Instituto Tecnológico
y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Estrategias Basadas en Datos para la inversión Inmobiliaria en la Era de Airbnb

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Presenta: **JUAN PABLO ACEVEDO ROBLES**

Director **DRA. ALMA NAYELI RODRÍGUEZ VÁZQUEZ**

Tlaquepaque, Jalisco, enero de 2024.

AGRADECIMIENTOS

A mi familia, por su apoyo constante, en los momentos complicados de estos últimos 3 años. Un agradecimiento especial a mi mamá, cuyos 40 años de trabajo en el ITESO me facilitaron enormemente continuar con mi educación.

Expreso mi profundo agradecimiento a la doctora Alma Rodríguez, cuya guía experta y consejo sabio han sido cruciales en la culminación de este trabajo. Su paciencia y dedicación son profundamente apreciadas.

Agradezco sinceramente a los profesores y colegas del departamento de matemáticas y física, cuya erudición y colaboración han enriquecido mi experiencia académica y han contribuido significativamente a mi crecimiento intelectual.

A mis amigos, compañeros de tantas jornadas, gracias por cada momento de alivio, cada risa compartida y por estar allí en cada paso del camino.

RESUMEN

En la era digital actual, las transformaciones tecnológicas han reinventado la industria del hospedaje. Una de las plataformas de alojamiento que se ha convertido en líder a nivel mundial es Airbnb. Con más de cuatro millones de anfitriones registrados en el año 2023, la plataforma de Airbnb presenta una oportunidad lucrativa para los propietarios de inmuebles que buscan incrementar sus ingresos mediante el alquiler a corto plazo de sus propiedades. Sin embargo, para un inversionista, elegir la ubicación y el tipo de propiedad sin información que oriente la toma de decisiones puede resultar en una inversión de alto riesgo. En este contexto, surge la necesidad de analizar los datos generados por la plataforma de Airbnb para crear modelos predictivos que sirvan como herramienta para asistir a los inversionistas en tomar decisiones estratégicas y rentables. Bajo ese contexto, en este proyecto se propone la utilización de modelos basados en árboles de decisión y redes neuronales para predecir la cantidad de noches reservadas en las diferentes zonas del área metropolitana de Guadalajara y estimar la utilidad anual. Con esta información, un comprador potencial podrá elegir una propiedad basándose en información que le permita identificar la ubicación y las características más idóneas para incrementar el rendimiento de su inversión. Ambos modelos son evaluados utilizando diferentes métricas de desempeño, demostrando su robustez y eficiencia en sus predicciones.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

1.1. Contexto

México es uno de los destinos turísticos más frecuentados a nivel mundial. En el año 2021, de acuerdo con el INEGI (INEGI, 2022), México atrajo a 55.3 millones de turistas internacionales, convirtiéndose en el segundo país más visitado del mundo. Esta afluencia turística, aunada con el turismo nacional, contribuyó con el 7.5% del Producto Interno Bruto (PIB) del país, destacando el rol crucial del turismo en la economía mexicana (INEGI, 2022).

Uno de los sectores económicos más beneficiados con el turismo es el hotelero. Sin embargo, aunque los hoteles solían ser lo opción de alojamiento predominante en el mercado, en la actualidad existen distintas opciones que brindan servicio de hospedaje gracias al desarrollo de las plataformas digitales. Las opciones de alojamiento para los turistas, que antes eran limitadas y poco asequibles para un gran sector de la población, ahora son variadas y contemplan un amplio rango de presupuestos. Al año 2023, los turistas tienen la posibilidad de explorar diversas opciones de alojamiento en el mercado, siendo Airbnb una de las plataformas más populares.

La plataforma Airbnb nace oficialmente en el año 2007 en la ciudad de San Francisco, cuando dos de sus fundadores, Brian Chesky y Joe Gebbia, estuvieron a punto de perder su departamento por un aumento en la renta (Airbnb, 2023). Con el objetivo de aumentar sus ingresos, Bryan y Joe decidieron ofrecer parte de su espacio como alojamiento a personas que estuvieran interesadas en visitar la ciudad. Ellos rentaban el espacio a un menor costo, lo cual resultaba atractivo para los visitantes. De ahí nace la idea que ha posicionado a Airbnb como una de las plataformas digitales que ha revolucionado el alojamiento tradicional.

El rápido crecimiento de Airbnb ha traído beneficios a inversionistas de todo el mundo. A marzo de 2023, Airbnb alberga a más de 4 millones de anfitriones globalmente. Esto ha supuesto beneficios económicos superiores a los $180,000 millones de dólares para los anfitriones, habiendo acogido a más de 1000 millones de huéspedes (Airbnb, 2023). Esta información subraya que Airbnb se ha consolidado como una plataforma innovadora, permitiendo a los anfitriones rentar sus propiedades de manera flexible, en lugar de limitarse a una cuota fija mensual. Sin embargo, el crecimiento de Airbnb no ha estado exento de

desafíos. En algunos lugares, la plataforma ha enfrentado críticas debido al incremento de los precios de las viviendas y problemas relacionados con el ruido y la inseguridad. Además, la industria hotelera tradicional también se ha visto afectada y ha tenido que adaptar sus estrategias para competir en este nuevo entorno.

La incorporación de la plataforma de Airbnb en México ha causado un impacto considerable en la economía del país. De acuerdo con el sitio web *alltherooms* (AllTheRooms, 2023), en el año 2021, México registró 391,612 propiedades, con una demanda de 13,252,133 noches y unos ingresos brutos de $1,256,876,562 dólares. Este volumen de actividad sitúa a México en el octavo lugar a nivel mundial dentro de la plataforma Airbnb. Estos datos subrayan el importante papel que juega Airbnb en la economía de México con la introducción del alojamiento compartido.

Las generaciones que están ejerciendo un impacto significativo en el mercado actual son los Millennials (personas nacidas entre 1981 y 1995) y la generación Centennial (personas nacidas entre el 2000 y 2010). Según la revista Expansión (Patiño, 2021), ya en el año 2021, estas generaciones representaban el 39.5% del mercado. Se estima que para 2025, los Millennials conformarán el 75% de la fuerza laboral en América Latina (Amieva, 2021). Estas generaciones, caracterizadas por su afinidad hacia la tecnología, prefieren reservar un alojamiento a través de una página web o mediante una aplicación para teléfonos inteligentes que mediante el método tradicional. Por tal motivo, la cantidad de usuarios de Airbnb se ha incrementado, ocasionando que se genere un gran volumen de información proveniente tanto de las búsquedas y reservaciones de los huéspedes, como de las características y amenidades de los alojamientos que se ofrecen.

El éxito de Airbnb ha motivado a muchos inversionistas a adquirir propiedades con el propósito de alquilarlas a través de la plataforma. Sin embargo, elegir la ubicación y el tipo de propiedad sin información que oriente la toma de decisiones puede resultar en una inversión de alto riesgo. En este contexto, surge la necesidad de analizar los datos generados por la plataforma de Airbnb para crear modelos predictivos que sirvan como herramienta para asistir a los inversionistas en tomar decisiones estratégicas y rentables.

Existe una amplia variedad de modelos predictivos que pueden utilizarse para simular el comportamiento y las tendencias de los usuarios en la reservación de alojamientos mediante la plataforma de Airbnb. Algunos modelos populares utilizados para realizar predicciones son los basados en árboles de decisión (Max Kuhn, 2016) y las redes neuronales (Nielsen, 2019). Estos modelos son de naturaleza distinta y tienen características diferentes. Por un lado, los basados en árboles de decisión ofrecen fácil implementación e interpretación sencilla, lo cual los hace ideales para comprender y comunicar la toma de decisiones. Por otro lado, las redes neuronales permiten crear relaciones complejas, lo cual las hace eficientes para modelar comportamientos sofisticados y complicados. A pesar de sus diferencias, ambos modelos han ganado popularidad por haber demostrado un alto desempeño en distintas aplicaciones (Carrillo, 2019).

Con el objetivo de asistir a los inversionistas en la toma de decisiones para la adquisición de propiedades de alquiler, en este proyecto se propone la utilización de modelos basados en árboles de decisión y redes neuronales para predecir la cantidad de noches reservadas en las diferentes zonas del área metropolitana de Guadalajara y estimar la utilidad anual del anfitrión a partir de estas reservaciones. Con esta información, un comprador potencial podrá elegir una propiedad basándose en información que le permita identificar la ubicación y las características más idóneas para incrementar el rendimiento de su inversión.

1.2. Justificación

México, con su atractivo turístico tanto a nivel nacional como internacional, representa un mercado altamente atrayente para propietarios de inmuebles y potenciales inversionistas que buscan incrementar sus ingresos mediante el alquiler a corto plazo de sus propiedades en plataformas como Airbnb. Sin embargo, tomar decisiones sin conocimiento sobre la ubicación y el tipo de propiedad ideal puede resultar un desafío. La inversión inmobiliaria es una decisión costosa que puede ser de alto riesgo si se hace sin información que oriente la toma de decisiones. En este contexto, surge la necesidad de analizar los datos que se generan mediante la plataforma de Airbnb y crear modelos predictivos para ayudar a los inversionistas a tomar decisiones estratégicas y rentables. Algunos de los beneficios que se obtendrían de este análisis son:

- Los inversionistas podrían tomar decisiones fundamentadas sobre dónde y qué tipo de propiedad comprar. Esto implica una asignación más eficiente de los recursos y la reducción de riesgos financieros.
- El análisis de los datos permitiría identificar patrones de demanda en diferentes ubicaciones y momentos del año. Esto ayudaría a los inversionistas a entender cuándo y dónde es más probable que sus propiedades sean alquiladas, lo que se traduce en una mayor ocupación y mayores ingresos.
- No todas las propiedades son igualmente adecuadas para el alquiler a corto plazo. Al utilizar datos de Airbnb, los inversionistas podrían identificar propiedades que tienen un alto potencial de rentabilidad. Esto incluye factores como la ubicación, las amenidades, el tamaño y las comodidades.
- Al realizar una inversión rentable, los inversionistas podrían contribuir al crecimiento económico y al turismo local. Esto puede ocasionar un aumento en la demanda de servicios locales y la generación de empleo.
- La aplicación de la ciencia de datos en este contexto puede ser de utilidad para aquellos sin un conocimiento técnico profundo sobre el análisis de datos y el modelado. Al simplificar y presentar la información de una manera accesible y fácil de

entender, este proyecto tiene el potencial de abrir la inversión en el sector de alojamiento compartido a una audiencia más amplia.

Esta propuesta se suma a la creciente necesidad de tomar decisiones basadas en datos y ofrece una solución para un sector económico en crecimiento. La solución propuesta, que consiste en procesar y analizar datos de Airbnb para identificar oportunidades de inversión y tendencias emergentes, tiene el potencial de ser altamente favorable y lucrativo para los inversionistas, así como de contribuir al desarrollo de la economía local.

1.3. Problema

Airbnb es una plataforma cuyo éxito a nivel mundial ha generado un creciente interés por invertir en propiedades para obtener beneficios de su uso como alojamiento de alquiler a corto plazo. Sin embargo, el uso de la plataforma a nivel global también representa un desafío: con miles de posibles ubicaciones para la inversión inmobiliaria y con la gran cantidad de posibles características que puede tener el inmueble, se convierte en una tarea difícil el determinar cuáles son las zonas más prometedoras y las comodidades más solicitadas por los huéspedes, especialmente si el inversionista no tiene conocimiento previo de la ciudad de interés y su cultura.

La necesidad de identificar con rapidez y sin esfuerzo las mejores zonas para invertir en una propiedad destinada a Airbnb, así como las características del inmueble, es un problema que requiere de una solución basada en datos. El análisis de la información mediante las técnicas de ciencia de datos permite identificar patrones sobre el comportamiento de las reservaciones que realizan los huéspedes. Con estos métodos, es posible generar modelos que permitan realizar predicciones para acelerar el proceso de selección y asistir a los inversionistas en la toma de decisiones para maximizar el rendimiento de sus inversiones.

1.4. Objetivos

1.4.1. Objetivo General:

Desarrollar modelos predictivos basados en árboles de decisión y redes neuronales mediante el análisis de datos de Airbnb para predecir la cantidad de noches reservadas por zona y estimar la utilidad anual del propietario a partir de las reservaciones estimadas.

1.4.2. Objetivos Específicos:

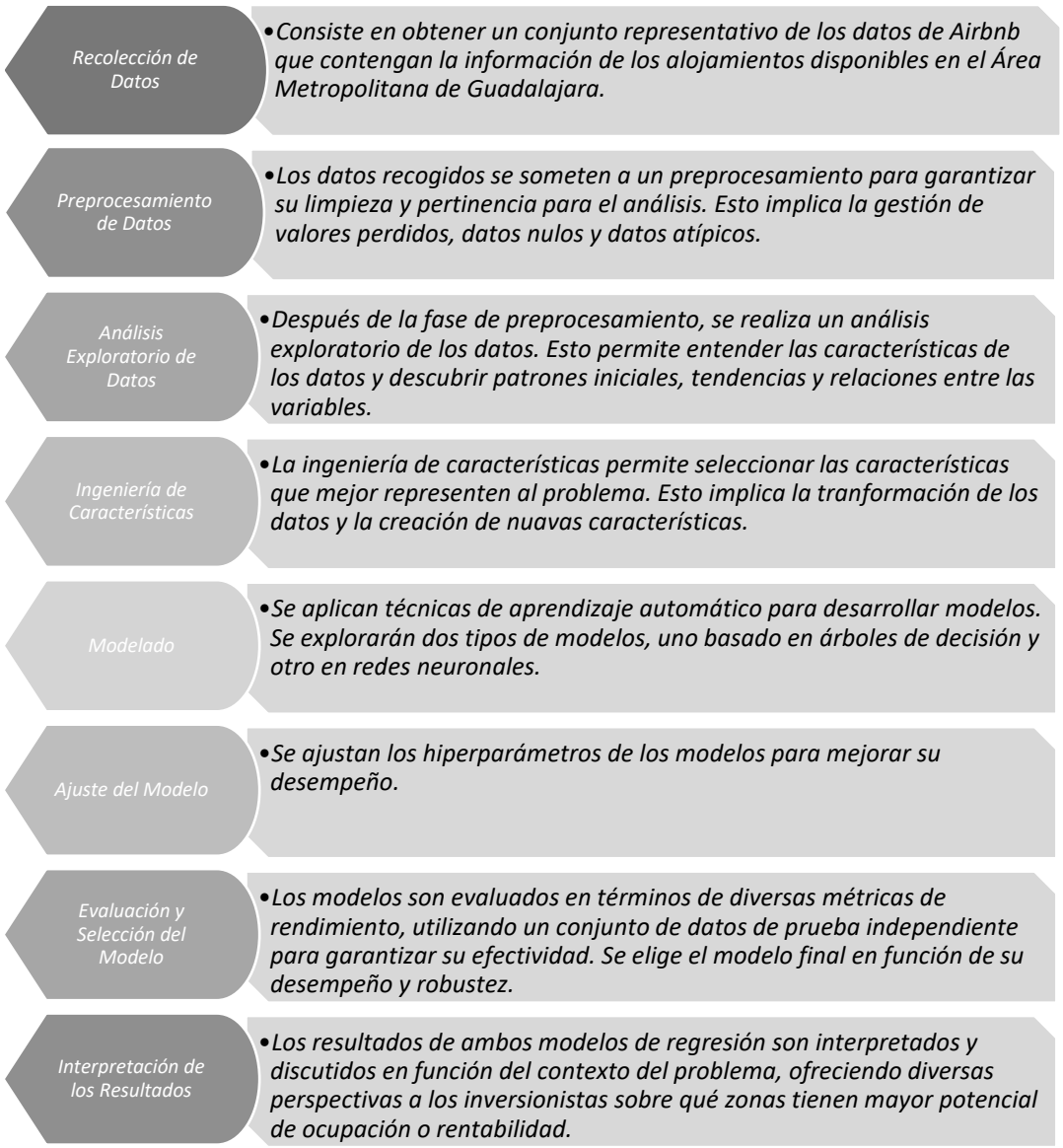
- Examinar bases de datos de alojamientos registrados en la plataforma de Airbnb de las diferentes zonas del área metropolitana de Guadalajara

mediante un análisis exploratorio de los datos para la visualización, preprocesamiento y limpieza de la información

- Seleccionar las características y amenidades de las propiedades listadas en Airbnb mediante la aplicación de ingeniería de características para transformar y determinar cuáles influyen más en la cantidad de noches reservadas y en la rentabilidad del propietario
- Desarrollar modelos de regresión utilizando árboles de decisión y redes neuronales para predecir la cantidad de noches reservadas por zonas en el área metropolitana de Guadalajara
- Desarrollar modelos de regresión utilizando árboles de decisión y redes neuronales para estimar la utilidad anual del propietario basado en la cantidad de noches reservadas
- Calcular diversas métricas para evaluar el desempeño de los diferentes modelos
- Examinar los resultados y comparar los modelos para seleccionar el de mejor desempeño
- Interpretar y explicar los resultados experimentales para resaltar los descubrimientos y concluir con el trabajo propuesto

2. METODOLOGÍA

En esta sección se presenta en detalle el desarrollo metodológico, el cual incluye los siguientes elementos:



2.1. Descripción de los datos

El conjunto de datos contiene información referente a los alojamientos que se ofrecen dentro de la zona metropolitana de Guadalajara mediante la plataforma de Airbnb. Los datos incluyen información de los alojamientos disponibles en los municipios de Guadalajara, Zapopan, Tonalá, Tlaquepaque, El Salto y Tlajomulco. Estos datos constan de 5126 observaciones con 48 características. Una muestra de 11 observaciones del conjunto de datos se ilustra en las Tablas 1, 2, 3 y 4.

Tabla 1. Muestra de 11 observaciones del conjunto de datos de los alojamientos Airbnb en la zona metropolitana de Guadalajara considerando las características de la 1 a la 12.

ID de listado	URL del anuncio	Título del anuncio	Tipo de propiedad	Tipo de arreglo	Longitud	Latitud	Ciudad	Habitaciones	Baños	Máximo de huéspedes	Calificación general
10014351	https://www.airbr	Habitacion Jr. Suite	boutique hotel	1	-103.3978	20.64976	M5	1	1	3	90
10057523	https://www.airbr	Excellent apartment	apartment	0	-103.419	20.63471	L5	1	1	4	90
10091650	https://www.airbr	Mini Depa	loft	0	-103.4362	20.66554	K6	1	1	2	100
10123233	https://www.airbr	Casa Azcarraga	house	1	-103.3726	20.6839	N7	6	3	12	90
10207548	https://www.airbr	Casa Luna, Tlaquepa	apartment	0	-103.3237	20.65303	Q6	1	1	2	100
10209243	https://www.airbr	** Quinta Tres Rein	house	0	-103.4718	20.82437	I9	5	5	16	90
10337699	https://www.airbr	Linda habitaciÃ³n pa	house	1	-103.368	20.62965	N4	1	2	1	100
10379603	https://www.airbr	Departamento - Gal	serviced apartmen	0	-103.3448	20.67629	P7	2	1	4	90
10468089	https://www.airbr	Nice Apartment in d	apartment	0	-103.3591	20.67757	O7	1	2	2	100
10732948	https://www.airbr	Departamento PLAI	apartment	0	-103.4266	20.673	L7	2	1	4	100
10771978	https://www.airbr	Historic Casa del M	villa	0	-103.3485	20.68336	O7	4	4	9	100

Tabla 2. Muestra de 11 observaciones del conjunto de datos de los alojamientos Airbnb en la zona metropolitana de Guadalajara considerando las características de la 13 a la 24.

Comodidades	Depósito de seguridad (DÓLARES)	Fecha de creación	Número de reseñas	Número de fotos	Aire acondicionado	Registro 24 horas	Jardín	Fumar	Acepta mascotas	Gimnasio y bienestar	Alberca
air conditioning	0	09/07/2016	18	12	1	0	0	0	0	0	1
24 hour check in,bedroom essentia		10/07/2016	29	19	0	1	0	0	0	0	0
24 hour check i	4658	11/07/2016	166	48	1	1	1	0	1	0	0
bedroom esser	124	09/07/2016	18	9	0	0	0	1	0	1	1
24 hour check i	94	22/07/2016	12	13	0	1	0	1	0	0	0
bbq,coffee,esse	98	08/07/2016	82	71	0	0	1	1	1	0	1
essentials,first aid kit,fridge,hange		14/02/2017	6	11	0	0	0	1	0	0	0
bedroom esser	0	09/07/2016	34	13	0	0	0	1	0	0	0
24 hour check i	221	22/04/2017	116	10	0	1	0	0	0	0	0
24 hour check i	0	08/07/2016	70	34	0	1	1	0	1	0	0
bedroom esser	500	08/07/2016	17	64	0	0	0	0	0	0	0

Tabla 3. Muestra de 11 observaciones del conjunto de datos de los alojamientos Airbnb en la zona metropolitana de Guadalajara considerando las características de la 25 a la 36.

Frente al agua	Estacionamiento	ID del anfitrión principal	Estrellas	Noches reservadas en 2022	Noches disponibles en 2022	Conteo de reservas en 2022	Tarifa diaria promedio en 2022	Duración media de la estancia en 2022	Noches reservadas en 2021	Noches disponibles en 2021	Conteo de reservas en 2021
0	1	51362138	4.56	0	223	0	0	0	12	292	3
0	1	49877447	4.54	8	8	0	60	60	272	378	2
0	1	32514989	4.89	390	526	110	768	58	628	730	160
0	1	51944583	4.29	78	161	9	91	48	123	364	5
0	1	52431922	4.92	0	75	0	0	0	53	365	1
0	1	52454591	4.68	210	510	74	3848	42	302	688	132
0	1	24295856	4.83	0	31	0	0	0	2	152	2
0	0	#N/D	#N/D	154	262	24	226	182	0	10	0
0	1	53943335	4.75	155	272	11	275	120	152	238	9
0	1	86966325	4.77	134	134	0	264	276	368	448	18
0	1	25556395	4.95	59	218	6	1547	37	105	147	0

Tabla 4. Muestra de 11 observaciones del conjunto de datos de los alojamientos Airbnb en la zona metropolitana de Guadalajara considerando las características de la 37 a la 48.

Tarifa diaria promedio en 2021	Duración media de la estancia en 2021	Noches reservadas en 2020	Noches disponibles en 2020	Conteo de reservas en 2020	Tarifa diaria promedio en 2020	Duración media de la estancia en 2020	Noches reservadas en 2019	Noches disponibles en 2019	Conteo de reservas en 2019	Tarifa diaria promedio en 2019	Duración media de la estancia en 2019
165	10	28	334	10	326	18	25	120	11	255	9
390	308	22	38	0	44	60	76	116	8	170	48
856	126	468	496	12	510	588	238	244	14	192	540
85	111	3	343	0	14	3	86	119	1	27	83
214	91	22	324	0	83	37	30	80	2	89	25
5126	52	286	666	112	4582	64	74	212	42	1342	12
28	2	16	184	1	46	33	12	122	0	10	22
0	0	34	285	11	93	8	19	96	5	106	14
261	147	141	261	6	215	132	83	120	13	139	23
512	340	234	322	0	116	296	16	28	2	58	12
1296	112	50	97	1	379	50	0	103	0	0	0

Las 48 características consideran información de cada alojamiento, tal como las amenidades, datos de identificación del alojamiento e información referente a las reservaciones y tarifas. Una descripción detallada de cada característica se muestra en la Tabla 5.

Tabla 5. Descripción de las características del conjunto de datos.

Característica	Descripción
ID de listado	Este es el código único de identificación asignado a cada propiedad listada en Airbnb.
URL del anuncio	Este es el enlace directo al anuncio de la propiedad en la plataforma de Airbnb.
Título del anuncio	Este es el encabezado promocional que describe el alojamiento en el sitio web de Airbnb.
Tipo de propiedad	Esta es una descripción concisa del alojamiento tal como aparece en Airbnb.
Comodidades	Este campo resume las facilidades y servicios ofrecidos en el alojamiento, tal como se presentan en la página de Airbnb.
Número de fotos	Este dato indica la cantidad total de imágenes que se muestran en el anuncio de la propiedad.
Fecha de creación	Esta es la fecha en la que se publicó inicialmente el anuncio en la plataforma de Airbnb.
Frente al agua	Este atributo señala si el alojamiento tiene vistas o acceso directo a un cuerpo de agua, como un lago o el mar.
ID del anfitrión	Código de identificación del anfitrión
Tipo de arreglo	Se refiere al tipo de alojamiento que se ofrece y se clasifica en cuatro categorías: casa completa, hotel, habitación privada y habitación compartida.
Ciudad	Indica la zona de la ciudad donde se encuentra la propiedad. Hay 78 categorías posibles para esta variable.
Mascotas permitidas	Denota si se permiten o no mascotas en el alojamiento.
Habitaciones	Indica el número de habitaciones disponibles en la propiedad.
Baños	Indican el número de baños disponibles en la propiedad.
Fumar	Indica si se permite o no fumar.
Depósito de seguridad (DÓLARES)	Señala la cantidad de dólares que se depositan como garantía por algún daño a la propiedad
Aire acondicionado	Señala si la propiedad está equipada con aire acondicionado.
Registro 24 horas	Informa si es posible registrase en la propiedad en cualquier momento del día o noche.
Jardín	Indica si la propiedad tiene jardín.
Gimnasio y bienestar	Muestra si la propiedad cuenta con opciones para hacer actividades físicas.

Alberca	Informa si la propiedad dispone de una alberca.
Estacionamiento	Señala si la propiedad cuenta con estacionamiento.
Máximo de huéspedes	Representa el número máximo de huéspedes que la propiedad puede alojar.
Calificación general	Es una puntuación general que refleja la valoración de los huéspedes que se han alojado en la propiedad.
Numero de reseñas	Representa la cantidad de comentarios que los huéspedes han dejado sobre la propiedad.
Longitud	Es la coordenada geográfica que indica la longitud de la ubicación de la propiedad.
Latitud	Es la coordenada geográfica que indica la latitud de la ubicación de la propiedad.
Estrellas	Representa la clasificación en estrellas de la propiedad, usualmente una medida de calidad y servicio.
Noches reservadas en 2019	Es el número de noches que fueron reservadas en la propiedad durante el año 2019.
Noches disponibles en 2019	Representa el número de noches que la propiedad estuvo disponible para ser reservada en el año 2019.
Tarifa diaria promedio en 2019	Es la tarifa diaria promedio para la propiedad durante el año 2019.
Duración media de la estancia en 2019	Es la duración promedio de las estadías en la propiedad durante el año 2019.
Conteo de reservas en 2019	Es el número total de reservas que se realizaron para la propiedad durante el año 2019.
Noches reservadas en 2020	Es el número de noches que fueron reservadas en la propiedad durante el año 2020.
Noches disponibles en 2020	Representa el número de noches que la propiedad estuvo disponible para ser reservada en el año 2020.
Tarifa diaria promedio en 2020	Es la tarifa diaria promedio para la propiedad durante el año 2020.
Duración media de la estancia en 2020	Es la duración promedio de las estadías en la propiedad durante el año 2020.
Conteo de reservas en 2020	Es el número total de reservas que se realizaron para la propiedad durante el año 2020.
Noches reservadas en 2021	Es el número de noches que fueron reservadas en la propiedad durante el año 2021.
Noches disponibles en 2021	Representa el número de noches que la propiedad estuvo disponible para ser reservada en el año 2021.
Tarifa diaria promedio en 2021	Es la tarifa diaria promedio para la propiedad durante el año 2021.
Duración media de la estancia en 2021	Es la duración promedio de las estadías en la propiedad durante el año 2021.
Conteo de reservas en 2021	Es el número total de reservas que se realizaron para la propiedad durante el año 2021.
Noches reservadas en 2022	Es el número de noches que fueron reservadas en la propiedad durante el año 2022.
Noches disponibles en 2022	Representa el número de noches que la propiedad estuvo disponible para ser reservada en el año 2022.
Tarifa diaria promedio en 2022	Es la tarifa diaria promedio para la propiedad durante el año 2022.
Duración media de la estancia en 2022	Es la duración promedio de las estadías en la propiedad durante el año 2022.
Conteo de reservas en 2022	Es el número total de reservas que se realizaron para la propiedad durante el año 2022.

De las Tablas 1, 2, 3 y 4 se puede observar que algunas características no aportan información relevante para el modelado debido a que no están relacionadas con el objetivo de predecir las noches reservadas y la utilidad anual. Tales características son “ID de listado”, “URL del anuncio”, “Título del anuncio”, “ID del anfitrión” y “Fecha de creación”. Estas características son etiquetas que sólo sirven para identificar a cada alojamiento. Por tal motivo, serán excluidas del análisis. Las características “Tipo de propiedad” y “Comodidades” también son excluidas del análisis debido a que contienen información redundante. La característica “Tipo de propiedad” contiene información similar a “Tipo de arreglo”, mientras que “Comodidades” incluye la misma información que “Estacionamiento”, “Aire acondicionado”, “Alberca”, y otras características referentes a las amenidades del alojamiento. En cuanto a la característica "Frente al agua", solo dos observaciones cumplen con esta condición. Por lo tanto, también se descarta esta variable para el estudio. Por otro lado, la característica “Número de fotos” presenta inconsistencias en sus valores ya que la mayoría de los alojamientos registran valores de hasta 469 fotos, lo cual es poco creíble y está fuera de lo normal. Por tal motivo, se pude considerar que esta característica presenta una gran cantidad de valores atípicos, por lo que también será desestimada en el análisis.

Las Tablas 6 y 7 muestran un resumen del conjunto de datos. El resumen considera las características generales de los datos, tal como los tipos de variables, la existencia de valores ausentes, así como los valores mínimos y máximos de cada variable. Este resumen permite realizar un análisis global de la información.

La descripción de cada columna de las Tablas 6 y 7 se detalla a continuación:

- Nombres: Contienen los nombres de las variables que se utilizarán en el proyecto.
- Tipo: Esta columna especifica el tipo de dato de las variables. Los tipos que se observan son int64 y float64, lo que indica que son variables numéricas.
- Valores faltantes: Esta columna destaca la cantidad de valores ausentes o faltantes en las variables. Un valor de cero indica que hay datos faltantes para esa variable específica. Este aspecto es crucial, ya que el tener datos ausentes puede afectar significativamente los resultados del análisis y, a menudo, requiere que se adopten técnicas de preprocesamiento adicionales para manejar estos valores ausentes de manera apropiada.
- Valores presentes: Esta columna es un complemento a los valores faltantes. Registra la cantidad de valores que sí están presentes en las diferentes variables del conjunto de datos.
- Valores únicos, Valor mínimo, Valor máximo: Estas tres columnas ayudan a entender la variabilidad en los datos al mostrar cuántos valores únicos existen, así como sus valores mínimos y máximos.

Tabla 6. Resumen del conjunto de datos considerando las características de la 1 a la 24.

Nombres	Tipo	Valores faltantes	Valores presentes	Valores únicos	Valor mínimo	Valor máximo
ID de listado unico	float64	0	5126	5109	14086	7.21911e+17
URL del anuncio	object	0	5126	5126	https://www.airbn…	https://www.airbnb.com/rooms 9998159
Título del anuncio	object	0	5126	5043	! HABITACIONES NU…	ðŸ‡²ðŸ‡½Casa Kali con desayu Barrio MÃ©xicoðŸ’™
Tipo de propiedad	object	0	5126	35	Santa Teresita."	villa
Tipo de arreglo	int64	0	5126	4	0	3
Longitud	float64	0	5126	4058	-103.66	-103.252
Latitud	float64	0	5126	3780	20.5459	20.9063
Ciudad	int32	0	5126	78	0	77
Habitaciones	int64	0	5126	15	0	21
Baños	int64	0	5126	17	0	21
Máximo de huéspedes	int64	0	5126	18	1	30
Calificación general	int64	0	5126	6	50	100
Mascotas permitidas	int64	0	5126	2	0	1
Comodidades	object	2	5124	4717	nan	nan
Depósito de seguridad (DÓLARES)	float64	2455	2671	193	0	5847
Fecha de creación	int64	0	5126	1636	41984	44829
Número de reseñas	int64	0	5126	264	3	458
Número de fotos	float64	12	5114	127	1	469
Aire acondicionado	int64	0	5126	2	0	1
Registro 24 horas	int64	0	5126	2	0	1
Jardín	int64	0	5126	2	0	1
Fumar	int64	0	5126	2	0	1
Gimnasio y bienestar	int64	0	5126	2	0	1
Alberca	int64	0	5126	2	0	1

En la Tabla 6 se aprecia que la característica “Depósito de seguridad (DÓLARES)” cuenta con 2455 valores ausentes, los cuales representan el 47.9% de información perdida. A pesar del alto porcentaje de valores ausentes, esta característica se mantiene como parte del estudio ya que, entre más alto del depósito, menor la probabilidad de reservación del alojamiento, lo que podría impactar directamente en los objetivos planteados para la estimación del número de reservaciones y la utilidad anual.

De forma similar, en la Tabla 7 se puede apreciar que las características relacionadas con años anteriores al 2022 presentan una cantidad significativa de valores ausentes. Tales características corresponden a las noches reservadas, noches disponibles, conteo de reservas, tarifa diaria promedio y duración media de la estancia. En particular, las características relacionadas con el año 2019 tienen un total de 3067 observaciones faltantes, lo que representa el 59.8% de información perdida. Por otra parte, las características relacionadas con el año 2020 y 2021 presentan 2267 y 1108 valores ausentes, respectivamente. Esta información perdida equivale al 44.2% para el año 2020 y 21.6% para el año 2021. Aunque las características relacionadas con el año 2021 no presentan un

porcentaje tan alto de valores ausentes como las del año 2019 y 2020, se decide excluir todas estas características del análisis ya que uno de los objetivos consiste en realizar la estimación anual de la utilidad y para ello se decide usar la información más actual y completa que se tiene, la cual corresponde a la del año 2022.

Finalmente, todas las observaciones que cuentan con información faltante son removidas del conjunto de datos.

Tabla 7. Resumen del conjunto de datos considerando las características de la 25 a la 48.

Nombres	Tipo	Valores faltantes	Valores presentes	Valores únicos	Valor mínimo	Valor máximo
Frente al agua	int64	0	5126	2	0	1
Estacionamiento	int64	0	5126	2	0	1
ID del anfitrión principal	object	0	5126	2274	#N/D	99908738
Estrellas	object	0	5126	125	#N/D	5
Noches reservadas en 2022	float64	1	5125	353	0	496
Noches disponibles en 2022	float64	1	5125	396	0	546
Conteo de reservas en 2022	float64	1	5125	118	0	202
Tarifa diaria promedio en 2022	float64	1	5125	1065	0	28520
Duración media de la estancia en 2022	float64	1	5125	319	0	1670
Noches reservadas en 2021	float64	1108	4018	424	0	722
Noches disponibles en 2021	float64	1108	4018	503	0	730
Conteo de reservas en 2021	float64	1108	4018	147	0	300
Tarifa diaria promedio en 2021	float64	1108	4018	1039	0	53872
Duración media de la estancia en 2021	float64	1108	4018	331	0	2216
Noches reservadas en 2020	float64	2267	2859	316	0	566
Noches disponibles en 2020	float64	2267	2859	495	0	732
Conteo de reservas en 2020	float64	2267	2859	103	0	182
Tarifa diaria promedio en 2020	float64	2267	2859	770	0	7890
Duración media de la estancia en 2020	float64	2267	2859	268	0	886
Noches reservadas en 2019	float64	3067	2059	159	0	244
Noches disponibles en 2019	float64	3067	2059	172	0	244
Conteo de reservas en 2019	float64	3067	2059	55	0	80
Tarifa diaria promedio en 2019	float64	3067	2059	440	0	5106
Duración media de la estancia en 2019	float64	3067	2059	135	0	1136

De este primer análisis se descartan algunas características y observaciones con base en la relevancia y pertinencia respecto a los objetivos generales, además de valores faltantes y valores atípicos, reduciendo el conjunto de datos de 48 a 24 variables y de 5126 a 2314 observaciones. La Tabla 8 muestra las 24 características que resultan de esta selección.

De la Tabla 8 se puede observar que las variables seleccionadas ofrecen una visión global de las características que contribuyen a la popularidad de un alojamiento, desde factores físicos, como la cantidad de habitaciones y baños, hasta servicios ofrecidos, como la posibilidad de traer mascotas o la disponibilidad de un jardín, lo cual abona directamente a los objetivos.

Tabla 8. Descripción de las características seleccionadas.

Características	Descripción
Tipo de arreglo	Se refiere al tipo de alojamiento que se ofrece y se clasifica en cuatro categorías: casa completa, hotel, habitación privada y habitación compartida.
Ciudad	Indica la zona de la ciudad donde se encuentra la propiedad. Hay 78 categorías posibles para esta variable.
Mascotas permitidas	Denota si se permiten o no mascotas en el alojamiento.
Habitaciones	Indica el número de habitaciones disponibles en la propiedad.
Baños	Indican el número de baños disponibles en la propiedad.
Fumar	Indica si se permite o no fumar.
Depósito de seguridad (DÓLARES)	Señala la cantidad de dólares que se depositan como garantía por algún daño a la propiedad
Aire acondicionado	Señala si la propiedad está equipada con aire acondicionado.
Registro 24 horas	Informa si es posible registrase en la propiedad en cualquier momento del día o noche.
Jardín	Indica si la propiedad tiene jardín.
Gimnasio y bienestar	Muestra si la propiedad cuenta con opciones para hacer actividades físicas.
Alberca	Informa si la propiedad dispone de una alberca.
Estacionamiento	Señala si la propiedad cuenta con estacionamiento.
Máximo de huéspedes	Representa el número máximo de huéspedes que la propiedad puede alojar.
Calificación general	Es una puntuación general que refleja la valoración de los huéspedes que se han alojado en la propiedad.
Numero de reseñas	Representa la cantidad de comentarios que los huéspedes han dejado sobre la propiedad.
Longitud	Es la coordenada geográfica que indica la longitud de la ubicación de la propiedad.
Latitud	Es la coordenada geográfica que indica la latitud de la ubicación de la propiedad.
Estrellas	Representa la clasificación en estrellas de la propiedad, usualmente una medida de calidad y servicio.
Noches reservadas en 2022	Es el número de noches que fueron reservadas en la propiedad durante el año 2022.
Noches disponibles en 2022	Representa el número de noches que la propiedad estuvo disponible para ser reservada en el año 2022.
Tarifa diaria promedio en 2022	Es la tarifa diaria promedio para la propiedad durante el año 2022.
Duración media de la estancia en 2022	Es la duración promedio de las estadías en la propiedad durante el año 2022.
Conteo de reservas en 2022	Es el número total de reservas que se realizaron para la propiedad durante el año 2022.

De las 24 características seleccionadas, se identifican dos variables de salida para atender los dos objetivos generales planteados para la estimación del número de reservaciones y la utilidad anual. Estas características son “Noches reservadas en 2022” y “Tarifa diaria promedio en 2022”, las cuales son seleccionadas ya que “Noches reservadas en 2022” indica el número de noches que los huéspedes eligieron pasar en un alojamiento específico durante el año 2022, lo que nos da una medida de su popularidad. Por otro lado, “Tarifa diaria promedio en 2022” es el cobro por la estadía en ese alojamiento durante el mismo año, lo cual nos da una medida de la rentabilidad. Sin embargo, “Tarifa diaria promedio en 2022” por sí sola no sirve para estimar la utilidad anual. Por tal motivo, esta variable de salida se sustituye con una nueva que resulta de multiplicar “Noches reservadas en 2022” por “Tarifa

diaria promedio en 2022”. Esta nueva variable podría considerarse como una medida de utilidad anual, ofreciendo un indicador del ingreso total que generó cada alojamiento durante el año 2022.

2.2. Análisis exploratorio

El conjunto de datos contiene características con valores categóricos y numéricos. Esto se puede apreciar en las Tablas 1, 2, 3, y 4. Un ejemplo de característica categórica es "Estacionamiento" ya que contine una etiqueta que informa si el alojamiento tiene o no tiene estacionamiento. Por otro lado, un ejemplo de una característica numérica es "Habitaciones" ya que contine un valor numérico que representa la cantidad de habitaciones con las que cuenta cada alojamiento. El listado de las características categóricas y la cantidad de categorías con las que cuenta cada una se muestra en la Tabla 9. Mientras que el listado de las características numéricas se puede observar en la Tabla 10, la cual incluye las variables de salida.

Tabla 9. Listado de características categóricas y la cantidad de categorías con las que cuenta cada una.

Variable	Número de categorías
Tipo de arreglo	4
Ciudad	78
Mascotas permitidas	2
Fumar	2
Aire acondicionado	2
Registro 24 horas	2
Jardín	2
Gimnasio y bienestar	2
Alberca	2
Estacionamiento	2

Algunas de las características categóricas del conjunto de datos se codificaron numéricamente ya que sus valores correspondían a texto. Por ejemplo, la disponibilidad de estacionamiento se representaba con las palabras "sí" o "no". Para poder utilizar estos datos en el modelado, fue necesario transformarlos en números, en este caso, 0 y 1.

Además, la variable "Ciudad" se codificó utilizando la librería *LabelEncoder*, lo que permitió asignar a cada observación un valor numérico manteniendo el orden de los datos.

Los datos sin procesar rara vez son útiles en su forma original, y el preprocesamiento puede implicar todo, desde la limpieza de datos y la imputación de valores faltantes, hasta la codificación de variables categóricas y la normalización de variables numéricas.

Es importante destacar que, en el contexto de las redes neuronales, se llevará a cabo una normalización de las variables, pero esta no será necesaria ni se realizará en el caso de los árboles de decisión.

Tabla 10. Listado de características numéricas.

Variable
Habitaciones
Baños
Longitud
Latitud
Máximo de huéspedes
Calificación general
Depósito de seguridad (DÓLARES)
Número de reseñas
Estrellas
Noches disponibles en 2022
Conteo de reservas en 2022
Noches reservadas en 2022
Tarifa diaria promedio en 2022
Duración media de la estancia en 2022

Dada la naturaleza intrínsecamente diferente de estos dos tipos de variables, se analizarán por separado, ya que cada tipo requiere un tratamiento distinto.

Una parte esencial del análisis exploratorio es la visualización los datos, en este caso, a través de histogramas. Las Figuras 1, 2 y 3 muestran los histogramas correspondientes a todas las variables categóricas.

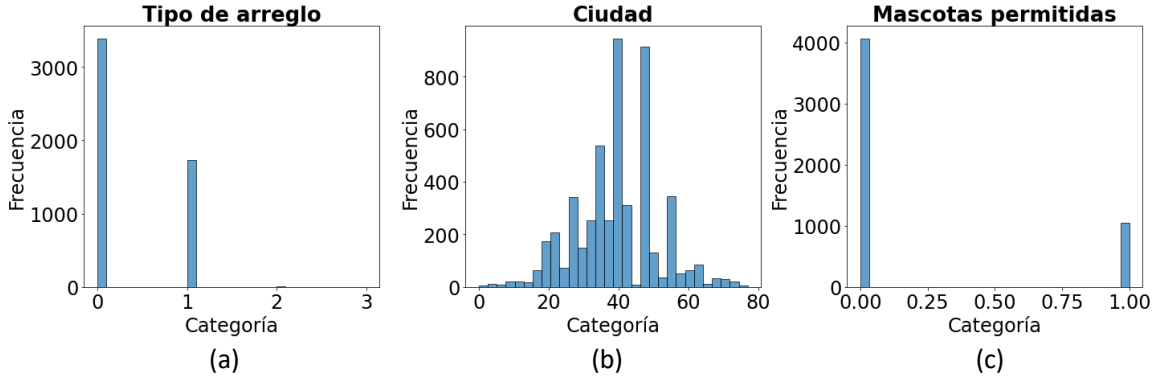


Figura 1. Histogramas de las características: (a) "Tipo de arreglo", (b) "Ciudad", y (c) "Mascotas permitidas".

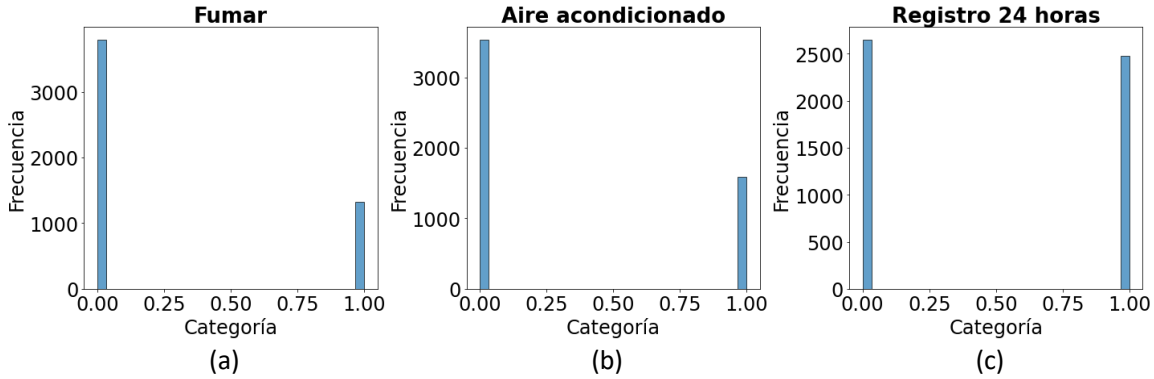


Figura 2. Histogramas de las características: (a) "Fumar", (b) "Aire acondicionado", y (c) "Registro 24 horas".

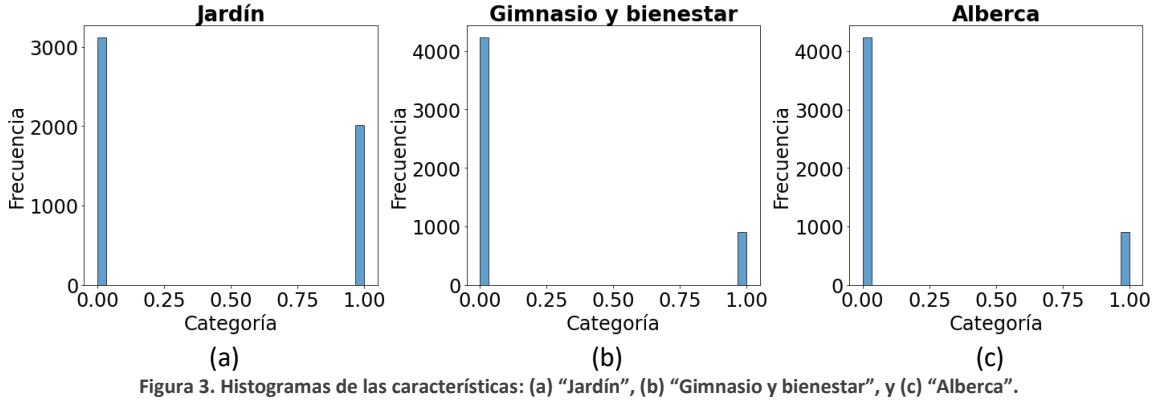


Figura 3. Histogramas de las características: (a) "Jardín", (b) "Gimnasio y bienestar", y (c) "Alberca".

Los histogramas son herramientas valiosas en este contexto. Algunos modelos de aprendizaje automático, como las redes neuronales y las Máquinas de Vectores de Soporte (SVM), son sensibles a los desequilibrios en los datos. En esencia, buscamos que haya un número similar de observaciones para cada categoría, lo que se puede apreciar fácilmente en un histograma: idealmente, las barras tendrían alturas similares. Por ejemplo, la Figura 2 (c), es como se esperaría ver una variable con un balance correcto. Aunque hay más alojamientos que no permiten el registro durante las 24 horas, la diferencia con los que sí lo permiten no es drástica. Sin embargo, en términos generales, la mayoría de las variables muestran un desequilibrio considerable. Este conocimiento es valioso, ya que podría indicar que un modelo basado en árboles de decisión podría ser más apropiado para estos datos, o podría sugerir la necesidad de aplicar técnicas de submuestreo.

A continuación, se aplica el mismo proceso de análisis para las variables numéricas, presentando sus gráficos de la Figura 4 a la 8.

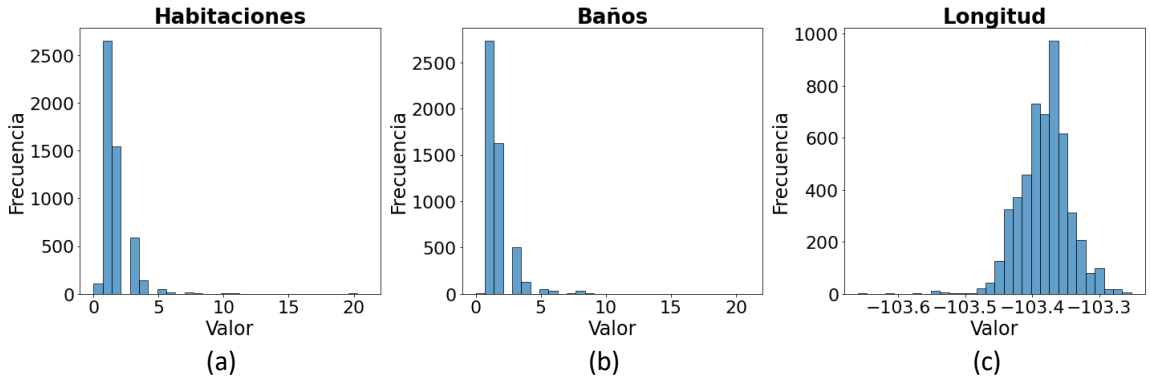


Figura 4. Histogramas de las características: (a) "Habitaciones", (b) "Baños", y (c) "Longitud".

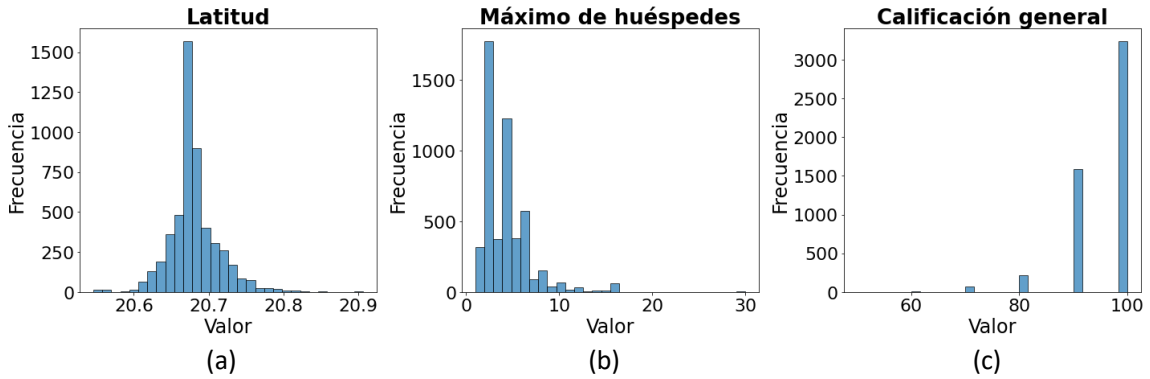


Figura 5. Histogramas de las características: (a) "Latitud", (b) "Máximo de huéspedes", y (c) "Calificación general".

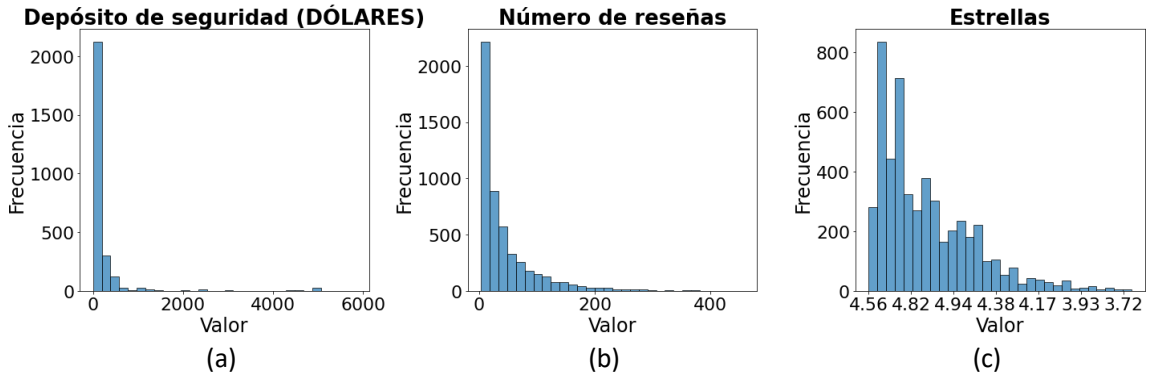


Figura 6. Histogramas de las características: (a) "Depósito de seguridad (DÓLARES)", (b) "Número de reseñas", y (c) "Estrellas".

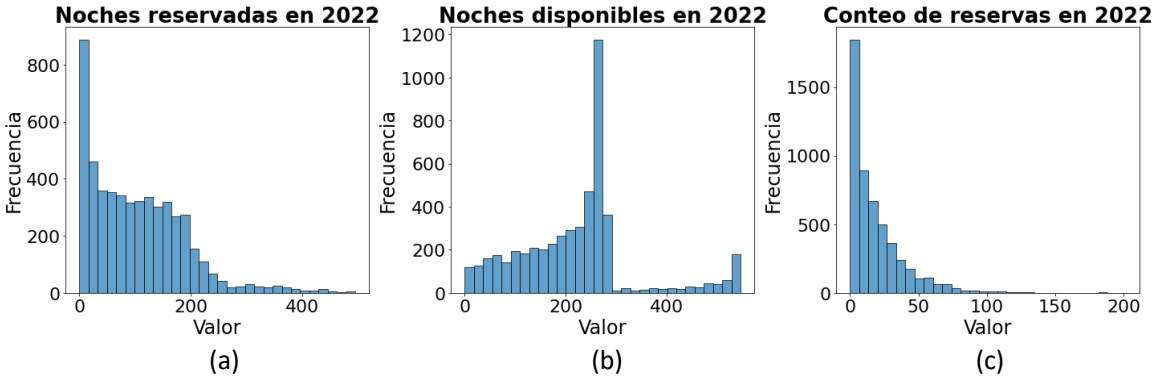


Figura 7. Histogramas de las características: (a) "Noches reservadas en 2022", (b) "Noches disponibles en 2022", y (c) "Conteo de reservas en 2022".

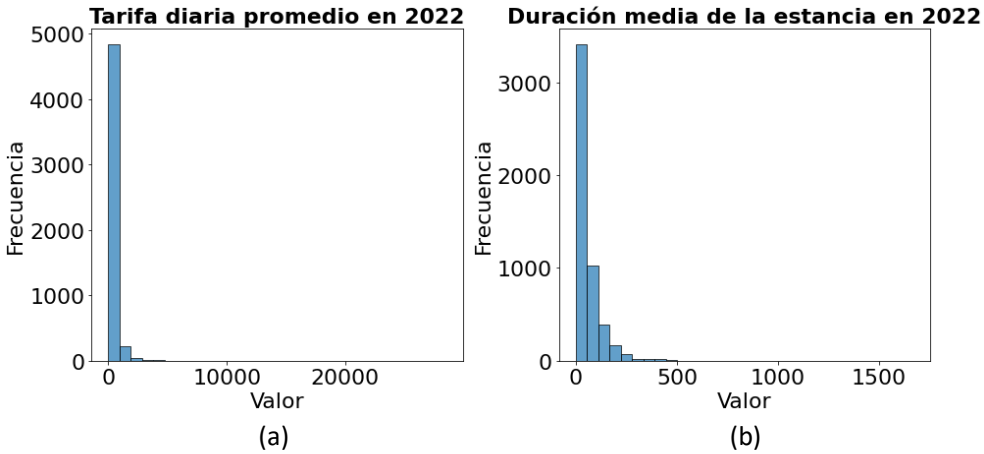


Figura 8. Histogramas de las características: (a) "Tarifa diaria promedio en 2022" y (b) "Duración media de la estancia en 2022".

Desde los histogramas, de las Figuras 4 a 8, se destaca que hay características predominantes dentro de las variables en relación con sus valores. Por ejemplo, en la Figura 8 (a), se observa que la mayoría de las observaciones se concentran alrededor de un valor numérico específico.

Además de los histogramas, otro método de análisis que resulta útil para las variables numéricas es la matriz de correlación. Esta herramienta busca identificar la variabilidad entre las variables del modelo. En caso de que dos o más variables presenten una alta correlación, podríamos optar por eliminar alguna de ellas para evitar la multicolinealidad. La Figura 9 muestra el mapa de calor con los resultados de este análisis.

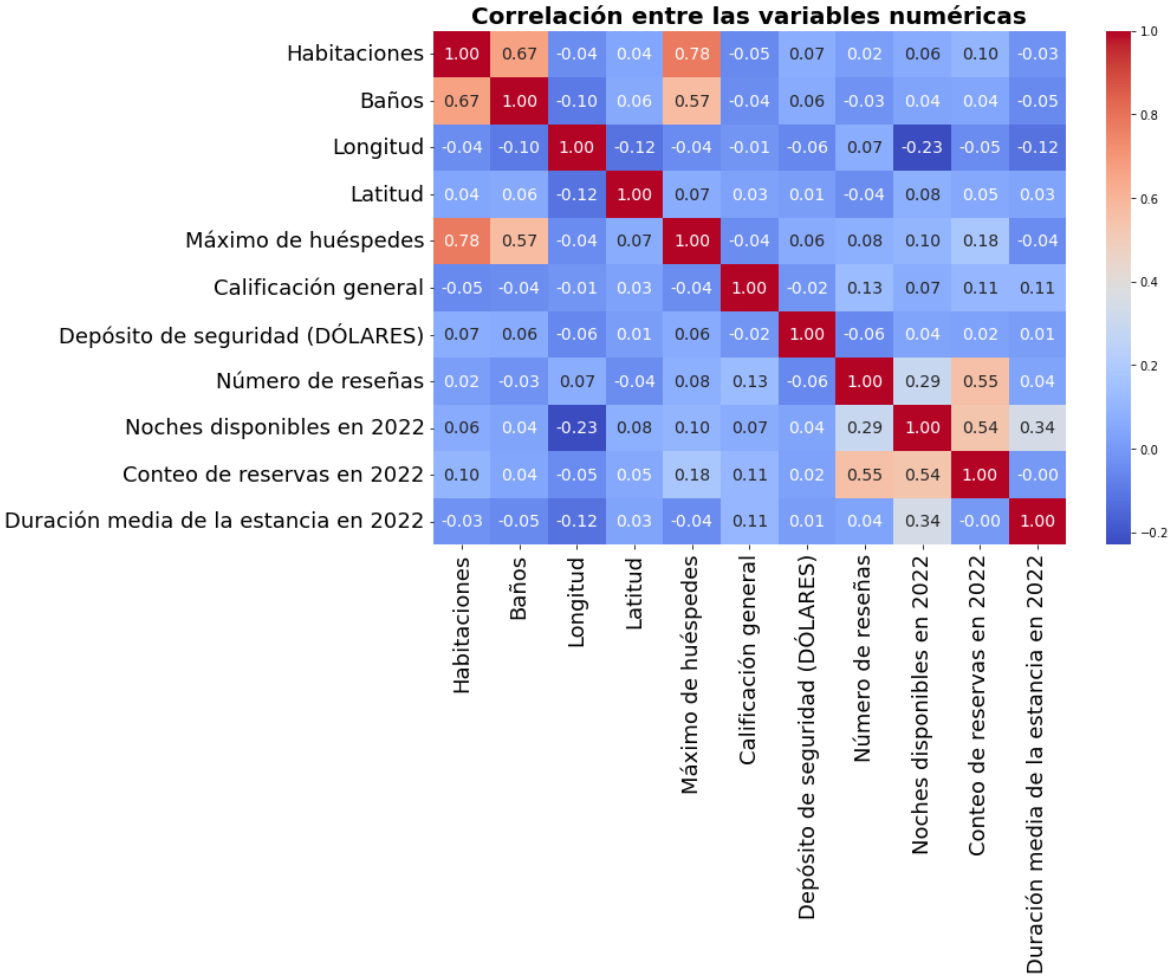


Figura 9. Mapa de calor de las características numéricas.

Los valores en el análisis de correlación oscilan únicamente entre -1 y 1. Cuanto más cercano sea el valor a cualquiera de estos extremos, mayor será la correlación. El mapa de calor proporciona un soporte visual para esta información, enfatizando las variables con mayor similitud mediante el uso de colores más intensos. De acuerdo con la Figura 9, las variables que destacan por su correlación son:

- Habitaciones y Baños, con un valor de 0.67
- Máximo de huéspedes y Habitaciones, con un valor de 0.78

No hay una regla estricta que determine cuándo es preferible eliminar una variable debido a su alta correlación con otra. No obstante, para este proyecto, se decidió que si la correlación entre dos variables superaba el valor de 0.8, se descartaría una de ellas. En nuestro caso, el valor más alto alcanzado fue 0.78, por lo que, por ahora, se decidirá mantener todas las variables.

Para analizar las relaciones entre variables categóricas, se utiliza un enfoque específico conocido como Análisis de Correspondencias Múltiples (ACM). Esta técnica estadística sirve para explorar y entender las asociaciones entre dichas variables. La Figura 10 muestra el mapa de correspondencias.

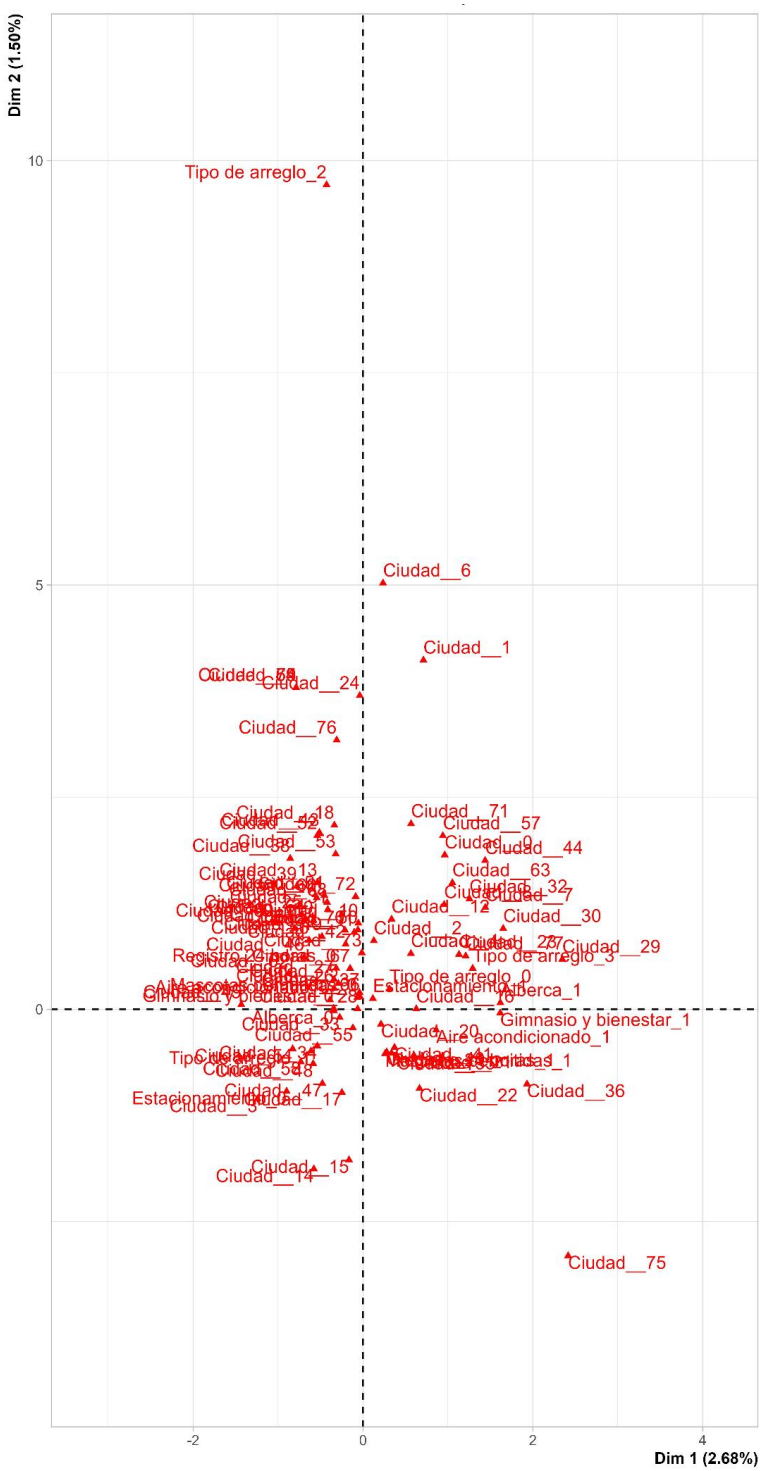


Figura 10. Mapa de correspondencias múltiples de las características categóricas.

La presencia de numerosas categorías dentro de una sola variable puede dificultar la claridad del mapa de correspondencias, aunque no necesariamente su interpretación. En general, cuanto más cercanas se encuentran las categorías entre sí en el mapa, más fuerte es la asociación entre las variables. En la Figura 10, tres categorías aparecen notoriamente separadas, lo cual podría indicar dos cosas: que estas categorías no tienen una asociación significativa con las demás o que cuentan con muy pocas observaciones. En este caso, el histograma la Figura 1 (a) y (b) confirman que tienen poca información, lo que las hace menos relevantes para el estudio.

Figura 11. Mapa de correspondencias múltiples de las características categóricas sin considerar la característica "Ciudad".

Para obtener una visión más detallada de la mayoría de las características, se decide eliminar la variable "Ciudad" del estudio. El mapa resultante, mostrado en la Figura 11, es más fácil de interpretar, ya que las relaciones entre las categorías se visualizan con mayor claridad.

En el eje inferior derecho de la Figura 11 se observa información interesante ya que dos categorías se encuentran agrupadas. Estas características agrupadas son “Alberca_1” y “Gimnasio y bienestar_1”. Esto podría sugerir que, cuando las personas buscan alojamiento, estas dos características tienden a seleccionarse conjuntamente, indicando que cuando se busca un alojamiento con alberca, también se busca que tenga gimnasio.

Hasta ahora, el análisis se ha centrado únicamente en las variables individuales. Sin embargo, también es crucial entender la interacción entre las variables predictivas y la variable objetivo. Como se mencionó anteriormente, planeamos desarrollar dos modelos: uno para predecir el número de noches reservadas y otro para predecir un componente de utilidad. Las Figura 12 a 15 muestran los gráficos de dispersión, considerando como variable de salida Noches reservadas en 2022.

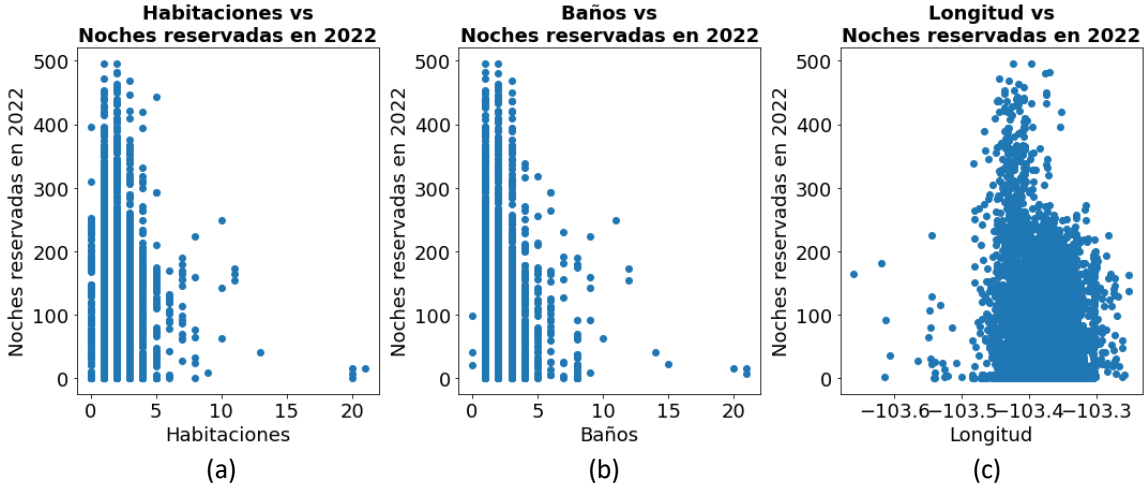


Figura 12. Gráfica de dispersión: (a) “Habitaciones” vs “Noches reservadas en 2022”, (b) “Baños” vs “Noches reservadas 2022”, y (c) “Longitud” vs “Noches reservadas en 2022”.

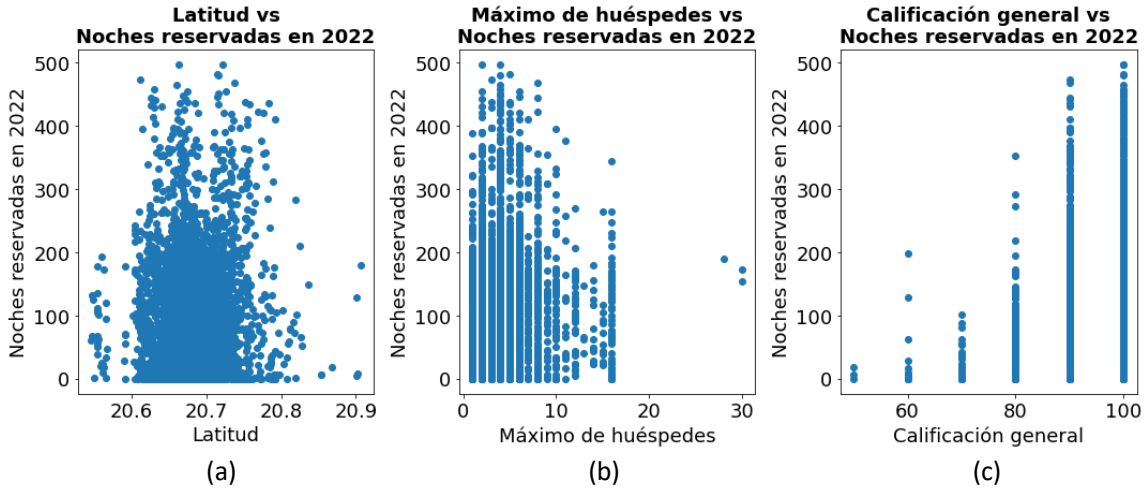


Figura 13. Gráfica de dispersión: (a) “Longitud” vs “Noches reservadas en 2022”, (b) “Máximo de huéspedes” vs “Noches reservadas 2022”, y (c) “Calificación general” vs “Noches reservadas en 2022”.

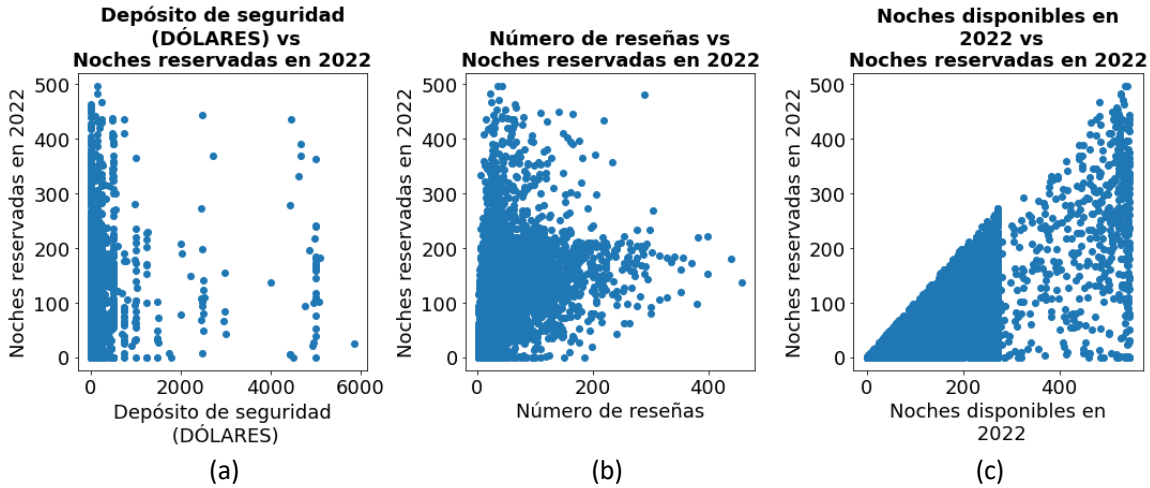


Figura 14. Gráfica de dispersión: (a) "Depósito de seguridad (DÓLARES)" vs "Noches reservadas en 2022", (b) "Número de reseñas" vs "Noches reservadas 2022", y (c) "Noches disponibles en 2022" vs "Noches reservadas en 2022".

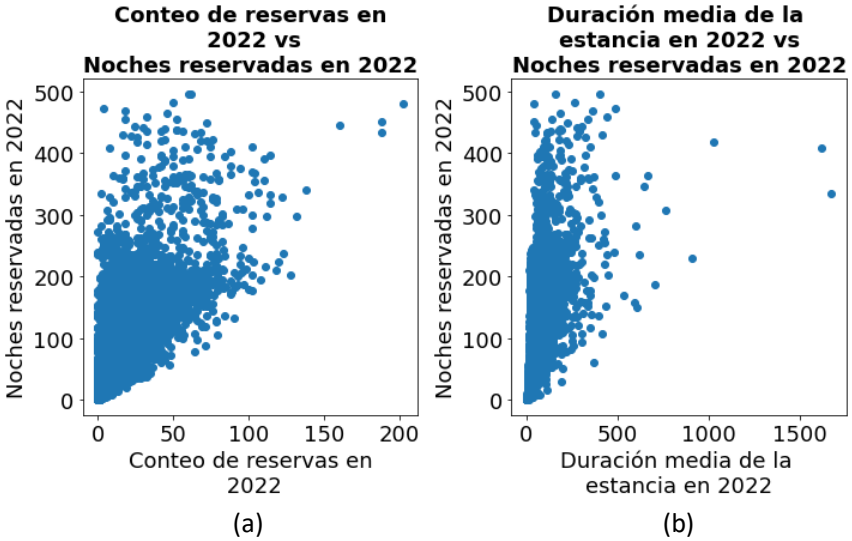


Figura 15. Gráfica de dispersión: (a) "Conteo de reservas en 2022" vs "Noches reservadas en 2022" y (b) "Duración media de la estancia en 2022" vs "Noches reservadas 2022".

Es importante observar que sólo se utilizan las variables numéricas en estos gráficos, ya que las variables categóricas requerirán un tipo diferente de visualización.

Para el caso de estudio de las variables, solo vemos un comportamiento lineal en "Calificación general", Figura 13 (c) y "Noches reservadas en 2022", Figura 14 (c). Las demás variables no parecen seguir ningún patrón perceptible a simple vista. Esto sugiere que, dependiendo del modelo de regresión seleccionado, podríamos necesitar realizar alguna transformación en los datos.

Para examinar la interacción de las variables categóricas, se selecciona un diagrama de caja para observar la relación de las noches reservadas en relación con las categorías de estas variables. Para mejorar la legibilidad de los gráficos, se separa la variable "Ciudad", la cual

cuenta con 78 categorías. La Figura 16 muestra el diagrama de caja para la característica "Ciudad", mientras que las Figuras 17, 18 y 19 muestran los diagramas para las demás variables categóricas.

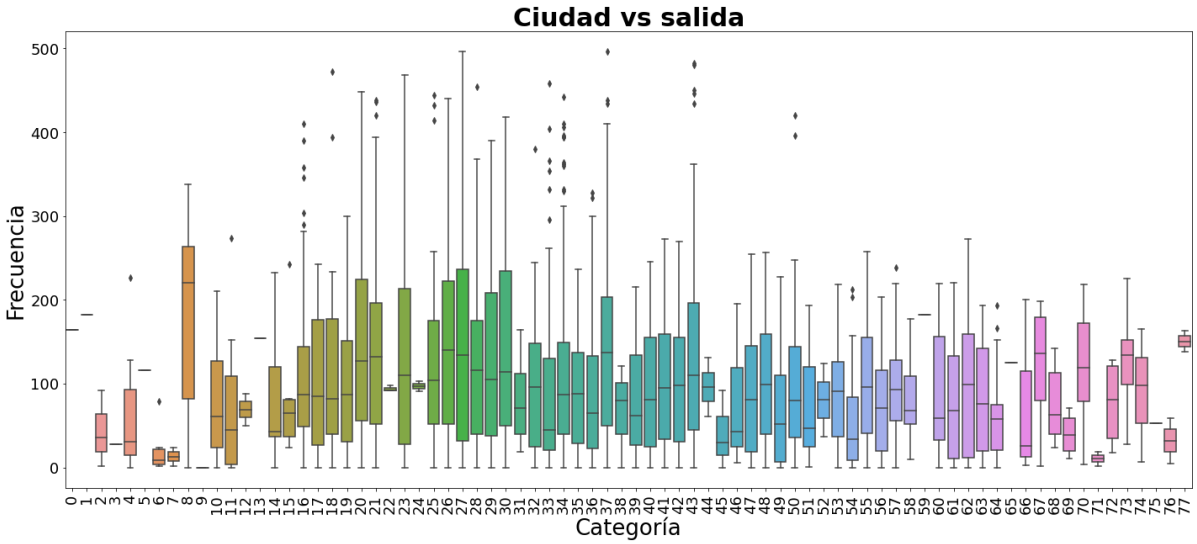


Figura 16. Diagrama de caja considerando la característica "Ciudad" vs "Noches reservadas en 2022".

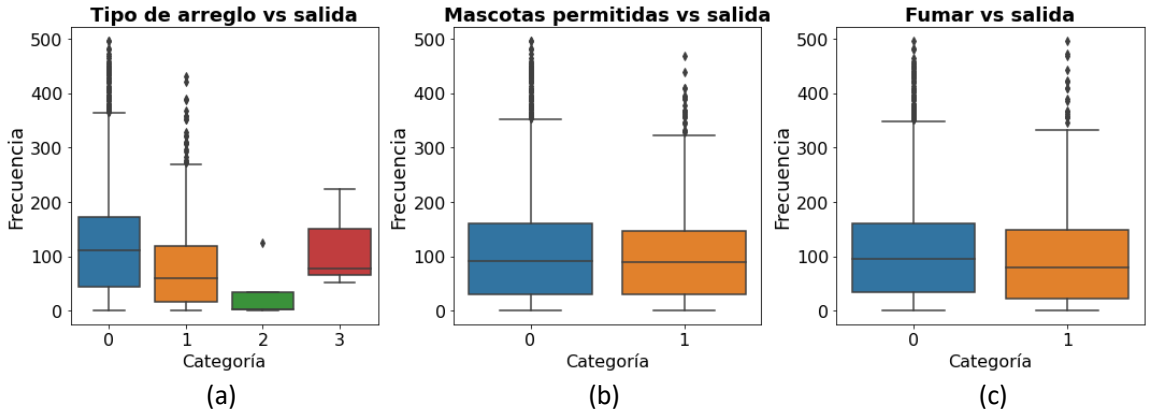


Figura 17. Diagrama de caja de la característica: (a) "Tipo de arreglo", (b) "Mascotas permitidas", y (c) "Fumar".

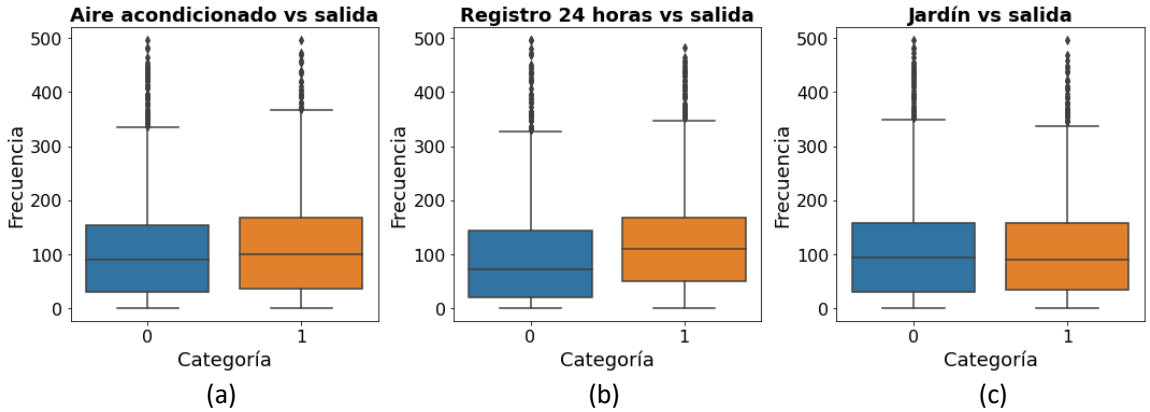


Figura 18. Diagrama de caja de la característica: (a) "Aire acondicionado", (b) "Registro 24 horas", y (c) "Jardín".

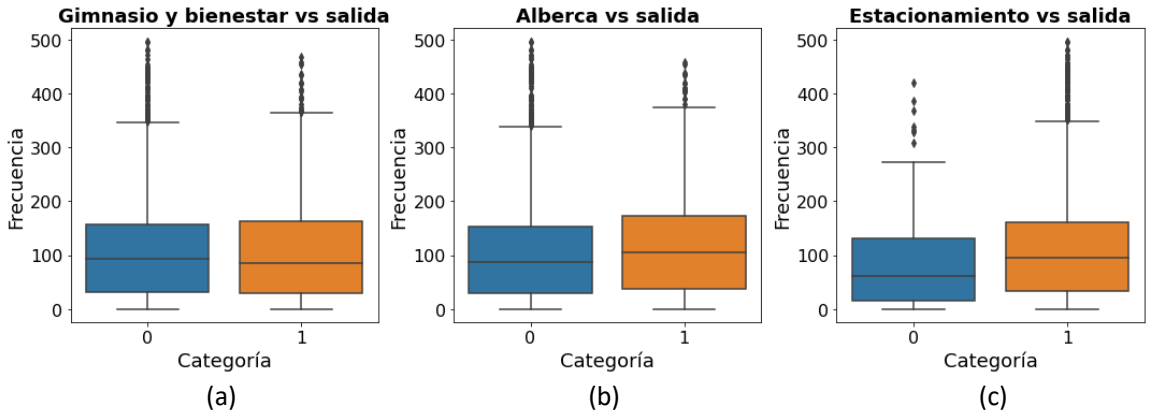


Figura 19. Diagrama de caja de la característica: (a) "Gimnasio y bienestar", (b) "Alberca", y (c) "Estacionamiento".

De la Figura 16 no se observa que una categoría en particular sea dominante, existe algunas con medianas más bajas respecto a las demás, como pueden ser las categorías, 6, 7 y 71, las cuales son particularmente bajas. Se puede observar también un amplio margen de datos atípicos, mayormente en las Figuras 17, 18 y 19. Sin embargo, no se descarta esa información para el análisis.

Ya que se tiene pensado analizar para el proyecto 2 tipos de variables de salida, también es necesario hacer el mismo estudio para la faltante, la cual se recuerda que es una combinación de Noches reservadas en 2022 multiplicada por Tarifa diaria promedio en 2022, en lo que se considerará un modelo para predecir la utilidad.

Las Figuras 20, 21, 22 y 23 muestran los diagramas de dispersión de las variables numéricas con respecto a la nueva variable de salida llamada "Utilidad 2022".

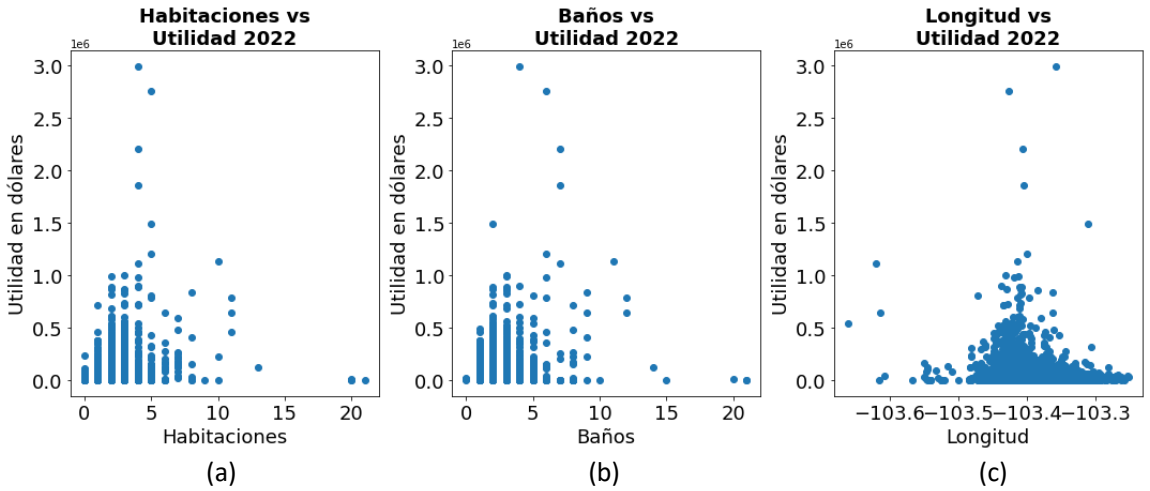


Figura 20. Gráfica de dispersión: (a) "Habitaciones" vs "Utilidad 2022", (b) "Baños" vs "Utilidad 2022", y (c) "Longitud" vs "Utilidad 2022".

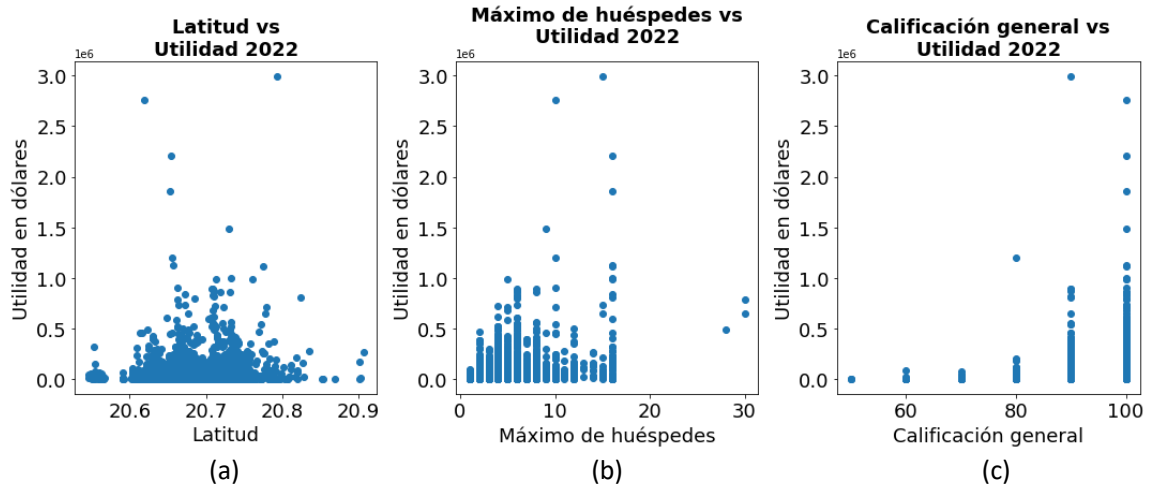


Figura 21. Gráfica de dispersión: (a) "Latitud" vs "Utilidad 2022", (b) "Máximo de huéspedes" vs "Utilidad 2022", y (c) "Calificación general" vs "Utilidad 2022".

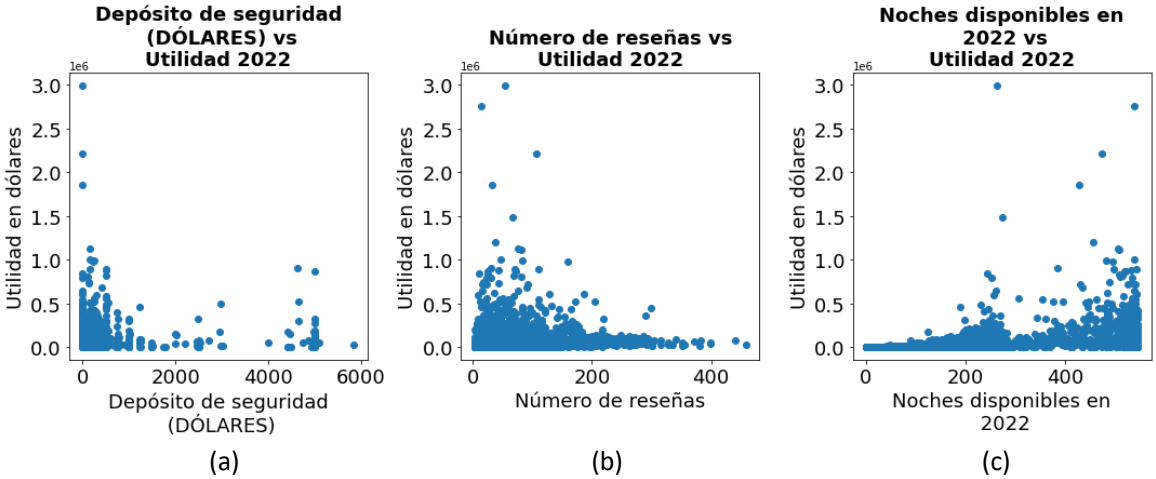


Figura 22. Gráfica de dispersión: (a) "Depósito de seguridad (DÓLARES)" vs "Utilidad 2022", (b) "Número de reseñas" vs "Utilidad 2022", y (c) "Noches disponibles" vs "Utilidad 2022".

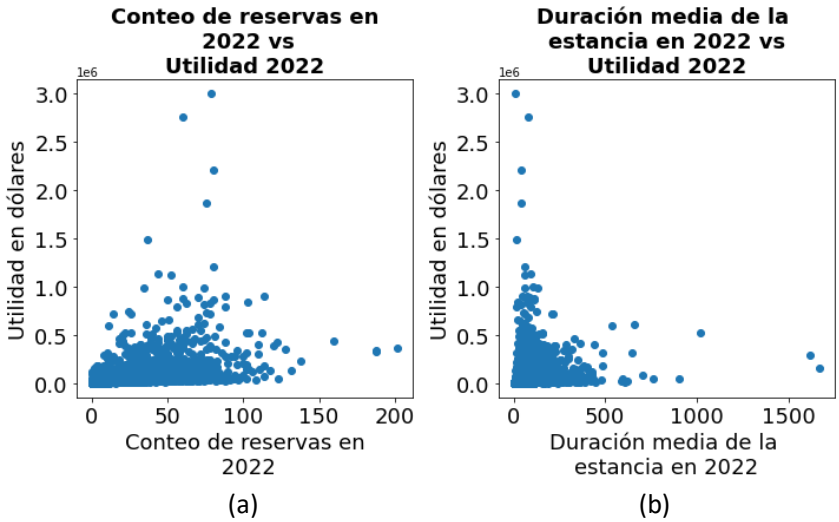


Figura 23. Gráfica de dispersión: (a) "Conteo de reservas en 2022" vs "Utilidad 2022" y (b) "Duración media de la estancia en 2022" vs "Utilidad 2022".

En este caso en particular, es aún más complicado encontrar algún componente lineal para las variables.

Se analizan los datos para la nueva variable de “Utilidad 2022”, la cual multiplica las noches reservadas por el precio promedio por noche. La Figura 24 muestra el diagrama de caja para la característica “Ciudad”, mientras que las Figuras 25, 26 y 27 muestran los diagramas para las demás variables categóricas.

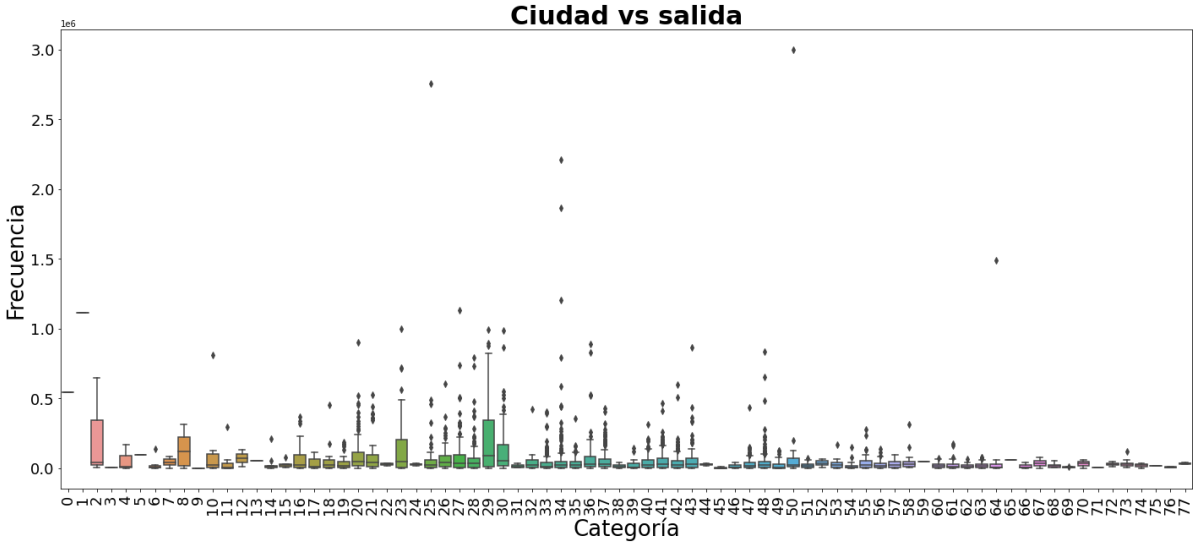


Figura 24. Diagrama de caja considerando la característica “Ciudad” vs “Utilidad 2022”.

La primera observación es que aumenta considerablemente el número de valores atípicos, con esto se pierde algo de interpretabilidad, ya que dificulta analizar las cajas en las gráficas, esto se debe a que se marca en mayor medida la diferencia entre la utilidad que genera cada uno de los alojamientos de Airbnb del estudio.

Para disminuir el efecto que se produce por la diferencia en los valores y tener un mayor detalle del análisis, se podría aplicar una transformación logarítmica, ya que se podría decir que cambia la escala de los valores y los vuelve más próximos entre sí. Una consideración antes de usar una transformación de este tipo es que los logaritmos son exclusivos para números positivos, esto descarta también cualquier valor en 0. Por lo que habría una disminución en el número de réplicas.

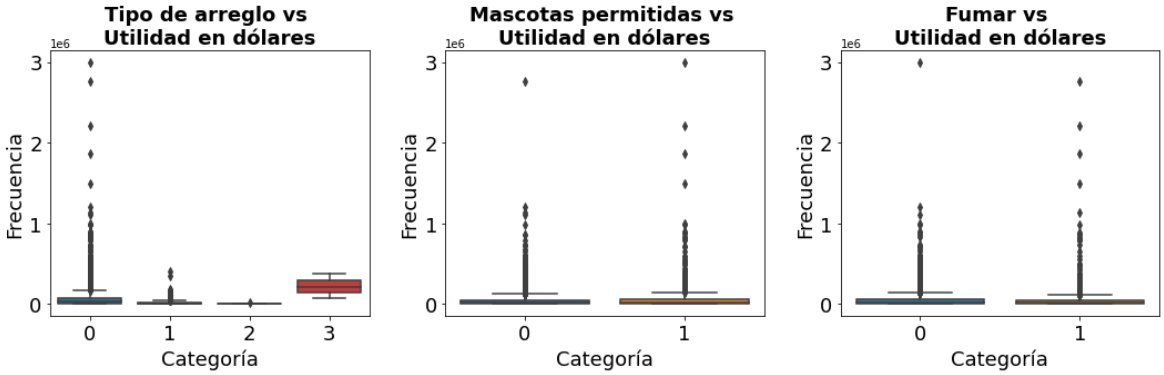


Figura 25. Diagrama de caja de la característica: (a) "Tipo de arreglo", (b) "Mascotas permitidas", y (c) "Fumar".

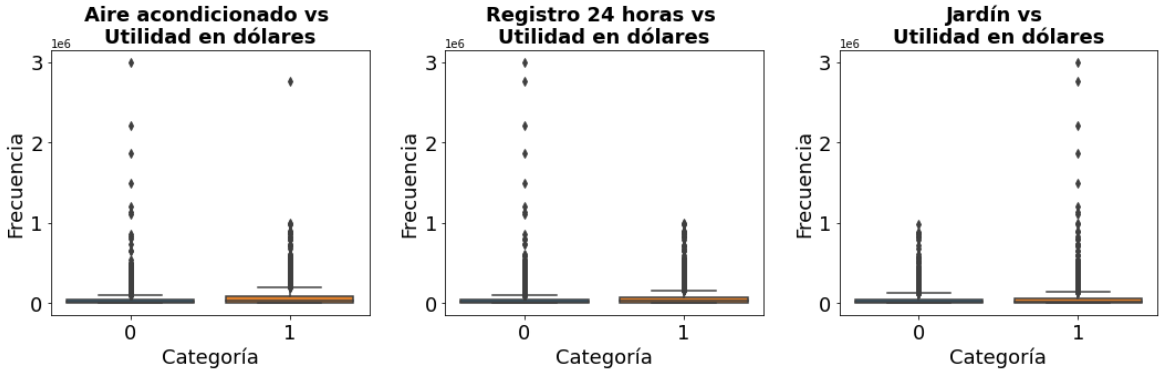


Figura 26. Diagrama de caja de la característica: (a) "Aire acondicionado", (b) "Registro 24 horas", y (c) "Jardín".

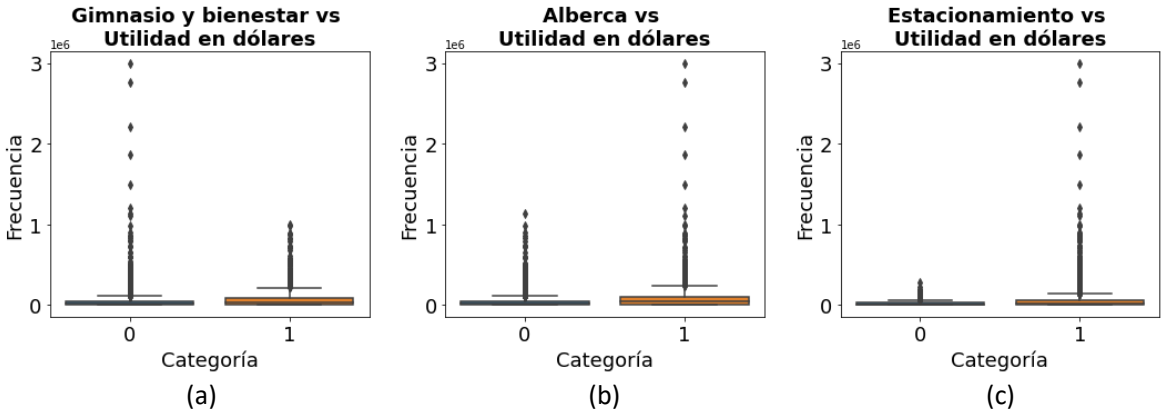


Figura 27. Diagrama de caja de la característica: (a) "Gimnasio y bienestar", (b) "Alberca", y (c) "Estacionamiento".

Las principales características de las variables del conjunto de datos incluyen:

- Variables categóricas desbalanceadas: Hay un desequilibrio notable en las categorías, lo cual puede sesgar los resultados. Las técnicas como el sobre muestreo de la clase

minoritaria o el submuestreo de la clase mayoritaria pueden ayudar a mitigar este efecto.

- Presencia de valores atípicos: Nuestro conjunto de datos contiene una gran cantidad de valores atípicos, lo que puede distorsionar los análisis. Estos pueden ser tratados mediante la eliminación, o mediante transformaciones como la raíz cuadrada o logarítmicas, que reducen su impacto.
- Falta de patrones lineales en variables numéricas: La ausencia de relaciones lineales en las variables numéricas puede complicar la aplicación de modelos lineales. Aquí, distintas transformaciones podrían ser exploradas para descubrir patrones subyacentes que puedan mejorar el rendimiento en una regresión.

Dadas estas características, una alternativa viable podría ser un modelo de regresión basado en árboles de decisión o uno basado en redes neuronales. Estos enfoques son flexibles y capaces de manejar tanto el desbalance en las clases como las relaciones no lineales, ofreciendo una solución robusta para el conjunto de datos.

2.3. Descripción de los modelos

Árboles de decisión

Los modelos de árboles y reglas son apreciados en la comunidad de modelización. Ofrecen una interpretación sencilla e implementación fácil y pueden manejar una variedad de predictores sin preprocesamiento. A diferencia de la regresión lineal, no requieren que se defina una relación específica entre predictores y respuesta. También gestionan bien los datos faltantes y seleccionan características de forma implícita, lo que es útil en muchos contextos. A pesar de sus fortalezas, estos modelos tienen ciertas debilidades. Pueden ser inestables, ya que pequeñas alteraciones en los datos pueden cambiar mucho la estructura del modelo, y su capacidad predictiva puede ser insatisfactoria si la relación entre predictores y respuesta no se ajusta a sus regiones rectangulares definidas. En tales casos, otros modelos podrían tener un error de predicción menor. Para combatir estos problemas, los investigadores desarrollaron métodos de conjunto que combinan muchos árboles o modelos basados en reglas en un solo modelo. Los conjuntos tienden a tener un rendimiento predictivo mucho mejor que los árboles individuales, y esto generalmente es cierto también para los modelos basados en reglas. Finalmente existen diferentes versiones de árboles de decisión (Max Kuhn, 2016).

Boosting

Boosting nació en los años 90, influenciado por teorías de aprendizaje. Originalmente se trataba de combinar clasificadores débiles (que son apenas mejores que una elección al azar) para crear un clasificador fuerte con un error de mala clasificación generalizada mejorado.

El algoritmo AdaBoost, desarrollado por Freund y Schapire, marcó un hito en la implementación práctica de esta teoría, transformando un aprendizaje débil en uno fuerte. AdaBoost mostró ser una herramienta de predicción poderosa, superando a muchos modelos individuales en diversas aplicaciones como la expresión genética, quimiometría e identificación de géneros musicales.

Investigadores, después del éxito de AdaBoost, conectaron este algoritmo con conceptos estadísticos como funciones de pérdida, modelado aditivo y regresión logística. La interpretación de boosting como un modelo aditivo que minimiza la pérdida exponencial permitió su generalización a problemas de clasificación y regresión. Friedman fue fundamental en esta transición, creando las "máquinas de boosting de gradiente" que abordaron tanto la clasificación como la regresión (Max Kuhn, 2016).

Gradient Boosting

Basándose en una función de pérdida (por ejemplo, error cuadrado para regresión) y un aprendizaje débil (como árboles de regresión), busca un modelo aditivo que minimice dicha función. Se inicia con la mejor estimación de la respuesta y se ajusta un modelo a los residuos para minimizar la pérdida, sumando este modelo al anterior, repitiendo este proceso varias veces. Cualquier técnica con parámetros de ajuste puede ser un aprendizaje débil, y los árboles son ideales para el *boosting*. Se pueden generar rápidamente, agregarse fácilmente y ajustarse para ser aprendices débiles limitando su profundidad. Para la regresión, los parámetros de ajuste principales en boosting son la profundidad del árbol y el número de iteraciones (Max Kuhn, 2016).

Redes neuronales

Una red neuronal es un sistema inspirado en las conexiones neuronales del cerebro humano. Para iniciar exploración en este campo, es esencial mencionar un modelo fundamental de neurona artificial: el perceptrón. Esta estructura fue concebida en la mitad del siglo XX por Frank Rosenblatt, basándose en investigaciones previas de figuras como Warren McCulloch y Walter Pitts. Aunque en la actualidad hay tendencia a utilizar modelos neuronales más avanzados, como la neurona sigmoide, es crucial entender los perceptrones para captar la evolución y la base de estas tecnologías. Un perceptrón se caracteriza por recibir múltiples entradas binarias y emitir una salida binaria. Rosenblatt propuso un sistema en el que cada entrada se multiplica por un peso, siendo la suma total de estos productos lo que determina la salida según un valor umbral específico (Nielsen, 2019).

Cuando nos enfocamos en problemas de regresión con redes neuronales, es común adoptar una función de error de suma de cuadrados de la forma:

$$E = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 \qquad (1)$$

Para simplificar, consideramos solo una salida, aunque extender esto a múltiples salidas es directo. Con base en esto, podemos representar la matriz Hessiana como:

$$H = \nabla\nabla E = \sum_{n=1}^{N} \nabla y_n \nabla y_n + \sum_{n=1}^{N} (y_n - t_n)\nabla\nabla y_n \tag{2}$$

Si la red ya ha sido entrenada con un conjunto de datos y sus salidas resultan estar muy cerca de los valores objetivos, el segundo término en la ecuación anterior será mínimo, por lo que a menudo se desestima. Sin embargo, una justificación más general para desestimar este término es la siguiente: recordemos que la función óptima que minimiza la pérdida de suma de cuadrados es el promedio condicional de los datos objetivo. La diferencia entre la salida y el valor objetivo es, por lo tanto, una variable aleatoria con media cero. Si suponemos que esta diferencia no está correlacionada con el término de segunda derivada, entonces este término en su totalidad promediará a cero en la suma.
Al descartar el segundo término, obtenemos la aproximación Levenberg–Marquardt, también conocida como aproximación del producto externo:

$$H \approx \sum_{n=1}^{N} b_n b_n^T \tag{3}$$

Donde la derivación y evaluación son directas gracias a la función de activación para las unidades de salida. Es crucial resaltar que esta aproximación es válida principalmente para redes que han sido entrenadas adecuadamente. Para una red en general, los términos de segunda derivada no serán despreciables.

Para la función de error de entropía cruzada en una red con funciones de activación de unidad de salida sigmoide logística, la aproximación correspondiente es:

$$H \approx \sum_{n=1}^{N} y_n(1 - y_n) b_n b_n^T \tag{4}$$

De manera similar, se puede obtener un resultado análogo para redes de múltiples clases con funciones de activación de unidad de salida *Softmax* (Bishop, 2006).

2.4. Descripción de las métricas

Coeficiente de determinación R^2

Una forma común de resumir cuán bien se ajusta un modelo de regresión lineal a los datos es a través del coeficiente de determinación, o R^2. Esto se puede calcular como el cuadrado

de la correlación entre los valores observados y y los valores predichos $\hat{y}$. Alternativamente, también se puede calcular como:

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2} \quad (5)$$

donde las sumas son sobre todas las observaciones. Por lo tanto, refleja la proporción de variación en la variable de pronóstico que se explica mediante el modelo de regresión.

En la regresión lineal simple, el valor de R^2 también es igual al cuadrado de la correlación entre y y x (siempre que se haya incluido una intersección). Si las predicciones están cerca de los valores reales, esperaríamos que R^2 esté cerca de 1. Por otro lado, si las predicciones no están relacionadas con los valores reales, entonces R^2=0 (nuevamente, suponiendo que hay una intersección). En todos los casos, R^2 se encuentra entre 0 y 1.

El valor R^2 se utiliza con frecuencia, aunque a menudo de manera incorrecta, en la predicción. El valor de R^2 nunca disminuirá al agregar un predictor adicional al modelo, y esto puede llevar a un sobreajuste. No hay reglas establecidas para lo que es un buen valor de R^2, y los valores típicos de R^2 dependen del tipo de datos utilizados. Validar el rendimiento de pronóstico de un modelo en los datos de prueba es mucho mejor que medir el valor R^2 en los datos de entrenamiento (Hyndman, 2021).

RMSE

El error cuadrático medio o *Root Mean Squared Error* (RMSE), es similar al error estándar de estimación en la regresión lineal, excepto que se calcula sobre los datos de validación en lugar de sobre los datos de entrenamiento. Tiene las mismas unidades que la variable de resultado.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad (6)$$

RAE

El erro absoluto relativo o *Relative Absolute Error* (RAE) es otra métrica muy utilizada (Cichosz, 2015). Para evaluar adecuadamente la utilidad práctica de un modelo en una aplicación específica, no solo debemos considerar las medidas de rendimiento basadas en residuales, sino también incorporar una descripción de la distribución de la función objetivo, tal como se observa en el conjunto de datos utilizado. Esta incorporación ayuda a determinar si los residuales del modelo son comparables con los valores o la variabilidad inherente de la función objetivo. Un indicador esencial en este contexto es el RAE, que muestra la relación entre el residual medio y la desviación media de la función objetivo respecto a su media. Este último valor puede verse como el error absoluto medio que tendría un modelo de predicción

basado simplemente en el valor medio. Idealmente, el RAE debería ser considerablemente menor a 1 para modelos razonables y lo más cercano a 0 posible.

$$\mathrm{RAE} = \frac{\sum_{x \epsilon S} |f(x) - h(x)|}{\sum_{x \epsilon S} |f(x) - m_S(f)|} \tag{7}$$

MAPE

MAPE (Mean Absolute Percentage Error). Esta medida proporciona un porcentaje que indica cuánto se desvían las predicciones (en promedio) de los valores reales (Shmueli, 2017).

$$\mathrm{MAPE} = 100 \times \frac{1}{n} \sum_{i=1}^{n} |e_1 / y_i| \tag{8}$$

Coeficiente de correlación de Pearson

La covarianza es sensible a los cambios de unidades en los datos; por ejemplo, la covarianza será más baja si la altura de un árbol se mide en metros en comparación con centímetros. Para corregir esto, la covarianza se normaliza dividiéndola por el producto de las desviaciones estándar de ambas variables, x e y, obteniendo así la correlación de Pearson:

$$r = \frac{\mathrm{cov}(x, y)}{S_x S_y} \tag{9}$$

La correlación de Pearson mide la fuerza y dirección de la relación lineal entre dos variables, con valores posibles desde -1 hasta +1. Un valor de -1 implica una relación lineal negativa perfecta, 0 significa que no hay relación lineal, y +1 representa una relación lineal positiva perfecta (Michael H. Herzog, 2019).

2.5. *Descripción de los experimentos / simulaciones*

En los experimentos, se realizarán dos análisis utilizando las variables “Noches reservadas en 2022” y “Tarifa diaria promedio en 2022” de formas diferentes. En el primer análisis se usará únicamente “Noches reservadas en 2022” para predecir qué tan atractivo podría ser un alojamiento basándonos en cuántas noches se reservó. En el segundo, se creará una nueva variable combinando “Noches reservadas en 2022” y “Tarifa diaria promedio en 2022”. Esta nueva variable podría considerarse como una medida de utilidad, ya que multiplica las noches que se reservaron por la tarifa promedio, ofreciendo un indicador del ingreso total que generó cada alojamiento.

Para el adecuado procesamiento y análisis de la base de datos, se llevaron a cabo varias transformaciones y ajustes:

Codificación de la variable “Ciudad”

La variable “Ciudad”, representada por una combinación de números y letras, no es directamente utilizable en muchos modelos de aprendizaje automático. Por lo tanto, fue necesario convertirla en un formato más manejable. Se utilizó la función *LabelEncoder* de la biblioteca *sklearn*, que permite codificar automáticamente etiquetas categóricas en valores numéricos. La Tabla 11 ilustra el resultado después de aplicar esta codificación.

Tabla 11. Ejemplo de codificación numérica de la variable Ciudad mediante *LabelEncoder*.

Ciudad
20
41
10
48
28
48

Limpieza de datos

Se optó por eliminar las filas que contienen datos faltantes. Además, en lo que respecta a las variables de salida, se eliminaron aquellas que registran un valor de cero. Dado que uno de los análisis se basará en una función logarítmica, esta no puede procesar adecuadamente valores cero. Aunque esta decisión reduce el número total de observaciones, se consideró necesaria. No se cuenta con claridad sobre la razón de la falta de información en ciertos registros, y trabajar con datos completos asegura una interpretación más precisa del comportamiento subyacente.

Análisis de la regresión

Una vez realizadas las limpiezas previamente mencionadas, el tamaño del conjunto de datos se redujo a 2,314 observaciones. A pesar de que representa una pérdida significativa de datos, lo que queda es información de calidad que puede reflejar de manera más fidedigna las tendencias reales.

Preprocesamiento de datos

Antes de dividir los datos en conjuntos de entrenamiento y prueba, fue necesario transformar las variables categóricas utilizando la técnica de *one hot encoding*. Esta técnica descompone cada categoría en diferentes columnas y utiliza valores binarios para indicar la presencia o ausencia de la categoría correspondiente en cada observación. Ya que se usaron diferentes modelos para poder hacer comparaciones entre las métricas, para los modelos

basados en redes neuronales es necesario escalar los datos antes de entrenarlos. Los modelos basados en árboles no son particularmente sensibles a la variabilidad en magnitudes entre características.

Gradient Boosting Regressor

El *Gradient Boosting* es una técnica de aprendizaje automático que construye un modelo predictivo en forma de una serie de árboles de decisión. Lo hace de manera secuencial, donde cada nuevo árbol intenta corregir los errores cometidos por los árboles anteriores. Y cuenta con los siguientes hiperparámetros:

- Numero de estimadores: Se refiere al número de árboles que se añadirán al modelo. Es importante mencionar que añadir más árboles puede incrementar la precisión, pero también el tiempo de cómputo.
- Tasa de aprendizaje: Es un factor que se multiplica por la contribución de cada árbol. Una tasa de aprendizaje más baja significa que cada árbol influirá menos en la predicción final, lo que puede requerir más árboles, pero podría resultar en un modelo más robusto.
- Estado aleatorio: Garantiza la reproducibilidad de los resultados. Fijar un valor garantiza que los resultados serán consistentes en diferentes ejecuciones, lo cual es crucial para comparar modelos o para colaboraciones.

La Tabla 12 muestra los valores que se utilizaron como hiperparámetros del *Gradient Boosting*.

Tabla 12. Hiperparámetros seleccionados para la regresión por Gradient Boosting.

Número de estimadores	Tasa de aprendizaje	Estado aleatorio
1000	0.1	42

Validación cruzada

La validación cruzada es una técnica para evaluar el desempeño de modelos de aprendizaje automático. Divide el conjunto de datos en un número especificado de partes (denominadas pliegues), entrena el modelo en todas las partes excepto una, y luego evalúa el modelo en la parte restante. Este proceso se repite para cada pliegue, asegurando que cada parte se utilice tanto para el entrenamiento como para la prueba.

- Número de pliegues: Es el número de veces que el modelo será entrenado y evaluado, usando una parte diferente como conjunto de prueba cada vez.

- Métrica: El coeficiente de determinación R^2 es una métrica que indica qué tan bien las predicciones del modelo se ajustan a los valores reales. Un valor de 1 indica un ajuste perfecto, mientras que un valor de 0 indica que el modelo no realiza predicciones mejores que simplemente tomar el promedio de los valores objetivos.
- Núcleos del procesador: Utilizar todos los núcleos del procesador permite que el proceso de validación cruzada se lleve a cabo de manera paralela, haciendo que la evaluación sea más rápida.

Los hiperparámetros para la validación cruzada se muestran en la Tabla 13.

Tabla 13. Hiperparámetros seleccionados para la validación cruzada por Gradient Boosting.

Número de pliegues	Métrica	Núcleos del procesador
8	R^2	-1

Redes neuronales

Las redes neuronales son modelos computacionales inspirados en la forma en que las neuronas biológicas en el cerebro humano se comunican y procesan la información. Están compuestas por capas de nodos, llamados "neuronas", que transforman los datos de entrada en una salida. Las redes neuronales son especialmente útiles para tareas donde las relaciones entre las variables no son lineales. A continuación, se describen algunos de los hiperparámetros más relevantes para las redes neuronales:

- Número de capas y neuronas: Determina la profundidad y anchura de la red. Una red con más capas y/o neuronas puede representar funciones más complejas, pero también es más susceptible al sobreajuste y requiere más datos para entrenar.
- Función de activación: Define la transformación que se aplica a la salida de cada neurona. Las funciones comunes incluyen *ReLU*, *sigmoid*, *tanh*, y *LeakyReLU*.
- Optimizador: Es el algoritmo utilizado para actualizar los pesos de la red durante el entrenamiento. Ejemplos populares son *SGD*, *Adam*, y *RMSprop*.
- Tasa de aprendizaje: Controla la rapidez con la que el modelo se adapta a los datos. Una tasa muy alta puede hacer que el entrenamiento no converja, mientras que una tasa muy baja puede hacer que el entrenamiento sea demasiado lento.
- Tamaño de lote: Es el número de muestras que se utilizan para calcular una actualización de los pesos en cada iteración del entrenamiento.
- Épocas: Representa el número de veces que el algoritmo de aprendizaje trabajará en todo el conjunto de entrenamiento.

- Regularización: Ayuda a evitar el sobreajuste añadiendo una penalización a los pesos de la red. Las técnicas comunes incluyen L1 y L2.
- Descarte (*dropout*): Es una técnica de regularización que consiste en apagar aleatoriamente un porcentaje de neuronas durante el entrenamiento.

La Tabla 14 describe las características de la red neuronal utilizada.

Tabla 14. Estructura y características detalladas de la red neuronal.

Características	Descripción
Número total de capas	3 capas, 2 capas densas y una para normalización del lote
Capa de entrada	64 neuronas
Capa de normalización	No tiene neuronas, es una capa de procesos (BatchNormalization)
Capa intermedia	LeakyReLU, capa de activación, no tiene neuronas propias
Capa de salida	1 neurona. Función de activación linear
Número total de neuronas	65
Funciones de activación	ReLU para la capa de entada, linear para la capa de salida y LeakyReLU después de la normalización
Optimizador	Adam
Función de pérdida	Error cuadrático medio (MSE)
Número de épocas	100
Tamaño de lote	32
Datos de validación	Sí
Verbosidad	1, para observar el progreso del entrenamiento

3. RESULTADOS Y DISCUSIÓN

3.1. Resultados y discusión

Considerando primero como variable de salida las noches reservadas en 2022 multiplicada por la Tarifa diaria promedio en 2022, se compara una regresión hecha por un modelo de redes neuronales y uno de los árboles de decisión por *gradient boosting*, se comparan las mismas métricas para ambos modelos. Los resultados se muestran en la Tabla 15.

Tabla 15. Análisis comparativo de métricas entre los modelos evaluados para la variable de salida "Noches reservadas en 2022" multiplicada por "Tarifa diaria promedio en 2022"

Métrica	Red neuronal entrenamiento	Árboles entrenamiento	Red neuronal prueba	Árboles Prueba
R^2	0.8881	0.9934	0.7780	0.9251
RMSE	0.5365	0.1296	0.7772	0.4514
RAE	0.3554	0.0097	0.6774	0.0308
MAPE	4.5142	0.0101	6.2356	0.0341
Coeficiente Pearson	0.9668	0.9967	0.9122	0.9620

En general, los datos obtenidos en este estudio muestran resultados prometedores. En particular, la métrica R^2, que refleja la cantidad de variabilidad en la variable dependiente explicada por las variables independientes, alcanzó un valor destacado de 0.9934. Este alto valor se observó principalmente en los modelos basados en árboles, seguido por las redes neuronales con 0.8881, cuando se consideran los datos de entrenamiento. Sin embargo, al evaluar con los datos de prueba, el ajuste disminuyó, siendo especialmente notable en los modelos basados en árboles, donde R^2 descendió a 0.7780, siendo un 7% menor en el modelo de árboles. Al contrastar los resultados entre entrenamiento y prueba en estos modelos, la significativa discrepancia en los valores de R^2 para los árboles sugiere un posible sobreajuste en dicho modelo esto ya que los valores son considerablemente elevados.

El RMSE, que evalúa la magnitud de los errores de las predicciones, reflejó un comportamiento similar. En el entrenamiento, el modelo de árboles arrojó un valor de 0.1296, considerablemente mejor que el 0.5365 de la red neuronal. Sin embargo, al analizar el conjunto de prueba, el error se incrementó para ambos modelos, siendo más pronunciado en la red neuronal con un RMSE de 0.7772 en contraste con el 0.4514 de los árboles.

En cuanto al RAE y MAPE, métricas que nos proporcionan una perspectiva sobre el error en términos absolutos y porcentuales, respectivamente, se reitera la tendencia observada anteriormente: el modelo basado en árboles supera a la red neuronal en el conjunto de entrenamiento, pero ambos ven disminuir su eficacia al enfrentarse al conjunto de prueba.

A pesar de los descensos mencionados, es crucial subrayar que el coeficiente de Pearson sigue siendo sólido en el conjunto de prueba, con un valor de 0.9122 para la red neuronal y 0.9620 para el modelo basado en árboles, lo que indica una correlación lineal fuerte entre las predicciones y los valores reales.

El análisis gráfico de los residuales constituye una parte esencial del proceso de evaluación, permitiéndonos determinar la confiabilidad de las predicciones. Estas representaciones visuales ofrecen una capacidad intuitiva para inferir rápidamente la calidad y características de nuestros modelos. Las Figuras 28 a 30 ilustran los resultados para el modelo basado en árboles, mientras que las Figuras 31 a 33 muestran los resultados del modelo de redes neuronales.

La Figura 28 presenta una comparativa entre las predicciones generadas con el modelo basado en árboles y los valores reales del conjunto de prueba. Dado el elevado ajuste que exhibe el modelo, la representación gráfica muestra una notable linealidad entre ambas variables. Este comportamiento es precisamente el deseado y esperado en este tipo de análisis gráfico, subrayando la eficacia del modelo basado en árboles en su capacidad predictiva.

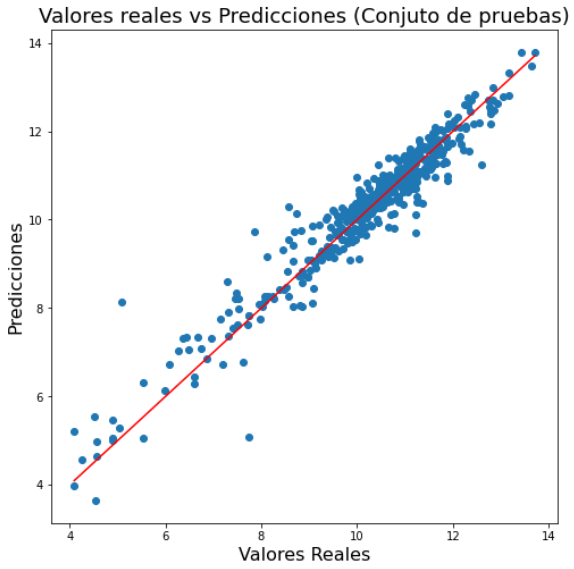


Figura 28. Comparativa entre las predicciones generadas por el modelo basado en árboles y los valores reales del conjunto de prueba.

La Figura 29 muestra un gráfico de los residuales frente a las predicciones. Idealmente, en este tipo de representaciones, se aspira a que los puntos en el diagrama de dispersión se distribuyan de forma aleatoria, sin evidenciar ningún patrón específico. La ausencia de

patrones discernibles refuerza la confianza en la idoneidad del modelo y la pertinencia de los datos seleccionados para este análisis.

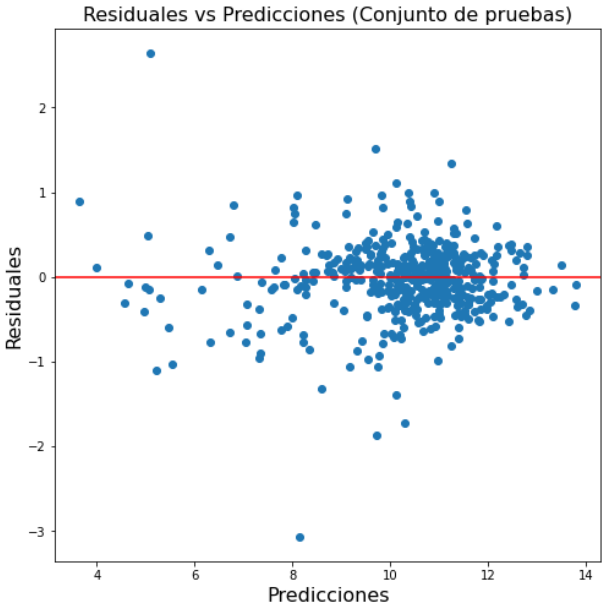


Figura 29. Gráfica de dispersión de los residuales con respecto a las predicciones del modelo basado en árboles de decisión.

Para la Figura 30, se presenta el histograma de normalidad de los residuales. Aunque no muestra una distribución perfectamente normal, se aproxima lo suficiente como para considerarla aceptable para muchos análisis estadísticos. La normalidad de los residuos es crucial ya que, si se cumple, refuerza la confianza en las inferencias realizadas a partir del modelo.

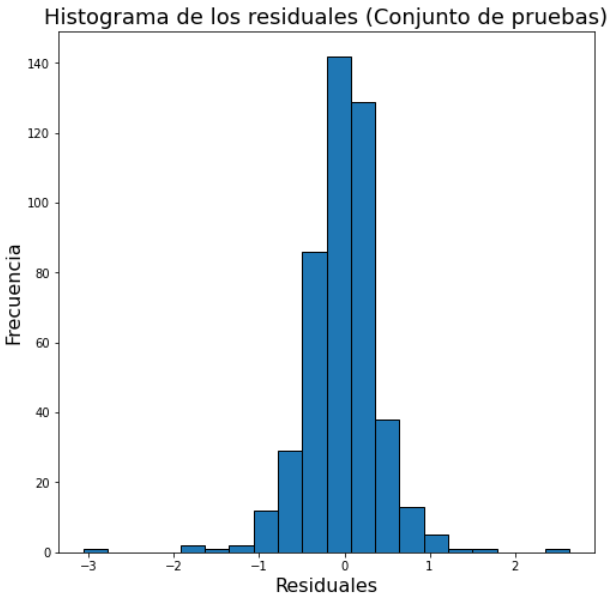


Figura 30. Histograma de los residuales considerando el modelo basado en árboles de decisión.

La Figura 31 correspondiente a los resultados generados con la red neuronal e ilustra la comparación entre los valores reales y las predicciones. Aunque se distingue cierta tendencia lineal, esta no es tan marcada como la que se manifiesta en la Figura 28. Esta observación gráfica se reafirma, lo que se podría anticipar al analizar las métricas de R^2 presentadas en la Tabla 15.

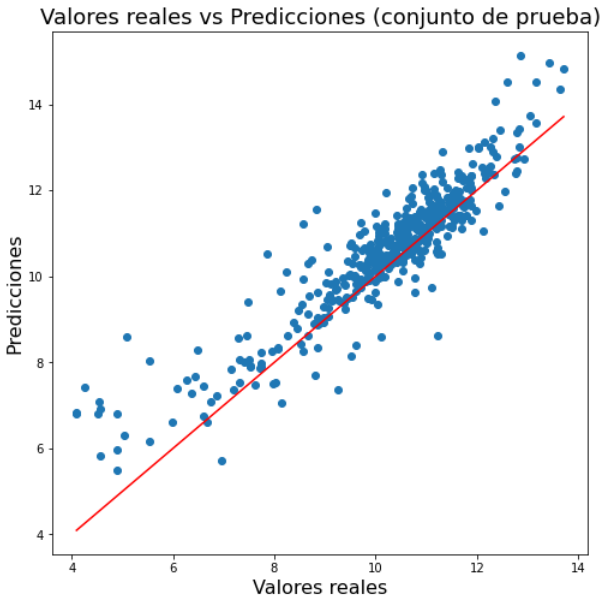


Figura 31. Comparativa entre las predicciones generadas por el modelo basado en redes neuronales y los valores reales del conjunto de prueba.

La Figura 32 se emplea para evaluar si los residuales se ajustan a una distribución normal. Si bien el histograma muestra indicios de normalidad, no es una coincidencia perfecta con la distribución teórica.

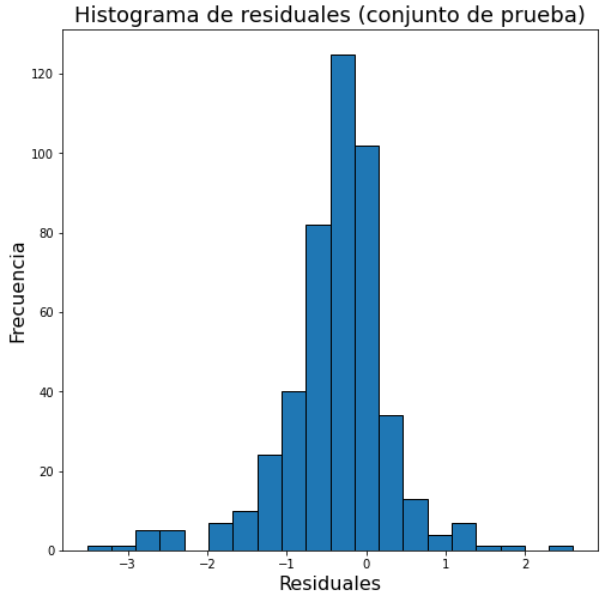


Figura 32. Histograma de los residuales considerando el modelo basado en redes neuronales.

Finalmente, la Figura 33 muestra los residuales contra las predicciones. Al igual que en la Figura 29, en esta figura no se puede observar algún patrón en el diagrama de dispersión.

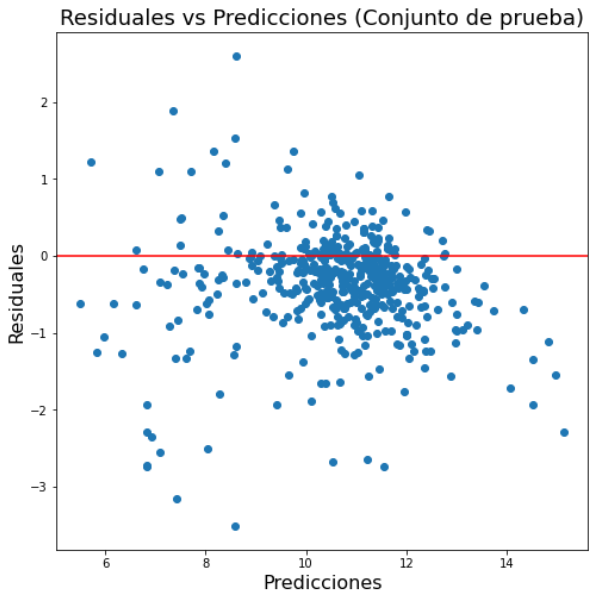


Figura 33. Gráfica de dispersión de los residuales con respecto a las predicciones del modelo basado en redes neuronales.

De manera general, los residuales para ambos modelos, el basado en redes neuronales y el de árboles de decisión, muestran un comportamiento acorde a lo anticipado. Este ajuste es indicativo de la adecuación del modelo al conjunto de datos seleccionado. Manteniendo la tendencia en la que los árboles de decisión parecen ser un mejor modelo para el proyecto.

Uno de los beneficios de utilizar árboles de decisión radica en su capacidad para identificar y destacar las variables más significativas del análisis. En la Figura 34, se presenta un histograma que destaca las 20 variables de mayor relevancia para el modelo junto con su nivel de importancia correspondiente. Entre estas, es notable la prominencia de variables como el número de reservas, la duración promedio de estancia y la cantidad de noches disponibles a lo largo del año.

Otro objetivo primordial del proyecto era determinar y prever las zonas más propicias. Esta regresión, gracias a su naturaleza descriptiva, enfatiza ciertas áreas urbanas, ya que se categorizan como variables significativas. A modo de ejemplo, la Tabla 16 detalla diversas zonas de la ciudad:

Tabla 16. Listado de zonas urbanas clave en relación con la variable "Noches reservadas en 2022" multiplicada por "Tarifa diaria promedio en 2022".

Código	Zona de la ciudad
29	Plaza Andares
19	Av. El Coli
35	La Minerva
30	Acueducto

Con base en el modelo de árboles de decisión, las zonas reflejadas en la Tabla 16 son reconocidas y frecuentadas tanto por habitantes longevos de la ciudad de Guadalajara como por aquellos con cierta familiaridad con la misma. Este reconocimiento no solo respalda la validez del modelo, sino que también se alinea con las expectativas previas del proyecto.

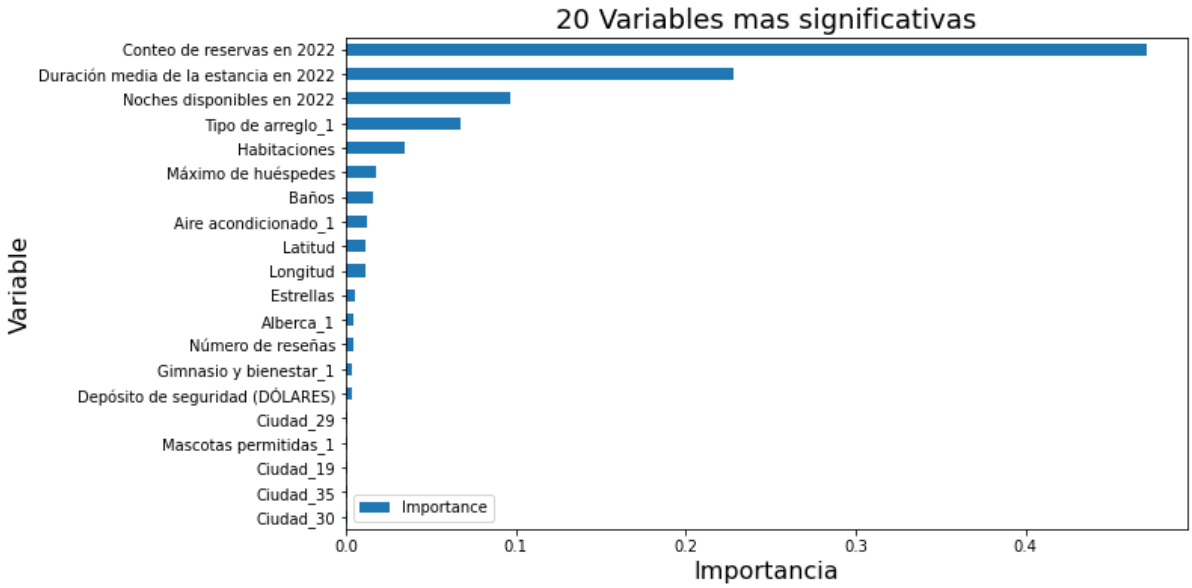


Figura 34. Histograma de las 20 variables de mayor relevancia y su nivel de importancia de acuerdo con el modelo basado en árboles de decisión.

Ahora se presenta el mismo análisis, pero tomando como variable de salida únicamente a las noches reservadas en 2022. Se presenta en la Tabla 17 las métricas para el nuevo modelo generado.

Tabla 17. Análisis comparativo de métricas entre los modelos evaluados para la variable de salida "Noches reservadas en 2022".

Métrica	Red neuronal entrenamiento	Árboles entrenamiento	Red neuronal prueba	Árboles Prueba
R^2	0.9416	0.9954	0.8292	0.9491
RMSE	0.2288	0.0639	0.3965	0.2163
RAE	0.2528	0.0105	0.3849	0.0316
MAPE	Inf	167730831502.9111	inf	5042058472415.7680
Coeficiente Pearson	0.9798	0.9977	0.9230	0.9725

En términos generales, y al igual que con las variables de salida analizadas anteriormente, el análisis de las métricas sugiere que ambos modelos, redes neuronales y árboles de decisión por *Gradient Boosting*, han obtenido buenos resultados, aunque con diferencias notables entre ellos. Específicamente, el coeficiente R^2, que refleja cuánta variabilidad de la variable dependiente es explicada por las variables independientes, muestra un excelente ajuste para el modelo de árboles con un valor de 0.9954 durante el entrenamiento. Por otro lado, las

Redes Neuronales también mostraron un ajuste fuerte, pero ligeramente inferior con un valor de 0.9416. Sin embargo, cuando se analizaron con los datos de prueba, el R^2 para las Redes Neuronales disminuyó a 0.8292, mientras que el de los árboles de decisión descendió a 0.9491. Esta reducción es una señal de que el modelo de árboles, a pesar de su fuerte rendimiento en el entrenamiento, podría estar ligeramente sobre ajustado.

El RMSE, que cuantifica el tamaño de los errores de las predicciones, revela un patrón similar. Durante el entrenamiento, el modelo de árboles tuvo un error significativamente menor de 0.0639 en comparación con el 0.2288 de la red neuronal. Pero, al evaluar con el conjunto de prueba, ambos modelos aumentaron el valor del error, aunque fue más pronunciado en la red neuronal con un RMSE de 0.3965, en comparación con el 0.2163 obtenido con el método de árboles de decisión.

En cuanto a las métricas RAE y MAPE, que ofrecen una visión del error en términos absolutos y porcentuales, se observa nuevamente una superioridad del modelo de árboles de decisión en el conjunto de entrenamiento. Pero, al enfrentarse a los datos de prueba, ambos modelos presentaron un aumento en estos errores. Es importante mencionar que los valores de MAPE muestran anomalías, posiblemente por cálculos problemáticos o valores cercanos a cero en los datos.

Por último, el coeficiente de Pearson, que mide la correlación lineal entre las predicciones y los valores reales, se mantiene fuerte en el conjunto de prueba para ambos modelos. Las redes neuronales obtuvieron un valor de 0.9230, mientras que el modelo de árboles logró un valor de 0.9725, reafirmando la robustez de sus predicciones en relación con los valores verdaderos.

Para los nuevos modelos se hace también el análisis de residuales. Las Figuras 35 a 37 contienen la información de los árboles de decisión, mientras que las Figuras 38 a 40 la generada por el modelo basado en redes neuronales.

La Figura 35 representa una comparativa entre los valores reales y las predicciones para el conjunto de pruebas considerando el modelo de árboles de decisión. En esta gráfica, es notable una tendencia lineal bastante clara, lo que concuerda con el alto grado de ajuste observado previamente. Esta coherencia entre los valores reales y las predicciones subraya la capacidad del modelo de árboles de decisión para generar estimaciones cercanas a los valores reales.

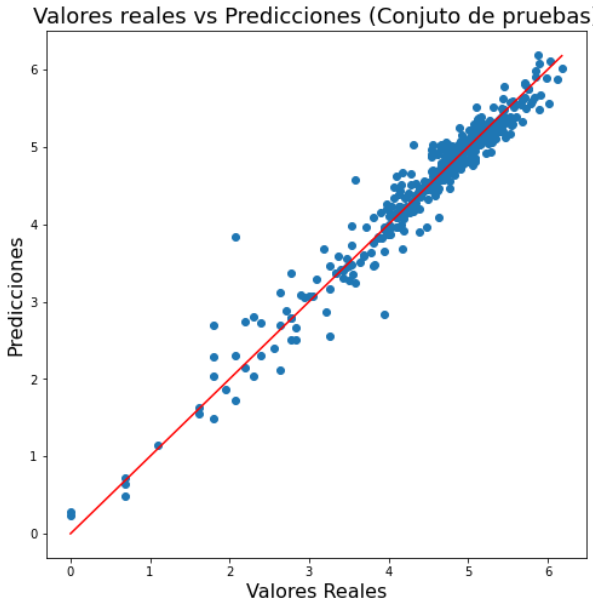


Figura 35. Comparativa entre los valores reales y las predicciones para el conjunto de pruebas considerando el modelo de árboles de decisión.

La Figura 36 presenta el histograma de los residuos. Aunque se observa una tendencia hacia una distribución normal, no se puede afirmar con certeza que los residuos sigan una distribución perfectamente normal.

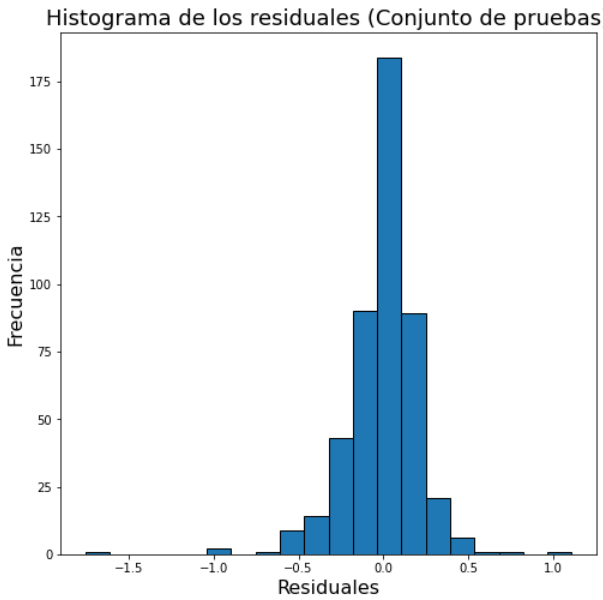


Figura 36. Histograma de los residuales considerando el modelo de árboles de decisión.

La Figura 37 muestra un diagrama de dispersión de las predicciones. No se distingue un patrón definido y hay una notable dispersión entre los puntos.

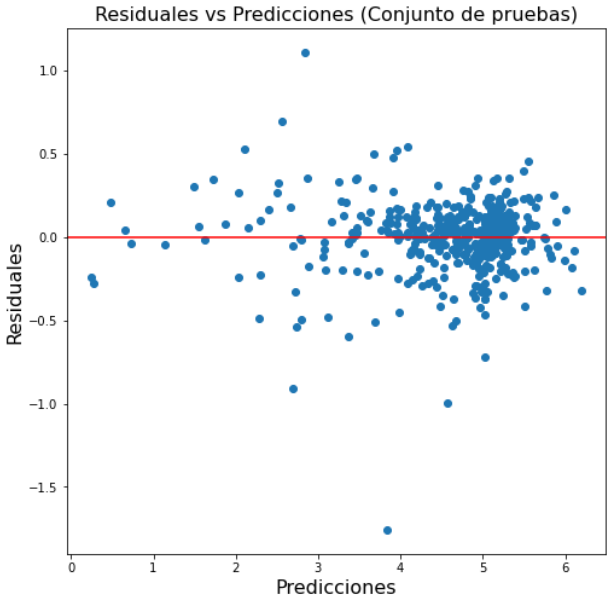


Figura 37. Gráfica de dispersión de los residuales con respecto a las predicciones del modelo basado en árboles de decisión.

La Figura 38 muestra una comparativa entre los valores reales y las predicciones para el conjunto de pruebas considerando el modelo de redes neuronales. Se puede observar que la Figura 38 no exhibe la misma linealidad observada en la Figura 35. En esta, se percibe una mayor linealidad para valores más altos en comparación con los más bajos.

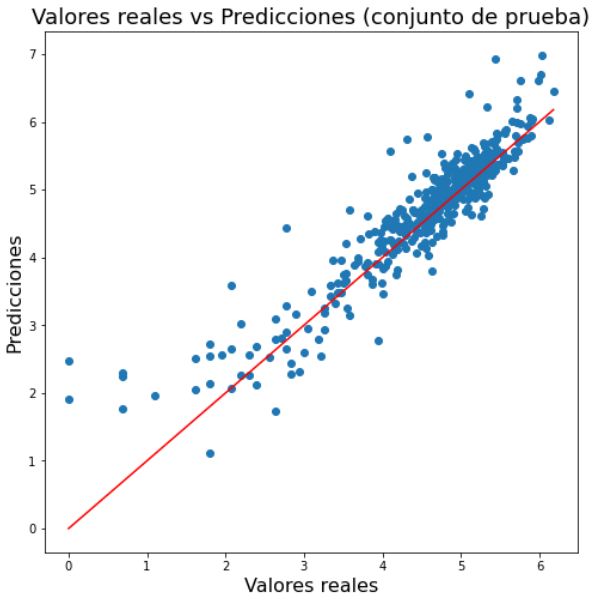


Figura 38. Comparativa entre los valores reales y las predicciones para el conjunto de pruebas considerando el modelo de redes neuronales.

Por otro lado, el histograma de la Figura 39 sugiere un patrón que tiende a la normalidad; sin embargo, al compararlo con el de la Figura 36, se desvía más de lo anticipado.

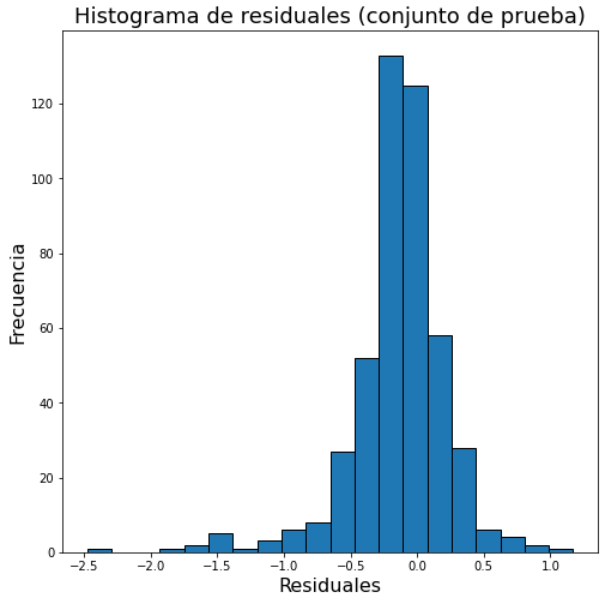


Figura 39. Histograma de los residuales considerando el modelo de redes neuronales.

Finalmente, para la Figura 40, no se observa ningún patrón en el diagrama de dispersión, considerando esto como algo positivo.

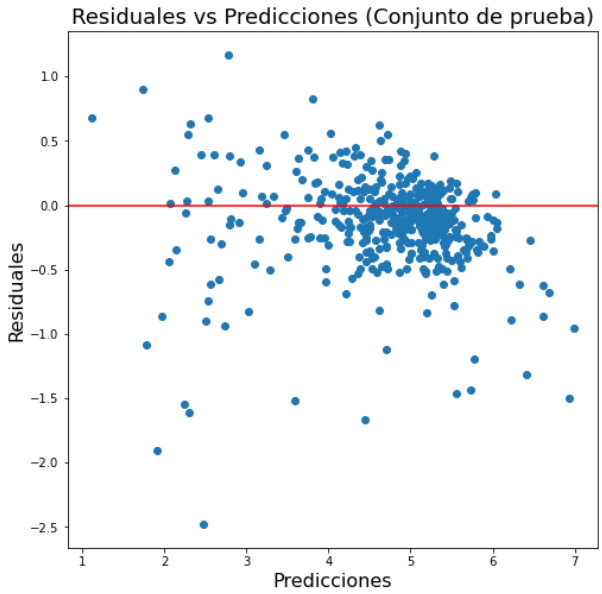


Figura 40. Gráfica de dispersión de los residuales con respecto a las predicciones del modelo basado en redes neuronales.

Teniendo en cuenta las observaciones de las Figuras 38, 39 y 40, es evidente que la aplicación de redes neuronales en este contexto presenta resultados mixtos. Por un lado, la Figura 38 indica una variabilidad en la linealidad dependiendo del rango de valores, mientras que la Figura 39 muestra una tendencia hacia una distribución normal, aunque no se alinea perfectamente con las expectativas basadas en la Figura 36. La ausencia de patrones

discernibles en el diagrama de dispersión de la Figura 40 es alentadora, ya que sugiere una aleatoriedad en los errores, lo cual es deseable en modelos predictivos.

En el nuevo modelo, donde se utiliza como variable de salida las noches reservadas en 2022, se han generado las 20 variables más significativas para los árboles de decisiones. Estas variables se visualizan en la Figura 41. Además, se destacarán las zonas de la ciudad que se incluyeron en esta categoría, las cuales están detalladas en la Tabla 18.

Tabla 18. Listado de zonas urbanas clave en relación con la variable "Noches reservadas en 2022".

Código	Zona de la ciudad
29	Andares
43	Seattle
21	Galerías
42	Country

Es notable que, una vez más, la variable más significativa en relación con las zonas es la cercana a Plaza Andares. Este hecho resulta intrigante, considerando que Plaza Andares es reconocida por ser una zona de alto costo. Adicionalmente, es interesante destacar que esta es la única zona que se mantiene constante en ambos estudios.

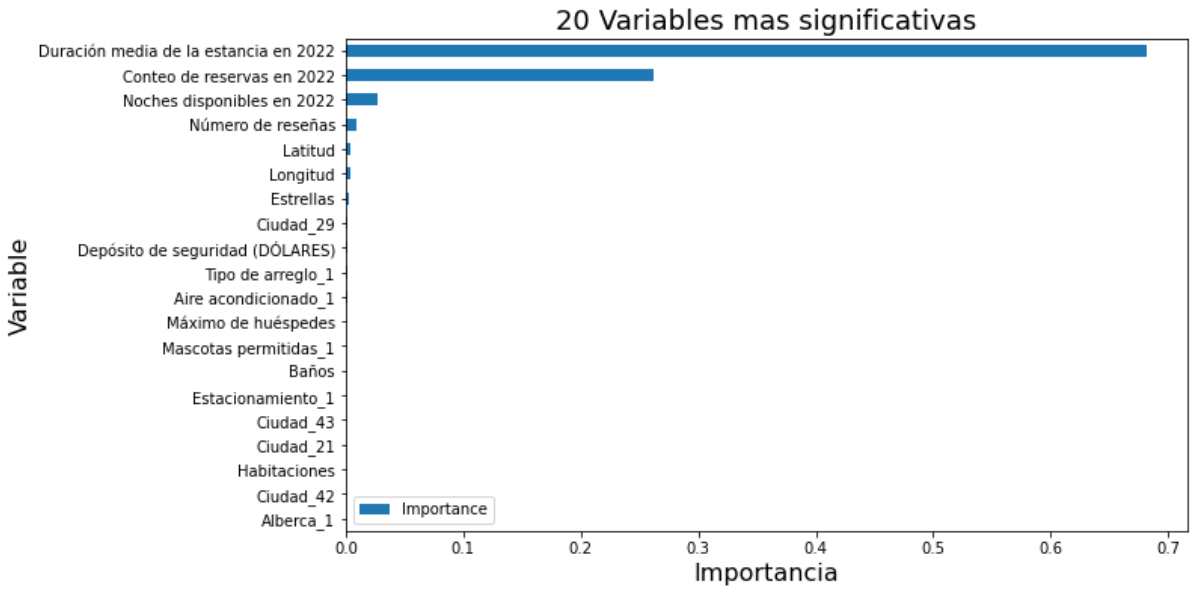


Figura 41. Histograma de las 20 variables de mayor relevancia y su nivel de importancia de acuerdo con el modelo basado en redes neuronales.

4. CONCLUSIONES

4.1. *Conclusiones*

Con respecto al primer objetivo, que es desarrollar un modelo para predecir el número de noches que se reservaría una locación de Airbnb basándose en las características presentes en el conjunto de datos, los modelos utilizados mostraron resultados positivos. Las métricas obtenidas son una clara evidencia de ello. Los árboles de decisión son una herramienta prometedora para este tipo de datos, aunque muestra una fuerte tendencia al sobreajuste.

Los resultados con las redes neuronales mostraron ser positivos. Dada la naturaleza de los datos con los que se contaba, no era de esperarse obtener ajustes sobresalientes. Cabe mencionar que el único preprocesamiento que se realizó fue una normalización. A pesar de la limitación en la cantidad de datos, los resultados muestran que es posible hacer predicciones con bastante precisión, lo cual fue un hallazgo importante.

En relación con el segundo objetivo, que busca predecir la utilidad que se podría obtener de un alojamiento utilizando las mismas variables de entrada, los resultados fueron favorables para ambos modelos. Sin embargo, vale la pena destacar que el valor obtenido para la métrica MAPE resultó ser excepcionalmente alto. Esto podría señalar que la fórmula utilizada para calcular este error no es adecuada para los datos utilizados en el entrenamiento. Sin embargo, causa conflicto que no se haya observado este comportamiento en el primer modelo. Esto sugiere que se debería analizar en detalle el MAPE con respecto a estos datos. A pesar de este inconveniente, ambos modelos mostraron un buen rendimiento. Aunque no se descarta el potencial sobreajuste en los árboles de decisión, en general, los resultados son alentadores.

Una de las ventajas destacadas de los árboles de decisión es la capacidad de identificar las variables más relevantes. Esta cualidad es especialmente valiosa cuando se trata de identificar las amenidades más atractivas para futuros anfitriones. Por ejemplo, al considerar la ubicación de las propiedades como una variable clave, el modelo puede discernir eficazmente cuáles son las zonas más relevantes. Esta información es de gran utilidad, ya que facilita la identificación de las localidades más prometedoras para enfocar inversiones en el futuro.

4.2. *Trabajo Futuro*

Los resultados que se obtuvieron en este proyecto son prometedores y refuerzan la idea de que el análisis que se ha propuesto podría ser una estrategia eficaz para dirigir inversiones en Airbnb. Durante el desarrollo del proyecto resultaron varios desafíos, uno de ellos fue la reducción de la base de datos a la mitad de las observaciones debido a los valores faltantes. Ésta es un área de oportunidad para futuros trabajos. Implementar algún método para completar la información faltante de manera que continúe siendo representativa para el estudio sería una mejora que podría incrementar la confiabilidad de los modelos predictivos desarrollados.

Otra área de oportunidad sería explorar más el modelo basado en redes neuronales ya que ha demostrado gran capacidad para manejar datos de diferentes naturalezas y problemas complejos logrando disminuir el sobreajuste. Además, se pueden considerar otras métricas de desempeño para validar el modelo.

Existen diversos sitios en internet que contienen análisis de datos de Airbnb e incluyen mapas de las zonas analizadas, lo que proporciona una perspectiva más clara de los resultados. Un trabajo futuro sería adoptar esta técnica y agregar mapas detallados al análisis que ya se hizo en el desarrollo del proyecto. Esto permitiría ofrecer a los clientes una visión más amplia y detallada de las oportunidades de mercado.

Finalmente, falta llevar el proyecto a bases de datos de otros lugares, para hacer los ajustes necesarios y saber si los resultados se siguen manteniendo, con la metodología establecida en este proyecto.

5. BIBLIOGRAFÍA

Airbnb. (6 de 10 de 2023). Obtenido de https://news.airbnb.com/es/about-us/#:~:text=Airbnb%20naci%C3%B3%20en%202007%2C%20cuando,todos%20los%20pa%C3%ADses%20del%20mundo.

AllTheRooms. (6 de 10 de 2023). Obtenido de https://alltherooms.com/resources/articles/get-airbnb-statistics-for-any-market/

Amieva, A. U. (07 de 07 de 2021). *El Economista*. Obtenido de https://www.eleconomista.com.mx/capitalhumano/Millennials-los-lideres-de-la-nueva-normalidad-del-mundo-del-trabajo-20210707-0003.html

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Carrillo, G. (6 de Noviembre de 2019). *medium*. Obtenido de https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a

Cichosz, P. (2015). *Data Mining Algorithms Explainded Using R.* Wiley.

Hyndman, R. &. (2021). *Forecasting: principles and practice.* Melbourne: OTexts.

INEGI. (15 de Diciembre de 2022). Obtenido de https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2022/CST/CST2021.pdf

INEGI. (26 de 09 de 2022). *INEGI*. Obtenido de https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_TURISMO_22.pdf

Max Kuhn, K. J. (2016). *Applied Predictive Modeling.* Groton: Springer.

Michael H. Herzog, G. F. (2019). *Understanding Statistics and Experimental Desing How to Not Lie with Statistics.* Ankara: Springer.

Nielsen, M. (1 de Diciembre de 2019). *Neuronal Networks and Deep Learning.* Obtenido de http://neuralnetworksanddeeplearning.com/

Patiño, D. (25 de 03 de 2021). *Expansión*. Obtenido de https://expansion.mx/economia/2021/03/25/millennials-y-centennials-patrocinaran-futura-reforma-fiscal

Shmueli, G. (2017). *Data Mining For Business Analytics Concepts, Techniques.* Hoboken: Wiley.