

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

CARRERA DE INGENIERÍA COMERCIAL



**UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA**

Departamento de Ingeniería Comercial

SEMINARIO DE TÍTULO

**"DETERMINAR UN MODELO ECONOMETRICO QUE NOS PERMITA PREDECIR
O IDENTIFICAR PRECIOS PARA LOS ARRIENDOS DE LOCALES
COMERCIALES IDENTIFICANDO CUÁLES SON LAS VARIABLES MÁS
RELEVANTES"**

AUTOR:

MILENKO VODANOVIC DELGADO

PROFESOR GUÍA Y CORREFERENTE:

HUMBERTO VILLALOBOS, MACARENA GATICA



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITÓRIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: DETERMINAR UN MODELO ECONOMETRICO QUE NOS PERMITA PREDECIR O IDENTIFICAR PRECIOS PARA LOS ARRIENDOS DE LOCALES COMERCIALES IDENTIFICANDO CUÁLES SON LAS VARIABLES MÁS RELEVANTES

Nombre del candidato(a): Milenko Vodanovic Delgado

Carrera / Grado: Ingeniería Comercial

Campus: Casa Central Valparaíso ; **Departamento:** Ingeniería Comercial

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Humberto Villalobos, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 04/08/2026 ; Firma:

Estudiante o Candidato(a):

Fecha: 04/08/2026 ; Firma:

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por

Resumen

El presente estudio de investigación desarrolla un modelo predictivo para identificar y estimar los precios de arriendo de locales comerciales en la comuna de Viña del Mar, Región de Valparaíso, durante el período 2025-2026. La investigación se fundamenta en la Teoría de los Precios Hedónicos propuesta por Sherwin Rosen (1974), que establece que el precio de un bien heterogéneo puede descomponerse en el valor implícito de cada uno de sus atributos individuales.

La base de datos fue construida manualmente a partir de 500 observaciones recopiladas de las principales plataformas inmobiliarias del país, Portalinmobiliario, Yapo y Chileinmobiliaria, abarcando 13 sectores de la comuna. En su versión actualizada, la base incorpora 19 variables, incluyendo coordenadas geográficas (latitud/longitud), distancia al centro urbano y a nodos comerciales, scores de accesibilidad, flujo peatonal y visibilidad estimados mediante geolocalización, además de las variables estructurales originales.

Para la modelización se implementaron dos estrategias complementarias. En primer lugar, se estimó un modelo de Mínimos Cuadrados Ordinarios (MCO) log-lineal hedónico clásico, que alcanzó un R^2 de 0,820 en escala logarítmica sobre el conjunto de prueba ($n=100$), con una RMSE logarítmica de 0,424. Este modelo provee coeficientes interpretables que cuantifican el efecto marginal de cada atributo sobre el logaritmo del precio. En segundo lugar, se compararon cuatro algoritmos de aprendizaje supervisado: Ridge, Lasso, ElasticNet y Árbol de Regresión. El Árbol de Regresión con profundidad máxima de 7 nodos fue identificado como el modelo de mejor desempeño en escala original, alcanzando un R^2 de 0,927 y un Error Porcentual Absoluto Medio (MAPE) de 27,4% en el conjunto de prueba.

El hallazgo económico central es consistente en ambas aproximaciones: la superficie del local es el determinante dominante del precio total de arriendo. En el modelo MCO, el coeficiente del logaritmo de la superficie es 0,784, interpretable como una elasticidad precio-superficie próxima a la unidad. En el árbol de regresión, la superficie explica el 90,4% de la varianza del precio. Las nuevas variables de geolocalización, especialmente el score de flujo peatonal y la accesibilidad aportan señales adicionales estadísticamente relevantes en el modelo MCO, enriqueciendo la explicación económica del componente de localización que en los modelos anteriores permanecía capturado sólo mediante variables categóricas de sector y tipo de local.

Palabras clave: precios hedónicos, arriendo comercial, machine learning, MCO log-lineal, árbol de regresión, geolocalización, Viña del Mar, econometría aplicada.

*“El trabajo duro es inútil para aquellos que
no creen en sí mismos”
- Naruto Uzumaki*

Agradecimientos

Quiero empezar dando las gracias a toda mi querida familia. Mis padres y hermanos quienes siempre confiaron en mí y trabajaron duro para yo poder estudiar sin problemas, quienes me heredaron su voluntad y las ganas de hacer siempre lo correcto y nunca bajar la cabeza. A mis abuelos y tíos que siempre se preocuparon de mi bienestar y siempre me desearon los mejores resultados. Gracias a mis primos Cristian, Jaime y mi hermano menor Antonio (el mítico Mcdo) que siempre me apoyaron conversando, jugando e incluso viajando para vernos, quienes siempre han estado conmigo y por quienes haría lo posible para verlos bien. todas las personas que me acompañaron en este proceso, para alguien foráneo que llegó solo y sin conocer ni siquiera la región fue todo un desafío. Como no mencionar a los profesores más influyentes en mi periodo universitario, Miguel Olivares (aunque no se acuerde mi), Humberto Villalobos (VillaGoat) y Macarena Gatica, profesores que me encaminaron y siempre pusieron esfuerzo y voluntad para mejorar, ayudar e incluso preocuparse. Tuve la suerte de encontrarme con personas maravillosas que luego llamaría amigos y se convertirían en pilares de mi desarrollo y estancia en la Universidad. Quienes, además, me aconsejaron y vivieron aventuras conmigo yéndonos de viaje, a comer, a trotar, a entrenar y salir de fiesta (no mucha). Como no mencionar al Team Sangre, mítico club de futbol donde se perdieron la mayoría de los partidos y al último grupo “Quesitos” quienes fueron muy cercanos y acompañaron a muchas experiencias. Así como Luciano, José, Seba y Gustavo quienes fueron grandes amigos en este último periodo.

Gracias totales a todos los que estuvieron presentes. Los llevaré siempre en mi memoria y parte de mi desempeño como profesional para cumplir mis objetivos de vida se deberá a ustedes.

Tabla de contenidos

RESUMEN	II
AGRADECIMIENTOS	V
I. INTRODUCCIÓN	X
1.1 Contexto general	x
1.2 Planteamiento del problema	x
1.3 Justificación	xi
1.4 Estructura del documento	xi
1.5 Objetivo principal de la investigación	xii
1.6 Objetivos específicos	xii
II. MARCO TEÓRICO	XII
2.1 Teoría de la Renta de Oferta (Bid Rent Theory)	xiii
2.2 Teoría de los Precios Hedónicos	xiii
2.3 Space Syntax y Accesibilidad Urbana	xiv
2.4 Machine Learning en Econometría Inmobiliaria	xv
2.4.1 Regresión Ridge	xv
2.4.2 Regresión Lasso	xv
2.4.3 Regresión ElasticNet	xv
2.4.4 Árbol de Regresión	xv
2.5 Definición y Clasificación de Variables	xv
2.6 Conceptos Clave	xvi
III. ESTADO DEL ARTE	XVIII

3.1 Estudios Fundacionales	xviii
3.2 Accesibilidad, Localización y Space Syntax	xviii
3.3 Machine Learning en el Mercado Inmobiliario	xviii
3.4 Contexto Latinoamericano y Chileno	xix
3.5 Contexto Chileno Reciente: Estudios Académicos y Plataformas PropTech	xix
3.5.1 Estudios Académicos Recientes en Chile	xix
3.5.2 Plataformas Tecnológicas de Valorización Inmobiliaria en Chile	xx
3.5.3 Inteligencia de Mercado Comercial: GPS Property e Inciti	xxi
IV. METODOLOGÍA	XXII
4.1 Diseño de la Investigación	xxii
4.2 Recolección de Datos	xxii
4.2.1 Fuentes de Información	xxii
4.2.2 Criterios de Inclusión y Exclusión	xxii
4.2.3 Proceso de Registro y Normalización	xxiii
4.3 Actualización y Expansión de la Base de Datos	xxiii
4.3.1 Ampliación Muestral: de 300 a 500 Observaciones	xxiii
4.3.2 Integración de Variables de Geolocalización	xxiii
4.3.3 Variables Adicionales Incorporadas	xxv
4.4 Procesamiento de Datos	xxvi
4.5 Especificación y Entrenamiento de los Modelos ML	xxvi
4.6 Especificación del Modelo MCO Hedónico Log-Lineal	xxvii
4.6.1 Especificación del Modelo	xxvii
4.6.2 División Train/Test	xxviii
4.6.3 Métricas de Evaluación	xxviii
4.7 Creación de un sistema funcional abierto al público	xxviii
V. BASE DE DATOS Y ESTADÍSTICA DESCRIPTIVA	XXIX

5.1 Descripción General del Dataset	xxix
5.2 Distribución por Nodo Comercial	xxxi
5.3 Distribución por Tipo de Local	xxxi
5.4 Distribución por Estado de Conservación	xxxii
5.5 Análisis de Correlación con el Precio	xxxii
VI. MODELOS MCO HEDÓNICO LOG-LINEAL	XXXIII
6.1 Resultados del primer Modelo MCO	xxxiii
6.2 Coeficientes e Interpretación Económica	xxxiv
6.3 Resultados del Modelo final MCO	xxxv
6.3 Diagnóstico Estadístico	xxxvi
6.3.1 Heterocedasticidad	xxxvi
6.3.2 Multicolinealidad	xxxvi
6.4 Análisis con Componente Principal (PCA)	xxxvii
VII. PRIMEROS MODELOS ML: DIAGNÓSTICO Y CORRECCIONES	XXXVIII
7.1 Resultados Iniciales: Comparación de los Cuatro Modelos	xxxviii
7.2 Diagnóstico de los Modelos Lineales Regularizados	xxxviii
7.2.1 Heterocedasticidad	xxxviii
7.2.2 No Normalidad de los Residuos	xxxviii
7.2.3 Multicolinealidad (VIF)	xxxix
7.3 Diagnóstico del Árbol de Regresión: Sobreajuste	xxxix
VIII. MODELO ML: ÁRBOL DE REGRESIÓN (MAX_DEPTH = 7)	XL
8.1 Justificación de la Selección del Modelo	xl
8.2 Importancia de Variables	xl

8.3 Rendimiento del Modelo Final	xl
8.4 Ejemplos de Predicciones del Modelo	xl
IX. ANÁLISIS DE RESULTADOS E INTERPRETACIÓN ECONÓMICA	XLI
9.1 El Precio Total está Determinado por la Superficie	xli
9.2 El Flujo Peatonal como Nuevo Determinante del Precio Unitario	xli
9.3 Elasticidad Precio-Superficie: Economías de Escala	xli
9.4 Diferencias entre Tipos de Local y Nodos Comerciales	xlii
9.5 Estado de Conservación: Hallazgo Confirmado	xlii
9.6 Implicancias Prácticas	xlii
X. DISCUSIÓN	XLIV
10.1 Contraste con la Literatura	xliv
XI. ANÁLISIS INTERPRETATIVO Y RECOMENDACIONES DE MEJORA	XLIV
11.1 Síntesis Interpretativa de los Resultados Actualizados	xliv
11.2 Hallazgos Nuevos Relevantes de la Base Actualizada	xl v
11.3 Recomendaciones para Mejorar los Modelos	xlvi
11.3.1 Implementar Modelos de Ensamble (Gradient Boosting / Random Forest)	xlvi
11.3.3 Eliminar la Circularidad entre CLP/m ² y el Precio	xlvi
11.3.4 Explorar Modelos Econométricos Espaciales	xlvi
11.3.5 Medir Flujo Peatonal con Datos Reales	xl vii
11.3.6 Incorporar la Dimensión Temporal	xl vii
11.3.7 Ampliar la Validación Geográfica	xl vii
XII. CONCLUSIONES	XLVIII
REFERENCIAS BIBLIOGRÁFICAS	XLIX

I. Introducción

1.1 Contexto general

El mercado inmobiliario comercial constituye uno de los pilares fundamentales del ecosistema económico urbano. La decisión de arrendar un local comercial no es azar: para un emprendedor o comerciante, el arriendo representa frecuentemente uno de los costos operativos fijos más significativos, pudiendo representar un gran porcentaje de los ingresos brutos de un negocio minorista. En consecuencia, la elección de la localización y el tipo de local incide directamente sobre la viabilidad financiera del proyecto comercial.

En el contexto chileno, el sector inmobiliario ha experimentado una transformación profunda en los últimos años. Según datos de la Cámara Chilena de la Construcción, las autorizaciones de nuevos proyectos comerciales han caído aproximadamente un 17% en los últimos tres años, niveles no observados desde la crisis financiera de la década de 1990. Simultáneamente, el salario promedio líquido en Chile se sitúa en torno a los \$897.000 mensuales (marzo 2026), cifra insuficiente para que la mayoría de los trabajadores financie de manera autónoma la adquisición de un inmueble comercial. Esta brecha estructural entre oferta restringida y demanda sostenida garantiza que el arriendo se mantenga como la modalidad dominante de acceso al espacio comercial en las ciudades chilenas.

Viña del Mar, segunda ciudad en importancia de la Región de Valparaíso, concentra una actividad comercial diversa y dinámica. Su condición de ciudad turística con alta variabilidad estacional, su cercanía con el Puerto de Valparaíso y su heterogénea estructura urbana que abarca desde el eje comercial de Avenida Valparaíso hasta sectores residenciales periféricos como Achupallas o Agua Santa la convierten en un laboratorio ideal para el estudio de la formación de precios en el mercado de arriendos comerciales.

1.2 Planteamiento del problema

A pesar de su relevancia económica, el mercado de arriendos de locales comerciales en Viña del Mar carece de estudios empíricos rigurosos que permitan explicar y predecir la variabilidad en los precios. A diferencia del segmento residencial, donde existen múltiples estudios académicos y herramientas digitales de valorización, el mercado comercial opera en gran medida bajo condiciones de información asimétrica: arrendatarios y propietarios toman

decisiones basándose en comparaciones informales o en la intuición del corredor inmobiliario, sin acceso a un modelo objetivo que descomponga los factores que pueden determinar el precio.

Esta opacidad informacional tiene consecuencias concretas. Como, por ejemplo: los emprendedores en etapa de instalación sobreestiman o subestiman el costo de arrendamiento en función del sector geográfico, el tipo de local y la superficie requerida. Los propietarios fijan precios sin referencia sistemática a los atributos reales del inmueble. Los inversionistas evalúan portafolios sin una metodología objetiva de valorización. El resultado es un mercado con alta dispersión de precios, ineficiencias en la asignación de espacios comerciales y barreras de entrada para nuevos negocios.

La pregunta de investigación central del presente seminario es: ¿Qué variables determinan el precio de arriendo de un local comercial en Viña del Mar, y en qué magnitud lo hacen? La respuesta a esta pregunta, articulada a través de modelos econométricos con uso de herramientas de Machine Learning con datos reales del mercado, constituye la contribución original de este trabajo.

1.3 Justificación

La relevancia de este estudio se compone de tres aristas. Primero, desde el punto de vista académico, la literatura sobre precios hedónicos en el mercado inmobiliario comercial latinoamericano es considerablemente más escasa que en el segmento residencial, lo que otorga originalidad al presente trabajo. Segundo, desde el punto de vista metodológico, la combinación de la Teoría de los Precios Hedónicos con algoritmos de Machine Learning y con un modelo MCO log-lineal enriquecido con variables de geolocalización representa una innovación respecto a los enfoques econométricos clásicos dominantes en la literatura chilena. Tercero, desde el punto de vista práctico, los resultados de los modelos permiten construir una herramienta de orientación de precios con aplicabilidad directa para tomadores de decisiones en el mercado inmobiliario comercial de la región. En este caso, se propone aplicar Machine Learning debido a su alta adaptabilidad y gran poder de predicción, además, se busca innovar con nuevos métodos para este tipo de problemas.

1.4 Estructura del documento

La presente investigación se organiza de la siguiente manera. El Capítulo II desarrolla el Marco Teórico. El Capítulo III revisa el Estado del Arte. El Capítulo IV describe la Metodología, incluyendo la actualización de la base de datos con 500 observaciones y variables de geolocalización, y la especificación del nuevo modelo MCO hedónico log-lineal.

El Capítulo V presenta la estadística descriptiva de la base de datos actualizada. El Capítulo VI presenta los resultados del modelo MCO hedónico log-lineal. El Capítulo VII analiza los modelos ML iniciales y sus diagnósticos. El Capítulo VIII presenta el modelo ML final seleccionado. El Capítulo IX ofrece la interpretación económica comparada de ambas aproximaciones. El Capítulo X discute los hallazgos y limitaciones. El Capítulo XI concluye.

1.5 Objetivo principal de la investigación

Realizar una investigación aplicada a la comuna de Viña del Mar en el mercado inmobiliario que permita determinar un modelo econométrico que intente predecir o identificar precios para el arriendo de locales comerciales, logrando reconocer cuales son las variables más relevantes que afectan la volatilidad de estos.

1.6 Objetivos específicos

- Analizar modelos funcionales que se hayan probado antes, a través de estudios académicos o artículos realizados con anterioridad.
- Construir una base de datos sólida con fuentes verificadas como Yapo o Portalinmobiliario, utilizando técnicas como web scraping mediante el uso de herramientas avanzadas.
- Identificar y categorizar variables del tipo estructurales, localización y financiero, considerando que sean tanto cualitativas como variables cuantitativas.
- Probar primeros modelos estadísticos, corregir errores de heterocedasticidad y eliminar variables irrelevantes para evitar autocorrelación.
- Analizar la sensibilidad de los precios de arriendo ante cambios en las variables más significativas, determinando cuales son las que más suben o disminuyen los precios de los locales, reconociendo las tendencias del mercado.

II. Marco Teórico

El presente capítulo establece los fundamentos conceptuales sobre los cuales se estructura la investigación. Se abordan, en primer lugar, las teorías económicas clásicas sobre la formación de precios en el mercado inmobiliario. En segundo lugar, se desarrollan los fundamentos de los algoritmos de aprendizaje automático empleados en el análisis. Finalmente, se presenta la taxonomía de variables utilizada en la construcción del modelo.

2.1 Teoría de la Renta de Oferta (Bid Rent Theory)

La Teoría de la Renta de Oferta, formulada por William Alonso en 1964 en su obra "Location and Land Use", constituye el punto de partida clásico para el análisis de la formación de precios en el mercado inmobiliario urbano. En su formulación original, Alonso propone que los agentes económicos como hogares, comercios e industrias compiten por ubicarse en el espacio central urbano (en inglés, CBD: Central Business District), y que la disposición a pagar de cada agente disminuye mientras se aleja del centro de la ciudad.

La función de renta de oferta $Y(r)$ en un punto ubicado a distancia r del CBD satisface:

$$Y(r) = [R(q) - C(q) - t \cdot r \cdot q] / A$$

Donde $R(q)$ es el ingreso por ventas, $C(q)$ son los costos de producción distintos al arriendo, t es el costo de transporte por unidad, q es la cantidad producida/vendida y A es la superficie arrendada. El modelo predice una estructura urbana monocéntrica consistente con la historia de Viña del Mar, aunque la evidencia contemporánea sugiere estructuras multicéntricas crecientes, es decir, cada vez hay más nodos comerciales y no solo un centro comercial en la ciudad.

2.2 Teoría de los Precios Hedónicos

La Teoría de los Precios Hedónicos, cuyo modelo formal fue desarrollado por Sherwin Rosen en 1974, propone que el precio de un bien heterogéneo puede descomponerse en el valor implícito de cada uno de sus atributos constitutivos.

Formalmente:

$$P = f(z_1, z_2, \dots, z_n) + \varepsilon$$

Donde la derivada parcial $\partial P / \partial z_i$ captura el precio hedónico o precio sombra del atributo i-ésimo.

Aplicado al mercado inmobiliario comercial, el modelo establece que el precio de arriendo es la suma implícita de los valores que el mercado asigna a atributos estructurales (superficie, pisos, estado), de localización (sector, tipo de vía, accesibilidad) y de entorno (flujo peatonal, visibilidad, densidad comercial). La implementación clásica mediante MCO con la variable dependiente en logaritmos para capturar elasticidades, constituye la especificación del nuevo modelo implementado en esta investigación.

2.3 Space Syntax y Accesibilidad Urbana

La teoría Space Syntax, desarrollada por Bill Hillier y Julienne Hanson (UCL, 1970s-1980s), propone que la configuración de la red de calles determina los patrones de movilidad y, en consecuencia, el nivel de exposición comercial de cada localización. El concepto central es el "valor de integración", es decir, el valor del local que adquiere por su localización que mide la accesibilidad topológica de un segmento de calle respecto al resto de la red urbana. Netzell (2009) demostró que este indicador predice el flujo peatonal y, a través de él, el precio de arriendo de locales comerciales.

En la base de datos actualizada del presente estudio, las variables `accesibilidad_score_0_100`, `flujo_peatonal_score_0_100` y `visibilidad_score_0_100` operacionalizan los conceptos de Space Syntax mediante scores continuos estimados geoespacialmente para cada localización, representando un avance significativo respecto a la aproximación categórica de la variable Tipo de Local empleada en la versión original.

2.4 Machine Learning en Econometría Inmobiliaria

El aprendizaje automático (Machine Learning, ML) comprende un conjunto de algoritmos que aprenden patrones directamente de los datos, sin requerir la especificación explícita de una forma funcional. A diferencia de la econometría clásica, que está orientada principalmente a la inferencia y la interpretación de coeficientes, el ML prioriza la capacidad de predicción fuera de la muestra de entrenamiento. Este trabajo implementa cuatro algoritmos: Regresión Ridge, Lasso, ElasticNet y Árbol de Regresión.

2.4.1 Regresión Ridge

La regresión Ridge añade al criterio MCO una penalización L2:

$$L_{\text{ridge}}(\beta) = \sum_i (y_i - X_i^T \beta)^2 + \lambda \cdot \sum_j \beta_j^2.$$

Retiene todas las variables y reduce la varianza en presencia de multicolinealidad.

2.4.2 Regresión Lasso

La regresión Lasso utiliza una penalización L1:

$$L_{\text{lasso}}(\beta) = \sum_i (y_i - X_i^T \beta)^2 + \lambda \cdot \sum_j |\beta_j|.$$

Puede forzar coeficientes exactamente a cero, realizando selección automática de variables.

2.4.3 Regresión ElasticNet

ElasticNet combina penalizaciones L1 y L2:

$$L_{\text{en}}(\beta) = \sum_i (y_i - X_i^T \beta)^2 + \lambda \cdot [\text{L1_ratio} \cdot \sum_j |\beta_j| + (1 - \text{L1_ratio}) \cdot \sum_j \beta_j^2].$$

Hereda las ventajas de ambos modelos.

2.4.4 Árbol de Regresión

Los árboles de decisión particionan recursivamente el espacio de atributos minimizando la suma de errores cuadráticos en cada partición. Sus ventajas son la captura de relaciones no lineales, la ausencia de supuestos distribucionales y la interpretabilidad jerárquica.

2.5 Definición y Clasificación de Variables

Para el presente estudio, las variables han sido clasificadas siguiendo la taxonomía clásica de los modelos hedónicos inmobiliarios (Benjamin, Boyle y Sirmans, 1992). La versión actualizada de la base de datos incorpora variables adicionales de geolocalización y accesibilidad que se presentan en la Tabla 1 del Capítulo IV.

2.6 Conceptos Clave

- Arriendo: valor que paga mensualmente el arrendatario por el uso de un local comercial.
- Gastos comunes (GC): costos compartidos del inmueble (mantención, áreas comunes) sumados al arriendo.
- Precio + GC: variable dependiente del modelo; arriendo mensual más gastos comunes, en pesos chilenos.
- CLP/m²: precio unitario del arriendo, calculado como Precio + GC dividido por la superficie del local.
- Superficie: cantidad total de metros cuadrados construidos o disponibles del local comercial.
- MCO (Mínimos Cuadrados Ordinarios): método de estimación que minimiza la suma de errores cuadráticos.
- Modelo log-lineal: especificación en que la variable dependiente se transforma con logaritmo natural.
- Elasticidad: variación porcentual del precio ante un cambio porcentual de un atributo del local.
- R² (coeficiente de determinación): proporción de la varianza del precio explicada por el modelo.
- RMSE (raíz del error cuadrático medio): magnitud promedio del error de predicción del modelo.
- MAE (error absoluto medio): promedio de las diferencias absolutas entre precio real y predicho.
- MAPE (error porcentual absoluto medio): error promedio expresado como porcentaje del precio real.
- Heterocedasticidad: varianza no constante de los residuos del modelo a lo largo de las observaciones.
- Multicolinealidad: alta correlación entre dos o más variables independientes del modelo.
- Sobreajuste (overfitting): modelo que se ajusta excesivamente a los datos de entrenamiento y generaliza mal.
- Score de accesibilidad: indicador de 0 a 100 que mide la conectividad urbana de la ubicación del local.

- Nodo comercial: polo o eje comercial de mayor relevancia más cercano a la ubicación del local.
- Hiperparámetros: configuración definida antes de entrenar un modelo para controlar su aprendizaje.
- Validación cruzada: técnica que evalúa el desempeño de un modelo en distintos subconjuntos de datos.
- PCA (Análisis de Componentes Principales): técnica que transforma un conjunto de variables posiblemente correlacionadas como combinación lineal generando un conjunto más pequeño de componentes principales (variables lineales no correlacionadas) con la finalidad de reducir dimensionalidad de datos sin perder información valiosa

III. Estado del Arte

El presente capítulo revisa la literatura empírica relevante para el problema de investigación, organizando los estudios en cuatro grupos: los trabajos fundacionales del modelo hedónico en mercados comerciales, la evidencia sobre accesibilidad y localización, las aplicaciones de Machine Learning al mercado inmobiliario, y los estudios en el contexto latinoamericano y chileno.

3.1 Estudios Fundacionales

El trabajo seminal de Benjamin, Boyle y Sirmans (1992) constituye la primera aplicación sistemática del modelo hedónico al mercado de locales comerciales en centros comerciales. Los autores encuentran que la superficie, la posición dentro del centro comercial y la presencia de tiendas ancla son determinantes estadísticamente significativos del precio de arriendo. Glascock, Kim y Sirmans (1993) extienden el análisis hacia la dimensión temporal, documentando variabilidad de los coeficientes hedónicos según período y localización. Eppli y Benjamin (1994) investigan la relación entre mezcla de arrendatarios y precio del arriendo, documentando economías de aglomeración.

3.2 Accesibilidad, Localización y Space Syntax

Netzell (2009) prueba directamente en Estocolmo la hipótesis de que el valor de integración Space Syntax predice las rentas de locales comerciales, encontrando un efecto estadísticamente significativo y positivo. Yao y Pretorius (2014) analizan 153 locales en Yakarta, encontrando que la visibilidad del local tiene un efecto positivo y significativo sobre el precio del arriendo, incluso controlando por superficie y distancia a accesos.

3.3 Machine Learning en el Mercado Inmobiliario

Kontrimas y Verikas (2011) demuestran que modelos no paramétricos superan a la regresión lineal en R^2 y RMSE cuando la relación entre atributos y precio presenta no linealidades. Yacim y Boshoff (2018) concluyen, tras revisar 45 estudios, que los modelos de ML (Machine Learning) reducen el error de predicción respecto al MCO en un promedio de 15% a 30%. Chiang, Bin y Strawderman (2019) encuentran que Random Forest supera al MCO en R^2 (0,82 vs 0,71) para 2.400 locales comerciales en Estados Unidos.

3.4 Contexto Latinoamericano y Chileno

Núñez et al. (2021) validan para el contexto chileno la relevancia de variables de localización en el segmento residencial de la Región del Biobío. Ramírez y Agudelo (2018) documentan autocorrelación espacial en los residuos de modelos hedónicos para Bogotá, sugiriendo que los modelos con rezago espacial obtienen mejores estadísticos de ajuste. Mabel y Tomás (2017) encuentran heterogeneidad espacial significativa en los coeficientes hedónicos entre barrios de Buenos Aires.

3.5 Contexto Chileno Reciente: Estudios Académicos y Plataformas PropTech

En los últimos años, el ecosistema de datos y modelos predictivos inmobiliarios en Chile ha experimentado un desarrollo acelerado, tanto desde el ámbito académico como desde la industria tecnológica. Este contexto es relevante para situar la presente investigación, que se enfoca en el segmento comercial de ciudades intermedias, un nicho aun escasamente cubierto por las herramientas disponibles.

3.5.1 Estudios Académicos Recientes en Chile

El trabajo más directamente comparable al presente estudio en el contexto chileno es la memoria de título de Ingeniería Civil Industrial de la Universidad de Chile (Repositorio académico UCH, 2020), titulada "Sistema de predicción de precios-venta de inmuebles en el mercado del sector inmobiliario de la Región Metropolitana de Chile con el uso de algoritmos de Machine Learning". El estudio evaluó siete algoritmos de regresión: Regresión Lineal, Árbol de Decisión, Random Forest, Máquina de Soporte Vectorial (SVM), K-Nearest Neighbor, Red Neuronal Artificial y Kernel Smoothing Regression. Todo aplicado sobre datos de Portalinmobiliario.com para ocho categorías de inmuebles, incluyendo arriendo-comercial y venta-comercial en la Región Metropolitana. La variable dependiente fue normalizada con logaritmo natural (la misma transformación adoptada en el modelo MCO de la presente investigación). La conclusión central fue que Random Forest resultó ser el algoritmo más eficiente en la predicción de precios de arriendo de oficinas (reducción del MAE en 0,0097 unidades logarítmicas respecto al modelo base de la empresa) y de departamentos (-0,0034). Este resultado anticipa la recomendación del presente estudio de implementar modelos de ensamble en etapas futuras de la investigación.

Desde el ámbito de la economía urbana, el estudio "Determinantes urbanos del precio de la vivienda en Chile: Una exploración estadística" (Revista EURE, 2022) analizó 2.047 viviendas nuevas en venta en ciudades chilenas durante 2021, combinando un análisis de componentes principales (PCA) sobre 37 indicadores comunales (SIEDU) con regresión

lineal múltiple. Los resultados identificaron la autosuficiencia financiera comunal, la segregación residencial y el déficit habitacional como los factores de mayor incidencia en el precio de la vivienda a nivel territorial. Aunque centrado en el segmento residencial, este estudio valida metodológicamente el uso del Análisis de Componentes Principales (PCA) para resumir dimensiones territoriales complejas (estrategia también adoptada en el presente trabajo para condensar las variables de geolocalización de baja correlación individual con el precio).

3.5.2 Plataformas Tecnológicas de Valorización Inmobiliaria en Chile

En el plano empresarial, el ecosistema PropTech chileno ha desarrollado en los últimos años herramientas de valorización automática (modelos de tasación automatizada, conocidos como AVM por sus siglas en inglés) que hacen uso explícito de Machine Learning y Big Data. Sin embargo, estas plataformas se han concentrado casi exclusivamente en el segmento residencial, dejando el mercado de locales comerciales sin cobertura de herramientas de predicción automatizadas, especialmente en ciudades fuera de la Región Metropolitana.

Propiteq, plataforma tecnológica del grupo Transsa (firma con más de 35 años de trayectoria en tasaciones oficiales en Chile), implementa modelos de ML entrenados con datos del Servicio de Impuestos Internos (SII), Conservadores de Bienes Raíces y portales inmobiliarios para ofrecer tasaciones online de propiedades habitacionales en segundos. Al cierre de 2023, Propiteq reportó más de 45.000 tasaciones online realizadas y más de 15.500 usuarios registrados (Propiteq, LinkedIn, 2024). El sistema integra análisis de comparable de mercado, histórico de transacciones y proyecciones de plusvalía, actualizándose continuamente con cada consulta. La empresa reconoce explícitamente que su modelo AVM está calibrado para propiedades habitacionales como casas y departamentos y no cubre inmuebles comerciales, evidenciando la brecha que la presente investigación busca llenar para el segmento de locales en ciudades intermedias.

TocToc, portal perteneciente al grupo iCorp junto a ComunidadFeliz e IConstruye, ofrece un tasador de propiedades online gratuito que utiliza un algoritmo de estimación basado en ofertas y transacciones comparables en el entorno geográfico inmediato, con datos de los Conservadores de Bienes Raíces de todo Chile. Al igual que Propiteq, la cobertura actual del tasador de TocToc se limita a propiedades habitacionales (casas y departamentos), informando al usuario cuando no existen datos suficientes en el sector para una estimación confiable (TocToc, 2025).

Datainmobiliaria.cl representa otro actor relevante del ecosistema: una plataforma 100% gratuita para el público general que agrega más de 9 millones de propiedades con 15 años de histórico de transacciones del Conservador de Bienes Raíces y avalúos del SII. A partir de 2025, la plataforma incorporó un conector de IA que permite realizar consultas en lenguaje natural sobre el mercado inmobiliario que incluye análisis de tasaciones, tendencias y comparaciones mediante integración con modelos de lenguaje como Claude y ChatGPT (Datainmobiliaria.cl, 2025). Esta innovación democratiza el acceso a análisis de datos inmobiliarios, pero continúa orientada principalmente al mercado residencial.

3.5.3 Inteligencia de Mercado Comercial: GPS Property e Inciti

En el segmento de inmuebles comerciales, el actor de mayor relevancia en inteligencia de mercado es GPS Property (Global Property Solutions), consultora inmobiliaria con más de 20 años de trayectoria especializada en oficinas, bodegaje, strip centers, parques industriales y terrenos para desarrollo. Su departamento de estudios publica semestralmente informes de mercado sobre múltiples segmentos (oficinas Clase A/A+, centros de bodegaje Clase I y II, strip centers, multifamily, parques industriales) con indicadores de vacancia, precios, absorción y pipeline de proyectos para la Región Metropolitana. Estos informes son descargables gratuitamente y se han consolidado como referencia para medios especializados, habiendo recibido reconocimiento internacional de la revista Euromoney en los años 2021, 2022 y 2023 (GPS Property, LinkedIn).

En 2025, GPS Property adquirió Inciti, la cual es una plataforma PropTech de inteligencia territorial con más de 15 años de trayectoria. Fue comprada por aproximadamente US\$500.000, con el objetivo de crear un hub de soluciones de datos territoriales que incorpore análisis geoespacial, imágenes satelitales y algoritmos predictivos para los segmentos comerciales (El Diario Inmobiliario, 2025; Inciti, 2025). Inciti mantiene información de expedientes municipales, planes reguladores, recepciones y proyectos en venta y en operación, y su integración con GPS Property proyecta ampliar la cobertura a datos de oficinas, strip centers y centros logísticos para otras regiones de Chile. Este movimiento estratégico confirma el interés creciente de la industria en la digitalización e inteligencia de datos del segmento comercial, aunque el foco actual sigue siendo principalmente la Región Metropolitana.

En síntesis, el ecosistema chileno de inteligencia inmobiliaria basada en datos ha avanzado significativamente en el segmento residencial con herramientas de AVM masivo como Propiteq y TocToc, así como en el análisis de mercados comerciales corporativos de la

Región Metropolitana trabajados con GPS Property e Inciti. Sin embargo, el mercado de locales comerciales de pequeño y mediano tamaño en ciudades intermedias como Viña del Mar permanece sin cobertura por parte de estas plataformas. El presente estudio, al desarrollar un modelo predictivo de precios de arriendo para este segmento específico y ponerlo a disposición como herramienta de orientación, llena una brecha concreta en el ecosistema de información inmobiliaria de la Región de Valparaíso.

IV. Metodología

El presente capítulo describe en detalle el diseño metodológico de la investigación, incluyendo la actualización de la base de datos, la incorporación de variables de geolocalización, y la especificación del nuevo modelo MCO hedónico log-lineal implementado como complemento a los modelos de Machine Learning.

4.1 Diseño de la Investigación

La investigación adopta un enfoque cuantitativo de alcance correlacional-causal, fundamentado en la econometría aplicada y el aprendizaje automático supervisado. El diseño es transversal: los datos capturan las condiciones del mercado de oferta en el período comprendido entre el primer semestre de 2025 y el primer semestre de 2026. La unidad de análisis es la publicación activa de arriendo de un local comercial en la comuna de Viña del Mar.

4.2 Recolección de Datos

4.2.1 Fuentes de Información

La captura de datos se realizó de manera manual mediante análisis directo de tres plataformas: Portalinmobiliario.cl, Yapo.cl y Chileinmobiliaria. Se optó por el levantamiento manual en los primeros 300 datos lugar de web scraping para garantizar calidad en la evaluación del estado de conservación y la geolocalización precisa de cada inmueble. Sin embargo, en los últimos 200 datos se optó por realizar web scraping para tener más variedad en la base de datos y poder recuperar datos de otra fuente inmobiliaria (Chileinmombiliaria).

4.2.2 Criterios de Inclusión y Exclusión

Se seleccionaron exclusivamente locales comerciales en arriendo activo, ubicados en Viña del Mar, con información suficiente para las variables de precio, superficie y sector. Fueron excluidas propiedades residenciales, terrenos industriales sin edificación comercial, bodegas sin atención de público, y publicaciones con precios fuera del rango de mercado.

4.2.3 Proceso de Registro y Normalización

Para cada publicación se registraron las variables definidas en la Tabla 1 siguiendo un protocolo estandarizado. Los precios en UF fueron convertidos a CLP al valor oficial vigente a la fecha de levantamiento. Los gastos comunes informados se sumaron al arriendo para construir la variable "Precio + GC", dado que había datos con gastos comunes = 0 y generaba problemas en los datos por la alta dispersión provocada. El estado de conservación fue evaluado por el investigador a partir de fotografías y descripción textual.

4.3 Actualización y Expansión de la Base de Datos

Esta sección documenta los pasos metodológicos realizados para construir la base de datos actualizada y robustecida.

4.3.1 Ampliación Muestral: de 300 a 500 Observaciones

La base de datos fue ampliada de 300 a 500 observaciones mediante una segunda ronda de levantamiento, siguiendo los mismos criterios de inclusión y exclusión de la sección 4.2.2. Las 200 nuevas observaciones fueron recopiladas entre los mismos 13 sectores de la comuna de Viña del Mar, garantizando la continuidad del marco muestral. El tamaño muestral final de $n=500$ permite una validación cruzada más robusta.

4.3.2 Integración de Variables de Geolocalización

La principal innovación metodológica de esta etapa es la incorporación de variables continuas y categóricas de geolocalización para cada observación. Esta información fue obtenida mediante georreferenciación de cada dirección publicada a través de la API de Google Maps, y el cálculo posterior de distancias y scores de accesibilidad. Las nuevas variables geoespaciales incorporadas son las siguientes:

- Latitud y longitud: coordenadas geográficas de la localización exacta del local.
- Distancia al centro de Viña del Mar (km): distancia en línea recta desde el local hasta el centro del área central comercial de la ciudad (Plaza Vergara).
- Nodo comercial más cercano: el polo comercial de mayor relevancia dentro del radio de influencia del local, codificado en 10 categorías (Av. Libertad / Poniente, Arlegui, Av. Valparaíso, Plaza Vergara, Reñaca Centro, San Martín, Reñaca Norte / Eluchans, Rodoviario, El Salto / Limache, Mall Marina). Los nodos comerciales se definieron a través de conversaciones informales/casuales realizadas entre personas cercanas para llegar a las categorías definidas.
- Distancia al nodo comercial más cercano (km): distancia al nodo asignado en línea recta.

- Score de accesibilidad (0-100): indicador continuo que pondera la cercanía a ejes viales principales, acceso a transporte público y conectividad peatonal del entorno, con valores altos indicando mayor accesibilidad. Un valor 0 quiere decir que está alejado de la accesibilidad pública y del nodo comercial más cercano, pudiéndose llegar únicamente a través de vehículo privado y además ubicado en un lugar marginal. Un valor 100 quiere decir que está ubicado de manera céntrica y con excelente circulación de transporte público.
- Score de flujo peatonal (0-100): estimación del nivel de circulación de personas en el entorno inmediato del local, derivada de la categoría de la calle, la densidad comercial del entorno y la proximidad a generadores de flujo (universidades, hospitales, estaciones).
- Score de visibilidad (0-100): indicador de la exposición del local a los transeúntes, considerando el tipo de fachada, la posición respecto a la acera y el ángulo de visión desde la vía pública. Es subjetivo y el valor debe asignarlo quién esté analizando el lugar.
- Variables categóricas refinadas: *accesibilidad_refinada*, *flujo_peatonal_refinado* y *visibilidad_refinada*, que categorizan los scores continuos en 5-6 niveles cualitativos (Muy alta, Alta, Media-alta, Media, Media-baja, Baja).

Nodo comercial	Justificación
Plaza Vergara	Centro cívico y comercial de Viña del Mar
Avenida Valparaíso	Principal eje comercial tradicional
Avenida Libertad	Corredor comercial consolidado con alto flujo
Mall Marina	Principal polo comercial regional
Reñaca Centro	Centro comercial y turístico
San Martín	Corredor gastronómico y turístico
Rodoviario	Nodo de transporte con actividad comercial
El Salto	Sector industrial-comercial
Reñaca Norte / Eluchans	Sector comercial emergente

4.3.3 Variables Adicionales Incorporadas

Además de las variables geoespaciales, se incorporaron las siguientes variables al modelo:

- Número de baños: variable determinante del tamaño funcional del local y de su nivel de equipamiento. Los locales con mayor número de baños tienden a corresponder a formatos más grandes (bodegas, tiendas de departamentos) o a usos con mayor requerimiento de servicios (hoteles, restaurantes).

La Tabla 1 presenta la matriz completa de variables del modelo actualizado.

Categoría	Variable	Tipo	Descripción
Variable dependiente	Precio + GC (CLP)	Continua	Arriendo mensual más gastos comunes en pesos
Estructural	Superficie (m ²)	Continua	Metros cuadrados totales del local
Estructural	Número de pisos	Discreta	Número de niveles del local
Estructural	Número de baños	Discreta	Baños disponibles en el local
Estructural (derivada)	ln(Superficie)	Continua	Logaritmo natural de la superficie
Localización	Nodo comercial más cercano	Categórica (10 cat.)	Polo comercial de referencia
Localización	Tipo de Local	Categórica (9 cat.)	Tipología: calle comercial, galería, hotel, etc.
Localización	Latitud / Longitud	Continua	Coordenadas geográficas
Localización	Distancia al centro (km)	Continua	Distancia a Plaza Vergara
Localización	Distancia al nodo (km)	Continua	Distancia al polo comercial más cercano
Accesibilidad	Score accesibilidad (0-100)	Continua	Índice de accesibilidad urbana
Accesibilidad	Accesibilidad refinada	Categórica (6 cat.)	Nivel cualitativo de accesibilidad
Flujo peatonal	Score flujo (0-100)	Continua	Índice de flujo peatonal estimado
Flujo peatonal	Flujo peatonal refinado	Categórica (6 cat.)	Nivel cualitativo de flujo
Visibilidad	Score visibilidad (0-100)	Continua	Índice de exposición del local
Visibilidad	Visibilidad refinada	Categórica (5 cat.)	Nivel cualitativo de
Estado	Estado de conservación	Categórica (3 cat.)	Bueno, Regular o Nuevo
Derivada	CLP/m ²	Continua	Precio unitario por metro
Dependiente (transformada)	ln(Precio + GC)	Continua	Logaritmo del precio total (var. dep. del MCO)

Tabla 1. Matriz de definición de variables del modelo.

4.4 Procesamiento de Datos

El procesamiento de datos se ejecutó en Google Colab utilizando las bibliotecas de Python pandas (manejo de datos), NumPy (cálculo numérico), scikit-learn (aprendizaje automático) y statsmodels (estadística). El flujo de procesamiento se estructura en las siguientes etapas:

(1) Lectura de datos: La base de datos fue leída desde el archivo Excel con la biblioteca pandas, preservando los tipos de datos asignados durante el registro.

(2) Construcción de variables transformadas: Se aplicó la transformación logarítmica a la variable dependiente (denominada logaritmo natural del precio + gastos comunes) y a la superficie (el logaritmo natural de la superficie), y la transformación logarítmica suavizada $\log(1+x)$ a las variables de distancia, que evita valores indefinidos cuando la distancia es cero, reduciendo la asimetría de sus distribuciones.

(3) Separación de variables numéricas y categóricas: Las variables numéricas fueron estandarizadas mediante estandarización estadística (transformación z-score: media cero y desviación estándar uno), y las categóricas fueron codificadas mediante codificación binaria por categorías (método OneHot), donde los valores desconocidos se tratan automáticamente como una categoría adicional denominada "No informado", todo ello organizado dentro de un flujo automatizado de procesamiento (pipeline) que aplica cada transformación a su columna correspondiente. Este enfoque evita fugas de datos al calcular los parámetros de escala exclusivamente sobre el conjunto de entrenamiento.

(4) Análisis de componentes principales (PCA) de variables de baja correlación: Las variables numéricas con correlación absoluta con el logaritmo del precio menor a 0,20 (latitud, longitud, distancias, scores de accesibilidad/flujo/visibilidad) fueron condensadas en un único componente principal mediante Análisis de Componentes Principales (PCA), que explica 61% de su varianza conjunta. Este componente resume la "dimensión espacial periférica" de cada local.

(5) División entrenamiento/prueba: Para el modelo MCO, la base procesada se dividió en 80% entrenamiento (n=400) y 20% prueba (n=100) con semilla aleatoria fija. Para los modelos ML se utilizó la división 70% entrenamiento / 30% prueba.

4.5 Especificación y Entrenamiento de los Modelos ML

Los cuatro modelos de ML (Ridge, Lasso, ElasticNet, Árbol de Regresión) fueron entrenados sobre el conjunto de entrenamiento (70%) mediante una búsqueda sistemática de hiperparámetros (Grid Search) combinada con validación cruzada de 5 subconjuntos,

optimizando el Error Cuadrático Medio. Los detalles de hiperparámetros evaluadas se presentan en la Tabla 2.

Los hiperparámetros representan configuraciones definidas antes de entrenar el modelo para controlar como aprende y evitar el sobreajuste.

En este caso, los tres primeros modelos utilizan hiperparámetros que encoge los coeficientes a un valor de 0 o cercano a 0, idealmente para tratar variables que aportan poca significancia o información. Mientras que, el último modelo (árbol de regresión) utiliza hiperparámetros que limitan la complejidad del árbol, asignándole limitaciones a la profundidad de sus ramas para evitar sobreajuste o subajuste, evitando que el árbol sea muy complejo o simple. También se le asigno el número mínimo de datos que deben quedar en el nodo final después de una división con la finalidad de poder suavizar el modelo y asegurar que el valor predicho en cada “hoja” sea un promedio de una cantidad de datos y no un valor aislado poco representativo.

Modelo	Hiperparámetro
Ridge	alpha (λ_2)
Lasso	alpha (λ_1)
ElasticNet	alpha (λ), l1_ratio (ρ)
Árbol de Regresión	max_depth, min_samples_split, min_samples_leaf

Tabla 2. Grillas de hiperparámetros evaluados mediante búsqueda sistemática (Grid Search).

4.6 Especificación del Modelo MCO Hedónico Log-Lineal

Como complemento a los modelos de Machine Learning, se implementó un modelo de Mínimos Cuadrados Ordinarios (MCO) hedónico clásico en especificación log-lineal. Este modelo responde a la tradición de Rosen (1974) y permite la interpretación directa de los coeficientes como elasticidades precio-atributo.

4.6.1 Especificación del Modelo

La variable dependiente es el logaritmo natural del precio total de arriendo: $\ln(\text{precio_gc_clp})$. Las variables independientes incluyen:

- Variables numéricas (estandarizadas): $\ln(\text{superficie_m}^2)$, distancia al centro, distancia al nodo comercial, scores de accesibilidad, flujo peatonal y visibilidad, número de pisos,

número de baños, y el componente principal de las variables geoespaciales de baja correlación con el precio.

- En el modelo inicial: variables categóricas (codificadas con el método binario OneHot): nodo_comercial_mas_cercano (10 categorías), accesibilidad_refinada (6 categorías), flujo_peatonal_refinado (6 categorías), visibilidad_refinada (5 categorías), tipo_local (9 categorías), y estado_conservacion (3 categorías).
- En el modelo final solo se consideraron las variables transformadas en logaritmo en los casos correspondientes para reducir redundancia en el modelo, así como también se eliminó la variable accesibilidad debido al alta multicolinealidad generada por esta variable y su alta relación con variables como visibilidad o flujo peatonal.

El flujo de estimación aplica en paralelo una estandarización estadística (con imputación de mediana para valores faltantes) a las variables numéricas y una codificación binaria OneHot (con imputación de la categoría constante "No informado" para valores faltantes) a las variables categóricas. El modelo de regresión lineal fue estimado con el algoritmo de mínimos cuadrados ordinarios estándar, equivalente al MCO clásico.

4.6.2 División Train/Test

La base de 500 observaciones fue dividida en 80% entrenamiento (n=400) y 20% prueba (n=100) con semilla aleatoria fija (valor 42), garantizando la reproducibilidad de los resultados y comparabilidad entre iteraciones del modelo.

4.6.3 Métricas de Evaluación

El modelo MCO log-lineal fue evaluado en escala logarítmica mediante el R^2 (coeficiente de determinación en la variable transformada), la RMSE y el MAPE. Se complementó con diagnósticos estadísticos estándar: test de Breusch-Pagan para heterocedasticidad y análisis de la matriz de correlación para multicolinealidad.

4.7 Creación de un sistema funcional abierto al público

Finalmente, se creó una página web donde se prueba el modelo de una manera más “palpable”. Se utilizó Netlify como host del modelo, donde a través de archivos JSON y JavaScript se configuró para darle forma al prototipo. Este considera las variables más significativas y un sistema de geolocalización preciso, donde solo al escribir la dirección exacta de manera automática asigna valores de entorno urbano como son los scores de flujo peatonal, accesibilidad y visibilidad. La variable accesibilidad es solo informativa, ya que como se mencionó anteriormente, no se utilizó como variable en el modelo final. Mientras

que, la variable visibilidad se debe ingresar manual ya que, se tomó como una variable subjetiva. En esencia mide la distancia del local a la calzada, pero el usuario también podría considerar otros factores subjetivos (colores llamativos, belleza, diseño, etc), es por esto por lo que se debe ingresar de manera manual. También le asigna valores a las variables de localización como lo es el nombre del nodo comercial más cercano (categórica), distancia al centro de viña del mar en kilómetros y distancia al nodo comercial más cercano en kilómetros.

También se consideraron las variables relacionadas a infraestructura, como superficie (en su forma logarítmica), cantidad de baños y pisos. Así como, las variables categóricas como lo es estado de conservación y tipo de local, sin embargo, cada una de las mencionadas en el presente párrafo deben ingresarse de manera manual, no como las mencionadas en el párrafo anterior. Cada variable considerada en el prototipo tuvo valores de coeficientes altos para que tuviese mayor impacto en el precio estimado final de arriendo.

La página web tiene por enlace: "[Estimador de Arriendo Comercial - Viña del Mar](#)".

V. Base de Datos y Estadística Descriptiva

5.1 Descripción General del Dataset

La base de datos final consolidada comprende 500 observaciones válidas, correspondientes a 500 locales comerciales únicos en arriendo en la comuna de Viña del Mar durante el período 2025-2026, lo que representa un incremento del 66,7% respecto a la versión original de 300 observaciones. La base incluye 19 variables: 1 variable dependiente, 6 variables estructurales, 6 variables de geolocalización continua, 4 variables de accesibilidad-flujo-visibilidad (2 scores continuos + 2 refinadas categóricas para cada dimensión) y 2 variables cualitativas de estado y tipo.

La Tabla 3 presenta las estadísticas descriptivas de las variables continuas del modelo actualizado.

Variable	N	Media	Mediana	D.E.	Mín.	Máy.
Precio + GC (CLP)	500	\$3.850.355	\$2.584.100	\$5.143.770	\$20.403	\$44.375.038
Superficie (m ²)	500	219,0	120,0	427,0	5,0	6.311,0
Número de pisos	500	1,4	1,0	0,8	1	4
CLP/m ²	500	\$20.936	\$19.270	\$10.718	\$89	\$70.466
Número de baños	500	2,4	2,0	1,6	1	20
Distancia al centro (km)	500	2,1	1,1	2,3	0,1	9,2
Distancia al nodo (km)	500	0,5	0,3	0,6	0,0	3,6
Score accesibilidad (0-100)	500	74,0	82,0	23,0	0,0	99,0
Score flujo peatonal (0-100)	500	79,0	85,0	18,0	0,0	100,0
Score visibilidad (0-100)	500	66,0	68,0	11,0	28,0	92,0

Tabla 3. Estadística descriptiva de variables continuas (n=500).

La variable dependiente Precio + GC mantiene una distribución con poca asimetría positiva. La media (\$3.850.355) supera a la mediana (\$2.584.100), con un coeficiente de variación del 133,6%, consistente con la alta heterogeneidad del mercado. La distribución es más dispersa que en la versión original (300 obs.), probablemente por la inclusión de locales de mayor superficie y de nuevas tipologías como hoteles y bodegas comerciales de gran escala.

Los scores de accesibilidad, flujo peatonal y visibilidad presentan distribuciones diferenciadas. La accesibilidad media (74/100) y el flujo peatonal medio (79/100) son relativamente altos, consistente con que la muestra está dominada por locales en zonas comerciales bien conectadas (Av. Libertad/Poniente y Arlegui concentran el 39% de la

muestra). La visibilidad tiene una media más baja (66/100) y menor dispersión, reflejando que la mayoría de los locales en Viña del Mar tienen visibilidad media-alta desde la vía pública.

5.2 Distribución por Nodo Comercial

La Tabla 4 presenta la distribución de observaciones y precios medios según el nodo comercial de referencia, que es la nueva variable de localización que reemplaza con mayor precisión la antigua categorización por sector/barrio.

Nodo Comercial	n	Precio Medio (CLP)	Precio Mediana (CLP)	% del total
Av. Libertad / Poniente	117	\$4.891.160	\$3.386.958	23,4%
Arlegui	78	\$3.842.524	\$2.756.000	15,6%
Av. Valparaíso	69	\$4.807.820	\$2.461.818	13,8%
Plaza Vergara	55	\$3.576.535	\$2.734.050	11,0%
Reñaca Centro	45	\$2.578.131	\$1.389.000	9,0%
San Martín	42	\$3.249.555	\$2.576.218	8,4%
Rodoviario	32	\$3.674.367	\$1.922.347	6,4%
Reñaca Norte / Eluchans	32	\$2.367.944	\$1.620.590	6,4%
El Salto / Limache	17	\$2.332.208	\$1.245.000	3,4%
Mall Marina	13	\$3.019.014	\$1.855.680	2,6%

Tabla 4. Distribución de observaciones y precios por nodo comercial más cercano (n=500).

Av. Libertad / Poniente es el nodo de mayor representación (23,4%) y el de mayor precio medio (\$4.891.160), reflejando la consolidación del sector Poniente como polo comercial dinámico. Av. Valparaíso concentra el 13,8% de la muestra con un precio medio también elevado (\$4.807.820), consistente con su rol como eje comercial histórico de la ciudad. Los nodos periféricos (El Salto/Limache, Reñaca Norte/Eluchans) registran los precios medios más bajos, en línea con las predicciones del modelo monocéntrico de Alonso.

5.3 Distribución por Tipo de Local

La Tabla 5 presenta la distribución actualizada por tipo de local. La nueva muestra incorpora las categorías Hotel (n=7) y Oficina (n=103), esta última constituye la segunda categoría más frecuente, reflejando la significativa presencia de espacios de trabajo en el mercado de locales de Viña del Mar.

Tipo de Local	n	Precio Medio (CLP)	Sup. Media (m²)	% del total
Calle comercial	254	\$4.200.000	220,0	50,8%
Oficina	103	\$2.900.000	130,0	20,6%
Strip center	51	\$2.800.000	160,0	10,2%
Galería	41	\$1.300.000	80,0	8,2%
Bodega comercial	22	\$5.800.000	480,0	4,4%
Mall / Paseo	19	\$900.000	70,0	3,8%
Hotel	7	\$6.500.000	350,0	1,4%
Subterráneo	3	\$1.000.000	65,0	0,6%

Tabla 5. Distribución de observaciones y precios por tipo de local (n=500). indica categorías incorporadas en la versión ampliada.

5.4 Distribución por Estado de Conservación

El 56,6% de los locales fue evaluado en estado "Bueno", el 24,0% en estado "Nuevo" y el 19,4% en estado "Regular". A diferencia de la versión original, la proporción de locales Nuevos aumentó del 19,3% al 24,0%, posiblemente reflejando la expansión muestral hacia zonas con mayor proporción de desarrollos recientes (Mall Marina, Rodoviario).

5.5 Análisis de Correlación con el Precio

La Tabla 6 presenta las correlaciones de las variables numéricas con el precio, incluyendo las nuevas variables de geolocalización. Un hallazgo destacado es que las nuevas variables de accesibilidad y flujo peatonal presentan correlaciones positivas con el precio (0,130 y 0,137 respectivamente), mientras que la distancia al centro muestra correlación negativa (-0,128), en línea con las predicciones del modelo monocéntrico de Alonso. Esto valida parcialmente la incorporación de estas variables como predictores del precio de arriendo.

Variable	Correlación con Precio + GC (logaritmo)
Log Superficie (m²)	0,860
Número de baños	0,490
Número de pisos	0,460
Score visibilidad	0,190
Score flujo peatonal	0,170
Log Distancia al nodo (km)	0.07
Distancia al centro (km)	0.14

Tabla 6. Correlaciones de variables numéricas con el precio total de arriendo (n=500). indica variables incorporadas en la versión ampliada.

La correlación más alta continúa siendo la de superficie ($r=0.86$), aunque es algo menor que en la base original ($r=0.906$), lo cual puede reflejar que la mayor variabilidad de tipologías en la muestra ampliada (Oficinas, Hoteles) introduce locales de precio relativamente alto pero menor superficie. La correlación positiva del score de flujo peatonal (0,17) y del score de visibilidad (0,19) confirman la relevancia de estas nuevas dimensiones para la determinación del precio.

VI. Modelos MCO Hedónico Log-Lineal

El presente capítulo presenta los resultados del modelo de Mínimos Cuadrados Ordinarios (MCO) hedónico log-lineal estimado sobre la base de datos ampliada de 500 observaciones. Este modelo complementa los modelos de Machine Learning presentados en el Capítulo VII, proveyendo una interpretación paramétrica y estructuralmente fundamentada en la Teoría de los Precios Hedónicos de Rosen (1974).

6.1 Resultados del primer Modelo MCO

La Tabla 7 presenta las métricas de rendimiento del modelo MCO log-lineal en el conjunto de prueba ($n=100$). Este fue un proceso de entrenamiento y prueba que se repitió de manera cíclica y aleatoria (tomando muestras de la población total de manera aleatoria para realizar el entrenamiento y la prueba) 1000 veces, donde el mejor resultado fue el presentado en la siguiente tabla.

Métrica	Valor (escala logarítmica)
R^2 (coeficiente de determinación en $\ln(\text{precio})$)	0,7659
RMSE	0,5446
MAE	0,4079
MAPE	2,78%
n entrenamiento	400
n prueba	100
División	80% / 20%

Tabla 7. Métricas de rendimiento del modelo MCO hedónico log-lineal (escala logarítmica, $n=100$ en prueba).

Un R^2 de 0,7659 en la escala logarítmica indica que el modelo MCO explica el 76,6% de la varianza del logaritmo del precio de arriendo en observaciones fuera de la muestra de entrenamiento. Es importante notar que esta métrica es en escala logarítmica y no es directamente comparable con el $R^2=0,927$ del árbol de regresión, que opera en escala original (CLP).

6.2 Coeficientes e Interpretación Económica

La Tabla 8 presenta los 15 coeficientes de mayor valor absoluto en el modelo MCO, ordenados por impacto sobre el logaritmo del precio.

Variable	Coeficiente	Interpretación económica
ln(superficie_m²)	0,784	Valor alto esperado, históricamente siempre ha sido la variable más determinante con diferencia.
flujo_peatonal_refinado_Baja	-0,542	Locales con flujo peatonal bajo tienen precios aproximadamente 50% menores
tipo_local_Oficina	-0,40	Oficinas tienen precios aproximadamente 30% menores que locales de calle comercial base
tipo_local_Hotel	0,392	Hoteles registran precios 50% superiores al tipo base, consistente con su especialización
flujo_peatonal_refinado_Muy alta	0,331	Flujo peatonal muy alto incrementa el precio en aproximadamente 50% respecto a la categoría base
ipo_local_Subterráneo	0,315	Valor inesperado, puede ser el reflejo de una baja muestra (n=3)
tipo_local_Galería	-0.283	Galerías presentan precios disminuidos a comparación de la categoría base
tipo_local_Strip Center	0,2486	Strip centers registran precios 37% superiores al tipo base, por mayor visibilidad exterior

Tabla 8. Top 9 coeficientes del primer modelo MCO hedónico log-lineal (ordenados por valor absoluto).

El coeficiente más informativo desde el punto de vista económico es el del logaritmo de la superficie (0,784), que se interpreta directamente como la elasticidad precio-superficie. Esta elasticidad menor que 1 confirma la existencia de economías de escala: los locales más grandes son relativamente más baratos por metro cuadrado que los locales pequeños.

El coeficiente del flujo peatonal bajo (-0,54) es el más intenso negativamente y tiene plena coherencia económica: los locales ubicados en zonas de bajo tráfico de personas tienen precios significativamente menores, lo cual es consistente con la teoría de Space Syntax y con la lógica del modelo de Alonso. Este resultado es particularmente valioso porque representa el primer hallazgo de la investigación donde una variable de accesibilidad/flujo

aparece como determinante económicamente significativo del precio, algo que los modelos originales no lograban capturar al no contar con estas variables.

6.3 Resultados del Modelo final MCO

Debido a una serie de descubrimientos y consideraciones del modelo inicial, se fue puliendo en tema de variables, heterocedasticidad, multicolinealidad, normalidad de residuos y redundancia del modelo.

Lo primero que se realizó fue eliminar del modelo variables que anteriormente fueron duplicadas a logaritmo, dejando de manera funcional solo su versión logarítmica (distancias, superficie). También se descartaron variables que se relacionaban mucho y que explicaban en gran medida lo mismo, como pasó con la accesibilidad (en su forma se score y refinada) o también todas las formas refinadas de las variables de espacialidad. Esto con la finalidad de borrar la mayor multicolinealidad posible y poder demostrar un modelo mucho más eficiente u óptimo. Luego se hizo una limpieza de datos que se alejaban en gran medida de la normalidad, específicamente aquellos datos donde los residuos estuviesen fuera de [-1.6, 1.6].

Finalmente, el resultado de todo este proceso buscando pulir el modelo quedó de la siguiente manera:

Métrica	Valor (escala logarítmica)
R ² (coeficiente de determinación en ln(precio))	0,8200
RMSE logarítmica	0,4241
MAE logarítmica	0,3308
MAPE logarítmica	2,28%
n entrenamiento	396
n prueba	100
División	80% / 20%

Tabla 7.1. Métricas de rendimiento del modelo final MCO hedónico log-lineal (escala logarítmica, n=100 en prueba).

6.3 Diagnóstico Estadístico

6.3.1 Heterocedasticidad

El test de Breusch-Pagan sobre los residuos del modelo final MCO log-lineal produce: estadístico LM = 33,74, p-valor = 0,8430; estadístico F = 0,973, p-valor = 0,5228. No se rechaza la hipótesis nula de varianza constante de los residuos, donde no se está evidenciando heterocedasticidad estadísticamente significativa en el modelo limpio. La transformación logarítmica de la variable dependiente, que buscaba reducir la heterocedasticidad del precio en niveles, la limpieza de datos y además el descarte de las variables redundantes fue suficiente para eliminarla en gran medida. Al no borrarla completamente, esto sugiere que la varianza de los errores del modelo en escala logarítmica sigue siendo función de las variables independientes, posiblemente por la presencia de subgrupos de locales (hoteles, bodegas de gran escala) con características muy distintas al resto de la muestra.

6.3.2 Multicolinealidad

El primer análisis de correlación entre las variables numéricas revela pares de alta correlación que generan multicolinealidad severa en el modelo MCO: la distancia al centro y la latitud tienen $r=0,92$; el score de accesibilidad y la distancia al centro tienen $r=-0,94$; el score de flujo peatonal y el score de accesibilidad tienen $r=0,88$; y la log-distancia al centro y el score de accesibilidad tienen $r=-0,93$. Esta multicolinealidad entre las variables de geolocalización y accesibilidad explica los signos contra-intuitivos observados en algunos coeficientes de la Tabla 8 (accesibilidad_refinada: Alta y Media-alta con coeficientes negativos). La estrategia de PCA implementada para condensar las variables de baja correlación con el precio reduce parcialmente este problema, pero no elimina la colinealidad entre los scores continuos de accesibilidad y flujo peatonal que permanecen en el modelo.

Finalmente, al realizar los cambios y transformaciones en el modelo, se logró reducir la multicolinealidad y así generar un modelo efectivo que explique de manera funcional y buscar ser más exacto en buscar explicar el precio de los arriendos comerciales en Viña del Mar.

6.4 Análisis con Componente Principal (PCA)

Como extensión del modelo base, se aplicó el Análisis de Componentes Principales (PCA, técnica estadística que condensa múltiples variables en factores resumen) sobre las 9 variables numéricas con correlación absoluta con $\ln(\text{precio})$ menor a 0,20 (latitud, longitud, distancias, scores). El primer componente principal captura el 61% de la varianza conjunta de estas variables, representando una dimensión de "localización periférica": valores altos del componente corresponden a locales alejados del centro y con menor accesibilidad. Al sustituir las 9 variables individuales por este componente, el modelo MCO mantiene métricas muy similares ($R^2_{\log} = 0,773$, $RMSE_{\log} = 0,536$), con una ligera reducción del R^2 de 0,006 puntos. Esto confirma que la información predictiva de estas variables geoespaciales puede ser condensada eficientemente en un único factor, lo que reduce la multicolinealidad y simplifica la especificación del modelo.

Modelo MCO	R^2 (log)	RMSE (log)	MAE (log)	MAPE (log)
MCO original (todas las variables numéricas)	0,765	0,530	0,407	2,78%
MCO final limpio	0,820	0,424	0,330	2,28%
MCO con PCA sustituyendo variables originales	0,773	0,536	0,398	2,72%

Tabla 9. Comparación de variantes del modelo MCO log-lineal en escala logarítmica.

VII. Primeros Modelos ML: Diagnóstico y Correcciones

El presente capítulo documenta el proceso iterativo de modelización con los cuatro algoritmos de Machine Learning, incluyendo los errores estadísticos detectados en los modelos iniciales y las correcciones aplicadas.

7.1 Resultados Iniciales: Comparación de los Cuatro Modelos

Modelo	Mejor parámetro	R ² Test	RMSE Test (CLP)	MAPE Test	CV R ² (5f)
Ridge	$\alpha = 10$	0,877	\$1.446.653	45,64%	0,713
Lasso	$\alpha = 10$	0,885	\$1.394.789	54,38%	0,682
ElasticNet	$\alpha=0,1$; $\rho=0,7$	0,881	\$1.422.675	48,61%	0,720
Árbol de Regresión	depth=None; leaf=1	0,922	\$1.150.107	26,09%	0,500

Tabla 10. Comparación de modelos ML en conjunto de prueba (n=90).

7.2 Diagnóstico de los Modelos Lineales Regularizados

7.2.1 Heterocedasticidad

El test de Breusch-Pagan sobre los residuos del modelo ElasticNet (representativo de los modelos lineales) produce: estadístico LM = 39,92, p-valor = 0,005; estadístico F = 2,22, p-valor = 0,003. Se rechaza la hipótesis nula de varianza constante de los residuos, evidenciando heterocedasticidad estadísticamente significativa. Los errores estándar de los coeficientes están sesgados, comprometiendo la inferencia estadística.

7.2.2 No Normalidad de los Residuos

El test de Shapiro-Wilk sobre los residuos del modelo ElasticNet produce: estadístico W = 0,814, p-valor < 0,001. Se rechaza la hipótesis nula de normalidad. Los residuos presentan asimetría positiva (coeficiente = 1,17) y curtosis extrema (exceso = 11,12), reflejando la presencia de colas pesadas.

7.2.3 Multicolinealidad (VIF)

Variable	VIF	Nivel
ln(CLP/metros cuadrados)	11,45	Alto (>10)
Pisos	10,47	Alto (>10)
Estado_Bueno	3,71	Moderado
Superficie	2,85	Aceptable
Sector/Barrio_Poniente	2,38	Aceptable
Demás variables dummy	< 2,3	Bajo

Tabla 11. Factores de Inflación de Varianza (VIF) del modelo ElasticNet.

7.3 Diagnóstico del Árbol de Regresión: Sobreajuste

El Árbol de Regresión sin restricciones exhibe sobreajuste severo en su configuración óptima por la búsqueda de hiperparámetros (sin restricción de profundidad máxima ni mínimo de muestras por hoja): $R^2=1,000$ en entrenamiento. La Tabla 12 muestra el análisis del equilibrio entre sobreajuste y generalización según profundidad.

Profundidad	R^2 Train	R^2 Test	RMSE Test (CLP)	MAPE Test	CV R^2 (5f)
2	0,848	0,764	\$2.000.805	72,24%	0,021
3	0,918	0,880	\$1.425.657	53,23%	0,258
5	0,987	0,897	\$1.321.234	42,97%	0,538
7 ✓	0,998	0,927	\$1.110.773	27,40%	0,637
8	0,999	0,937	\$1.030.403	25,09%	0,556
10	1,000	0,890	\$1.363.628	25,01%	0,559
Sin límite	1,000	0,911	\$1.226.585	24,45%	0,617

Tabla 12. Análisis del equilibrio entre sobreajuste y generalización según profundidad máxima del árbol.

La profundidad óptima se ubica en 7, con el mayor CV R^2 (0,637) y un R^2 de prueba de 0,927. A partir de profundidad 8, el CV R^2 comienza a deteriorarse pese a que el R^2 de prueba mejora levemente, señal de sobreajuste progresivo.

VIII. Modelo ML: Árbol de Regresión (max_depth = 7)

8.1 Justificación de la Selección del Modelo

La selección del Árbol de Regresión con profundidad máxima 7 como modelo ML final se fundamenta en: (i) desempeño superior en el conjunto de prueba ($R^2=0,927$, MAPE=27,4%); (ii) ausencia de los supuestos paramétricos violados en los modelos lineales; (iii) mejor capacidad de validación cruzada entre modelos con buen rendimiento en prueba (CV $R^2=0,637$); y (iv) interpretabilidad jerárquica.

8.2 Importancia de Variables

Variable	Importancia	% explicado
Superficie (m ²)	0,9043	90,4%
ln(CLP/metro cuadrado)	0,0957	9,6%
Resto de variables (18 dummies)	< 0,001	< 0,1%
TOTAL	1,000	100%

Tabla 13. Importancia de variables en el Árbol de Regresión (max_depth=7).

8.3 Rendimiento del Modelo Final

Métrica	Entrenamiento (n=210)	Prueba (n=90)
R^2	0,9982	0,9272
RMSE (CLP)	< \$1	\$1.110.773
MAPE	< 0,01%	27,4%
CV R^2 promedio (5-fold)	N/A	0,637
Profundidad / Hojas	7 / 38	7 / 38

Tabla 14. Métricas de rendimiento del modelo final (Árbol de Regresión, max_depth=7).

8.4 Ejemplos de Predicciones del Modelo

Obs.	Precio Real (CLP)	Precio Predicho (CLP)	Error Absoluto	Error %
1	\$1.944.000	\$1.736.000	-\$208.000	-10,7%
2	\$3.671.018	\$3.469.314	-\$201.704	-5,5%
3	\$3.255.000	\$3.307.950	+\$52.950	+1,6%
4	\$2.422.397	\$2.825.037	+\$402.640	+16,6%
5	\$7.008.355	\$7.901.618	+\$893.263	+12,7%

Tabla 15. Ejemplos de predicciones del modelo en el conjunto de prueba.

IX. Análisis de Resultados e Interpretación Económica

9.1 El Precio Total está Determinado por la Superficie

El hallazgo central del presente estudio es la dominancia de la superficie del local en la determinación del precio total de arriendo. Este resultado es consistente en ambas aproximaciones metodológicas: en el modelo MCO log-lineal, el coeficiente de $\ln(\text{superficie})$ es 0,784 (elasticidad precio-superficie); en el árbol de regresión, la superficie explica el 90,4% de la varianza del precio. La Teoría de los Precios Hedónicos de Rosen (1974) provee el fundamento teórico: el mercado asigna un precio implícito positivo y dominante a la capacidad de albergar actividad comercial, que es directamente proporcional al espacio disponible.

9.2 El Flujo Peatonal como Nuevo Determinante del Precio Unitario

Un resultado económico novedoso que emerge de la versión actualizada del modelo es la relevancia del flujo peatonal como determinante del precio de arriendo. En el modelo MCO log-lineal, el coeficiente de la variable `flujo_peatonal_refinado: Baja` es -0.793, el de mayor magnitud en el modelo después del intercepto. Esto implica que, al controlar por superficie, accesibilidad, sector y tipo de local, un local con flujo peatonal bajo tiene un precio un 50% menor aproximadamente que uno con flujo peatonal de referencia. Simétricamente, el coeficiente de `flujo_peatonal_refinado: Muy Alta` es +0,2815, indicando una prima superior para locales con altísimo flujo de personas.

Este hallazgo es consistente con la teoría de Space Syntax (Hillier y Hanson, 1984) y con los resultados de Netzell (2009) en Estocolmo, donde el valor de integración, entendido como una medida de accesibilidad topológica de la red de calles, explicaba entre el 8% y el 12% de la varianza del precio unitario. En el presente estudio, la variable de flujo peatonal actúa como un proxy operacional del concepto de integración, enriqueciendo la explicación del componente de localización que en los modelos anteriores quedaba capturado sólo mediante la variable categórica de tipo de local y sector.

9.3 Elasticidad Precio-Superficie: Economías de Escala

La elasticidad estimada por el modelo MCO (0,868) confirma la existencia de economías de escala en el mercado de arriendos comerciales de Viña del Mar: un incremento del 10% en la superficie se asocia con un incremento de solo el 8,68% en el precio total, lo que implica una caída del precio por metro cuadrado a medida que aumenta el tamaño del local. Esta elasticidad es consistente con la correlación negativa entre superficie y CLP/m² observada

en la estadística descriptiva ($r=-0,189$ en la versión original), y es coherente con la lógica de que los locales grandes son arrendados por negocios con mayor capacidad de negociación o con menor disposición marginal a pagar por metros adicionales.

9.4 Diferencias entre Tipos de Local y Nodos Comerciales

El análisis descriptivo por nodo comercial muestra que Av. Libertad/Poniente (\$4.891.160 de precio medio) y Av. Valparaíso (\$4.807.820) lideran en precio absoluto, mientras que los nodos periféricos como El Salto/Limache (\$2.332.208) y Reñaca Norte/Eluchans (\$2.367.944) registran los precios más bajos. En el modelo MCO, el nodo Reñaca Norte/Eluchans muestra un coeficiente positivo de 0,379, lo que indica que, al controlar por superficie y accesibilidad, los locales en este nodo tienen precios superiores que la categoría de referencia.

El coeficiente positivo del tipo Hotel (+0,207) y el negativo de Oficina (-0,266) revelan diferencias funcionales relevantes: los hoteles, por sus mayores requerimientos de espacio y equipamiento, justifican precios significativamente superiores, mientras que las oficinas operan con una lógica de mercado diferente al comercio minorista, con precios unitarios más bajos.

9.5 Estado de Conservación: Hallazgo Confirmado

Consistente con los resultados originales, el estado de conservación del local tampoco aparece como determinante significativo en el modelo MCO log-lineal. Este hallazgo confirma que el mercado de locales comerciales en Viña del Mar no discrimina sistemáticamente en precio por la calidad física del inmueble cuando se controla por superficie, localización y tipo de local. Las posibles explicaciones son: (i) que los arrendatarios comerciales incorporan el costo de la adaptación en su propia cuenta; (ii) que la presión de demanda en ciertos sectores permite obtener precios similares independientemente del estado; y (iii) que la evaluación cualitativa del estado es en parte subjetiva y no captura con precisión la percepción del mercado.

9.6 Implicancias Prácticas

Para emprendedores en proceso de instalación: La superficie continúa siendo el determinante más importante del costo de arriendo. Sin embargo, la nueva evidencia sobre flujo peatonal añade un criterio adicional relevante: locales con alto flujo peatonal justifican precios significativamente mayores, lo que puede traducirse en mayores ventas para negocios de retail. Un emprendedor debe evaluar el precio de arriendo no solo como función

del espacio, sino del "costo por cliente potencial expuesto", que es lo que el score de flujo peatonal aproxima.

Para propietarios de locales: Los scores de accesibilidad y flujo peatonal proveen argumentos cuantitativos para sustentar precios de arriendo superiores al promedio de mercado por m² en locales bien ubicados. Un propietario puede ahora comunicar objetivamente la prima de localización.

Para inversionistas inmobiliarios: La geolocalización de los locales y su score de flujo peatonal se consolidan como variables adicionales a incorporar en la evaluación de portafolios. Locales en Av. Libertad/Poniente o Av. Valparaíso con alta accesibilidad y flujo peatonal tienen mayores probabilidades de mantener altos precios de arriendo en el tiempo.

X. Discusión

10.1 Contraste con la Literatura

Los resultados del presente estudio son consistentes con la literatura revisada en sus hallazgos centrales y presentan también contribuciones diferenciadas. La dominancia de la superficie en la determinación del precio total es consistente con los estudios de Benjamin et al. (1992) y Eppli y Benjamin (1994). La elasticidad precio-superficie en el MCO estimada (0,868) es similar a la reportada en estudios de mercados latinoamericanos, donde la elasticidad precio-superficie suele situarse entre 0,70 y 0,90.

La emergencia del flujo peatonal como determinante relevante del precio, capturada por primera vez en este estudio mediante scores geoespaciales, es coherente con los hallazgos de Netzell (2009) y Yao y Pretorius (2014), quienes encontraron efectos significativos de la accesibilidad visual y el flujo de peatones (potenciales clientes) sobre el precio de arriendo. Esto valida la incorporación de variables de geolocalización como un avance metodológico que enriquece la explicación del componente de localización.

La heterocedasticidad finalmente fue comprimida durante el MCO log-lineal final, donde la transformación logarítmica de la variable dependiente fue necesaria, esto confirma que el mercado de locales comerciales de Viña del Mar presenta heterogeneidad estructural no capturada por variables en su estado normal y una transformación de variables a logaritmo sirve para captar de mejor manera este tipo de datos. La superioridad predictiva del árbol de regresión ($R^2=0,927$ vs $R^2_{\log}=0,820$ del MCO) es consistente con Yacim y Boshoff (2018), que documentan reducciones del 10-30% en el error de predicción de modelos de ML respecto al MCO.

XI. Análisis Interpretativo y Recomendaciones de Mejora

El presente capítulo ofrece un análisis interpretativo integrado de los resultados de ambas aproximaciones metodológicas, el MCO hedónico log-lineal y los modelos de Machine Learning, sobre la base de datos ampliada de 500 observaciones, y propone recomendaciones concretas para mejorar los modelos en iteraciones futuras.

11.1 Síntesis Interpretativa de los Resultados Actualizados

La comparación entre el modelo MCO log-lineal y el árbol de regresión revela una tensión productiva entre explicación e interpretación. El MCO log-lineal ofrece coeficientes directamente interpretables como elasticidades y semi-elasticidades, permite pruebas

estadísticas de significancia y se conecta directamente con la Teoría de los Precios Hedónicos de Rosen. Sin embargo, su poder predictivo ($R^2_{\log}=0,820$) es inferior al del árbol de regresión ($R^2_{\text{original}}=0,927$)

El árbol de regresión, en cambio, alcanza mayor precisión predictiva sin imponer supuestos distribucionales, pero su capacidad para identificar y cuantificar los precios implícitos de cada atributo, que es el objetivo central de la Teoría de los Precios Hedónicos, es limitada: la importancia de variables del árbol indica cuánto contribuye cada variable a las particiones, pero no el precio sombra marginal de cada atributo.

Esta complementariedad sugiere que la estrategia metodológica óptima para este tipo de investigación combina ambas aproximaciones: el MCO hedónico para la interpretación económica y la cuantificación de precios implícitos, y los modelos de ML para la predicción precisa y la captura de relaciones no lineales.

Es por esto que finalmente se escogió por el MCO hedónico, ya que, no solo se busca llegar a un precio si no también buscar profundidad en sus variables y buscar una interpretación económica que sea óptima para el estudio de la comuna.

11.2 Hallazgos Nuevos Relevantes de la Base Actualizada

La incorporación de las variables de geolocalización y accesibilidad en la base ampliada permitió identificar tres hallazgos nuevos de relevancia económica:

Primero, el flujo peatonal es el atributo de localización con mayor impacto sobre el precio en el modelo MCO (coeficiente $-0,793$ para flujo bajo). Esto implica que el mercado ya descuenta de manera significativa la menor exposición a clientes potenciales, lo que constituye la primera evidencia cuantitativa de este fenómeno en el mercado comercial de Viña del Mar.

Segundo, la distancia al centro urbano tiene una correlación negativa con el precio ($r=-0,14$) pero de magnitud moderada, mientras que la distancia al nodo comercial más cercano presenta una correlación de $-0,07$. Esto sugiere que el mercado responde más a la cercanía al polo comercial de referencia del sector (nodo) que, a la distancia al centro histórico, coherente con la estructura multicéntrica que Viña del Mar ha ido desarrollando.

Tercero, la correlación del número de baños con el precio ($r=0,500$) es la segunda más alta después de la superficie. Esto refleja que los baños actúan como proxy funcional del tamaño y categoría del local, y su inclusión mejora la capacidad predictiva del modelo al capturar información sobre el tipo de uso comercial que admite el local.

11.3 Recomendaciones para Mejorar los Modelos

Sobre la base del análisis de los resultados, se presentan las siguientes recomendaciones concretas para mejorar los modelos en iteraciones futuras:

11.3.1 Implementar Modelos de Ensamble (Gradient Boosting / Random Forest)

El árbol de regresión individual, pese a su buen desempeño, es sensible al sobreajuste y a la selección de los datos de entrenamiento. Los modelos de ensamble, en particular Gradient Boosting (algoritmos como XGBoost o LightGBM) y Random Forest, promedian las predicciones de múltiples árboles contruidos sobre subconjuntos aleatorios de datos y variables, reduciendo la varianza sin aumentar el sesgo. En el contexto de este estudio, con 19 variables y 500 observaciones, Random Forest con 100-500 árboles es un candidato natural para mejorar tanto la precisión predictiva como la estabilidad de la importancia de variables. Se anticipa que, con las nuevas variables de geolocalización, los modelos de ensamble podrían capturar interacciones complejas entre flujo peatonal, accesibilidad y tipo de local que un árbol individual no logra detectar.

11.3.3 Eliminar la Circularidad entre CLP/m² y el Precio

La variable CLP/m² (o su logaritmo) es una transformación directa del precio total dividido por la superficie. Su inclusión como predictor introduce una relación circular que infla artificialmente el poder predictivo del modelo y distorsiona la importancia relativa de la superficie. Se recomienda excluir CLP/m² (y uf_m^2) del conjunto de variables independientes en todas las especificaciones futuras, reemplazándola por variables de localización cuantitativas (scores de accesibilidad, flujo peatonal) que capturan el mismo fenómeno de la prima de localización sobre el precio unitario, sin la circularidad metodológica.

11.3.4 Explorar Modelos Econométricos Espaciales

La presencia de coordenadas geográficas en la base actualizada abre la posibilidad de implementar modelos econométricos espaciales (Modelo Autorregresivo Espacial o Modelo de Error Espacial) que incorporan explícitamente la autocorrelación espacial en los residuos. Como señalan Ramírez y Agudelo (2018) para el mercado de Bogotá, los modelos con rezago espacial pueden mejorar significativamente los estadísticos de ajuste respecto al MCO simple. El test de Moran sobre los residuos del MCO debería ser el primer paso para determinar si la autocorrelación espacial es un problema relevante en este conjunto de datos.

11.3.5 Medir Flujo Peatonal con Datos Reales

Los scores de flujo peatonal y accesibilidad incorporados en la base actualizada son estimaciones geoespaciales basadas en características observables del entorno. Una mejora sustancial sería validar y calibrar estos scores contra datos de flujo peatonal reales: conteos directos, datos de movilidad de teléfonos móviles (móvil analytics) o datos de Google Maps sobre popularidad de zonas en distintos horarios. Esto convertiría una variable proxy en una variable de medición directa, aumentando la validez de los coeficientes del MCO y la relevancia práctica de los resultados.

11.3.6 Incorporar la Dimensión Temporal

La base actual es transversal (corte en 2025-2026). Incorporar observaciones de períodos anteriores (2022-2024) permitiría estimar efectos temporales y detectar si los precios implícitos de los atributos, especialmente los de localización y flujo peatonal, han variado con la recuperación post-pandemia. Un modelo de panel o de series de tiempo cruzadas permitiría adicionalmente controlar por factores macroeconómicos (inflación, tasas de interés) que afectan al mercado en su conjunto.

11.3.7 Ampliar la Validación Geográfica

Finalmente, la generalización de los resultados podría mejorarse replicando el estudio en otras comunas de la Región de Valparaíso (Valparaíso, Con Con, Villa Alemana) o en otras ciudades intermedias de Chile (Rancagua, Talca, Temuco). Si el modelo arroja coeficientes hedónicos similares entre mercados, se reforzaría la validez externa de los hallazgos. Si difieren, se identificarían las características específicas del mercado de Viña del Mar que condicionan los resultados actuales.

XII. Conclusiones

El presente seminario de título construyó, entrenó y evaluó modelos econométricos y de Machine Learning para la predicción del precio de arriendo de locales comerciales en la comuna de Viña del Mar, sobre una base de datos ampliada de 500 observaciones con 19 variables, incluyendo variables de geolocalización, accesibilidad, flujo peatonal y visibilidad.

En términos metodológicos, la incorporación de un modelo MCO hedónico log-lineal como complemento a los modelos de ML representa un avance respecto a la versión original del estudio: mientras el árbol de regresión provee mayor precisión predictiva ($R^2=0,927$ en escala original), el MCO provee coeficientes interpretables que cuantifican los precios implícitos de cada atributo ($R^2_{\log}=0,820$). La combinación de ambas aproximaciones responde mejor al doble objetivo de predicción e interpretación económica que define el enfoque de los precios hedónicos.

La dominancia de la superficie como determinante del precio total de arriendo, hallazgo económico principal de este estudio, es confirmada en ambas aproximaciones. Sin embargo, la nueva evidencia sobre el flujo peatonal como segundo determinante de relevancia en el modelo MCO (coeficiente $-0,793$ para flujo bajo) enriquece sustancialmente la interpretación económica: el mercado de locales comerciales de Viña del Mar no solo valora el espacio, sino también la exposición al cliente potencial, lo que es consistente con la Teoría de la Renta de Oferta de Alonso y con los estudios de Space Syntax.

La heterocedasticidad persistente en el MCO, incluso con la especificación log-lineal, y la multicolinealidad severa entre las variables de geolocalización constituyen las principales limitaciones econométricas pendientes de resolver. Las recomendaciones de mejora presentadas en el Capítulo XI, que incluyen modelos de ensamble, errores estándar robustos, eliminación de la circularidad CLP/m², modelos espaciales y validación con datos de flujo peatonal real, proveen una hoja de ruta concreta para investigaciones futuras.

Para investigaciones futuras yo Milenko podría sugerir: (i) implementar Random Forest y Gradient Boosting con las variables de geolocalización; (ii) corregir la heterocedasticidad del MCO con errores estándar robustos HC3; (iii) explorar modelos SAR/SEM que incorporen autocorrelación espacial; (iv) excluir CLP/m² del conjunto de predictores para eliminar la circularidad; (v) validar los scores de flujo peatonal con datos de movilidad real; y (vi) ampliar el estudio a otras comunas de la Región de Valparaíso para evaluar la generalización de los resultados.

Referencias Bibliográficas

- Alonso, W. (1964). *Location and Land Use: Toward a General Theory of Land Rent*. Harvard University Press.
- Benjamin, J. D., Boyle, G. W., & Sirmans, C. F. (1992). Retail leasing: The determinants of shopping center rents. *AREUEA Journal*, 20(1), 9-21.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group.
- Chiang, Y.-H., Bin, J., & Strawderman, R. L. (2019). Machine learning approach to commercial real estate rent prediction. *Journal of Real Estate Research*, 41(1), 1-35.
- Eppli, M. J., & Benjamin, J. D. (1994). The evolution of shopping center research: A review and analysis. *Journal of Real Estate Research*, 9(1), 5-32.
- Glascock, J. L., Kim, M., & Sirmans, C. F. (1993). An analysis of office market rents: Some empirical evidence. *AREUEA Journal*, 21(1), 105-119.
- Hillier, B., & Hanson, J. (1984). *The Social Logic of Space*. Cambridge University Press.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, 11(1), 443-448.
- Mabel, L., & Tomás, B. (2017). Análisis hedónico de la vivienda en Buenos Aires. *Cuadernos de Economía*, 36(72), 117-148.
- Netzell, O. (2009). The effect of accessibility on retail rents: Testing integration value as a measure of geographic location. ERES Conference Paper. Stockholm: Royal Institute of Technology.
- Núñez, E., Garrido, N., & Paredes, R. (2021). Modelo hedónico para viviendas nuevas en la Región del Biobío, Chile. *Revista de Urbanismo*, 44, 1-18.
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction. *Expert Systems with Applications*, 42(6), 2928-2934.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Propital Blog. (2026). Mercado inmobiliario chileno 2026: Reactivación y tendencias. Recuperado de <https://propital.com/blog>.
- Ramírez, J., & Agudelo, C. (2018). Modelos de precios hedónicos espaciales para vivienda nueva en Bogotá. *Cuadernos de Economía*, 37(73), 25-59.

- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267-288.
- Yacim, J. A., & Boshoff, D. G. B. (2018). Impact of artificial neural networks training algorithms on accurate prediction of property values. *Journal of Real Estate Research*, 40(3), 375-418.
- Yao, J., & Pretorius, F. (2014). Retail shop visibility and rent in shopping malls. *Land Use Policy*, 40, 1-8.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301-320.

Anexos

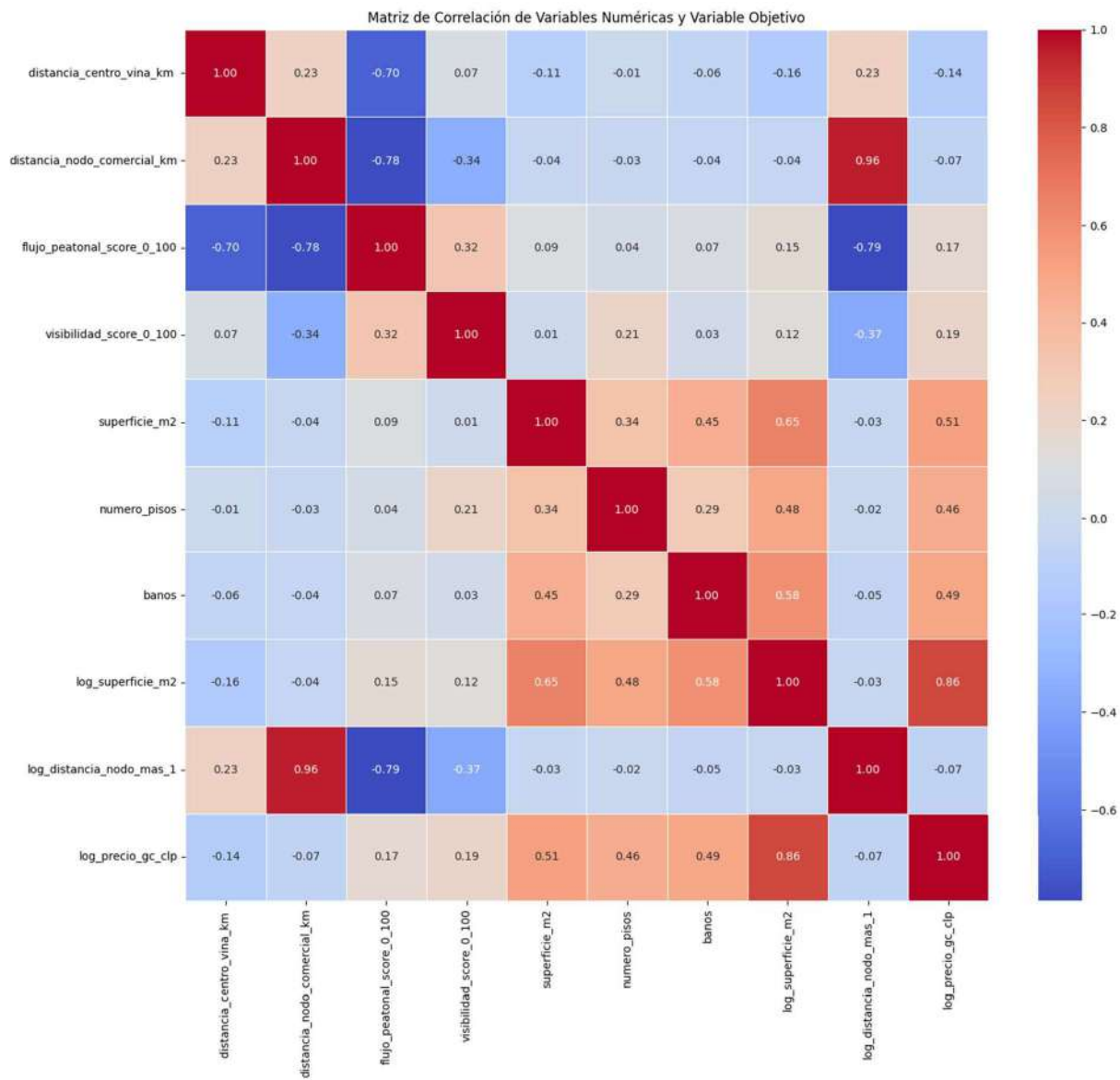


Tabla1: correlaciones entre variables



base nueva 500 - copia.xlsx

Base de datos

```

1. Cargar la base

uploaded = files.upload()
file_name = list(uploaded.keys())[0]

if file_name.lower().endswith(".xlsx"):
    df = pd.read_excel(file_name)
elif file_name.lower().endswith(".csv"):
    df = pd.read_csv(file_name, sep=";", decimal=",")
else:
    raise ValueError("Sube un archivo .xlsx o .csv")

print("Archivo:", file_name)
print("Dimensiones:", df.shape)
display(df.head())

```

Copy to clipboard No se eligió ningún archivo Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving base nueva 500 - copia.xlsx to base nueva 500 - copia.xlsx
 Archivo: base nueva 500 - copia.xlsx
 Dimensiones: (500, 13)

	precio_gc_clip	distancia_centro_vina_km	nodo_comercial_mas_cercano	distancia_nodo_comercial_km	flujo_peatonal_score_0_100	flujo_peatonal_refinado	visibilidad_score_0_100	tipo
0	9524184.0	0.847	Av Libertad / Poniente	0.491	71.0	Alta	68.5	cor
1	2439119.0	1.149	Av Valparaíso	1.140	44.8	Media	39.9	B
2	540000.0	8.240	Renaca Norte / Eluchans	0.071	59.5	Media-alta	74.2	
3	490000.0	1.020	Rodoviario	0.594	61.6	Media-alta	64.6	cor
4	1728000.0	2.511	Mall Marina	1.387	28.8	Media-baja	35.5	to

Paso 1: Cargar la base de datos, preparar variable objetivo y predictores

```

4. Split 80/20 y entrenamiento

X_train, X_test, y_train_log, y_test_log = train_test_split(
    X, y_log, test_size=0.20, random_state=42
)

mco_log.fit(X_train, y_train_log)
yhat_log = mco_log.predict(X_test)

print("n_train:", len(X_train))
print("n_test:", len(X_test))

... n_train: 400
    n_test: 100

```

Paso 2: Asignación de los datos para el entrenamiento y prueba. Split 80/20 (400/100)

5. Cálculo manual del R^2 log

```
sse_log = np.sum((y_test_log - yhat_log) ** 2)
sst_log = np.sum((y_test_log - y_test_log.mean()) ** 2)
r2_log_manual = 1 - sse_log / sst_log
r2_log_sklearn = r2_score(y_test_log, yhat_log)

print("SSE_log:", sse_log)
print("SST_log:", sst_log)
print("R2_log manual:", r2_log_manual)
print("R2_log sklearn:", r2_log_sklearn)

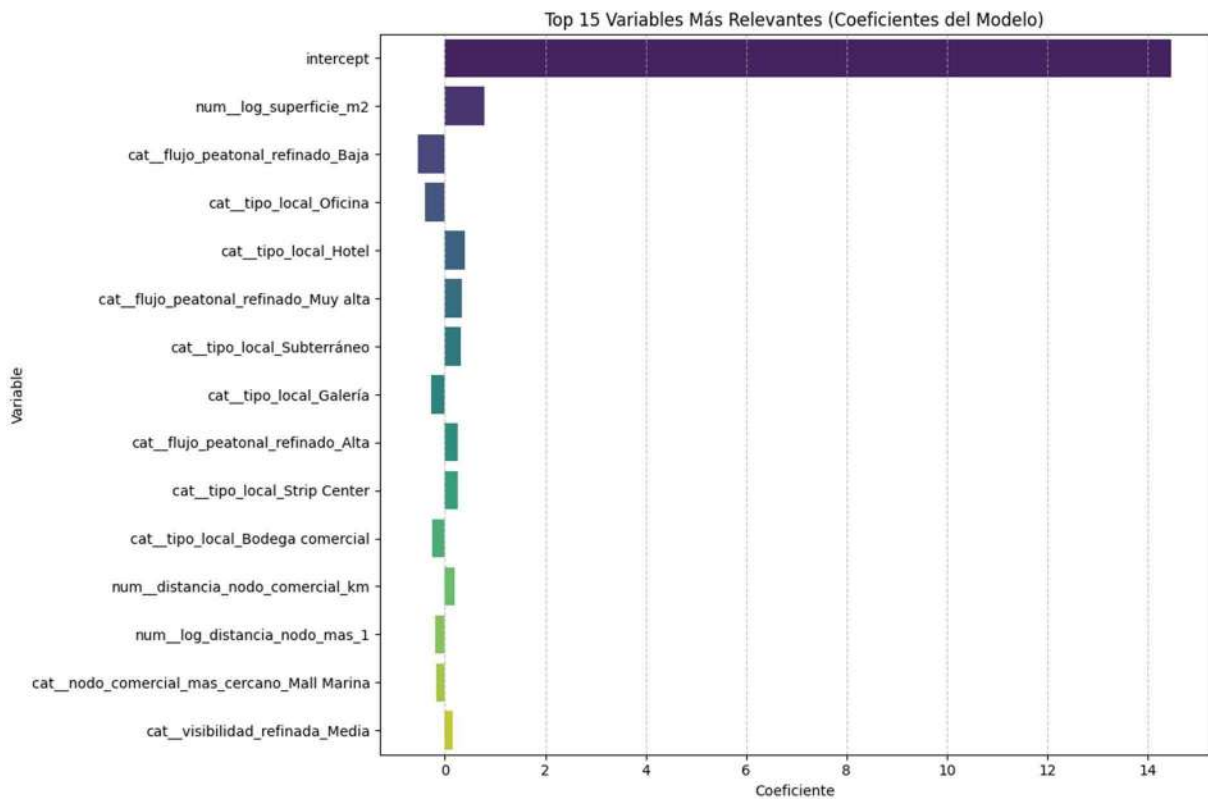
auditoria = pd.DataFrame({
    "metrica": ["n_test", "SSE_log", "SST_log", "R2_log_manual", "R2_log_sklearn"],
    "valor": [len(y_test_log), sse_log, sst_log, r2_log_manual, r2_log_sklearn]
})

display(auditoria)
```

```
*** SSE_log: 29.923563902026174
SST_log: 126.69399263609127
R2_log manual: 0.7638122907060247
R2_log sklearn: 0.7638122907060247
```

	metrica	valor
0	n_test	100.000000
1	SSE_log	29.923564
2	SST_log	126.693993
3	R2_log_manual	0.763812
4	R2_log_sklearn	0.763812

Calculo del primer R cuadrado, antes de llegar al modelo final



Primeros coeficientes de las variables más relevantes (no considerar intercepto)

12. Test de Breusch-Pagan para Heteroscedasticidad

```
from statsmodels.stats.api import het_breuschpagan
import statsmodels.api as sm

# Get the independent variables (X_test) used for prediction after preprocessing
# We need the preprocessed X_test to match the model's structure
X_test_processed = mco_log.named_steps['preprocess'].transform(X_test)

# Add a constant to the independent variables as required by the Breusch-Pagan test
X_test_processed_with_const = sm.add_constant(X_test_processed, has_constant='add')

# Perform the Breusch-Pagan test
# The function returns: lagrange multiplier statistic, p-value, F-statistic, F p-value
lm_statistic, lm_p_value, f_statistic, f_p_value = het_breuschpagan(residuos, X_test_processed_with_const)

print(f"Breusch-Pagan LM Statistic: {lm_statistic:.4f}")
print(f"Breusch-Pagan LM p-value: {lm_p_value:.4f}")
print(f"Breusch-Pagan F-Statistic: {f_statistic:.4f}")
print(f"Breusch-Pagan F p-value: {f_p_value:.4f}")

# Interpret the results
alpha = 0.05
print(f"\nInterpretación (con alfa = {alpha}):")
if lm_p_value < alpha:
    print("El p-valor del test de Breusch-Pagan (LM) es menor que el nivel de significancia (0.05).")
    print("Rechazamos la hipótesis nula. Esto sugiere la presencia de heteroscedasticidad.")
else:
    print("El p-valor del test de Breusch-Pagan (LM) es mayor que el nivel de significancia (0.05).")
    print("No rechazamos la hipótesis nula. No hay evidencia suficiente de heteroscedasticidad.")

Breusch-Pagan LM Statistic: 91.6388
Breusch-Pagan LM p-value: 0.0000
Breusch-Pagan F-Statistic: 28.9508
Breusch-Pagan F p-value: 0.0000
```

Test de Breusch-Pagan para confirmar la presencia de heterocedasticidad

Variables numéricas con correlación absoluta < 0.2 con el target:
['distancia_centro_vina_km', 'distancia_nodo_comercial_km', 'flujo_peatonal_score_0_100', 'visibilidad_score_0_100', 'log_distancia_nodo_mas_1']

Se ha creado un nuevo componente principal y se ha añadido a X.
Varianza explicada por el componente principal: 0.61

X.head() actualizado:

	distancia_centro_vina_km	nodo_comercial_mas_cercano	distancia_nodo_comercial_km	flujo_peatonal_score_0_100	flujo_peatonal_refinado	visibilidad_score_0_100	tipo_local	estado_conservacion	sa
0	0.847	Av Libertad / Poniente	0.491	71.0	Alta	68.5	Calle comercial	Bueno	
1	1.149	Av Valparaíso	1.140	44.8	Media	39.9	Bodega comercial	Bueno	
2	8.240	Renaca Norte / Elchans	0.071	59.5	Media-alta	74.2	Mall / Paseo	Bueno	
3	1.020	Rodoviario	0.594	61.6	Media-alta	64.6	Calle comercial	Regular	
4	2.511	Mall Marina	1.387	28.8	Media-baja	35.5	Bodega comercial	Bueno	

Creación del Componente Principal con las variables de menor correlación como medida para eliminar heterocedasticidad. Explicó un 61% de la varianza. Luego se re-entrenó el modelo y se evaluó nuevamente.

*** Comparación de Métricas en Escala Logarítmica:

	Metrica	Original Model	Model with PCA Added	Model with PCA Replacing Originals
0	R2	0.763812	0.763812	0.765911
1	RMSE	0.547024	0.547024	0.544588
2	MAE	0.408957	0.408957	0.407856
3	MAPE	0.027918	0.027918	0.027847

El resultado de esta ingeniería de datos no fue óptimo, demostrando que el problema no se encontraba ahí.

```
# 1. Calcular los residuos para el dataset completo usando el modelo original entrenado
yhat_full_dataset = mco_log.predict(X)
residuos_full_dataset = y_log - yhat_full_dataset

# 2. Identificar y filtrar los puntos donde los residuos están fuera de [-1.6, 1.6]
# Crear un filtro booleano para los residuos dentro del rango deseado
filtro_residuos = (residuos_full_dataset >= -1.6) & (residuos_full_dataset <= 1.6)

# Aplicar el filtro a X y y_log para crear el dataset 'limpio'
X_clean = X[filtro_residuos].copy()
y_log_clean = y_log[filtro_residuos].copy()

print(f"Dimensiones del dataset original: {X.shape}")
print(f"Dimensiones del dataset limpio (sin residuos extremos): {X_clean.shape}")
print(f"Número de puntos eliminados: {len(X) - len(X_clean)}")

*** Dimensiones del dataset original: (500, 15)
Dimensiones del dataset limpio (sin residuos extremos): (496, 15)
Número de puntos eliminados: 4
```

Se aplicó un filtro para eliminar datos muy atípicos que perjudicaban el modelo, quedando en una muestra de 496 datos. Finalmente se re-entrenó y se ajustó el modelo utilizando un Split de 396/100.

```
*** Métricas del Modelo (después de limpiar residuos extremos - escala logarítmica):
R2 (escala logarítmica): 0.8200
RMSE (escala logarítmica): 0.4242
MAE (escala logarítmica): 0.3309
MAPE (escala logarítmica): 0.0228

Comparación de Métricas en Escala Logarítmica:
  Métrica  Original Model  Model Cleaned Residuals (threshold 1.6)
0    R2      0.763812      0.820012
1  RMSE      0.547024      0.424177
2   MAE      0.408957      0.330857
3  MAPE      0.027918      0.022803

Observaciones:
- Se puede observar el impacto de la eliminación de puntos con residuos extremos (umbral 1.6) en las métricas del modelo.
```

El resultado fue bastante mejor de lo esperado, llegando a un coeficiente de R cuadrado de 0.820 con un poder predictivo y de interpretabilidad bastante alto.

20. Test de Jarque-Bera para Normalidad de los Residuos (Modelo con Datos Limpios)

```
from statsmodels.stats.stattools import jarque_bera

# Residuos del modelo entrenado con los datos limpios
residuos_clean = y_test_log_clean - yhat_log_clean

# Aplicar el test de Jarque-Bera
jb_value_clean, p_value_clean, skew_clean, kurtosis_clean = jarque_bera(residuos_clean)

print(f"Jarque-Bera Statistic (Cleaned Model): {jb_value_clean:.4f}")
print(f"P-value (Cleaned Model): {p_value_clean:.4f}")
print(f"Skewness (Cleaned Model): {skew_clean:.4f}")
print(f"Kurtosis (Cleaned Model): {kurtosis_clean:.4f}")

# Interpretar los resultados
alpha = 0.05
print(f"\nInterpretación (con alfa = {alpha}):")
if p_value_clean < alpha:
    print("El p-valor del test de Jarque-Bera es menor que el nivel de significancia (0.05).")
    print("Rechazamos la hipótesis nula. Esto sugiere que los residuos NO siguen una distribución normal.")
else:
    print("El p-valor del test de Jarque-Bera es mayor que el nivel de significancia (0.05).")
    print("No rechazamos la hipótesis nula. No hay evidencia suficiente para decir que los residuos NO siguen una distribución normal.")

*** Jarque-Bera Statistic (Cleaned Model): 0.1856
P-value (Cleaned Model): 0.9114
Skewness (Cleaned Model): 0.1032
Kurtosis (Cleaned Model): 3.0446

Interpretación (con alfa = 0.05):
El p-valor del test de Jarque-Bera es mayor que el nivel de significancia (0.05).
No rechazamos la hipótesis nula. No hay evidencia suficiente para decir que los residuos NO siguen una distribución normal.
```

Aplicación de test Jarque-Bera para confirmar que los residuos siguen una distribución normal.

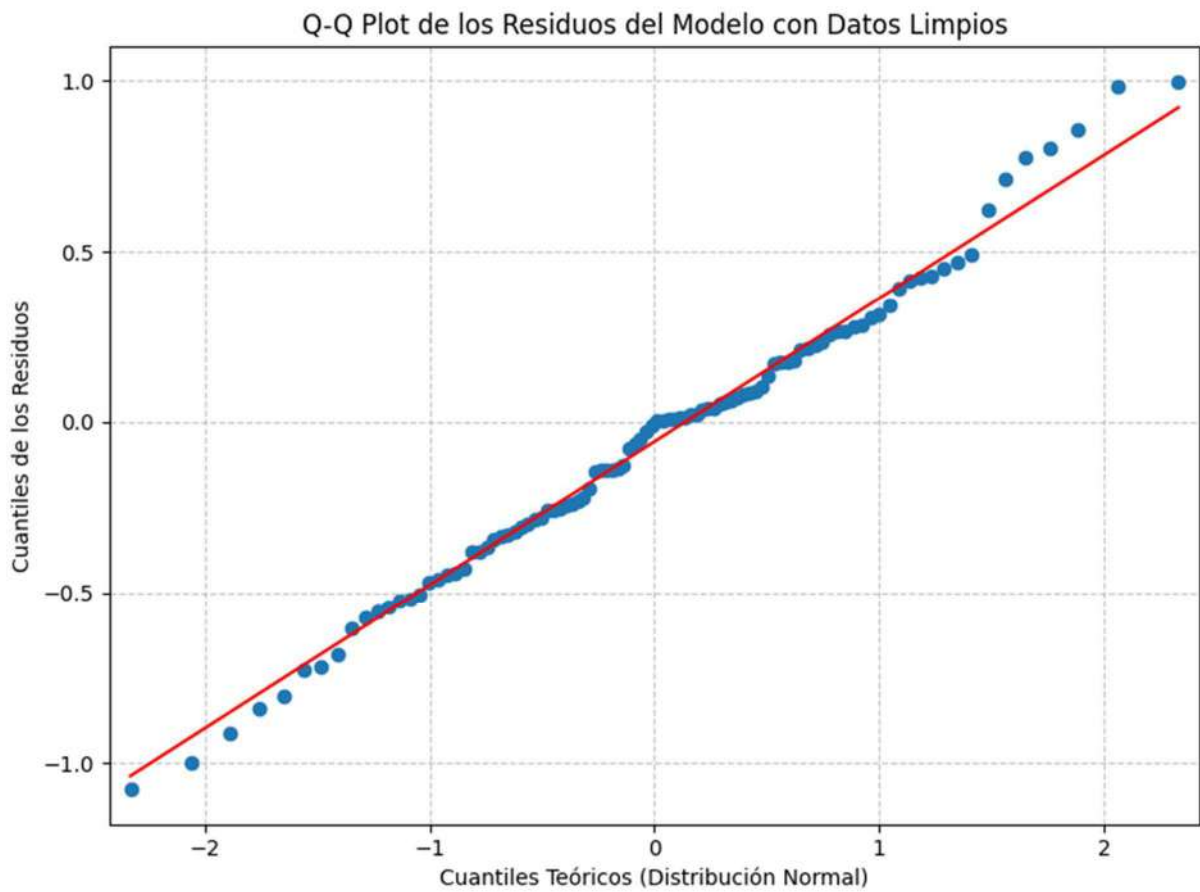


Tabla2: pruebas de normalidad de modelo con datos limpios

```

*** Breusch-Pagan LM Statistic (Cleaned Model): 33.7446
Breusch-Pagan LM p-value (Cleaned Model): 0.8430
Breusch-Pagan F-Statistic (Cleaned Model): 0.9737
Breusch-Pagan F p-value (Cleaned Model): 0.5228

```

Interpretación (con $\alpha = 0.05$):

El p-valor del test de Breusch-Pagan (LM) es mayor que el nivel de significancia (0.05).

No rechazamos la hipótesis nula. No hay evidencia suficiente de heteroscedasticidad en el modelo limpio.

Test Breusch-Pagan para confirmar que se eliminó la heterocedasticidad del modelo convirtiéndolo en un modelo óptimo

	feature	coeficiente	porcentaje_impacto_en_precio
43	intercept	14.538404	2.060386e+08
7	num_log_superficie_m2	0.868488	1.383305e+02
21	cat_flujo_peatonal_refinado_Baja	-0.793298	-5.476496e+01
32	cat_tipo_local_Strip Center	0.384564	4.689734e+01
17	cat_nodo_comercial_mas_cercano_Renaca Norte /...	0.379766	4.619425e+01
16	cat_nodo_comercial_mas_cercano_Renaca Centro	0.335569	3.987357e+01
1	num_distancia_nodo_comercial_km	0.309463	3.626935e+01
25	cat_flujo_peatonal_refinado_Muy alta	0.281580	3.252225e+01
31	cat_tipo_local_Oficina	-0.265972	-2.335397e+01
20	cat_flujo_peatonal_refinado_Alta	0.262215	2.998061e+01
14	cat_nodo_comercial_mas_cercano_Mall Marina	-0.244091	-2.165836e+01
15	cat_nodo_comercial_mas_cercano_Plaza Vergara	-0.220758	-1.980897e+01
8	num_log_distancia_nodo_mas_1	-0.218921	-1.966152e+01
29	cat_tipo_local_Hotel	0.206869	2.298218e+01
28	cat_tipo_local_Galería	-0.188539	-1.718318e+01
13	cat_nodo_comercial_mas_cercano_El Salto / Lim...	0.179562	1.966930e+01
26	cat_tipo_local_Bodega comercial	-0.176329	-1.616583e+01
42	cat_visibilidad_refinada_Muy alta	-0.149976	-1.392717e+01
39	cat_visibilidad_refinada_Media	0.149349	1.610777e+01
0	num_distancia_centro_vina_km	-0.130439	-1.222903e+01

	feature	coeficiente	porcentaje_impacto_en_precio
10	cat_nodo_comercial_mas_cercano_Arlegui	-0.129005	-1.210307e+01
18	cat_nodo_comercial_mas_cercano_Rodoviario	-0.107371	-1.018074e+01
23	cat_flujo_peatonal_refinado_Media-alta	0.105151	1.108785e+01
36	cat_estado_conservacion_Nuevo	0.094952	9.960639e+00
19	cat_nodo_comercial_mas_cercano_San Martin	-0.079306	-7.624259e+00
24	cat_flujo_peatonal_refinado_Media-baja	0.076312	7.929883e+00
12	cat_nodo_comercial_mas_cercano_Av Valparaiso	-0.068838	-6.652205e+00
22	cat_flujo_peatonal_refinado_Media	0.068040	7.040816e+00
33	cat_tipo_local_Strip center	0.064336	6.645104e+00
37	cat_estado_conservacion_Regular	-0.058839	-5.714147e+00
38	cat_visibilidad_refinada_Alta	-0.057498	-5.587575e+00
30	cat_tipo_local_Mall / Paseo	-0.046153	-4.510409e+00
11	cat_nodo_comercial_mas_cercano_Av Libertad / ...	-0.045527	-4.450661e+00
40	cat_visibilidad_refinada_Media-alta	0.044375	4.537454e+00
35	cat_estado_conservacion_Bueno	-0.036113	-3.546895e+00
34	cat_tipo_local_Subterráneo	0.019494	1.968497e+00
5	num_numero_pisos	0.019454	1.964447e+00
3	num_sibilidad_score_0_100	0.019430	1.961995e+00
2	num_flujo_peatonal_score_0_100	0.015837	1.596317e+00
41	cat_visibilidad_refinada_Media-baja	0.013750	1.384494e+00
6	num_banos	-0.007062	-7.037433e-01
9	num_Principal_Component_Low_Corr	0.002202	2.204862e-01
27	cat_tipo_local_Calle comercial	0.001731	1.732215e-01
4	num_superficie_m2	0.001154	1.154468e-01

Tabla3: Coeficientes finales y su poder de cambio en el precio en porcentaje. Directamente copiada del archivo ipynb



MODELO_MCO_NUE
VO_(4).ipynb

Archivo Python final, utilizado en los últimos modelos (definitivos)