Santiago de Cali, junio de 2024

Doctor
Diego Luis Linares
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "Construcción de un modelo para predecir ventas de unidades nuevas de vivienda en Cali por medio de técnicas de aprendizaje estadístico", el cual será realizado por el (los) estudiante (s) _Jorge Hernán Mora García, Leidy Lorena Conde Chavarro_ con código (s) __8980436, 8974558__ pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección del profesor David Arango.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

David Arango Londoño

David Arango Londoño
C.C. 1.130 586.950 de Cali, Valle del Cauca.

Jorge Hernán Mora García
C.C. 73.194.666 de Cartagena, Bolívar

Leidy Lorena Conde Chávarro
C.C. 1.075.312.298 de Neiva, Huila

CONSTRUCCIÓN DE UN MODELO PARA PREDECIR VENTAS DE UNIDADES NUEVAS DE VIVIENDA EN CALI POR MEDIO DE TÉCNICAS DE APRENDIZAJE ESTADÍSTICO

Jorge Hernán Mora García- Código 8980436
Leidy Lorena Conde Chávarro- Código 8974558

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
Maestría en Ciencia de datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: **“CONSTRUCCIÓN DE UN MODELO PARA PREDECIR VENTAS DE UNIDADES NUEVAS DE VIVIENDA EN CALI POR MEDIO DE TÉCNICAS DE APRENDIZAJE ESTADÍSTICO”**

1. ÉNFASIS: Construcción
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Sector construcción
4. ESTUDIANTE (S): Jorge Hernán Mora García, Leidy Lorena Conde Chávarro.
5. CORREO ELECTRÓNICO:jmoragarcia@javerianacali.edu.co , leidyconde@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 28a 9-29 B/Cámbulos Neiva-Huila. 3213639209 - 3182412246
7. DIRECTOR: David Arango Londoño
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Cátedra
9. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica): No aplica
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): No aplica
12. OTROS GRUPOS O EMPRESAS: No aplica
13. PALABRAS CLAVE (al menos 5): Predicción ventas, Google Trends, Modelación estadística, Vivienda, Modelo de predicción.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): ODS 9
15. FECHA DE INICIO (Desarrollo del proyecto): Abril 2023
16. RESUMEN (máximo 400 palabras).

El proyecto "Construcción de un modelo para predecir ventas de unidades nuevas de vivienda en Cali por medio de técnicas de aprendizaje estadístico" tiene como objetivo mejorar la precisión en la predicción de las transacciones de vivienda nueva en el futuro. Actualmente, la determinación de estas transacciones se basa en encuestas y sondeos de percepción de mercado, lo que limita la captura de información completa y actualizada sobre la situación real del mercado y la conducta de los posibles compradores. El proyecto consiste en la construcción de un modelo que utilice información secundaria para predecir las ventas de unidades nuevas de vivienda en el área del

Distrito de Cali. Esta información secundaria incluyó análisis de tendencias en Google Trends y variables macroeconómicas relevantes, como la inflación, el desempleo, las tasas de interés e indicadores agregados de percepción del consumidor y de desempeño de la economía. El enfoque se basó en técnicas de modelación estadística y métodos de aprendizaje automático supervisados, considerando que todos los datos son series temporales.

El modelo realizado proporcionó un método eficaz para obtener predicciones en tanto en el volumen como en la tendencia de venta de nuevas unidades de vivienda, respaldando así la toma de decisiones de política. Al utilizar técnicas de aprendizaje estadístico, se logró una mejor comprensión de los factores que influyen en las ventas de viviendas nuevas y, por lo tanto, se mejoró la capacidad de predecir las transacciones futuras. La modelación elaborada permite una planificación más eficiente de los recursos y una mejor comprensión de las dinámicas del mercado de viviendas nuevas en Cali. El proyecto propuso un modelo predictivo con técnicas de aprendizaje estadístico y datos secundarios que predice las ventas de unidades de viviendas nuevas en Cali, proporcionando así información más actualizada y precisa para respaldar la toma de decisiones en el sector de la construcción y servicios públicos, mejorando así la planificación y la comprensión del mercado.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La predicción de las ventas de unidades nuevas de vivienda en Cali se basa en encuestas de percepción del mercado realizadas a potenciales compradores y sondeos de intención de compra. Sin embargo, esta aproximación presenta limitaciones importantes. Conlleva el despliegue de recursos económicos considerables los cuales son fundamentales para garantizar la precisión y fiabilidad de las predicciones, estos comprenden los recursos para la obtención de datos y el análisis detallado de la información, al igual que el tiempo dedicado por profesionales cualificados en todas las etapas del estudio. De igual manera, la asignación de recursos se restringe a etapas específicas del desarrollo de cada investigación, lo cual puede derivar en una falta de captura exhaustiva de información sobre la situación real del mercado donde estas encuestas solo revelan la conducta de los potenciales compradores en el momento específico en el que se lleva a cabo la captura de los datos.

Para entidades como las administraciones municipales, los gremios de la construcción y los prestadores de servicios públicos, contar con datos e información más actualizados es esencial para la toma de decisiones fundamentadas. Por lo tanto, se requiere una metodología más eficiente y a menor costo que permita contrarrestar las limitaciones de las estimaciones económicas actuales.

Se propuso la construcción de un modelo para predecir las ventas de unidades nuevas de vivienda en Cali, utilizando técnicas de aprendizaje estadístico. El modelo se basa en información secundaria tales como consultas en Google Trends y variables macroeconómicas relevantes, como inflación, desempleo, tasas de interés. La modelación estadística y los métodos de aprendizaje automático supervisados o no supervisados permitió la aplicación de un método que proporcionó datos más actualizados y a menor costo.

Se abordó el desafío de la predicción de ventas de unidades nuevas de vivienda en Cali, para proporcionar a las entidades pertinentes una herramienta eficiente y actualizada para la toma de decisiones en el sector de la construcción de viviendas.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Actualmente, determinar el volumen de transacciones de vivienda nueva o su tendencia es relevante para diversos actores del sector inmobiliario, desde desarrolladores, entidades financieras y autoridades locales o nacionales. Si bien las encuestas de percepción de mercado y los sondeos de intención de compra son herramientas ampliamente utilizadas para este fin, es importante evaluar críticamente su efectividad y considerar alternativas, reflejando como señala CAMACOL "la oferta se corrige por la demanda, lo que significa que si la comercialización baja, así lo harán los lanzamientos en los meses subsiguientes, y viceversa" (Cámara Colombiana de la Construcción, CAMACOL, 2022) .

La planeación de estos elementos y el despliegue de los seguimientos requieren recursos económicos y se restringe a periodos determinados. Esto puede resultar en la falta de captura de información completa de la situación del mercado y, además, revela la conducta de los potenciales compradores solo cuando se realiza la fase de captura de datos. Para las administraciones municipales, gremios de la construcción y prestadores de servicios públicos, reviste mayor valor poder contar con datos e información más actualizada para la toma de decisiones.

Para abordar eficazmente la predicción de ventas de unidades de vivienda nueva es necesario comprender la dinámica del mercado en Colombia. Según los datos de la Superintendencia Financiera de Colombia, "en términos reales a noviembre de 2023, las tasas reales de construcción de vivienda No VIS es de 6,7%, de construcción de vivienda VIS de 6,8%, de adquisición de vivienda No VIS de 6,3% y de construcción de vivienda VIS de 5,1%" (Banco de la República, 2023) . Por ende, se resalta la necesidad de desarrollar una herramienta o método con datos más

actualizados y a menor costo, que permita generar estimaciones económicas utilizando modelación estadística y métodos de aprendizaje estadístico automático.

El efecto de las variables macroeconómicas como inflación, desempleo, tasas de interés o los agregados de percepción sobre la economía, así como los niveles de popularidad relativa de consultas asociadas al mercado de vivienda suministradas por Google Trends, se relacionan estrechamente con el sector inmobiliario, dado que "este sector se encuentra afectado por algunas variables macroeconómicas en diferentes mercados, haciendo que muchos elementos de la construcción como los precios de venta y producción, las importaciones, los precios en la mano de obra entre otras sean de gran repercusión para el sector" (Rodríguez Niño, 2022, 2023).

Comprendiendo estos factores se mitiga el riesgo de enfrentar consecuencias significativas en el mercado inmobiliario de Cali y en la economía en general ya que, la falta de precisión en las estimaciones de ventas podría llevar a una planificación inadecuada, resultando en un exceso o escasez de oferta de vivienda, lo que a su vez podría impactar negativamente en los precios de la vivienda y en la estabilidad del mercado.

1.2 FORMULACIÓN DEL PROBLEMA

De acuerdo con el Estudio de oferta y demanda de Vivienda en el Valle del Cauca para el año 2022, se desataca que "Para el III trimestre de 2022 el PIB del sector de la construcción registró un crecimiento anual del 13,4% en su serie original, mostrando así una recuperación de 8,7 p.p respecto al primer trimestre del año. Este comportamiento se explica por el buen desempeño del componente de construcción de edificaciones residenciales y no residenciales, el cual, registró una variación anual de 19,3% en su serie original" (CAMACOL VALLE, 2022). Estos datos subrayan la relevancia de comprender las variables que inciden en la decisión de compra de vivienda nueva

en Cali, lo cual constituye un aspecto crucial para orientar estrategias en el mercado inmobiliario local.

El problema por abordar en este proyecto radica en la carencia de un método actualizado y económico para predecir ventas de unidades nuevas de vivienda en Cali. La propuesta es construir un modelo basado en técnicas de aprendizaje estadístico y datos secundarios, que permita a las partes interesadas en el sector de la construcción contar con información más precisa y actualizada para la toma de decisiones estratégicas.

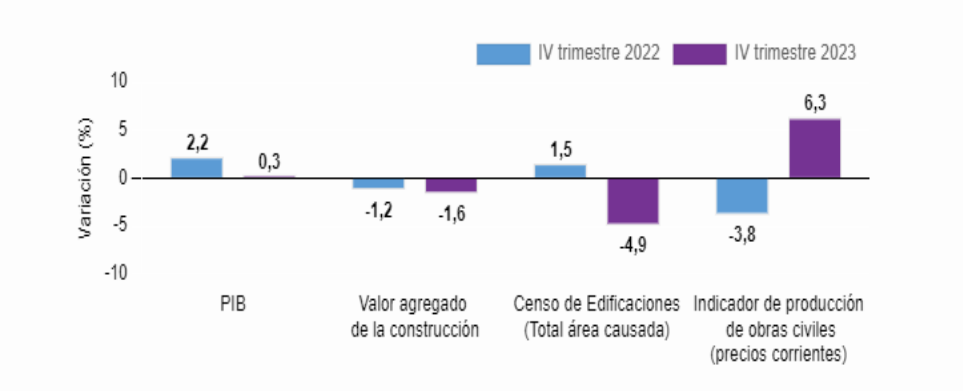


Figura 1. Variación anual de los indicadores de coyuntura del sector de la construcción IV trimestre (2022-2023)[1]

¿Se puede predecir la venta de nuevas unidades de vivienda a partir del análisis de datos provenientes de Google Trends y de indicadores macroeconómicos (inflación, desempleo, tasa de interés, tasa de cambio, entre otros) en Cali?, y, además, ¿Es posible a partir del uso de técnicas de aprendizaje automático obtener un pronóstico más preciso y expedito de la venta de unidades

[1] Tomado de: https://www.dane.gov.co/index.php/estadisticas-por-tema/construccion/indicadores-economicos-alrededor-de-la-construccion

nuevas de vivienda en Cali, frente al tradicional mecanismo de captura de información primaria con encuestas?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Construir un modelo de predicción que utilice técnicas de aprendizaje estadístico y datos secundarios para predecir las ventas de unidades nuevas de vivienda en Cali.

2.2 OBJETIVOS ESPECÍFICOS

- Integrar fuentes secundarias y realizar el preprocesamiento de los datos necesarios para la construcción del modelo de predicción de ventas de unidades nuevas de vivienda en Cali utilizando técnicas de aprendizaje estadístico.
- Desarrollar un modelo estadístico utilizando técnicas de aprendizaje automático supervisado o no supervisado.
- Evaluar la eficacia del modelo de predicción desarrollado.
- Ofrecer una herramienta o método que proporcione datos más actualizados y a menor costo para la toma de decisiones en el sector de la construcción de viviendas en Cali.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

Se consideran los fundamentos teóricos y metodológicos necesarios para desarrollar un modelo estadístico que pueda pronosticar ventas de unidades nuevas de vivienda en Cali, con técnicas de aprendizaje estadístico. Este marco abarca las dimensiones teóricas que envuelven el análisis de datos grandes y su aplicación práctica dentro de la industria de la construcción y el mercado inmobiliario.

Se revisan diferentes conceptos vinculados a la probabilidad y las estadísticas, modelos supervisados y no supervisados y la relevancia de las diferentes fuentes de datos para la etapa de recolección, como consultas de Google Trends, Indicadores macroeconómicos y otros, haciendo uso de la información accesible donde "se busca extraer de sitios web información sobre el mercado de la vivienda para la generación de indicadores más precisos como el número de nuevas ofertas publicadas o la fluctuación de los precios a lo largo del tiempo para las ofertas existentes" (Rosso-Mateus, Montilla-Montilla, & Garzon-Martínez, 2022). Finalmente, los enfoques previos utilizados para problemas similares se discuten; esto proporciona un contexto histórico y técnico para respaldar el diseño y la implementación del modelo propuesto.

El mercado inmobiliario en Colombia ha experimentado una evolución significativa en las últimas décadas influenciada por factores económicos, sociales y políticos, donde se va formando la oferta y demanda de vivienda según esos componentes. "Colombia se destaca como uno de los países de Latinoamérica donde el sector inmobiliario experimenta un fuerte crecimiento. Esta dinámica ha generado un mercado en constante movimiento y ha permitido un mayor grado de compromiso por parte de las constructoras que buscan ampliar cada vez más su participación en el mercado"

(Vargas, 2023). De esta manera, se buscan estrategias para mantener el desarrollo inmobiliario en el país, siendo este sector determinante para el crecimiento de la economía, la captación de inversión y la generación de empleo.

La adaptación de los procesos tecnológicos con el sector inmobiliario está impulsando un cambio positivo, con el objetivo de crear mecanismos más eficientes y útiles. “El desarrollo de la tecnología en los últimos años ha ido cambiando y evolucionando los modelos de negocios de los diferentes sectores. Y lejos de quedar exento de esta transformación, el sector inmobiliario ha apostado por la tecnología de la propiedad con un objetivo claro; impulsar sus oportunidades de negocio en beneficio de sus usuarios” (Vargas, 2023).

Ahora bien, es importante tener en cuenta que existen restricciones que el mercado inmobiliario puede presentar a la oferta de vivienda, donde es importante destacar que las constructoras no poseen la libertad de desarrollar sus proyectos en el lugar que deseen. Paciorek establece que en primer lugar que la “regulación”, que implica la obtención de permisos (licencias) de construcción, y, en segundo lugar, las restricciones geográficas asociadas a los terrenos disponibles para construir, tales como cuerpos de agua o pendientes elevadas, impactan a la oferta de este mercado por la vía de un mayor costo de las viviendas (Paciorek, 2013). Adicionalmente, este tipo de restricciones moldean la ubicación espacial de la oferta de vivienda a algunos sectores, que para el caso del Distrito de Cali y su área de influencia para el periodo 2009 – 2015, se han venido concentrando en la zona sur de la ciudad y en los municipios de Palmira y Jamundí. (Guerrero, 2016).

Por lo anterior, se analizan las tendencias actuales del mercado inmobiliario de Cali, considerando factores económicos, sociales y políticos para mejorar las predicciones del modelo y proporcionar

una herramienta integral para la planificación estratégica y la toma de decisiones informadas.

3.1.1. Fundamentos de modelos predictivos

Los modelos predictivos son herramientas matemáticas y estadísticas que nos permiten percibir eventos futuros. Al analizar datos históricos y reconocer patrones ocultos, estos modelos identifican correlaciones entre variables diversas, brindándonos información para la toma de decisiones en múltiples campos.

Según Timón & Fontes en el campo médico, los modelos predictivos permiten anticipar la evolución de enfermedades basándose en síntomas y datos demográficos de los pacientes, esto facilita diagnósticos más precisos, tratamientos personalizados y una mejor gestión de recursos en el sistema de salud. Las finanzas también se benefician del poder predictivo por medio del análisis de datos históricos y tendencias del mercado, estas herramientas ayudan a evaluar riesgos de crédito, predecir fluctuaciones del mercado y optimizar estrategias de inversión.

"Esta tendencia al uso del Análisis Predictivo es consecuencia de la nueva cultura que se ha generalizado con respecto a los datos. La capacidad real de almacenar y procesar grandes conjuntos de datos, ligada a los avances experimentados por las TI, ha permitido generar archivos masivos de datos de todo tipo, susceptibles de ser analizados en busca de tendencia" (Timón & Fontes, 2017).

La construcción de un modelo predictivo para el mercado inmobiliario de Cali acarrea varias implicaciones de relevancia como la transformación que requieren de las organizaciones desde la planificación precisa y fundamentada para desarrollar la oferta, anticipándose a la demanda futura y a las variaciones económicas que puedan afectar el mercado inmobiliario. Por lo tanto, este

modelo ofrece a desarrolladores, planificadores urbanos y entidades financieras una avanzada herramienta que facilita la oferta de alternativas y la toma de decisiones basada en la información actualizada.

Además, los modelos predictivos contribuyen al desarrollo sostenible del sector inmobiliario al facilitar la creación de una oferta inmobiliaria alineada con las necesidades reales del mercado "Para el análisis de datos se usan diferentes modelos, los predictivos estiman valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos llamadas variables independientes o predictivas" (Cuecha, Hernández, & Rodríguez, 2022). Proporcionan información valiosa sobre la demanda del mercado, el diseño y la construcción de edificios, la gestión de activos y la economía circular, estos modelos empoderan a los actores del sector para tomar decisiones más responsables.

3.1.2. Análisis de tendencias y datos del mercado inmobiliario en Colombia

Al comprender las tendencias y preferencias de los consumidores, las empresas pueden desarrollar proyectos inmobiliarios que sean ambientalmente responsables, socialmente inclusivos y económicamente viables. Los principales compradores de vivienda nueva consideran algunas restricciones, sin dejar de lado que, en Colombia, la vivienda es un componente importante del patrimonio de un hogar, y el hecho de que un hogar sea propietario, representa un factor de bienestar social, de acuerdo con Caicedo, Morales y Pérez (Caicedo, Morales-Mosquera, & Pérez-Reyna, 2010).

La primera de las restricciones identificadas en la demanda se basa en el precio de la vivienda nueva, la cual se puede ver afectada por la evolución del poder adquisitivo. El mercado inmobiliario está ligado con la cantidad de bienes a los que puede acceder cada persona y por consiguiente esto

genera un impacto significativo en los precios de la vivienda nueva a causa de factores como la inflación, tasas de interés y capacidad adquisitiva.

De igual manera, existen políticas de planificación urbana y las regulaciones de zonificación también pueden limitar la disponibilidad de terrenos para desarrollar nuevos proyectos de viviendas nuevas, ya que estas regulaciones pueden también imponer restricciones en las características de las construcciones permitidas, lo que podría aumentar los costos de desarrollo y por ende, los precios finales de las viviendas. Así, la única opción de adquirir vivienda es el acceso al crédito, el cual representa una limitación al momento de realizar el proceso de compra de vivienda nueva debido a los requisitos para acceder al mismo. Siendo este un elemento fundamental que determina la dinámica del mercado inmobiliario y la posibilidad de que las personas accedan a la propiedad de vivienda, es necesario implementar políticas inclusivas que promuevan el acceso al crédito, especialmente para aquellos grupos de población que enfrentan mayores dificultades para obtener financiamiento adecuado.

Además, el costo de financiación representa también un impedimento que retracta a la población de adquirir créditos de vivienda nueva en Colombia debido a tasas de interés, seguros, condiciones hipotecarias. De acuerdo con el informe de financiación de vivienda, emitido por el Departamento Administrativo Nacional de Estadística DANE (Departamento Administrativo Nacional de Estadística - DANE, 2023), se registraron desembolsos por valor de 4.13 billones de pesos, que se distribuyeron de la siguiente forma: 3.19 billones por modalidad de crédito de vivienda y 0.94 billones por modalidad de leasing habitacional para todo tipo de vivienda, nueva y usada. Sin embargo, estas cifras revelan una reducción del 31% en el total del monto desembolsado de créditos del primer trimestre de 2023 respecto al mismo periodo de 2022.

Los montos de financiación de vivienda y el número de créditos desembolsados hacen parte de la evaluación teórica del estudio, se consideran aspectos como: ¿Cómo impactan el aumento reciente de las tasas de interés a la decisión de los compradores de vivienda nueva? ¿Cómo impacta la reducción de los subsidios de vivienda otorgados a este mercado, especialmente en el mercado de Vivienda de Interés Social (VIS)?. Obsérvese que otros de los aspectos de importancia en la dinámica de la demanda de este mercado, se basa en las subvenciones a la población de menores ingresos (Gonzales-Arrieta, 2005). Adicional, se considera importante la cuantificación de la conducta de algunas variables macroeconómicas, tales como el empleo, el crecimiento económico o las expectativas de la inflación. (Schwab, 1982).

3.1.3. Implicaciones del proyecto

Como la investigación se basa en la construcción de un modelo para predecir ventas de vivienda en el Distrito de Cali mediante técnicas de aprendizaje estadístico, es importante considerar las restricciones de la oferta y de la demanda. Al respecto, una de las fuentes de información relevante puede ser las búsquedas registradas en Google Trends en algunas palabras clave como “subsidios”, “hipotecas”, “constructora XYZ” o sobre los proyectos de vivienda existentes en el ámbito espacial definido.

“Google Trends (tendencias en español) es un servicio gratuito de Google que cuantifica las búsquedas realizadas en este motor. La información se encuentra disponible en informes semanales a partir de 2004. Su principal ventaja es su interfaz: resulta muy simple de usar, intuitiva y da la posibilidad de graficar términos de búsqueda más populares del pasado reciente. A su vez, es posible exportar los resultados a un archivo csv”. (Blanco, 2014)

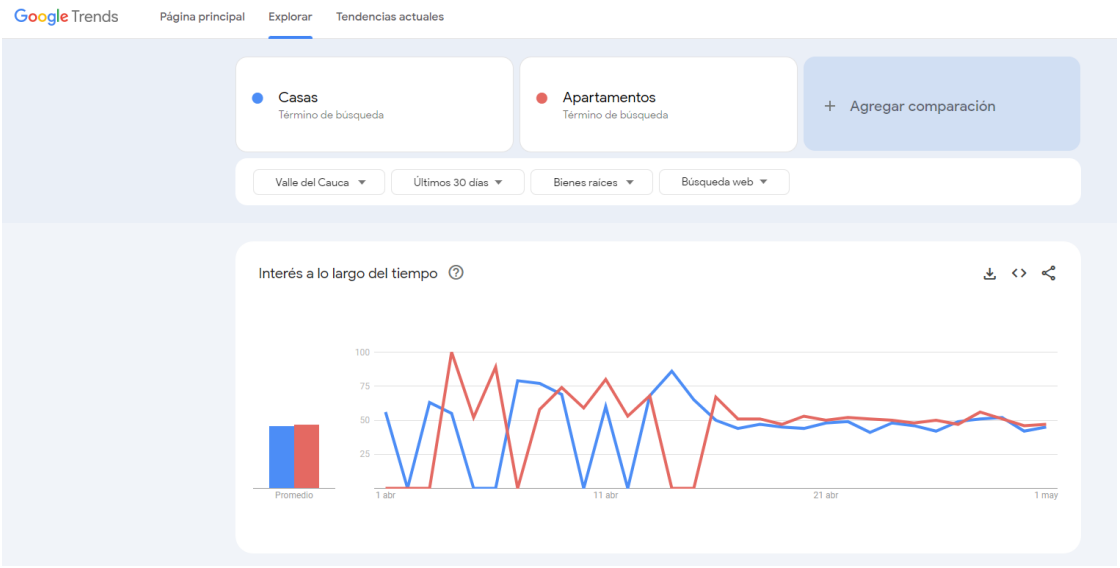


Figura 2 Ejemplo de búsqueda *Tomado de:*

https://trends.google.es/trends/explore?cat=29&date=today%201-m&geo=CO-VAC&q=Casas,Apartamentos&hl=es-419

Algunos estudios evidencian la importancia del análisis de tendencias en Google Trends como predictor de variables importantes en el desarrollo de proyectos que requieren eficiencia y una actualización constante optimizando los recursos y haciendo uso de la tecnología. Wijnhoven y Plant concluyeron, en un estudio de venta de vehículos en Países Bajos, que el análisis de datos a partir de Google Trends se perfilaba como un buen predictor de la venta de modelos específicos (Wijnhoven & Plant, 2017). A una conclusión similar llegaron Carrière y Labbé en un estudio similar adelantado en Chile donde se resalta este enfoque ya que ofrece una forma más rápida y precisa de predecir las tendencias del mercado, lo que es valioso para los responsables de las políticas y planificadores económicos (Carrière-Swallow & Labbé, 2013).

En consecuencia, aunque el mercado de vivienda tiene diferencias en estructura con el mercado de vehículos, en ambos casos los bienes sujetos a las transacciones (casas y vehículos) se enmarcan

en una categoría similar en cuanto a definición de bienes duraderos. Esto implica que algunas restricciones pueden ser comunes en ambos mercados, como el acceso a créditos, o porque la decisión de compra de este tipo siempre es compleja.

3.2. ANTECEDENTES

En el ámbito de la predicción de ventas de unidades nuevas de vivienda, se han realizado diversos estudios y aplicaciones que abordan problemas similares al planteado en el proyecto "Construcción de un modelo para predecir ventas de unidades nuevas de vivienda en Cali por medio de técnicas de aprendizaje estadístico". A continuación, se presentan algunos de estos trabajos junto con un resumen de su enfoque, la referencia o enlace donde se encuentran los detalles y cómo contribuyen al desarrollo del proyecto actual:

El proyecto "Análisis espacial de los cambios y determinantes de la oferta de vivienda nueva no VIS en Cali para el periodo 2009-2015" constituye un antecedente importante para el desarrollo del presente trabajo, debido a su aporte de estudio y análisis del mercado inmobiliario, resaltando que “El departamento del Valle del Cauca ocupa la tercera posición en el mercado de la vivienda nueva, presentando una participación del 13% sobre las ventas en lo corrido a agosto 2016. Desde el 2009 la oferta de vivienda en la ciudad se ha concentrado en el segmento No VIS, con la característica de presentarse en estado preventa (sobre planos) y de tipo apartamento” (Guerrero, 2016). De esta manera, comparte enfoque y objetivos relacionados con el mercado inmobiliario en Cali con el uso de datos secundarios e indicadores económicos, además del aporte a la evolución de la oferta de vivienda en Cali y las variables que influenciaron esos cambios, proporcionando así un conocimiento sólido sobre la dinámica del mercado inmobiliario en la ciudad. “El sector de la

construcción (SC) es uno de los sectores con mayor impacto en las economías del mundo. La construcción es una industria y, por lo tanto, la relación de esta actividad con la economía de un país o de una región es tanto cercana como sensible a los cambios en el ciclo de producción de la industria. La investigación genera las siguientes preguntas: ¿cuál es el comportamiento mensual de monitoreo del SC (IMSC) y cuáles son los ciclos económicos que se producen en el Valle del Cauca? y ¿cómo se comporta la trayectoria histórica del SC, frente a otros indicadores nacionales y regionales?, y se destacan los hechos relevantes a los cambios en los ciclos económicos motivados por la experiencia del indicador mensual de la actividad económica (IMAE) del Valle del Cauca"(Garay-Rodríguez, Vidal, & Cerón-Ordóñez, 2023). El análisis del sector de la construcción y su interacción con la economía regional es una pieza fundamental para comprender el pulso económico de cada región, de esta forma al hacer una evaluación profunda nos brinda información vital sobre la salud financiera, la generación de empleo, el impacto en otros sectores, la planificación urbana y el desarrollo sostenible. Cuando comprendemos esta dinámica, se toman decisiones que pueden anticipar crisis y crear planes para una debida gestión económica que beneficia a todo el sector.

Del mismo modo el nuevo proyecto aprovecha los métodos y la necesidad de crear un modelo de predicción que permita reducir los costos de la investigación e incluye la integración de datos de consultas de Google Trends para capturar las preferencias actuales de los consumidores o el uso de algoritmos más sofisticados para mejorar la precisión de las predicciones, analizando su relación con las variables macroeconómicas que afecten la decisión de venta y compra de vivienda. El artículo "Use of Google Trends to Predict the Real Estate Market: Evidence from the United Kingdom" (Bulczak, 2021) revela las ventajas de analizar los datos provenientes de Google Trends para mejorar el poder de predicción de los indicadores del mercado inmobiliario (Precios, ventas),

resaltando el análisis de las consultas “house for sale”, “mortgage” y “rentals”. Este estudio analiza la utilidad de Google Trends para comprender las fluctuaciones del mercado inmobiliario del Reino Unido durante la crisis financiera. La hipótesis es que la frecuencia de búsquedas relacionadas con la compraventa de viviendas refleja la actividad económica y el comportamiento del mercado inmobiliario, donde se busca validar la utilidad de Google Trends para predecir la demanda en este sector especialmente en tiempos de crisis y se discuten estudios previos sobre el potencial de Google Trends para predecir la actividad económica y el comportamiento del consumidor, y se proporciona una base para la aplicación de técnicas de análisis de datos de búsqueda en la predicción del mercado inmobiliario, lo que respalda la viabilidad del uso de técnicas de aprendizaje estadístico para predecir la venta de unidades nuevas de vivienda en Cali.

En este entendido, es relevante estudiar el uso de modelos de aprendizaje estadístico automático para estudiar conjuntos de datos masivos y obtener predicciones y resultados que ayuden a la toma de decisiones. “Otro problema por destacar en el ámbito inmobiliario, aplicable a métodos econométricos, pero también a los intensivos en inteligencia artificial, es que no se están aprovechando suficientemente las tecnologías emergentes para la recolección, la limpieza, el procesamiento, el análisis y la visualización de datos 15-16. Por el contrario, la mayoría de los trabajos emplean tecnologías tradicionales (Ej: software basado en botones, ejecución por pasos que depende de la manipulación humana, métodos de análisis limitados para abordar datos masivos,…)” (Rave, 2019)

La relevancia del aprendizaje estadístico para predecir las ventas de viviendas en Cali se evidencia en los estudios previos donde el análisis espacial de la oferta de vivienda nueva en la ciudad proporcionó información valiosa sobre la dinámica del mercado inmobiliario local. Ese mismo aporte es realizado por el trabajo “Metodología para obtención y análisis de datos inmobiliarios

usando fuentes alternativas: estudio de caso en tres ciudades intermedias de Colombia". Esta información fue de utilidad para construir un modelo de predicción en el nuevo proyecto, ya que permite aprovechar los métodos desarrollados como el análisis de tendencias de precios y factores económicos que afectan la demanda de vivienda. "El estudio realizado perfiló una ruta para la gestión y el análisis automático de la información inmobiliaria y de su entorno, teniendo como fuentes diferentes sitios web, que complementan las fuentes oficiales al brindar disponibilidad de datos donde se carezca de estos; lo anterior gracias a un proceso metodológico que permitió identificar las características propias y las variables externas que afectan el precio de los inmuebles en el mercado" (Rosso-Mateus, Montilla-Montilla, & Garzon-Martínez, 2022)

El proyecto actual busca combinar elementos de estos trabajos previos al utilizar información secundaria, como datos de popularidad relativa de consultas Google Trends, y agregados macroeconómicos, para mejorar la precisión en la predicción de transacciones de unidades nuevas de vivienda en Cali. A diferencia de los trabajos mencionados, esta propuesta se delimita en el contexto de Cali y se enfoca en la construcción de un modelo con estas fuentes de datos adicionales para dar información más actualizada y precisa a las partes interesadas en la toma de decisiones.

3.3. MODELOS PREDICTIVOS

3.3.1. Redes Neuronales Recurrentes – LSTM y GRU

Las Redes Neuronales Recurrentes (RNN, por sus siglas en inglés) son una clase de redes neuronales diseñadas para procesar secuencias de datos, lo que las hace particularmente útiles en tareas donde el orden y la temporalidad de los datos son importantes, como el procesamiento de lenguaje natural, la traducción automática y la predicción de series temporales. A diferencia de las

redes neuronales feedforward, las RNN poseen conexiones recurrentes que permiten mantener un estado interno que captura información sobre secuencias anteriores, lo que las capacita para manejar dependencias a corto plazo.

Sin embargo, las RNN tradicionales presentan problemas al intentar capturar dependencias a largo plazo debido a la aparición del problema del gradiente desvaneciente. Este problema impide que las RNN aprendan eficientemente patrones que se encuentran a gran distancia temporal en las secuencias de datos. Para solucionar esta limitación, se desarrollaron las Long Short-Term Memory (LSTM), una variante de las RNN introducida por Hochreiter y Schmidhuber en 1997, y posteriormente, las Gated Recurrent Units (GRU), introducidas por Cho, van Merrienboer, Bahdanau y Bengio en 2014.

Las LSTM mitigan el problema del gradiente desvaneciente mediante la introducción de una estructura de células de memoria con tres puertas reguladoras: la puerta de entrada, la puerta de salida y la puerta de olvido. Estas puertas permiten a la red controlar el flujo de información y mantener información relevante durante largos intervalos de tiempo. Estos componentes de la red se definen de la siguiente forma:

1. Puerta de entrada (input gate): Determina cuánta información de la entrada actual debe ser guardada en la célula de memoria.
2. Puerta de olvido (forget gate): Decide qué información de la célula de memoria debe ser descartada.
3. Puerta de salida (output gate): Controla cuánto de la información de la célula de memoria se utiliza para generar la salida de la LSTM.

La arquitectura de las LSTM ha demostrado ser efectivas especialmente en la predicción de series de tiempo. las LSTM son capaces de capturar patrones complejos y relaciones temporales a largo

plazo que otras técnicas, como los modelos ARIMA o las redes neuronales feedforward, no pueden (Hochreiter & Schmidhuber, 1997). Ahora bien, sin dejar de lado las restricciones computacionales, se evaluarán en la predicción de venta de nuevas unidades de vivienda en Cali, dado que son resistentes a los atributos de las series de tiempo como la estacionalidad y la tendencia, que generan ruido en otras técnicas estadísticas de predicción. Además, funcionan bien en conjuntos de datos pequeños.

Por otro lado, se evaluará una arquitectura alternativa, las redes recurrentes con capas Gated Recurrent Units (GRU). Las GRU están diseñadas para manejar dependencias a largo plazo de manera más eficiente y con menos complejidad computacional que las LSTM. Las GRU simplifican la arquitectura de las LSTM al combinar las puertas de entrada y de olvido en una sola "puerta de actualización", y eliminar la puerta de salida. Esta estructura simplificada permite a las GRU capturar dependencias temporales con un menor costo computacional, lo que puede ser ventajoso en aplicaciones donde los recursos son limitados (Cho, Merrienboer, Bahdanau, & Bengio, 2014).

Los componentes de estas redes se describen a continuación:

1. Puerta de actualización (update gate): Controla la cantidad de información de la entrada actual que se guarda en el estado y la cantidad de información antigua que se olvida.
2. Puerta de reinicio (reset gate): Determina cuánta información pasada se olvida al calcular el nuevo estado.

En el contexto de la predicción de ventas de unidades de vivienda en Cali, las GRU son adecuadas para manejar las complejidades y variabilidades de las series temporales, especialmente cuando se integran múltiples fuentes de datos como Google Trends y variables macroeconómicas.

Al igual que las LSTM, las GRU pueden aprender patrones complejos y relaciones temporales a largo plazo en los datos. Sin embargo, su arquitectura más simple puede ofrecer ventajas en términos de velocidad de entrenamiento y eficiencia computacional.

3.3.2. Random Forest

El Random Forest es un algoritmo de aprendizaje automático ampliamente utilizado para tareas de clasificación y regresión. Introducido por Breiman en 2001, este método se basa en la creación de un conjunto de árboles de decisión durante el entrenamiento y la salida de la clase que es el modo de las clases (clasificación) o la media de las predicciones (regresión) de los árboles individuales. La principal ventaja de Random Forest es su capacidad para mejorar la precisión y reducir el sobreajuste mediante el promedio de múltiples árboles de decisión, lo que lo hace robusto frente a la variabilidad en los datos.

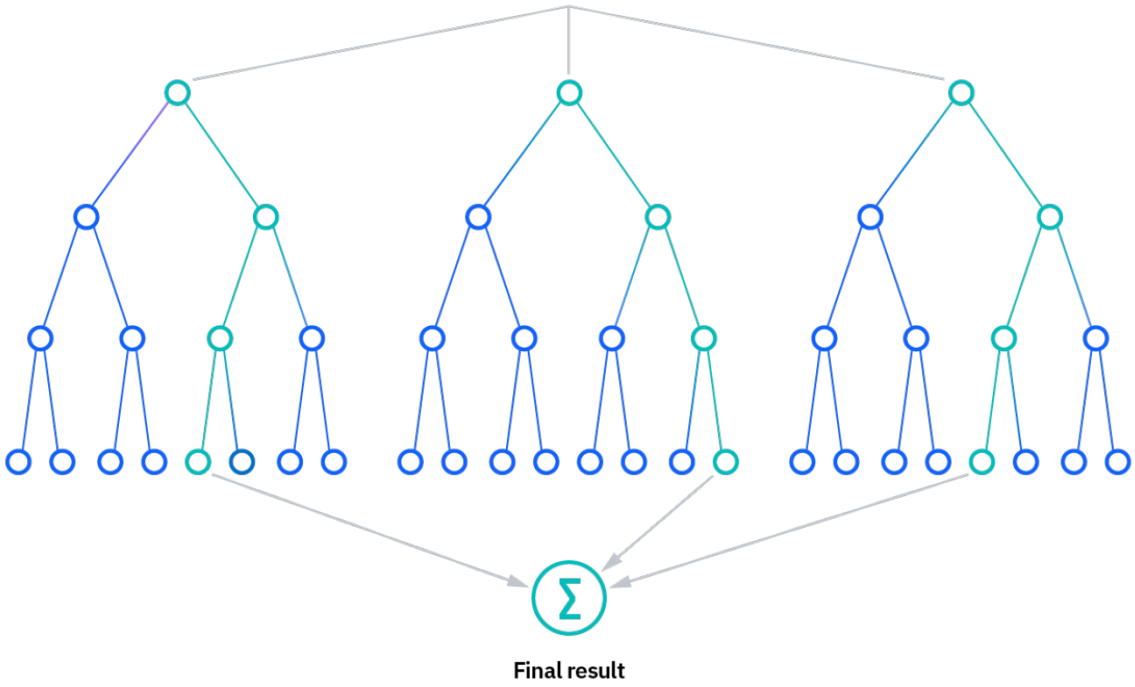


Figura 3. Esquema de Bosque Aleatorio. Tomado de https://www.ibm.com/mx-es/topics/random-forest

En el contexto de la predicción de series temporales, como la venta de unidades de vivienda en Cali, el Random Forest tiene aplicabilidad. Aunque originalmente no fue diseñado específicamente para datos temporales, se pueden adaptar varias estrategias para su uso en este tipo de datos. Al crear un modelo de Random Forest para series temporales, se requiere transformar los datos de manera que las características temporales sean capturadas adecuadamente. Esto generalmente implica la creación de características basadas en el tiempo, como rezagos, diferencias y tendencias. El uso de Random Forest en la predicción de ventas de unidades de vivienda en Cali puede beneficiarse de la inclusión de múltiples fuentes de datos, como la popularidad relativa de términos de búsqueda en Google Trends y variables macroeconómicas. Estos datos adicionales pueden mejorar significativamente la precisión del modelo al capturar una variedad de factores que influyen en las ventas de viviendas. Por ejemplo, términos de búsqueda relacionados con la compra

de vivienda pueden indicar cambios en el interés del mercado, mientras que las variables macroeconómicas, como las tasas de interés y el empleo, reflejan las condiciones económicas generales que afectan la capacidad de los consumidores para comprar viviendas. La estructura típica de este tipo de modelos tiene en cuenta lo siguiente:

1. Construcción del modelo: Se generan múltiples árboles de decisión utilizando diferentes subconjuntos de datos y características. Cada árbol se entrena en una muestra aleatoria del conjunto de datos de entrenamiento.
2. Combinación de predicciones: Las predicciones de todos los árboles se combinan para formar una predicción final. En regresión, esto se hace promediando las predicciones de todos los árboles.
3. Reducción de la varianza: Al combinar múltiples árboles, el Random Forest reduce la varianza de las predicciones y mejora la generalización del modelo.

Una ventaja significativa del Random Forest es su capacidad para manejar grandes conjuntos de datos con muchas características y detectar automáticamente las interacciones no lineales entre las variables. En el caso de la predicción de ventas de viviendas, esto significa que el modelo puede identificar patrones complejos y relaciones entre las variables de búsqueda en Google Trends, las variables macroeconómicas y la variable objetivo: ventas de viviendas.

Además, Random Forest proporciona una medida de la importancia de las características, lo que puede ser invaluable para entender qué factores tienen el mayor impacto en las predicciones de ventas. Esta información puede ser utilizada por las empresas para ajustar sus estrategias de mercado y tomar decisiones informadas sobre la gestión de inventarios y las proyecciones de ventas.

3.3.3. XGBoost

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje automático basado en árboles de decisión que ha ganado popularidad debido a su rendimiento superior en una variedad de tareas de predicción. Desarrollado por Chen y Guestrin en 2016, XGBoost se destaca por su eficiencia, velocidad y capacidad para manejar grandes conjuntos de datos con alta dimensionalidad. Utiliza un enfoque de boosting que construye árboles de decisión secuencialmente, donde cada nuevo árbol intenta corregir los errores de los árboles anteriores, mejorando iterativamente la precisión del modelo (Chen & Guestrin, 2016). Algunas de las características de esta técnica son las siguientes:

1. Construcción del modelo: XGBoost crea árboles de decisión en secuencia, donde cada árbol nuevo intenta corregir los errores cometidos por los árboles anteriores mediante la optimización del gradiente de la función de pérdida.
2. Regularización: XGBoost incluye parámetros de regularización para evitar el sobreajuste, lo que mejora la generalización del modelo a nuevos datos.
3. Importancia de las características: XGBoost proporciona métricas de importancia de las características, lo que ayuda a identificar las variables más influyentes en las predicciones.

Una ventaja significativa de XGBoost, al igual que Random Forest, es su capacidad para manejar interacciones no lineales entre múltiples variables, pero además presenta una mejora en su eficiencia computacional, lo que permite el manejo de grandes volúmenes de datos. En el caso de la predicción de ventas de viviendas, al igual que el algoritmo de Random Forest, XGBoost puede

identificar patrones complejos y relaciones entre las variables de búsqueda en Google Trends, las variables macroeconómicas y las ventas de viviendas.

Además, XGBoost ofrece interpretabilidad a través de las métricas de importancia de características, permitiendo a las empresas comprender mejor los factores clave que impulsan las ventas de viviendas.

3.3.4. Regresión lineal

La regresión lineal es uno de los métodos estadísticos más simples y ampliamente utilizados para modelar la relación entre una variable dependiente y una o más variables independientes (Granados, 2016). En el contexto de la predicción de series temporales, como la venta de unidades de vivienda en Cali, la regresión lineal puede ser una herramienta poderosa para identificar y cuantificar las relaciones entre las ventas de viviendas y diversas variables predictoras.

La regresión lineal se basa en el supuesto de que la relación entre las variables independientes (predictoras) y la variable dependiente (respuesta) es lineal. El modelo de regresión lineal simple se expresa como:

$$y=\beta_0+\beta_1 x_1+\epsilon$$

donde y es la variable dependiente, x_1 es la variable independiente, β_0 es el intercepto, β_1 es el coeficiente de la pendiente que representa el cambio en y por unidad de cambio en x_1, y ϵ es el término de error.

En el caso de la predicción de ventas de unidades de vivienda en Cali, la regresión lineal puede ser extendida para incluir múltiples variables independientes, convirtiéndose en una regresión lineal múltiple:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\epsilon$$

donde $x_1,x_2,\ldots,x_p$ representan diferentes variables predictoras, como la popularidad relativa de términos de búsqueda en Google Trends y varias variables macroeconómicas.

1. **Variables predictoras**: En este contexto, las variables predictoras podrían incluir términos de búsqueda relacionados con la compra de vivienda (datos de Google Trends), tasas de interés, tasas de empleo, índices de precios de vivienda, entre otros. Estas variables pueden proporcionar una visión integral de los factores que influyen en las ventas de viviendas.
2. **Entrenamiento del modelo**: Se utiliza un conjunto de datos históricos para ajustar el modelo de regresión lineal, determinando los coeficientes que minimizan la suma de los cuadrados de los errores (diferencia entre las predicciones del modelo y los valores observados).
3. **Predicción**: Una vez entrenado, el modelo puede ser utilizado para predecir las ventas futuras de viviendas en función de los valores de las variables predictoras.

La regresión lineal tiene la ventaja de ser fácil de interpretar y rápida de entrenar, lo que la hace adecuada para aplicaciones donde se requiere una comprensión clara de las relaciones entre las variables. En el caso de la predicción de ventas de viviendas, la regresión lineal puede proporcionar una base sólida para entender cómo los cambios en la popularidad de ciertos términos de búsqueda y las condiciones macroeconómicas afectan las ventas de viviendas.

Por ejemplo, un aumento en las búsquedas de términos relacionados con la compra de vivienda en Google Trends podría estar asociado con un aumento en las ventas de viviendas en los meses siguientes. Del mismo modo, variables macroeconómicas como una disminución en las tasas de interés podrían tener un impacto positivo en las ventas de viviendas al hacer más accesible la financiación para los compradores.

Aunque la regresión lineal tiene limitaciones, especialmente en su capacidad para capturar relaciones no lineales y complejas entre las variables, es el método más simple, menos complejo computacionalmente y se pondrá a prueba como método de control respecto a los resultados que arrojen las técnicas más avanzadas.

4. PREPROCESAMIENTO DE LOS DATOS E INTEGRACIÓN DE FUENTES SECUNDARIAS

La construcción del set de datos para entrenar a los modelos de aprendizaje estadístico automático descrito fue el primer paso. La metodología consistió en la compilación de un gran set de datos inicial, compuesto por la variable objetivo, el número total de unidades de vivienda vendidas, con un total de 179 observaciones mensuales entre enero de 2008 y noviembre de 2022.

Por su parte, se obtiene el valor de la popularidad relativa de las consultas de Google trends, para todo el periodo el en que se encuentran disponibles utilizando una consulta de pytrends sobre una notebook en entorno de Google Colab. En este aspecto se aclara que la delimitación geográfica se circunscribe a Valle del Cauca, como unidad mínima de desagregación para la herramienta. Los datos obtenidos para todas las consultas que se describen más adelante, se compilaron para el

periodo comprendido entre enero de 2004 hasta enero de 2024, con un total de 241 observaciones de la popularidad relativa para cada una de las consultas.

Por otro lado, y en cuando a la búsqueda de variables macroeconómicas, se extrajeron un total de 25 variables asociadas al mercado laboral, es decir, no necesariamente a la tasa de desempleo, para estudiar su conducta y la viabilidad de integrarlas como predictores en el ejercicio. La fuente de estos datos es el Departamento Administrativo Nacional de Estadística – DANE. Para algunas de las variables existe disponibilidad desde enero de 2004 hasta enero de 2024, con 241 observaciones, pero las específicas para el mercado laboral del Distrito de Cali se encuentran disponibles desde marzo de 2007 para un total de 203 observaciones.

Para el caso de la inflación, se extraen 4 atributos, la variación del IPC nacional, la variación acumulada de los últimos 12 meses a nivel nacional y estos dos mismos atributos para Cali específicamente. La fuente de estos datos es el DANE y los datos está disponibles desde enero de 2004 hasta enero de 2024.

También se evalúan los indicadores de la Encuesta de Opinión del Consumidor de Fedesarrollo. Un total de 6 posibles predictores entre los que se encuentran el Índice de Confianza del Consumidor ICC, el Índice de Condiciones Económicas – ICE, el Índice de Expectativas del Consumidor – IEC y los índices de Disposición a Comprar Vivienda o IDCV. En el caso del ICC y el IDCV, se logra consolidar parte de los datos específicamente para Cali. En este caso la disponibilidad de los datos no es total para el ICC específico para Cali y los Índices de Disponibilidad a Comprar Vivienda. Para subsanar este inconveniente después de extraer los datos de las publicaciones elaboradas por Fedesarrollo en su página web, se efectúan cálculos de interpolación lineal para completar las series.

Por su parte, se estudiaron 8 series de tiempo asociadas a las tasas de cambio. La fuente de estos datos es el Banco de La República, y comprenden desde tasas nominales a fin de mes, promedio del mes y los índices de tasas de cambio reales deflactados tanto para exportaciones tradicionales como no tradicionales. En cuanto a tasas de interés, cuya fuente es igualmente Banrep, se estudiaron cuatro series para créditos de vivienda VIS y no VIS y modalidad UVR y Pesos. En el caso de UVR, el valor reportado es la tasa de financiación real, es decir, el diferencial después de la variación de la UVR. Para estas series publicadas por banrep (Tasas de Cambio y Tasas de Interés) se logró disponibilidad completa en el periodo comprendido entre enero de 2004 y enero de 2024, 241 datos. Finalmente se estudia el Índice de Seguimiento a la Economía ISE, global y específico para el sector de la Construcción. Este dato calculado y publicado por el DANE está disponible desde enero de 2005, para un total de 229 registros mensuales.

4.1. DESCRIPCIÓN DE LOS CONJUNTOS DE DATOS

Tal y como se definió en la sección anterior, el set de datos se compone de series temporales, encabezadas por la variable objetivo, “VENTAS” que se refiere valor mensual total de unidades de vivienda vendidas, de acuerdo con datos suministrados por Camacol. Así mismo, se compone de los resultados devueltos por el aplicativo de Google Trends y las variables macroeconómicas asociadas a desempleo, inflación, tasa de cambio, tasa de interés y las percepciones del consumidor, enlistando adicionalmente el Índice de Seguimiento a la Economía. Dado que el set de datos inicial se compuso de 63 variables y la disponibilidad de datos de la variable objetivo es de sólo 179 observaciones, se ejecuta un ejercicio de exploración visual, a partir de las bibliotecas matplotlib y seaborn, además de analizar las correlaciones tanto a nivel como a las primeras diferencias con el

objetivo de depurar los atributos o predictores que se utilizaron en el entrenamiento de los modelos de aprendizaje utilizados. A continuación, se presentan los resultados del análisis exploratorio y descriptivo tanto de la variable objetivo como del grupo de predictores:

4.1.1. Unidades vendidas de vivienda

Según datos suministrados por Camacol, se registraron ventas de 278.666 viviendas entre enero de 2008 y noviembre de 2022 en el Distrito de Cali. Los datos de unidades vendidas son los que se encuentran disponibles en el desarrollo del presente proyecto, y presentan a primera vista, una tendencia al alza y un nivel de estacionalidad, lo que indica que dicha serie puede no ser estacionaria.

La serie de tiempo señalada marca un mínimo histórico con 345 unidades vendidas en el periodo de noviembre de 2008, mientras que el máximo histórico corresponde a noviembre de 2021, con 4034 unidades vendidas. Así mismo se visualiza un posible cambio estructural del nivel de ventas a partir del segundo semestre del 2020, que vino precedido de una caída hasta un mínimo de 876 unidades en abril de 2020, durante el periodo de aislamiento por la pandemia COVID-19, y que fue escalando progresivamente hasta un máximo de 3403 unidades vendidas reportadas para octubre de ese año.

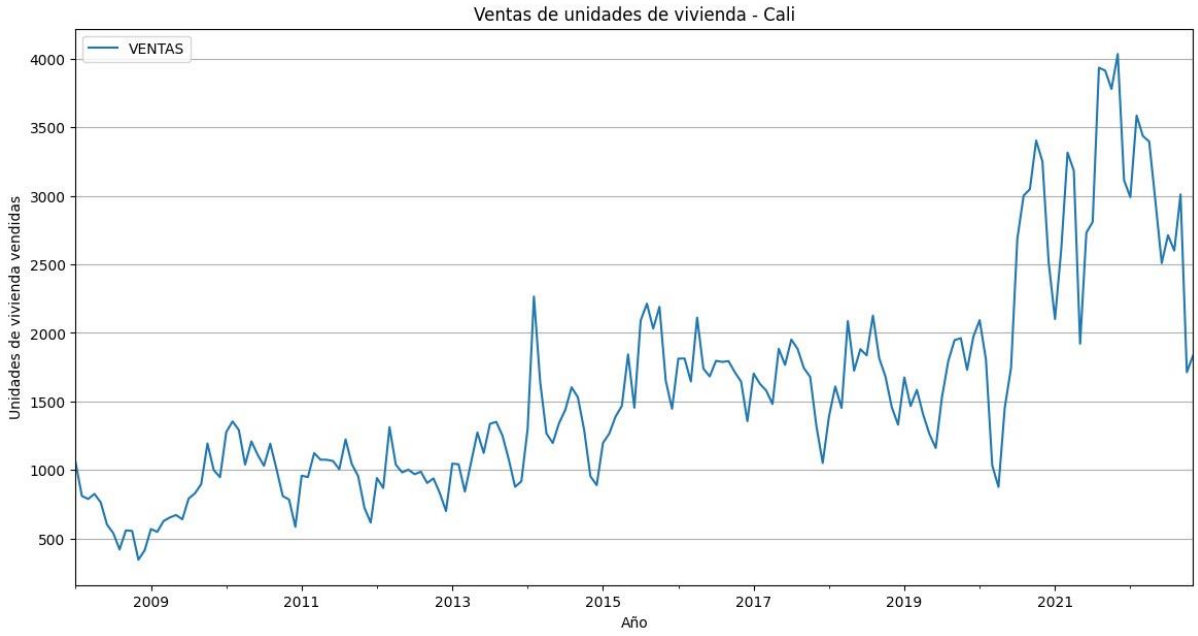


Figura 4. Serie de tiempo de "VENTAS". Ventas de unidades de vivienda en Cali.

El nivel total de ventas anuales de unidades de vivienda marca una tendencia al alza desde 2008, con un aumento sostenido de 2008 a 2010, con una reducción en 2011 y 2012. Luego aumenta y estabiliza a un promedio de alrededor de 20.000 unidades vendidas entre 2015 y 2019 para luego aumenta a un máximo de 37.445 unidades vendidas en el año 2021. Ahora bien, se aclara que los datos para el año 2022 se encuentran con corte a noviembre de 2022.

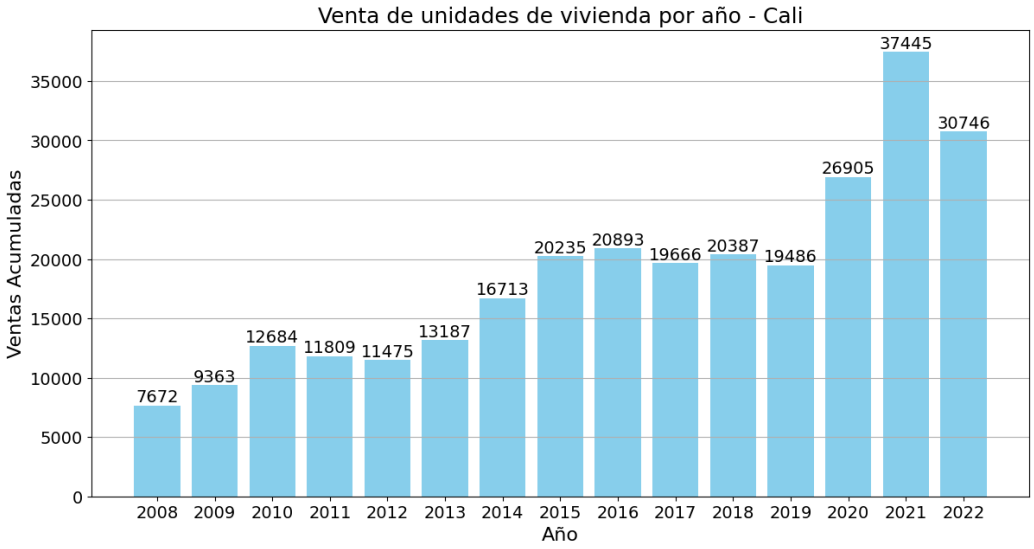


Figura 5. Evolución de las ventas anuales de nuevas unidades de vivienda en Cali.

Para el análisis de descomposición estacional del nivel de ventas de unidades de vivienda, se utilizó la función seasonal_decompose de la biblioteca statsmodels (Seabold & Perktold, 2010). Esta técnica se basa en los métodos clásicos de descomposición de series temporales, como se discute en (Chatfield, 2003).

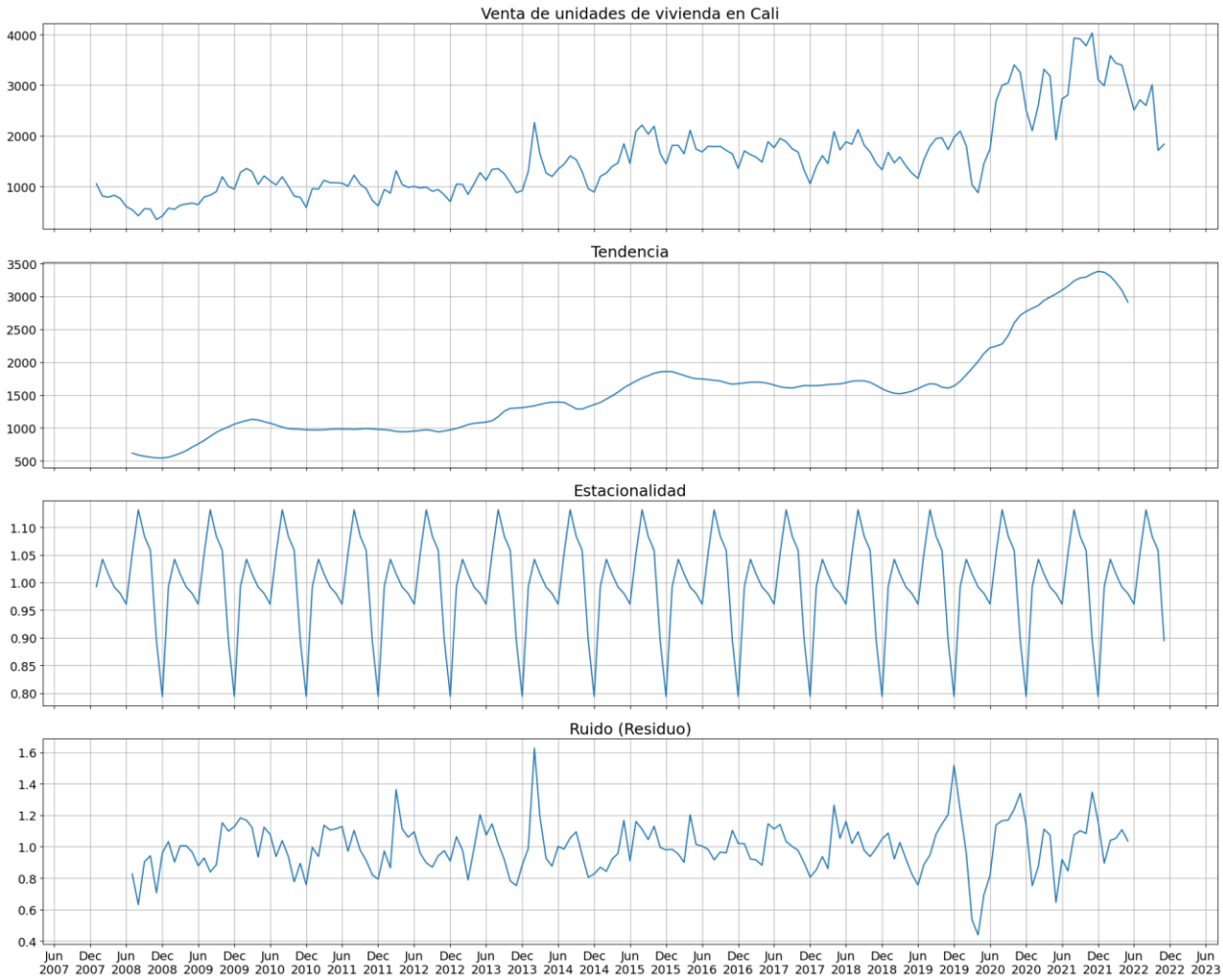


Figura 6. Descomposición estacional y de tendencia de la Serie "VENTAS"

Se confirma que, en efecto, la serie muestra una tendencia creciente, y una estacionalidad marcada que muestra una reducción en la venta de unidades de vivienda al finalizar el año, con dos picos estacionales marcados a lo largo del año. El primero de ellos en febrero y el más alto en agosto. Estos dos picos estacionales se registran al segundo mes del semestre calendario, para después caer al último mes del semestre, aunque la caída en la estacionalidad es más marcada para el segundo semestre del año, siendo diciembre y no junio, como marca esta conducta.

3.1.2 Consultas de Google Trends

De acuerdo con lo publicado en su portal web, Google Trends proporciona acceso a una muestra de búsquedas realizadas en Google, organizadas de manera anónima y agregada, lo que permite el análisis del interés por temas específicos a nivel global o local. Existen dos tipos de muestras disponibles: datos en tiempo real, que comprenden los últimos siete días, y datos históricos, que abarcan desde 2004 hasta 72 horas antes de la consulta. Los datos se normalizan para facilitar comparaciones entre términos y regiones, reflejando el volumen relativo de búsquedas mediante un índice que varía de 0 a 100. Este proceso excluye términos de búsqueda poco frecuentes, consultas duplicadas y aquellas con caracteres especiales.

Para el presente estudio, se definieron una serie de consultas relacionadas en la página https://trends.google.es/home con una serie de términos acotados para la zona geográfica de Valle del Cauca. Los resultados indican la popularidad de cada término respecto al volumen total de búsquedas, lo que perfilaría un aumento o disminución del interés de los usuarios en obtener información disponible en internet sobre dicho tópico.

Se evaluó la consistencia de los resultados de las búsquedas intentando escribir el mismo término de diferentes formas alternando uso de mayúsculas o minúsculas, tal como “casas”, “CASAS”, “Casas” o “CaSaS”, obteniendo resultados similares. Para efectos del presente estudio, los términos de búsqueda fueron escritos en mayúsculas, y fueron los siguientes, junto con las consideraciones del caso:

Tabla 1. Términos de Búsqueda de Google Trends

Término de búsqueda	**Observaciones**
"APARTAMENTOS"	Consulta básica en Google para obtener información general sobre apartamentos, un tipo de infraestructura para vivienda. Puede mostrar interés de cualquier tipo de transacción de compraventa o arriendo en la zona geográfica de todo el departamento.
"CASAS"	Consulta básica en Google para obtener información general sobre casas, un tipo de infraestructura para vivienda. Puede mostrar interés de cualquier tipo de transacción de compraventa o arriendo en la zona geográfica de todo el departamento.
"APARTAMENTOS CALI VENTA"	Consulta más específica que se haría para posibles operaciones de compraventa del tipo apartamentos, delimitados geográficamente para la zona del municipio de Cali.
"CASAS CALI VENTA"	Consulta más específica que se haría para posibles operaciones de compraventa del tipo casas, delimitados geográficamente para la zona del Distrito de Cali.
"PROYECTOS VIVIENDA CALI"	Consulta general para proyectos de vivienda de cualquier tipo, que muestra intención posible compraventa para

	cualquier tipo de infraestructura delimitado geográficamente al Distrito de Cali.
“SUBSIDIO VIVIENDA”	Para el segmento de viviendas de interés social, o para un grupo con restricción de ingresos, el interés o popularidad de la búsqueda de subsidios de vivienda se podría materializar en una posible operación de compraventa.
“CREDITO VIVIENDA”	Consulta que puede manifestar un interés en una operación de compraventa de vivienda mediante la evaluación de requisitos para acceder a un crédito de vivienda.
“HIPOTECA”	Consulta con la misma orientación de crédito de vivienda, entendiendo que, por lo general, este tipo de operaciones dejan como garantía de pago la titularidad del inmueble.
“CONSTRUCTORA BOLIVAR”	Consulta que puede revelar interés frente a proyectos de “Constructora Bolívar” en el área geográfica del departamento del Valle del Cauca. Se usa el nombre completo, dado que “Bolívar” por sí solo puede revelar interés en otros temas, tales como la divisa de Venezuela.
“CONSTRUCTORA MELENDEZ”	Consulta que puede revelar interés frente a proyectos de “Constructora Bolívar” en el área geográfica del departamento del Valle del Cauca. Se busca con “constructora” intentando eliminar el efecto de la palabra

	"Meléndez" que puede representar intereses por otros temas relacionados con este apellido.
"MARVAL"	Consulta que puede revelar interés frente a proyectos de "Constructora Marín Valencia" en el área geográfica del departamento del Valle del Cauca. Se usa "MARVAL", en vista que las consultas estructuradas con este término sólo harían referencia a dicha constructora.
"JARAMILLO MORA"	Consulta que puede revelar interés frente a proyectos de "Constructora Jaramillo Mora" en el área geográfica del departamento del Valle del Cauca. Se usa la combinación "JARAMILLO MORA" asumiendo que la mayoría de las búsquedas relacionadas con este término harían referencia a información de esta constructora.

Fuente: Elaboración propia

Por otro lado, es importante señalar que el algoritmo que genera el nivel de popularidad relativa de una consulta cambia constantemente, por lo que, para efectos del presente proyecto, se ejecutó una consulta usando la librería pytrends a fecha 9 de junio de 2024. Esta consulta se exportó a una tabla en formato xslx (Excel) y se integró al dataset. Así mismo, para manejar la consistencia en los datos, se toma solo la serie para donde existen datos disponibles en la variable objetivo.

3.1.2.1 Consultas "APARTAMENTOS" y "CASAS"

Ambas consultas muestran una conducta similar en cuanto a tendencia y estacionalidad.

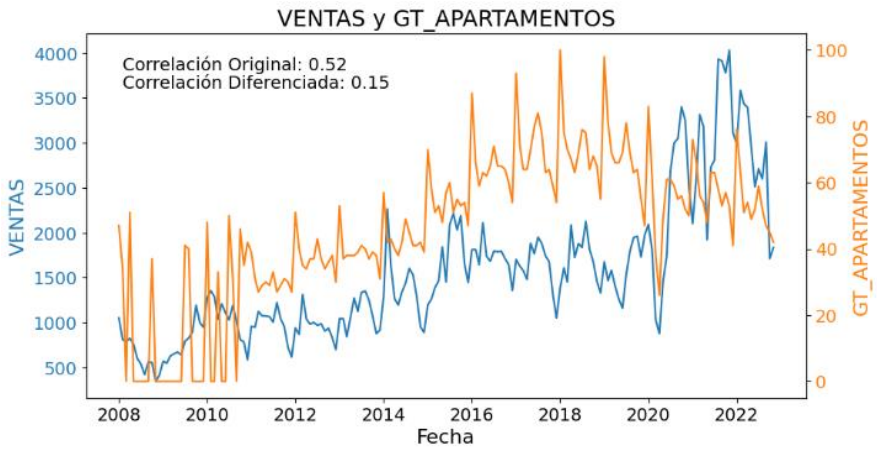


Figura 7. Correlación entre "VENTAS" y "GT_APARTAMENTOS"

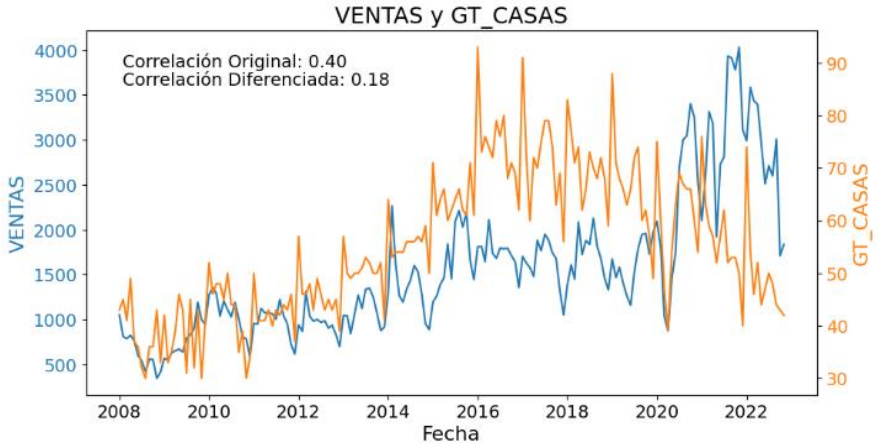


Figura 8. Correlación entre "VENTAS" y "GT_CASAS"

La popularidad relativa de ambas búsquedas muestra una tendencia creciente hasta aproximadamente 2016 para "CASAS" y hasta 2018 para "APARTAMENTOS", para estabilizarse y en ambos casos mostrar tendencia decreciente a partir de 2020. La consulta "APARTAMENTOS" muestra una tendencia más estable a partir de la caída de 2020, que se podría asociar a la pandemia COVID-19, pero en ambos casos, aplicando la descomposición estacional, se muestra que ambas series marcan estacionalidad valle en diciembre y pico en enero. Esto se puede deber a que a

principios de año se puede estar registrando una dinámica importante en las búsquedas asociadas a “APARTAMENTOS” o “CASAS”, sea por renovación de contratos de arriendo o por un interés de este tipo de consultas en épocas de vacaciones de principios de año. Así mismo, se destaca que la consulta “APARTAMENTOS” registra valores con oscilaciones muy altas en los tres primeros años de la serie, lo que no ocurre con la consulta “CASAS”.

Finalmente, en ambos casos, se calcula la correlación entre las series VENTA y las dos series en análisis, “APARTAMENTOS” y “CASAS”, presentando correlaciones de 0.52 y 0.40 respectivamente a nivel y de 0.15 y 0.18 para todas las series en primeras diferencias. El método de primeras diferencias se utiliza para eliminar la estacionalidad y la tendencia, lo que permite eliminar una relación espuria marcada por una conducta temporal determinada (Granger & Newbold, 1974). Los signos de correlación son los teóricamente esperados, aunque hay que considerar que estas consultas como se señaló en la tabla 1 pueden mostrar interés de cualquier tipo de transacción de compraventa o arriendo en la zona geográfica de todo el departamento del Valle del Cauca.

3.1.2.2 Consultas “APARTAMENTOS CALI VENTA” y “CASAS CALI VENTA”

Esta consulta se diferencia de las anteriores “APARTAMENTOS” y “CASAS” dado que podría delimitar el interés a sólo operaciones de compraventa de bienes inmuebles al interior del distrito de Cali. Al igual que las consultas anteriores registran relevancias bajas los tres primeros años de la serie, y en el caso de “CASAS CALI VENTA” mantiene el mismo nivel de baja en la popularidad relativa a partir del año 2020.

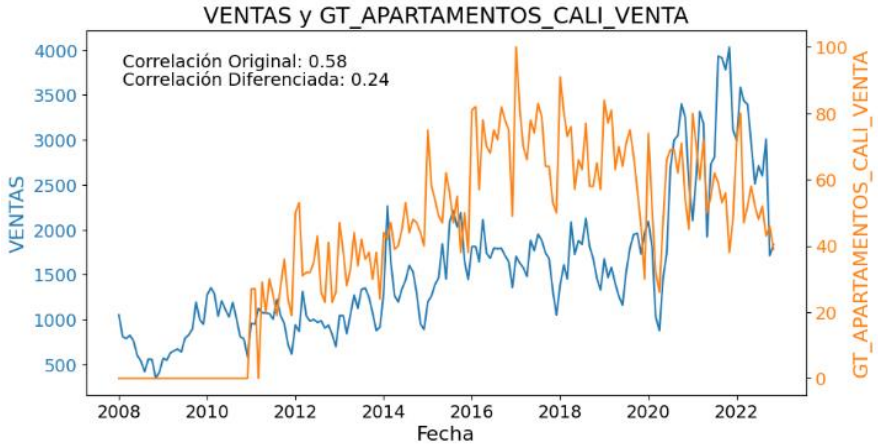


Figura 9. Correlación entre "VENTAS" y GT_APARTAMENTOS_CALI_VENTA"

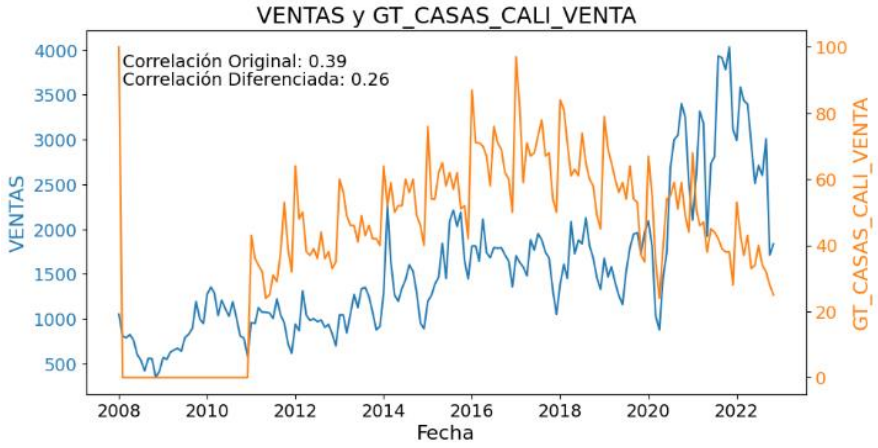


Figura 10. Correlación entre "VENTAS" y "GT_CASAS_CALI_VENTA"

Ambas consultas registran estacionalidad, siendo diciembre los de menor relevancia, para presentar un pico durante enero. Finalmente, se destaca que las correlaciones de las series en primeras diferencias respecto a la variable objetivo "VENTAS" son positivas en 0.24 y en 0.26 respectivamente. Siguen siendo correlaciones débiles, más altas que las consultas básicas "APARTAMENTOS" y "CASAS", pero con la relación o signo positivo, lo que era lo teóricamente esperado.

3.1.2.3 Consultas “PROYECTOS VIVIENDA CALI” y “SUBSIDIO VIVIENDA”

Las dos consultas son estructuralmente diferentes a las presentadas anteriormente. Aquí se sondea el potencial interés en la compra específicamente de vivienda nueva dentro del Distrito de Cali, y se adiciona el interés por el subsidio de vivienda, que generalmente aplicaría para compra de unidades nuevas de vivienda del segmento VIS (Vivienda de Interés Social):

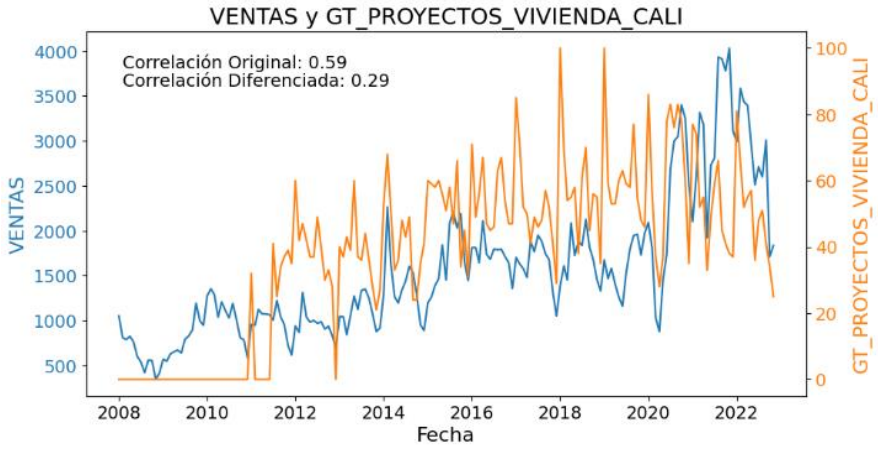


Figura 11. Correlación entre “VENTAS” y “GT_PROYECTOS_VIVIENDA_CALI”

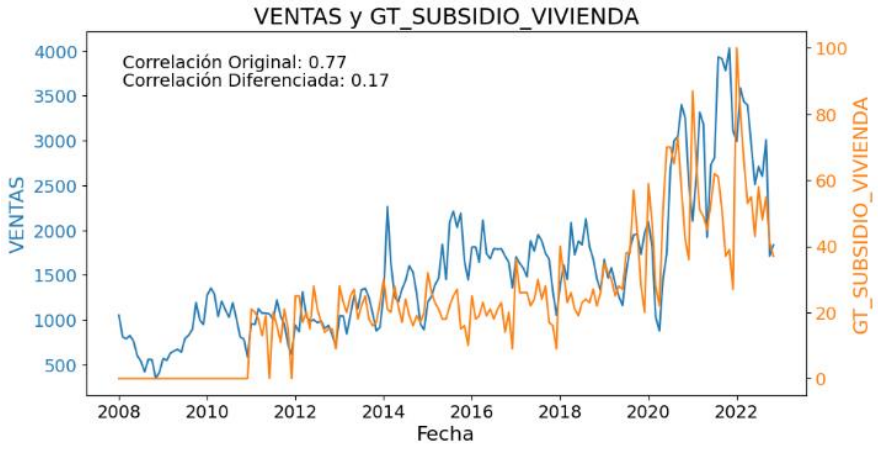


Figura 12. Correlación entre “VENTAS” y “GT_SUBSIDIO_VIVIENDA”

Tanto la consulta “PROYECTOS VIVIENDA CALI” junto con “SUBSIDIO DE VIVIENDA” presentan correlaciones positivas respecto a la variable de “VENTAS”, con valores de 0.29 y 0.17

en las series de tiempo de primeras diferencias, lo que es teóricamente lo esperado. Además, tienen en común con las cuatro consultas anteriores, que la estacionalidad muestra picos en los meses de enero y valle en el mes de diciembre.

3.1.2.4 Consultas “CREDITO VIVIENDA” e “HIPOTECA”

La popularidad relativa a términos que describan la potencial financiación para la adquisición de una vivienda puede considerarse dentro de las opciones para la predicción de las ventas. Aquí entran en juego las consultas “CREDITO VIVIENDA” e “HIPOTECA”, en el mismo sentido de (Bulczak, 2021). Los resultados respecto a las ventas son los siguientes:

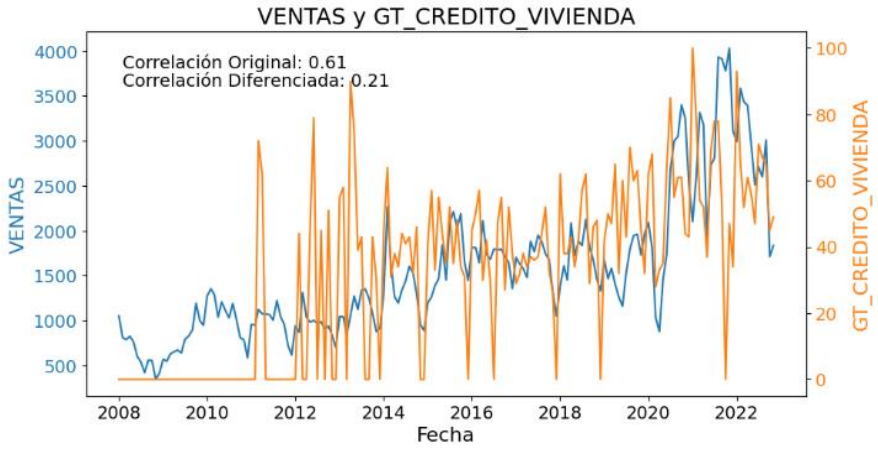


Figura 13. Correlación entre “VENTAS” y “GT_CREDITO_VIVIENDA”

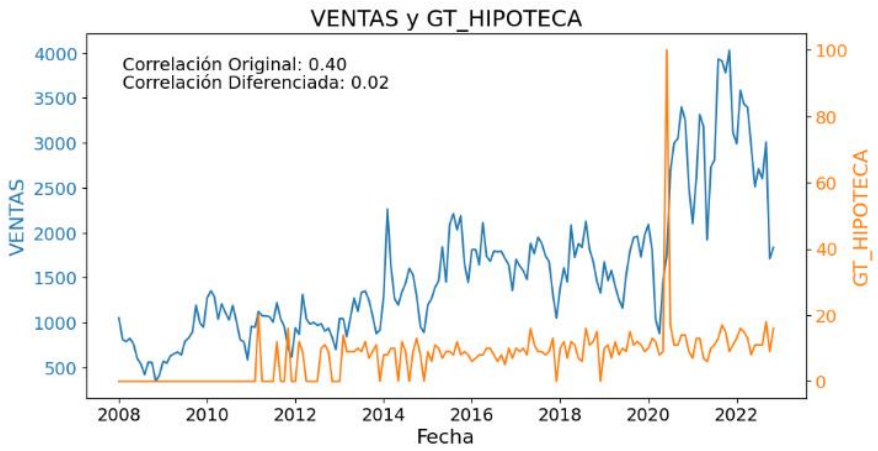


Figura 14. Correlación entre "VENTAS" y "GT_HIPOTECA"

El término "CREDITO VIVIENDA" para la zona geográfica de estudio es mucho más relevante que "HIPOTECA". Se observa que existe una correlación positiva entre "CREDITO VIVIENDA" respecto a la serie de "VENTAS" de 0,21 (Series en primeras diferencias), mientras que esta relación con el término de búsqueda "HIPOTECA" es casi que irrelevante, 0.02 para la serie en primeras diferencias. Por este motivo, este término de búsqueda podría ser el primero en quedar fuera de la evaluación para los modelos.

En cuanto a la estacionalidad, es marcado el pico estacional en el mes de febrero para "CREDITO DE VIVIENDA" y en junio para "HIPOTECA" mientras que el valle estacional para la popularidad relativa de ambos términos de consulta es, al igual que el de todas las consultas evaluadas hasta ahora, en el mes de diciembre.

3.1.2.5 Consultas referentes a empresas constructoras que operan en el distrito de Cali

Según lo expresado anteriormente, otro método que aproxima el interés en la adquisición de vivienda nueva puede asociarse a la popularidad relativa del nombre de empresas constructoras de

proyectos inmobiliarios. Los términos de consulta “CONSTUCTORA BOLIVAR”, “CONSTRUCTORA MELENDEZ”, “MARVAL” y “JARAMILLO MORA” se evalúan con los siguientes resultados:

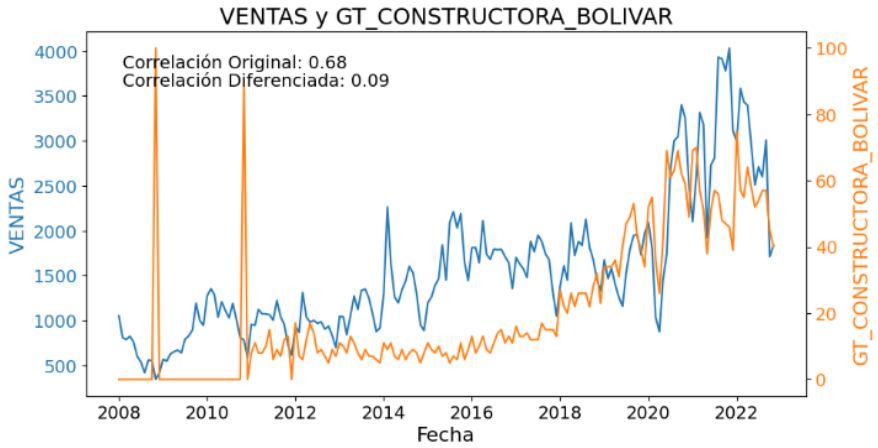


Figura 15. Correlación entre “VENTAS” y “GT_CONSTRUCTORA_BOLIVAR”

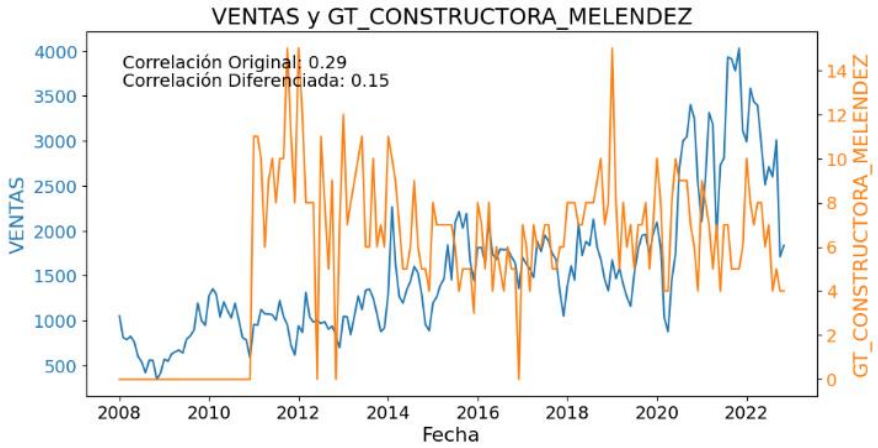


Figura 16. Correlación entre “VENTAS” y “GT_CONSTRUCTORA_MELENDEZ”

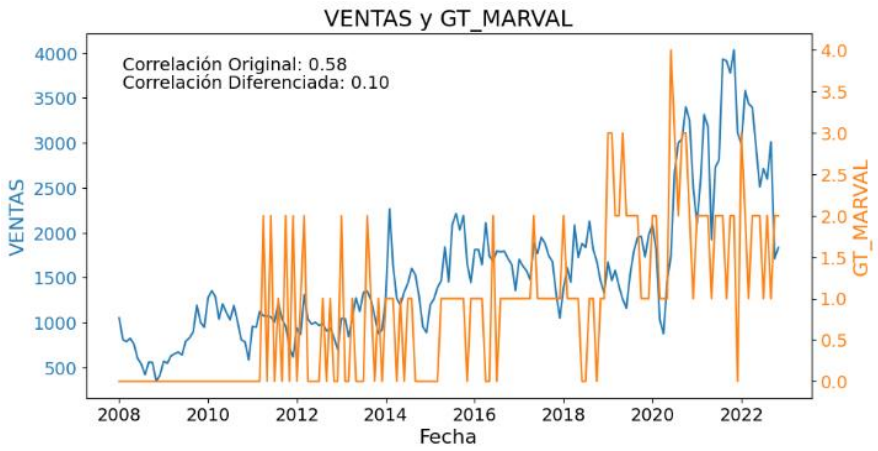


Figura 17. Correlación entre "VENTAS" y "GT_MARVAL"

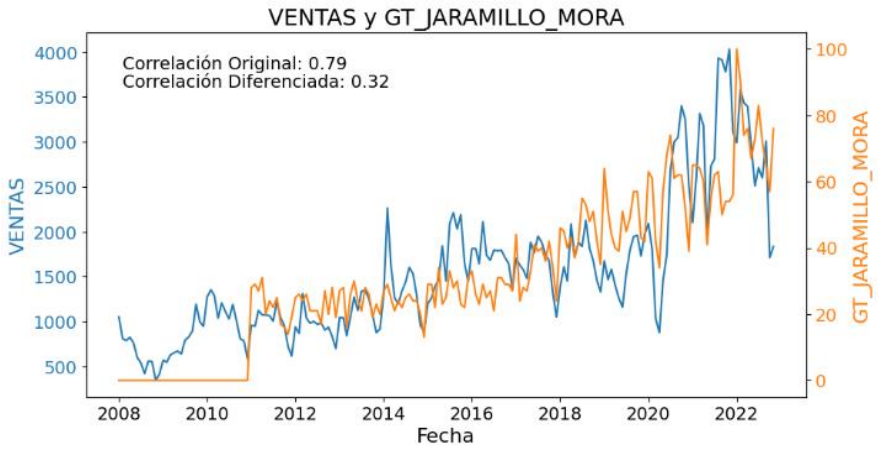


Figura 18. Correlación entre "VENTAS" y "GT_JARAMILLO_MORA"

Desde el punto de vista visual, se puede percibir una mejor correlación positiva entre la popularidad relativa de los términos de búsqueda, especialmente de "CONSTRUCTORA BOLIVAR" y "JARAMILLO MORA" respecto al nivel de "VENTAS". Los términos "CONSTRUCTORA MELENDEZ" y "MARVAL" tienen un nivel bajo en la popularidad relativa y correlaciones en las series diferenciadas de 0.15 y 0.1 respectivamente. Ahora bien, sólo el término de búsqueda de

"JARAMILLO MORA" tiene una correlación respecto a la variable objetivo, que es interesante para incluirla en los modelos predictivos.

Así mismo, vale la pena señalar que, en todos los casos, la estacionalidad valle ocurre al igual que el resto de las consultas en el mes de diciembre, mientras que la estacionalidad pico ocurre en el mes de enero.

3.1.3 Desempleo

La relación entre el mercado de vivienda y el desempleo ha sido estudiada en un enfoque diferente a una de las hipótesis presentadas en el presente proyecto. (Oswald, 1996) argumenta que las altas tasas de propiedad de vivienda pueden aumentar el desempleo debido a la menor movilidad de los propietarios de viviendas en comparación con los inquilinos. No obstante, se considera que la relación directa entre el desempleo y el crecimiento económico, en el sentido de la Ley de Okun (Okun, 1962). Este precepto establece que por cada 1% de aumento en la tasa de desempleo, el producto interno bruto (PIB) de un país se reduce aproximadamente un 3% por debajo de su potencial.

Lo anterior, ya que se plantea que la dinámica en la venta de unidades de vivienda responde positivamente al crecimiento económico, y por el efecto que el desempleo pueda tener sobre él, se busca que funcione como predictor, dado que el DANE mide mensualmente a partir de la Gran Encuesta Integrada de Hogares (GEIH).

Al respecto se analiza la tasa de ocupación, que de acuerdo con el DANE se define como las personas ocupadas sobre la PET (Población en Edad de Trabajar), y la tasa de desempleo, que se define como la población desempleada sobre la PEA (Población Económicamente Activa). Ambos

indicadores están altamente correlacionados entre ellos. Visualmente se aprecia de la siguiente forma:

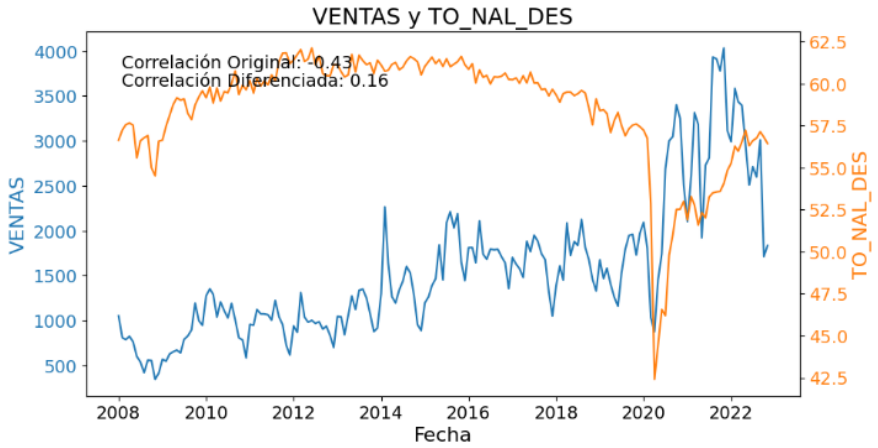


Figura 19. Correlación entre "VENTAS" y "TO_NAL_DES" (Tasa de Ocupación)

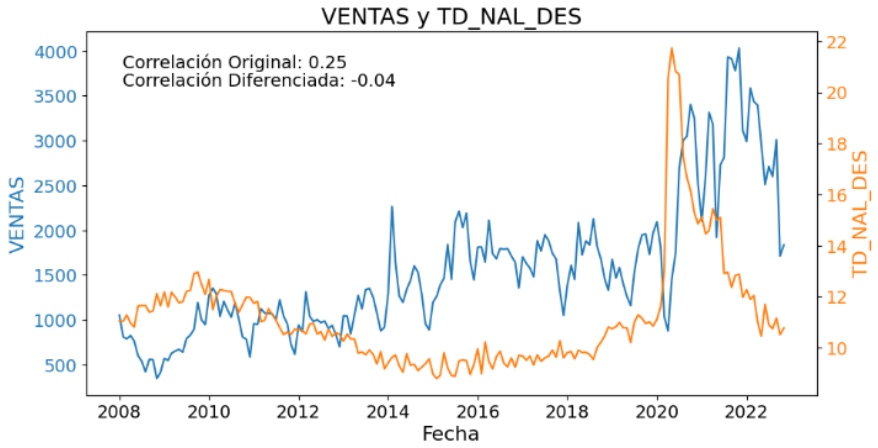


Figura 20. Correlación entre "VENTAS" y "TD_NAL_DES" (Tasa de Desempleo)

Analizando las correlaciones de las primeras diferencias versus la variable objetivo "VENTAS", se registra una correlación débil de 0.16 para la tasa de ocupación, mientras que la tasa de desempleo muestra un nivel de correlación más débil aún, de -0.04, pero por lo menos, el signo es teóricamente correcto.

3.1.4 Inflación

La inflación, definida como la variación de los índices de precios al consumidor, puede afectar negativamente al mercado de vivienda. La inflación afecta negativamente la capacidad adquisitiva y aumenta el costo vía tasas de interés real, impactando negativamente la intención de un potencial comprador de una nueva unidad de vivienda (Himmelberg, Mayer, & Sinai, 2005).

Para el presente proyecto se utiliza la variación del IPC para Cali tanto del periodo mensual como de la variación acumulada de los últimos 12 meses también para la misma área.

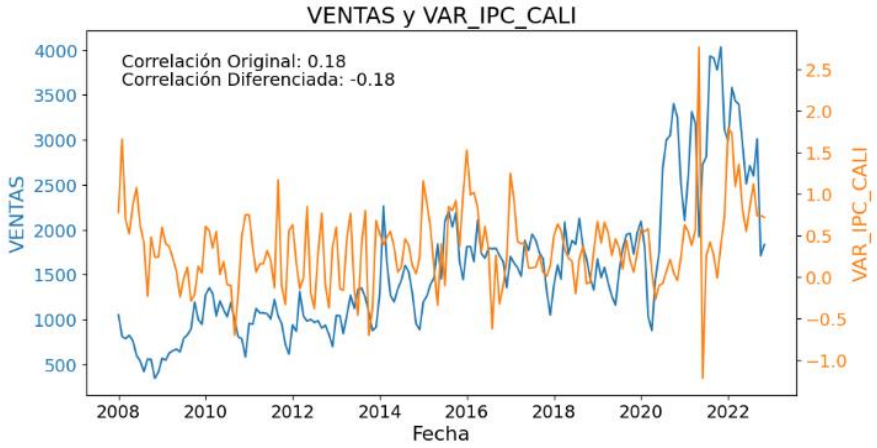


Figura 21. Correlación entre "VENTAS" y "VAR_IPC_CALI" (Variación mensual IPC para Cali)

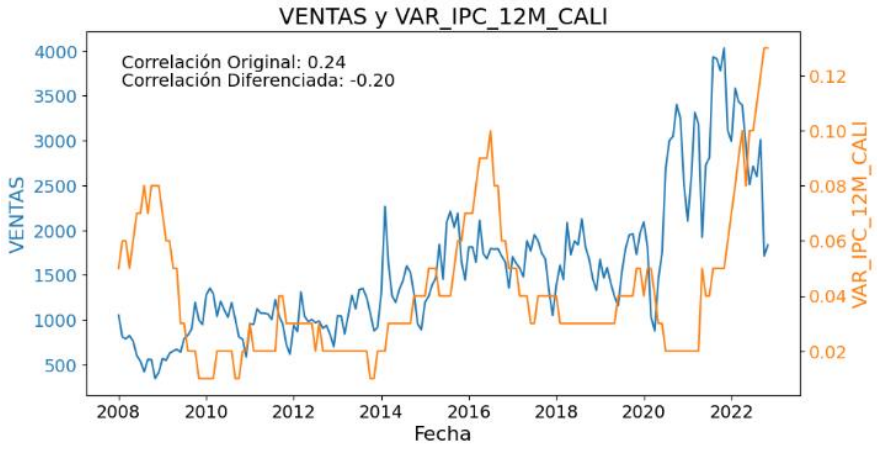


Figura 22.. Correlación entre "VENTAS" y "VAR_IPC_12M_CALI" (Variación acumulada últimos 12 meses IPC para Cali)

La correlación usando ambas series en primeras diferencias, respecto a las unidades de vivienda vendidas “VENTAS”, es de -0.18 para la variación mensual del IPC y de -0.2 para la variación acumulada del IPC de los últimos 18 meses. Aunque las correlaciones son débiles, estos posibles predictores están bien orientados teóricamente, dado el signo negativo. La variación del IPC registra un pico estacional en enero y un valle estacional en octubre, mientras que la variación acumulada a 12 meses registra picos en abril y valles en noviembre. Esto tiene sentido observando la dinámica de variación del IPC, siempre mayor durante el primer trimestre del año y con un comportamiento cíclico que muestra una caída hasta noviembre, donde se reactiva a partir de la temporada decembrina.

3.1.5 Percepción sobre la economía – ICC e IDCV

El Índice de Confianza del Consumidor (ICC) de Fedesarrollo mide la percepción de los consumidores colombianos sobre la situación económica actual y las expectativas a futuro. Este índice se obtiene de la Encuesta de Opinión del Consumidor (EOC), basada en varias preguntas que indagan sobre las condiciones económicas del hogar y del país. El ICC se calcula a partir del balance de respuestas a estas preguntas, donde el balance es la diferencia entre el porcentaje de respuestas positivas y negativas, proporcionando una medida neta de la confianza del consumidor.

Por su parte, el más específico Índice de Disposición a Comprar Vivienda (IDCV) mide la percepción de los consumidores colombianos sobre si es un buen o mal momento para comprar vivienda. Este índice se obtiene de la Encuesta de Opinión del Consumidor (EOC), basada en la pregunta: "¿Cree usted que este es un buen momento o un mal momento para comprar vivienda?". El resultado es el balance entre las respuestas positivas y negativas, proporcionando una medida neta de disposición a comprar vivienda (Fedesarrollo - Centro de Investigación Económica y

Social, 2017).

El ICC y el IDCV para este estudio pueden ser de utilidad y predictor, ya que anticipar el comportamiento del mercado de unidades de vivienda, a partir de dar una idea de la orientación de las expectativas de los consumidores.

De acuerdo con los datos publicados por Fedesarrollo en su portal web[2], se procedió a extraer la mayor cantidad de datos posibles de los informes en publicados en formato PDF para el ICC y el IDCV específicos para Cali. No obstante, las publicaciones son discontinuas y en aras de obtener una serie con mayor cantidad de datos, se procedió a efectuar una imputación utilizando el método de interpolación lineal, obteniendo el siguiente resultado:

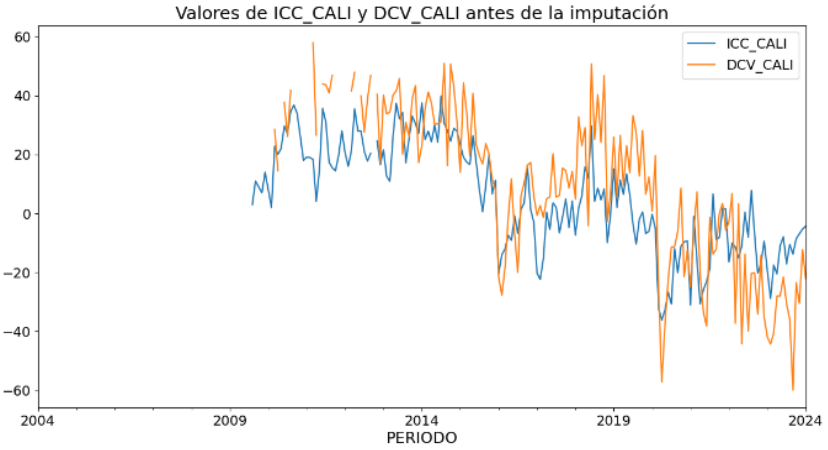


Figura 23. Valores de los agregados ICC y IDVC para Cali, antes de imputar datos.

[2] https://www.fedesarrollo.org.co/p/publicaciones

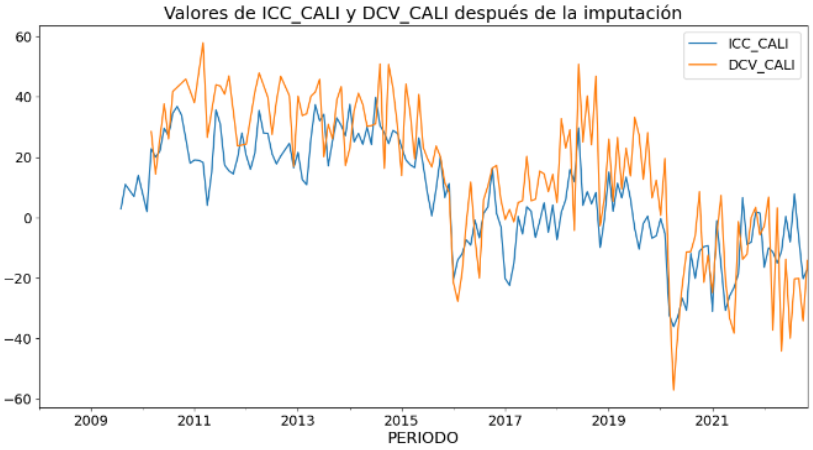


Figura 24. Valores de los agregados ICC y IDVC para Cali, después de imputar datos con interpolación lineal

Así las cosas, el ICC y el IDCV específico para Cali, se grafican junto con la evolución de las unidades de vivienda vendidas "VENTAS" obteniendo el siguiente resultado:

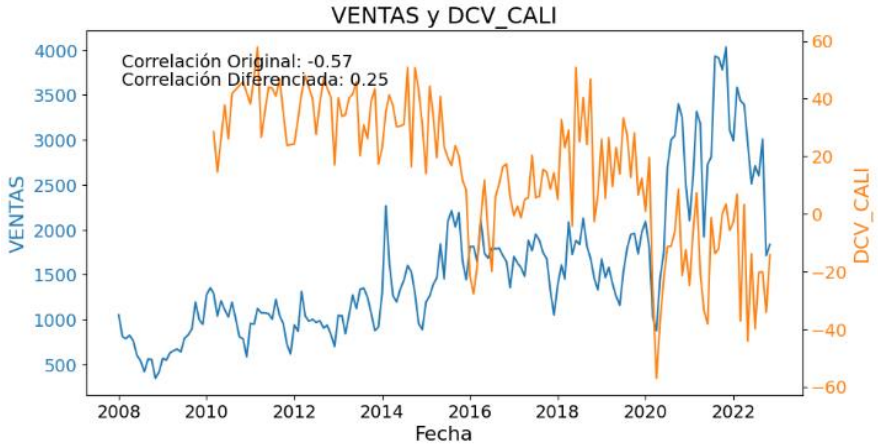


Figura 25. Correlación entre "VENTAS" y "DCV_CALI" (Índice de Disposición a Comprar Vivienda – Cali)

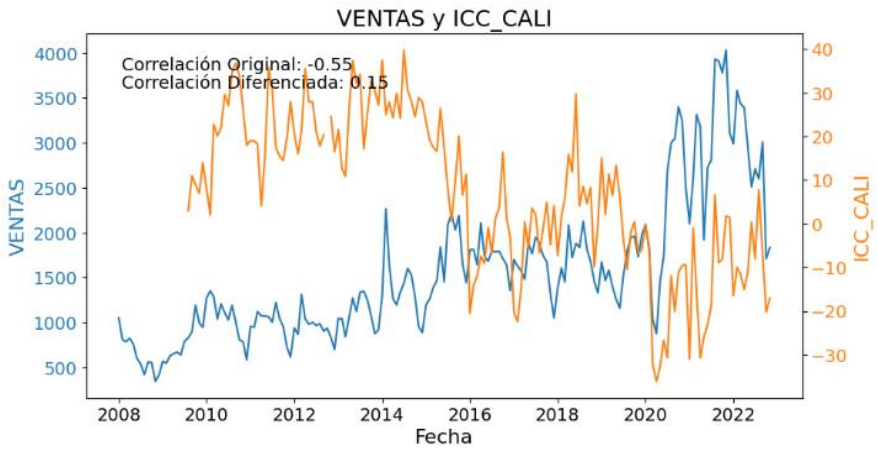


Figura 26. Correlación entre "VENTAS" e "ICC_CALI" (Índice de Confianza del Consumidor – Cali)

Al principio no tendría sentido teórico, ya que la disposición a comprar vivienda cae en indicador nacional y el de Cali, pero al evaluar las series a primeras diferencias, aparece una correlación positiva que puede aumentar la calidad de este indicador como predictor en los modelos a desarrollar. Como elemento adicional, además de la tendencia, ambas series muestran una estacionalidad marcada. Ahora bien, el uso de estos predictores puede no ser determinante por la poca cantidad de datos. Por ejemplo, el IDCV para Cali contaba con solo 144 datos después de la imputación por interpolación líneal, 35 datos menos que la variable objetivo, motivo por el que se usarán con cuidado en el vector de determinantes de cada modelo.

3.1.6 Tasas de cambio y tasas de interés

La evolución en el valor de la tasa de cambio real puede tener varios efectos en el mercado de vivienda. En primer lugar, las depreciaciones del tipo de cambio pueden aumentar la competitividad de la oferta de vivienda en la ciudad respecto al resto del mundo, aunque también podría fijarse como una inversión financiera por parte de un agente económico extranjero que valora a la tasa de cambio directamente con el costo financiero de su inversión. (Rodríguez &

Bustillo, 2010). Sin embargo, también podrían impactar en el costo de los insumos de la construcción, afectar el precio de la vivienda al alza y al tiempo reducir su venta.

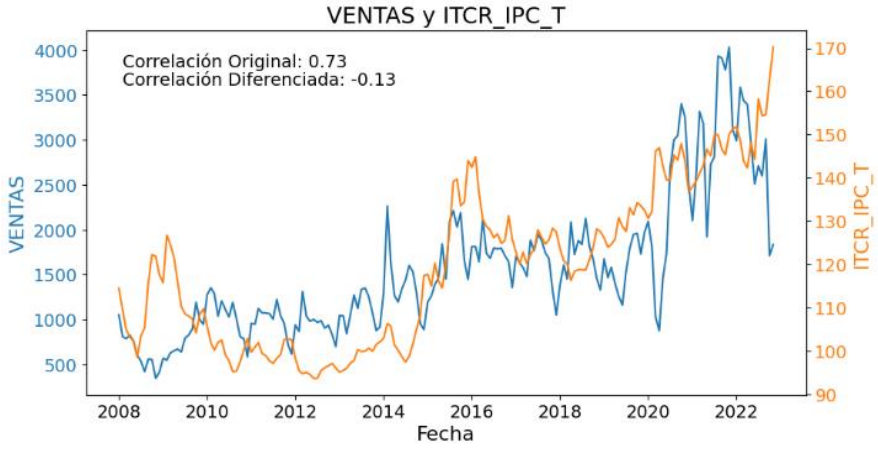


Figura 27. Correlación entre "VENTAS" e "ITCR_IPC_T" (Tasa de Cambio Real)

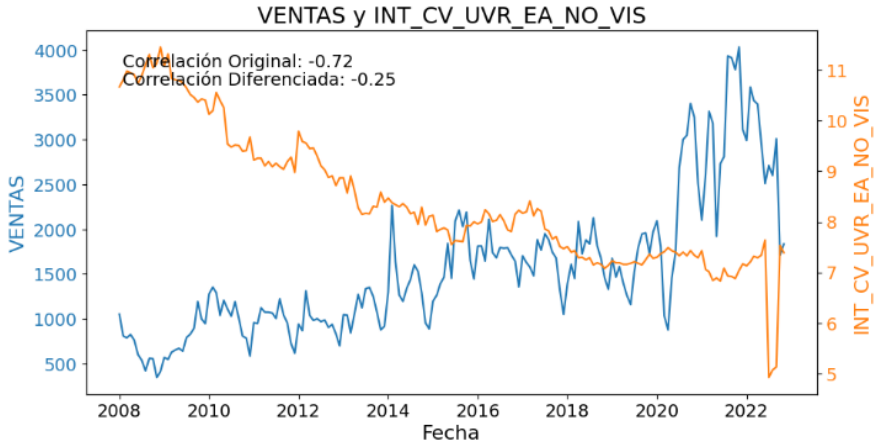


Figura 28. Correlación entre "VENTAS" e "INT_CV_UVR_EA_NO_VIS" (Tasas de Interés)

En el caso de ITCR, se nota una correlación negativa respecto a la venta de unidades de vivienda, por lo que se podría considerar que el efecto es una baja en la venta ante una depreciación del tipo de cambio real.

3.1.7 Índice de Seguimiento a la Economía.

El Indicador de Seguimiento a la Economía (ISE) es una medida mensual del comportamiento agregado de la actividad económica en Colombia, desarrollado por el Departamento Administrativo Nacional de Estadística (DANE). Metodológicamente, el ISE se basa en el marco de las Cuentas Nacionales Trimestrales (CNT) y utiliza 111 indicadores mensuales que representan las nueve actividades económicas principales. Estos indicadores son ponderados según su contribución al valor agregado de la economía (Departamento Administrativo Nacional de Estadística (DANE), 2016).

Las variaciones este Índice, tanto agregado para la economía como para l sector de la construcción se presenta a continuación:

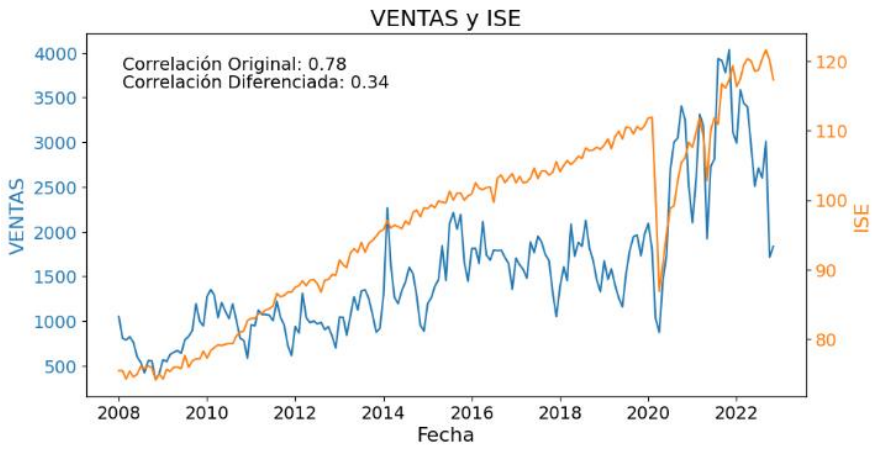


Figura 29. Correlación entre "VENTAS" e "ISE" (Índice de Seguimiento a la Economía)

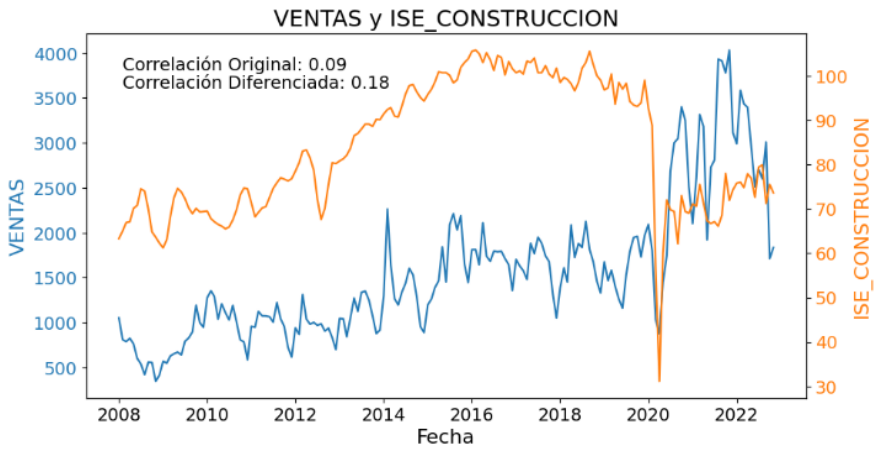


Figura 30. Correlación entre "VENTAS" e "ISE_CONSTRUCCIÓN" (ISE para el sector Construcción)

En el caso del ISE agregado, este mantiene una tendencia creciente, con una caída muy pronunciada en el primer trimestre de 2020, asociado a la contracción económica derivada del aislamiento por la pandemia de COVID-19. Por otro lado, el ISE para la construcción muestra una tendencia decreciente a partir de 2016 y la pronunciada caída en el primer trimestre de 2020 de la cual no se ha podido recuperar. Las correlaciones con las ventas de unidades de vivienda son positivas en ambos casos, con 0.34 y 0.18 para ambos índices respectivamente, lo cual está acorde con lo teóricamente esperado.

3.1.8 Selección de variables.

Al ser series de tiempo, visualmente se puede creer que un predictor afecte la venta de nuevas unidades de vivienda cuando puede no ser así, dado que, fortuitamente, el componente de tendencia y de estacionalidad pueden ser similares.

Tabla 2. Listado de variables preseleccionadas para entrenamiento de modelos

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 179 entries, 2008-01-01 to 2022-11-01
Data columns (total 15 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   VENTAS                              179 non-null    float64
 1   GT_APARTAMENTOS_CALI_VENTA_LAG_0    179 non-null    float64
 2   GT_CASAS_CALI_VENTA_LAG_0           179 non-null    float64
 3   GT_PROYECTOS_VIVIENDA_CALI_LAG_0    179 non-null    float64
 4   GT_CREDITO_VIVIENDA_LAG_0           179 non-null    float64
 5   GT_JARAMILLO_MORA_LAG_0             179 non-null    float64
 6   TD_13_CIUD_DES_LAG_0                179 non-null    float64
 7   TO_13_CIUD_DES_LAG_0                179 non-null    float64
 8   VAR_IPC_12M_CALI_LAG_0              179 non-null    float64
 9   ICC_CALI_LAG_0                      159 non-null    float64
 10  DCV_CALI_LAG_0                      153 non-null    float64
 11  ITCR_IPC_T_LAG_0                    179 non-null    float64
 12  INT_CV_UVR_EA_NO_VIS_LAG_0          179 non-null    float64
 13  ISE_LAG_0                           179 non-null    float64
 14  ISE_CONSTRUCCION_LAG_0              179 non-null    float64
dtypes: float64(15)
memory usage: 22.4 KB
```

La totalidad de los datos de las series de tiempo son numéricas. Así mismo, se efectúa una preselección de 14 variables predictoras además de la variable objetivo. Dentro del preprocesamiento, se segmentaron las observaciones de todos los predictores a la disponibilidad de los datos de "VENTAS". Dentro de este set de datos se encuentran 5 consultas de Google Trends (Precesidas por "GT_"), dos asociadas al desempleo, la tasa de ocupación y la tasa de desempleo a nivel nacional, una de inflación que es la variación acumulada 12 meses para el área de Cali, dos variables de percepción del consumidor, tanto el ICC como el IDCV para el área de Cali, Una de Tasa de Cambio, en este caso el Indice de Tasa de Cambio Real deflactado por el IPC para exportaciones tradicionales, una de Tasa de Interés, que corresponde al diferencial para créditos en UVR de vivienda no VIS, y finalmente, el Índice de Seguimiento a la Economía ISE agregado y el específico para el sector de la construcción. El rango de tiempo del set de datos es desde enero de 2008 hasta noviembre de 2022. Todas las series de tiempo tienen 179 observaciones a excepción de ICC e ICDV para Cali, con 159 y 153 observaciones respectivamente. Se destaca que todos los

predictores cuentan con datos para el periodo comprendido entre diciembre de 2022 y enero de 2024, mas no la variable objetivo. Esto último permitiría ejecutar un ejercicio de predicción para cada método estadístico seleccionado, lo que permitirá tener luces sobre la orientación de la herramienta a utilizar.

Las correlaciones a nivel de todas las variables del set de datos depurado, son las siguientes:

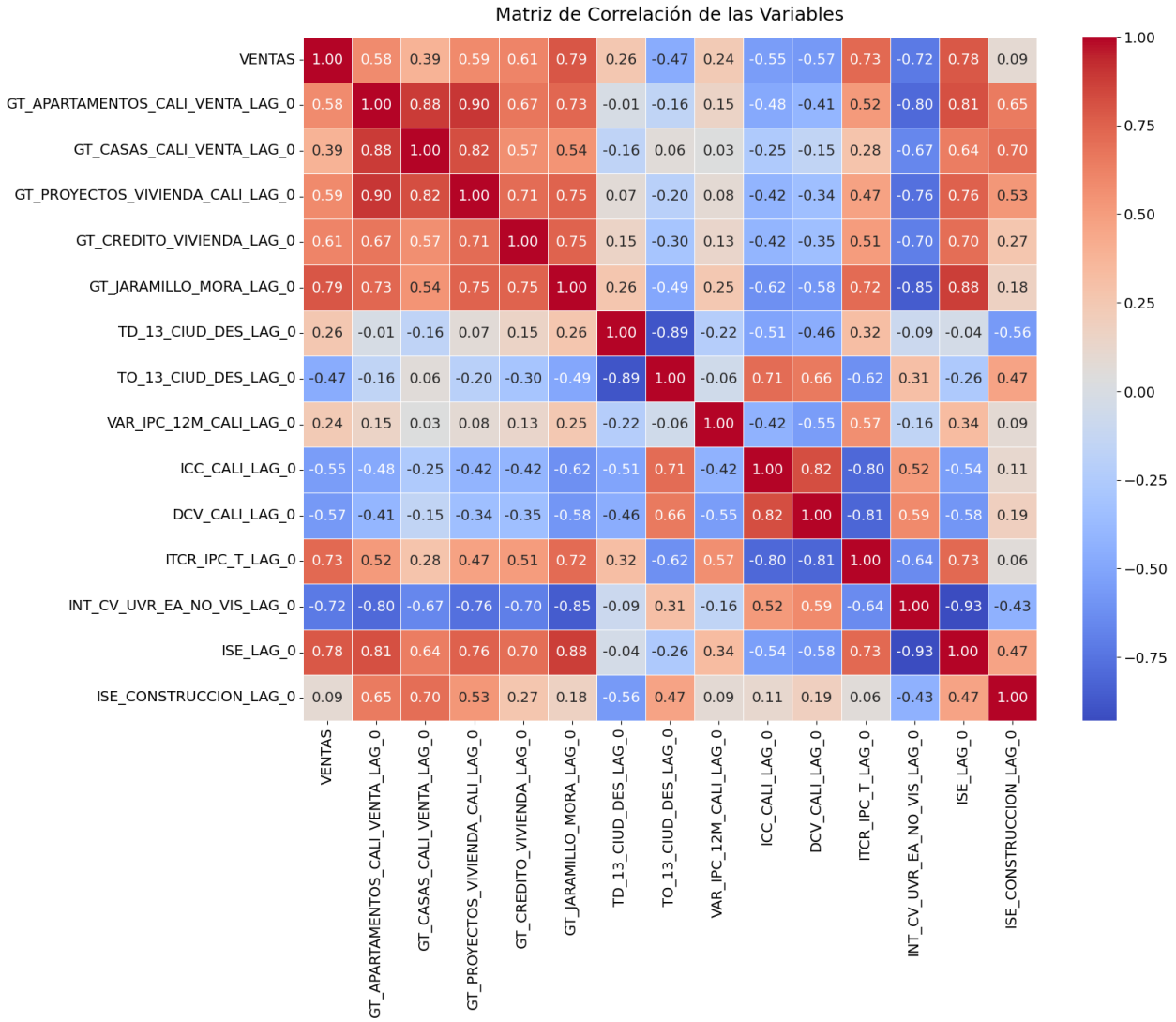


Figura 31. Matriz de correlaciones de variables preseleccionadas a nivel

Se notan altas correlaciones en las búsquedas de Gogole Trends, así como entre las tasas de ocupación y de desempleo. Ahora bien, tal y como se verificó visualmente en los gráficos, el ISE tiene una correlación positiva considerable tanto con las ventas de unidades de vivienda como con las variaciones de la popularidad de los términos de consultas de trends. Esto lo explica el fuerte componente de tendencia que reviste esta serie de tiempo. La presente investigación hace foco más sobre el efecto que pueden tener las variables de consulta sobre las ventas, pero no deja de lado el efecto que pueden llegar a tener las variables macroeconómicas seleccionadas. Se ubican todas las series en primeras diferencias, y se vuelven a calcular las correlaciones, en un intento de eliminar el componente tendencial y estacional, con el siguiente resultado:

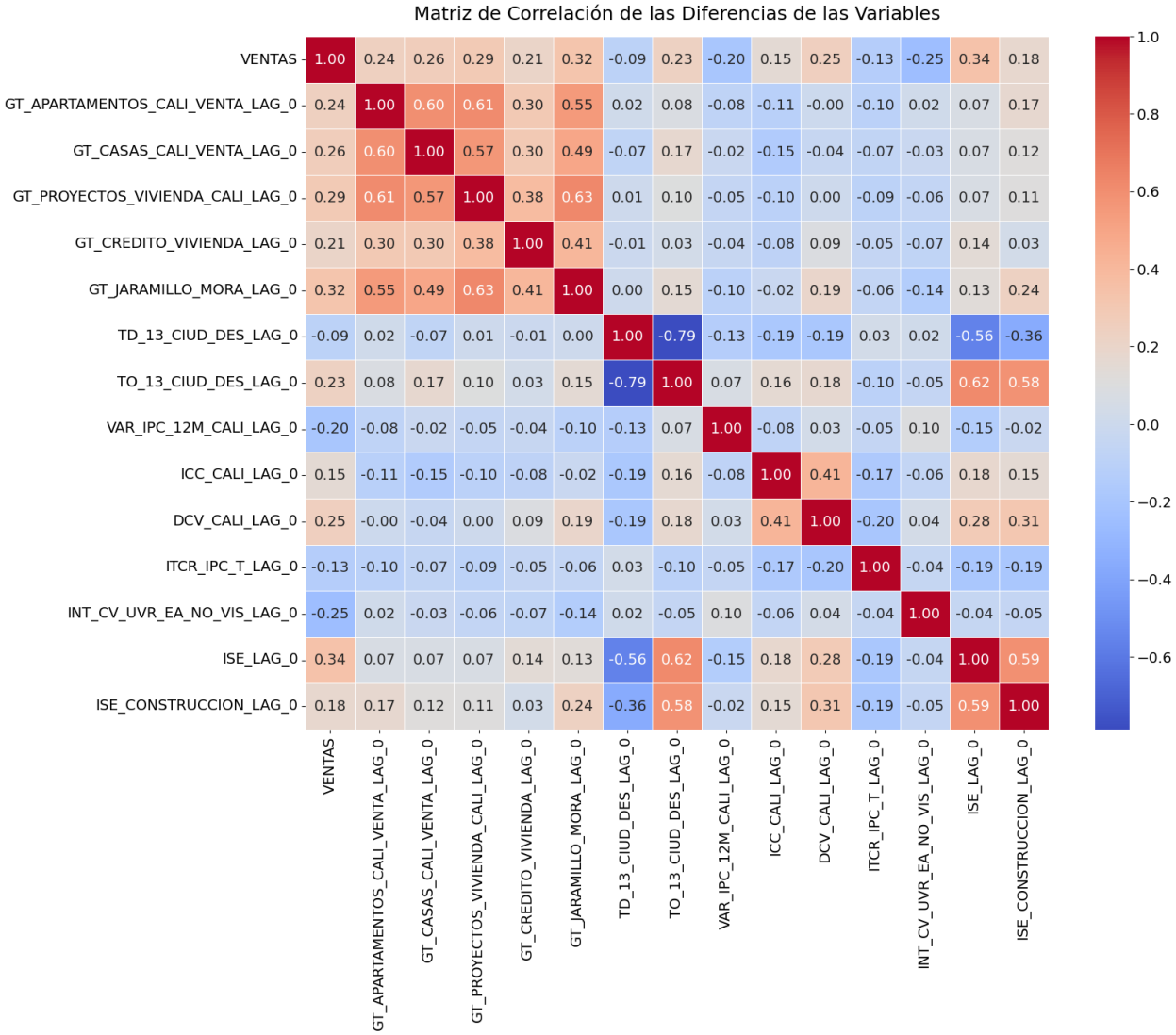


Figura 32. Matriz de correlaciones de variables preseleccionadas en primeras diferencias

Aquí las consultas GT siguen marcando una correlación relativamente alta, pero este gráfico ya perfila que algunos determinantes deben entrar con cuidado dado que tienen un nivel de correlación alto entre ellos, especialmente la tasa de ocupación y la tasa de desempleo.

Por otro lado, en vista de que se cuenta con pocos datos, las variables ICC_CALI y DCV_CALI posiblemente deban dejarse de lado en algunas simulaciones, en búsqueda de un mejor nivel de

entrenamiento de los modelos. Aquí se perfilaría uno de los primeros trabajos futuros y es correr las simulaciones con los datos completos y a diferentes niveles de rezagos.

También se descarta la posibilidad de rezagar las variables para no perder grados de libertad, frente a la restricción en la cantidad de datos que se disponen.

5. DESARROLLO DE LOS MODELOS ESTADÍSTICOS

Para el desarrollo de todos los modelos, se construyó un notebook en lenguaje Python utilizando como como IDE a Google colab. El set de datos, se compiló manualmente en formato Excel, para luego ser exportado a csv, cargado en Google Drive, publicado e integrado al notebook usando "requests".

Además de las bibliotecas pandas y numpy, para el manejo numérico y de manipulación y limpieza de los dataframes temporales creados a lo largo del ejercicio, se utilizaron las siguientes bibliotecas:

Tabla 3. Bibliotecas utilizadas y el uso en el proyecto.

Biblioteca	**Uso en el proyecto.**
matplotlib y seaborn	matplotlib es una biblioteca para la visualización de datos en Python. El módulo pyplot proporciona una interfaz similar a MATLAB para crear gráficos y visualizaciones de datos. seaborn es una biblioteca de visualización de datos basada en matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos. Se usaron como base para visualización de resultados de todos los modelos.
tensorflow y keras	tensorflow es una biblioteca de código abierto para el aprendizaje automático. keras es una API de alto nivel para construir y entrenar modelos de aprendizaje profundo, que ahora se incluye como parte de tensorflow. Se utilizaron para el entrenamiento y validación de

	los modelos con RNN, tanto para LSTM como para GRU
keras-tuner	keras-tuner es una biblioteca para optimizar hiperparámetros en modelos de aprendizaje profundo construidos con keras. Se utilizó para búsqueda de Hiperparámetros en los modelos RNN LSTM y GRU.
sklearn	scikit-learn es una biblioteca para el aprendizaje automático en Python. Proporciona herramientas simples y eficientes para la minería de datos y el análisis de datos. Se utilizaron MinMaxScaler para la normalización, RandomForestRegressor para el entrenamiento del modelo RF, y funciones de evaluación como mean_squared_error y r2_score, además de RandomizedSearchCV para la búsqueda de Hiperparámetros.
xgboost	xgboost es una biblioteca optimizada para el aumento de gradientes, que proporciona un rendimiento y eficiencia de cálculo superiores para tareas de clasificación y regresión. Se utilizó para entrenar el modelo XGboost junto con RandomizedSearchCV de sklearn, para búsqueda de Hiperparámetros.
statsmodels	La biblioteca statsmodels es una herramienta robusta para la estimación y evaluación de modelos estadísticos. Se utiliza principalmente para análisis econométrico y series de tiempo. Base para la estimación del modelo de regresión lineal.

5.1. REDES NEURONALES RECURRENTES RNN

El primero de los modelos en entrenarse fueron los de Redes Neuronales.Ambos modelos se configuraron con Keras Tuner para optimizar sus hiperparámetros, permitiendo la búsqueda automática del número óptimo de unidades en las capas LSTM o GRU y la selección del optimizador más adecuado. La primera capa LSTM del modelo tiene un número de unidades que varía entre 50 y 500, en incrementos de 50. Esta capa está configurada para devolver secuencias completas (return_sequences=True), lo que significa que cada paso de tiempo en la secuencia de entrada produce una secuencia de salida. La segunda capa LSTM o GRU también tiene entre 50 y 500 unidades, pero no devuelve secuencias (return_sequences=False), lo que significa que produce una sola salida en lugar de una secuencia. Después de las capas LSTM, se añaden dos capas densas: la primera con 100 neuronas y la segunda con una sola neurona que corresponde a la salida del modelo.

El modelo se compila utilizando un optimizador seleccionado a partir de una lista de opciones ('adam' y 'rmsprop') definidas por la función hp.Choice. La función de pérdida utilizada es el error cuadrático medio (mean_squared_error), comúnmente empleada en tareas de regresión para minimizar la diferencia entre los valores predichos y los valores reales. Este enfoque de ajuste de hiperparámetros permite una mayor flexibilidad y potencial para mejorar el rendimiento del modelo, adaptándose mejor a las características específicas del conjunto de datos. La combinación de capas LSTM o GRU y densas, junto con la capacidad de ajustar dinámicamente los hiperparámetros, tiene la expectativa de capturar las secuencias temporales y lograr un resultado favorable.

A partir de esta descripción, se debe tener en cuenta que son dos modelos. Las dos capas iniciales son LSTM o GRU. Es una arquitectura sencilla, pero se resume de la siguiente manera, gracias al comando plot_model:

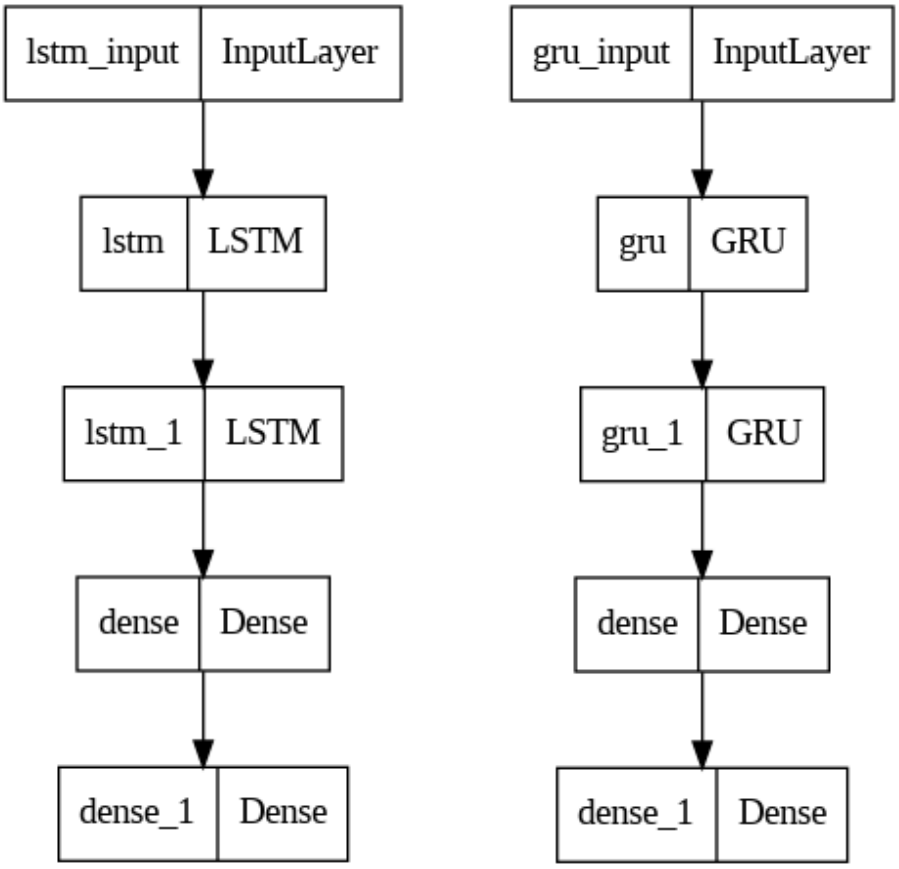


Figura 33. Arquitectura de los modelos de tipo RNN utilizados

El set de datos de entrenamiento para ambos modelos omitió las variables de la tasa de ocupación, por la alta correlación con la tasa de desempleo y los índices ICC e IDCV, por la falta de datos. Cuando se incluyen variables en estos modelos sin datos, los resultados son errados. El set de datos utilizado en ambos modelos es el siguiente:

Tabla 4. Variables seleccionadas para el entrenamiento del modelo de RNN

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 179 entries, 2008-01-01 to 2022-11-01
Data columns (total 12 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   VENTAS                          179 non-null    float64
 1   GT_APARTAMENTOS_CALI_VENTA_lag0 179 non-null    float64
 2   GT_CASAS_CALI_VENTA_lag0        179 non-null    float64
 3   GT_PROYECTOS_VIVIENDA_CALI_lag0 179 non-null    float64
 4   GT_CREDITO_VIVIENDA_lag0        179 non-null    float64
 5   GT_JARAMILLO_MORA_lag0          179 non-null    float64
 6   TD_13_CIUD_DES_lag0             179 non-null    float64
 7   VAR_IPC_12M_CALI_lag0           179 non-null    float64
 8   ITCR_IPC_T_lag0                 179 non-null    float64
 9   INT_CV_UVR_EA_NO_VIS_lag0       179 non-null    float64
 10  ISE_lag0                        179 non-null    float64
 11  ISE_CONSTRUCCION_lag0           179 non-null    float64
dtypes: float64(12)
memory usage: 18.2 KB
```

Finalmente, después del entrenamiento y la validación (20% de los datos), el optimizados para ambos modelos fue adam (Adaptative Moment Estimation), pero para LSTM, el mejor modelo es el de 400 unidades o neuronas, mientras que para GRU, fue el que tenía 450 unidades.

El resultado para la red LSTM se presenta a continuación:

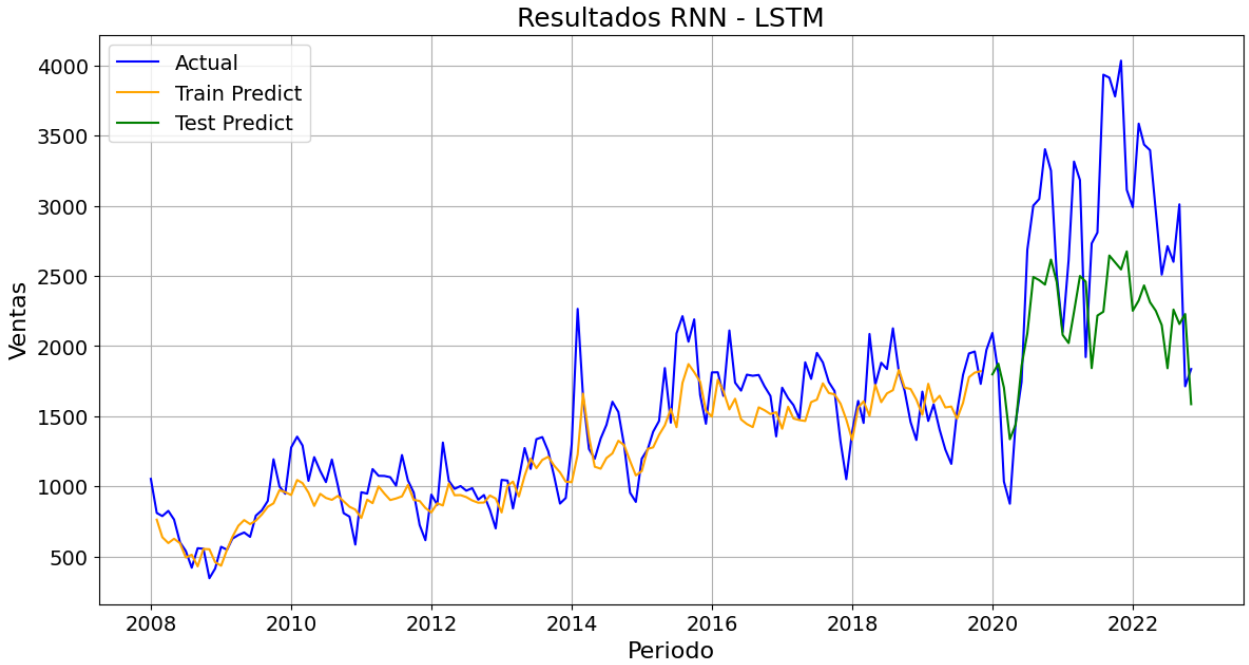


Figura 34. Resultados de las predicciones en los conjuntos de entrenamiento y validación – LSTM

En el caso de la red GRU, los resultados son muy similares:

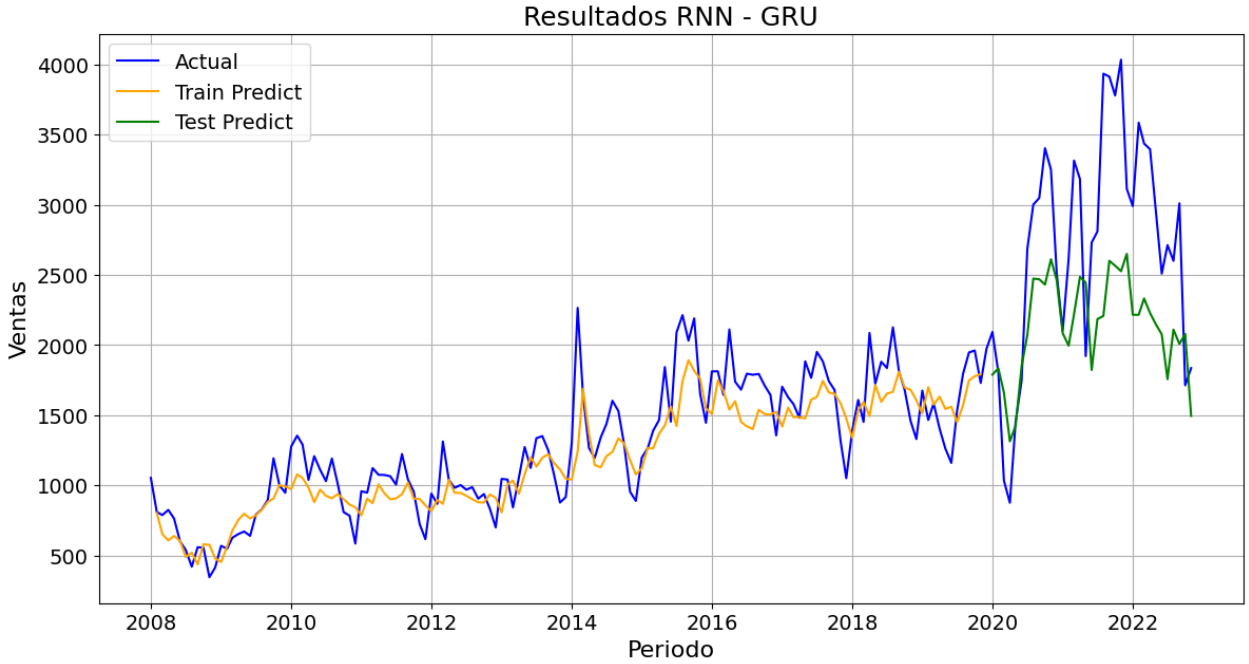


Figura 35. Resultados de las predicciones en los conjuntos de entrenamiento y validación – GRU

Ambos modelos tienen una conducta aceptable en el conjunto de entrenamiento. No obstante, al momento de generalizar, capturan correctamente las variaciones, pero no alcanzan a ser precisos en la predicción del valor exacto. Esto es un avance positivo, porque el sólo capturar la tendencia de para donde va la variable objetivo ya genera confianza sobre la estructura, robustez y orientación metodológica de los modelos.

5.2. RANDOM FOREST

El modelo usa un set de datos similar al que se utilizó para las redes neuronales. No obstante se siguieron los siguientes pasos en la elaboración del mismo:

- Se utiliza MinMaxScaler para normalizar los datos en un rango de [0, 1]. Esta técnica de escalado asegura que todas las características contribuyan de manera equitativa al modelo de aprendizaje y mitiga un poco el dimensionamiento de la escala y la tendencia en las series de tiempo.
- Se define una función create_dataset que crea secuencias de datos (X) y etiquetas (Y) basadas en un valor de look_back de 1. Esto significa que, para cada punto de datos, se usa la observación anterior como entrada para predecir la siguiente observación.
- Se define un rango de hiperparámetros para la búsqueda, incluyendo el número de árboles (n_estimators), características a considerar para la mejor división (max_features), profundidad máxima del árbol (max_depth), número mínimo de muestras requeridas para dividir un nodo (min_samples_split), número mínimo de

muestras por hoja (min_samples_leaf), y si usar bootstrap para muestrear los datos de entrenamiento (bootstrap).

- Se utiliza RandomizedSearchCV con validación cruzada de series temporales (TimeSeriesSplit) para buscar la mejor combinación de hiperparámetros. Se realizan 100 iteraciones de búsqueda aleatoria para encontrar el modelo óptimo.

Los resultados de este modelo son los siguientes:

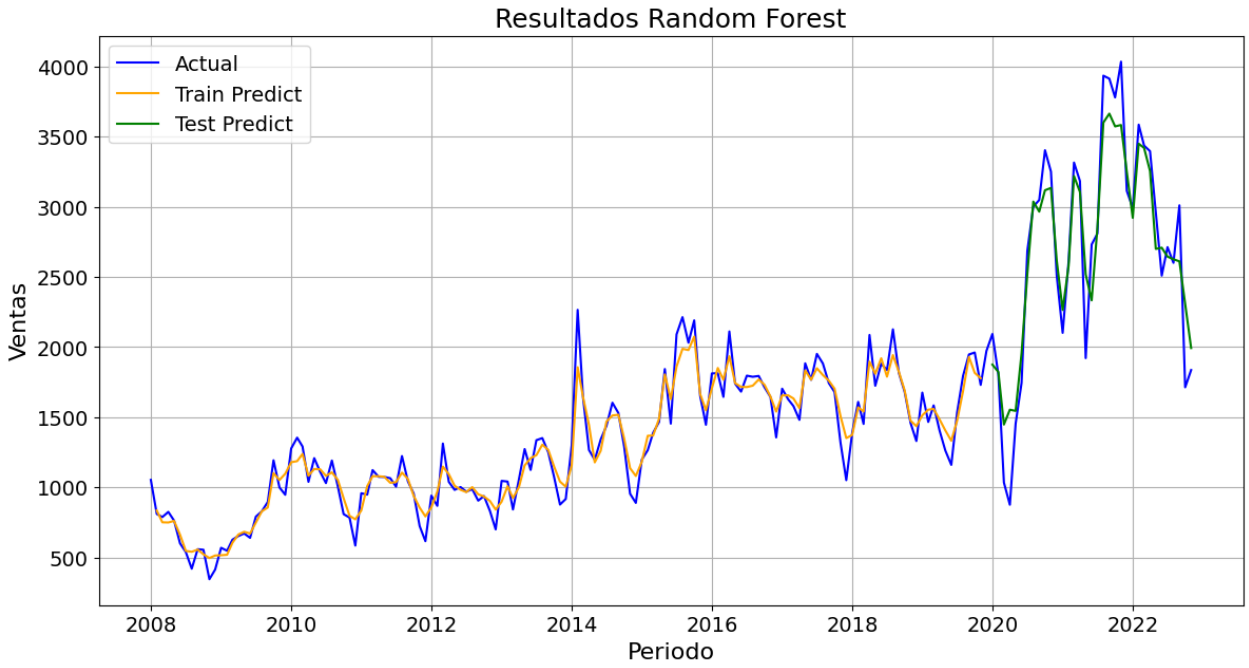


Figura 36. Resultados de las predicciones en los conjuntos de entrenamiento y validación – Random Forest

A primera vista, esta estructura de modelo no solo captura bien los cambios de tendencia en el conjunto de entrenamiento sino que generaliza mucho mejor que las redes neuronales al acercarse mucho más en el conjunto de validación.

5.3. XGBOOST

Se utiliza la misma estructura de set de datos con la que se entraron las redes recurrentes y el modelo de Random Forest. Teniendo en cuenta que este tipo de modelos tiene una arquitectura similar a la del Random Forest, se describen los siguientes pasos:

- Los datos fueron normalizados utilizando la técnica MinMaxScaler para escalar las características al rango [0, 1]. Esto ayuda a mejorar la eficiencia del entrenamiento y a evitar que características con magnitudes más grandes dominen el modelo.}
- Se creó una función para generar conjuntos de datos con características rezagadas (lag features). Esto permite que el modelo XGBoost utilice información de periodos anteriores para hacer predicciones.
- Se definió una gama de valores posibles para los hiperparámetros del modelo, tales como el número de árboles (n_estimators), la profundidad máxima de los árboles (max_depth), la tasa de aprendizaje (learning_rate), y la proporción de muestras y características usadas en cada árbol (subsample y colsample_bytree).
- Se utilizó `RandomizedSearchCV` con validación cruzada temporal (`TimeSeriesSplit`) para evaluar distintas combinaciones de hiperparámetros y seleccionar la mejor configuración.
- El modelo fue evaluado utilizando un conjunto de prueba separado 80% para conjunto de entrenamiento y 20% para conjunto de pruebas. Las predicciones se realizaron tanto para el conjunto de entrenamiento como para el de prueba, y se desnormalizaron para compararlas con los valores reales.

Los resultados de este modelo fueron los siguientes:

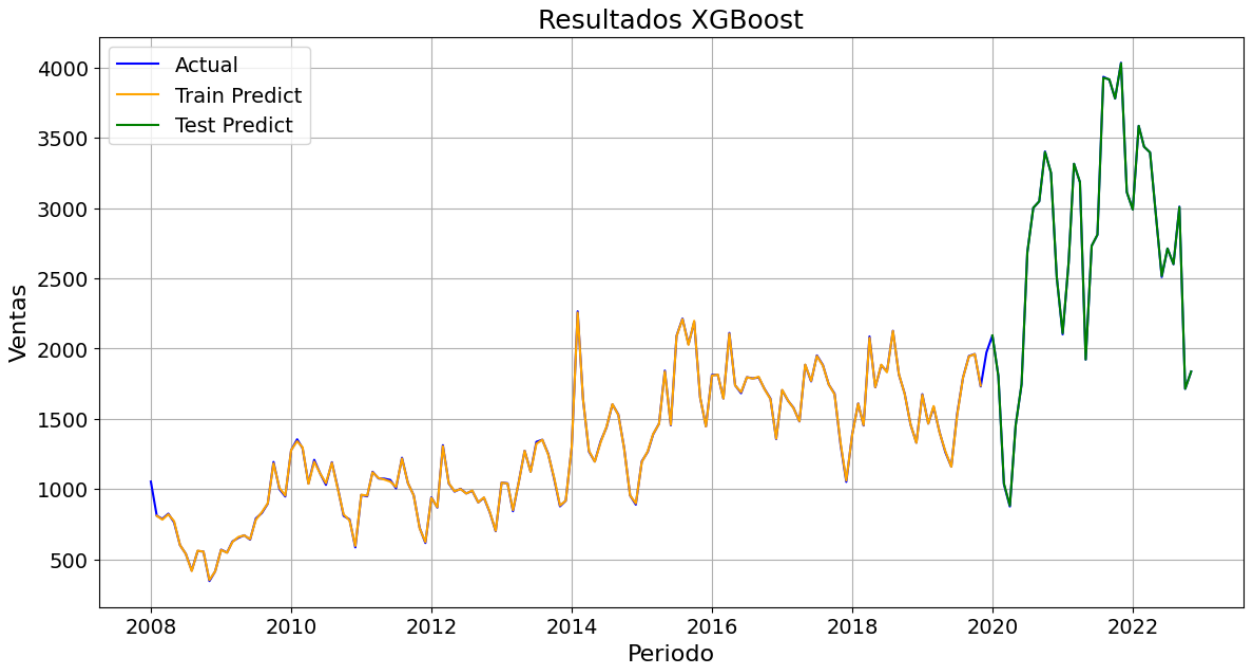


Figura 37 Resultados de las predicciones en los conjuntos de entrenamiento y validación – XGBoost

De todos los modelos, el ajuste tanto al conjunto de entrenamiento como para el conjunto de validación, son casi perfectos, resultados que pueden llevar a observarlos a mayor detalle. En todo caso, este modelo que tiene un ajuste casi perfecto a ambos conjuntos, se podrá a prueba en la predicción.

5.4. REGRESIÓN LINEAL

A diferencia de los demás modelos, el de regresión lineal, el set de datos se truncó al uso de 4 predictores buscando maximizar las métricas del modelo. Los predictores selecionados fueron los siguientes:

-GT_PROYECTOS_VIVIENDA_CALI: Consulta Trends para “PROYECTO VIVIENDA CALI”

- VAR_IPC_12M_CALI: Variación acumulada 12 meses del IPC área Cali.

- INT_CV_UVR_EA_NO_VIS: Interés para crédito de vivienda en UVR vivienda no VIS.

- ISE: Índice de Seguimiento de la Economía.

Lo anterior, ante la imposibilidad de ejecutar una búsqueda de Hiperparámetros para esta sencilla técnica. Adicional a lo anterior, se tuvo en cuenta lo siguiente:

- En cuanto a tratamiento de datos, para eliminar tendencias y hacer los datos más estacionarios, se aplicó la diferenciación a las variables. Adicionalmente, se separó el conjunto de entrenamiento y de pruebas en el mismo sentido (80/20).
- Los datos fueron normalizados usando MinMaxScaler para reescalar las características al rango [0, 1]. Esto mejora la eficiencia del entrenamiento y asegura que las diferentes escalas de las variables no afecten el modelo.

Los resultados fueron los siguientes:

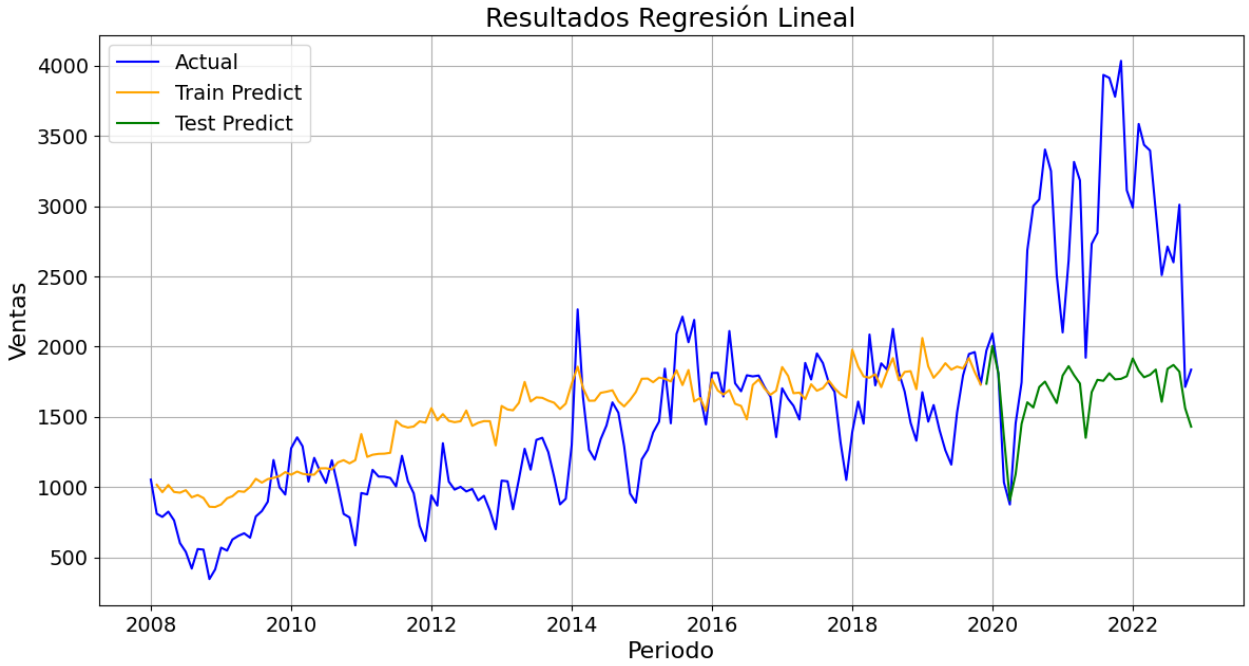


Figura 38. Resultados de las predicciones en los conjuntos de entrenamiento y validación – Regresión Lineal

El modelo de regresión lineal no tiene la misma capacidad que sus predecesores para capturar las variaciones, o por lo menos lo hace levemente. Posiblemente, existan relaciones no lineales entre los determinantes y la evolución de la variable objetivo ventas de unidades de vivienda, motivo por el cual, el modelo de regresión lineal no tenga la misma capacidad que los otros métodos.

6. EVALUACIÓN DEL MODELO DE PREDICCIÓN

A continuación, se presentan las métricas de rendimiento RMSE (Root Mean Squared Error); MAE (Mean Absolute Error) y MAPE (Mean Absolute Percentage Error), junto con los principales hiperparámetros resultantes resultantes de tanto la búsqueda aleatoria para los modelos configurables:

Tabla 5. Hiperparámetros seleccionados y métricas de rendimiento de los modelos

Modelo	Hiperparámetros	RMSE	MAE	MAPE
RNN – LSTM	units: 400 optimizer: adam LSTM layers: 2 Dense layers: 2	Train: 226.86 Test: 785.36	Train: 168.53 Test: 668.41	Train: 0.1342 Test: 0.2412
RNN – GRU	units: 450 optimizer: adam GRU layers: 2 Dense layers: 2	Train: 225.96 Test: 820.62	Train: 168.59 Test: 698.01	Train: 0.1351 Test: 0.2488
Random Forest	n_estimators: 700 min_samples_split: 5 min_samples_leaf: 4 max_features: sqrt max_depth: 100 bootstrap: false	Train: 101.39 Test: 270.62	Train: 76.31 Test: 207.71	Train: 0.0684 Test: 0.1023
XGBoost	subsample: 1.0 n_estimators: 1000 max_depth: 3 learning_rate: 0.2 colsample_bytree: 0.6	Train: 5.2292 Test: 3.3593	Train: 4.0069 Test: 2.4229	Train: 0.0037 Test: 0.0012
Regresión Lineal	N.A.	Train: 214.52 Test: 480.78	Train: 161.91 Test: 369.42	Train: 2.0525 Test: 1.0264

Los modelos estadísticos con mejor rendimiento son el XGboost y el Random Forest. De hecho, es el XGboost el único modelo que tiene mejores métricas en la predicción de los valores del conjunto de validación, el cual estructuralmente se mostraba complejo por el alto nivel de variabilidad que registró respecto a la evolución normal del conjunto de entrenamiento.

El próximo paso es intentar predecir el valor de las ventas de unidades de vivienda. Se debe considerar que el set de datos solo tiene información entre enero de 2008 y noviembre de 2022 para

la variable de venta de vivienda. Para los predictores seleccionados, existen datos entre diciembre de 2022 y enero de 2024, lo que permitiría hacer una prueba de validación efectiva prediciendo los 14 datos de ese rango de tiempo para la variable de venta de viviendas.

7. MÉTODO PROPUESTO

A partir de los modelos desarrollados, se efectuó las predicciones para la variable "VENTAS" que se refiere a la venta de unidades de vivienda en el periodo comprendido entre diciembre de 2022 y enero de 2024.

Los resultados fueron los siguientes:

Tabla 6. Resultados de la predicción de las ventas de nuevas unidades de vivienda de cada modelo (14 periodos)

Periodo	LSTM	GRU	XGBoost	Random Forest	Linear Regression
dic-22	2,109	2,067	1,097	1,586	2,747
ene-23	3,292	3,168	1,436	2,262	3,098
feb-23	2,241	2,287	1,200	1,774	1,900
mar-23	1,866	1,900	1,321	1,567	1,989
abr-23	1,695	1,698	1,178	1,569	2,048
may-23	2,033	2,035	1,105	1,618	2,157
jun-23	1,254	1,220	1,212	1,586	2,266
jul-23	1,236	1,207	1,012	1,639	2,136
ago-23	1,431	1,404	908	1,382	2,155
sep-23	1,584	1,581	831	1,376	2,341
oct-23	1,068	1,030	1,258	1,576	2,118
nov-23	803	725	835	1,354	2,288
dic-23	474	330	932	1,270	2,327
ene-24	1,616	1,476	967	1,513	2,413
Total	**22,702**	**22,130**	**15,293**	**22,072**	**31,983**

Como dato interesante, se obtiene que, en el rango comprendido entre diciembre de 2022 y enero de 2024, las redes neuronales y el modelo de Random Forest predicen una venta de alrededor de 22 mil unidades de nuevas viviendas en este periodo. Por su parte, el modelo con mejores métricas de rendimiento, el XGboost, es a su vez el más cauto o más pesimista al momento de predecir, con un total de 15 mil unidades de vivienda, mientras que la técnica de regresión lineal es la más optimista con una venta propuesta de casi 32 mil unidades de vivienda.

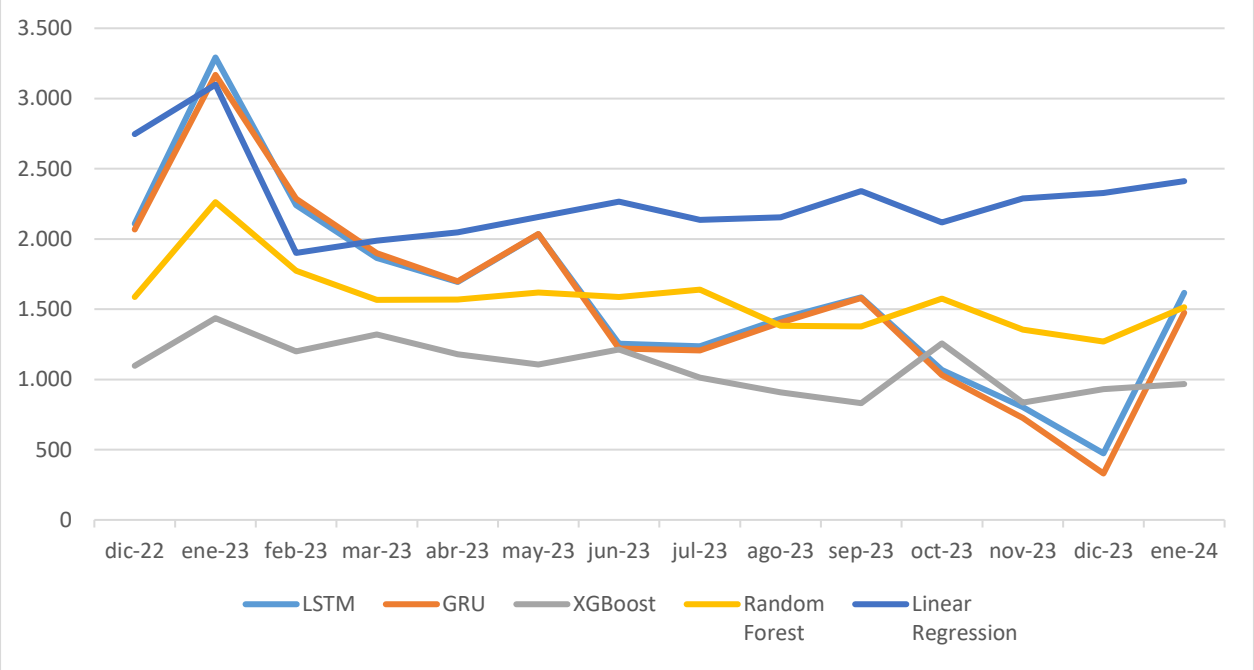


Figura 39. Gráfico de la predicción de las ventas de unidades de vivienda de cada modelo (14 periodos)

Para tener una mejor perspectiva sobre la dimensión de la predicción fuera del set de datos inicial para la variable ventas, se grafican los resultados observados de la variable ventas respecto a la predicción que se realizó con cada una de las técnicas:

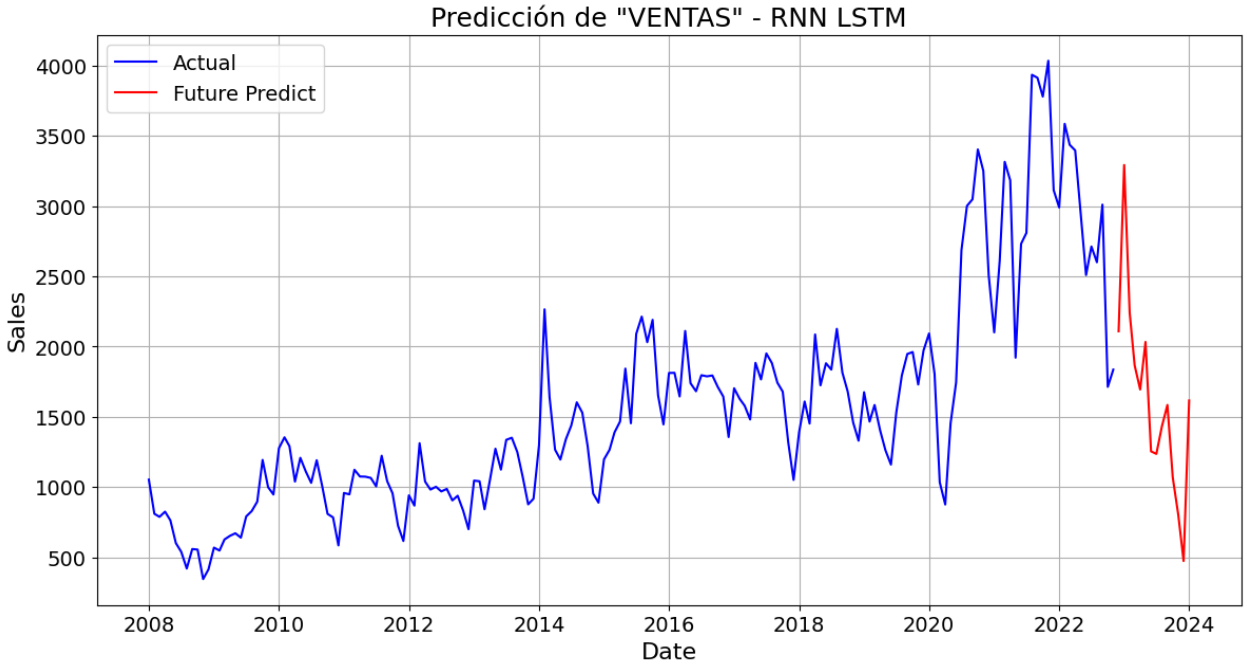


Figura 40. Valores observados y predicción de venta de unidades de vivienda. Modelo LSTM

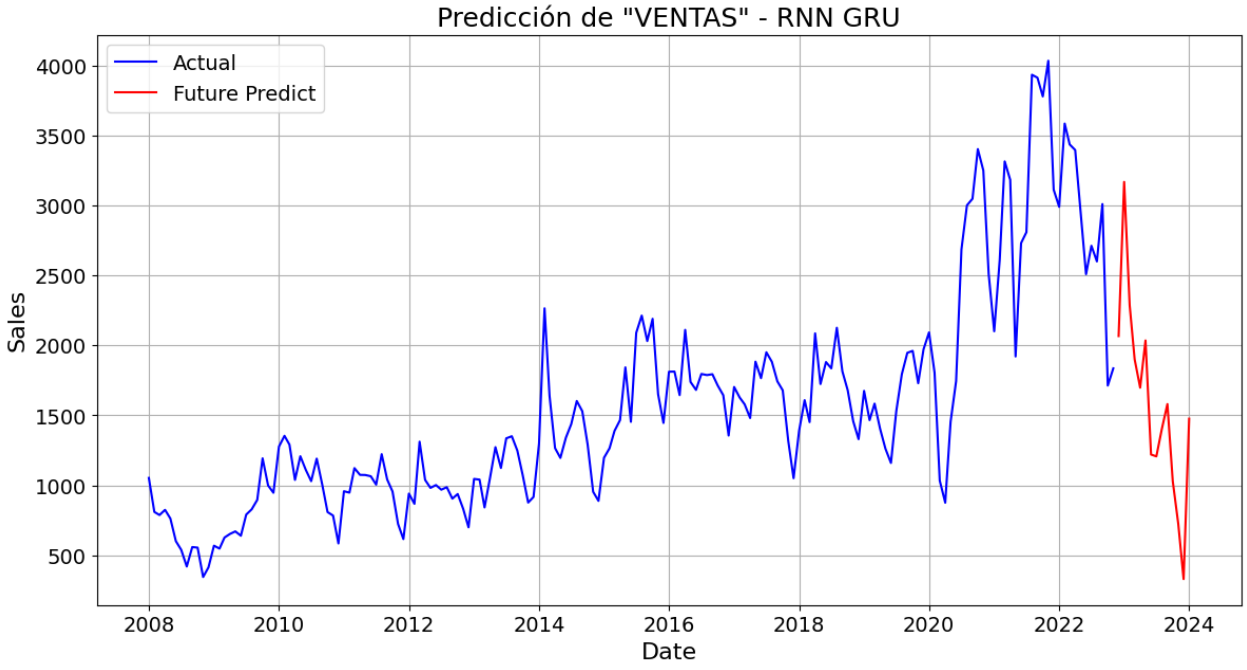


Figura 41. Valores observados y predicción de venta de unidades de vivienda. Modelo GRU

En el caso de los modelos de redes neuronales recurrentes, las predicciones tienen una orientación similar, con picos altos y una clara tendencia decreciente hasta llegar a un mínimo histórico.

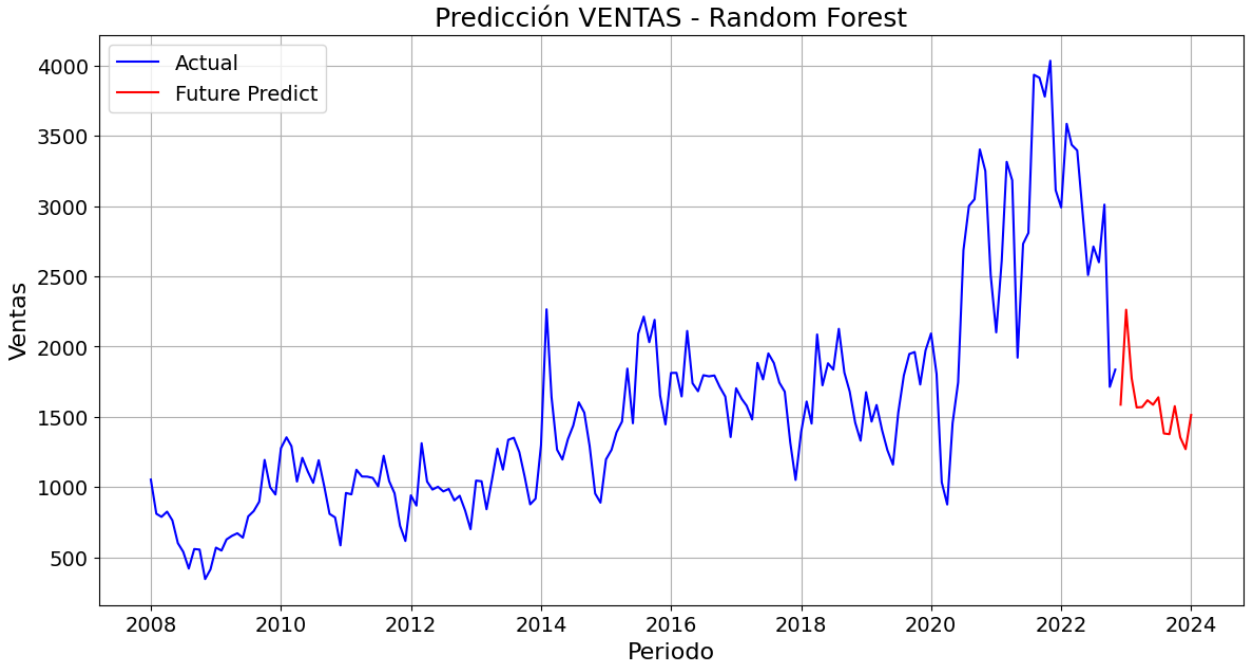


Figura 42. Valores observados y predicción de venta de unidades de vivienda. Modelo Random Forest

El modelo de Random Forest predice una cantidad similar de viviendas que los modelos de redes neuronales, con un pico inicial y una tendencia marcadamente decreciente. No obstante muestra el mínimo para el periodo de diciembre de 2023 con 1270 unidades. Por su parte los modelos de redes neuronales predicen los mínimos históricos de las series para este periodo, con 474, el modelo LSTM y 330 el modelo con capas GRU.

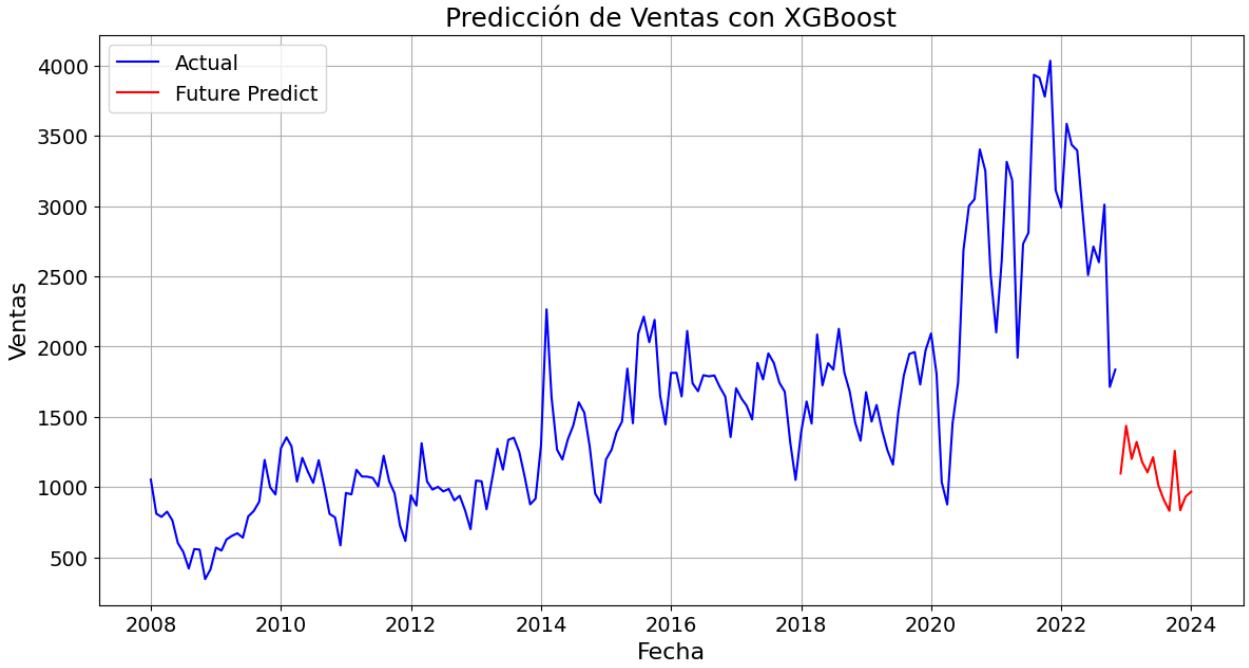


Figura 43. Valores observados y predicción de venta de unidades de vivienda. Modelo XGBoost

Ahora, al momento de evaluar el modelo de mejor rendimiento en el entrenamiento, el XGboost, la predicción fuera del rango de tiempo muestra una tendencia decreciente, aunque estructuralmente diferente a las observaciones históricas. Este modelo predice un mínimo de 831 unidades vendidas en septiembre de 2023.

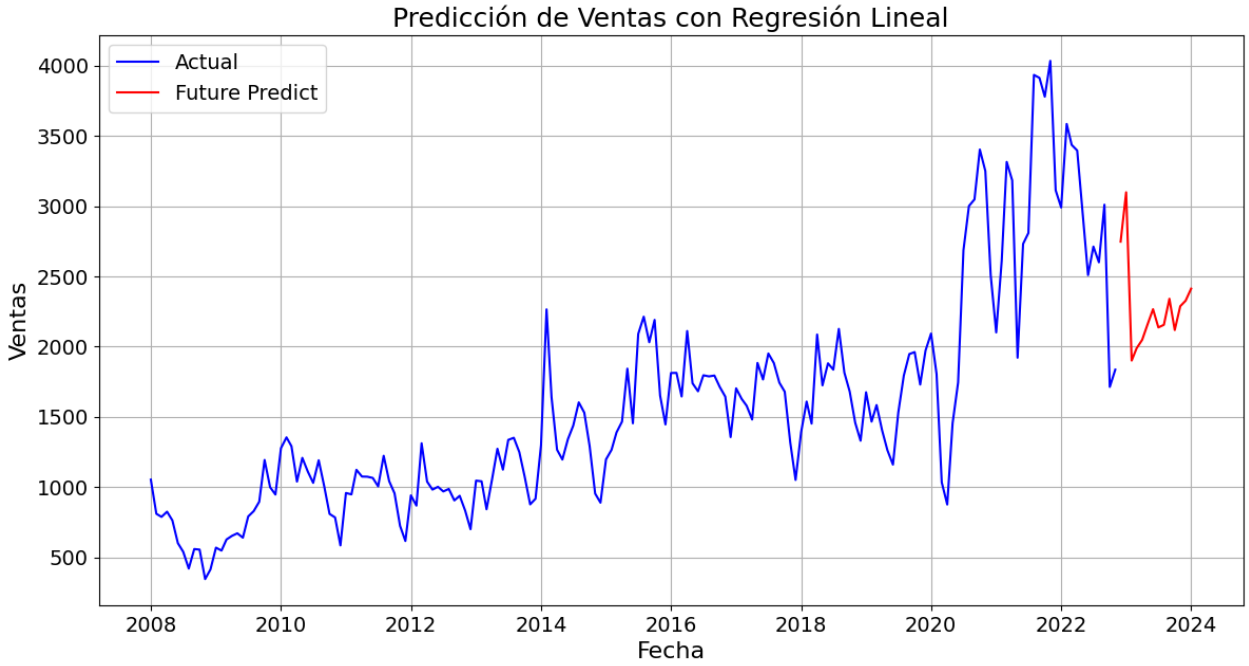


Figura 44. Valores observados y predicción de venta de unidades de vivienda. Modelo de Regresión Lineal

Finalmente, la predicción elaborada con la regresión lineal es la única que cae en picado los dos primeros meses, pero posteriormente tiene una tendencia creciente. Esta predicción marca en los 14 meses una venta total de casi 32 mil unidades de vivienda. No obstante, metodológicamente no es comparable con los demás dado que se basa en un conjunto de sólo cuatro predictores.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES

El presente proyecto demuestra que es viable utilizar como determinantes los niveles de popularidad relativa de consultas de Google Trends e indicadores macroeconómicos para predecir las ventas de unidades de vivienda en el Distrito de Cali.

Se ha logrado integrar exitosamente múltiples fuentes de datos, incluyendo consultas de Google Trends y variables macroeconómicas como inflación, desempleo, tasas de interés e indicadores de percepción del consumidor. Esta integración permitió capturar una visión más completa y actualizada del mercado inmobiliario.

Además, se implementaron y evaluaron diversos modelos predictivos, incluyendo Random Forest, XGBoost, y redes neuronales recurrentes (LSTM y GRU). Cada uno de estos modelos fue optimizado y validado utilizando la búsqueda de hiperparámetros y la validación cruzada temporal. También se incluyó un modelo de regresión lineal para efectos comparativos.

Los modelos desarrollados demostraron capacidad para predecir las ventas de unidades nuevas de vivienda. Las métricas de evaluación, como el RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) y MAPE (Mean Absolute Percentage Error), indicaron que los modelos lograron una precisión adecuada en sus predicciones, especialmente las técnicas de Random Forest y XGboost. Adicional a lo anterior, se observó que el modelo XGBoost, en particular, ofreció un equilibrio óptimo entre precisión y eficiencia computacional, destacándose como una herramienta que puede usarse como base en el modelo predictivo por defecto para la predicción de ventas en el

sector inmobiliario.

El análisis de tendencias en Google Trends demostró ser una herramienta valiosa para predecir el comportamiento del mercado de viviendas. Las consultas relacionadas con la compra de viviendas, como "APARTAMENTOS CALI VENTA" y "CASAS CALI VENTA", mostraron correlaciones significativas con las ventas de unidades nuevas. La inclusión de datos de Google Trends en los modelos predictivos mejoró la precisión de las predicciones, respaldando la viabilidad de su uso como indicador adelantado de la demanda en el mercado inmobiliario.

El uso efectivo de datos secundarios representó una innovación significativa en el análisis y predicción del mercado inmobiliario. La integración de estas fuentes de información enriquece los modelos predictivos, proporcionando herramientas más robustas y versátiles para la toma de decisiones estratégicas y el éxito en el sector. La introducción de un nuevo enfoque para el análisis del mercado no solo beneficia al sector inmobiliario de Cali, sino que también abrió un camino para el desarrollo de herramientas de análisis más precisas y replicables en diferentes contextos. Este avance contribuye a una mejor comprensión del mercado, una toma de decisiones más informada y un desarrollo inmobiliario más sostenible a nivel global.

8.2. TRABAJOS FUTUROS

- En vista de la limitación de datos, se podrían incorporar rezagos a las predictoras para evaluar los resultados. En el mismo sentido, vale la pena explorar los resultados con los datos completos de los Índices de Confianza al Consumidor – ICC y el Índice de Disposición a Comprar Vivienda – IDCV de la Encuesta de Opinión del Consumidor de Fedesarrollo.
- La variable de "VENTAS" corresponde al total de unidades nuevas de vivienda vendida de todo tipo, VIS, no VIS. Se podría considerar desagregar este indicador para modelar la predicción por segmentos.
- Replicar este modelo predictivo realizado a otras ciudades y regiones, ajustando sus variables y parámetros a las particularidades de cada mercado. Esta iniciativa permitirá validar su robustez y aplicabilidad en diversos contextos, ampliando su utilidad para la toma de decisiones a nivel nacional o regional.
- Para fortalecer el modelo predictivo, se podría enriquecer evaluando nuevas fuentes de datos secundarios. Esto, para potenciar su precisión incorporando información proveniente de redes sociales, plataformas inmobiliarias y otras fuentes relevantes. Con esta iniciativa se tienen el potencial de ampliar significativamente la utilidad del modelo para la toma de decisiones estratégicas a nivel nacional o regional.
- Se propone desarrollar un sistema de predicciones en tiempo real que permita actualizaciones constantes y responda ágilmente a los cambios del mercado. Con esta iniciativa, se obtienen beneficios en cuanto a adaptabilidad y toma de decisiones oportunas la convierten en una línea de trabajo atractiva.

8. REFERENCIAS

Banco de la República. (2023). *Análisis de la cartera y del mercado inmobiliario en Colombia.* Obtenido de https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/10755/Informe_Especial_vivienda_2023-II.pdf?sequence=5&isAllowed=y

Blanco, E. (2014). *Herramientas de Big Data: øPodemos aprovechar Google Trends para pronosticar algunas variables macro relevantes?* Obtenido de https://d1wqtxts1xzle7.cloudfront.net/47729849/Herramientas_big_data-libre.pdf?1470156512=&response-content-disposition=inline%3B+filename%3DHerramientas_de_Big_Data_Podemos_aprovec.pdf&Expires=1714536692&Signature=Oc8-mFfltlfgj9amq7xExJfij~3BoVe6WOlo9Rdt

Breiman, L. (08 de Noviembre de 2022). Random Forests. *Machine Learning*, 5-32. doi:10.1023/A:1010933404324

Bulczak, G. M. (2021). " Uso de Google Trends para predecir el mercado inmobiliario: evidencia del Reino Unido ". *International Real Estate Review, Global Social Science Institute*, págs. vol. 24(4), páginas 613-631.

Caicedo, S., Morales-Mosquera, M., & Pérez-Reyna, D. (2010). «Un análisis de sobrevaloración en el mercado de la vivienda en Colombia». *Banco de la República de Colombia*.

CAMACOL VALLE. (2022). *https://camacolvalle.org.co/.* Obtenido de https://camacolvalle.org.co/wp-content/uploads/2024/02/Estudio-de-Oferta-y-Demanda-2022.pdf

Cámara Colombiana de la Construcción, CAMACOL. (2022). *"Tendencias de la construcción, economía y coyuntura sectorial.* Obtenido de Cámara Colombiana de la Construcción, CAMACOL: https://camacol.co/sites/default/files/descargables/TENDENCIAS%2025%20DICIEMBRE%2013%20DE%202022-PARA%20WEB.pdf, 2022.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an Emerging Market. *Journal of Forecasting*, 289-298.

Chatfield, C. (2003). *The Analysis of Time Series, An Introduction.* Nueva York: Chapman and Hall/CRC. doi:https://doi.org/10.4324/9780203491683

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco: ACM.

Cho, K., Merrienboer, B. v., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *CoRR*, 1409.1259.

Cuecha, D., Hernández, M., & Rodríguez, D. (2022). *Modelo de direccionamiento estratégico para la empresa inmobiliaria Colombiana de Finca Raíz.* Obtenido de Universidad EAN: https://repository.universidadean.edu.co/bitstream/handle/10882/12373/RodriguezDavid2022.pdf?sequence=1

Departamento Administrativo Nacional de Estadística - DANE. (2023). «Financiación de vivienda (FIVI)». Bogotá.

Departamento Administrativo Nacional de Estadística (DANE). (2016). *Metodología General Indicador de Seguimiento a la Economía (ISE).* Bogotá.

Fedesarrollo - Centro de Investigación Económica y Social. (2017). *Metodología de la Encuesta*

de Opinión del Consumidor. Bogotá.

Garay-Rodríguez, S., Vidal, A. P., & Cerón-Ordóñez, J. (07 de Noviembre de 2023). *Scielo.* Obtenido de El monitoreo del sector de la construcción en el Valle del Cauca: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-30532023000100237

Gonzales-Arrieta, G. (2005). «El crédito hipotecario y el acceso a la vivienda para los hogares de menores ingresos en América Latina». REVISTA DE LA CEPAL 85.

Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 111-120. doi:https://doi.org/10.1016/0304-4076(74)90034-7

Guerrero, S. (2016). *Análisis espacial de los cambios y determinantes de la oferta de vivienda nueva no VIS en Cali para el periodo 2009 - 2015.* Medellín: Universidad Eafit.

Himmelberg, C., Mayer, C., & Sinai, T. (2005). Assessing High House Prices: Bubbles, Fundamentals and Misperceptions. *Journal of Economic Perspectives*, 67-92. doi:10.1257/089533005775196769

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, 1735-1780.

Okun, A. (1962). Potential GNP: Its Measurement and Significance. *Proceedings of the Business and Economics Section. American Statistical Association.*, 98-103.

Oswald, A. J. (1996). A Conjecture on the Explanation for High Unemployment in the Industrialized Nations: Part I. *Warwick Economic Research Papers*.

Paciorek, A. (2013). «Supply constraints and housing market dynamics,» . *Journal of Urban Economics*, vol. 77, pp. 11-26.

Rave, J. I. (marzo de 2019). *Scielo.* Obtenido de Statihouse ® : desarrollo tecnológico basado en ciencia de datos para explorar estadísticamente el sector inmobiliario:

https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052019000100113

Rodríguez Niño, C. (2022, 2023). *Escuela Técnica Superior de Ingeniería de Edificación.* Obtenido de Universitat Politécnica de Valéncia: https://riunet.upv.es/bitstream/handle/10251/190779/Rodriguez%20-%20Influencia%20de%20las%20variables%20macroeconomicas%20en%20el%20mercado%20inmobiliario%20de%20Colombia%20Im....pdf?sequence=1&isAllowed=y

Rodríguez, C., & Bustillo, R. (2010). Modelling Foreign Real Estate Investment: The Spanish Case. *The Journal of Real Estate Finance and Economics*, 354-367.

Rosso-Mateus, A. E., Montilla-Montilla, Y. M., & Garzon-Martínez, S. C. (2022). "Metodología para obtención y análisis de datos inmobiliarios usando fuentes alternativas: estudio de caso en tres ciudades intermedias de Colombia". vol. 27, no. 3.

Schwab, R. M. (1982). «Inflation Expectations and the Demand for Housing». En *The American Economic Review* (págs. vol. 72, nº 1, pp. 143-153).

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference* (págs. 92-96). Austin: Stefan van der Walt and Jarrod Millman .

Timón, C. E., & Fontes, X. M. (16 de enero de 2017). *Trabajo de Fin de Grado "Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso".* Obtenido de https://openaccess.uoc.edu/bitstream/10609/59565/6/caresptimTFG0117mem%C3%B2ria.pdf

Vargas, R. G. (2023). *CARACTERIZACIÓN DE ESTRATEGIAS DE MARKETING DIGITAL, EN EL SECTOR INMOBILIARIO DE CALI.* Obtenido de

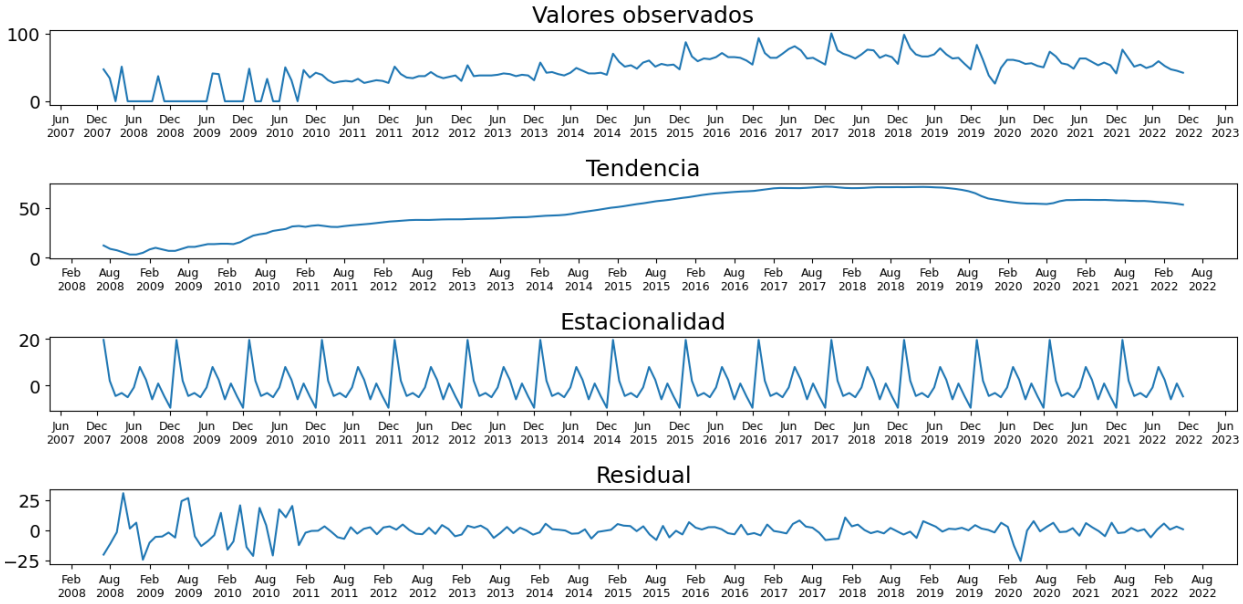
https://red.uao.edu.co/server/api/core/bitstreams/7000f701-1a3c-4b56-9f32-bda50c88d4b8/content

Wei, W. (2006). *Time Series Analysis: Univariate and Multivariate Methods.* Pearson.

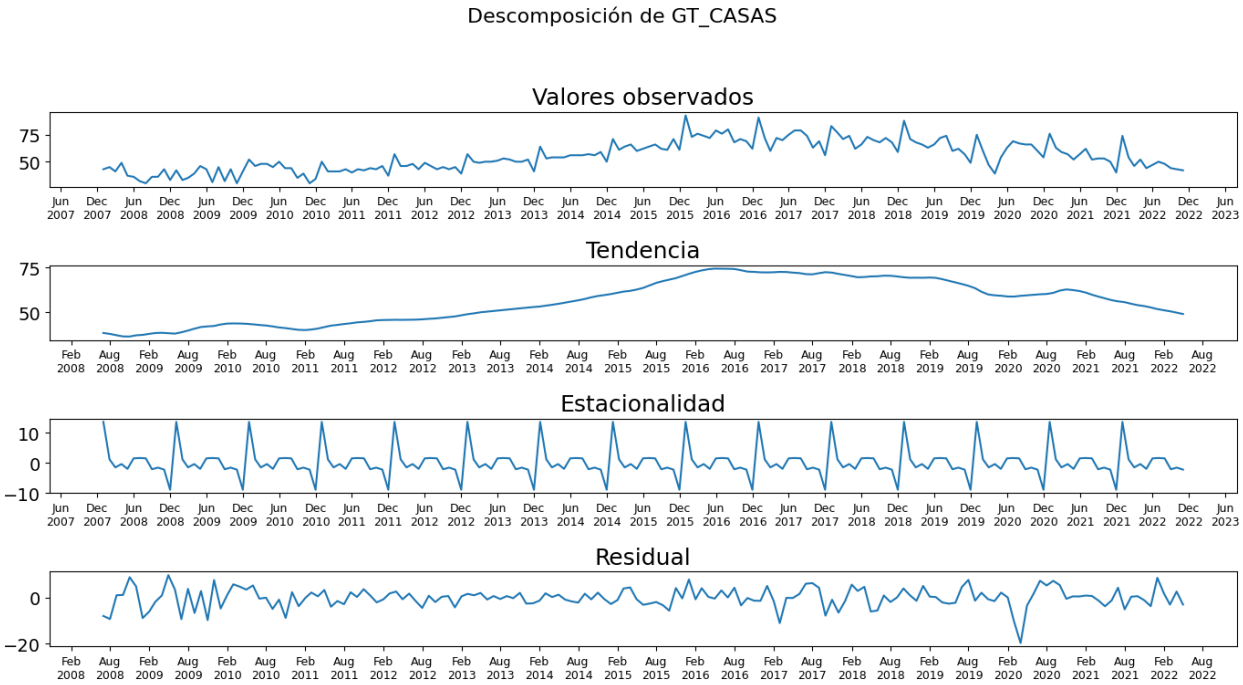
Wijnhoven, F., & Plant, O. (2017). Sentiment Analysis and Google Trends Data for Predicting Car Sales.

9. ANEXOS

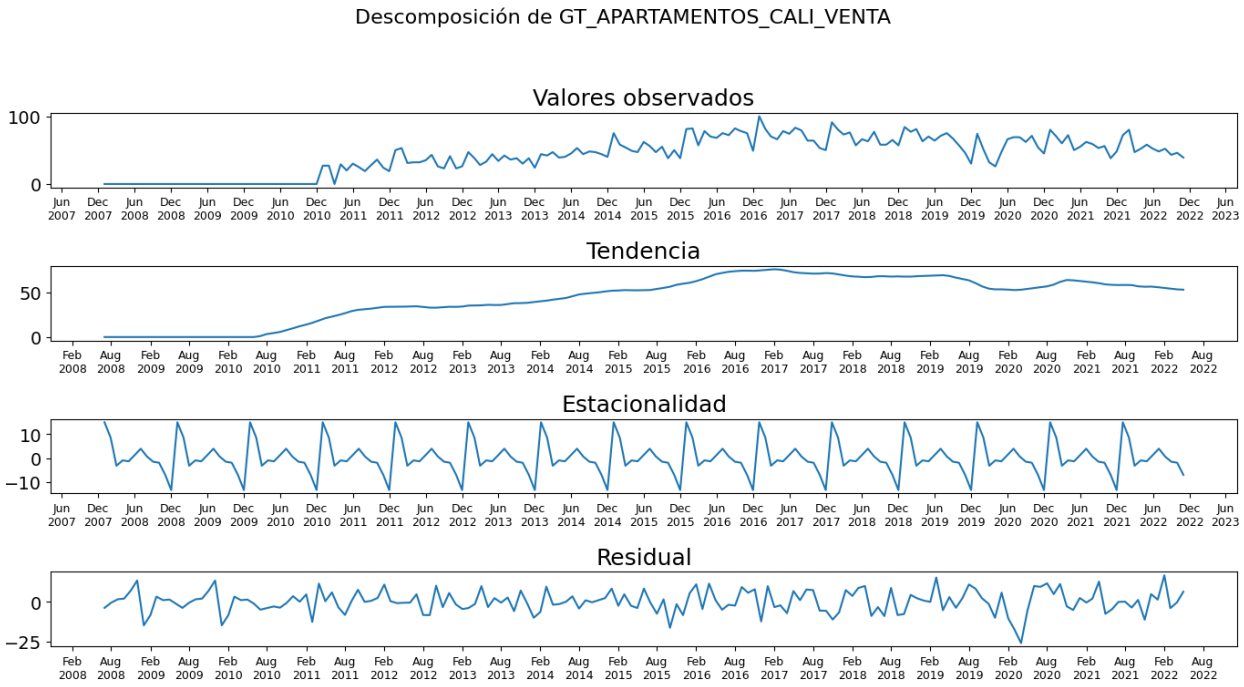
Anexo 1. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "APARTAMENTOS"



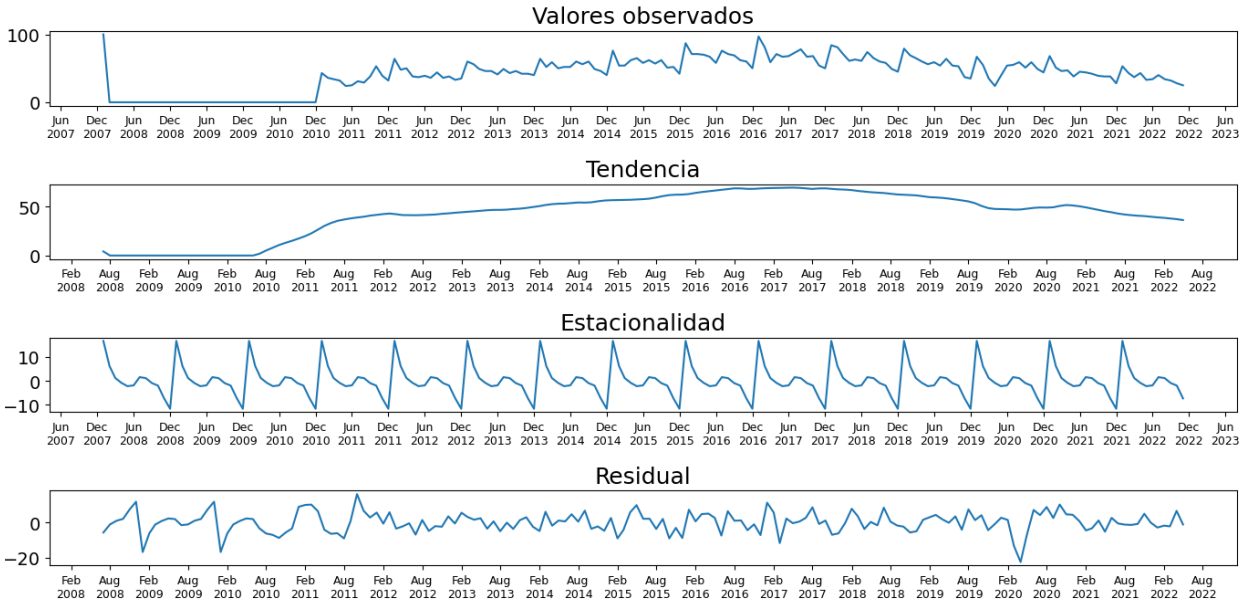
Anexo 2. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "CASAS"



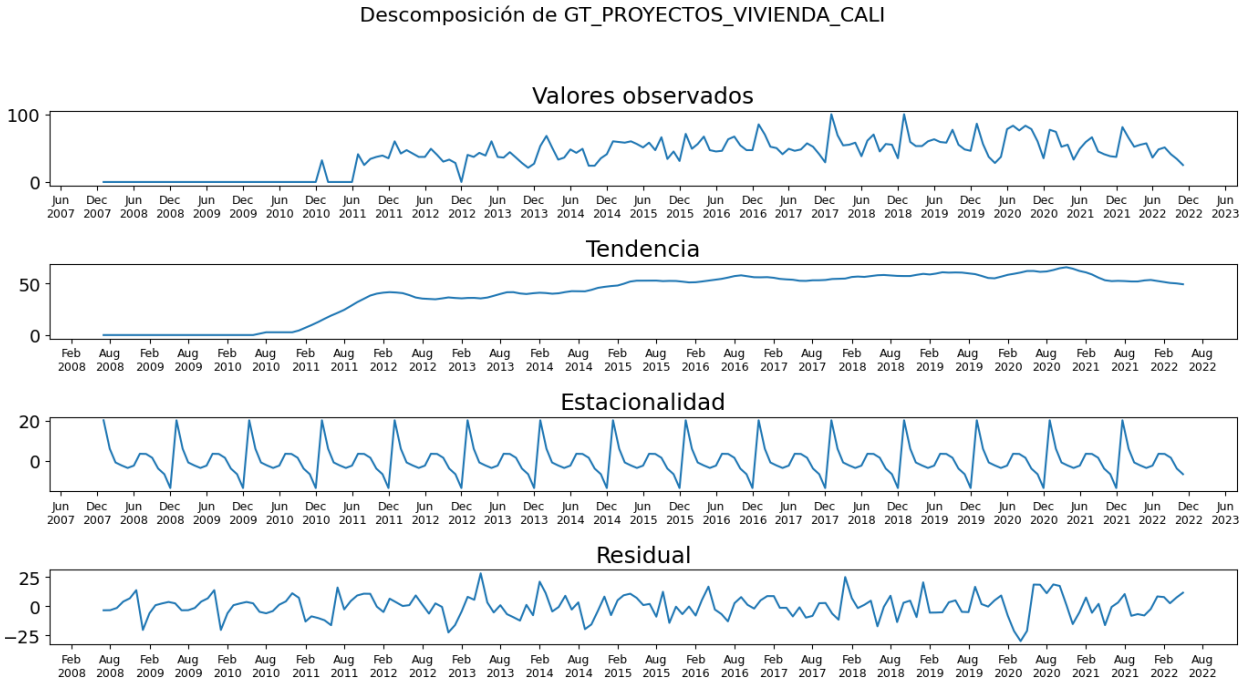
Anexo 3. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "APARTAMENTOS CALI VENTA"



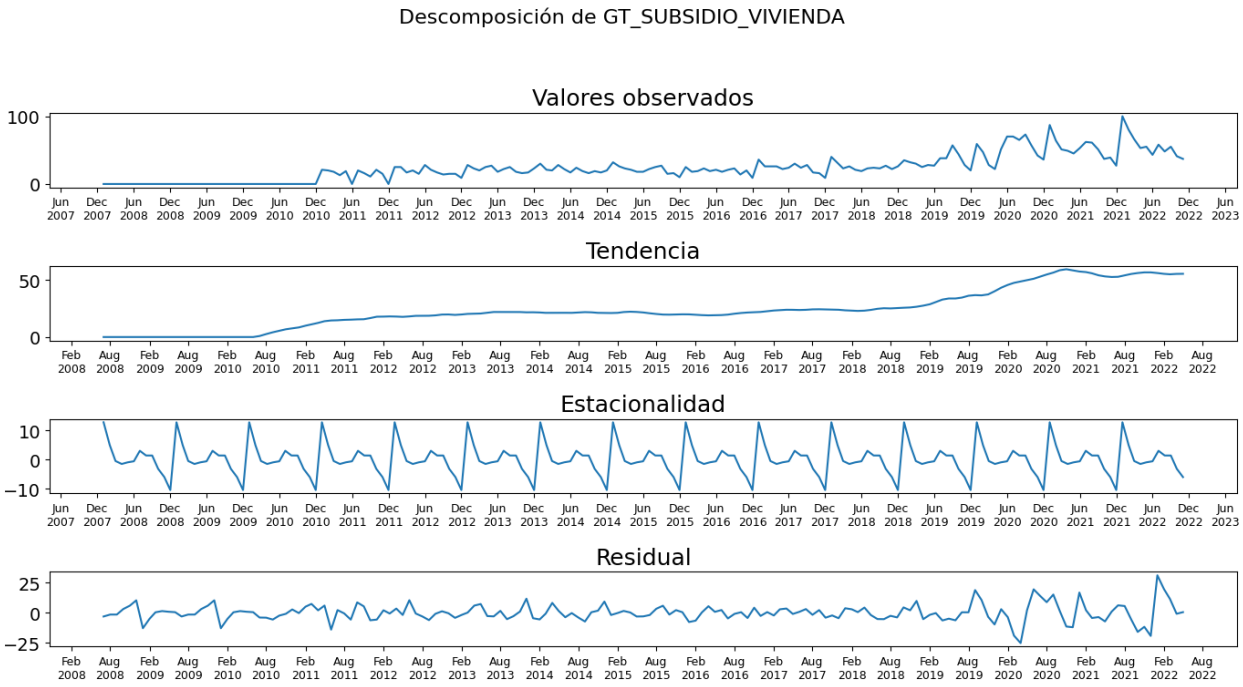
Anexo 4. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "CASAS CALI VENTA"



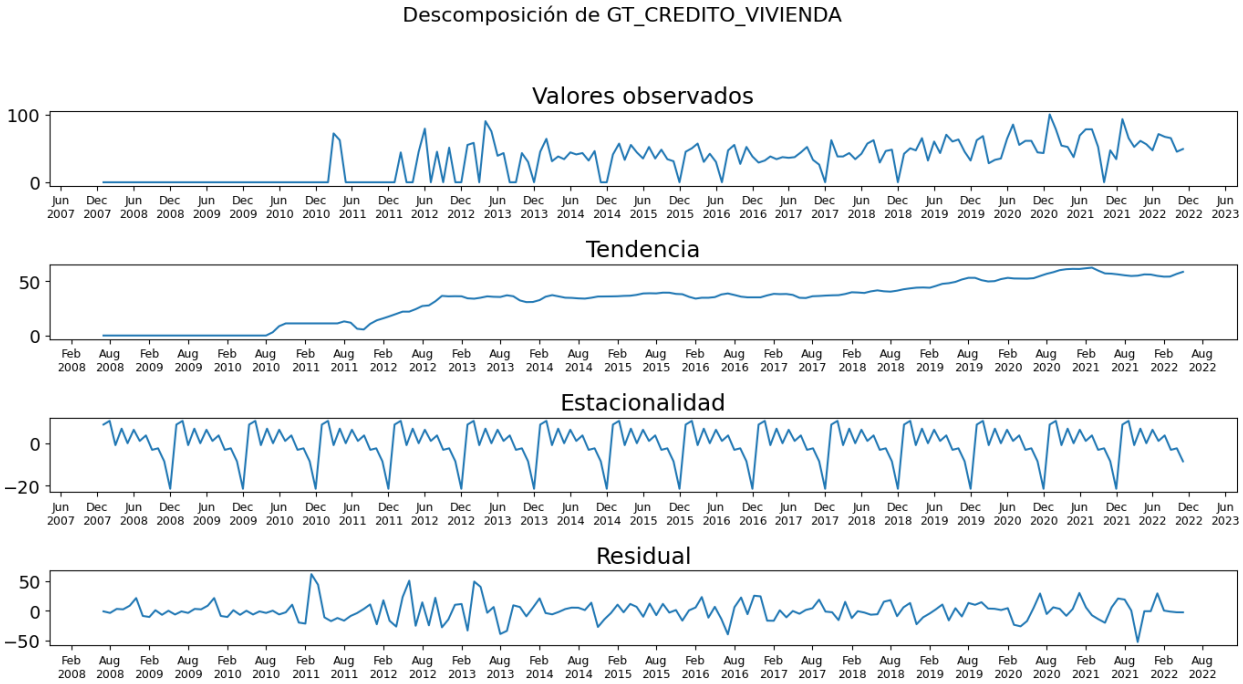
Anexo 5. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "PROYECTOS VIVIENDA CALI"



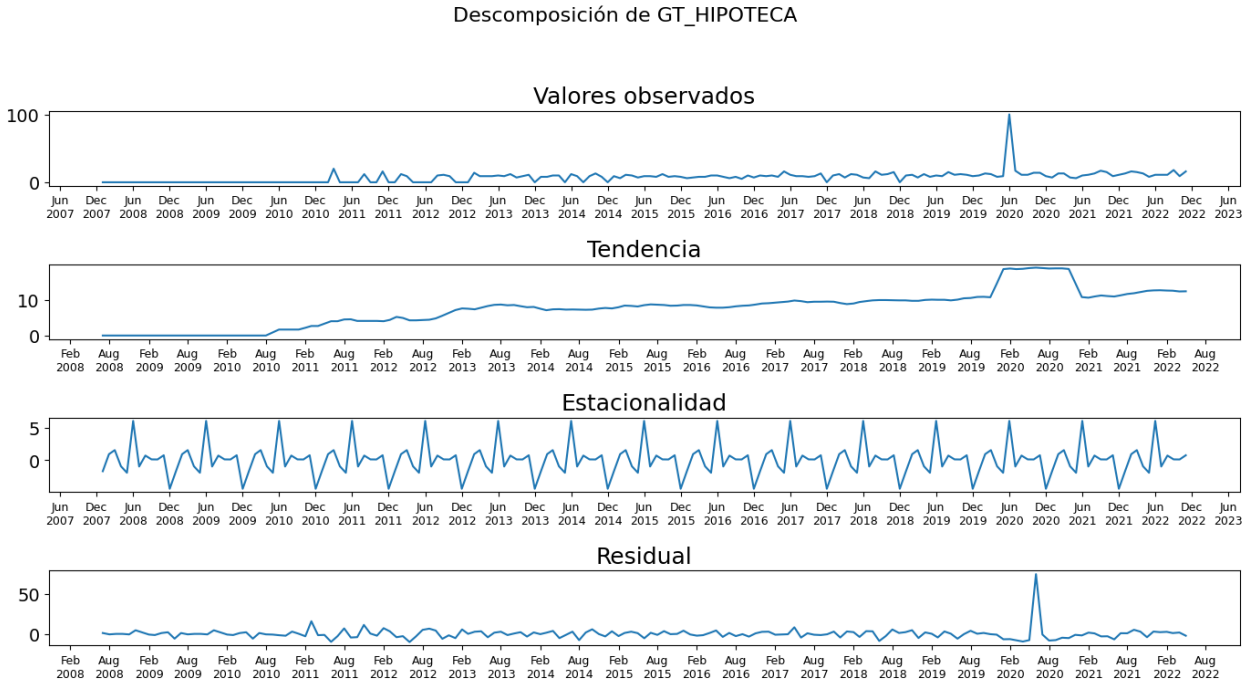
Anexo 6. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "SUBSIDIO VIVIENDA"



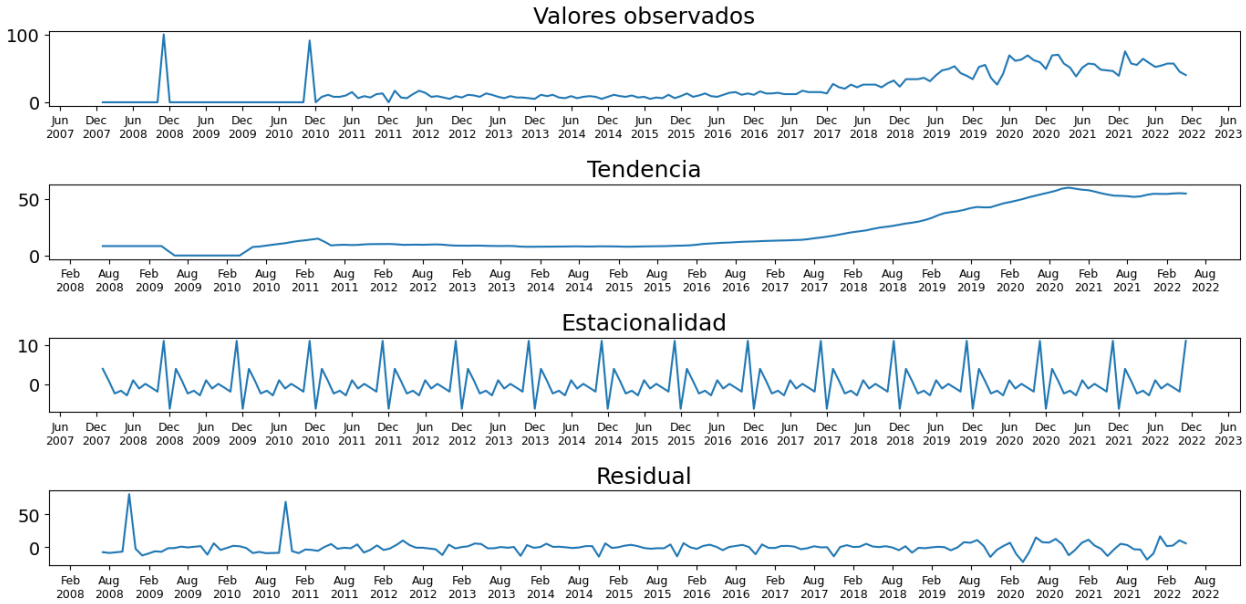
Anexo 7. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "CREDITO VIVIENDA"



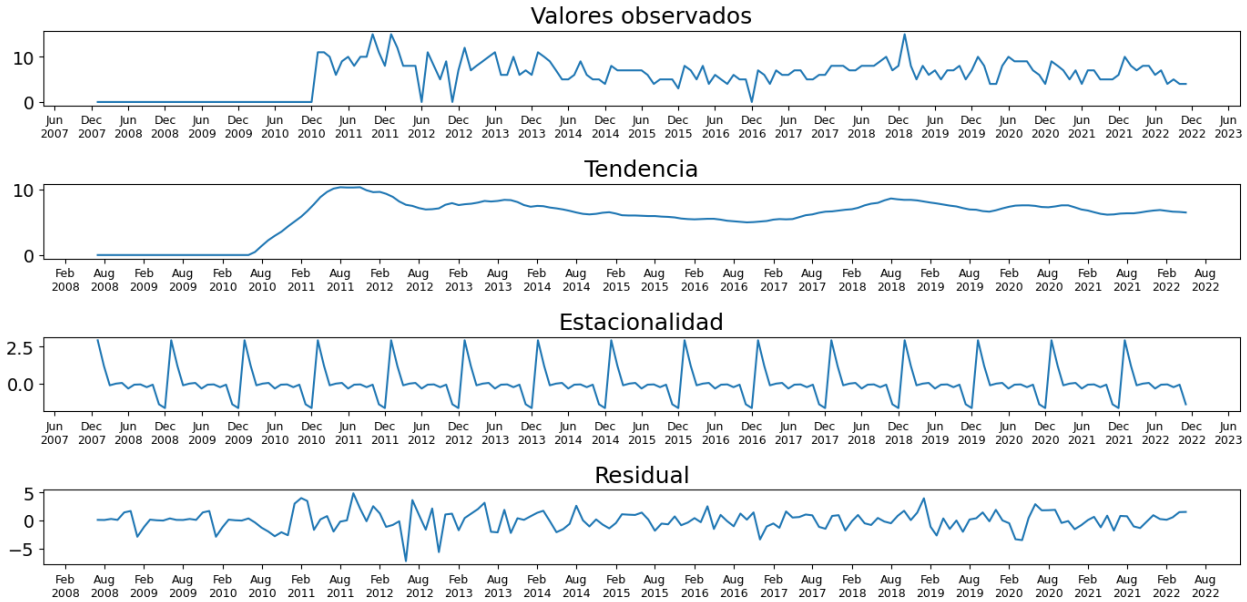
Anexo 8. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "HIPOTECA"



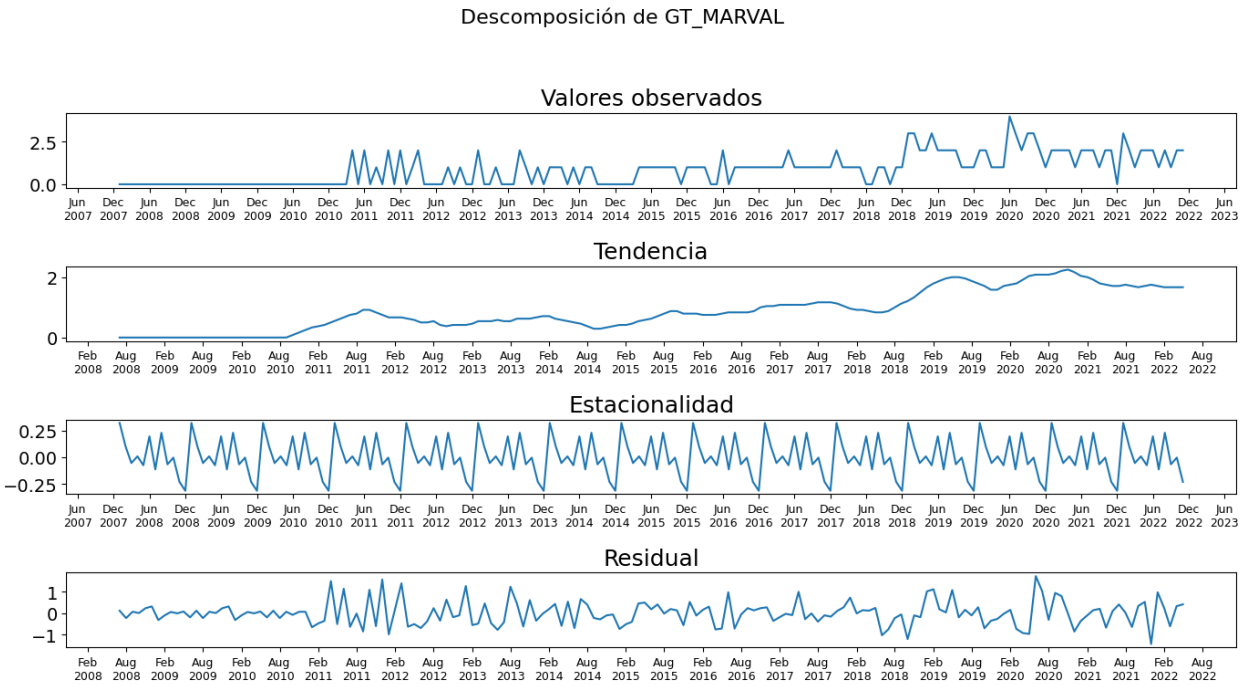
Anexo 9. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "CONSTRUCTORA BOLIVAR".



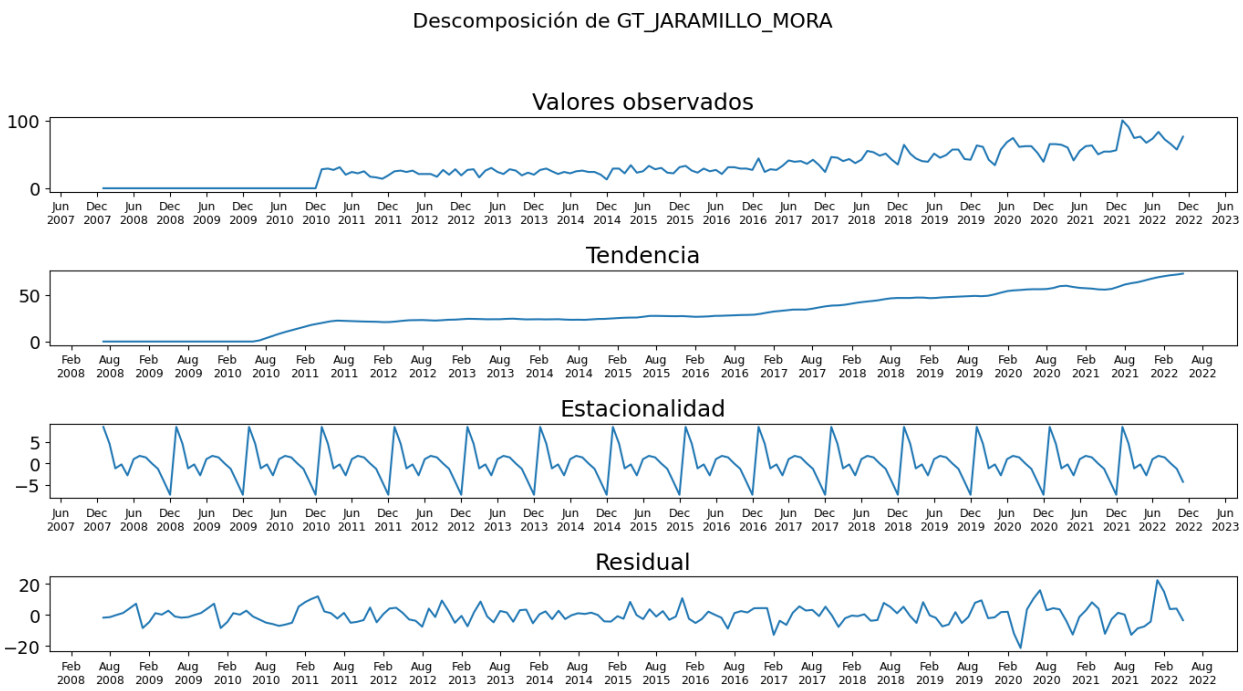
Anexo 10. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "CONSTRUCTORA MELENDEZ"



Anexo 11. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "MARVAL"



Anexo 12. Descomposición estacional y de tendencia de la popularidad relativa de la consulta "JARAMILLO MORA"



Anexo 13. Resumen de la arquitectura del modelo RNN – LSTM

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 1, 400)            660800

 lstm_1 (LSTM)               (None, 400)               1281600

 dense (Dense)               (None, 100)               40100

 dense_1 (Dense)             (None, 1)                 101

=================================================================
Total params: 1982601 (7.56 MB)
Trainable params: 1982601 (7.56 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Anexo 14. Resumen de la arquitectura del modelo RNN – GRU

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 gru (GRU)                   (None, 1, 450)            626400

 gru_1 (GRU)                 (None, 450)               1217700

 dense (Dense)               (None, 100)               45100

 dense_1 (Dense)             (None, 1)                 101

=================================================================
Total params: 1889301 (7.21 MB)
Trainable params: 1889301 (7.21 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Anexo 15. Resumen del modelo de Regresión lineal

OLS Regression Results

Dep. Variable:	VENTAS	R-squared:	0.252
Model:	OLS	Adj. R-squared:	0.235
Method:	Least Squares	F-statistic:	14.60
Date:	Fri, 28 Jun 2024	Prob (F-statistic):	2.70e-10
Time:	03:36:15	Log-Likelihood:	-1255.6
No. Observations:	178	AIC:	2521.
Df Residuals:	173	BIC:	2537.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.0924	21.555	-0.422	0.674	-51.637	33.452
GT_PROYECTOS_VIVIENDA_CALI	4.9720	1.308	3.801	0.000	2.390	7.554
VAR_IPC_12M_CALI	-5734.0685	3082.488	-1.860	0.065	-1.18e+04	350.059
INT_CV_UVR_EA_NO_VIS	-215.6051	66.357	-3.249	0.001	-346.578	-84.632
ISE	48.5332	10.930	4.440	0.000	26.960	70.106

Omnibus:	18.385	Durbin-Watson:	2.306
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.236
Skew:	0.449	Prob(JB):	3.02e-09
Kurtosis:	5.117	Cond. No.	2.37e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.37e+03. This might indicate that there are
strong multicollinearity or other numerical problems.