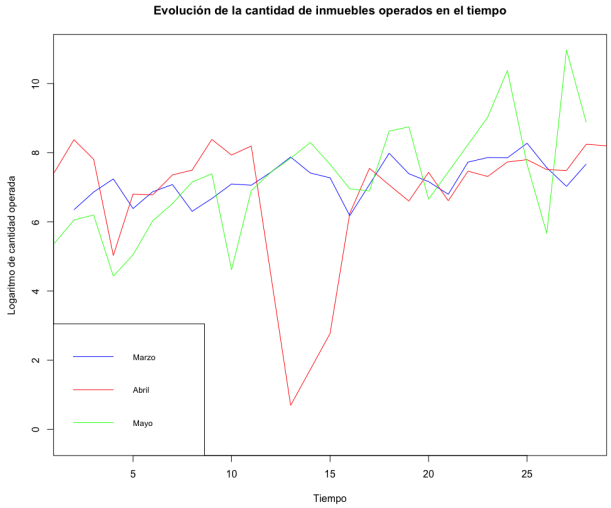

Figura 3.8: Evolución (en días mes) de `ln(cantidad operada de inmuebles)` en la base completa - Fuente: Elaboración Propia

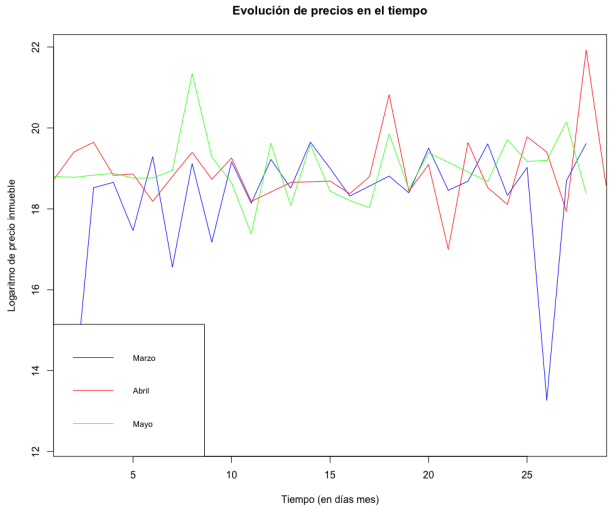

Figura 3.9: Evolución (en días mes) de `ln(promedio de precios de venta)` en la base completa - Fuente: Elaboración Propia

- **MARZO - ABRIL**: $\rho = -0,2224741$ y p-valor $= 0,2747$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y abril.
- **ABRIL - MAYO:** $\rho = -0,03106764$ y p-valor $= 0,8802$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre abril y mayo.
- **MARZO - MAYO:** $\rho = -0{,}2308834$ y p-valor $= 0{,}2565$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y mayo.

Nuevamente se evidencia la inexistencia de una relación entre los meses utilizados. Esto es un elemento en contra a la hora de tener que definir si escoger (o no) la atemporalidad como elemento característico de la base.
Se podría pensar que, si no funciona con los meses, debiera - por granularidad del problema

- funcionar “mejor” con los días de semana. No obstante, realizar la operación no es posible *a priori*.
Recordando la definición de correlación de Pearson, los vectores que contengan la distribución de la variable aleatoria deben tener *igual tamaño.* Esto no ocurre en los meses; y se visualiza en la siguiente tabla,

Diferencia en la cantidad de observaciones entre los meses pertenecientes a la base de datos							
Mes | Día	LUNES	MARTES	MIÉRCOLES	JUEVES	VIERNES	SÁBADO	DOMINGO
MARZO	7533	5253	5633	4809	3712	3737	2824
ABRIL	5803	13201	8426	3709	1409	3671	5941
MAYO	34913	10869	2841	11162	33289	8424	5850

Tabla 3.5: Tabla conteo de observaciones por día en los meses de análisis - Fuente: Elaboración Propia

lo importante son las columnas que formalizan cada día. Es evidente que ninguna de ellas coincide con su mes vecino. Lo que impide la creación de correlaciones de Pearson. Pero se puede arreglar formulando una tabla resumen que agrupe todos los valores correspondientes al mes y día que se tengan. Para orientar al lector, las primeras correlaciones se crearon a partir de una tabla resumen que tenía como criterio de agrupación los días de `fechascrap`. Comenzando desde el 02 de marzo del 2019, y terminando el 28 de mayo del 2019 (75 observaciones = 75 días). Es ahí entonces en donde se crearon vectores con el tamaño de cada mes, y se realizaron las correlaciones. En este caso, se procederá tal y como se muestra en el gráfico. Se utilizará como criterio de agrupación el día de semana. Con él se tendrán 21 observaciones en total, partisionadas de a 3 (un valor por cada mes analizado). Manifestando lo que se muestra en la figura 3.10. En síntesis, 3 meses en total, con el valor - logarítmico - de la cantidad de inmuebles transados en cada uno de los días de semana de los mismos.

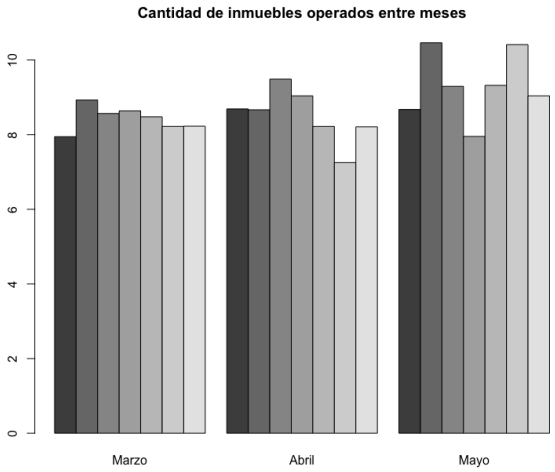

Figura 3.10: Evolución (en días semana) de `ln(cantidad operada de inmuebles)` por cada mes en la base completa - Fuente: Elaboración Propia

Re-calculando para los 3 meses de análisis se tiene,

- **MARZO - ABRIL**: $\rho = 0,6257929$ y p-valor $= 0,1328$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y abril.
- **ABRIL - MAYO:** $\rho = -0,03394431$ y p-valor $= 0,9424$. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre abril y mayo.
- **MARZO - MAYO:** $\rho = -0,1598702$ y p-valor = 0,7321. Dado el valor de p-valor, no se rechaza la hipótesis nula. Sigue que no hay correlación entre marzo y mayo.

Este resultado era de esperar. Por tanto, a modo de arreglo, debiera evaluarse en todas las características observables dentro de cada inmueble clasificado según tipo-operación como se vio anteriormente. Pues, técnicamente en este gráfico, se están mezclando inmuebles diferentes. Un ejemplo de ellos sería el par: inmueble comercial en arriendo con un inmueble de departamento en venta. Valores de distinta índole que terminan sesgando la muestra. Ideal sería considerar la correlación de Pearson en inmuebles clasificados por tipo-operación dentro de cada comuna. No obstante ello, existe un problema de costo de oportunidad. Clasificar por: operación (2) - tipo (4) - comunas (51) implica una separación de: $2 \times 4 \times 51 = 408$ gráficos. Además, surgen 2 posibles inconvenientes técnicos: ¿cuál debería ser el criterio para confiar en una correlación? ¿únicamente mayor a 0,5; o tomando una cota de 0,8 de forma adicional?. También está la construcción de la correlación. Como se expuso en su definición, las variables a utilizar *deben ser cuantitativas y continuas*. Por tanto, para que los 408 subclasificaciones funcionen - y así se pueda analizar sus correlaciones - todos deben tener la misma cantidad de `fechascrap` utilizadas, y en las mismas posiciones. Basta con que una fecha no exista en algún elemento del par evaluado, para que la correlación de Pearson no funcione.

Entonces, y en adición a la consideración de captura de shocks aleatorios dentro de la economía inmobiliaria más el peso de las variables a través del tiempo, se decide no trabajar con una base de datos caracterizada como temporal. Siendo definida de manera definitiva como una base de datos atemporal para el trabajo dentro de la presente memoria.
No obstante, el aprendizaje de máquinas puede predecir valores[12] cercanos a lo correcto dentro de juegos con shocks aleatorios, siempre y cuando se escoja el par correcto de función pérdida-métrica de error [7]. Finalmente se detalla al lector que predecir precios de venta a posterior pierde sentido por la estructura de compra-venta del sector inmobiliario.
Suponiendo la creación futura de inmuebles de interés en las cercanías de un dato ya localizado dentro de la base de datos, ya se sabría su valor cercano de venta por encontrarse "en verde". Un interesado ya tendría un valor referencial. Por tanto no existe la necesidad a priori de saber futuros precios, por la facilidad con que estos se pueden conseguir. Como último punto está el trabajo de los algoritmos de Machine Learning. En la consulta de un precio de venta se incorporan la localización (latitud y longitud) de su ubicación; pero el analista desconoce a priori cómo estas son tratadas. Por lo que referenciar una unidad de interés en el perímetro se dificulta por no se saber qué tan cercana esta debería estar para considerarla al momento de trazar una región factible de análisis. ¿Cuándo se considera vecino? ¿un elemento localizado en el extremo del perímetro comunal, o lisa llanamente en las cercanías de otra

[12]Considerando una adaptación a random shocks + status quo pro tempore. Prediciendo para un período posterior. Forecasting ML: comandos no existentes en "R" aún, por ser papers nacidos desde el 2017 en adelante.

comuna, afecta el valor de mercado de dicho elemento? Estas son preguntas sin resolver, y que el analista tiene por objetivo definir - entre otras cosas - para dar cabida a la predicción del valor de mercado en inmuebles futuros.

Capítulo 4

Evaluación

A continuación se presentarán los resultados obtenidos de la predicción de los modelos de regresión descritos con anterioridad. Más, se señala al lector que el desarrollo inicial del presente capítulo se encuentra visible en el apéndice C. En él podrá encontrar la metodología y los resultados previos realizados para llegar a la quinta etapa del presente capítulo.

En total serán 6 etapas. La primera mostrará a los algoritmos (sin meta-parámetros asociados) una tabla de 3 (métricas) x 8 (bases de información) que contendrán en cada una de las celdas, el valor promedio de todas las submuestras de cada métrica perteneciente a cada una de las bases de información. En la segunda etapa, los algoritmos (con parámetro) tendrán 3 tablas (métricas). En cada tabla se tendrán P filas (parámetros utilizados) x 8 columnas (bases de información) mostrando el valor promedio de la métrica referente a la tabla respectiva. En la tercera etapa se seleccionarán los mejores P (parámetros) de cada tabla, según lo esperado por cada métrica elegida. Estos criterios son: (1) valor R^2 más alto, (2) valor RMSE más bajo y (3) valor MAPE más bajo. Resumiendo así, en una única tabla (por algoritmo) con tamaño: 3 (métricas) x 8 columnas (bases de información) con los parámetros óptimos correspondientes (en caso de existir) a cada uno de los algoritmos.
En la cuarta etapa, se elegirá el mejor parámetro perteneciente a cada algoritmo, según cuál se repita más en cada una de las métricas correspondientes a la sub-base correspondiente. La quinta etapa es comparar cada uno de los algoritmos (pertenecientes a cada una de las bases) eligiendo el mejor con respecto al modelo de la empresa. Finalmente se reportará al lector cuáles algoritmos (de esta memoria) son mejores al algoritmo de la empresa con las métricas: R^2 y RMSE. Dicho eso, finalmente se verán los valores MAE correspondientes, y se calculará si la empresa gana (o pierde) con esto.

4.1. Resultados Obtenidos

4.1.1. Quinta Etapa

Utilizando la información provista en el apéndice C, se procederá en comparar cada uno de los algoritmos optimizados con sus mejores parámetros contra el modelo de la empresa Generalized Boosted Regression, utilizando los datos contenidos en cada una de las sub-bases diseñadas. Serán 4 algoritmos[1] aplicados en cada una de las 8 bases de testeo con información extraída a inicios del análisis computacional; con 20 % de la información total de la base (limpiada) de datos otorgados por la empresa.

Los resultados por base se presenta a continuación.

Base 01	**RF**	**GB**
R2	0.295	0.296
RMSE	1.037	1.079
MAPE	0.046	0.045

Tabla 4.1: Tabla resultado de aplicación etapa 06 base 01 - Random Forest VS Generalized Boosted Regression

Observando las métricas R^2 y RMSE se tiene que no hay conclusión posible sobre qué algoritmo es mejor. No obstante ello, se recuerda al lector, que la base 01 ha sido **eliminada de análisis.**

Base 02	**LR**	**GB**
R2	0.785	0.681
RMSE	0.357	0.496
MAPE	0.017	0.014

Tabla 4.2: Tabla resultado de aplicación etapa 06 base 02 - Linear Regression VS Generalized Boosted Regression

Replicando el procedimiento anterior, se procede en confirmar que el algoritmo óptimo **para la base 02 - referente a casas en arriendo - es Linear Regression sin parámetros.**

Continuando, se observa que el algoritmo óptimo **para la base 03 - representando a los departamentos en arriendo - es Random Forest con node size = 7.**

Siguiendo con el proceso, se tiene que el algoritmo óptimo para analizar **la base 04 - mostrando a las oficinas en arriendo - es Random Forest con node size = 6.**

[1]Linear Regression (1) + KNN (1) + Decision Tree (1) + Random Forest (1) + SVM (1)

Base 03	**RF**	**GB**
R2	0.863	0.862
RMSE	0.222	0.224
MAPE	0.009	0.009

Tabla 4.3: Tabla resultado de aplicación etapa 06 base 03 - Random Forest VS Generalized Boosted Regression

Base 04	**RF**	**GB**
R2	0.825	0.807
RMSE	0.183	0.192
MAPE	0.007	0.007

Tabla 4.4: Tabla resultado de aplicación etapa 06 base 04 - Random Forest VS Generalized Boosted Regression

Base 05	**LR**	**GB**
R2	0.022	0.382
RMSE	1.223	1.279
MAPE	0.045	0.035

Tabla 4.5: Tabla resultado de aplicación etapa 06 base 05 - Linear Regression VS Generalized Boosted Regression

Con el algoritmo óptimo para la **base 05 - representando a los bienes comerciales en venta - no es posible de definir.** Esto pues la discrepancia generada entre el resultado métrico de R^2 y RMSE no finaliza con un candidato único para considerar. No obstante, se reitera que **la base 05 a sido eliminada de todo análisis.**

Base 06	**RF**	**GB**
R2	0.860	0.863
RMSE	0.312	0.308
MAPE	0.008	0.008

Tabla 4.6: Tabla resultado de aplicación etapa 06 base 06 - Random Forest VS Generalized Boosted Regression

Procediendo con encontrar el algoritmo óptimo para la **base 06 - representando a las casas en venta - se encuentra a el algoritmo de la empresa como el más indicado, Generalized Boosted Regression.**

Base 07	**RF**	**GB**
R2	0.860	0.864
RMSE	0.312	0.308
MAPE	0.008	0.008

Tabla 4.7: Tabla resultado de aplicación etapa 06 base 07 - Random Forest VS Generalized Boosted Regression

El algoritmo óptimo para la **base 07 - representando a los departamentos en venta - es nuevamente el algoritmo de la empresa: Generalized Boosted Regression.** La diferencia es pequeña, por tanto se puede evaluar aceptar las dos.

Base 08	**RF**	**GB**
R2	0.875	0.866
RMSE	0.256	0.2698
MAPE	0.006	0.007

Tabla 4.8: Tabla resultado de aplicación etapa 06 base 08 - Random Forest VS Generalized Boosted Regression

Finalmente **para la base 08 - representando a las oficinas - el mejor algoritmo es Random Forest con node size = 7.** Concluyendo que la presente memoria le otorga a la empresa una mejora en 4 (de las 6) bases totales. 2 de ellas mantienen el uso del algoritmo de la empresa. El detalle a continuación,
Si se desea obtener resultados similares a los vistos en la presente sección, se recomienda

Base	Representación	Algoritmo	Parámetro Óptimo
1	Arriendo Comercial	ELIMINADA	
2	Arriendo Casa	Linear Regression	NaN
3	Arriendo Departamento	Random Forest	node size = 7
4	Arriendo Oficina	Random Forest	node size = 6
5	Venta Comercial	ELIMINADA	
6	Venta Casa	Generalized Boosted Regression	Variables default
7	Venta Departamento	Generalized Boosted Regression	Variables default
8	Venta Oficina	Random Forest	node size = 7

Tabla 4.9: Tabla resultado final de algoritmos seleccionados con parámetros óptimos

al lector utilizar el paso a paso implementado; y mostrado a inicios de este capítulo de Evaluación. Se debe, de forma obligada, considerar las variables: `PRECIOS; DORMITORIOS; BAÑOS; METROS MÍNIMOS; METROS MÁXIMOS; ESTACIONAMIENTO` y `COMUNA` para el caso de Linear Regression. Y para los demás, en adición de los anteriores, las variables: `LATITUD` y `LONGITUD.`

Tal y como se expuso en la subsección 1.1.2 de Marco Institucional, se procede en calcular el MAE de todos los algoritmos escogidos anteriormente junto al perteneciente al algoritmo de la empresa Generalized Boosted Regression en cada una de las sub-bases de información. Teniendo en consideración que se eliminaron las sub-bases número 1 y 5, los cálculos de MAE son los siguientes

Base	Algoritmo	Parámetro Óptimo	MAE GBoost Regression	MAE Algortimo Seleccionado	Diferencia
1	SUB-BASE DE DATOS ELIMINADA				
2	LR	NaN	0.205	0.242	**-0.037**
3	RF	node size = 7	0.124	0.121	**0.003**
4	RF	node size = 6	0.099	0.089	**0.009**
5	SUB-BASE DE DATOS ELIMINADA				
6	Gboosted Regression	Default	0.162	0.162	**0.0**
7	Gboosted Regression	Default	0.127	0.127	**0.0**
8	RF	node size = 7	0.119	0.093	**0.026**

Tabla 4.10: Tabla esultado de modelos finales VS modelo empresa en métrica MAE

De la definición de las métricas utilizadas, se extrae que la penalización de valores outliers será mucho mayor en R^2 y en RMSE, que en MAE. Lo que sugiere que el impacto de outliers en MAE es mucho menor. Sigue que R^2 y RMSE son más sensibles a valores atípicos que el MAE.
Al ya haber normalizado los valores de precio de venta exante al análisis con la incorporación de logaritmo natural, más la clasificación de tipo-operación creando las 8 sub-bases; se hace confusa la sugerente existencia de outliers. Sin importar ello, se confiesa al lector la mantención del R^2 y RMSE como aquellas métricas que definen los mejores modelos entre los diseñados en la presente memoria y los usados por la empresa.

Observando la tabla 4.10 destaca que, en la base número 02, el mejor modelo elegido es mejor al modelo de la empresa; pero es a su vez, es peor en términos de MAE. Haciendo la conexión con lo explicado con anterioridad, se puede suponer que existieron *pocos* outliers dentro de la base. Pero aún así, la diferencia tampoco es significativa como para amputar el resultado. Se reconoce la existencia de esta anomalía, por tanto se procede en no considerarla como parte del cálculo final de ganancia neta por parte de la empresa. Finalmente se recomienda a la empresa trabajar con el modelo sugerido en la presente memoria cuando no haya una dispersión significativa en los datos analizados.

Por último se señala al lector que, al tener los mejores parámetros de los modelos; al momento de proceder a calcular los errores, se debe recordar el uso de la función logaritmo natural `ln()`. Pues, en vista de que ya se ha normalizado, se tiene el deber de volver a la escala original. Es por ello que a los datos de testeo y a los datos predichos se les debe operar con la exponencial `exp()` exante las métricas; para volver a sus unidades iniciales y - recién ahí - calcular las respectivas métricas de interés.

4.1.2. Ganancia Monetaria de Memoria

Como ya se explicó con anterioridad, la ganancia total momentaria (y semestral[2]) obtenida con los mejores modelos de la presente memoria, estará dada entonces por

$$G_{nueva-empresa} = 1.000.000.000 \times 0,25 \times \left[0,2 + \sum_{I} \Delta_{+}\right] G_{memoria} = 1.000.000.000 \times 0,25 \times \sum_{I} \Delta_{+}$$

En donde se recuerda que la suma del factor Δ_+ es la diferencia positiva total-acumulada de ganancia (medida en MAE) de cada una de las categorías finales de los inmuebles. Si la suma total es equivalente a 0,01; entonces la empresa habrá aumentado en 1 % su renta obtenida de las consultas de precio de venta por parte de sus clientes.
Al sumar las diferencias que reportarán outcome positivo[3] (en verde), se obtiene lo siguiente

$$\sum_{I} \Delta_{+} = 0,00335917 + 0,00967817 + 0,02614297 = 0,03918031 \sim 0,039 = 3,9\,\%$$

Traducido en términos financieros: *"por cada X % que la memoria logre disminuir el MAE general, se sumará ese Δ_+ a la rentabilidad del cliente.* Lo cual, tal cual se muestra en la ecuación de ganancia (adicional) otorgada por la memoria $G_{memoria}$, se otorga un beneficio tangible-monetario equivalente a

$$1.000.000.000 \times 0,25 \times 0,039 = 9.750.000$$

Por tanto, la presente memoria otorga $9.750.000 semestrales de beneficio financiero a la empresa durante el primer período de evaluación. No se consideran proyecciones con algún uso de un interés compuesto semestral. ■

[2]La empresa percibe aproximadamente $1.000.000.000 de ganancia semestral bruta.
[3]No se considerará el valor negativo por lo expuesto a fines de la sub-sección anterior.

Conclusión

4.2. Cumplimiento y Resultados

Como se manifestó en un comienzo, es importante estimar los precios venta de los inmuebles para poder dar al cliente solicitante de dicho elemento, una referencia que le permita decidir si comprar-vender-arrendar el inmueble en cuestión. Esto sea como oferente o demandante, teniendo en consideración maximizar su utilidad individual. De esta forma el cliente gana una referencia para distinguir un valor cercano al precio de tasación, y la empresa gana clientela fidedigna con la constante solicitud de predicciones.
Dados estos puntos, se plantean las siguientes preguntas de investigación: ¿cuál será el valor del inmueble solicitado, teniendo en consideración que este posee múltiples factores y características que pueden influir en ello? ¿qué modelo predictivo se debe utilizar en cada una de las sub-categorías[4] de los inmuebles de la base, para predecir de la manera más ajustada posible sin entrar en riesgo de underfitting o overfitting?

En el desarrollo de este trabajo se aborda el tema de cuál es el mejor modelo para estimar los precios de venta en cada categoría de diferenciación de los inmuebles dentro de la base. Esta diferenciación - de tipo-operación - se concretó en la creación de 8 sub-bases de datos analizadas a lo largo de la investigación. Estas fueron: (1) arriendo-comercial, (2) arriendo-casa, (3) arriendo-departamento, (4) arriendo-oficina, (5) venta-comercial, (6) venta-casa, (7) venta-departamento, (8) venta-oficina. En cada una de ellas se aplicó la misma metodología de *Data Preparation, Modeling* y *Evaluation.*

Los 192.179 datos de la Región Metropolitana, que datan de marzo del 2019 a mayo del mismo año; pertenecen al registro de la página web: *www.portalinmobiliario.com*. En ella se encuentran las características usadas como variables independientes. De las 18 columnas establecidas inicialmente en la base de datos de la empresa, se terminó trabajando con 62. Esto producto, entre otras cosas, de la creación de 52 variables dummies que representan las comunas reportadas en cada una de las observaciones.

Se testearon 7 modelos en cada una de los 8 bases. Se descartaron 2: Artificial Neural Networking y Kernel Smoothing Regression, por alto costo computacional e incertidumbre sobre cuando se tendría el resultado. Con las 5 restantes, salvo en Linear Regression, se probaron 5 valores de complejidades para Decission Tree, 5 valores de tamaño de nodo para Random Forest, 3 valores de costo y factor gamma para Support Vector Machine, y 10 valores de

[4]Creadas a partir de la segmentación combinadas de las características: tipo y operación de los inmuebles.

vecinos cercanos (de 3 a 12) para K-nearest neighbor. En cada uno de ellos se optimizó buscando los mejores parámetros según los criterios correspondientes a las métricas R^2, RMSE y MAPE. Para la primera métrica el mejor algoritmo (con parámetro optimizado) tendría el mayor valor del conjunto; mientras que para los otros se buscaría el valor mínimo de la misma forma. Aún así, se reportaron fallas de consistencia con la última métrica, decidiendo eliminarla de todo análisis posterior.

En virtud de lo detallados en la sub-sección 1.3.2., se reportan las siguientes conclusiones respecto a los objetivos específicos planteados.

I) **Corrección de Anomalías:** En la base de datos otorgada por la empresa se corrigieron anomalías tales como:

- **Valores outliers:** En algunos casos, dentro de las características observables, se presentaban valores contradictorios. Tales como que los metros mínimos (construidos) eran más grande que los metros máximos (del terreno)[5]. Se corrigió con la permutación de valores.
- **Valores de signo:** Se presentó el escenario donde inmuebles poseían metros construidos negativos. Se corrigió transformando todos los valores a positivo.
- **Inmuebles sin región-comuna asociada:** Como la presente memoria era sobre la Región Metropolitana, estuvo la duda permanente sobre el origen de dichos inmuebles no identificados. Se terminó por no considerar estos valores.

II) **Elección de atributos:** De un inicio se eliminaron 3 variables irrelevantes sin información, con la ubicación se extrajo la comuna; y en cada una de ellas una variable binaria asociada. Finalmente se trabajó con las variables independientes: `DORMITORIOS; BAÑOS; METROS MÍNIMOS; METROS MÁXIMOS; ESTACIONAMIENTO y COMUNA`, y la variable dependiente `PRECIO`.

III) **Uso de bases correctas:** Teniendo en consideración que las métricas utilizadas para evaluar si los algoritmos eran útiles, fueron: R^2 y RMSE, al momento de calcular las regresiones se reportaron valores atípicos que salían de los margenes de aceptabilidad para el uso confiado de las siguientes sub-bases de información: venta y arriendo de inmuebles comerciales. Finalmente se eliminaron de todo análisis.

IV) **Construcción y evaluación:** De los 5 modelos seleccionados para trabajar, se terminó por tener únicamente 2 de ellos como mejores candidatos para recomendarle a la empresa. Por el lado métrico estuvieron: Linear Regression y Random Forest; pero únicamente Random Forest se mantuvo de forma segura ante las evaluaciones con la métrica MAE. Cabe señalar que el modelo de la empresa, Generalized Boosted Regression, también obtuvo buen rendimiento.
Lo anterior permite concluir que:

- **Random Forest** es el mejor modelo en el contexto de predicción de precios de venta dentro del mercado arriendo inmobiliario para las oficinas y para los departamentos. A su vez, destaca como el mejor en el mercado de venta de oficinas.

[5]Esta relación debe ser menor o igual.

Todo dentro de la Región Metropolitana de la República de Chile.

- **Generalized Boosted Regression** es el mejor modelo en el contexto de predicción de precios de venta dentro del mercado de venta en casas y departamentos dentro de la Región Metropolitana de la República de Chile.

V) **Mejor modelo:** Si bien - en teoría - ambos algoritmos "empataron" dentro de este análisis; es rescatable el hecho de que Random Forest, con sus parámetros optimizados, haya logrado superar al modelo de la empresa.
Teóricamente funcionan de forma distinta. Generalized Boosted Regression optimiza con el método de gradiente, y Random Forest predice promedios con sus nodos, y termina por sacar un cálculo final con el promedio general. Sin importar ello, Random Forest tiene la ventaja que no entra en overfitting si a este se le aumenta la cantidad de árboles. En palabras sencillas *pueden aumentar tantas veces la cantidad de árboles y convergerá a la misma predicción* [3]. Esto es ventajoso pues en el modelo de la empresa si se puede entrar en overfitting. Random Forest tiene la cualidad subyacente que se intentó realizar a principios de la presente memoria. Reduce la varianza de los resultados ya que es un modelo que calcula el promedio de muchos árboles de regresión simples y, por ende, se reduce la varianza. Finalmente este modelo puede reportar la importancia de las variables utilizadas, dando posibilidad al analista de ver qué variables usar-mantener-eliminar dentro de una eventual predicción [3].

VI) **Utilidad financiera de la memoria:** Dado el monto del último ingreso semestral reportado por la empresa; el cambio de metodología, y posterior implementación de modelos, ayudará en aumentar en un 0,975 % las ganancias extraídas de la rentabilidad del cliente como parte del trato formalizado. Implicando en un ahorro (o obtención de ganancia) de $9.750.000 adicionales a los que ya opera. Esto sin considerar el crecimiento que la entidad posee a lo largo de los años; implicando automáticamente en un beneficio mayor de crecimiento a una tasa semestral.

4.3. Trabajo Futuro

En cuanto al trabajo futuro se reconoce la permanente existencia de variables exógenas que estarán alterando el precio de venta del inmueble en cuestión. Considerando esto como el factor más relevante a considerar dentro de un trabajo a futuro, se debe guardar registro de aquellos características que no debiesen estar incluidas en las tasaciones físicas, pero que son igualmente relevantes por ser una cooperación implícita; perfecta herramienta para predecir y adelantarse a los tasadores. Ejemplo de ello pueden ser las rutas de buses RED en los alrededores. No solo si hay tantos paraderos, o buses pasando a un radio definido de metros. Sino las conexiones que estos enlazan con otros lugares que no están considerados dentro de la tasación; y que sea además, tiempo bajo en ruta. Tomando: (1) un bus, y (2) unos tantos metros más allá; (3) invirtiendo tanta cantidad de tiempo, (4) se llega a tal lugar, en donde (5) el nivel promedio de la comuna le da cierta valorización. Estos elementos podrían perfectamente ser incluidos en una base de datos futura, pero que dado el factor temporal asociado, debiese considerar una base de tipo temporal. Esto pues se estaría midiendo la evolución temporal del inmuebles conforme a sus características propias, y del entorno.

Eventualmente, cuando los comandos computacionales existan, se podría modelar el precio de mercado de forma dinámica-temporal, según los horarios punta-valle; lugares de beneficio público que evolucionan en el tiempo[6], o algún otro fenómeno temporal a predecir que afecte de manera significativa el valor de mercado del inmueble. En caso de realizar esta extensión, se invita al lector investigar sobre Random Forest Regressors o la predicción de series de tiempo usando Recurrent Neural Network y modelos de Vector Autoregressive.
Finalmente se señala evaluar la posibilidad de combinar los modelos más aceptables con aquellos que posean menos estructura. Como ejemplo de ello se pueden enunciar: (1) tasas del hipotecario, (2) inmuebles del entorno; entre otros.

[6]Como lo pueden ser estaciones de metro, malls aledaños, uso de ciclovía, etc.

Bibliografía

[1] Cifuentes A. Albagli P. and P. Hempel. Desarrollo e implementación de un índice de precios del sector inmobiliario para la Región Metropolitana de Santiago. pages 2–21–25, Mar 2018.

[2] C.M. Bishop. Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, 2006.

[3] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, Oct 2001.

[4] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth, 1984.

[5] Encuesta Casen. Precios de arriendo por quintil de ingresos y región. 2017.

[6] P. Chapman, Janet M Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, C. Russell H. Shearer, and Richard Wirthl. The CRISP-DM reference model. 1:13–15, 2000.

[7] Philippe Coulombe, Maxime Leroux, D Miroslav Stevanović, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting ? . 2019.

[8] Instituto Nacional de Estadística. Sectores Económicos Índices de Ventas de Servicios - base promedio año 2014=100. (246):2, May 2019.

[9] Instituto Nacional de Estadística. Sectores Económicos Índices de Ventas de Servicios - base promedio año 2014=100. (247):2, Jun 2019.

[10] Departamento de Ingeniería Civil Matemática. Apunte de Cálculo Diferencial e Integral. pages 93–95, 2017.

[11] M. De Nadai and B. Lepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment. pages 1–2, Aug 2018.

[12] Ministerio de Trabajo y Previsión Social. Sistema de información laboral, 2019.

[13] Servicio Nacional del Consumidor. Ley del consumidor, 2019.

[14] Gergios Drakos. How to select the right evaluation metric for machine learning models: Part 1 regression metrics, 2020.

[15] R. O. Duda and P. E. Hart. Pattern classification and scene analysis. John Wiley, New York, 1973.

[16] Changyong Feng, Wang Hongyue, Naiji Lu, Tian Chen, Hua He, Ying Lu, and Xin Tu. Log-transformation and its implications for data analysis. Shanghai archives of psychiatry, 26:105–9, 04 2014.

[17] D. Freedman and P Diaconis. On the histogram as a density estimator: L_2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiet, 57:453–476, 1981.

[18] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189–1232, 2000.

[19] Mankiw N. Gregory. Principios de la Economía. South-Western, 10 edition, 2007.

[20] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. Ann. Math. Statist., 36(3):1049–1051, 06 1965.

[21] Omar Martínez. Análisis Económico. Astra Ediciones, México, 2014.

[22] W. S. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, (5):115–133, 1943. Reprinted in Neurocomputing: Foundations of Research, ed. by J. A. Anderson and E Rosenfeld. MIT Press 1988.

[23] Canela Meffert. Informe de práctica profesional 3. Santiago, 2019.

[24] Johanna Alvear Orellana. Árboles de decisión y random forest. Universidad de Cuenca, Cuenca, 2018.

[25] Savan Patel. Machine learning 101, 2017.

[26] Saed Sayad. Artificial neural network, 2010-2019.

[27] Saed Sayad. An introduction to data science, 2010-2019.

[28] Tavish Srivastava. Introduction to k-nearest neighbors: A powerful Machine Learning Algorithm (with implementation in Python R), 2018.

[29] Vladimir N. Vapnik. The nature of statistical learning theory. Springer-Verlag New York, Inc., 1995.

[30] Hal Varian. Microeconomía Intermedia: Un Enfoque Actual. 3 edition, 1994.

Apéndice A

Imágenes

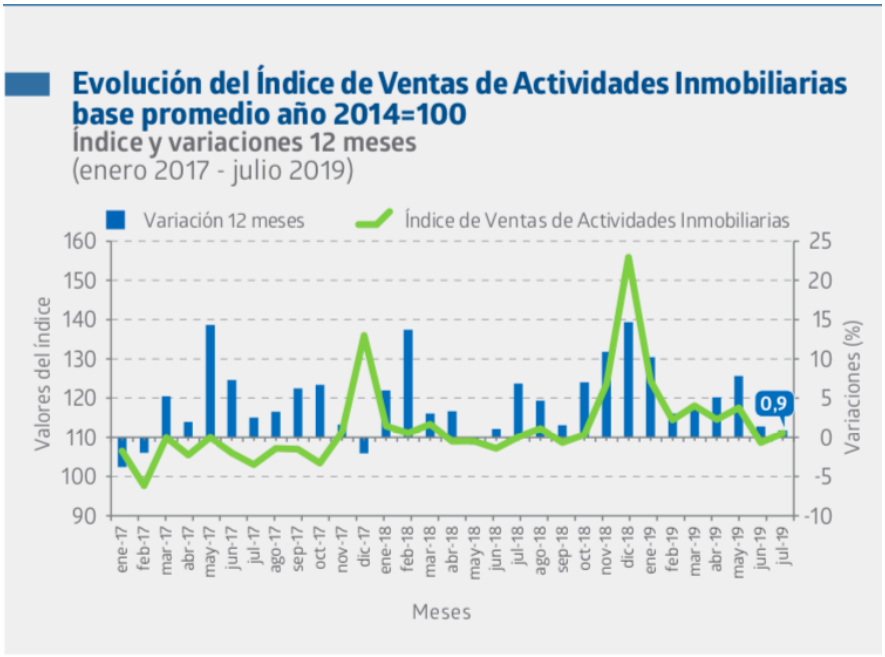

Figura A.1: Evolución del Índice de Ventas de Actividades Inmobiliarias Base Promedio año 2014=100 - Fuente: Instituto Nacional de Estadística

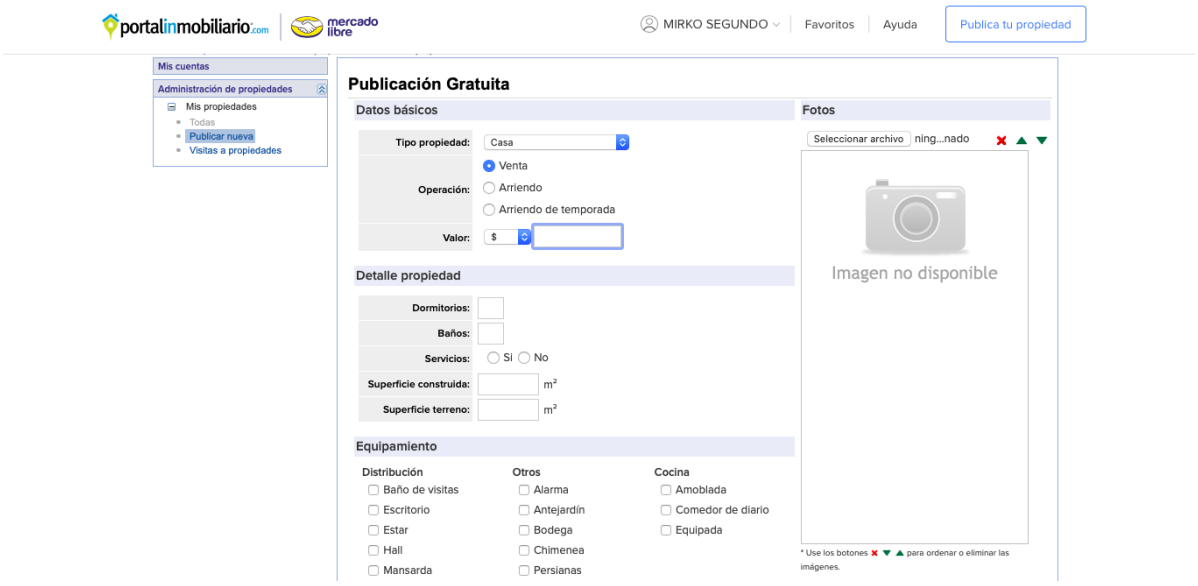

Figura A.2: Visualización portalinmobiliario de Usuario a Ofertar - Fuente: www.portalinmobiliario.com

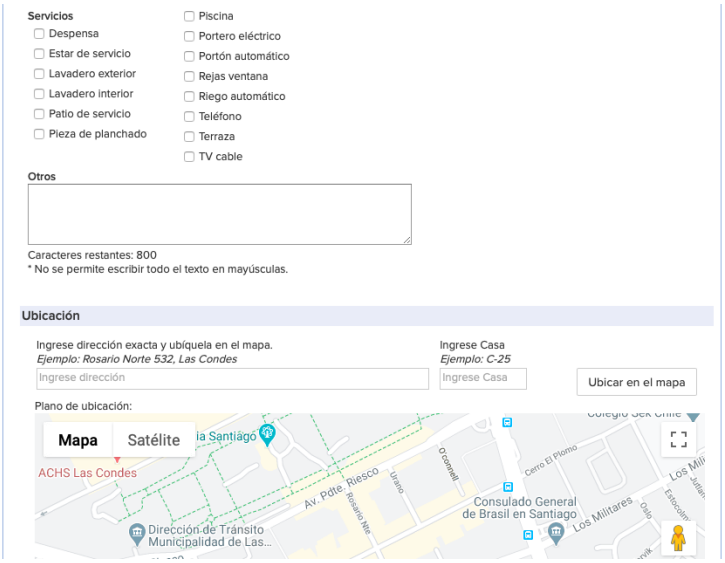

Figura A.3: Visualización portalinmobiliario de Usuario a Ofertar (continuación) - Fuente: www.portalinmobiliario.com

Apéndice B

Tablas

Métrica	B1	B2	B3	B4	B5	B6	B7	B8
R2	0.14	0.66	0.80	0.49	0.11	0.78	0.85	0.30
RMSE	1.27	0.49	0.27	0.34	1.47	0.39	0.29	0.35
MAPE	0.06	0.02	0.01	0.01	0.05	0.01	0.01	0.01

Tabla B.1: Tabla resultado de aplicación etapa 01 - Modelo: Regresión Lineal

Vecinos	B1	B2	B3	B4	B5	B6	B7	B8
$k = 3$	0.05	0.71	0.76	0.38	0.24	0.83	0.87	0.15
$k = 4$	0.10	0.72	0.78	0.42	0.22	0.83	0.87	0.20
$k = 5$	0.10	0.71	0.78	0.42	0.23	0.83	0.87	0.25
$k = 6$	0.12	0.71	0.79	0.42	0.23	0.84	0.88	0.26
$k = 7$	0.13	0.71	0.79	0.43	0.25	0.84	0.88	0.27
$k = 8$	0.13	0.71	0.80	0.43	0.24	0.84	0.88	0.27
$k = 9$	0.13	0.71	0.80	0.43	0.24	0.84	0.88	0.28
$k = 10$	0.13	0.71	0.80	0.44	0.22	0.84	0.88	0.28
$k = 11$	0.14	0.71	0.80	0.44	0.22	0.83	0.88	0.29
$k = 12$	0.14	0.71	0.80	0.44	0.21	0.83	0.88	0.28

Tabla B.2: Tabla resultado de aplicación etapa 02 - Modelo: KNN | Métrica: R2

Vecinos	B1	B2	B3	B4	B5	B6	B7	B8
$k = 3$	1.41	0.47	0.30	0.39	1.28	0.34	0.27	0.34
$k = 4$	1.38	0.46	0.29	0.39	1.31	0.34	0.26	0.34
$k = 5$	1.37	0.45	0.28	0.38	1.29	0.34	0.26	0.33
$k = 6$	1.36	0.45	0.28	0.38	1.29	0.34	0.26	0.33
$k = 7$	1.36	0.45	0.28	0.38	1.27	0.34	0.26	0.33
$k = 8$	1.36	0.45	0.28	0.37	1.28	0.34	0.26	0.32
$k = 9$	1.36	0.45	0.28	0.38	1.29	0.34	0.26	0.32
$k = 10$	1.37	0.45	0.28	0.38	1.28	0.34	0.26	0.32
$k = 11$	1.37	0.45	0.28	0.37	1.28	0.34	0.26	0.32
$k = 12$	1.37	0.45	0.28	0.37	1.28	0.34	0.26	0.32

Tabla B.3: Tabla resultado de aplicación etapa 02 - Modelo: KNN | Métrica: RMSE

Vecinos	B1	B2	B3	B4	B5	B6	B7	B8
$k = 3$	0.06	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 4$	0.06	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 5$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 6$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 7$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 8$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 9$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 10$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 11$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01
$k = 12$	0.07	0.02	0.01	0.01	0.04	0.01	0.01	0.01

Tabla B.4: Tabla resultado de aplicación etapa 02 - Modelo: KNN | Métrica: MAPE

Complejidad	B1	B2	B3	B4	B5	B6	B7	B8
$c = 0{,}015$	0.21	0.65	0.72	0.41	0.17	0.68	0.78	0.14
$c = 0{,}03$	0.18	0.60	0.67	0.39	0.16	0.61	0.74	0.13
$c = 0{,}0083$	0.22	0.67	0.73	0.42	0.17	0.72	0.81	0.16
$c = 0{,}006$	0.22	0.68	0.74	0.43	0.17	0.74	0.81	0.17
$c = 0{,}0005$	0.22	0.71	0.79	0.45	0.17	0.81	0.86	0.19

Tabla B.5: Tabla resultado de aplicación etapa 02 - Modelo: Árbol de Decisión | Métrica: R^2

Complejidad	B1	B2	B3	B4	B5	B6	B7	B8
$c = 0{,}015$	1.29	0.48	0.33	0.39	1.34	0.48	0.35	0.37
$c = 0{,}03$	1.31	0.51	0.35	0.39	1.35	0.52	0.37	0.37
$c = 0{,}0083$	1.28	0.47	0.32	0.38	1.34	0.44	0.33	0.36
$c = 0{,}006$	1.28	0.46	0.31	0.38	1.34	0.43	0.33	0.36
$c = 0{,}0005$	1.28	0.43	0.28	0.37	1.34	0.37	0.28	0.36

Tabla B.6: Tabla resultado de aplicación etapa 02 - Modelo: Árbol de Decisión | Métrica: RMSE

Complejidad	B1	B2	B3	B4	B5	B6	B7	B8
$c = 0{,}015$	0.06	0.02	0.01	0.01	0.05	0.02	0.01	0.01
$c = 0{,}03$	0.06	0.02	0.02	0.01	0.05	0.02	0.01	0.01
$c = 0{,}0083$	0.06	0.02	0.01	0.01	0.05	0.02	0.01	0.01
$c = 0{,}006$	0.06	0.02	0.01	0.01	0.05	0.01	0.01	0.01
$c = 0{,}0005$	0.06	0.02	0.01	0.01	0.05	0.01	0.01	0.01

Tabla B.7: Tabla resultado de aplicación etapa 02 - Modelo: Árbol de Decisión | Métrica: MAPE

Tamaño de Nodo	B1	B2	B3	B4	B5	B6	B7	B8
$nd\_size = 3$	0.42	0.76	0.84	0.54	0.47	0.89	0.91	0.49
$nd\_size = 4$	0.39	0.76	0.84	0.53	0.50	0.89	0.91	0.51
$nd\_size = 5$	0.39	0.76	0.84	0.55	0.47	0.89	0.91	0.50
$nd\_size = 6$	0.41	0.76	0.84	0.55	0.48	0.89	0.91	0.51
$nd\_size = 7$	0.40	0.77	0.84	0.54	0.47	0.89	0.91	0.53

Tabla B.8: Tabla resultado de aplicación etapa 02 - Modelo: Random Forest | Métrica: R^2

Tamaño de Nodo	B1	B2	B3	B4	B5	B6	B7	B8
$nd\_size = 3$	1.10	0.40	0.25	0.34	1.06	0.28	0.22	0.28
$nd\_size = 4$	1.12	0.39	0.25	0.34	1.04	0.28	0.22	0.28
$nd\_size = 5$	1.13	0.40	0.25	0.34	1.06	0.28	0.22	0.28
$nd\_size = 6$	1.11	0.40	0.25	0.34	1.05	0.28	0.22	0.28
$nd\_size = 7$	1.12	0.39	0.25	0.34	1.07	0.28	0.22	0.28

Tabla B.9: Tabla resultado de aplicación etapa 02 - Modelo: Random Forest | Métrica: RMSE

Tamaño de Nodo	B1	B2	B3	B4	B5	B6	B7	B8
$nd\_size = 3$	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01
$nd\_size = 4$	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01
$nd\_size = 5$	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01
$nd\_size = 6$	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01
$nd\_size = 7$	0.05	0.01	0.01	0.01	0.03	0.01	0.01	0.01

Tabla B.10: Tabla resultado de aplicación etapa 02 - Modelo: Random Forest | Métrica: MAPE

c	γ	B1	B2	B3	B4	B5	B6	B7	B8
4	0.1	0.1445101	0.7156901	0.7967837	0.4614989	0.3364349	0.8436168	0.8675215	0.4057889
	1	0.1388571	0.5575927	0.7303145	0.4213834	0.2788598	0.7809815	0.7772776	0.3374696
	10	0.08775411	0.30430832	0.55360338	0.32413940	0.04595533	0.59116356	0.60954924	0.23592234
8	0.1	0.1335222	0.7130932	0.7960914	0.4582752	0.3502564	0.8434816	0.8673945	0.3842789
	1	0.1399966	0.5503233	0.7282984	0.4144355	0.2748408	0.7762961	0.7770784	0.2930783
	10	0.10179390	0.29834840	0.54828097	0.31453571	0.04863496	0.58875586	0.60806703	0.22816663
16	0.1	0.1208112	0.7083354	0.7948911	0.4521331	0.3452298	0.8417442	0.8666613	0.3515848
	1	0.1380343	0.5429074	0.7248486	0.4001476	0.2634843	0.7724574	0.7763909	0.2672921
	10	0.09821535	0.28483230	0.53997245	0.30131561	0.04711830	0.58529299	0.60447583	0.21854709

Tabla B.11: Tabla resultado de aplicación etapa 02 - Modelo: SVM (parámetros variables) | Métrica: R2

c	γ	B1	B2	B3	B4	B5	B6	B7	B8
4	0.1	1.3409883	0.4332484	0.2767486	0.3690207	1.1902878	0.3305321	0.2688385	0.3089149
	1	1.3442221	0.5405367	0.3189925	0.3824509	1.2406474	0.3915814	0.3488416	0.3228890
	10	1.3847109	0.6779944	0.4106372	0.4128452	1.4300386	0.5347016	0.4621775	0.3436205
8	0.1	1.3492742	0.4352478	0.2772248	0.3701231	1.769829	0.3306761	0.2689639	0.3134925
	1	1.3414356	0.5448876	0.3201688	0.3847013	1.2435448	0.3957133	0.3490037	0.3313262
	10	1.3736193	0.6808438	0.4130707	0.4156735	1.4278440	0.5362596	0.4630527	0.3450682
16	0.1	1.3587010	0.4388638	0.2780479	0.3722119	1.1797526	0.3325131	0.2696993	0.3204598
	1	1.3413482	0.5493955	0.3222063	0.3893064	1.2531246	0.3991020	0.3495447	0.3362665
	10	1.3761997	0.6871527	0.4168471	0.4195975	1.4288963	0.5385334	0.4651742	0.3467418

Tabla B.12: Tabla resultado de aplicación etapa 02 - Modelo: SVM (parámetros variables) | Métrica: RMSE

c	γ	B1	B2	B3	B4	B5	B6	B7	B8
4	0.1	0.059329835	0.015734806	0.010228811	0.012327607	0.037052543	0.009591162	0.007652991	0.007237206
	1	0.062058633	0.022963897	0.012544947	0.013361270	0.042040261	0.011214583	0.009480216	0.007851744
	10	0.070766557	0.033911537	0.018335399	0.015957129	0.055786974	0.016901735	0.013336119	0.009214046
8	0.1	0.059111885	0.015825526	0.010240516	0.012387043	0.036404461	0.009569239	0.007648103	0.007336885
	1	0.061843951	0.023224567	0.012597010	0.013468825	0.042240428	0.011272886	0.009500596	0.008098117
	10	0.070542704	0.034038883	0.018526810	0.016122226	0.055764657	0.016953577	0.013424699	0.009315577
16	0.1	0.059600549	0.016010163	0.010271362	0.012492497	0.036297455	0.009592085	0.007661056	0.007471301
	1	0.061420449	0.023516516	0.012702323	0.013648886	0.042586615	0.011373987	0.009543726	0.008304027
	10	0.070710310	0.034272399	0.018900584	0.016370437	0.055793573	0.017037601	0.013639216	0.009383087

Tabla B.13: Tabla resultado de aplicación etapa 02 - Modelo: SVM (parámetros variables) | Métrica: MAPE

Parámetro	B1	B2	B3	B4	B5	B6	B7	B8
k	8	11	12	12	7	6	9	9

Tabla B.14: Tabla resultado de aplicación etapa 04 - Modelo: KNN | Parámetro óptimo por base

Parámetro	B1	B2	B3	B4	B5	B6	B7	B8
`cp`	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Tabla B.15: Tabla resultado de aplicación etapa 04 - Modelo: Decision Tree | Parámetro óptimo por base

Parámetro	B1	B2	B3	B4	B5	B6	B7	B8
`nodesize`	3	7	7	6	4	3	3	7

Tabla B.16: Tabla resultado de aplicación etapa 04 - Modelo: Random Forest | Parámetro óptimo por base

Parámetro Fijo	Parámetro Móvil	B1	B2	B3	B4	B5	B6	B7	B8
$\gamma = 0{,}1$	costo	4	4	4	4	4	4	4	4
$\gamma = 1$		8	4	4	4	4	4	4	4
$\gamma = 10$		4	4	4	4	4	4	4	4

Tabla B.17: Tabla resultado de aplicación etapa 04 - Modelo: SVM | Parámetro fijo: `gamma`

Parámetro Fijo	Parámetro Móvil	B1	B2	B3	B4	B5	B6	B7	B8
costo = 4	gamma	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
costo = 8		1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
costo = 16		1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Tabla B.18: Tabla resultado de aplicación etapa 04 - Modelo: SVM | Parámetro fijo: `costo`

Apéndice C

Desarrollo de Evaluación

C.1. Resultados Obtenidos

C.1.1. Primera Etapa

Esta etapa será exclusiva de aquellos modelos que esencialmente no poseen meta-parámetros a calibrar. Como lo es el caso de Regresión Lineal Multivariada.

• **Linear Regression:**

★ **Discusión:** Observando la Tabla B.1, y usando los criterios de selección señalados en la sub-sección 1.3.5., se evidencian diferencias significativas entre los resultados de cada una de las bases. En particular, el modelo resulta ser muy útil para predecir bases que contengan inmuebles del tipo[1] casa, y tipo[2] departamento. No así para los tipo[3] comerciales en donde evidencia una incorrecta predicción. Finalmente para las oficinas no es posible aseverar nada, puesto que los valores no son concluyentes.

C.1.2. Segunda Etapa

Siguiendo la estructura de la subsección anterior, se debe recordar que lo que se expondrá a continuación hace referencia a aquellos modelos que tuvieron dos o más parámetros a utilizar. Las tablas informativas[4] correspondientes a cada una de las métricas a utilizar serán reportadas en el mismo punto de discusión.

• **K-Nearest Neighbor:**

[1]Bases: B2 (arriendo) + B6 (venta).
[2]Bases: B3 (arriendo) + B7 (venta).
[3]Bases: B1 (arriendo) + B4 (venta).
[4]Como criterio numérico se dirá que una predicción es aceptable cuando la métrica $R^2 \sim 0{,}7$ (o mayor) + RMSE $< 0{,}4$. Para todo caso opuesto, será considerado no aceptable.

★ **Discusión:** La secuencia creciente de los vecinos k dentro del análisis, trata de responder a la pregunta: ¿cuál debe ser el valor k óptimo de vecinos elegidos, de modo que el R^2 sea máximo y las métricas RMSE y MAPE sean mínimas dentro de dicha secuencia respectiva en cada una de las bases? Pues, en esta oportunidad, dados los criterios de selección, es posible destacar las diferencias de magnitud (promedio) entre cada una de las bases en las métricas: R^2 y RMSE. En el caso de la tabla MAPE (tabla B.4), no se evidencian diferencias significativas. En las bases: B3, B4, B6, B7 y B8 se presentan niveles de los cuales se podría desconfiar, por sugerir una predicción sobre-ajustada. Un MAPE = 0.001 debe ser considerado una advertencia para el analista. Representando posiblemente una mala elección en el modelo de predicción utilizado.
No se cuestiona la calidad de información contenida en las bases, dado el comportamiento sistemático que poseen estas en las secuencias realizadas.

Considerando la métrica de R^2 (tabla B.2) como aquella primera preferencia, se observa que el modelo es muy útil para predecir bases que contengan inmuebles del tipo[5] casa, y tipo[6] departamento. No así para los tipo[7] comerciales en donde evidencia una incorrecta predicción. Finalmente para las oficinas[8] no es posible aseverar nada, puesto que los valores no son concluyentes.

Para la métrica de RMSE (tabla B.3), el escenario no es tan distinto. Se puede distinguir con claridad cuáles predicciones son más o menos ajustadas. Se reitera nuevamente el escenario en donde los inmuebles tipo comerciales poseen una predicción con bajo ajuste. Mientras que todas las demás categorías pasan a tener predicciones aceptables.

★ **Discusión General:** En virtud de las métricas anteriormente mencionadas, es claro ver que la métrica de MAPE está fallando por presagiar un posible sobre-ajuste. Por tanto, se deben considerar las otras métricas para dar una discusión general fidedigna del modelo.
Al observar los resultados de las métricas R^2 y RMSE se concluye que el modelo KNN es un predictor aceptable. Se postula de antemano que la dupla de métricas R^2 y RMSE arrojará la misma cantidad de vecinos óptimos en cada una de las bases. Se piensa que a futuro, dichas cantidades no tienen por qué ser iguales en todas las bases, por ser cantidades y patrones distintos. De esta forma, se permite una flexibilidad en el número de vecinos en cada una de estas.

• **Decision Tree:**

★ **Discusión:** Teniendo en cuenta la definición de parámetro de complejidad λ expuesto al final de la sub-sub-sección de Árboles de Regresión en sub-sección de Árboles de Decisión de la sección de Modelamiento en el capítulo de Marco Teórico, se tiene que este funciona como parámetro decisivo en el criterio de poda dentro del algoritmo. Determinando el trade-off entre minimizar el residuo cuadrático y la cantidad de nodos. Observando la métrica R^2 (tabla B.5), es posible ver similitudes con los resultados obtenidos en el modelo de regresión lineal y KNN. El modelo arroja predicciones aceptables para casas y departamentos. En contraste de los bienes comerciales y oficinas. Resaltando que las oficinas a venta, poseen peor predicción que las oficinas de arriendo;

[5]Bases: B2 (arriendo) + B6 (venta).
[6]Bases: B3 (arriendo) + B7 (venta).
[7]Bases: B1 (arriendo) + B5 (venta).
[8]Bases: B4 (arriendo) + B8 (venta).

en este último caso.

Finalmente, vislumbrando los valores de la tabla, podemos concluir que el modelo realiza predicciones aceptables para las casas y departamentos. Tanto en venta como en arriendo. No es posible replicar dicho argumento con los inmuebles comerciales, y de oficina. Siendo esto último peor en venta, que arriendo.

En lo que respecta a la métrica MAPE (tabla B.7) nuevamente presenta valores que incitan a dudar de la efectividad de esta métrica. Más detalladamente porque posee valores como 0,01 y 0,02 que simbolizan un posible sobre-ajuste. Considerando ello, se procederá como los algoritmos anteriores en donde se omite esta métrica como criterio de evaluación por incitar a la duda sobre la efectividad del algoritmo. Cabe señalar que se repite el patrón de la métrica anterior sobre los inmuebles de tipo comerciales.

Finalmente con la métrica RMSE (tabla B.6) se pueden utilizar las mismas conclusiones observadas en la métrica R^2, con la única salvedad sobre los inmuebles de oficina. Estos últimos poseen predicción aceptable considerando RMSE.

★ **Discusión General:** Se observan diferencias entre bases, y dentro de ellas, al cambiar los valores del parámetro de complejidad. Ello permite conjeturar: *"pequeñas variaciones en los valores de complejidad afectan notoriamente la estimación de las predicciones."* Más, estas se mantienen dentro de un intervalo razonable. Es decir, las fluctuaciones en los valores internos de cada base se mantienen cercanos al primer decimal del valor promedio. Permitiendo concluir sobre una estimación de base completa, sin distinción de complejidad.

Nuevamente se presenta la dupla R^2-RMSE como métricas coincidentes en la gran mayoría de los resultados. Implicando en que se deberían utilizar las mismas complejidades en todas las bases. Si comparamos las diferencias entre el presente modelo con el anterior (en ambas métricas), se notan leves diferencias entre los valores (promedio) arrojados. Diferencias mínimas que serán manifestadas al final de la presente etapa.

• **Random Forest:**

★ **Discusión:** El "node size" se define como aquel tamaño mínimo de nodos terminales. Establecer este número más grande hace que se cultiven árboles más pequeños, tomando menos tiempo. Considerando que el algoritmo necesita especificar la cantidad de árboles a generar, resulta muy conveniente variar este parámetro. A mayor cantidad de árboles, mayor cantidad de tiempo; por el tiempo computacional demandado en el procesamiento.

Observando la métrica R^2 (tabla B.8), se notan los mismos comportamientos correspondientes a los algoritmos anteriores en cuanto a predicciones aceptables. Nuevamente - con Random Forest - es útil predecir valores de casas y departamentos. Tanto en venta, como en arriendo. En cambio, no es recomendable nuevamente predecir inmuebles comerciales. Estos no poseen un número que obligue al analista rechazar esta idea por completo, ya que posee valores de R^2 dentro del intervalo $[0,4; 0,5]$ que pueden ser aceptables en distintos puntos de vista. Para un analista puede ser aceptable, pero teniendo en consideración los demás resultados; en la presente memoria no serán catalogados como tal. Finalmente se presenta una situación de poca certeza para las oficinas, pues poseen niveles aceptables según distintos puntos de vista (valores de arriendo y venta cercanos al $0,5$) pero que no dan - a juicio del autor - suficiente confianza para una

afirmación. Se deja a criterio del lector si aceptar, o no, estos valores.

En la métrica MAPE (tabla B.10) se presentan nuevamente valores monótonos dentro, y entre, las bases. En donde además, los datos que son distintos, siguen siendo poco creíbles en el contexto de crítica de selección. Por tanto, esta métrica no sirve para este algoritmo en particular.

Sobre la métrica de RMSE (tabla B.9) se puede observar que las casas, departamentos y oficinas pueden ser bien predichas mediante regresión. Tanto en venta como arriendo, las predicciones son aceptables. Nuevamente no se puede concluir lo mismo para los inmuebles comerciales.

★ **Discusión General:** Si bien los números arrojados son altos - dando el presagio de un posible over-fitting - no debe ser de preocupación para el lector tales números. Como es un patrón común hasta el momento, no es problema del algoritmo per se. Sino más bien en la estructura de los datos analizados. Se recuerda al lector que se partisionó en 5 sub-bases de testeos diferentes (en cada una de las bases) para realizar las predicciones. Presagiando de antemano, que el analista no debe predecir valores correspondientes a los bienes comerciales en ningún caso, con de los algoritmos presentes en esta memoria.

En lo que respecta a la variación del parámetro (node size) se tiene que pequeñas variaciones no afectan significativamente los resultados de las métricas. Más, estos se mantienen técnicamente constantes en cada una de las bases analizadas.

Por último, nuevamente la dupla R^2-RMSE son consistentes en sus resultados, manteniendo las características de los valores máximos en R^2 con las características de valores mínimos de RMSE en los resultados internos de cada base analizada.

• **Support Vector Machine:**

★ **Discusión:** A diferencia de los algoritmos anteriores, en SVM se optimizaron 2 parámetros pertenecientes a este. Los parámetros fueron "costo" y "gamma". Para definir el costo se debe tener presente la idea principal de la optimización en SVM. El algoritmo corta (con un hiperplano) las observaciones para definir cuáles entran dentro de una (o la otra) categoría. Espacios producidos por el corte del hiperplano. Entonces, el costo es aquel componente que da peso a las observaciones dentro del componente divisor (margen) para ver cuáles entran, o no, en como elementos útiles (en este caso) de predicción. A menor costo, las muestras dentro del margen se penalizan menos. Y viceversa. En tanto, el parámetro "gamma" define qué tan lejos llega la influencia de un set de entrenamiento en el ejercicio del SVM. Para valores pequeños, la influencia es grande. Y viceversa. En resumen, el inverso del radio de la influencia de las muestras seleccionadas por el modelo como vectores de soporte.

Se presentará a continuación una crítica de 3 niveles (diferentes valores gamma) manteniendo el parámetro costo como valor fijo. Si la tendencia es similar una con otra, se reportará de forma resumida sin necesidad de reiterar el mismo análisis. Finalmente, si hay similitudes muy parecidas con excepciones mínimas, nuevamente se reportará la relación con el análisis ya existente. Ahorrando en la caracterización.

Métrica R^2 (tabla B.11):
De manera general se puede partir concluyendo que a mayor valor de gamma, peor es el ajuste. A su vez, se puede aseverar nuevamente que los bienes comerciales no poseen - en ninguno de los 2 casos (venta y arriendo) - una predicción aceptable. Más, las

casas y departamentos son aquellos bienes en donde la predicción resulta con mejores resultados. Por último, para el caso de las oficinas nuevamente las predicciones deben ser analizadas en detalle. Los valores reportados indican que para el caso de oficina:

• **Evaluando para γ variable:** Estos convergen a $0,5$ en un solo caso (todos los costos con $\gamma = 0{,}1$). Para todo los demás valores de gamma, las oficinas poseen predicciones deficientes. Ergo, no poseen una predicción aceptable.

• **Evaluando por arriendo-venta:** Con este criterio las predicciones son mejores en arriendo que en venta.

Métrica MAPE (tabla B.13):

En este escenario, las métricas son significativamente distintas a las métrica. Pero siguen siendo bajas para - según el criterio de la presente memoria - ser aceptado. El máximo valor entre todo el conjunto de bases y variaciones paramétricas es 0,009. Por tanto los datos están repartidos en un intervalo de $[0; 0{,}009]$ reportando nuevamente valores para desconfiar. Finalmente la métrica de MAPE no brinda seguridad a la hora de evaluar criterios de selección con el uso de la métrica MAPE.

Métrica RMSE (tabla B.12):

Replicando la idea anterior se puede concluir que, a mayor valor de gamma, peor es el ajuste. A diferencia de la métrica R^2 ahora las oficinas proceden a tener un valor más confiable. Ergo, sus predicciones - según este criterio - pasan a ser aceptables. Incluso mejor que las casas en ambas operaciones. Comparten los departamentos como bienes que tienen predicciones razonables. Más, estas últimas son las mejores.

★ **Discusión General:** A estas alturas los datos empiezan a presentar un patrón claro en la relación R^2-RMSE. Dando el permiso para validar la hipótesis planteada también en KNN; y que es reiterada en los demás algoritmos: "con el uso de selección vía R^2-RMSE los parámetros óptimos serán iguales en ambas métricas".

• **Kernel Smoothing Regression:** Se reporta al lector que, después de dejar listos los preparativos para regresionar el modelo, este demandó mucho tiempo. En primera instancia, se utilizó una muestra representativa de 80 % de la base de datos de entrenamiento pertenecientes a cada base. Después de 5 días de análisis, el computador en donde se realizaban las predicciones se congeló. Después se utilizó el 40 % en donde - si bien el computador no se congeló - se invirtieron 4 días de procesamiento. En este caso se debe recordar que, de las 8 bases, cada una se partisionó en 5 sub-bases generadas por el cross-validation; transformándolo en un análisis general de 40 instancias (o iteraciones). El parámetro del algoritmo a optimizar es el bandwidth (ancho de banda); encargado de optimizar la varianza y sesgo en los resultados. Por tanto, el algoritmo - al igual que los otros - debía realizar 40 instancias para resultar en un ancho de banda óptimos. El punto está en que los algoritmos anteriores recibían únicamente una fórmula del tipo $Y \sim \sum_{j \in J} X_j$; ajustando sus regresiones solamente al resultado de dichas variables. En cambio, en el Kernel Smoothing Regression, se encarga de estimar *el peso total de la suma cada una de los valores* Y_{i} para un valor de predicción dado. Considerando ello en una base con 59 variables y una cantidad relativa de 375 a 27.846 datos (para arriendo), y 256 a 58.984 datos (para venta) se hace técnicamente costoso estimar para un resultado final. Repetir I veces con tantas características, termina saturando el cálculo. Es más, el creador del algoritmo - en el programa computacional R - es explicito en advertir al usuario que es recomendable usar bases de tamaño pequeño.

En el análisis computacional del problema se reporta que el último intento arrojo 3 resultados.

El primero de ellos demandó 301.700 segundos = 5028,333 minutos = 83,80556 horas = 3,491898 días; para el segundo resultado 396,5 segundos; y para el tercer resultado se tuvo un tiempo de 307.500 segundos = 5.125 minutos = 85,41667 horas = 3,559028 días. En total 7,050926 días de demanda para 3 resultados; de un total de 8 (bases).

Por lo anterior, se concluye que no es recomendable usar Kernel Smoothing Regression en contextos como este.

• **Artificial Neural Networking:** Al igual que el algoritmo anterior, Redes Neuronales posee un alto costo computacional. Pero algo más a diferenciar con Kernel Smoothing Regression, son sus componentes paramétricos a optimizar. Basándose en el capítulo de Metodología, se probó con optimizar la dupla de neuronas-capas en cada una de las iteraciones. Duplas que - con tan solo con un cambio de unidad en una de ellas - altera de manera considerable los resultados a obtener dentro de las predicciones. Marcando un cambio significativo en lo realizado con anterioridad con los algoritmos. Por permitir - en estos últimos - un análisis de sensibilidad robusto y con resultados cercanos, sin tanta varianza.

Redes Neuronales funciona como caja negra. Por tanto le es imposible al analista observar el funcionamiento paramétricos en la ejecución de la regresión.
Lo anterior trae un problema puesto que no es posible tampoco tener nociones "de lo que se está haciendo mal o correcto". Considerando ello, y lo anterior, no existen autores que hayan demostrado un mecanismo correcto a la hora de elegir una dupla perfecta según ciertas características del problema.

Sumado a lo anterior, está el tiempo de ejecución del comando para regresionar y tener resultados. No hay forma de medir el tiempo invertido en cada una de las bases. El único camino está en esperar el output con todos los resultados (de cada una de las bases) listo.
En este contexto, la prueba se cortó a los 6 días de corrido el algoritmo. Se tomó esa decisión por no tener herramientas de observación (de avance). Ergo, incertidumbre en saber cuándo terminaría.
Algunos analistas declaran que la elección de la cantidad de nodos, redes y kernel "son arte". Experiencia inexistente por parte del autor de esta memoria; por encontrarse en vías de postulación a su título profesional. En vista de que el escenario es incierto dado el trade-off entre costo computacional y calidad de predicción, se recomienda al lector no aplicar Redes Neuronales; en este contexto.

Finalmente todo se resume en las tablas de resultados generales de R^2 y RMSE mostradas a continuación,

Tal y como se observa en las tablas, las base 01 y 05 referentes a los inmuebles comerciales, posee los peores valores en todas las tablas señaladas. Tanto en arriendo como en ventas. Más aún, posee valores similares en todos los algoritmos utilizados[9] dando el permiso para postular la siguiente conjetura: *la prueba de bienes comerciales no tendrá un rendimiento tal, que permita a algún algoritmo poseer predicciones aceptables.* Desde ahora en adelante

[9]Salvo en Random Forest donde presenta una diferencia en comparación a los demás.

Algoritmo	Meta-Parámetro	B1	B2	B3	B4	B5	B6	B7	B8
KNN	k	0,1	0,7	0,8	0,4	0,2	0,8	0,9	0,3
DT	cp	0,2	0,7	0,7	0,4	0,2	0,7	0,8	0,2
RL	Null	0,1	0,7	0,8	0,5	0,1	0,8	0,9	0,3
	c = 4 | gamma = 0,1	0,1	0,7	0,8	0,5	0,3	0,8	0,9	0,4
	c = 4 | gamma = 1	0,1	0,6	0,7	04	0,3	0,8	0,8	0,3
	c = 4 | gamma = 10	0,1	0,3	0,6	0,3	0	0,6	0,6	0,2
SVM	c = 8 | gamma = 0,1	0,1	0,7	0,8	0,5	0,4	0,8	0,9	0,4
	c = 8 | gamma = 1	0,1	0,6	0,7	0,4	0,3	0,8	0,8	0,3
	c = 8 | gamma = 10	0,1	0,3	0,5	0,3	0	0,6	0,6	0,2
	c = 16 | gamma = 0,1	0,1	0,7	0,8	0,8	0,4	0,9	0,9	0,4
	c = 16 | gamma = 1	0,1	0,5	0,7	0,7	0,3	0,8	0,8	0,3
	c = 16 | gamma = 10	0,1	0,3	0,5	0,5	0	0,6	0,6	0,2
RF	node size	0,4	0,8	0,8	0,5	0,5	0,9	0,9	0,5

Tabla C.1: Tabla resultado de aplicación etapa 02 - Todos los modelos | Métricas: R^2

Algoritmo	Parámetro	B1	B2	B3	B4	B5	B6	B7	B8
KNN	k	1.37	0.45	0.28	0.38	1.28	0.34	0.26	0.33
DT	cp	1.28	0.47	0.33	0.39	1.34	0.48	0.35	0.36
RL	Null	1.27	0.49	0.27	0.34	1.47	0.39	0.29	0.35
	c = 4 | gamma = 0.1	1.34	0.43	0.28	0.37	1.19	0.42	0.37	0.74
	c = 4 | gamma = 1	1.34	0.54	0.32	0.38	1.24	0.8	0.8	0.3
	c = 4 | gamma = 10	1.38	0.68	0.41	0.41	1.43	0.53	0.46	0.34
SVM	c = 8 | gamma = 0.1	1.35	0.44	0.28	0.37	1.76	0.33	0.27	0.31
	c = 8 | gamma = 1	1.34	0.54	0.32	0.38	1.24	0.39	0.35	0.33
	c = 8 | gamma = 10	1.37	0.68	0.41	0.42	1.43	0.54	0.46	0.35
	c = 16 | gamma = 0.1	1.36	0.44	0.27	0.37	1.18	0.33	0.27	0.32
	c = 16 | gamma = 1	1.34	0.55	0.32	0.39	0.3	0.8	0.8	0.3
	c = 16 | gamma = 10	1.38	0.69	0.42	0.5	0	0.6	0.6	0.2
RF	node size	1.12	0.39	0.25	0.34	1.07	0.28	0.22	0.28

Tabla C.2: Tabla resultado de aplicación etapa 02 - Todos los modelos | Métricas: RMSE

se ignorará la base 01 en el análisis, dado lo anteriormente expuesto.

En lo que respecta a las base 02 y 06 representando a las casas, se evidencian cambios sustantivos al interior de Support Vector Machine en la zona de arriendo; teniendo valores (de R^2) que van desde $0,3$ a $0,7$, con un promedio de $0,52$. Los demás procedimientos realizados (con KNN, Decision Tree, Linear Regression y Random Forest) presentan valores similares ($0,7$ - $0,8$), que son aceptables según las condiciones expuestas en la presente memoria. Por tanto, SVM en algunas duplas de sus componentes, están sesgando los cálculos. Se concluye que la base 02 posee predicciones aceptables.
En ventas (base 06) los datos poseen un rendimiento de $0,8$ promedio, considerando todos los algoritmos utilizados. Por tanto, tanto en venta como en arriendo es recomendable predecir usando todos los métodos. Queda certificada la hipótesis planteada con anterioridad, pues de RMSE se pueden extraer las mismos conclusiones.

En las base 03 y 07 (casas), se observa el mismo fenómeno anómalo al interior de SVM. Altas diferencias en valores métricos de R^2 entre los diferentes resultados fluctuantes a causa de las variaciones pertenecientes a la dupla de parámetros costo-gamma.
Aún así, se repiten las mismas conclusiones del párrafo anterior. Destacando que la precisión es más fuerte en las casas que en los departamentos.

Por último, para las bases 04 y 08 (oficinas) se presenta un caso particular y diferente a los demás. En arriendo, con la métrica R^2, las oficinas se encuentran en niveles no aceptables. Siendo peor ventas que arriendo. No obstante, considerando la métrica RMSE es distinto. Las oficinas en arriendo son la 2da clase más aceptable de un total de 3 (casas-departamentos-oficinas). Mientras que las oficinas en venta son la 1ra clase aceptable dentro del trío anteriormente expuesto.
Por lo tanto, tenemos un conjunto aceptable, y otro conjunto no aceptable. Confesándole al lector que, si solo considera RMSE considera las oficinas; y si considera la métrica de R^2 entonces no proceda en utilizarla. Si utiliza ambas, queda a criterio del analista, teniendo en cuenta este parráfo de discusión.

Finalmente se concluye que los bienes comerciales poseen mala predicción. No es recomendable al analista utilizar dichas bases para analizar resultados, siempre y cuando considere utilizar los mismo algoritmos y métricas expuestas en la presente memoria. Sigue que no se considerarán las bases 01 y 05 para la siguiente etapa.
Las casas y departamentos (bases 02, 03, 06 y 07) brindan bastante información para lograr predecir - con alto grado de confianza - los bienes que posean dicha clasificación. Siendo más precisos los resultados que correspondan a departamentos.
Sobre las oficinas se tiene una respuesta no dependiente del autor de la memoria. Si bien, los resultados son confiables para una métrica; no es consistente con la otra utilizada. Fenómeno que si pasaba con las bases anteriores. Por tanto será decisión del analista decidir si predecir, o no, a los elementos clasificados como oficina. En la presente memoria se mantendrá el uso de estas bases; para dar una vista amplia sobre las conclusiones finales.

C.1.3. Tercera Etapa

Ahora se reportarán los parámetros óptimos (en caso de existir) en cada uno de los algoritmos aplicados según los criterios de métrica en cada una de las bases.

• KNN:
Como es de observar en la siguiente tabla, se confirma lo concluido en la etapa 02. Las métricas R^2 y RMSE son consistentes en sus resultados.

Métricas	B1	B2	B3	B4	B5	B6	B7	B8
R2	8	11	12	12	7	6	9	9
RMSE	8	11	12	12	7	6	9	9
MAPE	3	3	12	6	3	3	4	3

Tabla C.3: Tabla resultado de aplicación etapa 03 - Modelo: KNN | Número de vecinos óptimos

Teniendo en consideración la ignorancia que se debe tener sobre las bases 01 y 05, puede deducirse que los elementos de arriendo necesitan más vecinos que sus pares de venta. Resulta interesante destacar la similitud entre los bienes vecinos correspondientes al arriendo. Pues, a menor cantidad, más similitud (de comportamiento) entre las características observables de los elementos del conjunto. Por tanto se concluye que en arriendo (resp. venta) los valores son más (resp. menos) fluctuantes en el sector aledaño del inmueble analizado. Implicando mayor (resp. menor) volatilidad en los precios del sector de arriendo (resp. venta) en los inmuebles de la Región Metropolitana.

• Árbol de Decisión:
En la tabla presente a continuación - a diferencia del caso anterior - coinciden las tres métricas en los parámetros óptimos.

Métricas	B1	B2	B3	B4	B5	B6	B7	B8
R2	0.0083	5e-04	5e-04	5e-04	0.015	5e-04	5e-04	5e-04
RMSE	0.0083	5e-04	5e-04	5e-04	0.015	5e-04	5e-04	5e-04
MAPE	0.0083	5e-04	5e-04	5e-04	0.015	5e-04	5e-04	5e-04

Tabla C.4: Tabla resultado de aplicación etapa 03 - Modelo: Árbol de Decisión | Complejidades óptimas

Más particular aún es la coincidencia en un único resultado. Eliminando las bases 01 y 05 del análisis, se tiene que el parámetro óptimo para todos los casos, es $cp = 5\text{e} - 04$. El conjunto de valores probados en dicho parámetro fue: $\{0,0005; 0,0083; 0,006; 0,015; 0,03\}$. Indicando que es conveniente elegir un valor bastante pequeño a la hora de diseñar una predicción que use Árboles de Decisión.

• Random Forest:
En el caso de Random Forest se tiene la siguiente tabla,
Nuevamente R^2 y RMSE son consistentes en sus resultados como dupla métrica. En ambos

Métricas	B1	B2	B3	B4	B5	B6	B7	B8
R2	3	7	7	6	4	3	3	7
RMSE	3	7	7	6	4	3	3	7
MAPE	3	3	3	3	3	3	3	4

Tabla C.5: Tabla resultado de aplicación etapa 03 - Modelo: Random Forest | Tamaños de nodo óptimos

casos los resultados de "node size" se repiten, haciendo presagiar que los resultados de optimalidad están asegurados.
En lo que respecta al problema resulta interesante observar la diferencia de niveles en arriendo y venta. Para el primer caso son las casas y los departamentos aquellos bienes que poseen un mayor "node size" (con un valor de 7), a diferencia de las oficinas que poseen un valor de 6.
Una relación distante si se observan los mismos inmuebles en venta. Mientras las casas y departamentos tienen un "nodesize" de 3, las oficinas poseen un valor de 7. Una diferencia mayor entre la dupla casa-departamento y oficina; en comparación al primer caso.

Se debe tener un tamaño de nodo similar entre los elementos de arriendo y relativo en los elementos de venta. Más, la cantidad de nodos en la dupla casa-departamentos (en ambos casos) debiese ser similar. Hasta iguales, si así lo decide el analizador. Sobre las oficinas se propone únicamente considerar un número alto.

• **Support Vector Machine:**
A continuación[10] se muestran los parámetros óptimos más reiterados de los parámetros variables y parámetro fijo a analizar dentro de la información recopiladas.

- ★ **Gamma = 0.1:** Costo = 4.
- ★ **Gamma = 1:** Costo = 4.
- ★ **Gamma = 10:** Costo = 4.
- ★ **Costo = 4:** Gamma = 0.1.
- ★ **Costo = 8:** Gamma = 0.1.
- ★ **Costo = 16:** Gamma = 0.1.

En vista de los resultados se debe hacer una especial atención en los resultados expuestos. Pues, a pesar de tantas posibilidades (cantidad de parámetros × cantidad de números probados × cantidad de métricas consideradas × cantidad de bases analizadas), convergen a un elemento común en todos los casos. Lo cual resulta inaudita considerando las variaciones paramétricas que existían en los anteriores algoritmos analizados.

Se concluye que para toda base a analizar, sin importar el tipo u operación del inmueble involucrado, los parámetros óptimos para este contexto - en SVM - son costo = 4 y gamma = 0.1.

[10]Considerando como métricas de análisis: R^2 y RMSE; y como bases eliminadas de todo análisis: base 01 y base 05.

C.1.4. Cuarta Etapa

La idea central de esta etapa estaba en mostrar los parámetros óptimos en cada uno de los algoritmos; considerando todas las bases.
Como ya se ha observado, se han descartado los bienes comerciales (base 01 y base 05) para el análisis. Además de la métrica MAPE. Por tanto, esta etapa se resume tan solo con visualizar los valores resultantes de lo anterior, justificado en la dupla de métricas R^2-RMSE que - a pesar de ser distintas - arrojaron el mismo parámetro óptimo en cada uno de los algoritmos. Por tanto, la presente etapa se resume en la siguiente tabla.

Algoritmo	Parámetro Variable	B1	B2	B3	B4	B5	B6	B7	B8
KNN	k	8	11	12	12	7	6	9	9
DT	cp	0.0083	0.0005	0.0005	0.0005	0.015	0.0005	0.0005	0.0005
RL	Null	X	X	X	X	X	X	X	X
SVM	costo	4	4	4	4	4	4	4	4
	gamma	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
RF	node size	3	7	7	6	4	3	3	7

Tabla C.6: Tabla general de parámetros óptimos etapa 04 | Métricas: R^2-RMSE

Lo anterior es verificable en las siguientes tablas,

• **KNN:**
◦ **Número de Vecinos Óptimos:** Tabla B.14.

• **Árbol de Decisión:**
◦ **Complejidades óptimas:** Tabla B.15.

• **Random Forest:**
◦ **Tamaño del nodo:** Tabla B.16

• **Support Vector Machine:**
Al ser 2 parámetros se debe analizar considerando un parámetro móvil, y el otro fijo. Por tanto habrán 2 tablas que representen el valor de un parámetro móvil con respecto a uno filo. Para el presente algoritmo, primero será la tabla que contenga el parámetro fijo `gamma`, y después la tabla respectiva del parámetro `costo`. Estas se presentan a continuación,
◦ **Parámetro fijo = Gamma:** Tablas B.17.
◦ **Parámetro fijo = Costo:** Tablas B.18.