Prediciendo El Inidice De Precios De Vivienda Nueva Con Google Trends

Santiago Collante Villabona
Mateo Rios De Los Rios

Universidad Icesi
Facultad Ciencias Administrativas y Económicas
Cristian Camilo Hoyos Bermeo

Santiago de Cali
12 de noviembre de 2023

Prediciendo El Inidice De Precios De Vivienda Nueva Con Google Trends

Autores
Santiago Collante Villabona
Mateo Rios De Los Rios

Director del proyecto
Cristian Camilo Hoyos Bermeo

Facultad de ciencias administrativas y económicas
Administración de Empresas con énfasis en Negocios Internacionales y Finanzas
Economía y Negocios Internacionales

Santiago de Cali
2023

Tabla de Contenido

pág.

Lista de Tablas

pág.

Resumen

Este documento se centra en el desarrollo y validación de un modelo de nowcasting para pronosticar con precisión el Índice de Precios de Vivienda Nueva (IPVN) en Colombia hasta el año 2023. Utilizando algoritmos de machine learning en RStudio, el estudio persigue tres objetivos clave. En primer lugar, realiza un análisis exhaustivo de diversas fuentes de información para la construcción del modelo. Luego, diseña un modelo que transforma datos extraídos de búsquedas en motores de información relacionados con el mercado inmobiliario. Finalmente, valida rigurosamente la calidad y precisión de las estimaciones obtenidas. La propuesta, respaldada por datos recopilados hasta septiembre de 2023, incorpora tasas de intervención, tasas de desempleo y datos de búsquedas en línea.

Los resultados destacan la Regresión Lasso como el modelo más preciso para predecir el IPVN, superando alternativas como ARIMA. Este hallazgo resalta la importancia de los modelos de nowcasting basados en datos de búsqueda en Google Trends para decisiones informadas en el sector inmobiliario colombiano y sugiere la aplicabilidad de la Regresión Lasso en otros contextos económicos, proporcionando una base sólida para futuras aplicaciones y expansiones en Colombia y contextos similares.

1.1 Palabras Claves

Nowcasting, Índice de Precios de Vivienda Nueva, Economía colombiana, Google Trends

Abstract

This paper focuses on the development and validation of a nowcasting model to accurately forecast the New Housing Price Index (IPVN) in Colombia until 2023. Using machine learning algorithms in RStudio, the study pursues three key objectives. First, it performs an exhaustive analysis of various sources of information for the construction of the model. Then, it designs a model that transforms data extracted from searches in information engines related to the real estate market. Finally, it rigorously validates the quality and accuracy of the estimates obtained. The proposal, supported by data collected through September 2023, incorporates intervention rates, unemployment rates, and online search data.

The results highlight Lasso Regression as the most accurate model for predicting the IPVN, outperforming alternatives such as ARIMA. This finding highlights the importance of nowcasting models based on Google Trends search data for informed decisions in the Colombian real estate sector and suggests the applicability of Lasso Regression in other economic contexts, providing a solid foundation for future applications and expansions in Colombia and similar contexts.

1.2 Key Words

Nowcasting, New Home Price Index, Colombian Economy, Google Trends

2. Introducción

2.1 Justificación

En el ámbito de la economía, los indicadores desempeñan un papel crucial al brindar una visión objetiva de la realidad económica de un país. El Departamento Administrativo Nacional de Estadística (DANE) en Colombia es reconocido por ser la fuente primordial de información sobre la actividad económica del país, aportando datos esenciales como el Producto Interno Bruto (PIB), indicadores del mercado laboral, índices de seguimiento económico y tasas de inflación, entre otros. Dentro de este contexto, se destaca el Índice de Precios de Vivienda Nueva como una variable macroeconómica de gran relevancia.

La vivienda es un componente fundamental de la economía y de la vida de los ciudadanos. Su impacto se extiende a diversos aspectos, incluyendo la inversión, el empleo, la demanda de materiales de construcción y la movilidad geográfica. Además, se podría decir que los precios de la vivienda están estrechamente relacionados con el Índice de Precios al Consumidor (IPC), un indicador clave que mide la variación de precios en una canasta básica de bienes y servicios. Los cambios en la inflación podrían influir directamente en los precios de la vivienda, lo que a su vez afecta la política monetaria y fiscal del país. Por lo tanto, contar con herramientas precisas para pronosticar los precios de la vivienda nueva es esencial para anticipar tendencias económicas y tomar decisiones informadas en diferentes sectores.

No obstante, el proceso de recolección y publicación de los indicadores económicos tradicionales presenta un desafío temporal. Los resultados suelen publicarse con un rezago respecto al período al que hacen referencia. Esto significa que, para obtener información sobre el Índice de Precios de Vivienda Nueva de un mes específico, se debe esperar aproximadamente un mes después de ese período para que los datos sean divulgados. Durante este período de espera, la toma de decisiones puede verse limitada por la falta de acceso a información actualizada.

2.2 Planteamiento del Problema

En el contexto actual de la economía colombiana, la capacidad de medir y prever variables macroeconómicas de manera precisa y oportuna es esencial para la toma de decisiones estratégicas. Aunque el Departamento Administrativo Nacional de Estadística (DANE) proporciona valiosa

información económica a través de indicadores tradicionales como el Producto Interno Bruto (PIB), estos datos sufren de un retraso temporal debido a los procesos de recolección y divulgación. Este desfase limita la capacidad de los actores económicos para responder ágilmente a cambios en el entorno económico.

Dentro de esta perspectiva económica colombiana, el mercado de viviendas emerge como un pilar fundamental en la estructura financiera del país. El Índice de Precios de Vivienda Nueva (IPVN) es un indicador crucial en este sector, proporcionando una visión integral de la salud del mercado inmobiliario. Sin embargo, la publicación de estos datos se ve afectada por un retraso de aproximadamente un mes, lo que complica la toma de decisiones en tiempo real para empresas, entidades gubernamentales y hogares, privándolos de información actualizada para fundamentar sus elecciones financieras y estratégicas.

En este contexto, se origina la necesidad imperiosa de adoptar enfoques innovadores para superar el inconveniente temporal en la disponibilidad de datos económicos. Una solución prometedora radica en la técnica de "nowcasting", que emplea volúmenes de búsqueda en tiempo real para prever indicadores económicos antes de su publicación oficial. Esta metodología ha demostrado su eficacia en diversas áreas y ha sido ampliamente utilizada en estudios que emplean datos de búsqueda en línea para anticipar tendencias económicas con alta precisión y en tiempo real.

2.3 Objetivo General

Desarrollar y validar, para el año 2023, un modelo de *nowcasting* que permita pronosticar con precisión el Índice de Precios de Vivienda Nueva (IPVNBR) en Colombia mediante algoritmos de machine learning en RStudio.

2.4 Objetivos Específicos

- Realizar un análisis exhaustivo de las fuentes de información disponibles que puedan contribuir a la construcción y alimentación del modelo de nowcasting para el Índice de Precios de Vivienda Nueva en Colombia (IPVNBR).

- Diseñar y desarrollar un modelo que transforme y estructure los datos extraídos de los volúmenes de búsqueda en motores de información relacionados con el mercado inmobiliario y los precios de viviendas
- Validar de manera rigurosa la calidad y precisión de las estimaciones obtenidas a través del modelo de nowcasting propuesto.

3. ANTECEDENTES

3.1 Marco Teórico

En el próximo apartado, se procederá a la definición de los conceptos fundamentales de índole económica, gestión de datos y aprendizaje automático. Estos conceptos representan los pilares técnicos sobre los cuales se sustenta la presente investigación.

3.1.1 Conceptos económicos

Nowcasting

El concepto de Nowcasting en el ámbito económico se ha consolidado como una herramienta de suma relevancia en la toma de decisiones y el análisis de las tendencias económicas. Según la definición propuesta por Bańbura y Reichlin, el Nowcasting se refiere a "la predicción del presente, el futuro cercano y el pasado reciente". En esencia, el Nowcasting se basa en la aplicación de modelos analíticos que permiten anticipar el comportamiento de una variable económica en tiempo real o con un rezago muy reducido, utilizando información de otras variables con una periodicidad menor a la de la variable de interés. Dicho de otro modo, se busca estimar el presente, el pasado inmediato y el futuro próximo de una serie temporal económica a partir de indicadores y datos disponibles con mayor frecuencia y prontitud.

Esta disciplina es esencial para la monitorización temprana de la economía, ya que, debido a los habituales rezagos en la publicación de datos económicos, el Nowcasting se convierte en un valioso instrumento para anticipar tendencias económicas, tomar decisiones más informadas y entender la dinámica económica en tiempo real. El Nowcasting, siguiendo la definición de Cabria, implica la aplicación de diversas acciones analíticas que culminan en la estimación de variables económicas a partir de datos con una periodicidad menor, lo que proporciona una visión más precisa y oportuna de la situación económica actual y futura.

Indicador económico

Un indicador económico es un dato económico estadístico que proporciona una medida cuantitativa de la realidad económica en un territorio específico. Estos indicadores, como el Producto Interno Bruto (PIB), tasas de ocupación y desempleo, tasas de interés, inflación, tipo de cambio, balanza de pagos, y la confianza del empresario y del consumidor, se obtienen mediante un conjunto de datos recopilados en un periodo determinado. Su función fundamental radica en la vigilancia y evaluación de la actividad económica. Además, permiten realizar predicciones sobre posibles eventos económicos futuros, contribuyendo así a la comprensión y análisis de la coyuntura económica en un área geográfica.

Índice de precios de la vivienda nueva (IPVNBR)

El Índice de Precios de la Vivienda Nueva (IPVNBR) se erige como una herramienta esencial para comprender la dinámica de los precios en el mercado de la vivienda nueva en Colombia, focalizándose en las ciudades de Bogotá, Medellín, Cali y municipios circundantes. Su metodología de cálculo se sustenta en el índice tipo Fischer con base fija, referencia temporalmente al mes de diciembre de 2006. La construcción de este índice implica una cuidadosa aplicación de fórmulas específicas, considerando datos de precios y áreas de todos los inmuebles nuevos disponibles para la venta en un período determinado. La información necesaria se extrae de La Galería Inmobiliaria y, posteriormente, es procesada y analizada meticulosamente por el Banco de la República.

La formulación del Índice Tipo Fischer con base fija, denotado como It*It*, en el contexto del IPVNBR, se define de la siguiente manera:

$$I_t = \left(\Pi_{i=1}^{n} \frac{P_{i,t}}{P_{i,0}} \right)^{\frac{1}{n}} \times 100$$

Donde:

- I_t es el valor del índice en el período t.
- n es el número de componentes en el índice.
- $P_{i,t}$ es el precio del componente i en el período t.
- $P_{i,0}$ es el precio del mismo componente en el período de referencia, en este caso, diciembre de 2006.

Este enfoque de índice tipo Fischer permite comparar de manera efectiva los cambios relativos en los precios de la vivienda nueva, proporcionando una medida precisa y representativa de la variación en el tiempo. La utilización de esta metodología respalda la generación de información confiable y relevante para la toma de decisiones económicas y políticas públicas en el ámbito inmobiliario.

3.1.2 Conceptos de datos y machine learning

Google Trends

Google Trends, desarrollado por Google, es una herramienta pública que ofrece seguimiento al interés de los usuarios en diversos temas, basándose en una muestra de búsquedas realizadas en el motor de búsqueda de Google. Los datos recopilados se normalizan en una escala de 0 a 100, representando la proporción con respecto al total de búsquedas, lo que permite la comparación entre regiones geográficas. Esta valiosa herramienta proporciona información relevante a los usuarios al analizar y visualizar tendencias en búsquedas, lo que puede ser de gran utilidad en investigaciones y análisis en campos diversos, incluyendo el nowcasting y la economía.

Series de tiempo

Una serie de tiempo se refiere a un conjunto de observaciones de una variable particular realizadas en intervalos temporales consistentes, tales como tasas de desempleo mensuales, PIB trimestral, inflación mensual o rendimientos diarios de una acción. La estructura de una serie de

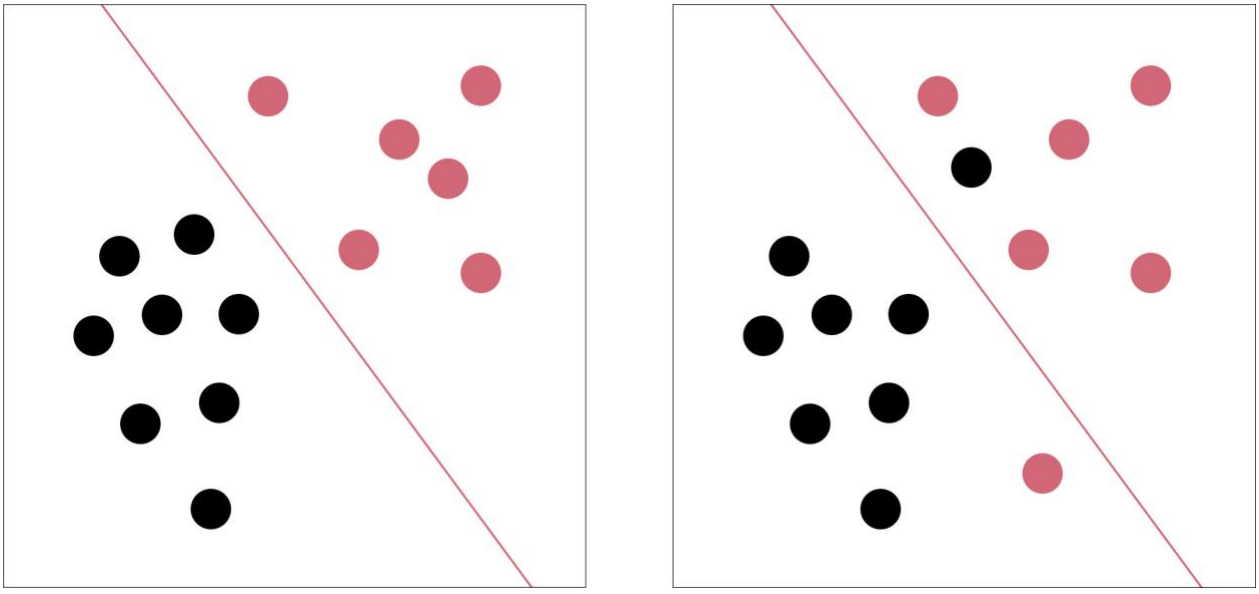
tiempo se representa de manera formal como $\{y_t | t \in 1,2,\dots,T\}$, donde y_t representa la observación en un tiempo específico t, y T denota el número total de observaciones. Estas series son fundamentales en el análisis económico y financiero, proporcionando información valiosa para comprender patrones, tendencias y ciclos temporales en datos económicos y financieros.

Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) representan una herramienta de gran relevancia en el ámbito del aprendizaje automático, hallando una amplia aplicación en tareas de clasificación y regresión de datos. Su propósito central reside en la instauración de hiperplanos en un espacio multidimensional, cuyo propósito fundamental consiste en la segregación y categorización de observaciones. En su forma más elemental, este hiperplano puede asumir la configuración de una línea recta, operando como una frontera de toma de decisiones que separa dos conjuntos de observaciones (Ilustración 1). La eficacia de esta segregación se mide a través de la distancia existente entre la línea de decisión y las observaciones más cercanas, que son comúnmente denominadas como vectores de soporte.

No obstante, en escenarios del ámbito real, la segregación lineal frecuentemente se evidencia como insuficiente debido a la complejidad presente en la distribución de los datos. Con el propósito de sortear esta limitación, las SVM permiten cierto grado de flexibilidad en la clasificación, lo que se conoce como *flexibilización*. En esta instancia, las observaciones ubicadas en las inmediaciones de la frontera de decisión, así como aquellas que se autoriza clasificar de manera incorrecta, desempeñan un papel esencial en la edificación de la máquina de soporte de vectores (Ilustración 1).

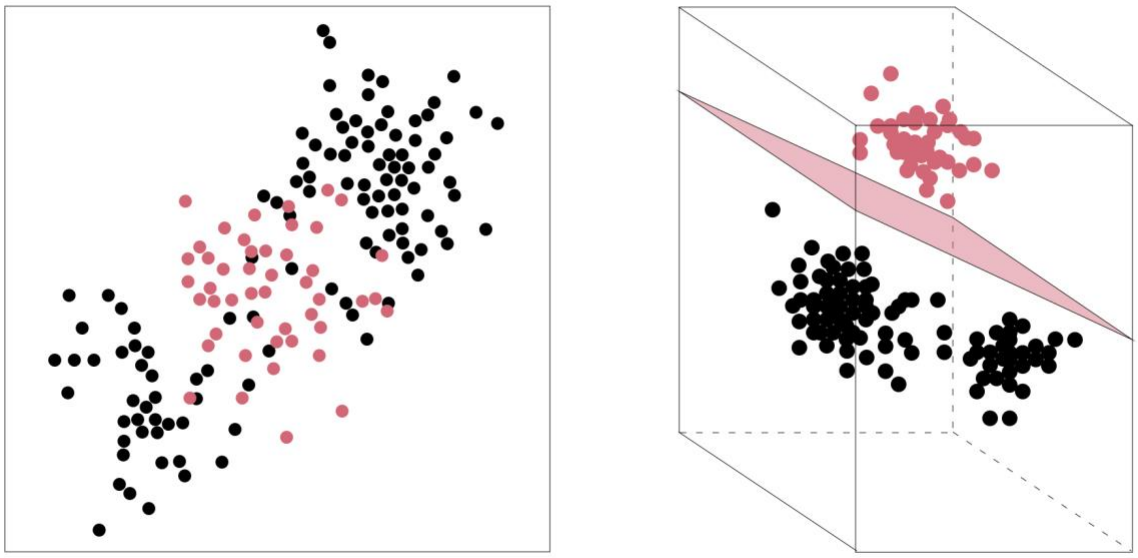
Ilustración 1: *Caso de separación con hiperplano lineal a partir de Máquinas de Vectores de Soporte (SVM)*



Nota. Elaboración propia

En situaciones en las cuales una segregación lineal persiste siendo inadecuada, las SVM exploran la expansión del grado del hiperplano de segregación con el propósito de investigar soluciones no lineales que se ajusten más adecuadamente a la estructura de los datos. Este proceso involucra la aplicación del kernel trick, una técnica que implica la proyección de los datos en un espacio dimensional superior, donde la segregación lineal puede resultar más eficaz (Ilustración 2).

Ilustración 2: *Caso de separación no lineal a partir de la técnica kernel trick*



Nota. Elaboración propia

Regresión Lasso

En el ámbito de la estadística y el aprendizaje automático, la Regresión Lasso, cuya denominación proviene de "Least Absolute Shrinkage and Selection Operator" (en español, operador de contracción y selección mínima absoluta), es una técnica de análisis de regresión que se utiliza para lograr dos objetivos clave: la selección de variables relevantes y la regularización de los coeficientes de un modelo de regresión. Esta técnica fue propuesta por Robert Tibshirani en 1996 y se basa en la idea de penalizar la magnitud de los coeficientes estimados para reducir la complejidad del modelo y, a su vez, mejorar la precisión y la interpretabilidad.

En esencia, la Regresión Lasso busca minimizar una función objetivo que consta de dos términos. El primero se relaciona con el error de predicción, que es la diferencia entre las observaciones reales y las predicciones del modelo. El segundo término incluye una penalización que depende de la magnitud de los coeficientes estimados, controlada por un parámetro λ (lambda).

A medida que λ aumenta, los coeficientes estimados se vuelven más pequeños, lo que puede llevar a la eliminación de algunas variables del modelo al forzar sus coeficientes a cero.

Una característica distintiva de la Regresión Lasso es su capacidad para realizar selección de variables al hacer que los coeficientes sean exactamente cero, lo que no es posible en otros métodos de regularización, como la Regresión Ridge. Esto la convierte en una herramienta poderosa para simplificar modelos y extraer características relevantes en situaciones en las que se tienen muchas variables predictoras. La elección del valor apropiado de λ es fundamental, ya que afecta el equilibrio entre la precisión y la esparcibilidad del modelo, lo que puede ser determinado a partir de los datos disponibles.

En el marco de la Regresión Lasso, partimos de la formulación básica de la regresión por mínimos cuadrados, donde se busca minimizar la suma de los cuadrados de los residuos (RSS):

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

La introducción de la regularización en Lasso se lleva a cabo mediante la inclusión del parámetro λ, transformando la función objetivo a:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Esta formulación incorpora una penalización basada en la magnitud absoluta de los coeficientes estimados

Red neuronal artificial

Las redes neuronales, pertenecientes al ámbito de la inteligencia artificial, son una herramienta fundamental en el campo del Deep Learning. En esencia, una red neuronal opera tomando una serie de variables como entrada, desarrollando una función no lineal y proporcionando una salida correspondiente. Aunque el proceso inicial parece similar a otros algoritmos, la distinción radica en su estructura característica.

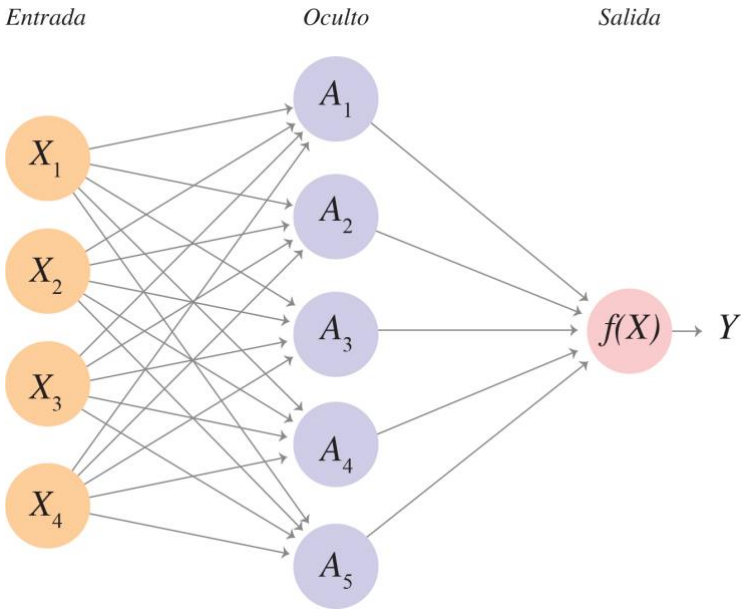
Matemáticamente, una red neuronal adopta la forma:

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X)$$

Donde $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$, siendo k el número de neuronas en la capa oculta, g una función de activación no lineal, y j el número de variables. De este modo, las activaciones A_k, se convierten en:

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k A_k$$

Ilustración 3: *Capas de una red neuronal*



Nota. Elaboración propia

Es evidente que este resultado guarda similitudes con una regresión lineal. Es importante mencionar que las funciones de activación pueden variar en su naturaleza, lo que influye directamente en la salida del modelo, otorgando a las redes neuronales una versatilidad excepcional para abordar distintos tipos de problemas analíticos.

Por otra parte, es relevante destacar que la inclusión de múltiples capas ocultas y salidas en una red neuronal constituye una estrategia plenamente factible, la cual substancialmente mejora su aptitud para afrontar desafíos de gran complejidad. Esto habilita un procesamiento más profundo y acentúa su versatilidad en lo concerniente a la representación de datos, lo que amplía su capacidad analítica.

Arboles de decisión

Los árboles de decisión son una poderosa técnica dentro del campo de la ciencia de datos y el aprendizaje automático que se utiliza para realizar decisiones basadas en reglas condicionales.

Estos árboles representan un modelo predictivo en forma de estructuras de árbol, donde cada nodo interno del árbol representa una característica o atributo, y cada rama representa una posible decisión o resultado asociado con esa característica. Las hojas del árbol contienen las predicciones o resultados finales. El proceso de construcción del árbol implica dividir recursivamente el conjunto de datos en función de las características más relevantes para maximizar la pureza de las hojas, lo que optimiza la capacidad del modelo para generalizar y tomar decisiones precisas en datos no vistos.

Boosting Regression

Boosting Regression es una técnica en el campo de la ciencia de datos que se utiliza para mejorar el rendimiento predictivo de modelos de regresión. A diferencia de métodos tradicionales que ajustan un único modelo a los datos, el *boosting* regresivo construye una secuencia de modelos débiles, donde cada nuevo modelo se enfoca en corregir los errores de los modelos anteriores. Este enfoque iterativo permite mejorar gradualmente la precisión de la predicción.

El algoritmo de *Gradient Boosting* es una combinación secuencial de modelos más simples, generalmente árboles de decisión poco profundos. La idea central radica en corregir los errores de los modelos anteriores, construyendo iterativamente modelos adicionales que se centren en los residuos del modelo anterior. Sea y_i la variable dependiente (la variable de valor continuo que queremos predecir) y x_i el vector de características correspondiente al i-ésimo ejemplo en nuestro conjunto de datos. Supongamos que ya tenemos un modelo $F_{m-1}(x)$ que representa la predicción acumulativa hasta la $(m-1)$-ésima iteración.

El objetivo es ajustar un nuevo modelo $h_m(x)$ para predecir los residuos restantes:

$$h_m(x) = y_i - F_{m-1}(x_i)$$

Luego, el modelo actualizado $F_m(x)$ se obtiene sumando este nuevo modelo ponderado a la predicción acumulativa anterior:

$$F_m(x) = F_{m-1}(x_i) + \eta \cdot h_m(x)$$

Donde η es la tasa de aprendizaje, un hiperparámetro que controla la contribución de cada modelo individual a la predicción final. La elección de este parámetro es crucial y debe ajustarse cuidadosamente para evitar sobreajuste o subajuste.

3.2 Estado del arte/trabajos relacionados

3.2.1 Revisión Internacional

Los indicadores económicos son herramientas esenciales para que economistas, empresas y responsables políticos analicen la situación económica actual y a corto plazo. Los primeros indicadores se crearon en la década de 1970 para dar señales tempranas de los puntos de inflexión de la actividad económica. El sistema de indicadores adelantados compuestos (IAC) de la OCDE se desarrolló para anticipar los puntos de inflexión del ciclo económico en unos siete meses. Recientemente, el uso del volumen de búsquedas de Google, a través de la herramienta Google Trends, ha permitido crear modelos capaces de predecir determinados comportamientos de la economía. Esta información puede ser utilizada por las empresas para ajustar sus estrategias de marketing y por los responsables políticos para tomar decisiones informadas.

Los estudios sobre la predicción de indicadores económicos y financieros con Google Trends han demostrado resultados prometedores, en especial en la predicción de precios de vivienda y permisos de construcción en varios países. Además, algunos estudios han encontrado una correlación positiva entre las búsquedas en Google Trends relacionadas con la compra de vivienda y el aumento de los precios de la vivienda en diferentes países (Bulczak, 2021) (Limnios & You, 2018). Sin embargo, es importante tener en cuenta que los resultados varían según los factores, como el país y las palabras clave utilizadas. Aunque se necesitan más investigaciones para determinar la precisión de estas predicciones, la investigación sobre la predicción de indicadores económicos y financieros con Google Trends sigue siendo un campo de investigación prometedor en constante evolución.

3.2.2 Revisión nacional

A la luz de la revisión bibliográfica efectuada, es relevante destacar el aumento del número de investigaciones que han empleado Google Trends como herramienta de análisis en el ámbito nacional. En particular, se han identificado investigaciones relacionadas con el sector de la salud, en las cuales se ha utilizado Google Trends para predecir la aparición de casos confirmados de Zika (Morsy et al., 2018) y COVID-19 (Ortiz Martínez et al., 2020), con el objetivo de apoyar la preparación del sistema sanitario. Asimismo, se han encontrado estudios relacionados con la política, en los cuales se han realizado predicciones sobre los resultados de las últimas elecciones en Colombia (A. Ospina y L. Caballero, 2019). Asimismo, se encontraron diversos artículos enfocados en la economía del país, que se han enfocado principalmente en la predicción en tiempo real de la tasa de desempleo del país (L. Cardona Rojas y J. Rojas Aguilera, 2017) (L. Trespalacios Cárdenas y A. García Suaza, 2021), mediante la utilización de modelos de nowcasting. Igualmente, se han creado modelos para predecir la llegada de turistas a Colombia (A. Correa, 2021). Sin embargo, es relevante destacar que no se ha encontrado ninguna investigación en Colombia que haya abordado la utilización de Google Trends para la predicción de indicadores económicos centrados en el sector inmobiliario.

4. Metodología

En el marco de este proyecto, la metodología propuesta para el pronóstico del Índice de Precios de Vivienda Nueva (IPVNBR) en Colombia mediante la técnica de nowcasting se fundamenta en un enfoque sistemático y preciso. Iniciamos con un profundo entendimiento del contexto económico, identificando el papel crucial del IPVNBR y su interconexión con otros indicadores macroeconómicos. A continuación, se realiza un minucioso análisis de las fuentes de datos disponibles, considerando la frecuencia y rezago temporal, así como la representación fiel de la realidad económica. La preparación de los datos se enfoca en garantizar la coherencia y calidad de los insumos del modelo, mientras que el modelamiento implica la selección y ajuste de algoritmos de machine learning en RStudio. Finalmente, la evaluación se centra en determinar el modelo

óptimo, no solo en términos de métricas cuantitativas, sino también en su coherencia con el sentido económico de los resultados. Este enfoque integral busca superar los desafíos temporales en la disponibilidad de datos, proporcionando un instrumento efectivo para anticipar tendencias en el dinámico mercado de viviendas colombiano.

Entendimiento del negocio

En esta etapa, se realizará una identificación exhaustiva del indicador económico clave para el nowcasting, en este caso, el Índice de Precios de Vivienda Nueva (IPVNBR) en Colombia. Se buscará comprender su relevancia en el contexto económico del país, su interpretación y la influencia potencial de factores económicos y sociales en su comportamiento. Se explorarán conexiones con otros indicadores macroeconómicos, especialmente aquellos vinculados al mercado inmobiliario y al Índice de Precios al Consumidor (IPC). Este entendimiento del negocio proporcionará una base sólida para la construcción y validación del modelo de nowcasting.

Entendimiento de los datos

En esta fase, se identificarán y examinarán las fuentes de datos disponibles necesarias para alimentar el modelo de nowcasting. Se analizará la frecuencia de actualización de estos datos y se considerará el rezago en su obtención, reconociendo la limitación temporal que se busca superar con el nowcasting. Se evaluará la representación de la realidad de los datos, buscando entender la variabilidad y la consistencia a lo largo del tiempo. Además, se explorarán posibles desafíos y limitaciones inherentes a las fuentes de datos seleccionadas.

Preparación de los datos

En esta etapa, se abordarán los requisitos de tratamiento de datos para asegurar la calidad y consistencia de los insumos del modelo de nowcasting. Se realizará la transformación y estructuración de los datos extraídos de las fuentes identificadas, adaptándolos al formato requerido por el modelo. Se determinarán las ventanas de tiempo óptimas para el entrenamiento y la

evaluación del modelo, considerando la naturaleza temporal de los datos y la necesidad de anticipar el comportamiento del IPVNBR en el corto plazo.

Modelamiento

En esta fase, se seleccionarán y estimarán los modelos de machine learning en RStudio que se utilizarán para realizar el nowcasting del IPVNBR. Se explorarán algoritmos adecuados para la tarea de pronóstico, ajustando sus parámetros para optimizar el rendimiento del modelo. Se llevará a cabo un análisis riguroso para garantizar que los modelos sean capaces de capturar patrones relevantes y proporcionar predicciones precisas en tiempo real.

Evaluación

En el último paso, se determinará el mejor modelo de nowcasting en función de métricas predefinidas, como precisión, error medio y capacidad de generalización. Además, se contrastarán los resultados obtenidos con el sentido económico del comportamiento del IPVNBR, asegurando que las predicciones sean coherentes con la realidad económica. Este proceso de evaluación garantizará la validez y utilidad práctica del modelo de nowcasting en el contexto específico del mercado de viviendas en Colombia.

La metodología implementada en este proyecto se destaca por su meticulosidad y la conjunción de datos económicos primordiales con información recabada en línea. Su finalidad reside en perfeccionar la exactitud y la oportunidad de las proyecciones relacionadas con el Índice de Precios de Vivienda Nueva (IPVNBR) en Colombia. El objetivo principal radica en proporcionar información de valor a los agentes económicos y las autoridades gubernamentales, a fin de fundamentar decisiones informadas en el ámbito económico y en el sector inmobiliario del país.

Para materializar el propósito de esta investigación, se llevará a cabo un minucioso estudio en el entorno de R Studio, haciendo uso de una serie de librerías especialmente diseñadas para la

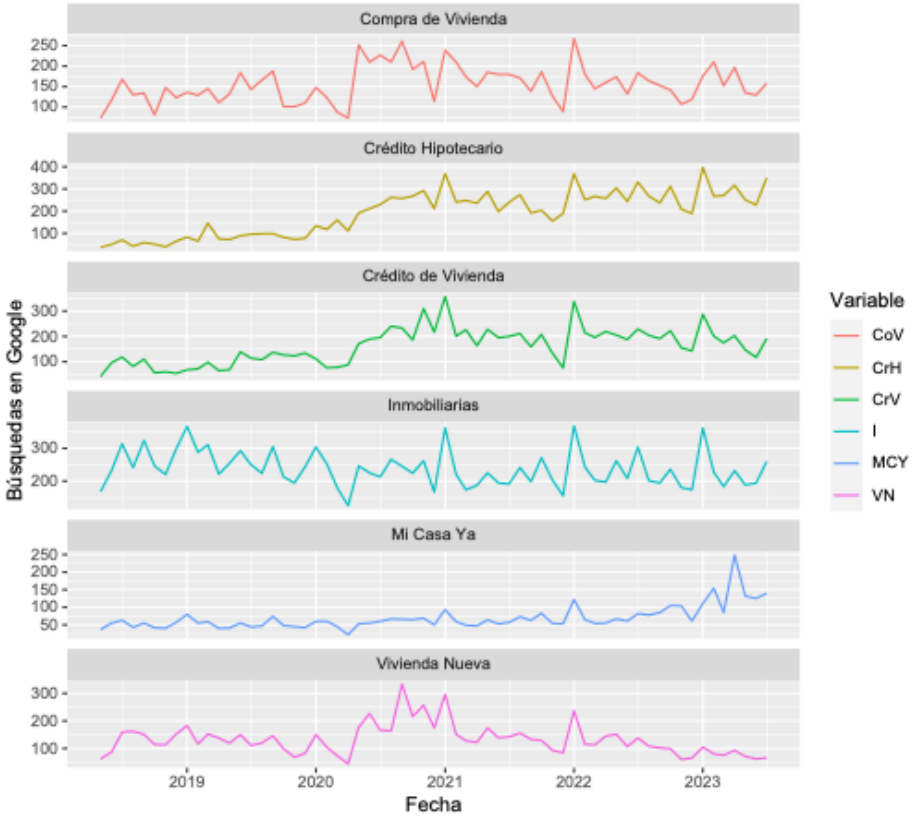
manipulación de datos. Entre las destacadas librerías que se emplearán, se encuentran: readxl, readr, dplyr, lubridate, zoo, plm, caret, randomForest, forecast, ggplot2, tidyr. Estas herramientas simplificarán el proceso de importación de bases de datos esenciales, el posterior procesamiento y depuración de los datos, la realización de análisis estadísticos y la construcción de modelos de aprendizaje automático.

Las bases de datos, cuya significación es crucial en esta investigación, comprenden las siguientes fuentes de información: las tasas de intervención del Banco de la República, la tasa de desempleo publicada por el Departamento Administrativo Nacional de Estadística (DANE), la tasa IPVN (Índice de Precios de Vivienda Nueva) publicada por el Banco de la República y los datos de búsqueda en Google Trends relacionados con el mercado inmobiliario. Estos datos desempeñarán un papel central en el análisis y la predicción del IPVNBR.

La recopilación de datos se efectúa desde diversas fuentes y en distintos formatos. En cuanto a las tasas de intervención, la tasa de desempleo y el IPVN, los datos se han registrado desde el 1 de enero de 2018. En lo que respecta a la tasa de intervención del Banco de la República, se ha recopilado información de forma diaria hasta el 11 de septiembre de 2023. Este conjunto de datos se erige como un indicador crítico para comprender la política monetaria del país y sus efectos en el mercado inmobiliario. En lo que respecta a la tasa de desempleo, se han recopilado datos de manera mensual hasta el 1 de julio de 2023. Estos datos proporcionan una perspectiva fundamental sobre el mercado laboral en Colombia y su influencia en el sector inmobiliario. El índice de precios de vivienda nueva (IPVN) se recopila de manera mensual, abarcando el periodo que llega hasta el 1 de julio de 2023. Este indicador ocupa una posición central en la investigación, ya que refleja las variaciones en los precios de la vivienda nueva en Colombia. Por último, se han adquirido datos relacionados con búsquedas en línea utilizando palabras clave como "Crédito hipotecario", "Crédito de vivienda", "Inmobiliarias", "Compra de vivienda", "Vivienda nueva" y "Mi casa ya". Estos datos se han recopilado desde el 16 de septiembre de 2018 hasta el 3 de septiembre de 2023, con una periodicidad semanal. Cabe destacar que, con la finalidad de facilitar los cálculos y análisis, todos los datos se han convertido a una periodicidad mensual. Esta conversión garantiza que todos los datos se encuentren en la misma escala temporal, simplificando de este modo la

comparación y el análisis de series temporales. En la ilustración [4] se presenta la serie de tiempo de cada uno de los términos de búsqueda en Google Trends.

Ilustración 4: Series de tiempo de índices de Google Trends



Nota. Elaboración propia

5. Presentación de la Propuesta

La propuesta concebida para afrontar la problemática delineada en las secciones precedentes se ha desarrollado minuciosamente, presentando un enfoque estructurado que se refleja de manera integral en la ilustración 6. Este esquema detallado encapsula cada elemento esencial de la estrategia propuesta, trazando una ruta clara desde la identificación de la problemática hasta la aplicación práctica de la solución. En esta representación gráfica, se destacan los puntos clave de la metodología, resaltando la interconexión entre las etapas cruciales del proceso. La figura sirve

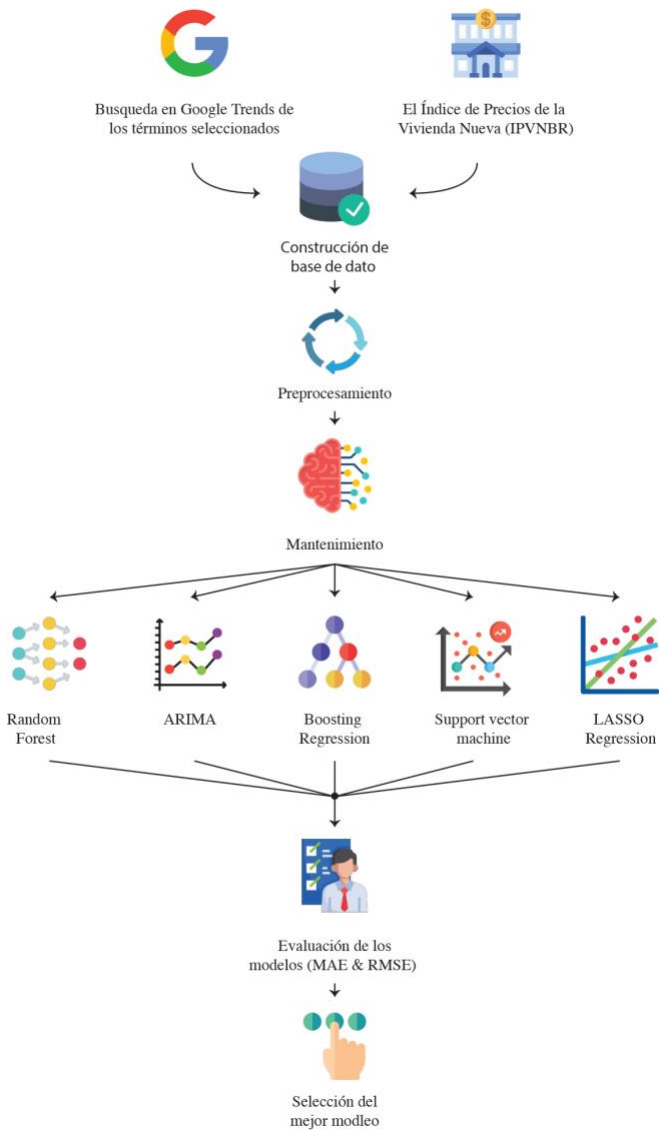
como una guía visual que facilita la comprensión y la comunicación efectiva de la propuesta, proporcionando una visión panorámica y detallada de la secuencia lógica de pasos que se seguirán para la implementación del modelo de nowcasting del Índice de Precios de Vivienda Nueva (IPVNBR) en el contexto económico colombiano.

5.1 Datos y software

La recopilación de datos se extiende a diversas fuentes y formatos. Las tasas de intervención, incluyendo la tasa de desempleo y el Índice de Precios de Vivienda Nueva (IPVN), se registran desde el 1 de enero de 2018. En cuanto a la tasa de intervención del Banco de la República, se recopila información diaria hasta el 11 de septiembre de 2023, siendo un indicador crítico para comprender la política monetaria y sus efectos en el mercado inmobiliario. La tasa de desempleo, capturada mensualmente hasta el 1 de julio de 2023, proporciona una perspectiva fundamental sobre el mercado laboral y su influencia en el sector inmobiliario. El IPVN se recopila mensualmente hasta el 1 de julio de 2023, ocupando una posición central al reflejar las variaciones en los precios de la vivienda nueva en Colombia. Además, se han adquirido datos relacionados con búsquedas en línea, desde el 16 de septiembre de 2018 hasta el 3 de septiembre de 2023, con palabras clave como "Crédito hipotecario", "Crédito de vivienda", "Inmobiliarias", "Compra de vivienda", "Vivienda nueva" y "Mi casa ya". Estos datos se han convertido a periodicidad mensual para facilitar comparaciones y análisis.

El tratamiento de datos y la modelación se llevaron a cabo exclusivamente en el entorno del software estadístico R, permitiendo realizar análisis robustos y aplicar

Ilustración 5: *Diagrama de la propuesta planteada*



Nota. Elaboración propia

El proceso de validación de los modelos propuestos se desglosa en dos vertientes. En primera instancia, se realiza un análisis comparativo de los modelos de regresión previamente citados: la regresión Lasso, Boosting Regressor, SVM y Random Forest. Este enfoque tiene como propósito discernir cuál de estos modelos se aproxima de manera más acertada y realista a las tendencias reales, fundamentado en los datos extraídos de Google Trends, detallados en secciones precedentes.

En segundo lugar, se llevará a cabo una evaluación de uno de los modelos ampliamente reconocidos en la predicción de indicadores económicos: el modelo ARIMA. Esta medida permitirá contrastar la efectividad de los modelos de regresión propuestos, basados en las búsquedas de Google Trends, frente al método tradicional de ARIMA en la predicción de indicadores económicos.

El análisis de estas validaciones se fundamentará en métricas clave, MAE (Mean Absolute Error) y RMSE (Root Mean Squared Error), esenciales en la evaluación de la precisión de los modelos predictivos. Es imperativo recordar que valores más bajos de MAE y RMSE reflejan una capacidad predictiva más sólida y confiable del modelo en cuestión. Este criterio será fundamental para la valoración y selección de los modelos más idóneos y eficientes en la predicción del Índice de Precios de Vivienda Nueva en el contexto colombiano.

6. Resultados Obtenidos

La evaluación de los modelos propuestos revela tendencias interesantes en su desempeño para predecir el Índice de Precios de Vivienda Nueva (IPVN). La tabla 1 presenta los valores de RMSE y MAE, esenciales para comparar la precisión de cada modelo.

Tabla 1:
Resultados sobre la muestra de evaluación de los modelos de regresión

Modelo	RMSE	MAE
Random Forest	3.745885	3.557661
Regresión Lasso	0.407449	0.3688524
SVM (Máquinas de Vectores de Soporte)	4.204172	3.825155
Boosting Regressor	6.917697	6.854578
ARIMA	1.052001	0.8086242

Nota. Elaboración propia

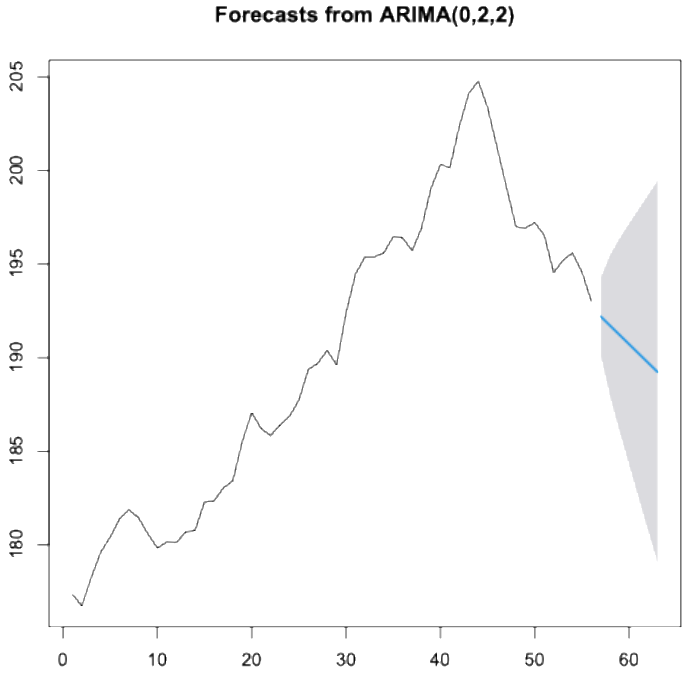
La regresión Lasso destaca con valores bajos en ambas métricas, indicando su alta precisión y rendimiento en la predicción del IPVN. En contraste, modelos como Random Forest y SVM exhiben métricas relativamente más altas, insinuando su menor eficiencia. No obstante, el Random Forest y el SVM sobresalen en la gestión de conjuntos de datos complejos y la mitigación del sobreajuste, mostrando su capacidad de lidiar con la complejidad de las tendencias.

La comparación entre estos modelos sugiere que, aunque los valores de Random Forest y SVM son más altos que el Boosting Regressor, todavía tienen potencial de optimización en la precisión para pronosticar el IPVN. Sin embargo, la Regresión Lasso parece ser el modelo más preciso al predecir el IPVN en base a los valores de RMSE y MAE.

Al comparar la Regresión Lasso con el ARIMA, se reafirma que la Regresión Lasso es más precisa en la predicción del IPVN, aunque el ARIMA no está lejos en términos de precisión. El ARIMA, que no indica una autocorrelación significativa en sus residuos según la prueba de Box-Ljung y mantiene su normalidad según la prueba de Shapiro-Wilk, demuestra que sus predicciones no difieren considerablemente de la realidad.

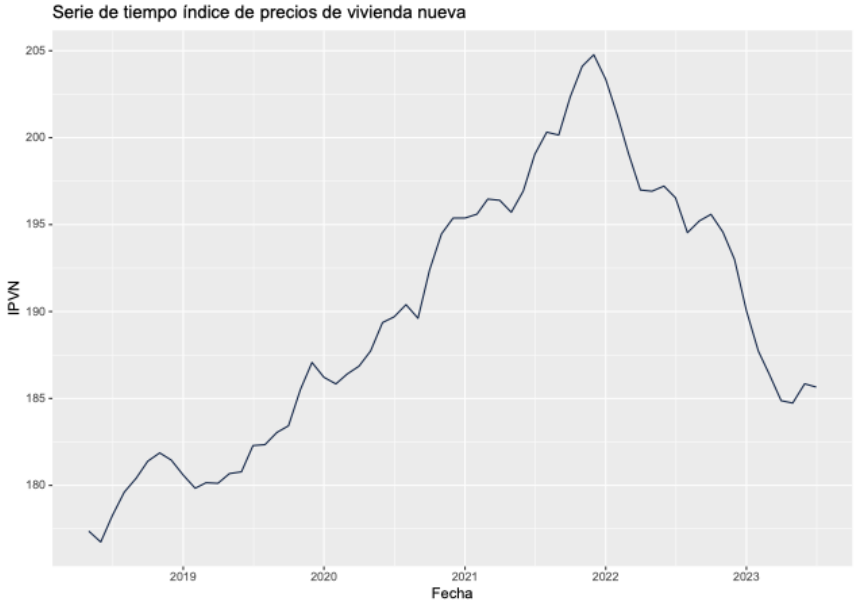
En la ilustración 6 los pronósticos de ARIMA para los próximos tres meses son presentados:

Ilustración 6: Pronóstico IPVN a tres meses.



Nota. Elaboración propia

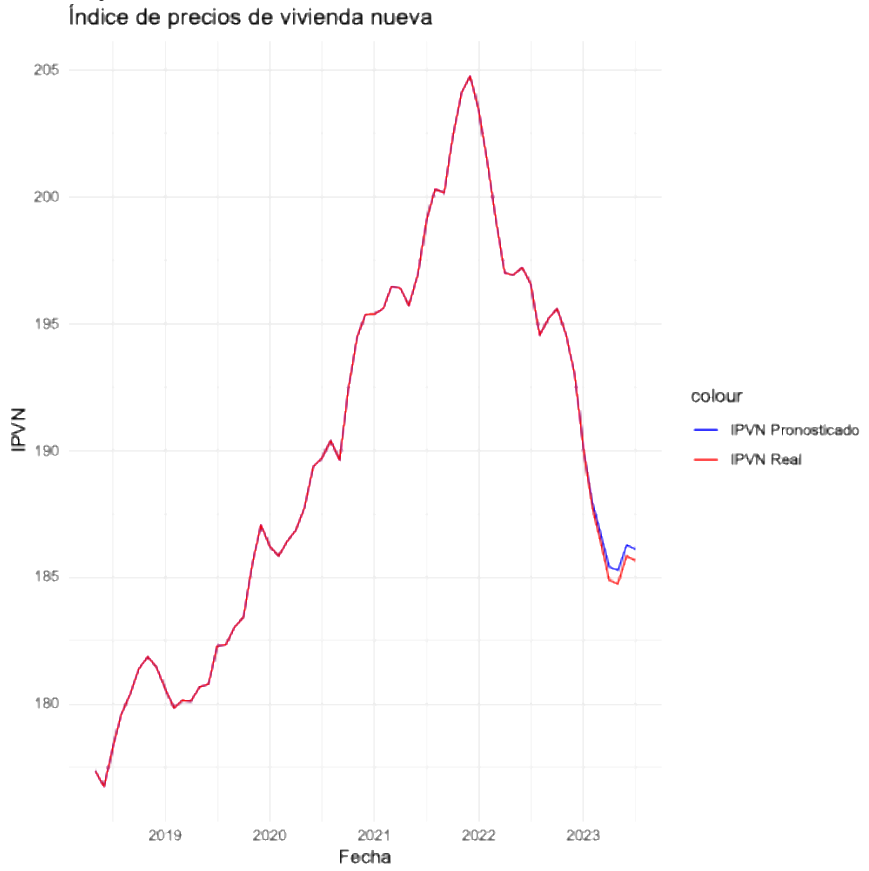
Ilustración 7: Series de tiempo índice de precios de vivienda nueva



Nota. Elaboración propia

Se evidencia que la tendencia de los datos pronosticados para los meses posteriores a la fecha final evaluada (1/12/2022) coinciden en gran medida con la serie de tiempo de los valores reales del IPVN, demostrando la eficiencia del modelo. Estos hallazgos apuntan a la eficiencia del ARIMA como un modelo aplicable y cercano a la realidad, a pesar de no ser tan preciso como la Regresión Lasso. Esto se evidencia de mejor forma en la ilustración [#]:

Ilustración 8: IPVN Real y Predicción



Nota. Elaboración propia

La representación gráfica anterior muestra una proximidad notable entre la predicción del Índice de Precios de Vivienda Nueva (IPVN) y el valor real del indicador. Esto evidencia la precisión del modelo de regresión Lasso, el cual proporciona pronósticos muy cercanos a la realidad, ofreciendo proyecciones ajustadas a futuros posibles escenarios basados en las búsquedas realizadas en Google Trends.

Teniendo estos resultados en cuenta, se evidencia que el ARIMA, al demostrar que sus residuos no presentan autocorrelación significativa y mantienen la normalidad, confirma la solidez de su desempeño en la predicción del IPVN. Mientras tanto, la Regresión Lasso, al ser un modelo de regresión lineal regularizado, destaca por su capacidad para interpretar de manera directa las relaciones entre las variables predictoras y el IPVN basado en las búsquedas de Google Trends. Este enfoque puede proporcionar una mayor comprensión de los factores y relaciones subyacentes que impulsan las tendencias del IPVN, lo que resulta valioso para la predicción precisa de este indicador económico.

7. Conclusiones

A lo largo de esta investigación, se ha llevado a cabo una evaluación rigurosa de distintos modelos predictivos para el Índice de Precios de Vivienda Nueva (IPVN) en Colombia. Este estudio representa una contribución significativa a la predicción de indicadores inmobiliarios bajo el paradigma del nowcasting, utilizando la regresión Lasso basada en datos obtenidos de Google Trends. La resolución del objetivo principal del estudio, que era determinar el modelo más preciso y confiable para pronosticar este indicador económico, destaca la importancia del modelo de nowcasting para la anticipación y la toma de decisiones en un entorno económico siempre cambiante.

En el contexto colombiano, donde la economía es altamente influenciable por factores internos y externos, contar con modelos predictivos precisos es fundamental. Al evaluar y comparar modelos de regresión, la regresión Lasso sobresale como el más preciso y efectivo, ofreciendo una predicción notablemente más certera del IPVN en comparación con otras alternativas de regresión lineal fundamentadas en datos de Google Trends. Esta precisión se refuerza con los resultados de la regresión Lasso en comparación con el modelo ARIMA, donde se evidenció que, si bien este último mostró una competencia cercana, la Regresión Lasso presentó valores más bajos en RMSE y MAE, sugiriendo una mayor precisión en la predicción del IPVN.

Este aporte significativo en la precisión de la predicción de un indicador económico crítico, como el IPVN, promete desempeñar un papel vital en el respaldo de decisiones informadas en el ámbito económico y el sector inmobiliario del país. Además, resalta la importancia de los modelos basados en datos de búsqueda en Google Trends para la predicción de tendencias del mercado inmobiliario, siendo una herramienta valiosa para actores económicos y autoridades gubernamentales en la toma de decisiones futuras, particularmente en un país con una economía en desarrollo donde el mercado inmobiliario es esencial para el crecimiento económico general.

Los resultados obtenidos brindan una base sólida para futuras aplicaciones. Al demostrarse la alta precisión del modelo de Regresión Lasso en la predicción del IPVN, se sugiere su uso en otros contextos económicos. Además, este estudio podría ampliarse, incorporando nuevas variables o indicadores económicos relevantes. Esta metodología, además de proyectar comportamientos económicos en diferentes sectores, se vislumbra como una herramienta valiosa para instituciones gubernamentales, inversores y planificadores urbanos, no solo en Colombia sino en otros contextos económicos similares.

8. Bibliografía

Askitas, N. (2016). Trend-Spotting in the Housing Market. Cityscape, 18(2), 165–178. http://www.jstor.org/stable/26328261

Banco de la República. (2023, 12 de noviembre). Índice de precios de la vivienda nueva (IPVNBR). Recuperado de https://www.banrep.gov.co/es/estadisticas/indice-precios-vivienda-nueva-ipvnbr

Banco de la República. (2023, 12 de noviembre). Tasas de interés de política monetaria. Recuperado de https://www.banrep.gov.co/es/estadisticas/tasas-interes-politica-monetaria

Bulczak, G. M. (2021). Use of Google Trends to Predict the Real Estate Market: Evidence from the United Kingdom. *INTERNATIONAL REAL ESTATE REVIEW*, 19.

Cardona Rojas, L. F., & Rojas Aguilera, J. A. (2017). Pronósticos para la tasa de desempleo en Colombia a partir de Google Trends. Archivos de Economía, 16050. Departamento Nacional de Planeación.

Coble, D., & Pincheira, P. M. (1 de febrero de 2017). Now-Casting Building Permits with Google Trends. SSRN. https://ssrn.com/abstract=2910165

Correa, Alexander. (2021). Prediciendo la llegada de turistas a Colombia a partir de los criterios de Google Trends. Lecturas de Economía, (95), 105-134. Epub September 29, 2021.https://doi.org/10.17533/udea.le.n95a343462

Correa Ospina, A., & Caballero Andrade, L. E. (2019). Modelo de Pronóstico para la Demanda de Turistas en Colombia a partir de Criterios de Búsqueda en Google: Una Aproximación Utilizando la Metodología MIDAS. Bogotá.

DANE. (2023, 12 de noviembre). Empleo y desempleo en Colombia. Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo

Limnios, C., & You, H. (11 de noviembre de 2018). Can Google Trends Actually Improve Housing Market Forecasts?. SSRN. https://ssrn.com/abstract=2886705

Lynn Wu, & Erik Brynjolfsson. (2015). The Future of Prediction How Google Searches Foreshadow Housing Prices and Sales. *National Bureau of Economic Research*.

Morsy S, Dang TN, Kamel MG, Zayan AH, Makram OM, Elhady M, Hirayama K, Huy NT (2018). Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. Epidemiology and Infection 146, 1625–1627. https://doi.org/10.1017/S0950268818002078

Ortiz-Martínez, Y., Garcia-Robledo, J. E., Vásquez-Castañeda, D. L., Bonilla-Aldana, D. K., & Rodriguez-Morales, A. J. (2020). Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia. Travel medicine and infectious disease, 37, 101703. https://doi.org/10.1016/j.tmaid.2020.101703

Rizun, N. (2021). *Can Web Search Queries Predict Prices Change on the Real Estate Market? .*

Trespalacios Cárdenas, L. M. (2021). Modelo de Nowcasting para Pronosticar la Tasa de Desempleo de Colombia Utilizando Google Trends. Universidad EIA, Ingeniería Financiera Envigado.