UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

TRABAJO FINAL DE GRADO PARA OBTENER EL TÍTULO DE LICENCIADO EN ESTADÍSTICA

Precio de oferta de apartamentos en Montevideo: Una aproximación desde la ciencia de datos.

Lucía Coudet y Alvaro Valiño

Tutora:

Dra. Natalia da Silva

Montevideo, Diciembre 2021.

El tribunal docente integrado por los abajo firmantes aprueba el Trabajo Final:

Título

Precio de oferta de apartamentos en Montevideo: Una aproximación desde la ciencia de datos.

Autor/es

Lucía Coudet

Alvaro Valiño

Tutora/Coordinadora

Dra. Natalia da Silva

Puntaje .....

Tribunal

Profesor Dr. Marco Scavino ..... (firma).

Profesor Dr. Leonardo Moreno ..... (firma).

Profesora Dra. Natalia da Silva ..... (firma).

Fecha .....

Abstract

El siguiente trabajo consiste en la implementación de técnicas de aprendizaje estadístico con el fin de realizar predicciones sobre el precio de oferta de los apartamentos a la venta en Montevideo, Uruguay. Para ello, se trabajó con una base de datos de elaboración propia a partir de los apartamentos que se ofertan a través de la plataforma *Mercado Libre*. La misma contiene información para el periodo comprendido entre el mes de Mayo al mes de Setiembre del año 2021. Esto último en la medida que no es posible la obtención de datos históricos. En cuanto a las técnicas de aprendizaje estadístico, se trabajó con algoritmos de *Árboles de Decisión* y métodos agregados de los mismos como ser *Random Forest* y *Boosting*, al igual que con métodos de *regresión robusta* (*Support Vector Regression*). Asimismo, se implementó un ajuste mediante un *Modelo de Regresión Lineal Múltiple* el cual es conocido en la literatura económica como *Modelo de Precios Hedónicos* y usualmente aplicado en problemas de este índole. Este último con el fin de comparar su capacidad predictiva en relación a las técnicas anteriores. Con respecto a la performance predictiva de los algoritmos, se observó que el ajuste por *Boosting* presentó una performance predictiva superior, esto último luego de haber ajustado los respectivos *parámetros de ajuste*. En cuanto al análisis de interpretabilidad de los modelos, se trabajó con métodos *modelo-agnósticos*, haciendo hincapié en las metodologías *importancia de las variables permutadas* y *gráficos de dependencia parcial*. De esta forma, se observó que la distancia entre el apartamento y la rambla este de Montevideo es la variable más importante en el ajuste, y a su vez esta presenta una relación inversa en cuanto a la predicción del precio de oferta del mismo.

Palabras claves: aprendizaje estadístico, *Árboles de Decisión*, métodos agregados, *regresión robusta*, parámetros de ajuste, métodos *modelo-agnósticos*.

Índice general

Agredecimientos

Agredecer en primera instancia a nuestra tutora Natalia da Silva por su dedicación y compromiso en cada instancia de este proceso. A la Facultad de Ciencias Económicas y de Administración de la UdelaR que ha acompañado nuestra formación universitaria y profesional a lo largo de muchos años. A nuestras familias, amigos y compañeros que fueron un pilar fundamental para poder estar hoy finalizando esta etapa de nuestras vidas. A quienes han sido nuestros docentes a lo largo de la carrera por la formación y conocimientos brindados.

Capítulo 1

Introducción

El presente trabajo tiene como principal objetivo el estudio e implementación de diferentes técnicas de aprendizaje estadístico mediante las cuáles realizar predicciones de una variable de interés. En particular, se trabajó con el precio de oferta en dólares estadounidenses de los apartamentos a la venta en el departamento de Montevideo, Uruguay, disponibles en *Mercado Libre* (precio de oferta en dólares) para el periodo Mayo a Setiembre del año 2021.

Para ello, se plantearon 3 objetivos específicos, siendo estos: 1) obtención y procesamiento de los datos para transformarlos en información, con énfasis en la automatización y reproducibilidad de los mismos; 2) implementación de técnicas de aprendizaje estadístico que permitan una mayor flexibilidad en la estimación con respecto a las técnicas clásicas de estimación y 3) una primera aproximación a un análisis de interpretabilidad enfocándose en los *métodos globales modelo-agnósticos* (Molnar, 2020) ([26]).

En lo que respecta al primer objetivo, los datos de oferta de los apartamentos fueron obtenidos a través de la interfaz de programación de aplicaciones (API) de *Mercado Libre*. Esto implicó la creación de un programa que permite una descarga automatizada de la información disponible. Asimismo, se trabajó con fuentes adicionales de información externas. En particular con la *Encuesta Contínua de Hogares* (ECH) (`https://www.ine.gub.uy/web/guest/encuesta-continua-de-hogares1`)y datos

obtenidos de la plataforma *Google My Maps* (`https://www.google.com/intl/es/maps/about/mymaps/`). De esta manera, fueron construídas variables adicionales que se consideraron de interés para el problema planteado.

En cuanto al segundo objetivo, las técnicas de aprendizaje estadístico implentadas se separan en dos grupos. En primer lugar, un conjunto de métodos de agregación basados en *Árboles de Decisión*, en particular *Random Forest* y *Boosting* (James, 2013) ([18]). En segundo lugar, se trabajó con la técnica *Support Vector Regression* la cual pertenece al grupo de los modelos de regresión robusta. (Vapnik, 1999) ([35])

En lo que respecta a las técnicas clásicas de estimación, se realizó la aplicación de un *Modelo de Regresión Lineal Múltiple* (Carmona, 2005) ([5]). Este último en la medida que es una técnica usualmente aplicada en problemas de este índole y conocido en la literatura económica como *Modelo de Precios Hedónicos*.

Una vez implementados los modelos mencionados en el punto anterior, con el fin de determinar el modelo con mejor capacidad predictiva, se llevó a cabo un análisis comparativo en función de diferentes métricas.

Por último, en la medida que los modelos implementados (con excepción del *Modelo de Regresión Lineal Múltiple*) se denominan como modelos de caja negra (Molnar, 2020) ([26]) se llevó a cabo un análisis de interpretabilidad aplicado al mejor modelo en términos de performance predictiva. Esto último fue abordado principalmente mediante la realización de un análisis gráfico.

Se destaca que las técnicas utilizadas implican un alto costo computacional. De esta forma, con el fin de obtener un mayor poder de cómputo se utilizó un enfoque desde la programación en paralelo. A su vez, todas las etapas fueron realizadas teniendo en cuenta la reproducibilidad de las mismas.

Los resultados obtenidos fueron a través del lenguaje y entorno de programación para análisis estadístico y gráfico, *R* ([32]), enfocándose en la optimización de todos los procesos principalmente mediante la programación en paralelo.

El trabajo se conforma de 8 capítulos. A continuación se presenta en el Capítulo 2 el detalle de la obtención y depuración de los datos utilizados. Luego, en el Capí-

tulo 3 se mencionan los principales antecedentes consultados. En lo que respecta al Capítulo 4, se detalla la metodología utilizada y el sustento teórico de la misma. Este último seguido por el Capítulo 5 donde se presentan los aspectos que conciernen a la reproducibilidad del trabajo. Posteriormente, en los Capítulos 6 y 7 se presenta el análisis exploratorio de los datos y los principales resultados obtenidos respectivamente. Por su parte, en el Capítulo 8 se presentan las principales conclusiones abordadas y principales líneas de trabajo para futuras investigaciones.

Capítulo 2

Datos

Como fue mencionado en el Capítulo 1, el principal objetivo del trabajo es el estudio e implementación de diferentes técnicas de aprendizaje estadístico mediante las cuales realizar predicciones de una variable de interés. En particular se trabajó con el precio de oferta en dólares estadounidenses de los apartamentos a la venta en Montevideo, Uruguay, para el periodo Mayo a Setiembre del año 2021.

A continuación se presenta en la Sección 2.1 la descripción del proceso de obtención de los datos. Luego, en la Sección 2.2 se detallan los principales criterios de depuración utilizados. Posteriormente, en la Sección 2.3 se describen las principales fuentes externas de información utilizadas. Por último en la Sección 2.4 se presentan las variables construídas en base a elaboración propia.

2.1. Obtención

La base de datos utilizada fue obtenida consultando la API de *Mercado Libre*. Donde con el fin de interactuar con la misma, es necesario seguir los siguientes pasos: 1) El usuario se registra en la web `https://developers.mercadolibre.com.uy/`; 2) se crea una aplicación con la cual se obtiene un identificador y una contraseña para interactuar con la API y 3) se obtiene una clave (*token*) para poder realizar las consultas.

En particular, para obtener la información de las publicaciones de apartamentos se realiza una consulta a la API especificando en la *url* la categoría *MLU1474*. Adicionalmente, para considerar solamente los apartamentos en el departamento de Montevideo, se especifica en la *url* el identificador de cada uno de los barrios de Montevideo. En el Anexo A.2 se presenta en la Tabla A.2 los barrios de Montevideo y su identificador.

Luego, con el fin de obtener un mayor número de variables explicativas, se accedió a los atributos específicos de cada publicación. Para ello, es necesario ingresar en la *url* cada identificador de cada publicación.

De esta manera fue posible obtener la información de todos los apartamentos a la venta en Montevideo disponible en la API de *Mercado Libre* y considerada de interés de manera automatizada.

Se destacan dos aspectos relevantes en cuanto al funcionamiento de la API de *Mercado Libre*.

En primer lugar, es posible obtener la información de las publicaciones vigentes a la fecha de consulta. Por lo tanto, se realizaron descargas mensuales abarcando el periodo comprendido entre los meses de Mayo a Setiembre del año 2021, ambos inclusive. Dado que las publicaciones mantienen los identificadores a través de los meses, para el caso de las publicaciones que se mantienen vigentes más de un mes, se optó por considerar los datos correspondientes al último mes de vigencia.

En segundo lugar, la API permite extraer hasta un límite determinado de publicaciones por consulta (10.000 consultas a la fecha de realización del trabajo). Debido a que las consultas fueron realizadas de forma iterativa por barrio, fue posible sobrellevar esta limitante.

2.2. Procesamiento y criterios de depuración

El proceso de depuración de los datos fue realizado diferenciando según la naturaleza de las variables.

En primer lugar, en lo que respecta a las variables cuantitativas el proceso de depuración se centró en el reconocimiento de valores erróneos, por ejemplo valores que repitan una secuencia de números.

Para ello se construyó una función auxiliar que es capaz de detectar cuando un valor tiene 3 o más números iguales repetidos. En ese caso se considera que el dato es erróneo y se lo asigna como valor faltante, con excepción de la variable precio de oferta en dólares donde la observación completa es eliminada.

Asimismo se decidió no considerar en el análisis las observaciones cuyo precio de oferta es inferior al valor del percentil 75 % entre las observaciones con precio inferior a 40,000 dólares ya que se consideran datos erróneos o de muy baja frecuencia. De manera similar y debido a la elevada presencia de valores atípicos, se eliminan todas las observaciones cuyos valores de la variable precio de oferta en dólares superan el percentil 95 % de la misma. De esta manera, no se consideraron en el análisis aproximadamente un 8 % de las observaciones.

A su vez, en lo que respecta a las variables área total y área cubierta se decidió asignar como valor faltante a todos los valores superiores a 1000 metros cuadrados, ya que se considera que podrían ser datos erróneos.

En segundo lugar, respecto a las variables cualitativas, existe un conjunto de variables asociadas a la publicación del inmueble que toman valor Si, No, o faltante. Debido a que en todos los casos se trata de campos no obligatorios para el usuario que realiza la publicación, no es posible diferenciar fehacientemente entre los valores faltantes y el valor No. Por lo tanto se trabajó bajo el supuesto de que los valores faltantes corresponden a el valor No y de esta forma se realizó la recodificación correspondiente.

En tercer lugar, en cuanto a las variables latitud y longitud que permiten georreferenciar a los apartmentos, se realizó un proceso de depuración atendiendo a la factibilidad del dato. Para ello se consideraron los valores de ambas variables correspondientes a coordenadas geográficas dentro del cuadrante donde se encuentra el polígono de Montevideo. De esta manera, valores de latitud inferiores a -35 y

superiores a -34.7, al igual que valores de longitud inferiores a -56.5 y superiores a -56 se consideraron datos erróneos ya que Montevideo se encuentra entre dichos valores de latitud y longitud. Los mismos fueron recodificados imputando los valores correspondientes al centroide del barrio donde se ubica el apartamento.

A su vez, se detectó otro tipo de error en las georreferencias siendo este el caso de aquellas que no se ubican dentro del polígono del barrio al cual pertenece el apartamento. Para estos casos, se supone que el dato correcto es el nombre del barrio y no la georreferencia específica, y se imputa el valor de latitud y longitud correspondiente al centroide del barrio.

Dada la complejidad que implica la detección de estas georreferencias erróneas, mediante un procedimiento de *trade-off* entre eficiencia y complejidad, se procedió de la siguiente forma: 1) se separó el mapa de Montevideo en dos regiones utilizando el corte de la calle Avenida Italia en continuación con la calle Avenida 18 de Julio; 2) se computó a qué región resultante de 1) pertenece el apartamento utilizando los valores de latitud y longitud; 3) se procedió de forma análoga a 2) considerando el centroide del barrio al cual pertenece el apartamento, 4) se evalua si 2) y 3) coinciden, en caso que difieran se le imputa al apartamento los valores de latitud y longitud correspondientes al centroide del barrio.

En cuarto lugar, se decidió considerar en el análisis las variables con porcentaje de valores faltantes inferior al 15 %. En la Tabla A.1 de la Sección del Anexo A.1 se presenta la proporción de valores faltantes para cada una de ellas.

Por último, en lo que respecta al tipo de cambio, todos los valores de la variable precio de oferta expresados en moneda nacional fueron convertidos a dólares estadounidenses utilizando el tipo de cambio a la fecha de obtención de los datos. Sobre éste punto, se destaca que la obtención del valor del tipo de cambio se realiza de manera automatizada realizando un procedimiento de *Web Scrapping* sobre la página web oficial del *Instituto Nacional de Estadística* (INE) (`https://www.ine.gub.uy/`).

2.3. Fuentes externas de información

En la medida que la base de datos construida contiene información sobre la latitud y longitud donde está ubicado cada apartamento, se consideró de interés la elaboración de variables geoespaciales.

La herramienta utilizada para construir mapas fue *Google My Maps*, el cual es un servicio de *Google* que permite a los usuarios crear mapas personalizados. De esta manera, es posible añadir puntos, líneas y formas sobre *Google Maps*. (`https://sites.google.com/mrpiercey.com/resources/geo/my-maps`).

Una vez que se construyen los mapas desde dicha plataforma, se exportan los archivos generados. Luego, estos son transformados a archivos ESRI Shapefile utilizando *QGIS* el cual es un *Sistema de Información Geográfica* (SIG).

El formato ESRI (Environmental Systems Research Institute, Inc.) Shapefile (SHP) es un formato vectorial de almacenamiento digital donde se guarda la localización de los elementos geográficos y los atributos asociados a ellos. Es un formato multiarchivo, es decir está generado por varios ficheros informáticos. El número mínimo requerido es de tres y tienen las extensiones siguientes:

- shp es el archivo que almacena las entidades geométricas de los objetos.
- shx es el archivo que almacena el índice de las entidades geométricas.
- dbf es la base de datos, en formato dBASE, donde se almacena la información de los atributos de los objetos.

(ESRI, 1998) ([9])

En lo que respecta a la geometría del departamento de Montevideo la misma fue obtenida a partir de los archivos Shapefile disponibles en la página web del *Instituto Nacional de Estadística* (INE) en la siguiente dirección: `https://www.ine.gub.uy/`.

Por otra parte, se utilizó también información de la encuesta contínua de hogares 2020 para la construcción de variables que consideren el nivel de ingreso por barrio.

2.4. Variables construidas

Como se menciona en la Sección 2.3, tener la georreferencia específica en la base de datos motivó la elaboración de variables geoespaciales. En particular, las variables construídas fueron: ubicación respecto a la calle Avenida Italia en continuación con la calle Avenida 18 de Julio (zona Avd. Italia), distancia a la rambla este de Montevideo (distancia a la rambla este), y distancia al centro comercial más cercano (distancia al centro comercial más cercano).

La metodología utilizada para el cálculo de las distancias fue a través de la fórmula de *Haversine* la cual permite el cómputo de la distancia mínima entre dos puntos que se encuentran en un cuerpo esférico utilizando latitud y longitud. Para ello en particular se trabajó con la función *distm* del paquete *geosphere* ([17]). En el Anexo A.3 se especifican los detalles sobre el cálculo.

En lo que respecta a la variable zona Avd. Italia, la misma toma valor norte o sur según el apartamento se encuentre al norte o al sur del corte de la calle Avenida Italia en continuación con la calle Avenida 18 de Julio. Para ello, se tomó el punto más cercano a la ubicación del apartamento en el mapa de líneas que compone la calle Avenida Italia en continuación con la calle Avenida 18 de Julio. En particular, se toma la diferencia mínima en términos de longitud entre el apartamento y los puntos que componen el mapa. Una vez obtenido el mínimo, se comparan las latitudes. En caso que la latitud del apartamento sea mayor que la del punto más cercano seleccionado, se asigna sur, y en otro caso se asigna norte.

Por otra parte, la variable distancia al centro comercial más cercano fue contruida calculando la distancia en metros entre el apartamento y el centro comercial más cercano. Para ello, fue necesaria la obtención de las coordenadas geográficas (latitud y longitud) de todos los centros comerciales ubicados en Montevideo (Montevideo Shopping Center, Punta Carretas Shopping, Tres Cruces Shopping, Nuevocentro Shopping y Portones Shopping).

Por último, la variable distancia a la rambla este fue construida utilizando la

distancia en metros entre el apartamento y la rambla este de Montevideo. Esta variable fue construida a través de los resultados observados mediante el ajuste de un *Árbol de Regresión* de la variable precio de oferta en dólares en función de la latitud y la longitud que corresponde a la ubicación de cada apartamento. En el mismo se observó que los apartamentos con predicciones superiores toman valores de latitud entre -34.92 y -34.89, y valores de longitud mayores o iguales a -56.17. En la Sección del Anexo A.4 se encuentra el gráfico del árbol ajustado.

Se destaca que los mapas utilizados para construir las variables zona Avd. Italia, distancia a la rambla este y distancia al centro comercial más cercano fueron obtenidos en base a elaboración propia a partir de *Google My Maps*, según detallado en la Sección 2.3. En la Figura 2.1 se presenta el mapa del departamento de Montevideo, Uruguay, con las variables construidas.

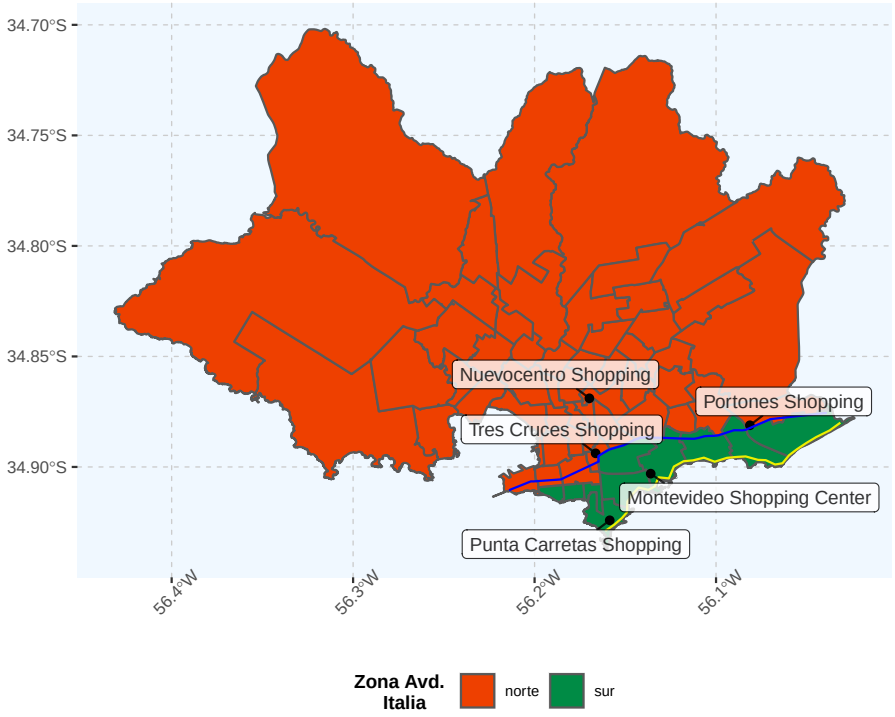

Figura 2.1: Mapa de Montevideo, Uruguay con la georreferencia de los centros comerciales, calle Avenida Italia en continuación con la calle Avenida 18 de Julio y rambla este de Montevideo . La línea azul indica la calle Avenida Italia en continuación con la calle Avenida 18 de Julio mientras que la línea amarilla indica la rambla este de Montevideo.

Por otra parte, se construyó la variable ingreo medio ECH utilizando la información de la encuesta contínua de hogares (ECH) del año 2020 disponible en `https://www.ine.gub.uy/`. En particular se utilizó la variable *HT11*: Ingreso total del hogar con valor locativo sin servicio doméstico (en pesos uruguayos). Se calculó el ingreso medio por barrio de los hogares de Montevideo y luego se asignó a cada observación, el nivel de ingreso medio que le corresponda según el barrio donde se encuentre el apartamento.

En la Figura 2.2 se presenta el mapa del ingreso promedio de los hogares por barrio de Montevideo.

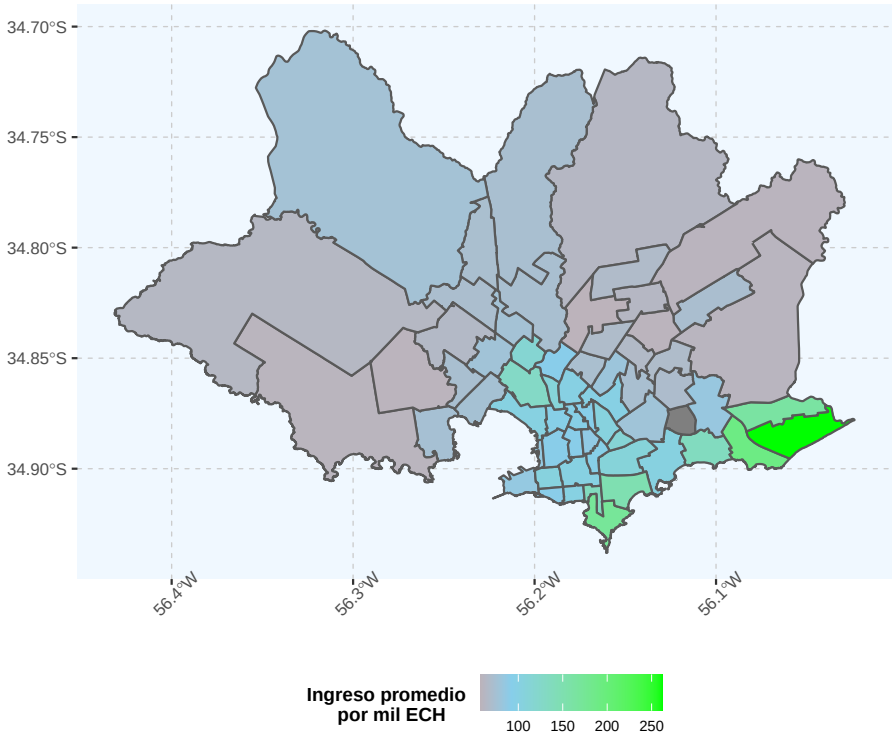

Figura 2.2: Mapa del departamento de Montevideo, Uruguay, por nivel de ingreso promedio por mil obtenido a partir de la Encuesta Contínua de Hogares 2020. El color celeste indica el ingreso promedio de los barrios de Montevideo.

Capítulo 3

Antecedentes

Como fue mencionado en el Capítulo 1, el presente trabajo tiene como principal objetivo la implementación y estudio de diferentes técnicas de aprendizaje estadístico multivariadas aplicadas sobre el precio de oferta de los apartamentos a la venta en Montevideo, Uruguay.

En este sentido, existen diversos artículos y documentos de trabajo que tratan en particular sobre la implementación de modelos predictivos para el precio de los inmuebles tanto en el mercado internacional como en el mercado uruguayo.

En lo que respecta a la bibliografía internacional, existen diversos estudios que realizan la implementación tanto de técnicas clásicas de estimación como de técnicas de aprendizaje estadístico. Por ejemplo se encuentra el trabajo de Mullainathan y Spiess (2017) que realiza una introducción a los modelos de aprendizaje estadístico utilizando datos de precios de viviendas. (Mullainathan, 2017) ([28])

Por otro lado, el trabajo de Baldominos et al (2018) consiste en predecir el precio de oferta de inmuebles ofertados en un sitio web de España y ubicados en la ciudad de Salamanca. Para ello plantearon un problema de regresión y adicionalmente trabajaron con técnicas de aprendizaje estadístico: *Árboles de Regresión*, *k-vecinos más cercanos*, *Support Vector Regression* y *Redes Neuronales*. (Baldominos, 2018) ([1])

En lo que respecta a la bilbiografía nacional, en primer lugar se tiene el trabajo de Ponce (2012) ([31]) donde propone un modelo de precios de fundamentos para

las viviendas el cual se basa en el hecho de que una vivienda puede ser considerada como un activo de inversión y como un activo que brinda servicios. De esta forma, el precio puede ser considerado como el resultado del mercado por los servicios de vivienda y como el resultado de equilibrio en un mercado de activos. Concluye, entre otros aspectos, que los precios de las viviendas fluctúan más que lo justificado por sus fundamentos lo cual implica periodos de subvaloración o sobrevaloración de los precios de las viviendas. (Ponce, 2012) ([31])

Posteriormente, Ponce y Tubio (2013) ([30]) realizan una aplicación de modelos hedónicos para el mercado uruguayo. Para ello utilizan una base de datos con información de inmuebles que se ofertan a través de la web. La misma contiene información de más de 500 inmobiliarias y más de 20 barrios de Montevideo, Uruguay. Se trata de una sistematización de metodologías existentes para la elaboración de índices de precios de inmuebles, en particular atendiendo a modelos que permitan evaluar el desvío de los precios corrientes con respecto a los fundamentos del mercado. (Ponce, 2013) ([30])

Luego, Landaberry y Tubio (2015) ([22]) con el fin de monitorear el mercado de viviendas en Uruguay proponen una serie de índices de precios. En particular para el caso de Montevideo proponen índices desde un enfoque hedónico ya que, entre otras cosas, permiten estimar precios sombras para los atributos de las viviendas. (Landaberry, 2015) ([22])

En el año 2017 Goyeneche et al ([10]) trabajaron en la construcción de un modelo predictivo del valor contado de un determinado inmueble, entendido como valor contado aquel que es asignado por el tasador. En línea con esto, recurrieron a la metodología de *Stacking* con el fin de lograr predicciones más precisas. En particular trabajaron con información de inmuebles en Montevideo, Uruguay. La base de datos utilizada fue proporcionada por el Banco Hipotecario del Uruguay (BHU). (Goyeneche, 2017) ([10])

Por último, en el año 2019 Picardo ([29]) presenta modelos predictivos para el precio de las viviendas. En particular trabajó mediante la implementación de *Mode-*

los de Regresión Lineal Múltiple, *Árboles de Regresión*, y el algoritmo *Random Forest* para inmuebles ubicados en Montevideo, Uruguay. La base de datos utilizada corresponde a elaboración propia obtenida a través de la web y registros administrativos de transacciones de la Dirección General de Registros (DGR).

Se destaca que éste último se considera como principal antecedente, donde en el presente trabajo se tiene como principal línea de investigación profundizar en 1) las técnicas de aprendizaje estadístico utilizadas y su implementación, 2) la obtención de una mayor medida de interpretabilidad de las mismas, 3) la reproducibilidad de los resultados obtenidos y 4) la optimización de recursos computacionales.

En lo que respecta a las técnicas estadísticas utilizadas, se consultaron diferentes literaturas. En este sentido, entre las bibliografías destacadas para los algoritmos de aprendizaje estadístico se encuentra el libro publicado por Gareth et al ([18]), al igual que el libro publicado por Hastie et al ([16]).

Por otra parte, en lo que respecta al detalle metodológico del *Modelo de Regresión Lineal Múltiple*, fue consultado principalmente el trabajo de Carmona ([5]).

Por último, para el análisis de interpretabilidad de los modelos de aprendizaje estadístico fueron consultados los trabajos de Molnar ([26]) y Greenwell et al ([13]).

Capítulo 4

Marco teórico y metodología

En este Capítulo se presenta los aspectos metodológicos más relevantes para la realización del trabajo. El mismo se encuentra conformado por cuatro secciones.

En primer lugar, en la Sección 4.1 se detalla la metodología de los modelos ajustados. Luego en las Secciones 4.2 y 4.3, se encuentra explicitada la metodología para evaluar la performance predictiva y realizar un análisis de interpretabilidad de los modelos ajustados respectivamente.

Por último, en la Sección 4.4 se presenta la metodología llevada a cabo para realizar el tratamiento de datos faltantes en la base de datos.

4.1. Análisis Supervisado y Aprendizaje Estadístico

Con el fin de obtener predicciones del precio de oferta de los apartamentos, fueron implementadas diferentes técnicas de aprendizaje estadístico. Estas consisten en modelar y analizar conjuntos de datos, mediante el aprendizaje de ejemplos, con el fin de predecir y estimar resultados en forma automática.

En este contexto, se realizó un análisis supervisado, en la medida de que se cuenta con una variable de salida (Y) y variables de entrada (X). Por lo tanto, se tiene que los posibles modelos son de la forma:

$$Y = f(X) + \epsilon$$

Siendo f una función desconocida y ϵ un error aleatorio independiente de X e Y con media 0. Se denota a la matriz X de dimensión $n \times p$ a la matriz de datos, donde se tiene n observaciones de entrenamiento y p variables.

La i-ésima fila se corresponde a la i-ésima observación (perteneciente al conjunto de entrenamiento) siendo de la forma $x_i = (x_{i1}, \ldots, x_{ip})^T$, con $x_i \in \mathbb{R}^p$. Por otro lado, se denota una nueva observación (o perteneciente al conjunto de testeo) como $x^* = (x^*_{i1}, \ldots, x^*_{ip})^T$, donde esta es un vector p-dimensional (al igual que x_i).

A la hora de estimar f, se realizó mediante métodos paramétricos, explicitados en la Sección 4.1.1, y métodos no paramétricos detallados en las Secciones 4.1.2, 4.1.3, 4.1.4 y 4.1.5.

En el primer caso, se asumió la forma funcional de f y se procedió a estimar sus respectivos parámetros. Por otro lado, en los métodos no paramétricos, no se asumió la forma funcional de f.

4.1.1. Modelo de Regresión Lineal Múltiple

El *Modelo de Regresión Lineal Múltiple* donde para el problema de aplicación seleccionado se conoce en la literatura como *Modelo de Regresión Lineal Múltiple de Precios Hedónicos* (o simplemente *Modelo de Precios Hedónicos*) parte del supuesto de que los precios observados de los productos se pueden desglosar en una suma de cantidades específicas de determinadas características asociadas al bien. De esta manera se define un set implícito de precios, también conocidos como *precios hedónicos*. (Rosen, 1974) ([33])

Este tipo de modelos fueron introducidos por Griliches en el año 1961 ([14]) para el precio de los automóviles y luego desarrollado y profundizado por Rosen en el año 1974 ([33]).

De esta manera, el precio del bien es regresado sobre las características del mismo, y utilizando técnicas clásicas de estimación se obtienen los anteriormente mencionados precios hedónicos.

En particular, este modelo puede aplicarse a los precios de los bienes inmuebles. Donde, para estos últimos se destaca que entre las características asociadas al inmueble pueden considerarse características que son propias del mismo así como también características asociadas a la localización geográfica en donde se encuentra ubicado, entre otras. (De Bruyne, 2013) ([8]).

Formalmente, el modelo se especifica como:

$$Y = f(X) + \epsilon$$

donde

$$f(X) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

con ϵ igual al vector de errores del modelo y a su vez ϵ_i independientes e identicamente distribuidos $N(0, \sigma^2)$. (Carmona, 2005) ([5]).

Siendo $(X_1, \cdots, X_p)$ el vector de las p características asociadas al bien y $(\beta_1, \cdots, \beta_p)$ el vector de los precios hedónicos. Es importante observar que el vector de precios hedónicos asociados a las características coincide con el vector de los parámetros de un modelo lineal clásico.

La estimación de los parámetros $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ se realiza por el método de los mínimos cuadrados. Para ello, se intenta hallar los valores $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ que minimicen la suma de cuadrados de los residuos $\boldsymbol{\epsilon'\epsilon} = \boldsymbol{(Y - X\beta)'(Y - X\beta)}$.

De esta manera, se obtiene que la estimación del vector β que minimiza la suma de los cuadrados de los residuos sujeto a que $X\beta$ pertenezca al espacio columna de X. La misma es solución de la siguiente ecuación:

$$\mathbf{X'X}\beta = \mathbf{X'Y}$$

$$\hat{Y} = \hat{f}(X) = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

En lo que respecta a la bondad de ajuste del modelo, se utilizaron el coeficiente de determinación (R^2) y coeficiente de determinación ajustado (R^2_a).

Se define al coeficiente de determinación (R^2) a través de la siguiente ecuación:

$$R^2 = 1 - \frac{SCR}{SCT}$$

donde:

- $SCR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- n el número de observaciones.

Por otro lado, se define el coeficiente de determinación ajustado mediante la siguiente expresión:

$$R^2_a = 1 - \big(\frac{n-1}{n-p-1}\big)(1 - R^2)$$

donde

- n es el número de observaciones
- p es el número de variables explicativas del modelo
- R^2 es el coeficiente de determinación.

En lo que respecta a la significación global del modelo, se utilizó la prueba *F de Fisher* (Prueba F) en donde la hipotesis nula implica la no significación global del modelo:

$$H_0)\beta_j = 0 \,\forall\, j = 1, \cdots, p$$

$$H_1)\beta_j \neq 0 \,\text{para al menos un}\, j = 1, \cdots, p$$

El estadístico del contraste (F) y su distribución bajo la hipótesis nula cierta tiene la siguiente forma:

$$F = \frac{(SCE - SCR)/p}{SCR/(n-p-1)} \sim F_{p,n-p-1}$$

donde

- SCE es la suma de los cuadrados explicada
- SCR es la suma de los cuadrados de los residuos
- n es el número de observaciones
- p el número de variables incluídas en el modelo

La hipótesis nula de no significación global del modelo se rechaza cuando $F > F_{p,n-p-1;1-\alpha}$, siendo α el nivel de significación del contraste. (Carmona, 2005) ([5])

Por otra parte, para la significación individual de las variables, se utilizó la prueba *t de Student* (Prueba t) en donde la hipotesis nula implica la no significación de la variable X_j para explicar a la variable de respuesta. De esta manera, el contraste se especifica de la siguiente forma:

$$H_0)\beta_j = 0$$

$$H_1)\beta_j \neq 0$$

El estadístico del contraste (t) y su distribución bajo la hipótesis nula cierta tienen la siguiente forma:

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma_j}(\hat{\beta}_j)} \sim t_{n-p-1}$$

donde

- $\hat{\beta}_j$ es el parámetro estimado asociado a la variable X_j
- $\hat{\sigma_j}(\hat{\beta}_j)$ es el desvío estimado del parámetro estimado asociado a la variable X_j

La hipótesis nula de no significación de la variable X_j se rechaza cuando $|t| > t_{n-p-1;1-\alpha}$. (Carmona, 2005) ([5])

Con el fin de testear los supuestos de normalidad y homocedasticidad de los residuos del modelo existen diferentes pruebas.

En particular, en este trabajo se realizó la prueba de normalidad de *Lilliefors* sobre los residuos del modelo. Esta prueba utiliza el estadístico *Kolmogorov-Smirnov* para el caso en que la media y la varianza poblacional son desconocidos. (Lilliefors, 1967) ([24])

El estadístico del contraste consiste en calcular la máxima diferencia absoluta entre la distribución empírica y la función de distribución acumulada hipotética:

$$D = \max_X |F^*(X) - S_N(X)|$$

donde $S_N(X)$ es la distribución acumulada en los datos mientras que $F^*(X)$ es la distribución acumulada de una variable aleatoria con distribución normal.

En el caso que D exceda un determinado valor crítico entonces se rechaza la hipótesis nula de distribución normal. (Lilliefors, 1967) ([24])

Para la aplicación de la prueba anterior se utilizó la función *lillie.test* del paquete *nortest* ([15]).

Por otro lado, se aplicó la prueba de homocedasticidad de *Breusch-Pagan* sobre los residuos del modelo, utilizando la función *bptest* del paquete *lmtest* ([37]). Esta prueba consiste en ajustar un modelo lineal para los residuos del modelo de regresión lineal mediante la siguiente expresión:

$$\frac{e^2}{\hat{\sigma}^2} = Z\alpha + \mu$$

donde:

- e es el vector de la estimación de los residuos en el modelo original
- $\hat{\sigma}$ la estimación del desvío de los residuos en el modelo original
- Z es las matriz de variables explicativas

- α es el vector de parámetros asociados a las variables explicativas
- μ es un vector aleatorio donde cada componente sigue una distribución normal, media cero y varianza constante

De esta forma, una vez ajustado el modelo de regresión sobre los residuos del modelo inicial, el contraste se especifica de la siguiente forma:

$$H_0)\sigma_i^2 = \sigma^2 \,\forall\, i = 1, \cdots, p$$

$$H_0)\sigma_i^2 \neq \sigma^2 \text{ para al menos un } i = 1, \cdots, p$$

El estadístico del contraste y su distribución asintótica bajo la hipótesis nula cierta son de la siguiente manera:

$$BP = \frac{SCE}{2} \sim \chi_p^2$$

De esta manera, se rechaza la hipótesis nula de homocedasticidad en el caso en que mucha varianza sea explicada por las variables explicativas. Por defecto, se utilizan en el contraste todas las variables explicativas del modelo inicial. (Breusch, 1979) ([4])

Se destaca que una de las ventajas más importantes de éste tipo de modelos es la fácil interpretación. No obstante suelen tener una mala performance predictiva en comparación a otros enfoques ya que pueden presentar problemas de heterocedasticidad, multicolinealidad, y variables omitidas. (James, 2013) ([18])

Por otra parte, el *Modelo de Regresión Lineal Múltiple* puede ser generalizado para el caso no lineal, lo cual no ha sido implementado en el presente trabajo.

4.1.2. Árboles de Regresión

Luego de realizar la implementación del *Modelo de Regresión Lineal Múltiple*, se procedió a modelar mediante un *Árbol de Decisión*. En la medida de que se cuenta con una variable de salida continua, se construyó un *Árbol de Regresión*.

A pesar de que en la literatura existen diversos enfoques para la construcción de estos modelos, se trabajó con el método *CART* el cual fue propuesto por *Breiman*, *Friedman*, *Olshen* y *Stone* en 1984. (Breiman, 2017) ([3])

Este método se caracteriza por la realización de particiones binarias recursivas del espacio de las variables de entrada. Mediante las mismas, se conforma una organización jerárquica en forma de árbol, donde en cada nodo interior se tiene una pregunta (dicotómica) sobre una variable de entrada y en cada nodo terminal (denominado "hoja") una decisión.

De esta forma, se procede a dividir el conjunto de los valores posibles de $X_1 \ldots, X_p$ (variables de entrada) en J regiones disjuntas $R_1, \ldots, R_J$.(James, 2013) ([18])

Luego para cada observación que se encuentra en la región R_j se realiza la misma predicción. Siendo esta, en el contexto de árboles de regresión, el promedio de la variable respuesta en dicha región.

En el momento de la construcción de las regiones $(R_1, \ldots, R_J)$ se realiza de tal forma que en cada subconjunto resultante (denominados como "nodos hijos") en cada iteración implique una disminución en la impureza de estos.

Para ello, se construyen las regiones $R_1, \ldots, R_j$ de forma tal que minimicen la suma de cuadrados de los residuos (SCR).

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

Siendo $\hat{y}_{R_j}$ el promedio de la variable respuesta en la j-ésima región.

Para lograr este cometido se utiliza una separación recursiva binaria de la siguiente forma. Se selecciona la variable X_j y el número s dividiendo el espacio en dos regiones $R_1(j,s) = \{X : X_j < s\}$ y $R_2(j,s) = \{X : X_j \geq s\}$ de forma tal que se haga mínimo

$$\sum_{i:x_i \in R_1(j,s)} (y - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y - \hat{y}_{R_2})^2$$

Una vez encontrada la mejor partición se separan los datos en las regiones resultantes y se repite el proceso en cada región. Es decir, se busca nuevamente la mejor

variable y el mejor punto de corte de forma que se incremente la disminución de la impureza en los nodos hijos.

El proceso continúa hasta que se satisfaga algún criterio de parada. Un criterio de parada puede ser por ejemplo que los nodos terminales tengan cierto número de observaciones.

Luego de definido el criterio de construcción de las regiones y el criterio de parada, se procede a realizar un proceso de poda en el árbol obtenido basado en un criterio de *costo-complejidad*. Esto en la medida de que si se deja crecer el árbol de forma indefinida se obtiene un modelo con un alto grado de sobreajuste (*overfitting*). Por su contraparte, un árbol muy "pequeño", posiblemente no logre capturar la estructura del conjunto de datos. (Hastie, 2001) ([16]).

El proceso de poda realizado, consiste en dejar crecer el árbol hasta que los nodos terminales tengan cierto número de observaciones (dicho árbol se denota como T_0). Luego se elige aquel subárbol el cual posee un menor error de predicción en el conjunto de testeo. En la medida de que un procedimiento de *validación cruzada* aplicado en cada posible subárbol es muy costoso en términos de "tiempo computacional", surge como alternativa el método basado en un criterio de *costo-complejidad*.

En dicho método se define a T_α como un subárbol obtenido al podar a T_0. De esta forma, para cada α se busca T_α que minimice la siguiente expresión:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

Donde se tiene que $|T|$ es igual al número de nodos terminales del árbol T, mientras que N_m es el número de observaciones en la región R_m. Por otro lado, la expresión $Q_m(T)$ consiste en la medida de impureza (SCR).

En cuanto al parámetro α, el mismo consiste en un parámetro de penalización aplicado a la complejidad (tamaño) del árbol. Donde valores altos de este, penalizan a árboles de gran tamaño. De esta forma, controla el compromiso entre la complejidad y la bondad de ajuste del modelo. Dicho parámetro se estima mediante *validación cruzada*.

4.1.3. Bagging y Random Forest

A pesar de que los *Árboles de Regresión* poseen un alto grado de interpretabilidad y sencillez en su implementación, estos poseen la gran limitante de ser inestables. Esto en el sentido de que pequeñas variaciones en el conjunto de entrenamiento y testeo generan grandes cambios en las estimaciones. (Hastie, 2001) ([16])

Por lo tanto, se emplearon diferentes métodos alternativos buscando estabilidad en las predicciones.

En primer lugar, se aplicó el método *Random Forest* desarrollado por *Breiman* en 1994. Este método consiste en construir un estimador combinando distintas versiones de estimadores.

En este contexto, estas nuevas versiones se construyen generando nuevos conjuntos de entrenamiento, mediante la técnica de remuestreo *bootstrap*. Esta técnica consiste en la generación de varias muestras con reemplazo, del conjunto de datos de entrenamiento, donde a cada observación se le asigna el mismo peso ($\frac{1}{n}$, siendo n el número de observaciones). Al número de muestras *bootstrap* se le suele denotar B.

A la hora de utilizar este método en problemas de regresión, se procede a tomar varias muestras *bootstrap*, donde a partir de cada una de ellas se construye un *Árbol de Regresión*. Luego, se le asigna a la observación el promedio de las respuestas de los *Árboles de Regresión* construidos en cada muestra.

Se destaca que el algoritmo a la hora de construir los diferentes estimadores (árboles), no considera en cada partición el total de variables, sino un subconjunto de estas elegido de forma aleatoria. Como primera aproximación se procedió a utilizar la parte entera de $\sqrt{p}$, siendo p el número de variables. En etapas posteriores del análisis, se modificó el valor del mismo con el fin de obtener un mayor poder predictivo.

Siendo este último punto lo que diferencia al algoritmo con su versión más simple denominada *Bagging* (también desarrollada por *Breiman*). En este último se

considera en cada división el total de las variables, por lo que resulta ser un caso particular del método *Random Forest*.

A su vez, a cada árbol ajustado no se le realiza un proceso de poda. Por lo que estos mismos presentan una gran varianza, pero bajo sesgo. Sin embargo, al predecir mediante un promedio de los B árboles, se logra una reducción considerable en la varianza del estimador y de esta forma se mejora la precisión de la predicción. (Hastie, 2001) ([16])

Por lo tanto, a modo de resumen el algoritmo utilizado para realizar el ajuste mediante el método *Random Forest* se expresa de la siguiente forma:

- 1) Sea B el número de muestras bootstrap, entonces para $b = 1, \ldots, B$

  (a) Se obtiene una muestra bootstap Z^b de tamaño n del conjunto de entrenamiento

  (b) Se ajusta un árbol a los datos obtenidos en Z^b, de forma recursiva mediante la repetición de los siguientes pasos en cada nodo terminal hasta que esten conformados por un número mínimo de observaciones:

  (b_1) De forma aleatoria se selecciona m variables de las posibles p variables

  (b_2) De las m variables obtenidas, se selecciona la mejor variable y el mejor punto de corte

  (b_3) Se realiza una partición en dicho nodo generado dos nodos hijos.

- 2) Se obtiene un conjunto de árboles $\{T_b\}_1^B$

De esta forma, en este método se tiene que el estimador toma la siguiente forma:

$$\hat{f}(X) = \frac{1}{B} \sum_{b=1}^{B} T(X, \Theta_b)$$

Donde Θ_b caracteriza el b-ésimo árbol aleatorio en términos de la muestra *bootstrap* utilizada, las variables de partición, los puntos de corte en cada nodo y los valores de predicción en cada nodo terminal.

Se optó por trabajar con el método *Random Forest* ya que se destaca sobre el método *Bagging* principalmente cuando se tiene que una variable es muy influyente. Esto se debe a que si en cada partición se consideran todas las variables a la hora de construir los diferentes B árboles, en la medida de que se tiene una variable muy influyente, posiblemente dichos árboles no difieran mucho entre sí. Esta limitante no se presenta en *Random Forest* en el sentido de que selecciona de forma aleatoria un subconjunto de los predictores en cada iteración. (Hastie, 2001) ([16])

Una característica relevante del método *Random Forest* (al igual que en *Bagging*), es que cada observación posee cierta probabilidad de ser seleccionada en cada remuestra realizada. De esta forma, se cuenta con un conjunto de observaciones las cuales no son utilizadas para construir el estimador.

Este conjunto de observaciones se denomina como *out of bag observations* (OOB). Por lo tanto, en cada iteración se procede a predecir dichas observaciones, mediante el estimador obtenido. Repitiendo este procedimiento para las n observaciones, se calcula el *error OOB*. Dicha medida se procedió a utilizar como una primera aproximación en cuanto a la performance predictiva del modelo.

A pesar de que el método anteriormente mencionado logra solucionar el problema de la inestabilidad por parte de los *Árboles de decisión*, este método se caracteriza por presentar una baja interpretabilidad.

4.1.4. Boosting

Como segunda alternativa para obtener un modelo de predicción estable, se trabajó con el método de agregación basado en *Árboles de Decisión*, *Boosting*. Este método, al igual que los presentados en la Sección 4.1.3, consiste en la combinación de la salida de varios estimadores con el fin de producir un estimador más preciso.

Sin embargo, el mismo difiere con los métodos anteriores en la forma de realizar este proceso. Mientras que en *Random Forest* (*Bagging*) se procede a construir varios árboles mediante diferentes conjuntos de entrenamiento y combinando la predicción de cada uno, en *Boosting* se trabaja mediante un enfoque secuencial. (James, 2013)

([18])

En este método, se construye una sucesión de estimadores, los cuales surgen de forma iterativa usando una modificación del conjunto de datos realizada a partir de la performance del estimador en el paso anterior. De esta forma, en cada iteración, se toma como variable de salida los residuos del modelo anterior y no a la variable de respuesta original (Y).

Esto último, con el fin de realizar un proceso de actualización de los residuos del modelo y por lo tanto mejorando la predicción del estimador en áreas donde el mismo no realiza un buen ajuste. (Hastie, 2001) ([16])

Generalmente, en *Boosting* cada árbol suele estar conformado por pocas particiones, por lo que el procedimiento de aprendizaje suele ser "lento". (James, 2013) ([18])

De forma resumida, el álgoritmo consiste en aplicar los siguientes pasos de forma iterativa:

1. Se establece $\hat{f}(X) = 0$ y $r = Y$ en el conjunto de entrenamiento.

2. Para cada $v = 1, 2, \cdots, V$ se repite:

a. Se ajusta un árbol $\hat{f}^v(X)$ con d particiones (es decir, $d+1$ nodos terminales) en los datos de entrenamiento (X,r).

b. Se actualiza $\hat{f}$ agregando el nuevo árbol en una versión reducida:

$$\hat{f}(X) \leftarrow \hat{f}(X) + \lambda \hat{f}^v(X)$$

c. Se actualizan los residuos

$$r \leftarrow r - \lambda \hat{f}^v(X)$$

3. Se genera el modelo

$$\hat{f}(X) = \sum_{v=1}^{V} \lambda \hat{f}^v(X)$$

Donde V denota el número de árboles utilizados en el algortimo. Se destaca que, a diferencia de *Random Forest*, *Boosting* puede generar un sobreajuste a los datos

en caso que V sea grande, a pesar de que éste sobreajuste ocurra de manera lenta. (James, 2013) ([18])

Por otro lado, el parámetro λ controla la tasa a la que aprende el algoritmo. Donde λ suele ser un número positivo pequeño, usualmente 0.01 o 0.001. A su vez, se tiene que generalemente, cuanto menor el valor de λ, mayor el número de árboles (V) necesarios. (Hastie, 2001) ([16])

Por último, la letra d denota el número de particiones en cada árbol, dicho parámetro controla la complejidad de cada estimador. Se observa que en el caso $d = 1$ implica que cada árbol tenga solamente una partición y de esta forma se cuenta con un modelo aditivo (cada término involucra una sola variable).

A su vez, el parámetro d puede ser intepretado también como el parámetro que controla el orden de interacción entre los modelos, ya que las d particiones pueden involucrar a los sumo d variables. (James, 2013) ([18])

Dichos parámetros (V, λ y d) se estiman mediante una metodología de *validación cruzada*.

4.1.5. Support Vector Regression (SVR)

Los modelos denominados *Support Vector Regression* (SVR), surgen como una generalización aplicada a problemas de regresión de los modelos *Support Vector Machine* (SVM). (Kuhn, 2013) ([21])

Por lo tanto, al ser una generalización de los SVM (para problemas de regresión), poseen características muy similares, principalmente la robustez en cuanto a observaciones atípicas. De esta forma, se tiene que los SVR pertenecen al grupo denominado *robust regression*, donde en estos métodos se busca minimizar el efecto de observaciones atípicas en la ecuación de regresión. (Kuhn, 2013) ([21])

Estos métodos surgen como alternativa a los modelos de regresión lineal, ya que estos últimos a la hora de estimar los parámetros buscan minimizar la suma de cuadrados residuales (SCR). Lo cual conlleva que una observación que no sigue la tendencia del resto, puede ser muy influyente. (Kuhn, 2013) ([21])

A pesar de que existen varios enfoques para llevar acabo SVR en este trabajo se centró en el denominado ϵ-*insensitive regression* (Kuhn, 2013) ([21]). En este contexto, a la hora de obtener las estimaciones de los parámetros del modelo, se define una nueva función de pérdida denominada ϵ-*insensitive loss function* (Vapnik, 1999) ([35]), siendo de la forma:

$$L(Y, f(X,\beta)) = L(|Y - f(X,\beta)|_\epsilon)$$

$$|Y - f(X,\beta)|_\epsilon = \begin{cases} 0, & \text{si } |Y - f(X,\beta)| \leq \epsilon \\ |Y - f(X,\beta)| - \epsilon, & \text{en otro caso} \end{cases}$$

En función a la ecuación anterior, se tiene que la perdida es igual a 0 si la discrepancia entre los predicho y lo observado es menor a ϵ, siendo ϵ un umbral establecido en forma previa. Por lo tanto se tiene que tanto las observaciones atípicas, como las observaciones que poseen un buen ajuste (residuos pequeños), no tienen efecto en la ecuación de regresión.

En este contexto, para estimar los parámetros del modelo, SVR utiliza la función de perdida anteriormente definida, pero a su vez considerando un parámetro de penalización. En dicho método se busca obtener los coeficientes que minimizan la siguiente expresión:

$$C \sum_{i=1}^{n} L(|y_i - f(x_i,\beta)|_\epsilon) + \sum_{j=1}^{P} \beta_j^2$$

Donde el parámetro C es un parámetro de penalización, el cual generalmente se estima mediante *validación cruzada*. En este contexto el parámetro C cumple un rol de indicar la complejidad del modelo. Conforme aumenta el valor de este el modelo obtiene mayor flexibilidad, en la medida que el efecto de los errores es aumentado. Por otro lado, al disminuir este parámetro el modelo se vuelve más rígido y con menor posibilidad de sobre ajustar a las observaciones.

Luego, se tiene que la solución al problema de minimización anteriormente mencionado, involucra el producto escalar entre las observaciones y no a las observaciones en si (Hastie, 2001) ([16]). De esta forma, se puede re expresar a la función de regresión mediante la siguiente expresión:

$$f(x^*) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x^*, x_i \rangle$$

Por lo tanto, se tiene que para evaluar $f(x^*)$ es necesario el cálculo del producto escalar entre la nueva observación (x^*) y cada una de las observaciones pertenecientes al conjunto de entrenamiento. A su vez, se cuenta con n parámetros α_i con $i = 1, \ldots, n$, donde cada uno corresponde a una observación de entrenamiento.

Sin embargo, en SVR se tiene la propiedad de que solo un subconjunto de los datos tiene un rol activo en la predicción de una nueva observación. Esto en la medida de que los parámetros α_i asociados a las observaciones de entrenamiento las cuales se encuentran a $\pm\ \epsilon$ de la recta de regresión (es decir se encuentran dentro del intervalo de longitud 2ϵ alrededor de la recta de regresión) son iguales a 0. (Kuhn, 2013) ([21])

A las observaciones las cuales determinan a la recta de regresión se les denomina *support vectors*. Además, en la medida de que el predictor se encuentra sujeto al producto escalar entre la nueva observación y las observaciones de entrenamiento (en particular solo aquellas que sean *support vectors*), se puede generalizar con el fin de captar relaciones no lineales entre las variables.

Para ello se utiliza una función denominada *kernel* la cual permite agrandar el espacio original de las variables, con el fin de obtener relaciones lineales en un nuevo espacio de mayor dimensión (James, 2013) ([18]). Esta función es una generalización del producto escalar y se denota de la siguiente forma:

$$K(x_i, x_j)$$

De esta forma el predictor queda expresado como:

$$f(x^*) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x^*, x_i)$$

A la hora de aplicar SVR existen diferentes *kernels* lo cuales se podrían utilizar. Uno de lo más utilizados en la bibliografía se denomina *radial kernel*. Este es de la forma:

$$K(x^*, x_i) = exp\left(-\gamma \sum_{j=1}^{p} (x_j^* - x_{ij})^2\right), \gamma > 0$$

En donde si la observación x^* se encuentra lejos de la observación x_i en terminos de distancia euclídea, entonces se tiene que $\sum_{j=1}^{p} (x_j^* - x_{ij})^2$ es una cantidad grande y por consecuente $K(x^*, x_i)$ es pequeño. Por lo tanto, x_i no va a tener un rol activo a la hora de predecir el valor de x^*.

Esto significa que el *radial kernel* posee un comportamiento local, en el sentido de que las observaciones de entrenamiento cercanas tienen un mayor efecto en la predicción del valor de una nueva observación.

Por otro lado, se tiene que γ es un parámetro de escala, el cual afecta la varianza en la estimación. Al igual que C, dicho parámetro generalmente se estima mediante *validación cruzada*.

Luego, se destacan dos aspectos de los modelos SVR. En primer lugar, en el caso de que la relación entre las variables sea realmente lineal (problemas de regresión), se recomienda usar un *linear kernel* (producto escalar) sobre un *radial kernel*. (Kuhn, 2013) ([21])

A su vez, en la medida de la ecuación de regresión ($f(X)$) se expresa a través del producto escalar entre las observaciones, se recomienda estandarizar las mismas con el fin de tener una misma unidad de medida. (Kuhn, 2013) ([21])

4.2. Validación cruzada y parámetros de ajuste

A la hora de evaluar la performance de los diferentes modelos planteados, se realizó un procedimiento de *validación cruzada*, particularmente *k-folds*.

El algoritmo consiste en dividir la muestra en k submuestras de igual tamaño. Luego $k-1$ submuestras se usan como datos de entrenamiento y la muestra restante k se usa para testear los datos.

A continuación, se procede a ajustar los datos de esa muestra con el modelo construido con las $k-1$ muestras. Donde el proceso se repite k veces, con cada una

de las k muestras. De tal forma que cada una de las k muestras es utilizada una sola vez como datos de testeo. (James, 2013) ([18])

De esta forma, todas las observaciones se usan tanto para entrenar (*train*) como para testear (*test*). A su vez, cada observación se usa para *test* una sola vez y para *train* $k-1$ veces. Los errores obtenidos en cada etapa se promedian para producir una sola estimación (error medio obtenido de los k análisis realizados).

Con el fin de medir el error de predicción del modelo en los modelos planteados anteriormente se consideró la siguiente medida:

$$RMSE = \frac{1}{k}\sum_{i=1}^{k} RMSE_i$$

Donde $RMSE_i$ es la raiz del error cuadratico medio en la i-ésima muestra.

$$RMSE_i = \sqrt{\frac{1}{n_i}\sum_{j=1}^{n_i}(y_j - \hat{y}_j)^2}$$

Siendo n_i es la cantidad de observaciones en la i-ésima muestra.

Asimismo, como medida de la performance predictiva se utilizó también el error absoluto medio:

$$MAE = \frac{1}{k}\sum_{i=1}^{k} MAE_i$$

Donde MAE_i es el error absoluto medio en la i-ésima muestra.

$$MAE_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(|y_j - \hat{y}_j|)$$

Nuevamente, siendo n_i la cantidad de observaciones en la i-ésima muestra.

Adicionalmente, para el modelo con mayor performance predictiva se obtuvo una medida del error de predicción en términos relativos. Para ello se computó el error porcentual absoluto medio (EPAM):

$$EPAM = \frac{1}{k}\sum_{i=1}^{k} EPAM_i$$

Donde $EPAM_i$ es el error porcentual absoluto medio en la i-ésima muestra.

$$EPAM_i = \frac{1}{n_i} \sum_{j=1}^{n_i} |\frac{(y_j - \hat{y}_j)}{y_j}|$$

Por otro lado, con el fin de obtener el modelo con la mejor performance predictiva, se realizó un proceso de selección de los *parámetros de ajuste*. El mismo consiste en obtener la mejor combinación de parámetros de ajuste posibles mediante una metodología de *validación cruzada*.

En este contexto, se define a un *parámetro de ajuste* como un valor necesario para ajustar un modelo el cual no se determina a partir de los datos sino que, por el contrario, es necesario que sea especificado previamente a la realización del ajuste. Dependiendo del algoritmo con el que se trabaje, el rol de los mismos puede variar.

4.3. Interpretabilidad

Una vez implementados los diferentes algoritmos de aprendizaje estadístico y ajustados los *parámetros de ajuste*, se procedió a explorar métodos para interpretar los resultados de los modelos de aprendizaje estadístico utilizados.

Para ello, se trabajó con *métodos globales modelo-agnósticos* de interpretación, aplicados a los modelos de aprendizaje estadístico y principalmente a aquel modelo con mejor performance predictiva, haciendo hincapié en el análisis gráfico.

Estos métodos de interpretación para modelos de caja negra consisten en describir el compartamiento promedio del modelo. Donde, los mismos generalmente se expresan mediante un valor esperado, basado en la distribución de los datos. (Molnar, 2020) ([26])

A pesar de que en la literatura existen diferentes métodos para llevar a cabo el análisis, se trabajó mediante dos aproximaciones.

Como primera aproximación, se implementó un análisis sobre la importancia de cada predictor, mediante la metodología de *importancia de las variables permutadas*. Luego, se realizó el análisis mediante los gráficos denominados *Partial Dependece Plot* (PDP).

En las siguientes subsecciones se detallan los principales aspectos teóricos de las metodologías, al igual que sus ventajas y limitantes.

4.3.1. Importancia de las variables permutadas

La metodología de *importancia de las variables permutadas* consiste en obtener una medida de importancia de determinada variable de entrada, calculando el aumento en el error de predicción obtenido al permutar los valores en dicha variable.

De esta forma, se tiene que si la variable es relevante, al permutar los valores de la misma de forma aleatoria en el conjunto de entrenamiento, se genera un aumento en el error de predicción (esto debido a que al permutar los valores de la variable se deshace cualquier relación entre la variable y la variable de respuesta) (Greenwell, 2020) ([13])

A la hora de llevar acabo esta metodología, en primer lugar se computa una métrica de la performance predictiva del modelo obtenida sin alterar el conjunto de entrenamiento. Luego, se computa dicha métrica permutando para cierta variable de interés, sus valores en el conjunto de entrenamiento. Esto con el fin de observar la diferencia entre el valor original de la métrica y el nuevo valor obtenido.

A modo de resumen, si denotamos $X_1, \ldots, X_j$ como las variables de interés, entonces el algortimo consiste en:

1) Para cada variable de interés X_i con $i = 1, \ldots, j$:

a) Se permuta los valores de la variable X_i en el conjunto de entrenamiento

b) Se calcula la performance predicitva mediante una métrica predefinida en a).

c) Se calcula la diferencia entre la medida de error original y la obtenida en b)

2) Se ordena a cada variable de forma decreciente en función de c)

Por ende, se tiene que la variable es relevante si al permutar sus valores, aumenta el error de predicción del modelo. Por su contraparte, una variable no es relevante, si al permutar sus valores el error del modelo permanece incambiado. (Greenwell, 2020) ([13])

Se destaca que si se cuenta con variables altamente correlacionadas, este método

posee una gran limitante. En la medida de que al permutar los valores de una sola variable, pueden generarse observaciones las cuales no son factibles. (Greenwell, 2020) ([13])

4.3.2. Gráficos de dependencia parcial (PDP)

Los *gráficos de dependencia parcial* (PDP) permiten observar el efecto marginal que una o dos variables tienen sobre la predicción de la variable de respuesta obtenida a través de un algoritmo de aprendizaje estadístico. (Molnar, 2020) ([26])

Estos gráficos pueden detectar cuando la relación entre la variable de respuesta y la variable predictora de interés es lineal, monótona, o bien cuando se trata de una forma funcional más compleja.

La *función de dependencia parcial* para el caso de regresión se define como:

$$\begin{aligned}\hat{f}_{X_S,PDP}(X_S) &= E_{X_C}\left[\hat{f}(X_S, X_C)\right] \\ &= \int_{X_C} \hat{f}(X_S, X_C)\mathbb{P}(X_C)dX_C\end{aligned}$$

En donde X_S son las variables para las cuales se quiere conocer el efecto sobre la predicción, mientras que X_C corresponde al resto de las variables utilizadas en el algoritmo de aprendizaje estadístico $\hat{f}$.

Ahora bien, la estimación para la función anterior se obtiene mediante la metodología *Monte Carlo* promediando sobre la muestra de entrenamiento:

$$\hat{f}_{X_S,PDP}(X_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

En donde $x_C^{(i)}$ son los valores en la base de datos para las variables que no son de interés y n el número de observaciones.

Se destaca que PDP es un método que permite determinar de manera global la relación entre una o dos variables predictoras sobre la variable de respuesta. Asimismo, interesa destacar que uno de los supuestos de PDP es que las variables en X_C y X_S no están correlacionadas. (Molnar, 2020) ([26])

De manera similar se obtiene una estimación PDP para el caso de variables categóricas en donde, para cada categoría, se realiza la estimación PDP forzando a que todas las observaciones tomen el valor correspondiente a dicha categoría. (Molnar, 2020) ([26])

Una de las ventajas principales de PDP es que es un método intuitivo y en el caso de variables incorrelacionadas, tiene una interpretación clara. El gráfico PDP permite observar cómo cambia la predicción promedio cuando la h-ésima variable predictora cambia. No obstante este resultado no es tan claro en el caso de correlación entre las variables con las que se construye el gráfico, ya que se construyen conbinaciones de las variables que son irreales o con probabilidad muy baja. Esto ocurrre ya que la función PDP en un punto en particular se obtiene forzando a que todos las observaciones tomen dicho valor en particular, lo cual puede no tener sentido en algunos casos. (Molnar, 2020) ([26])

Por último, se destaca que, para el caso de las variables de naturaleza contínua, debido al alto costo computacional que conlleva la realización de esta metodología, a la hora de obtener los resultados se procedió a trabajar con una grilla de tamaño 20 sobre el recorrido de cada variable utilizando puntos de corte equidistantes.

4.4. Tratamiento de datos faltantes

La base de datos construida contiene variables con diferentes grado de datos faltantes.

Este problema fue abordado siguiendo dos estrategias de imputación. En primera instancia se entrenan diferentes modelos imputando solamente a las variables numéricas que tienen proporción de valores faltantes inferior a 0.15, y se procede en esta primera instancia a realizar imputación por la media, lo cual es equivalente a asumir que el proceso generador de datos de estos valores es un proceso aleatorio. Es decir, que los datos faltantes son generados al azar.

Posteriormente, se realizó un proceso de imputación de valores faltantes mediante

un análisis supervisado. Para ello se trabajó de la siguiente manera: 1) Se ajustó un modelo tomando como variable de salida cada variable con datos faltantes, 2) en cada uno de ellos, se consideró como variables de entrada todas aquellas que originalmente no presentan datos faltantes, pero excluyendo la variable precio y 3) en cada variable, se sustituyó cada dato faltante por su predicción utilizando el modelo ajustado.

En particular, el algoritmo utilizado para ajustar los modelos fue *Random Forest*, mediante el paquete *missRanger* ([25]). En lo que respecta a los *parámetros de ajuste*, se destaca que debido al tiempo computacional que conlleva fueron utilizados en todos los casos los valores por defecto.

Como fue mencionado anteriormente el criterio genérico para seleccionar las variables a imputar fue según la proporción de valores faltantes (proporción inferior a 0.15 para las variables de tipo numéricas). Sin embargo, esta metodología de imputación, a diferencia de la anterior, permite realizar el proceso para variables de tipo cualitativas. De esta forma, se incluyó en el análisis la variable condición que indica si el apartamento es nuevo o usado. No obstante, si bien existen otras variables que podrían ser incluídas, se decidió dejarlas fuera del análisis en la medida que no se consideran relevantes para el mismo (principalmente por presentar grupos no balanceados).

En la Tabla A.1 de la Sección del Anexo A.1 se detalla la proporción de valores faltantes en las variables utilizadas a la hora de ajustar los modelos. Por otra parte, en la Tabla A.1 de la Sección del Anexo A.1 se presenta el listado de variables con el o los métodos de imputación de valores faltantes implementado.

Capítulo 5

Reproducibilidad

En este Capítulo se detalla brevemente las características tenidas en cuenta para lograr algunos aspectos de la reproducibilidad de los resultados obtenidos al igual que modificaciones de los mismos.

En primer lugar, en lo que respecta a los datos utilizados, se destaca que los mismos se encuentran disponibles en `https://drive.google.com/drive/folders/1uTIr9JVc5SZS2bOVG4faexfW0q4m9sVL?usp=sharing`. Asimismo, se trabajó con un repositorio remoto público disponible en `https://github.com/alvarovalinio/TFG`, donde se encuentra disponible el código necesario para la obtención de nuevos datos siguiendo los pasos detallados en la Sección 2.1.

A su vez, en dicho repositorio se encuentra disponible en diferentes scripts del lenguaje de programación *R* ([32]) el código necesario para realizar la implementación de todos los modelos estadísticos utilizados.

Por otro lado, como se detalla en el Capítulo 4 varias de las técnicas implementadas involucran un proceso de simulación de valores aleatorios. De esta forma se consideró apropiado trabajar con una única semilla.

Luego, según se menciona en el Capítulo 1, con el fin de obtener un mayor poder de cómputo en los algoritmos implementados, se utilizó un enfoque desde la programación en paralelo. Se destaca que este enfoque presenta ciertas limitantes en cuanto a la reproducibilidad de los resultados obtenidos. En este sentido, los

resultados presentados en la Sección 7.6 fueron obtenidos utilizando un total de dos núcleos y estableciendo una semilla en cada uno de ellos. Con el fin de llevar a cabo este cometido se trabajó con el paquete de *R doParallel* ([7]).

Por último, el informe se realizó utilizando el sistema *Sweave* que permite generar reportes dinámicos incorporando código de *R* con documentos de Latex. El mismo se encuentra también disponible en el repositorio remoto público `https://github.com/alvarovalinio/TFG`.

Capítulo 6

Análisis exploratorio de datos

En este Capítulo se presenta los principales resultados obtenidos a la hora de realizar el análisis exploratorio de los datos, luego de realizar el proceso de depuración de los mismos detallado en la Sección 2.2 . De esta forma, la base de datos está conformada por un total de 76667 observaciones y 49 variables.

Por otra parte, una vez realizado el proceso de tratamiento de datos faltantes especificado en la Sección 4.4 se mantiene un total de 24 variables en el caso de imputación de valores faltantes por la media, y 25 variables en el caso de imputación de valores faltantes por *Random Forest*.

En la Tabla A.1 disponible en la Sección del Anexo A.1 se detalla el listado de variables utilizadas para la implementación de los diferentes modelos en etapas posteriores del análisis.

Una vez obtenida la base de datos luego de realizar los procesos mencionados anteriormente, se procedió a analizar el comportamiento de la variable de respuesta precio de oferta en dólares. A continuación se presenta en la Tabla 6.1 las principales medidas de resumen:

Min	Q1	Mediana	Media	Q3	Max	Desvio	CV
29,000	115,000	145,500	163,572	186,000	450,000	76,788	0.47

Tabla 6.1: Medidas de resumen de la variable precio de oferta en dólares. En la medida que la media es superior a la mediana se tiene una asimetría hacia la derecha. Por otro lado, se observa un desvío de 76,788 dólares estadounidenses y un coeficiente de variación de 0.47.

En la Figura 6.1 se presenta gráficamente la distribución de la variable mediante un gráfico de histograma.

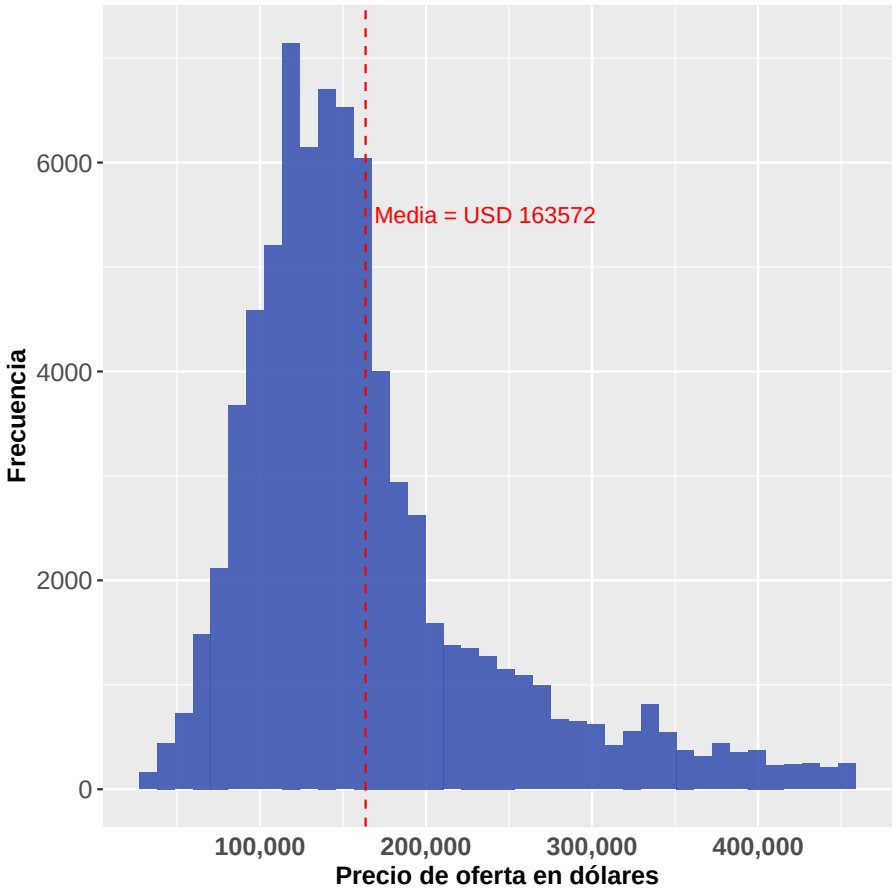

Figura 6.1: Histograma del precio de oferta en dólares de los apartamentos a la venta en Montevideo, Uruguay. El precio promedio es de 163572 dólares. Como fue comentado en la Tabla 6.1 se observa una asimetría a la derecha en la distribución.

Por otra parte, en la Figura 6.2 se presenta el mapa de Montevideo con los puntos dónde se encuentran ubicados los apartamentos, luego de haber implementado los criterios de depuración detallados en la Sección 2.2. El color del punto indica si el precio de oferta del apartamento es superior o inferior al precio promedio en los datos (163572 dólares), siendo este último de color celeste.

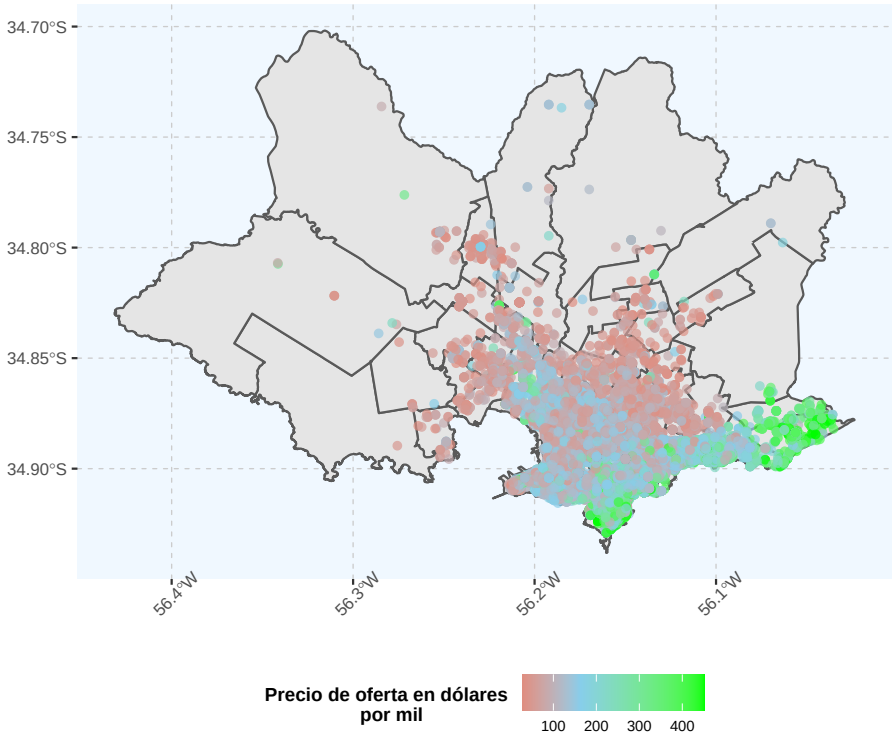

Figura 6.2: Mapa de puntos donde se encuentran ubicados los apartamentos en la base de datos obtenida de *Mercado Libre*. El color del punto indica si el precio de oferta del apartamento es superior (color verde) o inferior (color rojo) al precio promedio en los datos (163572 dólares), siendo este de color celeste.

Como se observa en la Figura 6.2, el precio de oferta en dólares de los apartamentos se incrementa conforme aumenta la proximidad de la ubicación de los mismos a la rambla este de Montevieo. En contraparte, se observa que el precio disminuye para los apartamentos que se encuentran en barrios ubicados hacia el norte de Montevideo. Por otro lado, se observa una mayor concentración de apartamentos en barrios ubicados hacia el sur de Montevideo.

En lo que respecta a las variables de entrada, en primer lugar se realizó un análisis exploratorio en cuanto a las variables de naturaleza cualitativa. A continuación se presenta en la Figura 6.3 la distribución de los niveles en los datos para un conjunto de variables mediante gráficos de barras.

Según se observa en la Figura 6.3 la variable zona Avd. Italia posee niveles balanceados en contraposición a las variables el edificio tiene piscina y cantidad de baños completos. Por su parte, la variable cantidad de dormitorios presenta mayor proporción de observaciones hacia los niveles centrales (uno y dos dormitorios).

Asimismo, con el fin de observar la covariación entre la variable precio de oferta en dólares y las variables presentadas en la Figura 6.3, se presentan en la Figura 6.4 los gráficos de caja y gráficos de violín correspondientes.

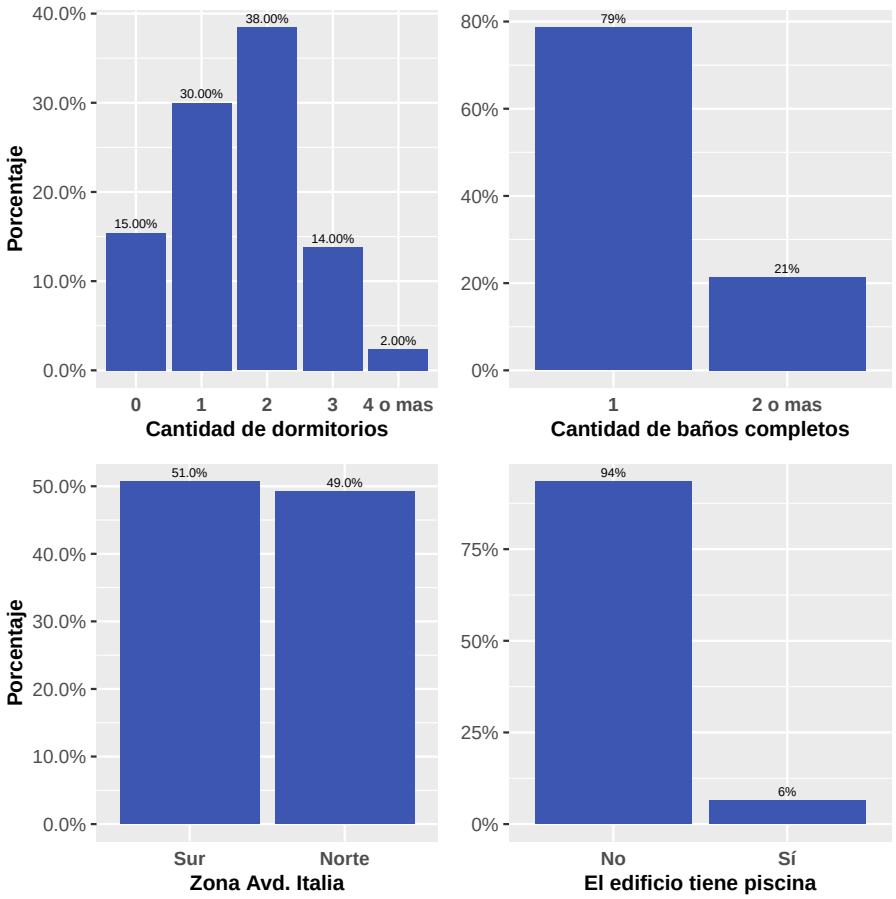

Figura 6.3: Gráfico de barras para diferentes variables cualitativas. En el panel superior izquierdo se encuentra graficada la variable cantidad de dormitorios mientras que en el panel superior derecho la variable cantidad de baños completos. Por otro lado, en el panel inferior izquierdo se encuentra la variable zona Avd. Italia mientras que en el panel inferior derecho la variable el edificio tiene piscina.

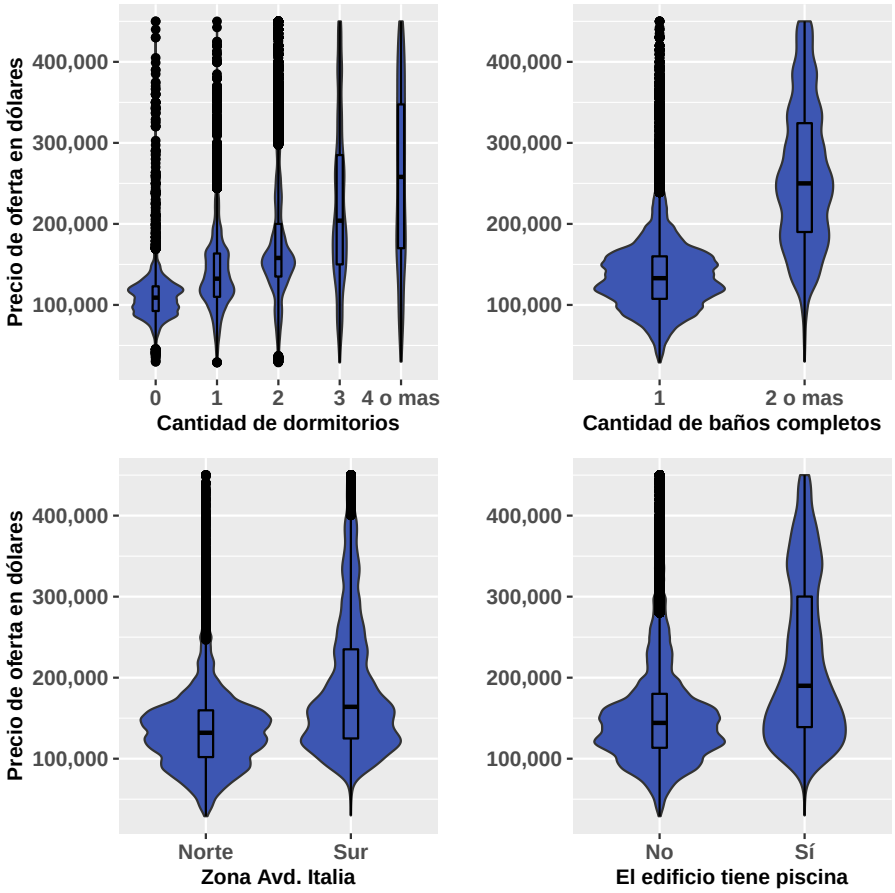

Figura 6.4: Gráfico de violín y gráfico de caja del precio de oferta en dólares según diferentes variables cualitativas. En el panel superior izquierdo se encuentra graficada la variable cantidad de dormitorios mientras que en el panel superior derecho la variable cantidad de baños completos. Por otro lado, en el panel inferior izquierdo se encuentra la variable zona Avd. Italia mientras que en el panel inferior derecho la variable el edificio tiene piscina.

Como se observa en la Figura 6.4, tanto la mediana como la dispersión del precio de oferta en dólares se incrementa conforme aumenta la cantidad de dormitorios del apartamento. Dicho comportamiento también se observa para la cantidad de baños completos.

Por otra parte, los apartamentos ubicados al sur de la calle Avenida Italia en continuación con la calle Avenida 18 de Julio presentan una mediana y dispersión superior del precio de oferta en dólares, respecto a los apartamentos ubicados al

norte.

En lo que respecta a los apartamentos con piscina en el edificio, de manera similar presentan una mediana y dispersión superior del precio de oferta en dólares, respecto a los apartamentos que no tienen piscina en el edificio.

Por último, se procedió a realizar un análisis gráfico bivariado en cuanto a la variable precio de oferta en dólares. De esta forma, en la Figura 6.5, se presenta la distribución de la misma teniendo en cuenta los efectos de las variables zona Avd. Italia y cantidad de baños completos. Mientras que en la Figura 6.6 de forma análoga, se observa los efectos de las variables distancia a la rambla este y cantidad de baños completos.

Se destaca que en lo que respecta a la variable distancia a la rambla este, se realizó la siguiente discretización para poder tener una aproximacón de su efecto sobre la variable precio de oferta en dólares:

- Inferior al primer cuantil: valores inferiores a 575 (metros).
- Entre primer y segundo cuantil: para valores en el intervalo [575, 2465) (metros).
- Entre segundo y tercer cuantil: para valores en el intervalo [2465, 3598) (metros).
- Igual o superior al tercer cuantil: valores iguales o superiores a 3598 (metros).

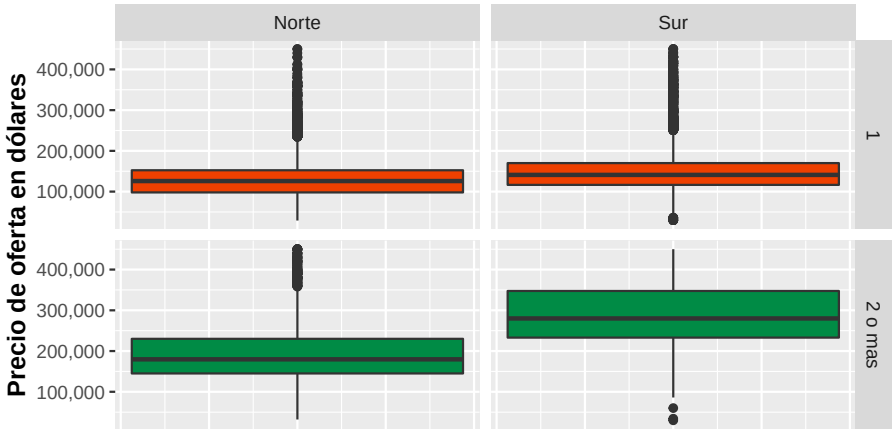

Figura 6.5: Gráfico de caja del precio de oferta en dólares según zona Avd. Italia y cantidad de baños completos. El gráfico sugiere que las diferencias en mediana observardas en la Figura 6.4 se mantienen cuando condicionamos en cantidad de baños completos.

En particular, la Figura 6.5 muestra que entre los apartamentos con dos o más baños completos, se observa una mediana superior para los apartamentos ubicados al sur respecto a los apartamentos ubicados al norte. En lo que respecta a los apartamentos con un solo baño completo, se observa que los apartamentos al sur poseen una mediana superior respecto a los ubicados al norte, si bien la diferencia es menor en este último caso.

En primer lugar, se observa en la Figura 6.6 que el precio de oferta en dólares de los apartamentos se ve diferenciado según la cantidad de baños completos. Esto último en la medida que, condicionando por todos los niveles de la variable distancia a la rambla este discretizada, la mediana de los apartamentos con dos o más baños completos es superior a aquellos con un solo baño completo.

Por otro lado, entre los apartamentos con dos o más baños completos, el precio se ve diferenciado si se considera la discretización de la variable distancia a la rambla este en función de los cuantiles de la misma. Esto último debido a que la mediana del precio de los apartamentos de los dos primeros niveles es mayor que la de los últimos.

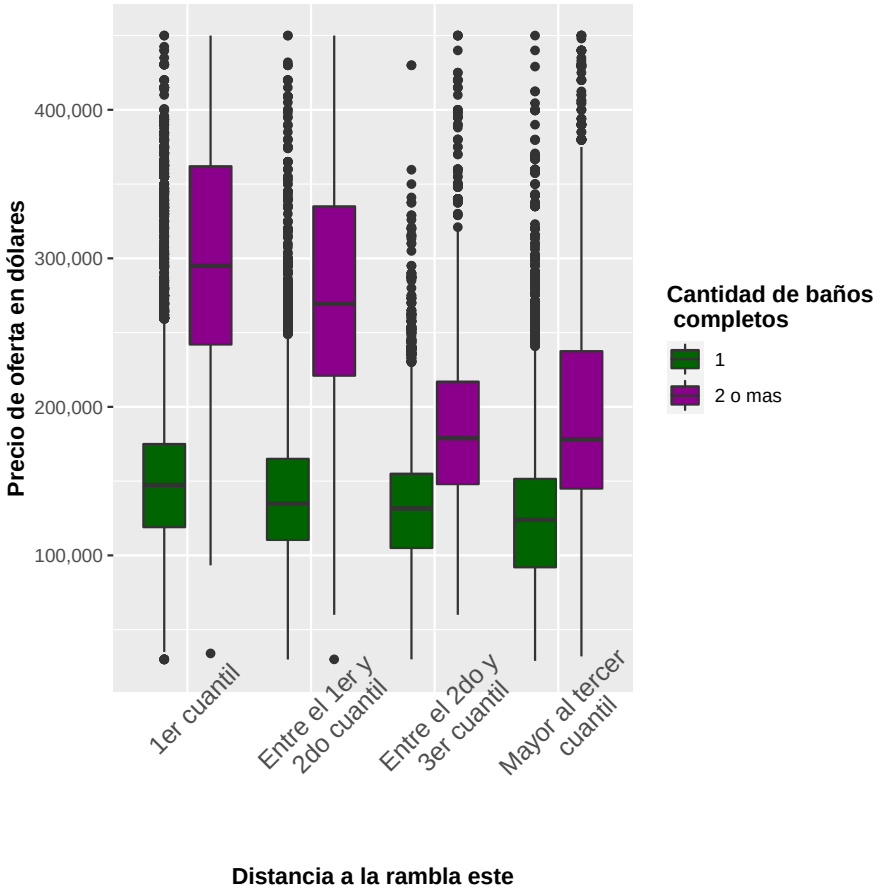

Figura 6.6: Gráfico de caja del precio de oferta en dólares según distancia a la rambla este y cantidad de baños completos utilizando una discretización de la variable distancia a la rambla este mediante los cuantiles.

Por otro lado, en lo que respecta al análisis exploratorio para las variables cuantitativas, se procedió a analizar la correlación lineal entre las mismas utilizando el *coeficiente de correlación lineal de Pearson*. Para ello, en la Figura 6.7 se presenta a modo de resumen la matriz de correlación de manera gráfica.

En primer lugar, en cuanto a la variable precio de oferta en dólares se observa en la Figura 6.7 que las variables área total, área cubierta y distancia a la rambla este se encuentran linealmente correlacionadas de forma moderada. Donde para las dos primeras esta relación es de carácter positivo mientras que para la última es de carácter negativo.

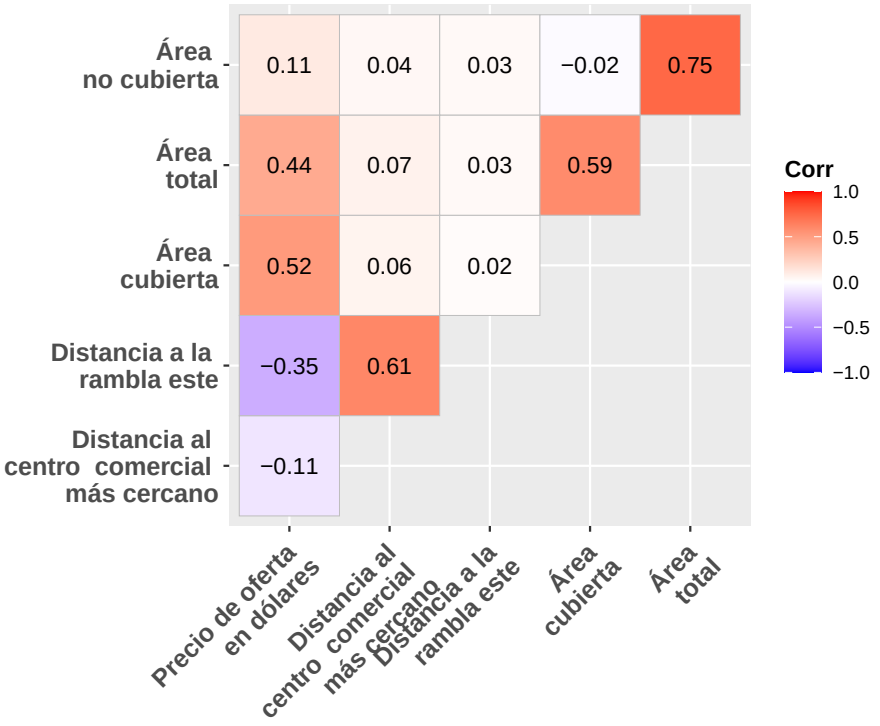

Figura 6.7: Gráfico de matriz de correlación de las variables precio de oferta en dólares, distancia al centro comercial más cercano, distancia a la rambla este, área total, área cubierta, y área no cubierta.

Por otro lado, se tiene que tanto distancia al centro comercial más cercano como distancia a la rambla este, no se encuentran linealmente correlacionadas con área total, al igual que el área cubierta y área no cubierta. Por último, se observan correlaciones lineales positivas las cuales son producto de la construcción misma de las variables. Por ejemplo, la correlación lineal positiva entre área total y área cubierta.

Capítulo 7

Resultados

En esta Sección se presentan los principales resultados obtenidos luego de realizar la implementación de las técnicas detalladas en el Capítulo 4.

En primer lugar, en la Sección 7.1 se tienen los resultados obtenidos luego de relizar un ajuste mediante un *Modelo de Regresión Lineal Múltiple*. Luego, en las Secciones 7.2, 7.3, 7.4 y 7.5 se exponen los resultados de la implementación de las técnicas de aprendizaje estadístico explicitadas en el Capítulo 4.

Con el fin de tener una primera aproximación de la performance predictiva de los modelos, se trabajó con una metodología de muestra de entrenamiento y muestra de testeo. Para ello, se consideró el 80 % de los datos como muestra de entrenamiento y el restante 20 % como muestra de testeo.

A continuación, en forma de resumen, se muestra en la Sección 7.6 un análisis comparativo de los diferentes modelos en función de su performance predictiva mediante la metodología de *k-folds* y el proceso de selección de los *parámetros de ajuste* especificada en la Sección 4.2. En particular, se trabajó con *5-folds*.

Por último, se presenta en la Sección 7.7 el análisis gráfico de interpretabilidad según lo detallado en la Sección 4.3 aplicado al mejor modelo en términos de performance predictiva resultante de la etapa anterior.

Se destaca que todas las etapas anteriores se aplicaron considerando las dos técnicas de imputación mencionadas en la Sección 4.4.

7.1. Modelo de Regresión Lineal Múltiple

En esta sección se presentan los principales resultados obtenidos luego de realizar un ajuste mediante un *Modelo de Regresión Lineal Múltiple*.

En lo que respecta a las estimaciones de los coeficientes del modelo, en la Sección del Anexo A.5 se expone en las Tablas A.6 y A.8 el resumen del ajuste para cada método de imputación de valores faltantes utilizado, con los coeficientes estimados, el error estándar de cada uno de ellos, el estadístico t de student y su p-valor asociado.

De esta forma, en función a lo explicitado en las Tablas A.6 y A.8, se tiene que todas las variables son significativas considerando un nivel de significación del 5 %. Sin embargo, se destaca que dicho resultado puede estar influenciado por la cantidad de observaciones utilizadas para ajustar los modelos.

A su vez se tiene que dado las demás variables constantes, el precio del apartamento disminuye a medida que aumenta la distancia a la rambla este. Esto en la medida de que la estimación del parámetro asociado a dicha variable se encuentra precedido por un signo negativo.

Por otro lado, se tiene que aquellos apartamentos con dos o más baños completos, dado las demás variables constantes, se caracterizan por presentar un precio mayor a aquellos con un solo baño completo. Esto debido a que la estimación del parámetro asociado al nivel de referencia (2 o más) de la variable cantidad de baños completos, se encuentra precedido de un signo positivo.

Por otra parte, en la Tabla 7.1 se presentan los resultados globales del *Modelo de Regresión Lineal Múltiple*.

Se destaca que el ajuste se realizó utilizando la función *lm* del paquete *stats* ([32]).

Imputación	**RMSE**	**MAE**	R^2	R_a^2	**Estadístico F**	**P-Valor**
Media	43,170	30,557	0.67	0.67	4,640	<2.2e-16
Random Forest	42,713	30,291	0.68	0.68	4,639	<2.2e-16

Tabla 7.1: Principales resultados de los *Modelos de Regresión Lineal Múltiple* ajustados según el método de imputación utilizado, considerando muestra de entrenamiento y muestra de testeo.

Según se observa en la Tabla 7.1 los modelos lineales ajustados tienen un error de predicción de aproximadamente 43,000 y 30,000 dólares si se utiliza como medida de error a la raíz cuadrada del error cuadrático medio y error absoluto medio respectivamente. Asimismo, los modelos explican aproximadamente un 70 % de la varianza en los datos y son globalmente significativos.

Por último, en lo que respecta al análisis de los residuos, en primer lugar se realiza un análisis gráfico de los mismos. En la Figura 7.1 se presentan los errores en función de los valores ajustados y la distribución de los mismos considerando ambos métodos de imputación.

El gráfico de dispersión de la Figura 7.1 sugiere heterocedasticidad en los residuos de los modelos ajustados. Por otra parte, el histograma de los residuos muestra que los mismos se encuentran en torno al valor 0.

Una vez realizado el análisis gráfico se procedió a realizar las pruebas sobre los supuestos de los residuos del modelo mencionadas en la Sección 4.1.1. Considerando ambos métodos de imputación y un nivel de significación del 0.05, se procedió a rechazar tanto el supuesto de normalidad como el supuesto de homocedasticidad de los residuos en la medida que se rechazan las hipótesis nulas de ambas pruebas. Esto último teniendo en cuenta que en ambos casos el p-valor asociado a la prueba es inferior al nivel de significación considerado.

En las Tablas A.9 y A.10 de la Sección del Anexo A.5 se encuentran los resultados de las pruebas aplicadas.

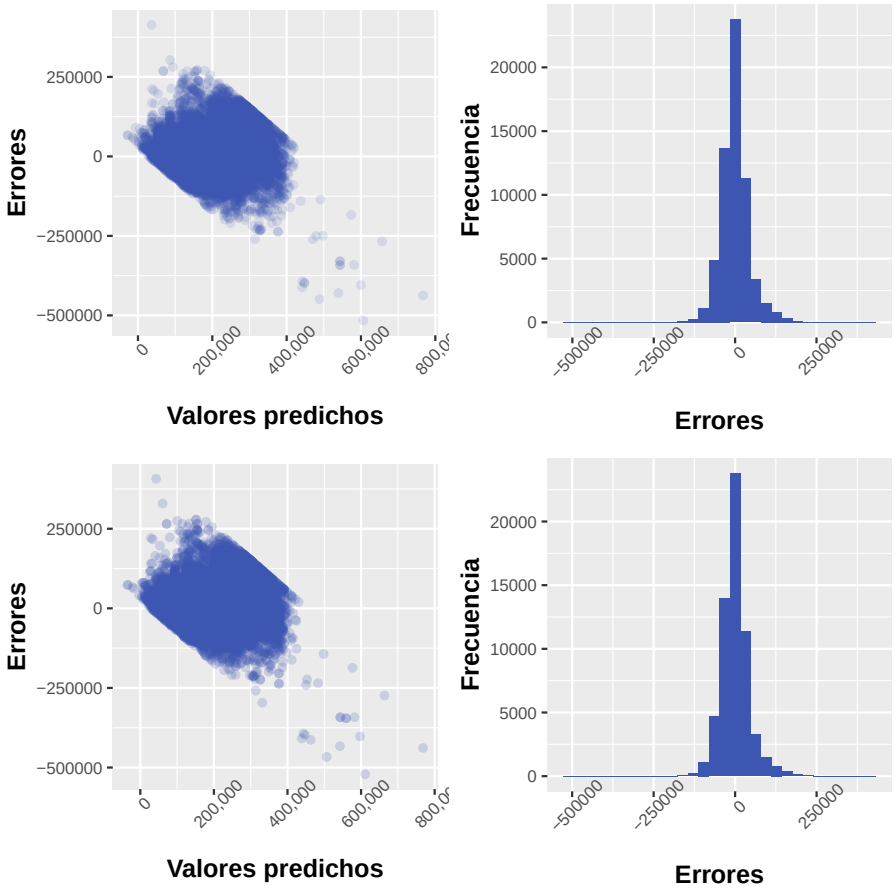

Figura 7.1: Gráficos de los residuos de los *Modelos de Regresión Lineal Múltiple* ajustados. En el panel superior se encuentran los resultados para el caso de imputación de valores faltantes por la media, mientras que en el panel inferior se encuentran los resultados para el caso de imputación de valores faltantes por *Random Forest*. A la izquierda se encuentra el gráfico de dispersión entre los errores del modelo y los valores predichos. A la derecha se encuentra el histograma de los mismos.

7.2. Árbol de regresión

A continuación se presentan los resultados del ajuste mediante un *Árbol de regresión* considerando ambas metodologías de imputación de valores faltantes.

En primer lugar se presenta, para cada modelo ajustado, el gráfico luego de realizar el proceso de poda mencionado en la Sección 4.1.2. Respecto a este punto, en la Sección del Anexo A.6 se presenta en las Tablas A.11 y A.12 los diferentes valores

del parámetro de costo-complejidad, su error mediante un proceso de *validación cruzada* asociado, al igual que el tamaño del árbol que dicho valor del parámetro implica y otras métricas de interés.

Se destaca que el ajuste se realizó utilizando la función *rpart* del paquete *rpart* ([34]).

Según se observa en la Figura 7.2 el árbol se conforma por 10 nodos terminales (hojas). Los porcentajes dentro de cada nodo indican el porcentaje del número de observaciones que se encuentran en el mismo. Además, se explicita la predicción de las observaciones pertenecientes al nodo. A su vez, se destaca que se realizan 9 particiones.

Asimismo, se tiene que la primera variable en realizar una partición binaria es la variable cantidad de baños completos. Las variables distancia a la rambla este, área total, cantidad de dormitorios e ingreso medio ECH son utilizadas en las siguientes particiones.

De esta forma, se tiene que el precio de oferta de un apartamento con más de un baño completo, que se encuentra a menos de 2,039 metros de distancia a la rambla este y con un área total mayor o igual a 107 metros cuadrados, es predicho por el modelo como 351,524 dólares.

Por su contraparte, un apartamento con un solo baño completo, con un área total menor a 55.2 metros cuadrados y que se encuentra a una distancia mayor o igual de 1,462 metros a la rambla este, el modelo realiza una predicción sobre el precio de oferta del mismo de 107,403 dólares.

Por otra parte, el gráfico correspondiente al árbol realizado mediante imputación de valores faltantes por *Random Forest* es presentado en la Figura A.2 de la Sección del Anexo A.6 debido a que no presenta diferencias sustanciales con el gráfico del árbol obtenido realizando imputación por la media.

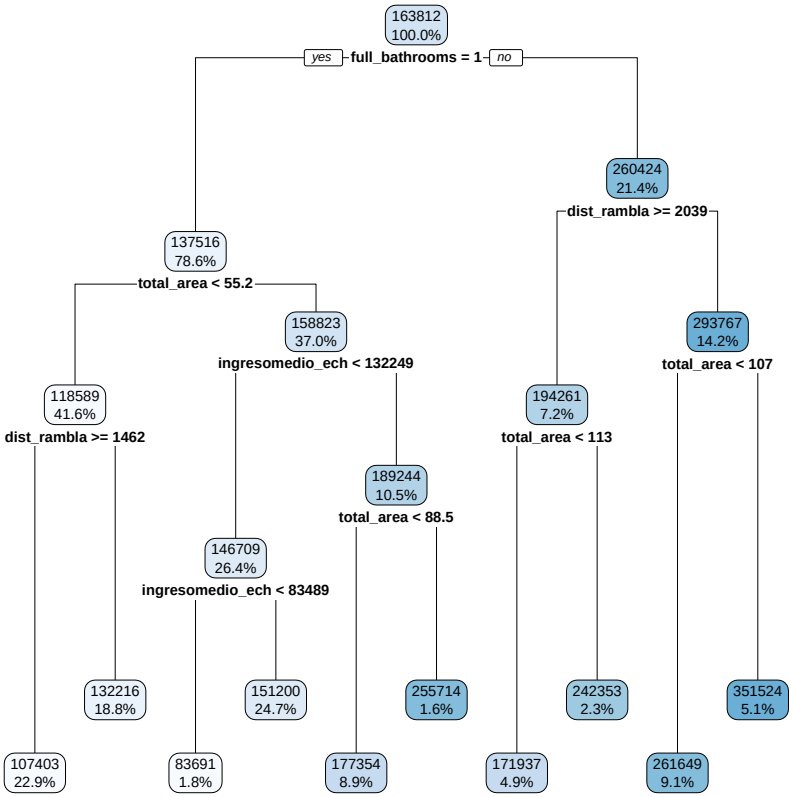

Figura 7.2: *Árbol de regresión* obtenido al ajustar la variable precio de oferta en dólares una vez realizado el proceso de poda con un valor de CP igual a 0.01. El método de imputación de valores faltantes utilizado en este caso es imputación por la media.

Una vez observada las principales características de los modelos, se procedió a obtener cierta medida de la performance predictiva en función de las métricas definidas en la Sección 4.2. Las mismas se presentan a continuación en la Tabla 7.2.

Imputación	**RMSE**	R^2	**MAE**
Media	43,204	0.67	31,053
Random Forest	43,175	0.67	31,036

Tabla 7.2: Principales medidas de resumen de los *Árboles de Regresión* ajustados según metodología de imputación utilizada, considerando muestra de entrenamiento y muestra de testeo. Se presentan la raíz cuadrada del error cuadrático medio (RMSE), el coeficiente de determinación (R^2), y el error absoluto medio (MAE).

Como se observa en la Tabla 7.2, ambos modelos tienen un error promedio de predicción de aproximadamente 43,000 y 30,000 dólares si se utiliza como medida de error a la raíz cuadrada del error cuadrático medio y el error absoluto medio respectivamente. Asimismo, los modelos explican aproximadamente un 70 % de la varianza en los datos.

7.3. Random Forest

En esta sección se presentan los resultados del ajuste por *Random Forest* considerando ambas metodologías de imputación.

Se destaca que el ajuste se realizó utilizando la función *ranger* del paquete *ranger* ([36]) y tomando los valores por defecto de los *parámetros de ajuste* detallados en la Sección 4.1.3 para ambos métodos de imputación.

De esta manera, en primer lugar se presenta en la Tabla 7.3 los valores utilizados.

Parámetro de ajuste	**Valor**
Número de árboles	500
Cantidad de variables	4
Min. obs.	5
Regla de partición	Min. SCR

Tabla 7.3: Valores por defecto de los *parámetros de ajuste* utilizados en los modelos ajustados por *Random Forest*. Se detalla el número de árboles, la cantidad de variables seleccionadas de manera aleatoria en cada partición, la cantidad de observaciones mínimas en un nodo terminal y la regla de partición.

Luego se presenta graficamente en la Figura 7.3 un análisis de la *importancia permutada* de las variables siguiendo la metodología detallada en la Sección 4.3.1.

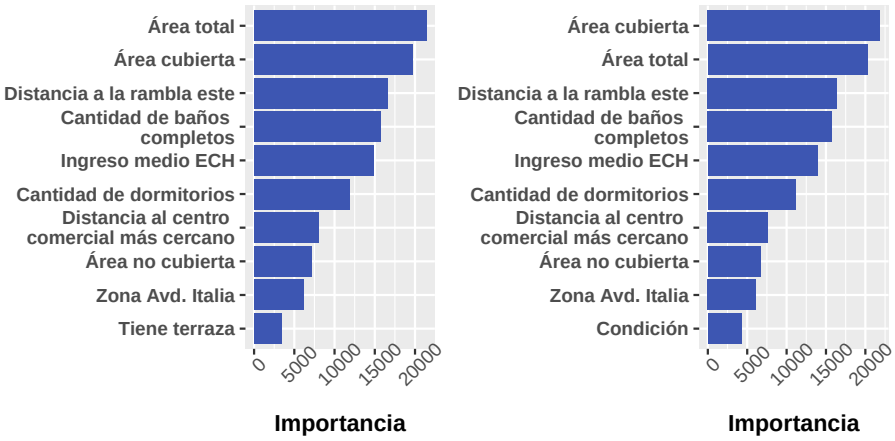

Figura 7.3: Gráfico de la *importancia permutada* de las 10 primeras variables en los modelos ajustados por *Random Forest*. Se observa que las variables área cubierta, área total y distancia a la rambla este son las 3 variables más importantes en ambos modelos.

Según se observa en la Figura 7.3 todas las variables geoespaciales construidas se encuentran entre las 10 variables más importantes para ambos casos. Por otro lado se destaca que, a la hora de realizar imputación de valores faltantes por *Random*

Forest, la variable condición se encuentra entre las 10 variables más importantes.

Luego se procedió a obtener cierta medida de la performance predictiva en función de las métricas definidas en la Sección 4.2. Las mismas se presentan a continuación en la Tabla 7.4.

Imputación	**RMSE**	R^2	**MAE**	**RMSE OOB**	R^2 **OOB**
Media	25,695	0.82	16,435	26,461	0.88
Random Forest	25,802	0.82	16,547	26,476	0.88

Tabla 7.4: Principales medidas de resumen de los modelos ajustados por *Random Forest* según metodología de imputación de valores faltantes utilizada, considerando muestra de entrenamiento y muestra de testeo. Se presentan la raíz cuadrada del error cuadrático medio (RMSE), el coeficiente de determinación (R^2), y el error absoluto medio (MAE), al igual que el error *out of bag* (RMSE OOB) y el coeficiente de determinación *out of bag* (R^2 OOB).

Según se observa en la tabla 7.4 ambos modelos tienen un error promedio de predicción de aproximadamente 26,000 y 17,000 dólares si se utiliza como medida de error a la raíz cuadrada del error cuadrático medio y al error absoluto medio respectivamente. Asimismo, los modelos explican aproximadamente un 82 % de la varianza en los datos. Por su parte, si se consideran las métricas de RMSE y R^2 *out of bag*, se tiene un error medio de predicción de aproximadamente 27,000 dólares y una varianza explicada aproximadamente del 90 %.

7.4. Boosting

En esta sección se presentan los resultados del ajuste por *Boosting* considerando ambas metodologías de imputación de valores faltantes.

Se destaca que el ajuste se realizó utilizando la función *gbm* del paquete *gbm* ([11]) y tomando los valores por defecto de los *parámetros de ajuste* detallados en la

Sección 4.1.4 para ambos métodos de imputación. De esta manera, en primer lugar se presenta en la Tabla 7.5 los valores utilizados.

Parámetro de ajuste	**Valor**
Número de árboles	100
Tasa de aprendizaje	0.1
Número de particiones en cada árbol	1

Tabla 7.5: Valores por defecto utilizados de los *parámetros de ajuste* en los modelos ajustados por *Boosting*. Se detalla el número de árboles, la tasa de aprendizaje y el número de particiones en cada árbol.

Luego, se presenta graficamente en la Figura 7.4 un análisis de la *importancia permutada* de las variables siguiendo la metodología detallada en la Sección 4.1.4.

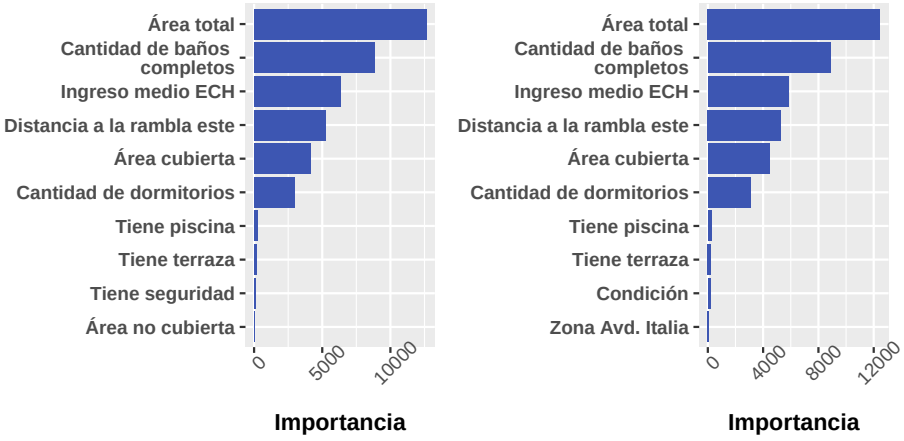

Figura 7.4: Gráfico de la *importancia permutada* de las 10 primeras variables en los modelos ajustados por *Boosting*. Se observa que las variables área total, cantidad de baños completos e ingreso medio ECH son las 3 variables más importantes en ambos modelos.

En lo que respecta a las variables geoespaciales construidas, se observa en la Figura 7.4 que la variable distancia a la rambla este se encuentra entre las 10 variables más importantes para ambos casos. Por otro lado se destaca que a la hora de realizar

la imputación por *Random Forest*, la variable condición se encuentra entre las 10 variables más importantes.

Luego, se procedió a obtener cierta medida de la performance predictiva en función de las métricas definidas en la Sección 4.2. Las mismas se presentan a continuación en la Tabla 7.6.

Imputación	**RMSE**	R^2	**MAE**
Media	39,793	0.62	27,928
Random Forest	39,800	0.62	28,031

Tabla 7.6: Principales medidas de resumen de los modelos ajustados por *Boosting* según metodología de imputación utilizada, considerando muestra de entrenamiento y muestra de testeo. Se presentan la raíz cuadrada del error cuadrático medio (RMSE), el coeficiente de determinación (R^2), y el error absoluto medio (MAE).

Según se observa en la Tabla 7.6 ambos modelos tienen un error promedio de predicción de aproximadamente 40,000 y 28,000 dólares si se utiliza como medida de error a la raíz cuadrada del error cuadrático medio y el error absoluto medio respectivamente. Asimismo, los modelos explican aproximadamente un 60 % de la varianza en los datos.

7.5. Support vector regression

En esta sección se presentan los resultados del ajuste por *Support Vector Regression* considerando ambas metodologías de imputación de valores faltantes.

Se destaca que el ajuste se realizó utilizando la función *ksvm* del paquete *kernlab* ([19]) y tomando los valores por defecto de los *parámetros de ajuste* detallados en la Sección 4.1.5 para ambos métodos de imputación.

De esta manera, en primer lugar se presenta en la Tabla 7.7 los valores de los *parámetros de ajuste* utilizados.

Parámetro de ajuste	Valor
Parámetro de complejidad	1
Umbral	0.1
Parámetro de escala	0.03085

Tabla 7.7: Valores por defecto utilizados de los *parámetros de ajuste* en los modelos ajustados por *Support Vector Regression*. Se detalla el parámetro de complejidad, el umbral y el parámetro de escala.

Por otro lado en la Figura 7.5 se presenta gráficamente un análisis de la *importancia permutada* de las variables siguiendo la metodología detallada en la Sección 4.3.1.

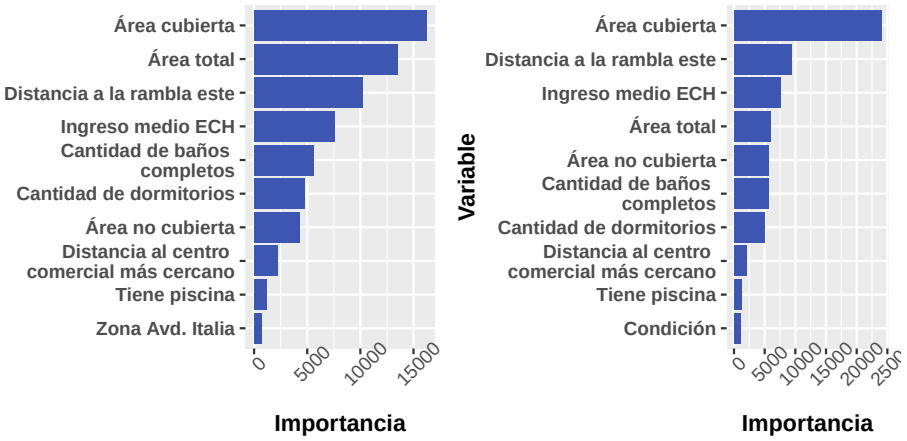

Figura 7.5: Gráfico de la *importancia permutada* de las 10 primeras variables en los modelos ajustados por *Support Vector Regression*. Se observa que las variables área cubierta, área total y distancia a la rambla este son las 3 variables más importantes en el ajuste utilizando método de imputación de valores faltantes por la media. En lo que respecta al ajuste con imputación de valores faltantes por *Random Forest*, las variables área cubierta, distancia a la rambla este e ingreso medio ECH son las 3 variables más importantes.

A su vez, en lo que respecta a las variables geoespaciales construidas, se observa

en la Figura 7.5 que la variable distancia al centro comercial más cercano e ingreso medio ECH se encuentran entre las 10 variables más importantes para los ajustes según ambos métodos de imputación de valores faltantes utilizados. Por otro lado, se destaca que a la hora de realizar la imputación de valores faltantes por *Random Forest*, la variable condición se encuentra entre las 10 variables más importantes.

Por último, se procedió a obtener cierta medida de la performance predictiva en función de las métricas definidas en la Sección 4.2. Las mismas se presentan a continuación en la Tabla 7.8.

Imputación	**RMSE**	R^2	**MAE**
Media	33,552	0.78	22,825
Random Forest	33,296	0.79	22,600

Tabla 7.8: Principales medidas de resumen de los modelos ajustados por *Support Vector Regression* según metodología de imputación utilizada, considerando muestra de entrenamiento y muestra de testeo. Se presentan la raíz cuadrada del error cuadrático medio (RMSE), el coeficiente de determinación (R^2), y el error absoluto medio (MAE).

La tabla 7.8 muestra que ambos modelos tienen un error promedio de predicción de aproximadamente 33,000 y 23,000 dólares si se utiliza como medida de error a la raíz cuadrada del error cuadrático medio y a el error absoluto medio respectivamente. Asimismo, los modelos explican aproximadamente un 80 % de la varianza en los datos.

7.6. Parámetros de ajuste

Una vez realizados los ajustes de los modelos detallados en el Capítulo 4 utilizando los valores por defecto, con el fin de mejorar la performance predictiva se procedió a realizar una búsqueda orientada sobre los posibles valores para los *parámetros de*

ajuste de cada modelo.

Para ello, para todos los ajustes se trabajó con la función *train* del paquete *CARET* ([20]), con excepción del ajuste realizado por *Boosting*. Este último se realizó mediante una función construída que permite realizar el proceso de selección de los *parámetros de ajuste*, a partir de la función *gbm* del paquete *gbm* y emulando la función *train*. Se realizó esta excepción en la medida que a la fecha del informe el paquete *CARET* presenta ciertos inconvenientes (error sobreestimado) a la hora de realizar un ajuste por el algoritmo *Boosting*.

Por otro lado, se destaca que la función *train* tiene especificada por defecto una grilla de valores para los *parámetros de ajuste* de cada modelo. De esta manera, para cada combinación de *parámetros de ajuste*, se realiza un ajuste y luego se evalúan en función de las métricas detalladas en la Sección 4.2. En la Sección del Anexo A.7 se detallan las grillas mencionadas en las Tablas A.13, A.14, y A.15.

Mediante los resultados obtenidos en estos ajustes iniciales, se realizó el proceso de búsqueda de combinaciones que puedan mejorar la performance predictiva en cada modelo.

Se destaca que algunos de los *parámetros de ajuste* detallados en el Capítulo 4 permanecen constantes a la hora de realizar los diferentes ajustes utilizando las grillas por defecto. Los mismos se detallan en la Tabla 7.9.

Modelo	**Hyperparámetro**	**Valor por defecto**
Ranfom Forest	Cantidad de árboles	500
Ranfom Forest	Min. Obs	5
Boosting	Tasa de aprendizaje	0.1
Boosting	Min. Obs	10
SVR	Parámetro de escala	0.03085
SVR	Umbral	0.1

Tabla 7.9: Tabla resumen de los valores de los *parámetros de ajuste* que utiliza por defecto la función *train* del paquete *CARET* y que se mantienen constantes para todos los ajustes.

En particular en la Tabla 7.9 se detalla la cantidad de árboles y cantidad de observaciones mínimas en cada nodo terminal para los ajustes por *Random Forest*. Asimismo, se detalla la tasa de aprendizaje y la cantidad mínima de observaciones en cada nodo terminal para los ajustes por *Boosting*. Por último, se especifican los valores para el parámetro de escala y el umbral tolerado para los ajustes por *SVR*.

De esta forma, en la Figura 7.6 se observa la evolución del error de predicción en función de los valores de los *parámetros de ajuste* utilizados para cada modelo.

En lo que respecta a los ajustes realizados por *Random Forest*, en la Figura 7.6 se observa el error de predicción en función de la cantidad de variables seleccionadas de manera aleatoria en cada partición, definidas en la grilla por defecto (2, 12, y 23). Se observa un comportamiento decreciente entre los valores 3 y 12 y un comportamiento creciente de menor magnitud, entre los valores 12 y 23. De esta forma, a la hora de realizar el proceso de selección de los *parámetros de ajuste* se consideraron todos los valores enteros entre los números 12 y 20.

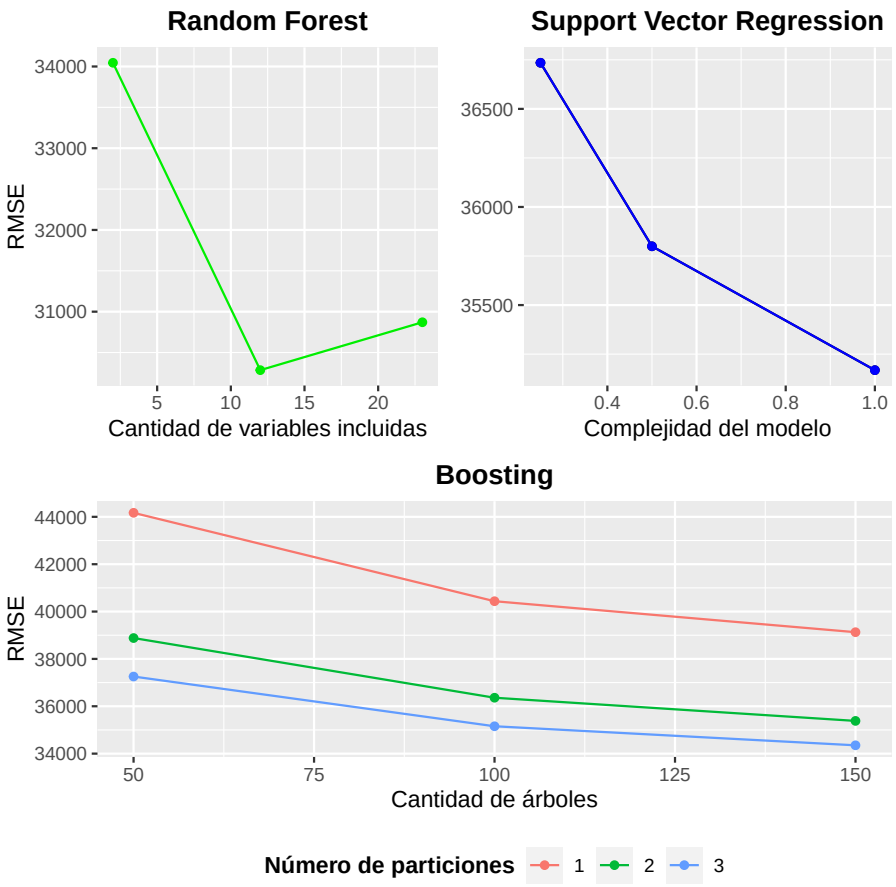

Figura 7.6: Gráfico de la evolución del error cuadrático medio de predicción (RMSE) en función de los valores de los *parámetros de ajuste* por defecto para el caso de imputación de valores faltantes por la media. En el panel suerior izquierdo se encuentran los resultados para los ajustes por *Random Forest*, en el panel superior derecho se encuentran los resultados para el ajuste por *Support Vector Regression*, y en el panel inferior izquierdo se encuentran los resultados para el ajuste por *Boosting*.

Por otro lado, en cuanto a los ajustes realizados por *Boosting*, se observa que el error de predicción es decreciente conforme crece el número de árboles. Asimismo, se observa que el mismo es decreciente en cuanto al número de particiones. De esta manera, en la medida que los comportamientos observados para dichos *parámetros de ajuste* son decrecientes conforme se incrementa el valor de los mismos, se tomaron valores del número de particiones entre 3 y 10; y cantidad de árboles entre 500 y 5000. Asimismo, se decidió variar la tasa de aprendizaje tomando valor 0.1 (valor

fijo por defecto) y 0.01.

Por último, sobre los ajustes realizados por *Support Vector Regression*, se observa que el error de predicción es decreciente a medida que aumenta el parámetro de complejidad. Por lo tanto, se trabajó con valores enteros entre 1 y 5. A su vez, se utilizaron valores de 0.05 y 0.01 del parámetro de escala.

Se destaca que en la medida que los resultados obtenidos son similares para el caso de imputación de valores faltantes por *Random Forest*, la grilla seleccionada para realizar el proceso de selección de los *parámetros de ajuste* es análoga para cada modelo. En la Figura A.3 de la Sección del Anexo A.7 se encuentran los gráficos correspondientes.

Un vez definida la grilla de *parámetros de ajuste* para cada modelo se procedió a implementar el proceso de selección de los *parámetros de ajuste*. En la Sección del Anexo A.7 se presenta en las Tablas A.17, A.7 y A.19 los resultados de cada ajuste para cada modelo según la combinación de *parámetros de ajuste* utilizada.

De esta forma, en función de los resultados obtenidos se procedió a seleccionar aquel ajuste con mejor performance predictiva para cada modelo y considerando ambos métodos de imputación. A modo de resumen, se presenta en la Tabla 7.10 los resultados obtenidos. Se destaca que en la misma se incluyen los resultados obtenidos sobre las métricas definidas en la Sección 4.2 para el caso de un ajuste mediante un *Modelo de Regresión Lineal Múltiple*.

Algoritmo	**Método de imputación**	**RMSE**	R^2	**MAE**
Boosting	Random Forest	26,624	0.89	17,669
Boosting	Media	26,758	0.88	17,752
Random Forest	Random Forest	30,118	0.85	19,563
Random Forest	Media	30,295	0.84	19,658
SVR	Random Forest	34,183	0.80	23,081
SVR	Media	34,764	0.80	23,494
Modelo Lineal	Random Forest	43,686	0.68	30,814
Modelo Lineal	Media	44,195	0.67	31,148

Tabla 7.10: Tabla con las principales medidas de resumen para los modelos con mejor poder predictivo de cada algortimo utilizado luego de haber realizado el proceso de selección de los *parámetros de ajuste*.

En función de lo observado en la tabla 7.10, para ambos métodos de imputación de valores faltantes, el modelo con mejor performance predicitiva con respecto a las tres métricas, es resultado de realizar un ajuste mediante el algoritmo *Boosting*, seguido por *Random Forest*. Por su contraparte, el modelo con menor desempeño en términos de performance predictiva con respecto a las tres métricas, es resultado de realizar un ajuste mediante un *Modelo de Regresión Lineal Múltiple*.

Se destaca que en todos los casos los ajustes considerando metodología de imputación de valores faltantes por *Random Forest* tienen un poder predictivo superior. No obstante, como fue mencionado en la Sección 4.4 estos ajustes incluyen adicionalmente a la variable condición.

De esta forma, con el fin de seleccionar al mejor modelo en términos de performance predictiva y menor complejidad, en lo que respecta a la cantidad de variables utilizadas en el ajuste y el método de imputación de valores faltantes utilizado, se optó por el ajuste que es resultado de aplicar el algoritmo *Boosting* con método de imputación de valores faltantes por la media.

Dicho ajuste se realizó utilizando los siguientes valores para los *parámetros de*

ajuste del mismo: 1) 5000 árboles, 2) 10 particiones, 3) tasa de aprendizaje 0.1 y 4) 10 observaciones como cantidad mínima en un nodo terminal.

A continuación, en la Tabla 7.11 a modo de resumen se presentan las métricas de perfomance predictiva definidas en la Sección 4.2, incluyendo el Error Porcentual Absoluto Medio (EPAM).

Algoritmo	**Método de imputación**	**RMSE**	R^2	**MAE**	**EPAM**
Boosting	Media	26,758	0.88	17,752	0.12

Tabla 7.11: Tabla con las principales medidas de resumen para el ajuste por *Boosting* con mejor poder predictivo y considerando método de imputación de valores faltantes por la media, luego de haber realizado el proceso de selección de los *parámetros de ajuste*.

7.7. Interpretabilidad

Una vez implementado el proceso de selección de los *parámetros de ajuste* y obtenido el modelo con mejor desempeño en términos de performance predictiva (ajuste por *Boosting*) se procedió a realizar el análisis de interpretabilidad detallado en la Sección 4.3.

En primer lugar se realizó un análisis sobre la importancia de cada predictor según se especifica en la Sección 4.3.1. De esta forma se presenta en la Figura 7.7 los resultados obtenidos utilizando la función *vip* del paquete *vip* ([12]).

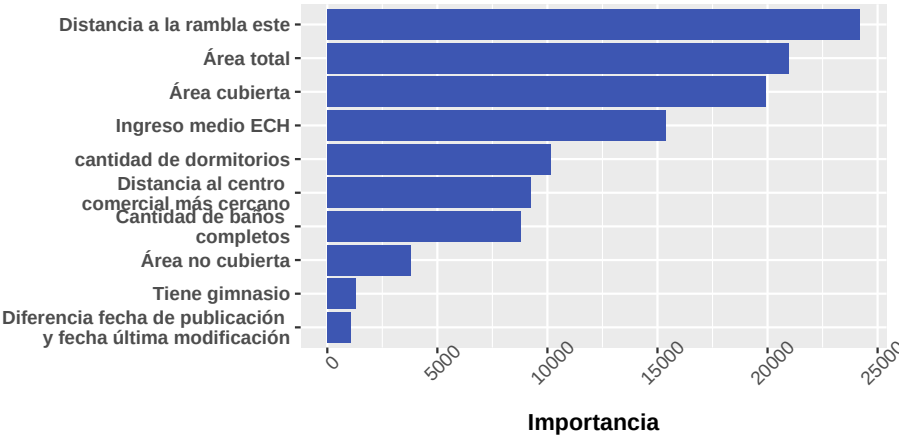

Figura 7.7: Gráfico de la *importancia permutada* de las 10 primeras variables en el modelo con la mejor performance predictiva según el error de predicción (ajuste por *Boosting*). Se observa que las variables distancia a la rambla este, área total y área cubierta son las 3 variables más importantes para dicho modelo, considerando ambos métodos de imputación.

En lo que respecta a las variables geoespaciales construidas, se observa en la Figura 7.7 que las mismas, con la salvedad de zona Avd. Italia, se encuentran entre las 10 variables más importantes. A su vez, se destaca que todas las variables de naturaleza contínua se encuentran entre las 10 variables más importantes.

Sin embargo, según se observa en la Figura 6.7 las variables área total, área cubierta, y área no cubierta presentan una alta correlación lineal. Esto último implicando que los resultados obtenidos en cuanto a la importancia de los mismos se ven afectados según se explicita en la Sección 4.3.1.

Posteriormente, según lo observado en la Figura 7.7 se procedió a obtener una medida del efecto marginal de algunas de las variables más relevantes sobre la predicción de la variable precio de oferta en dólares, según se detalla en la Sección 4.3.2. De esta manera, se presenta en la Figura 7.8 los resultados obtenidos. Se destaca que los resultados fueron obtenidos utilizando el paquete *iml* ([27]).

En primer lugar, se observa en la Figura 7.8 un comportamiento decreciente en la predicción del precio de oferta en dólares a medida que se incrementan los valores

de distancia a la rambla este. Este resultado se encuentra en concordancia con la correlación lineal negativa entre ambas variables observada en la Figura 6.7. Asimismo, el resultado se encuentra en línea con el signo de la estimación del coeficiente asociado a esta variable obtenido mediante el ajuste del *Modelo de Regresión Lineal Múltiple*, como se observa en las Tablas A.6 y A.8 que se encuentran en la Sección del Anexo A.5.

Por otra parte, en lo que respecta a la variable área total se observa en la Figura 7.8 un comportamiento creciente en la predicción del precio de oferta en dólares para los apartamentos con un área total inferor a 210 metros cuadrados. Luego, se tiene un comportamiento decreciente para los apartamentos con un área total de entre 210 metros cuadrados y 250 metros cuadrados. A partir de allí, se observa un comportamiento aproximadamente estable en la predicción. Sin embargo, debe considerarse que estos resultados están sujetos a la cantidad de observaciones en el rango de valores entre cada punto de corte definido en la Sección 4.3.2.

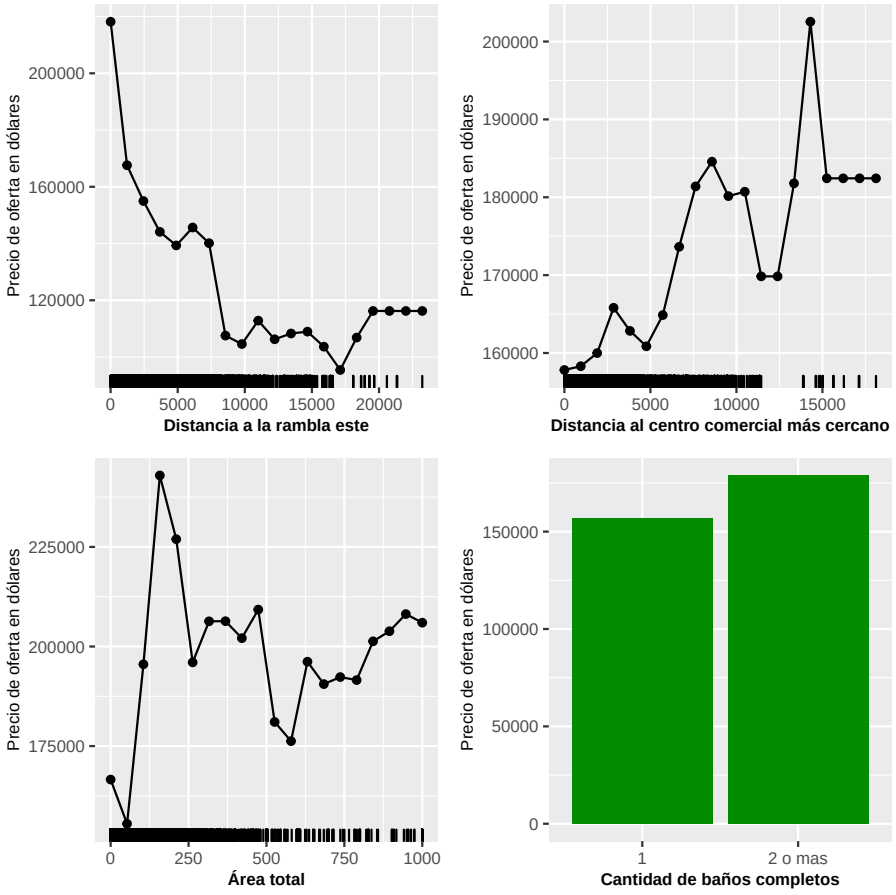

Figura 7.8: *Gráfico de dependencia parcial* para las variables distancia a la rambla este (panel superior izquierdo), distancia al centro comercial más cercano (panel superior derecho), área total (panel inferior izquierdo) y cantidad de baños completos (panel inferior derecho). Para las variables distancia a la rambla este, distancia al centro comercial más cercano y área total se incluye en el eje de las abscisas la distribución marginal de cada una de ellas.

Por su parte, para la variable distancia al centro comercial más cercano se observa en la Figura 7.8 un comportamiento creciente de la predicción del precio de oferta en dólares para los apartamentos con una distancia al centro comercial más cercano inferior a aproximadamente 8,600 metros. Para los apartamentos con una distancia al centro comercial más cercano superior a dicho valor, se observa un comportamiento irregular en la predicción del precio de oferta en dólares. Esto último puede deberse al reducido número de observaciones que se tiene en los datos para valores superiores

a 10,000 metros.

Por último, en la Figura 7.8 se observa que para aquellos apartamentos con dos o más baños completos el modelo realiza en promedio una predicción del precio de oferta en dólares de aproximadamente 22,000 dólares superior respecto a aquellos apartamentos con un solo baño completo.

Una vez realizados los análisis del efecto marginal para una sola variable, se procedió a obtener los efectos marginales para la interacción de dos variables. Para ello, en función de lo observado en las figuras 7.7 y 7.8 se realizó el analisis de intepretabilidad para las variables distancia a la rambla este y cantidad de baños completos, y área total e ingreso medio ECH. En las figuras 7.9 y 7.10 respectivamente se presentan los resultados obtenidos.

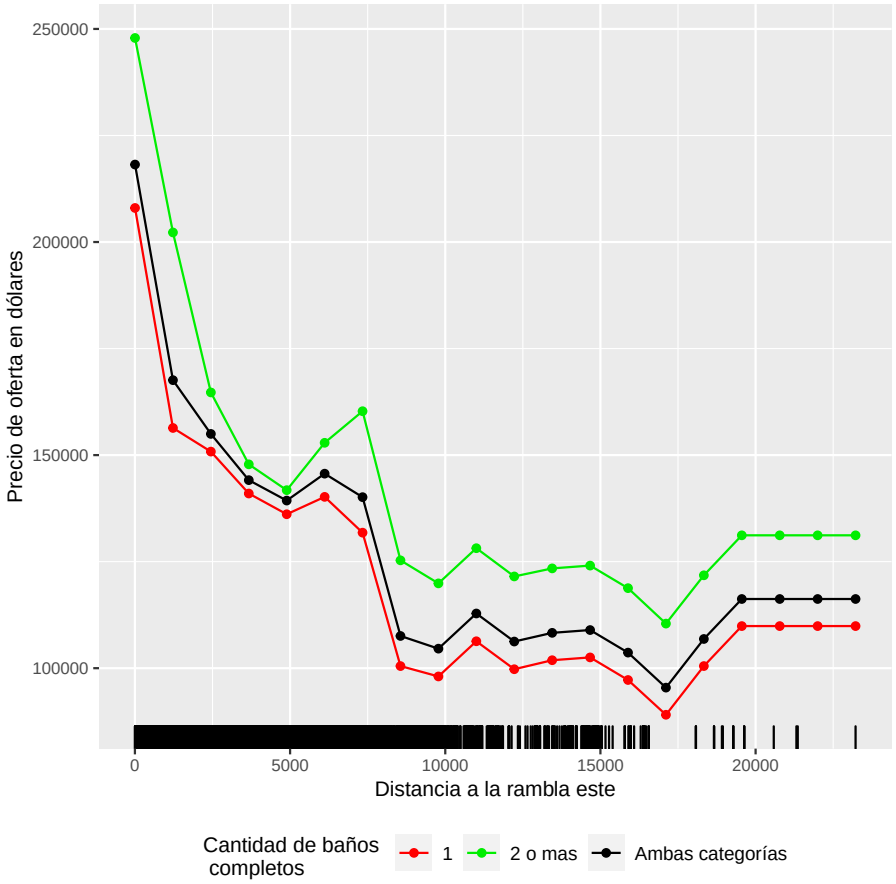

Figura 7.9: *Gráfico de dependencia parcial* para las variables distancia a la rambla este y cantidad de baños completos.

Según se observa en la Figura 7.9 para cada valor de distancia a la rambla este los apartamentos con dos o más baños completos tienen una predicción promedio del precio de oferta en dólares superior a los apartamentos con un solo baño completo. Donde esta diferencia se acentúa para los apartamentos ubicados a más de 5,500 metros de la rambla Este.

A su vez, a grandes razgos se observa en la Figura 7.9 que la predicción promedio del precio de oferta en dólares en función de la distancia a la rambla este sigue un mismo comportamiento tanto para los apartamentos con un solo baño completo como para los apartamentos con dos o más baños completos.

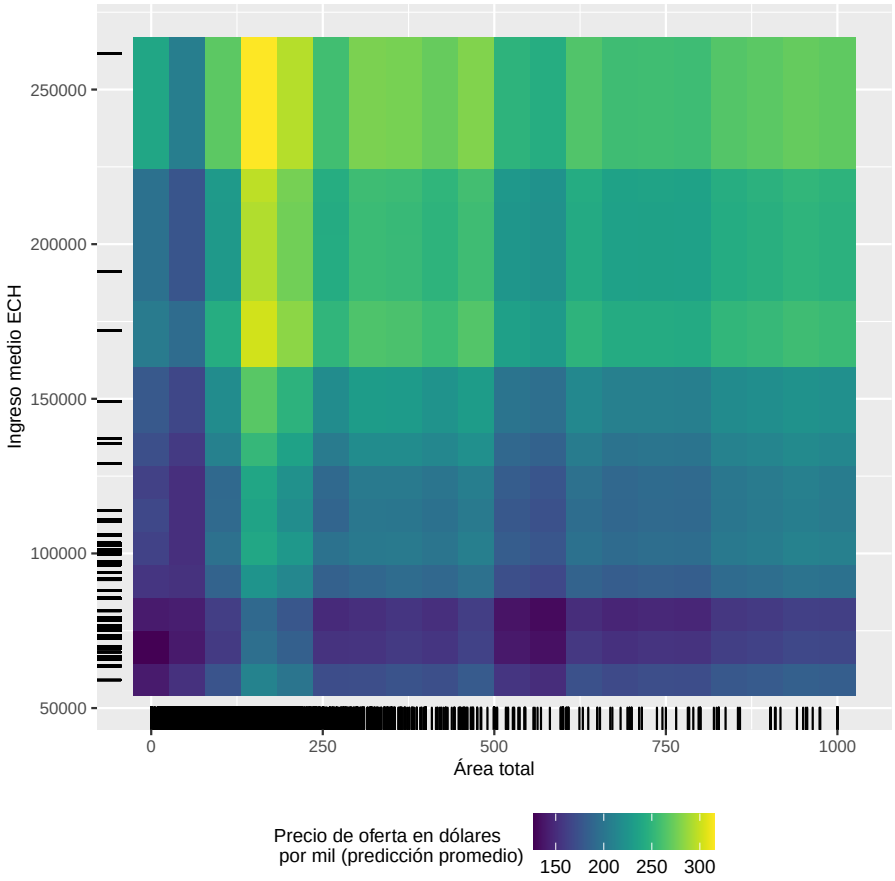

Figura 7.10: *Gráfico de dependencia parcial* para las variables área total e ingreso medio ECH. Para ambas variables se incluye la distribución marginal en sus respectivos ejes.

En primer lugar, como se observa en la Figura 7.10 para los apartamentos ubi-

cados en barrios con ingresos medios entre 50,000 y 70,000 pesos uruguayos, la predicción promedio del precio de oferta en dólares se mantiene en torno a los 150,000 dólares. Esto último para todos los valores de la variable área total, con excepción de los apartamentos de entre 100 y 250 metros cuadrados, donde se observa una predicción promedio superior.

Por otra parte, para los apartamentos con un área total de hasta aproximadamente 150 metros cuadrados, se observa a grandes razgos que la predicción del precio de oferta en dólares se incrementa a medida que aumenta el ingreso medio del barrio donde está ubicado el apartamento.

Por último, se observa que los valores de la predicción promedio para los apartamentos con un área entre 150 y 250 metros cuadrados y ubicados en barrios con ingreso medio superior a 150,000 pesos uruguayos son superiores a 200,000 dólares aproximadamente.

Capítulo 8

Comentarios finales

En este trabajo se presentaron distintos métodos de aprendizaje supervisado para predecir el precio de oferta de los apartamentos en Montevideo, Uruguay. A su vez, una parte muy importante del trabajo consistió en la generación y la transformación de datos para su posterior uso con énfasis en la reproducibilidad de los resultados.

En cuanto a la capacidad predictiva de los distintos modelos, se observó la incapacidad del *Modelo de Regresión Lineal Múltiple* a la hora de captar relaciones no lineales entre las variables predictoras y la variable de respuesta. Siendo este último el ajuste con menor capacidad predictiva.

Por otra parte, el ajuste con mayor performance predictiva fue mediante el algoritmo *Boosting* una vez realizado el proceso de selección de los *parámetros de ajuste*. En el cual se observó un error absoluto medio de aproximadamente 18,000 dólares y un error porcentual absoluto medio de aproximadamente un 12 %.

Sin embargo, se destaca que el ajuste previo a la realización de dicho proceso tuvo una performance predictiva inferior a la del resto de los modelos de aprendizaje estadístico utilizados, y similar a la obtenida a partir del ajuste por el *Modelo de Regresión Lineal Múltiple*. Por lo tanto, se recalca la relevancia de la realización del proceso de selección de los *parámetros de ajuste* para los modelos de aprendizaje estadístico.

Por otro lado, en cuanto a los métodos de imputación de valores faltantes imple-

mentados, no se observaron diferencias sustanciales en los resultados obtenidos. De esta forma, se consideró apropiado tomar como mejor modelización aquella obtenida mediante la realización de imputación de valores faltantes por la media, principalmente por su simplicidad de cálculo.

En lo que respecta a la importancia de las variables en el ajuste con mayor performance predictiva (ajuste mediante *Boosting*), se observó que todas las variables construidas mediante fuentes externas resultaron entre las 10 variables más importantes con excepción de la variable zona Avd. Italia. En particular, se destaca que la variable distancia a la rambla este se consideró la más importante en función de la metodología implementada (*importancia de las variables permutadas*).

Sobre éste punto, se sugiere para futuros trabajos replicar el análisis construyendo otras variables geoespaciales que puedan ser de interés para el problema planteado, tales como distancia a espacios verdes, ubicación respecto a la calle Bulevar Artigas, distancia a centros hospitalarios, entre otras.

A su vez, se destaca que todas las variables asociadas al tamaño (área) del apartamento se encontraron entre las variables más relevantes. No obstante como se menciona en la Sección 4.3 y según se observó en el Capítulo 6 este resultado puede estar influenciado por la correlación lineal entre las mismas.

En cuanto al rol que cumple la variable distancia a la rambla este en el algoritmo con mayor performance predictiva, se tiene que a medida que el apartamento se encuentra más cerca de la rambla este de Montevideo, mayor es su precio de oferta. Esto debido a que la misma tiene un efecto inverso y no lineal sobre la predicción promedio del precio de oferta del apartamento. Donde esta última disminuye a medida que se incrementa la distancia entre el apartamento y la rambla.

A su vez, se observó que en el modelo aquellos apartamentos con dos o más baños completos tienen un precio superior a los apartamentos con un solo baño completo, en la medida que la predicción promedio de estos últimos es menor con respecto a los primeros. Más aún, esta diferencia se mantiene en la predicción considerando la distancia a la rambla este.

Por otra parte, en lo que respecta al área del apartamento y el ingreso medio del barrio donde este se encuentra, se tiene un mayor precio de oferta a medida que se incrementa el ingreso medio del barrio, dada el área total del apartamento. Esto debido a que el modelo realiza una predicción promedio superior para barrios con mayores niveles de ingreso.

Asimismo, para los apartamentos con área total inferior a 150 metros cuadrados, se observó a grandes rasgos que, dado el ingreso del barrio, la predicción del precio de oferta se incrementa a medida que aumenta el área total del apartamento.

Sin embargo, la metodología considerada (*gráficos de dependencia parcial*) posee ciertas limitantes ya que implica el cálculo de los efectos marginales expresados a través de un promedio. Donde dicho resultado puede estar influenciado por observaciones individuales y por lo tanto no lograr captar la heterogeneidad en las predicciones. De esta forma, se propone para futuros trabajos complementar el análisis implementando una metodología de *metódos locales modelo-agnósticos* principalmente mediante la construcción de los gráficos denominados *expectativas individuales condicionales (ICE plots)*. (Molnar, 2021) ([26])

Más aún, como se mencionó en la Sección 4.3, la metodología de *gráficos de dependencia parcial* no logra captar correlaciones entre las variables utilizadas en el análisis. Donde en este caso en particular se observó la presencia de una alta correlación lineal en ciertas variables.

Con el fin de solucionar esta limitante, para futuros trabajos se propone la utilización de los denominados *gráficos de efectos locales acumulados (ALE plots)* que trabajan con distribuciones condicionales en lugar de con distribuciones marginales como es el caso de la metodología de *gráficos de dependencia parcial*. (Molnar, 2021) ([26])

En cuanto a la selección del modelo con mejor performance predictiva, se destaca que no se tomó en cuenta el costo computacional que conllevó el ajuste. En éste sentido, si bien se observó que el ajuste por *Random Forest* tiene una performance predictiva inferior, el mismo implica un costo computacional considerablemente me-

nor. De esta forma, dependiendo de los recursos computacionales con los que cuenta el investigador, se propone a *Random Forest* como una alternativa para el problema planteado.

Siguiendo en esta línea, se considera de interés replicar el análisis con un mayor poder de cómputo donde, a pesar de que se trabajó con procesamiento en paralelo, con el fin de realizar un análisis más exhaustivo (principalmente en el proceso de selección de los *parámetros de ajuste*) se considera de importancia aumentar el poder de cómputo.

A su vez, debido a la limitante que presenta la API de *Mercado Libre* en la medida que no permite la obtención de datos históricos, para futuras investigaciones se considera de interés realizar un procedimiento de integración contínua con el fin de obtener descargas periódicas de la información disponible. Donde dicho procedimiento a la fecha de realización del trabajo se realiza de forma manual.

Por último, a la hora de realizar los ajustes de los diferentes modelos de aprendizaje estadístico, se considera de interés trabajar con otros paquetes del lenguaje de programación *R* con mayor capacidad de análisis tales como *mlr* ([2]) y *h20* ([23]) .

Capítulo 9

Referencias bibliográficas

[1] Alejandro Baldominos y col. "Identifying real estate opportunities using machine learning". En: *Applied sciences* 8.11 (2018), pág. 2321.

[2] Bernd Bischl y col. "mlr: Machine Learning in R". En: *Journal of Machine Learning Research* 17.170 (2016), págs. 1-5. URL: https://jmlr.org/papers/v17/15-066.html.

[3] Leo Breiman y col. *Classification and regression trees*. Routledge, 2017.

[4] Trevor S Breusch y Adrian R Pagan. "A simple test for heteroscedasticity and random coefficient variation". En: *Econometrica: Journal of the econometric society* (1979), págs. 1287-1294.

[5] Francesc Carmona. "Modelos lineales". En: *Pub. Univ. de Barcelona, Barcelona* (2005).

[6] Nitin R Chopde y Mangesh Nichat. "Landmark based shortest path detection by using A* and Haversine formula". En: *International Journal of Innovative Research in Computer and Communication Engineering* 1.2 (2013), págs. 298-302.

[7] Microsoft Corporation y Steve Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.16. 2020. URL: https://CRAN.R-project.org/package=doParallel.

[8] Karolien De Bruyne y Jan Van Hove. "Explaining the spatial variation in housing prices: an economic geography approach". En: *Applied Economics* 45.13 (2013), págs. 1673-1689.

[9] Environmental Systems Research Institute (ESRI). "ESRI Shapefile Technical Description, an ESRI white paper". En: (1998).

[10] Juan José Goyeneche, Leonardo Moreno y Marco Scavino. "Predicción del valor de un inmueble mediante técnicas agregativas". En: *Serie DT (17/1)* (2017).

[11] Brandon Greenwell y col. *gbm: Generalized Boosted Regression Models*. R package version 2.1.8. 2020. URL: `https://CRAN.R-project.org/package=gbm`.

[12] Brandon M. Greenwell y Bradley C. Boehmke. "Variable Importance Plots—An Introduction to the vip Package". En: *The R Journal* 12.1 (2020), págs. 343-366. URL: `https://doi.org/10.32614/RJ-2020-013`.

[13] Brandon M Greenwell, Bradley C Boehmke y B Gray. "Variable Importance Plots-An Introduction to the vip Package." En: *R J.* 12.1 (2020), pág. 343.

[14] Zvi Griliches. "Hedonic price indexes for automobiles: An econometric of quality change". En: *The price statistics of the federal goverment*. NBER, 1961, págs. 173-196.

[15] Juergen Gross y Uwe Ligges. *nortest: Tests for Normality*. R package version 1.0-4. 2015. URL: `https://CRAN.R-project.org/package=nortest`.

[16] Trevor Hastie. *Tibshirani R. Friedman J.: The Elements of Statistical Learning*. 2001.

[17] Robert J. Hijmans. *geosphere: Spherical Trigonometry*. R package version 1.5-10. 2019. URL: `https://CRAN.R-project.org/package=geosphere`.

[18] Gareth James y col. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[19] Alexandros Karatzoglou y col. "kernlab – An S4 Package for Kernel Methods in R". En: *Journal of Statistical Software* 11.9 (2004), págs. 1-20. URL: http://www.jstatsoft.org/v11/i09/.

[20] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-88. 2021. URL: https://CRAN.R-project.org/package=caret.

[21] Max Kuhn, Kjell Johnson y col. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[22] María Victoria Landaberry, Magdalena Tubio y col. *Estimación de índice de precios de inmuebles en Uruguay*. Inf. téc. 2015.

[23] Erin LeDell y col. *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*. R package version 3.32.1.3. 2021. URL: https://CRAN.R-project.org/package=h2o.

[24] Hubert W Lilliefors. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown". En: *Journal of the American statistical Association* 62.318 (1967), págs. 399-402.

[25] Michael Mayer. *missRanger: Fast Imputation of Missing Values*. R package version 2.1.3. 2021. URL: https://CRAN.R-project.org/package=missRanger.

[26] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[27] Christoph Molnar, Bernd Bischl y Giuseppe Casalicchio. "iml: An R package for Interpretable Machine Learning". En: *JOSS* 3.26 (2018), pág. 786. DOI: 10.21105/joss.00786. URL: https://joss.theoj.org/papers/10.21105/joss.00786.

[28] Sendhil Mullainathan y Jann Spiess. "Machine learning: an applied econometric approach". En: *Journal of Economic Perspectives* 31.2 (2017), págs. 87-106.

[29] Pablo Picardo y col. *Predicción de precios de vivienda: Aprendizaje estadístico con datos de oferta y transacciones para la ciudad de Montevideo*. Inf. téc. 2019.

[30] Jorge Ponce, Magdalena Tubio y col. *Precios de inmuebles: aproximaciones metodológicas y aplicación empírica*. BCU, 2013.

[31] Jorge Ponce y col. *Precio de fundamentos para las viviendas en Uruguay*. BCU, 2012.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[33] Sherwin Rosen. "Hedonic prices and implicit markets: product differentiation in pure competition". En: *Journal of political economy* 82.1 (1974), págs. 34-55.

[34] Terry Therneau y Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. 2019. URL: https://CRAN.R-project.org/package=rpart.

[35] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[36] Marvin N. Wright y Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". En: *Journal of Statistical Software* 77.1 (2017), págs. 1-17. DOI: 10.18637/jss.v077.i01.

[37] Achim Zeileis y Torsten Hothorn. "Diagnostic Checking in Regression Relationships". En: *R News* 2.3 (2002), págs. 7-10. URL: https://CRAN.R-project.org/doc/Rnews/.

Apéndice A

Anexo

A.1. Variables utilizadas

Tabla A.1: Variables utilizadas en los modelos implementados. Se detalla el nombre de la variable en la base de datos, la fuente de obtención de cada una de ellas, la descripción con el nombre utilizado en el informe y la naturaleza.

Nombre	**Fuente**	**Descripción**	**Naturaleza**
bedrooms	API	Cantidad de dormitorios	cualitativa
covered_area	API	Área cubierta	cuantitativa
full_bathrooms	API	Cantidad de baños completos	cualitativa
has_air_conditioning	API	Tiene aire acondicionado	cualitativa
has_balcony	API	Tiene balcón	cualitativa
has_common_laundry	API	Tiene área de lavanderia	cualitativa
has_garden	API	Tiene jardin	cualitativa
has_gym	API	El edificio tiene gimnasio	cualitativa
has_heating	API	Tiene calefaccción	cualitativa
has_lift	API	El edificio tiene ascensor	cualitativa
has_party_room	API	El edificio tiene salón de fiestas	cualitativa
has_patio	API	Tiene patio	cualitativa
has_security	API	El edificio tiene seguridad	cualitativa
has_swimming_pool	API	El edificio tiene piscina	cualitativa
has_telephone_line	API	Tiene línea telefónica	cualitativa
has_terrace	API	Tiene terraza	cualitativa
item_condition	API	Condición del Item	cualitativa

price	API	Precio de oferta en dólares estadounidenses	cuantitativa
total_area	API	Área total	cuantitativa
dist_rambla	Elaboración propia	Distancia a la rambla	cuantitativa
dist_shop	Elaboración propia	Distancia al centro comercial más cercano	cuantitativa
f_dif_update	Elaboración propia	Diferencia fechas última modificación y publicación	cualitativa
ingresomedio_ech	Elaboración propia	Ingreso medio por barrio	cuantitativa
no_covered_area	Elaboración propia	Área no cubierta	cuantitativa
zona_avditalia	Elaboración propia	Zona respecto a Avd. Italia	cualitativa

Tabla A.2: Método de imputación utilizado para las variables de entrada de los modelos implementados. En la columna método de imputación se detalla el o los métodos utilizados, o se especifica que no requiere en caso que la variable no presente valores faltantes.

Nombre	**Descripción**	**Método de imputación**
bedrooms	Cantidad de dormitorios	No Requiere
covered_area	Área cubierta	Media - Random Forest
full_bathrooms	Cantidad de baños completos	No Requiere
has_air_conditioning	Tiene aire acondicionado	No Requiere
has_balcony	Tiene balcón	No Requiere
has_common_laundry	Tiene área de lavanderia	No Requiere
has_garden	Tiene jardin	No Requiere
has_gym	El edificio tiene gimnasio	No Requiere
has_heating	Tiene calefaccción	No Requiere
has_lift	El edificio tiene ascensor	No Requiere
has_party_room	El edificio tiene salón de fiestas	No Requiere
has_patio	Tiene patio	No Requiere
has_security	El edificio tiene seguridad	No Requiere
has_swimming_pool	El edificio tiene piscina	No Requiere
has_telephone_line	Tiene línea telefónica	No Requiere
has_terrace	Tiene terraza	No Requiere
item_condition	Condición del Item	Random Forest
price	Precio de oferta en dólares estadounidenses	No Requiere
total_area	Área total	Media - Random Forest
dist_rambla	Distancia a la rambla	No Requiere
dist_shop	Distancia al centro comercial más cercano	No Requiere

f_dif_update	Diferencia fechas última modificación y publicación	No Requiere
ingresomedio_ech	Ingreso medio por barrio	No Requiere
no_covered_area	Área no cubierta	Media - Random Forest
zona_avditalia	Zona respecto a Avd. Italia	No Requiere

Tabla A.3: Propoción de valores faltantes en las variables utilizadas en los modelos estadísticos implementados. Se detallan las variables con proporción estrictamente positiva.

Nombre	Descripción	Proporción de **valores faltantes**
item_condition	Condición	0.1171
no_covered_area	Área no cubierta	0.0355
total_area	Área total	0.0044
covered_area	Área cubierta	0.0021

A.2. Barrios de Montevideo

Tabla A.4: Barrios de Montevideo. En la columna identificador se especifica el código de cada barrio en la API de Mercado libre. En la columna Nombre ML se especifica el nombre asociado a cada identificador en Mercado libre. En la columna Nombre INE se especifica el nombre INE asociado a cada nombre en Mercado libre.

Identificador	**Nombre ML**	**Nombre INE**
TUxVQ0FHVWUwMzc3	Aguada	Aguada
TUxVQ0FJUjE1NDM	Aires Puros	Aires Puros
TUxVQ0FSUmJmZDVl	Arroyo Seco	Aguada
TUxVQ0FUQTYzNjc	Atahualpa	Atahualpa
TUxVQ0JFTDUyODE	Bella Vista	Reducto
TUxVQ0JFTDU0MjU0	Belvedere	Belvedere
TUxVQ0JPTDIxMDY	Bolivar	Mercado Modelo, Bolivar
TUxVQ0JSQWNhNzZl	Brazo Oriental	Brazo Oriental
TUxVQ0JVQzNlMDdl	Buceo	Buceo
TUxVQ0NBUDYxOTA	Capurro	Capurro, Bella Vista
TUxVQ0NBUmRhYWU0	Carrasco	Carrasco
TUxVQ0NFTjVjMTM	Centro	Centro
TUxVQ0NFUjI3NDg	Cerrito	Cerrito
TUxVQ0NFUjEwNDc	Cerro	Cerro
TUxVQ0NJVTk5MTU	Ciudad Vieja	Ciudad Vieja
TUxVQ0NPTDk2ZjYz	Colón	Colon Centro y Noroeste
TUxVQ0NPUjZmZjNm	Cordón	Cordon
TUxVQ0dPRTY1NDU	Goes	Aguada
TUxVQ0lUVTM0MjQ	Ituzaingó	Ituzaingo
TUxVQ0pBQzJiODI2	Jacinto Vera	Jacinto Vera
TUxVQ0pBUjMzOTE	Jardines Hipódromo	Jardines del Hipodromo
TUxVQ0xBWjk5YTE5	La Blanqueada	La Blanqueada
TUxVQ0xBQzU0NjU	La Comercial	La Comercial
TUxVQ0xBRjMyODM	La Figurita	La Figurita
TUxVQ0xBVDQ5ODI	La Teja	La Teja
TUxVQ0xBUzE0MjM	Las Acacias	Las Acacias
TUxVQ0xFWjQ0NTA	Lezica	Lezica, Melilla
TUxVQ01BTDE0YmY1	Malvin	Malvin
TUxVQ01BTDk4MDg	Malvin Norte	Malvin Norte
TUxVQ01BTjY0Mzc	Manga	Manga
TUxVQ01BUjczNTk	Maroñas	Maroñas, Parque Guarani

TUxVQ01BUjYwMDA	Maroñas, Curva	Maroñas, Parque Guarani
TVhYTWVyY2FkbyBNb2RlbG9UVXhWVUUxUFRsb	Mercado Modelo	Mercado Modelo, Bolivar
TUxVQ01PTjc2Nzc	Montevideo	Centro
TUxVQ05VRTk3MTk	Nuevo París	Nuevo Paris
TUxVQ1BBTDU0NzY	Palermo	Palermo
TUxVQ1BBUjVkNGE4	Parque Batlle	Pque. Batlle, V. Dolores
TUxVQ1BBUmU3Y2Nj	Parque Rodó	Parque Rodo
TUxVQ1BBUzc3MDM	Paso de la Arena	Paso de la Arena
TUxVQ1BBUzQzMDA	Paso Molino	Belvedere
TUxVQ1BF0TUxNDI	Peñarol	Peñarol, Lavalleja
TVhYUGVyZXogQ2FzdGVsbGFub3NUVXhWVUUxU	Perez Castellanos	Castro, P. Castellanos
TUxVQ1BJRWMyMjhl	Piedras Blancas	Piedras Blancas
TUxVQ1BPQzM5ZGRi	Pocitos	Pocitos
TUxVQ1BPQzIwNDU	Pocitos Nuevo	Pocitos
TUxVQ1BSQTYwOTJl	Prado	Prado, Nueva Savona
TUxVQ1BVRTI0NDA	Puerto Buceo	Buceo
TUxVQ1BVTjJmMjkx	Punta Carretas	Punta Carretas
TUxVQ1BVTjYxODI	Punta Gorda	Punta Gorda
TUxVQ1BVTjMxMTE	Punta Rieles	Pta. Rieles, Bella Italia
TUxVQ1JFRDg3YzQy	Reducto	Reducto
TUxVQ1NBWTkxODY	Sayago	Sayago
TUxVQ1RSRTg3OGM3	Tres Cruces	Tres Cruces
TUxVQ1VOSTVkOGFk	Unión	Union
TUxVQ1ZJTDE2MDU	Villa Biarritz	Punta Carretas
TUxVQ1ZJTDk1MTY	Villa Dolores	Pque. Batlle, V. Dolores
TUxVQ1ZJTDI2MzY	Villa Española	Villa Española
TUxVQ1ZJTDRkMTY	Villa Muñoz	Villa Muñoz, Retiro

A.3. Fórmula de Haversine

La fórmula de Haversine para el cálculo de distancias sobre un cuerpo esférico tiene la siguiente expresión:

$$d = 2\,r\,\mathrm{sen}^{-1}\left(\sqrt{\mathrm{sen}^2\,\tfrac{\phi_2-\phi_1}{2} + \cos(\phi_1)\cos(\phi_2)\,\mathrm{sen}^2\,\tfrac{\psi_2-\psi_1}{2}}\right)$$

siendo d la distancia entre dos puntos de longitud y latitud (ψ_1, ϕ_1) y (ψ_2, ϕ_2) respectivamente, y r el radio de la tierra. (Chopde, 2013) ([6])

A.4. Árbol de regresión de la variable precio de oferta en función de la latitud y longitud

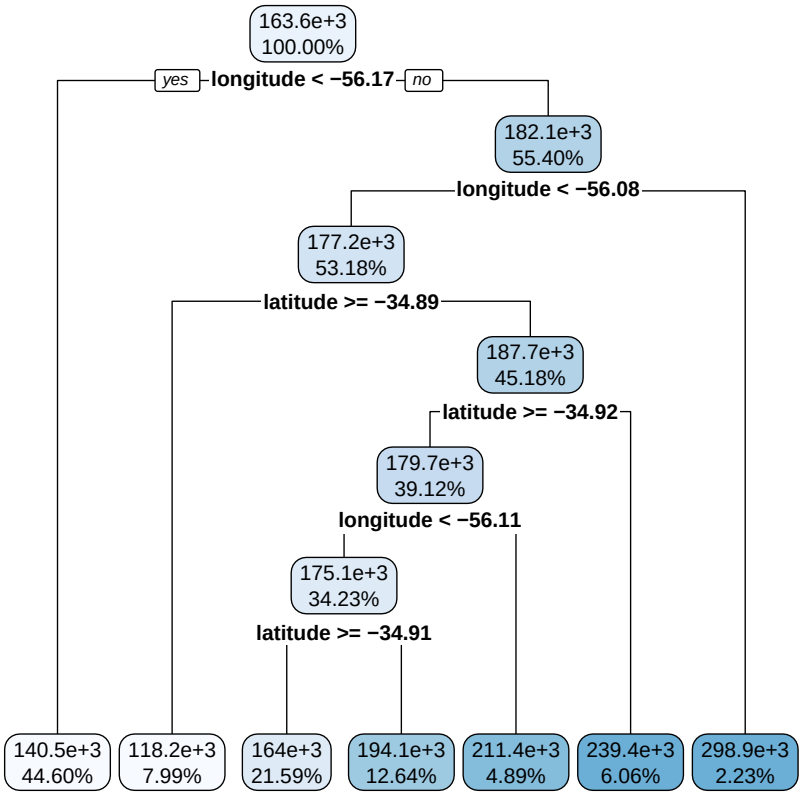

Figura A.1: Árbol de regresión obtenido al ajustar la variable precio de oferta en dólares mediante la base de datos construida en base a información obtenida de *Mercado Libre*, considerando como variables de entrada latitud y longitud del apartamento. A dicho árbol se le realizó el correspondiente proceso de poda, el cual queda conformado por 7 nodos terminales (hojas). Los porcentajes dentro de cada nodo indican el porcentaje del número de observaciones que se encuentran en el mismo. Además, se explicita la predicción de las observaciones pertenecientes al nodo. Donde a mayor intensidad del color azul, mayor la predicción en cuanto al precio de oferta del apartamento.

A.5. Modelo de Regresión Lineal Múltiple

Variable	Coeficientes estimados	Error estándar	Estadístico T	P-Valor
(Intercept)	10,291.8315273	1,183.6113269	8.695280	0.0000000
bedrooms1	31,849.2355274	577.6585333	55.135056	0.0000000
bedrooms2	56,154.0820466	624.6059422	89.903215	0.0000000
bedrooms3	60,695.5081339	859.1698830	70.644362	0.0000000
bedrooms4 o mas	63,505.5488681	1,494.6815303	42.487679	0.0000000
covered_area	286.0309743	11.4506571	24.979438	0.0000000
full_bathrooms2 o mas	57,917.3699464	605.5141574	95.649902	0.0000000
has_air_conditioningSí	576.1963331	474.9288274	1.213227	0.2250478
has_telephone_lineSí	-540.7877933	514.6436818	-1.050800	0.2933544
total_area	235.6874266	11.8770124	19.843999	0.0000000
has_balconySí	5,300.6290712	489.5129589	10.828373	0.0000000
has_common_laundrySí	4,095.9987941	590.0537797	6.941738	0.0000000
has_gardenSí	-7,126.2121341	1,144.2204302	-6.228006	0.0000000
has_gymSí	8,365.6322625	556.6842812	15.027606	0.0000000
has_heatingSí	6,871.7922835	597.4162051	11.502521	0.0000000
has_liftSí	-5,116.0906825	471.2157205	-10.857216	0.0000000
has_party_roomSí	-3,379.9836248	852.4796094	-3.964885	0.0000735
has_patioSí	-8,593.1066939	699.6443033	-12.282108	0.0000000
has_securitySí	7,878.6219144	491.9045509	16.016566	0.0000000
has_swimming_poolSí	21,732.2296915	833.1172992	26.085438	0.0000000
has_terraceSí	9,207.9863531	408.7508233	22.527138	0.0000000
f_dif_updateMORE 1 Year	9,165.2587167	463.1062695	19.790833	0.0000000
f_dif_updateNOW	-5,609.8263036	589.4069914	-9.517746	0.0000000
no_covered_area	-88.9715849	12.6695025	-7.022500	0.0000000
ingresomedio_ech	0.5798689	0.0078312	74.046197	0.0000000
zona_avditaliaSur	3,502.7939699	689.3476257	5.081317	0.0000004
dist_shop	0.8981036	0.2309495	3.888745	0.0001009
dist_rambla	-6.7236330	0.1884877	-35.671466	0.0000000

Tabla A.6: Tabla de resumen de los coeficientes del ajuste mediante el *Modelo de Regresión Lineal Múltiple* con imputación de valores faltantes por la media. Se presentan los coeficientes estimados, la estimación del error estándar de cada uno de ellos, el estadístico t de student y su p-valor asociado.

Variable	Coeficientes estimados	Error estándar	Estadístico T	P-Valor
(Intercept)	16,745.1507654	1,180.4959799	14.184843	0.0000000
bedrooms1	33,281.1356717	571.7861338	58.205566	0.0000000
bedrooms2	58,482.4409230	619.9169539	94.339154	0.0000000
bedrooms3	64,774.2155628	856.0672493	75.664868	0.0000000
bedrooms4 o mas	67,645.3006624	1,482.5163766	45.628704	0.0000000
covered_area	261.7967740	12.4182570	21.081604	0.0000000
full_bathrooms2 o mas	58,899.5725227	598.9988494	98.330026	0.0000000
has_air_conditioningSí	1,048.2984217	469.3579127	2.233473	0.0255213
has_telephone_lineSí	-1,321.5513508	508.7986235	-2.597396	0.0093956
total_area	268.8181274	13.1204221	20.488527	0.0000000
has_balconySí	3,216.4826009	486.4857097	6.611669	0.0000000
has_common_laundrySí	2,446.0384508	584.6121875	4.184036	0.0000287
has_gardenSí	-5,833.7811710	1,130.8276028	-5.158860	0.0000002
has_gymSí	4,167.9537392	560.5081743	7.436027	0.0000000
has_heatingSí	9,166.7379071	593.1289075	15.454883	0.0000000
has_liftSí	-1,097.1906899	477.1134646	-2.299643	0.0214718
has_party_roomSí	-3,586.0053751	842.1872545	-4.257967	0.0000207
has_patioSí	-7,621.5180070	691.8943673	-11.015436	0.0000000
has_securitySí	6,879.6196200	486.6635000	14.136297	0.0000000
has_swimming_poolSí	23,322.2401491	824.0624838	28.301543	0.0000000
has_terraceSí	7,580.6685097	406.0864002	18.667625	0.0000000
f_dif_updateMORE 1 Year	9,572.1948395	457.6133498	20.917648	0.0000000
f_dif_updateNOW	-4,427.0588665	583.0690337	-7.592684	0.0000000
no_covered_area	-132.9281945	13.8960723	-9.565883	0.0000000
ingresomedio_ech	0.5669746	0.0077438	73.216559	0.0000000
zona_avditaliaSur	4,835.3589664	682.0005720	7.089963	0.0000000
dist_shop	0.7160031	0.2281974	3.137648	0.0017039
dist_rambla	-6.3422995	0.1865249	-34.002431	0.0000000
item_conditionUsado	-15,627.9055712	404.8155646	-38.605002	0.0000000

Tabla A.8: Tabla de resumen de los coeficientes del ajuste mediante el *Modelo de Regresión Lineal Múltiple* con imputación de valores faltantes por *Random Forest*. Se presentan los coeficientes estimados, la estimación del error estándar de cada uno de ellos, el estadístico t de student y su p-valor asociado.

Método de imputación	**P-Valor de la prueba**	**Decisión**
Media	<2.2e-16	Se rechaza H0
Random Forest	<2.2e-16	Se rechaza H0

Tabla A.9: Resultados de la prueba de *Breusch-Pagan* para el análisis de homocedasticidad de los residuos de los *Modelos de Regresión Lineal Múltiple* ajustados según método de imputación de valores faltantes utilizado.

Método de imputación	P-Valor de la prueba	Decisión
Media	<2.2e-16	Se rechaza H0
Random Forest	<2.2e-16	Se rechaza H0

Tabla A.10: Resultados de la prueba de *Lilliefors* para el análisis de normalidad de los residuos de los *Modelos de Regresión Lineal Múltiple* ajustados según método de imputación de valores faltantes utilizado.

A.6. Árbol de regresión

CP	Nro particiones	Error de CV
0.4271076	0	1.0000448
0.0793501	1	0.5729712
0.0532923	2	0.4937803
0.0443657	3	0.4405051
0.0229089	4	0.3962241
0.0139919	5	0.3733362
0.0129397	6	0.3595139
0.0125827	7	0.3516558
0.0106675	8	0.3338580
0.0100000	9	0.3232129

Tabla A.11: Tabla resumen, indicadores para realizar el proceso de poda. Se tiene que CP es el parámetro de complejidad, donde se selecciona aquel valor el cual implique un menor error en el proceso de *validación cruzada*. La metodología de imputación de valores faltantes utilizada en este caso es imputación por la media.

CP	**Nro particiones**	**Error de CV**
0.4271076	0	1.0000448
0.0793501	1	0.5729712
0.0533506	2	0.4937803
0.0447131	3	0.4404471
0.0228827	4	0.3958197
0.0140741	5	0.3729577
0.0128948	6	0.3590529
0.0125830	7	0.3501465
0.0106681	8	0.3334427
0.0100000	9	0.3221150

Tabla A.12: Tabla resumen, indicadores para realizar el proceso de poda. Se tiene que CP es el parámetro de complejidad, donde se selecciona aquel valor el cual implique un menor error en el proceso de *validación cruzada*. La metodología de imputación de valores faltantes utilizada en este caso es imputación por *Random Forest*.

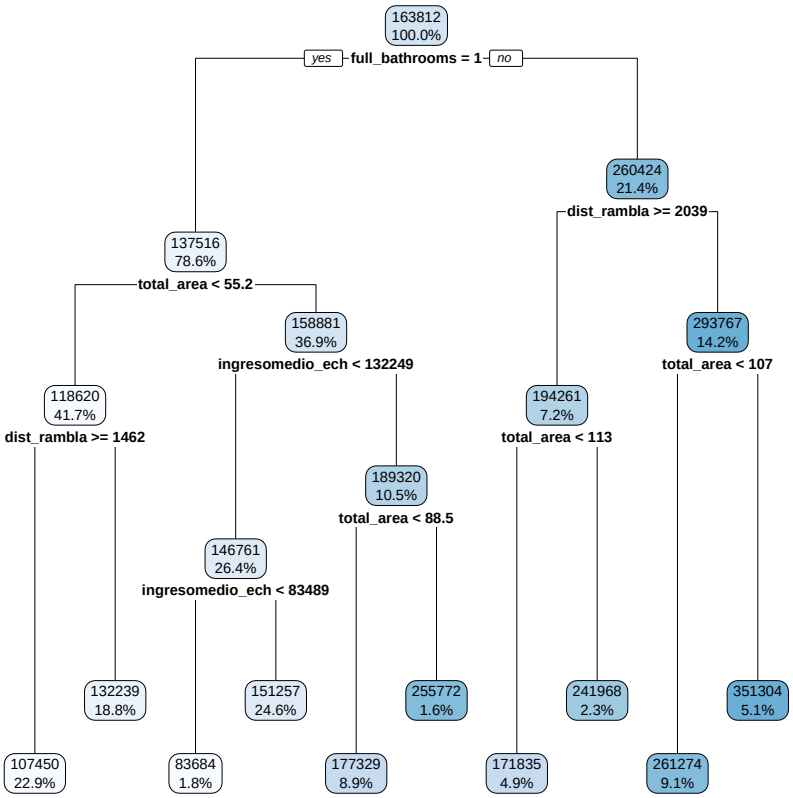

Figura A.2: Árbol de regresión obtenido al ajustar la variable precio de oferta en dólares mediante la base de datos construida en base a información obtenida de *Mercado Libre*, una vez realizado el proceso de poda con un valor de CP igual a 0.01. El método de imputación sobre valores faltantes es en éste caso imputación por *Random Forest*. El árbol se conforma por 10 nodos terminales (hojas). Los porcentajes dentro de cada nodo indican el porcentaje del número de observaciones que se encuentran en el mismo. Además, se explicita la predicción de las observaciones pertenecientes al nodo. A su vez, se destaca que se realizan 9 particiones.

A.7. Parámetros de ajuste

Regla de partición	Min. obs.	Can. de arboles	Cant. de variables
Variance	5	500	2
Variance	5	500	12
Variance	5	500	23

Tabla A.13: Grilla de *parámetros de ajuste* por defecto para los modelos ajustados por *Random Forest*. Se destaca que en todos los casos la regla de partición utilizada es la que minimiza la suma de cuadrados de los residuos y la cantidad de árboles utilizada en el ajuste es 500. Asimismo, la cantidad de observaciones mínimas en un nodo terminal se mantiene constante en 5.

Min. obs.	Cant. de árboles	Particiones	Tasa de aprendizaje
10	50	1	0.1
10	100	1	0.1
10	150	1	0.1
10	50	2	0.1
10	100	2	0.1
10	150	2	0.1
10	50	3	0.1
10	100	3	0.1
10	150	3	0.1

Tabla A.14: Grilla de *parámetros de ajuste* por defecto para los modelos ajustados por *Boosting*. Se destaca que el número de observaciones mínimas en cada nodo terminal se mantiene constante en 10.

Parámetro de complejidad	**Umbral**	**Parámetro de escala**
0.25	0.1	0.03085
0.50	0.1	0.03085
1.00	0.1	0.03085

Tabla A.15: Grilla de *parámetros de ajuste* por defecto para los modelos ajustados por *Support Vector Regression*. Se destaca que el valor del umbral se mantiene constante en 0.1.

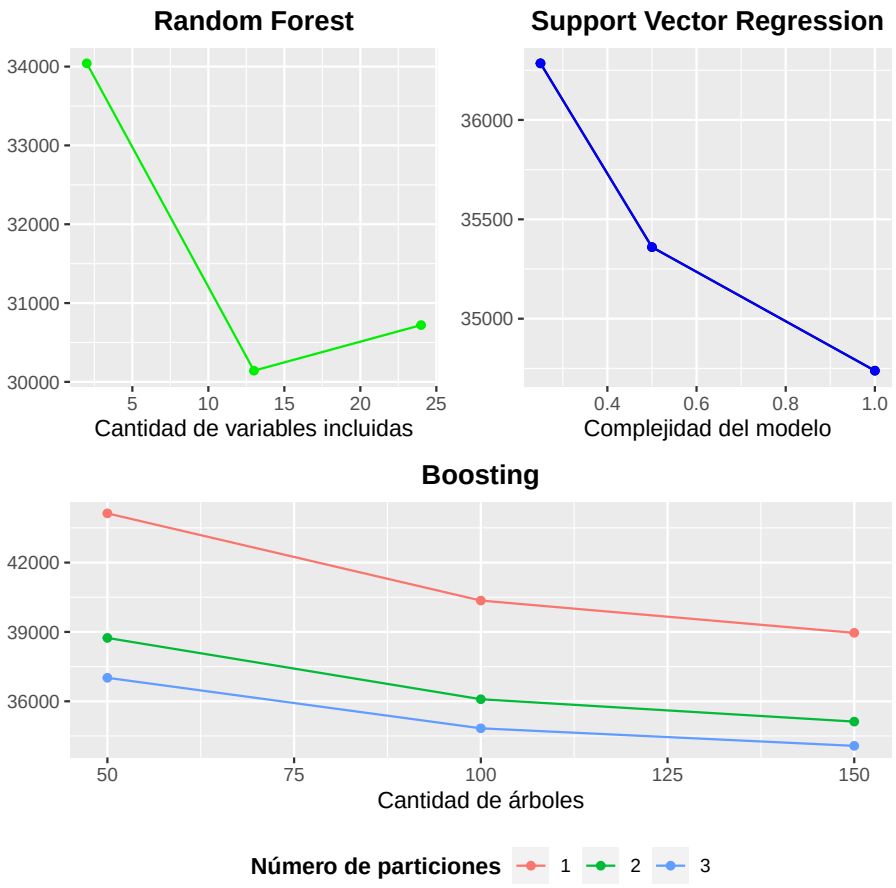

Figura A.3: Gráfico de la evolución del error cuadrático medio de predicción (RMSE) en función de los valores de los *parámetros de ajuste* por defecto para el caso de imputación por *Random Forest*. En el panel superior izquierdo se encuentran los resultados para los ajustes por *Random Forest*, en el panel superior derecho se encuentran los resultados para el ajuste por *Support Vector Regression*, y en el panel inferior izquierdo se encuentran los resultados para el ajuste por *Boosting*.

Imputación	**Regla part.**	**Obs. min.**	**Cant. de árboles**	**Particiones**	**RMSE**	R^2	**MAE**
Media	variance	5	500	12	30,295	0.84	19,658
Media	variance	5	500	13	30,323	0.84	19,661
Media	variance	5	500	14	30,355	0.84	19,671
Media	variance	5	500	15	30,400	0.84	19,686
Media	variance	5	500	16	30,427	0.84	19,698
Media	variance	5	500	17	30,477	0.84	19,718
Media	variance	5	500	18	30,534	0.84	19,741
Media	variance	5	500	19	30,578	0.84	19,759
Media	variance	5	500	20	30,638	0.84	19,788
Random Forest	variance	5	500	12	30,118	0.85	19,563
Random Forest	variance	5	500	13	30,147	0.85	19,565
Random Forest	variance	5	500	14	30,182	0.85	19,573
Random Forest	variance	5	500	15	30,220	0.85	19,580
Random Forest	variance	5	500	16	30,252	0.84	19,594
Random Forest	variance	5	500	17	30,292	0.84	19,612
Random Forest	variance	5	500	18	30,335	0.84	19,630
Random Forest	variance	5	500	19	30,384	0.84	19,652
Random Forest	variance	5	500	20	30,428	0.84	19,667

Tabla A.17: Principales medidas de resumen para los modelos ajustados por *Random Forest* según la grilla especificada en el proceso de selección de los *parámetros de ajuste* según la metodología de imputación de valores faltantes utilizada. Se destaca que en todos los casos la cantidad de árboles utilizada en el ajuste es 500. Asimismo, la cantidad de observaciones mínimas en un nodo terminal se mantiene constate en 5.

Tabla A.18: Principales medidas de resumen para los modelos ajustados por Boosting según la grilla especificada en el proceso de parámetros de ajuste según la metodología de imputación de valores faltantes utilizada.

Imputación	**Obs. min.**	**Cant. de árboles**	**Tasa de aprendizaje**	**Particiones**	**RMSE**	R^2	**MAE**
Media	10	5,000	0.100	10	26,758	0.88	17,752
Media	10	2,000	0.100	10	27,951	0.86	19,068
Media	10	5,000	0.100	5	28,164	0.86	19,182
Media	10	1,500	0.100	10	28,380	0.86	19,467
Media	10	2,000	0.100	5	29,487	0.84	20,380
Media	10	5,000	0.100	3	29,595	0.84	20,453
Media	10	1,500	0.100	5	29,912	0.84	20,757
Media	10	5,000	0.010	10	29,952	0.83	20,821
Media	10	500	0.100	10	30,205	0.83	21,004
Media	10	2,000	0.100	3	30,741	0.83	21,392
Media	10	1,500	0.100	3	31,145	0.82	21,718
Media	10	5,000	0.010	5	31,409	0.82	21,931
Media	10	2,000	0.010	10	31,494	0.81	21,951
Media	10	500	0.100	5	31,583	0.82	22,052
Media	10	1,500	0.010	10	31,952	0.81	22,276
Media	10	5,000	0.010	3	32,529	0.80	22,756
Media	10	500	0.100	3	32,615	0.81	22,842
Media	10	2,000	0.010	5	32,803	0.80	22,947
Media	10	1,500	0.010	5	33,243	0.79	23,262
Media	10	2,000	0.010	3	33,821	0.78	23,699
Media	10	500	0.010	10	34,032	0.74	23,730
Media	10	1,500	0.010	3	34,302	0.77	24,044
Media	10	500	0.010	5	35,685	0.70	25,008
Media	10	500	0.010	3	37,314	0.65	26,214
Random Forest	10	5,000	0.100	10	26,624	0.89	17,669
Random Forest	10	2,000	0.100	10	27,755	0.86	18,934
Random Forest	10	5,000	0.100	5	28,056	0.86	19,115
Random Forest	10	1,500	0.100	10	28,237	0.86	19,374
Random Forest	10	2,000	0.100	5	29,264	0.85	20,257
Random Forest	10	5,000	0.100	3	29,416	0.85	20,310
Random Forest	10	1,500	0.100	5	29,755	0.84	20,648
Random Forest	10	500	0.100	10	29,946	0.83	20,832
Random Forest	10	2,000	0.100	3	30,555	0.83	21,267
Random Forest	10	1,500	0.100	3	30,925	0.83	21,570
Random Forest	10	500	0.100	5	31,320	0.82	21,890
Random Forest	10	500	0.100	3	32,353	0.81	22,675

Random Forest	10	5,000	0.001	10	33,804	0.74	23,585
Random Forest	10	5,000	0.001	5	35,472	0.69	24,856
Random Forest	10	5,000	0.001	3	37,104	0.65	26,062
Random Forest	10	2,000	0.001	10	38,871	0.51	27,261
Random Forest	10	2,000	0.001	5	41,254	0.46	29,142
Random Forest	10	1,500	0.001	10	41,839	0.40	29,454
Random Forest	10	2,000	0.001	3	43,677	0.41	31,267
Random Forest	10	1,500	0.001	5	44,254	0.36	31,331
Random Forest	10	1,500	0.001	3	46,869	0.32	33,726
Random Forest	10	500	0.001	10	57,437	0.10	41,379
Random Forest	10	500	0.001	5	59,217	0.09	42,807
Random Forest	10	500	0.001	3	60,959	0.08	44,167

Imputación	**Umbral**	**Parámetro de escala**	**Parámetro de complejidad**	**RMSE**	R^2	**MAE**
Media	0.1	0.01000	1	36,718	0.77	24,956
Media	0.1	0.01000	3	35,963	0.78	24,425
Media	0.1	0.01000	5	35,687	0.78	24,225
Media	0.1	0.03085	1	35,156	0.79	23,798
Media	0.1	0.03085	3	34,764	0.80	23,494
Media	0.1	0.03085	5	34,830	0.79	23,514
Media	0.1	0.05000	1	35,246	0.79	23,726
Media	0.1	0.05000	3	35,096	0.79	23,614
Media	0.1	0.05000	5	35,321	0.79	23,753
Random Forest	0.1	0.01000	1	36,295	0.78	24,612
Random Forest	0.1	0.01000	3	35,522	0.79	24,068
Random Forest	0.1	0.01000	5	35,231	0.79	23,865
Random Forest	0.1	0.03085	1	34,539	0.80	23,351
Random Forest	0.1	0.03085	3	34,183	0.80	23,081
Random Forest	0.1	0.03085	5	34,259	0.80	23,120
Random Forest	0.1	0.05000	1	34,415	0.80	23,183
Random Forest	0.1	0.05000	3	34,319	0.80	23,120
Random Forest	0.1	0.05000	5	34,575	0.80	23,282

Tabla A.19: Principales medidas de resumen para los modelos ajustados por *Support Vector Regression* según la grilla especificada en el proceso de *parámetros de ajuste* y según la metodología de imputación de valores faltantes utilizada. Se destaca que el valor del umbral tolerado para los ajustes se mantiene constante en 0.1.