FORECASTING DE ÍNDICE DE CONFIANZA DEL CONSUMIDOR CON GOOGLE TRENDS

PROYECTO DE GRADO

LUISA MARÍA GARIZO PUERTO
MANUELA RAMÍREZ MARTÍNEZ

CRISTIAN CAMILO HOYOS BERMEO

FACULTAD DE CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS
ECONOMÍA Y NEGOCIOS INTERNACIONALES
MERCADEO INTERNACIONAL Y PUBLICIDAD
SANTIAGO DE CALI
2023

FORECASTING DE ÍNDICE DE CONFIANZA DEL CONSUMIDOR CON GOOGLE TRENDS

LUISA MARÍA GARIZO PUERTO
MANUELA RAMÍREZ MARTÍNEZ

CRISTIAN CAMILO HOYOS BERMEO

FACULTAD DE CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS
ECONOMÍA Y NEGOCIOS INTERNACIONALES
MERCADEO INTERNACIONAL Y PUBLICIDAD
SANTIAGO DE CALI
2023

TABLA DE CONTENIDO

pág.

Contenido

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

El índice de Confianza del Consumidor que se calcula actualmente en Colombia se obtiene con más de un mes de rezago, lo que conlleva implicaciones adversas para la toma de decisiones y obstaculiza la utilización oportuna de dicha información.

No obstante, el surgimiento de nuevas tecnologías y fuentes de información permite la complementación y mejora de las estimaciones convencionales del ICC. Un ejemplo de ello es Google Trends, una herramienta proporcionada por el navegador Google, que resulta valiosa para analizar y explorar las tendencias en las búsquedas de palabras clave a lo largo del tiempo. En consecuencia, en la presente investigación se busca desarrollar un modelo de nowcasting para el índice de confianza del consumidor mensual en Colombia.

Para la elaboración de la propuesta se utilizaron modelos de Support Vector Machine (SVM), Regresión de Lasso y Redes Neuronales. A partir de los resultados, se determina que el modelo que mejor logra predecir el comportamiento del ICC es el SVM Kernel Sigmoid, con un MAE de 3,16 y considerando el punto de referencia establecido, que fue el modelo ARIMA.

Con base a lo expuesto, se puede concluir que el modelo de nowcasting propuesto, que aprovecha las nuevas tecnologías y fuentes de información, constituye una predicción acertada. Esto representa un avance significativo en el ámbito económico, ya que brinda la posibilidad de obtener esta información de manera oportuna para la toma de decisiones.

1.1 Palabras Claves

Índice de Confianza del Consumidor, Google Trends, Nowcasting.

ABSTRACT

The Consumer Confidence Index currently calculated in Colombia is obtained with a lag of more than one month, which has adverse implications for decision making and hinders the timely use of such information.

However, the emergence of new technologies and sources of information allows the complementation and improvement of conventional estimates of the CCI. An example of this is Google Trends, a tool provided by the Google browser, which is valuable for analyzing and exploring trends in keyword searches over time. Consequently, this research seeks to develop a nowcasting model for the monthly consumer confidence index in Colombia.

Support Vector Machine (SVM), Lasso Regression and Neural Networks models were used to develop the proposal. From the results, it is determined that the model that best predicts the behavior of the CCI is the SVM Kernel Sigmoid, with a MAE of 3.16 and considering the established benchmark, which was the ARIMA model.

Based on the above, it can be concluded that the proposed nowcasting model, which takes advantage of new technologies and information sources, is an accurate prediction. This represents a significant advance in the economic field, since it offers the possibility of obtaining this information in a timely manner for decision making.

1.2 Key Words

Consumer Confidence Index, Google Trends, Nowcasting.

2. INTRODUCCIÓN

2.1 Contexto

El índice de confianza del consumidor (ICC) es un indicador crucial para comprender la percepción y expectativas de los consumidores en relación con la economía de un país. En el caso de Colombia, el ICC desempeña un papel significativo en la evaluación de la salud económica y en la toma de decisiones empresariales y políticas. Tradicionalmente, la recopilación de datos para calcular el ICC se ha basado en cuestionarios, para lo cual se requiere un tiempo de procesamiento y cálculo que resulta en boletines que no están actualizados. De hecho, el ICC en Colombia que se calcula actualmente se obtiene con más de un mes de rezago, lo cual trae implicaciones negativas como: reacciones tardías e ineficientes por parte de las empresas y los encargados de la formulación de políticas, toma de decisiones desinformadas, pérdida de oportunidades y dificultad para la evaluación de políticas.

No obstante, con el avance de las tecnologías digitales y el aumento en el uso de Internet, han surgido nuevas fuentes de información que pueden complementar e incluso mejorar las estimaciones tradicionales del ICC. Este es el caso de Google Trends, una herramienta del navegador Google, que es útil para el análisis y exploración de tendencias de búsqueda de palabras claves a través del tiempo. En consecuencia, el presente estudio busca construir un modelo de nowcasting para el índice de confianza del consumidor mensual en Colombia. El enfoque se centra en aprovechar los datos disponibles en Google Trends para actualizar el ICC de forma inmediata. Lo anterior, se realizará a través de un análisis de correlación y técnicas de modelado, por medio de las cuales se evaluará la relación entre el ICC y los términos relevantes

de búsqueda en Google Trends. Esto, permitirá identificar aquellos términos de búsqueda que puedan ser predictivos del ICC y seleccionar la mejor aproximación para estimar este indicador.

El texto se divide en siete secciones: la presente introducción; el marco teórico, donde están los conceptos claves abordados; la revisión bibliográfica, que reúne los principales estudios realizados hasta la fecha sobre esta técnica; la metodología, donde se incluye la explicación los modelos de regresión empleados y la base de datos construida; los resultados encontrados con los datos; la discusión y análisis de los modelos corridos; y, por último, las conclusiones de la investigación.

2.2 Planteamiento del Problema

En Colombia existen diversas instituciones tanto públicas y privadas que se encargan de la medición y publicación de múltiples indicadores económicos. Aun cuando estos indicadores son rigurosos y cuentan con mucha aceptación para la política pública e investigaciones privadas, todos cuentan con un rezago en la publicación que provoca que no sean reflejo fiel de la actualidad del país. Por ejemplo, el PIB se presenta trimestralmente, los indicadores del mercado laboral bimensualmente, el IPC cada mes y tasas de interés consumo cada tres meses. Esto genera fricciones en la exactitud y asertividad de las políticas que se basan en imágenes del estado del país antiguos.

En el caso específico del ICC, FEDESARROLLO, que es el único ente que lo calcula, se publican los resultados con un rezago de un mes y medio. Además, este índice es una representación de las expectativas de los agentes, las cuales se construyen a partir de cómo perciben el mercado laboral y el comportamiento de los precios, por ejemplo. Esto sugiere que el índice se construye sobre observaciones de la inflación y desempleo que están aún más

desactualizadas que el rezago de 45 días que tiene el índice. Así como lo menciona Hoyos: “Esto implica que estos agentes económicos tomen decisiones bajo incertidumbre, así como la posibilidad de creación de falsas expectativas” (Nowcasting de la tasa de desempleo en Colombia, usando información de Google Trends., 2021). Por lo que es de gran utilidad, buscar un método en el que se logre hacer una estimación con información actual y que se calcule de forma rápida, para que haya una publicación realmente oportuna.

2.3 Objetivo General

Construir un modelo de nowcasting para el índice de confianza del consumidor mensual en Colombia.

2.4 Objetivos Específicos

- Evaluar la correlación entre el ICC y los términos relevantes de las búsquedas en Google Trends.
- Identificar términos de búsqueda relacionados con la terminología económica que ayuden a pronosticar el ICC.
- Evaluar los modelos de aproximación para predecir el ICC con fuente de información de Google Trends.

3. ANTECEDENTES

3.1 Marco Teórico

En esta sección se presentará el marco teórico separado por Conceptos Económicos y Modelos de Regresión.

3.1.1 Conceptos Económicos

Confianza del Consumidor

La confianza del consumidor se refiere a la percepción individual y colectiva de los agentes económicos sobre el estado presente y futuro de la economía, así como sus propias condiciones económicas. Este concepto, introducido por Keynes (Keynes, 1936) a través del término "espíritus animales", abarca factores psicológicos como la confianza, la especulación y el entusiasmo, que influyen en las decisiones de consumo e inversión. Según Pickering et.al (Identification and Measurement of Consumer Confidence: Methodology and Some Preliminary Results, 1973) esta confianza se origina de una combinación de estímulos externos e internos que forman actitudes hacia la economía. Estas actitudes, ya sean optimistas, pesimistas o neutrales, moldean el comportamiento de los consumidores en términos de consumo e inversión. En síntesis, el objetivo de analizar esta confianza es determinar la percepción de los agentes sobre la economía, con el fin de conocer los patrones de consumo agregado en un futuro próximo. Las decisiones económicas de los agentes están influenciadas por factores emocionales, sociales y el contexto económico del momento.

Google Trends

Es una herramienta gratuita del navegador Google que es útil para el análisis y el descubrimiento de tendencias. Esta hace posible conocer datos del volumen de búsqueda para una palabra específica a lo largo de un periodo de tiempo específico y en regiones geográficas

específicas. Además, permite hacer comparaciones de la popularidad de diferentes términos y hacer predicciones de tendencias futuras con base en patrones de búsqueda anteriores. Esta herramienta, funciona a partir del análisis de los datos obtenidos de Google Search, estos se agregan y así se puede conocer la forma en la que las personas investigan sobre ciertos temas por medio de gráficas y tablas que evidencian el comportamiento a lo largo del tiempo.

Nowcasting

Se define como la predicción del presente, el futuro cercano y el pasado reciente. El término es una contracción del inglés para now (ahora) y forecasting (pronóstico) y se ha utilizado recientemente en economía (Giannone, Reichlin y Small, 2008). El nowcasting, a diferencia de otros modelos convencionales de predicción, no hace uso de datos históricos, sino que emplea los datos más recientes disponibles para la predicción inmediata. Lo anterior es relevante, pues esto suministra información importante para identificar el estado actual de la economía, que normalmente se suele obtener de forma rezagada, lo cual hace que las predicciones estén atrasadas y no sean tan útiles. “El nowcasting es relevante en economía porque las estadísticas clave sobre el estado actual de la economía están disponibles con un retraso significativo. Esto es particularmente cierto para los que se recopilan trimestralmente, siendo el Producto Interno Bruto (PIB) un ejemplo destacado... [Sin embargo,] la proyección inmediata también se puede aplicar significativamente a otras variables objetivo que revelen aspectos particulares del estado de la economía y seguidos de cerca por los mercados. Un ejemplo son las variables del mercado laboral.” (Bánbura, Giannone, Modugno, & Reichlin, 2013)

Índice de Confianza del Consumidor – Colombia

El ICC en Colombia es medido por la Fundación para la Educación Superior y el Desarrollo (FEDESARROLLO) -entidad privada sin ánimo de lucro-, quien lo calcula desde el

2001. Esta organización tiene como propósito contribuir al diseño, seguimiento y mejoramiento de las políticas públicas a través de la investigación en áreas políticas, sociales y económicas. Además, es la única entidad del país que se encarga de la construcción, medición y publicación del índice.

El ICC se construye mediante la Encuesta de Opinión del Consumidor (EOC), y su objetivo es obtener información sobre las percepciones actuales y las perspectivas de los hogares a un año vista. El indicador permite cuantificar la disposición de los consumidores a gastar en bienes durables (como vivienda, electrodomésticos y vehículos), la disposición a ahorrar y la solicitud de crédito formal e informal. La frecuencia de cálculo del índice es mensual, y se puede desagregar por principales ciudades (Bogotá, Medellín, Barranquilla, Cali y Bucaramanga) o por niveles socioeconómicos (alto, medio y bajo).

La metodología utilizada para el cálculo del ICC sigue la propuesta de la Universidad de Michigan con el Índice del Sentimiento del Consumidor, el cual es una medida mensual que evalúa y rastrea la confianza de los consumidores. Se basa en una encuesta telefónica a 500 personas seleccionadas al azar. El índice tiene dos componentes: el Índice de Condiciones Económicas, que refleja la percepción de los consumidores sobre su situación financiera actual y las condiciones económicas generales, y el Índice de Expectativas del Consumidor, que refleja las perspectivas futuras de los consumidores sobre la economía.

En el caso del ICC en Colombia, FEDESARROLLO realiza una encuesta que consta de 22 preguntas cualitativas, de las cuales se seleccionan 5 para la construcción del índice. El ICC se calcula como un promedio simple del Índice de Expectativa del Consumidor (IEC), que muestra las expectativas de los encuestados para el próximo año de su situación económica actual, y el Índice de Condiciones Económicas (ICE), que hace alusión a la valoración del estado económico

actual en comparación con el año anterior. El cálculo del índice se basa en los balances o saldos netos, que representan las diferencias entre los porcentajes de encuestados que dan respuestas favorables y desfavorables a cada pregunta (las respuestas neutras no se toman en cuenta).

Tabla 1: Resumen Cálculo de ICC por FEDESARROLLO

ÍNDICE DE CONFIANZA DEL CONSUMIDOR			
Índice de Condiciones Económicas ICE			
¿Cree usted que a su hogar le está yendo económicamente mejor o peor que hace un año?	(A) Balance = % mejor - % peor Rango: [-100,100]	Balance (A) + Balance (B)/2	Rango: [-100,100]
¿Cree usted que este es un buen momento para comprar cosas grandes como muebles y el ectrodomésticos?	(B) Balance = % bueno - % malo Rango: [-100,100]		
Índice de Ex ectativas del Consumidor IEC			
¿Piensa usted que dentro de un año a su hogar le estará yendo económicamente mejor, peor o lo mismo que ahora?	(C) Balance = % mejor - % peor Rango: [-100,100]		
¿Piensa usted que dentro de los próximos doce meses vamos a tener buenos o malos tiempos económicamente?	(D) Balance = % buenos % malos Rango: [-100,100]	[Balance (C)+ Balance (D)+ Balance (E)]/3	Rango: [-100,100]
¿Cree usted que las condiciones económicas del país estarán mejores o peores dentro de un año que hoy?	(E) Balance = % mejor - % peor Rango: [-100,100]		
ÍNDICE DE CONFIANZA DEL CONSUMIDOR ICC			
ICC = [Balance (A) + Balance (B) + Balance (C)+ Balance (D) + Balance (E)]/5 Rango: [-100,100]		Rango: [-100,100] Periodicidad: Mensual desde noviembre 2001	

Fuente: Elaboración Propia

3.1.2 Modelos de Regresión

Lasso

En presencia de muchas variables explicativas en los modelos de regresión, se es susceptible a que el modelo se ajuste demasiado a los datos de entrenamiento, capturando no solo los patrones subyacentes en los datos sino también el ruido aleatorio o las fluctuaciones. Esto se conoce como sobreajuste y se resuelve a través de métodos de regularización, los cuales consisten en penalizar los coeficientes grandes (reduciéndolos a cero) para minimizar el efecto de variables que no son importantes (Manwani, 2021). Para regularizar los modelos existe el método LASSO (*Least Absolute Shrinkage and Selection Operator*), por sus siglas en inglés. Este método consiste en incluir un nuevo término en la función de pérdida de Mínimos Cuadrados Ordinarios (MCO), en donde se añade el valor absoluto de los coeficientes multiplicado por un valor de

penalización, llamado hiperparámetro alfa o lambda, que es elegido según la necesidad de constreñir los coeficientes (ver ecuación 1) (Kumar, 2023).

Ecuación 1: MCO con regularización Lasso

$$J(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \alpha \sum_{j=1}^{p}\left|\beta_j\right|$$

Fuente: A Complete understanding of LASSO Regression

En concreto, la reducción del coeficiente se puede ver en la ecuación 2, donde se obtiene el coeficiente actualizado a partir del anterior (β_{viejo}) y se le substrae la penalización a través de los hiperparámetros (α y $learning\_rate$). Este proceso es reiterativo hasta que los valores que se obtengan con los nuevos coeficientes converjan para minimizar la función MCO. Este método tiene la ventaja de penalizar los coeficientes llevándolos exactamente a cero y deja los pocos, que son verdaderamente relevantes, como coeficientes no nulos (Rodrigo, 2016).

Ecuación 2: Resumen proceso de Lasso

$$\beta_{nuevo} = \beta_{viejo} - \ learning\_rate(\frac{dJ}{d\beta_{viejo}} + \ \alpha * \ sign(\beta_{viejo}))$$

Fuente: PennState Elberly College Statistical Thinking

Gradient Boosting

Según la publicación *Towards Data Science* de *Medium* el *Gradient Boosting* es un método de aprendizaje automático que combina múltiples modelos de regresión débiles para crear un modelo más fuerte y preciso. El algoritmo consiste en calcular una regresión como base a partir de las cuales se hace la predicción del modelo (F_0). Con esto se calculan los errores de predicción, F_0 (resta entre el valor observado y el predicho por el modelo). Posteriormente, se realiza una regresión que se ajusta a los residuos del modelo base, esto permite mejorar la

explicación del modelo al comprender el comportamiento de estos errores. Con este nuevo modelo se realiza una predicción sobre la muestra, que se le suma a la predicción del modelo base ($F_0(X)$) que es una función acumulativa y, después, se calculan otra vez los residuos de este modelo. Este proceso es iterativo y cada modelo se conoce como árbol de decisión, estas repeticiones lo que hacen es ajustar cada vez mejor el modelo inicial. Al final se hace una ponderación entre todas las predicciones para definir la estimación final del modelo, que tendrá un error en su predicción mucho más pequeña con estos estimadores (Masui, 2022).

Ecuación 3: Función para la predicción de estimadores

$$F(X) = F_0(X) + \alpha_1 h_1(X) + \alpha_2 h_2(X) + \ldots + \alpha_t h_t(X)$$

Fuente: Towards Data Science

ARIMA

El modelo ARIMA, es un modelo de análisis estadístico que emplea datos de series temporales. Este modelo es una forma de análisis de regresión que mide la dependencia existente entre los datos y tiene como objetivo la predicción de valores futuros a partir de valores pasados. Este modelo cuenta con 3 aspectos fundamentales que son los que forman su nombre por sus siglas en inglés: autoregresivo (AR), integrado (I) y promedio móvil (MA).

Este modelo es autoregresivo puesto que se toman en consideración valores pasados o previos para la predicción del valor actual. Integrado, ya que las series de tiempo pueden exhibir una tendencia a lo largo del tiempo, pero calcular la diferencia entre los valores presentes y pasados permite reducirla o eliminarla, en general, consiste en la diferenciación de la serie de tiempo. Y el promedio móvil, hace referencia a que se toman en consideración los errores de predicción anteriores para la predicción del valor actual. Es decir, que se tiene en cuenta la

diferencia entre las predicciones anteriores y los valores reales para mejorar las predicciones futuras.

Ecuación 4: Función para la predicción de estimadores

$$\left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\right)(1 - B)^d X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) Z_t$$

Fuente: Universidad Autónoma de Madrid

SVM

El SVM o Support Vector Machine es un conjunto de algoritmos de aprendizaje supervisado que hace parte del campo del aprendizaje automático y que se emplean principalmente para la clasificación y en menor medida para la regresión. En el ámbito de la clasificación, se pueden encontrar diferentes tipos de SVM según la manera en la que se realiza la separación entre las clases y el manejo de los datos:

- **SVM Kernel Sigmoid**

  El Kernel Sigmoid es uno de los núcleos empleados en las SVM y se emplea en gran medida en las redes neuronales para procesos de clasificación. Estos corresponden a funciones que hacen posible mapear los datos a un espacio dimensionalmente superior, lo que permite a las SVM separar aquellos datos que no son linealmente separables en el espacio original.

Ecuación 5: Función general de Kernel Sidmoid

$$K(x, x_i) = \tanh a\, x_i * x_i + x_j$$

Fuente: Análisis Vidhya

En la función x, x_i representan vectores que corresponden a 2 puntos de datos en el espacio original, x_j es una constante y a es un parámetro que ajusta la escala de la función.

- **SVM Kernel RBF**

  La función de base radial lineal es usada por las SVM para la transformación de datos y para hacer posible la clasificación de patrones que en su forma original no son linealmente separables. El Kernel RBF es útil ya que da flexibilidad para datos complejos al lograr manejar datos que no son linealmente separables, es versátil al acoplarse a diferentes problemas de clasificación y es menos sensible en comparación a otros Kernels a los outliers o datos atípicos.

Ecuación 6: Función Kernel RBF

$$K(x, x_i) = \exp\left(-\gamma * \| \mathrm{x} - x_i \|^2\right)$$

Fuente: Data Flair

En la función x, x_i corresponden a vectores que representan 2 puntos de datos y gamma (γ) es un parámetro positivo que controla la forma e influencia de la función RBF.

3.2 Estado del Arte

Los economistas se han visto enfrentados al reto de encontrar indicadores que revelen información oportuna, ya que gran parte de los datos económicos que son publicados presentan un retraso o dan cuenta de situaciones pasadas (Hellerstein & Middeldorp, 2004). A partir de esto, diferentes investigaciones académicas plantean que es posible hacer la predicción de indicadores económicos, o complementarlos, usando las búsquedas en internet como fuente de información directa y actual. Por ejemplo, se encuentra literatura relacionada a la predicción de

indicadores como el PIB, el consumo privado y el desempleo. Encontramos entre estas, la investigación *Nowcasting de la tasa de desempleo en Colombia, usando información de Google Trends* (Hoyos, 2021) que tiene como objetivo elaborar y validar un modelo de nowcasting para la tasa de desempleo mensual de Colombia, haciendo uso de herramientas como Google Trends para obtener una estimación temprana de un indicador económico.

No obstante, en Colombia, son pocos los estudios que se encuentran sobre nowcasting de indicadores económicos y, en particular, no existe ninguno sobre el índice de confianza del consumidor hasta el momento. Sin embargo, a continuación, se presentan los estudios más relevantes realizados respecto a este indicador en el mundo, que guían y dan sentido al desarrollo del presente trabajo.

En primer lugar, Hal Varian es considerado el pionero en la introducción de los datos de Google para la predicción de indicadores económicos y la difusión del término nowcasting como herramienta para predecir un panorama de la economía o tendencias de forma casi inmediata a través de los términos de búsqueda de Google Trends (Varian, 2009). Lo anterior sustentado en una fuerte correlación que encontró, entre algunas búsquedas en Google y los indicadores económicos más importantes. Sin embargo, él señala que como herramienta para la predicción de indicadores económicos, Google Trends, tiene algunas limitaciones y no siempre puede llevar a resultados exactos, pues la búsqueda de algunos términos puede depender de factores como: noticias del momento, tendencias, modas, entre otros, que pueden entorpecer la interpretación de los datos y las predicciones.

En la Investigación *Incorporating Google Trends Data in Predicting Consumer Confidence in Sri Lanka* de 2017 (Jayathilake, Wattegama, Gamage, & Dassanayak, 2017) hacen una predicción del ICC. En el artículo se logró determinar, a través de técnicas estadísticas, que

había una fuerte correlación entre las búsquedas de los consumidores en esta red y la confianza del consumidor de ese país. Debido a esta relación, construyeron un modelo de regresión lineal múltiple para predecir la confianza del consumidor empleando los datos de Google Trends. Con los resultados determinaron que el uso de Google Trends es beneficioso, por factores como: acceso a datos en tiempo real, mejora en la precisión en las predicciones, ahorro de tiempo y costos, y eliminación del retraso por la espera de la publicación de los datos de confianza del consumidor por parte de las empresas privadas.

En el mismo sentido, la publicación investigativa *Nowcasting with Google Trends in an emerging market* (Carrière-Swallow & Labbé, 2010) , expone cómo emplear las búsquedas de palabras claves en Google Trends para hacer una predicción del comportamiento del mercado de valores y de algunos indicadores económicos en Chila. Cabe resaltar que el artículo menciona que puede existir una limitación de precisión al usar los términos de Google Trends, no obstante, este instrumento es muy útil para analistas económicos que requieren precisamente datos actualizados, pues ayuda a tomar mejores decisiones. La metodología empleada consistía en la selección de términos de búsqueda en Google que tuvieran una correlación con la actividad económica en el mercado de interés.

Después, por medio de un modelo estadístico se analizó esta relación para obtener la predicción del comportamiento de los futuros indicadores económicos.

Los resultados obtenidos al combinar el GTAI (Google Trends Automotive Index) con modelos de nowcasting fueron superiores a especificaciones de referencia, es decir, que la incorporación de los datos obtenidos a partir de Google Trends mejoraron la precisión de los modelos de nowcasting. Estos datos son llamativos al ser derivados directamente de datos de usuario, abarcan

un gran número de usuarios de internet y se obtienen con una frecuencia alta en intervalos regulares.

De igual manera, en el artículo *Determinants of Consumer Confidence Index to Predict the Economy in Indonesia* (Tjandrasa & Budiawan, 2022), se resalta cómo se puede predecir la economía o el crecimiento económico del país empleando el ICC en Indonesia como indicador. Además, muestra que Google Trends permite medir la intención de los consumidores en relación con la economía, lo cual puede ayudar a predecir la demanda futura de bienes y servicios. Para esto, emplearon un modelo de regresión multivariable y la prueba t con un nivel de significancia del 5%. Los resultados de esta investigación dan cuenta entonces de que el ICC de este país está determinado por aspectos como la tasa de inflación, la tasa de desempleo, el tipo de cambio y el control de la corrupción, ya que esto forma una percepción de las condiciones futuras, lo cual afecta el nivel de confianza de las personas.

En 2010, Haifang Huang y Nicolas Della Penna en su artículo *"Constructing Consumer Sentiment Index for U.S. Using Google Searches*" argumentan que Google Trends es una herramienta que es útil para la construcción del Índice de sentimiento del consumidor en Estados Unidos, pues lo que las personas consultan en internet da cuenta de sus intereses, preocupaciones, necesidades y qué puede permitir de igual forma observar el cambio del comportamiento de los consumidores, y, en consecuencia, de indicadores económicos importantes. Para la construcción del ICC con la inclusión de Google Trends, se identificaron palabras clave que tuvieran correlación con la confianza del consumidor como bancarrota, muebles de oficina, artículos de lujo, energía y servicios públicos, vehículos híbridos y alternativos. Con esto se elaboró un modelo estadístico con el cual se predijo los valores del ICC. Finalmente, para validar el índice, se compararon las predicciones con las mediciones reales de la confianza del consumidor,

encontrando que google trends es útil para la construcción de un modelo que permita hacer predicciones sobre la confianza del consumidor.

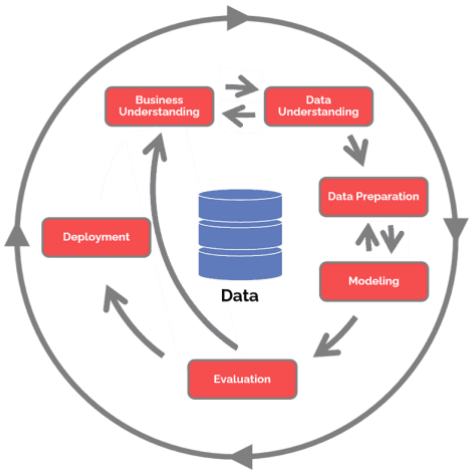
Además, en *Forecasting private consumption with GT data* (Woo & Owen, 2019) se presenta la forma en la que se puede usar la información obtenida con Google Trends para la predicción del consumo privado en la economía, de gran relevancia, al ser un indicador del crecimiento de la economía. Para el desarrollo de esta investigación, se empleó la técnica de análisis de regresión, para estudiar la relación entre los datos ofrecidos por esta herramienta y el consumo privado en una serie de países. Como resultado de esto, se obtuvo que la información de Google Trends es una forma de predicción significativa del consumo privado en la mayoría de los países. Además, se llegó a la conclusión de que al usar los datos de Google Trends se mejoraron las predicciones.

Finalmente, en *Google Search Trends Show How Customer Expectations Are Changing* (Southern, 2022), expone como la información que proporciona Google Trends puede ser empleada para tener una mejor comprensión de cómo evolucionan las expectativas de los consumidores con el tiempo y cómo estos responden a los cambios en la economía y en el mundo. Como ejemplo, propone los términos de búsquedas relacionadas con seguridad laboral y disponibilidad de empleo; los cuales pueden representar un indicador de los cambios en la confianza de los consumidores y con esto sus patrones de consumo. También, se menciona que la información de Google Trends sirve para la predicción de tendencias futuras en el comportamiento de los consumidores. Lo anterior, demuestra su gran utilidad para responsables de la política económica, pues, por ejemplo, permitiría entender mejor la evolución de las expectativas de los consumidores y así se podrían formular políticas que realmente apoyen al crecimiento y estabilidad de la economía.

4. METODOLOGÍA

Con el propósito de cumplir el objetivo de esta investigación, se hará uso de la metodología CRISP – DM (Cross Industry Standard Process for Data Mining). Esta es utilizada en el área de minería de datos y machine learning, y se compone de 6 fases que se muestran en la siguiente ilustración:

Ilustración 1: Ciclo de vida de minería de datos (CRISP-DM)



Fuente: Data Science Process Alliance

A continuación, se describirán cada una de las fases y las actividades que corresponden a cada una de estas. No obstante, es importante resaltar que en esta investigación no se desarrollará la fase de implementación o despliegue.

Fase I: Comprensión del negocio

En esta fase inicial se identifica el indicador económico de interés para el forecasting, su interpretación y los posibles factores que lo alteran.

Fase II: Comprensión de los datos

En este punto, se determinan los datos que son necesarios para el desarrollo del modelo. Además, se pretende familiarizarse con estos datos, observando la frecuencia con la que se obtienen, el

rezago que presentan, en qué medida los datos representan la realidad, entre otros. En general, esta fase se ocupa de la recolección, descripción, exploración y verificación de la calidad de los datos.

Fase III: Preparación de los datos

Para la preparación, se realiza la selección y limpieza de los datos que serán empleados. De igual forma, en esta fase se construyen algunos datos y se definen los periodos que se utilizarán para el entrenamiento y la evaluación.

Fase IV: Modelamiento

La fase de modelamiento consiste en la selección y estimación de modelos para realizar las predicciones que permitirán realizar el forecasting del indicador a estudiar. Adicionalmente, en esta fase se estiman los parámetros más apropiados para cada uno de los modelos.

Fase V: Evaluación

Por último, la fase de evaluación consiste en la selección del mejor modelo teniendo en cuenta las métricas definidas para hacer este análisis. Además, para esta fase se tiene en cuenta el contexto económico para el estudio de los resultados.

4.1 Presentación de la propuesta

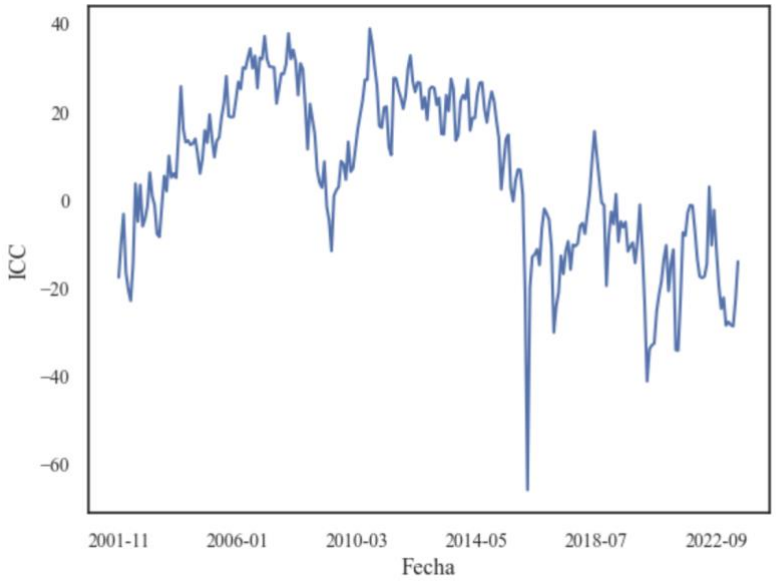
Datos y Software

Para la construcción de la base de datos se usaron como insumo dos fuentes principales de datos. La primera, consistió en descargar variables económicas cuantitativas que explicaran, según las fuentes revisadas en la bibliografía, el comportamiento del ICC. Dentro de esta primera clasificación se recuperaron: la tasa de desempleo y la inflación del DANE, la tasa representativa del mercado (TRM) del Banco de la República, la tasa de crédito al consumo de la Superfinanciera y el ICC de FEDESARROLLO (que es la variable por explicar). En segunda

instancia, se recuperaron de Google Trends los términos de búsqueda que se encontraron relevantes para predecir el comportamiento del índice, según investigaciones previas.

Los datos se tomaron de diferentes fuentes desde diciembre de 2017 hasta junio de 2023. Todos los datos están en una periodicidad mensual, excepto la TRM que es publicada diariamente, para lo cual se calculó la media de cada mes y esa fue la observación registrada en la base de datos. Además, para incluir el efecto autorregresivo a los modelos se rezagaron todas las variables explicativas en tres períodos. En la siguiente ilustración se mostrará el registro histórico del ICC, para referirse a registros de las demás variables ver Anexos.

Ilustración 2: Evolución del ICC desde su primera publicación

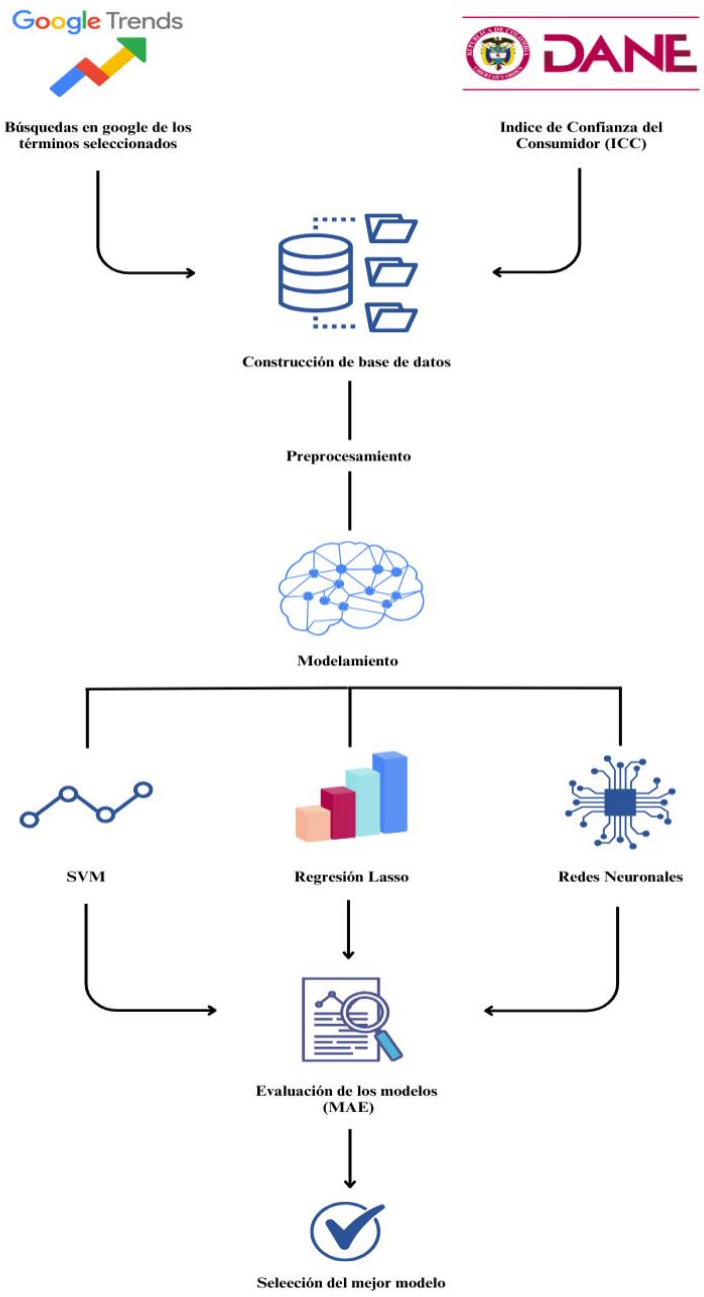


Fuente: Elaboración propia con datos de FEDESARROLLO

En cuanto a la construcción de los modelos se realizaron dos tipos de preprocesamiento: estandarización y análisis de componentes principales (PAC, por sus siglas en inglés). Por último, todo este análisis se llevó a cabo a través de dos herramientas para el Data Analysis, el software

estadístico R y el principal usado en esta investigación, el lenguaje de programación, Python. A continuación, se presentará un gráfico que resume en la Ilustración 2.

Ilustración 3: Recuento de la propuesta de para la construcción del modelo



Fuente: Nowcasting de la tasa de desempleo en Colombia, usando información de Google Trends

Modelamiento

Después del procesamiento de los datos, se realizaron 3 aproximaciones a través de modelos de regresión LASSO, Support Vector Machine (SVM) y redes neuronales. Todos estos modelos se pueden usar para realizar regresiones, no obstantes no están especificados para modelar series de tiempo tradicionales. Por lo tanto, en el trabajo se toman los datos como de corte transversal, para lo cual se incluye, tal como se indicó anteriormente, el rezago de las variables explicativas. Pues, para poder hacer la predicción debemos usar las variables independientes del mes anterior para pronosticar el valor del ICC del mes siguiente.

Evaluación

Con el fin de comprobar el poder de predicción de los modelos calculados se dividió en dos partes, muestra de entrenamiento y de evaluación, con el fin de poder analizar el poder predictivo del modelo construido. La muestra de entrenamiento recogió los datos hasta diciembre de 2022 y, en consecuencia, la de entrenamiento, recoge los datos correspondientes al año 2023. Además, sobre la muestra de entrenamiento se realizó una validación cruzada con 4 *folds*, lo que ayuda a obtener una estimación más exacta del rendimiento del modelo, sin importar que la muestra de datos sea limitada. Este método, también favorece a que se seleccionen los mejores parámetros que predicen cada modelo construido, lo que fortalece su capacidad de predicción.

El criterio de decisión para elegir entre los diferentes modelos construidos fue la métrica de la media de suma de los errores, MAE (*Mean Absolute Error*), por sus siglas en inglés. La fórmula para calcular el MAE es:

Ecuación 1: Fórmula para calcular el MAE

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Fuente: Elaboración Propia

Donde $\boldsymbol{y_i}$ son los datos observados en la muestra de evaluación y $\hat{y}_i$ son las predicciones arrojadas por los modelos. Esta medida permite determinar el porcentaje, en promedio, del error absoluto de las predicciones que arroja el modelo. Después de calcular todos los modelos, se seleccionó el que minimizara esta métrica. En este caso el modelo seleccionado para predecir el ICC, fue el SVM Kernel Sigmoid, pues es el que al final tuvo mejor desempeño en comparación con el modelo ARIMA.

4.2 Diseño del Modelo y Experimentación

Con el propósito de validar la propuesta descrita en la sección previa, se consideró algo fundamental y es la aplicación en la práctica. En este caso, la forma de verificar esto, es revisar si la propuesta permitía predecir los datos de forma anticipada a las publicaciones del ICC del mes correspondiente, lo cual se cumplió y fue desarrollado por medio de un modelo auto regresivo tradicional (ARIMA). Además, se calculó el MAE sobre la muestra de evaluación y se comparó con el modelo propuesto, lo cual hizo posible establecer si la aproximación es más apropiada que el cálculo tradicional para el ICC en Colombia.

5. DISCUSIÓN DE RESULTADOS

Después de ejecutar los modelos se encontró que el modelo que mejor logra predecir el comportamiento del ICC es el SVM Kernel Sigmoid. Pues, este modelo logra anticiparse mejor que el baseline que se tomó como el modelo ARIMA. Los resultados de los cuatro modelos se pueden resumir en la tabla 2, donde se encontró que el ajuste del modelo sigmoide es el que minimiza el error de predicción. En más detalle, este modelo quedó determinado por 9 componentes que lograron capturar y explicar mejor el comportamiento de los datos, a partir del PCA. Además, el nivel de regularización que mejoró el problema de sobreajuste fue de 5. El resto de los modelos tienen un peor desempeño que el ARIMA, por lo que no pueden tenerse en cuenta como forma para anticiparse al comportamiento del ICC de forma acertada, en la tabla 2 están organizados del que mejor se comportó al peor.

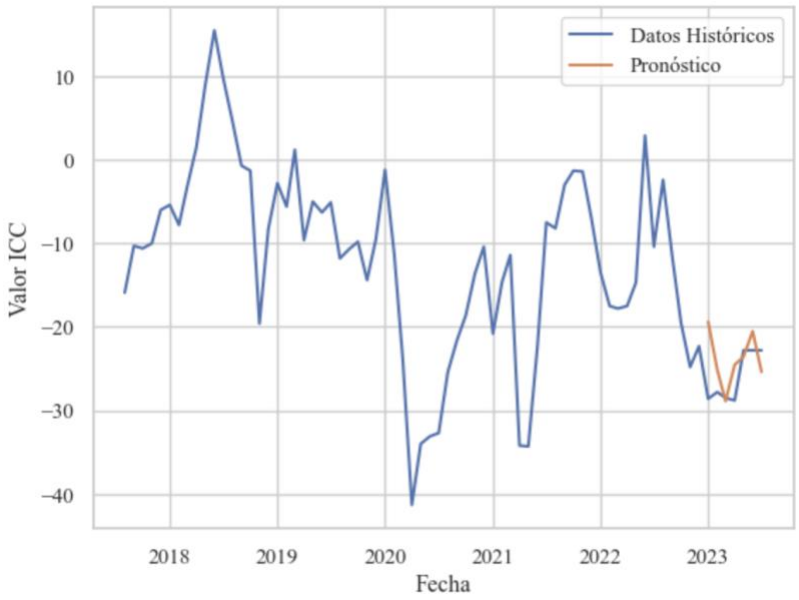
Tabla 2: Resumen de los resultados de los modelos

Modelo	MAE	RMSE	Especificaciones
SVM Kernel Sigmoid	3,16	4,18	Parámetro: C = 5 Preprocesamiento: Estandarización Gamma = 0,01
Baseline ARIMA	5,17	6,93	
Gradient Boosting	5,89	7,48	Hiperparámetro de máxima profundidad: 3
LASSO	6,57	7,26	Preprocesamiento: Estandarización ALFA = 1,9
SVM Kernel RBF	12,12	12,69	Parámetro: C = 10 Preprocesamiento: Estandarización Gamma = 0,09

Fuente: Elaboración Propia

En la imagen 3 se resume el resultado del modelamiento de los datos con SVM Kernel Sigmoid junto con los datos históricos del ICC. Se evidencia que aunque no hay un ajuste perfecto de los datos, el modelo sí logra comportarse en el mismo sentido del ICC y hace aproximaciones cercanas a los valores reales.

Ilustración 4: Gráfico Evolución del ICC y Pronóstico por SVM



Fuente: Elaboración Propia

6. CONCLUSIONES

El modelo final obtenido a través de SVM Sigmoid permitió hacer la mejor aproximación al ICC, para lo que fue clave la inclusión de los términos de búsqueda de Google Trends, pues, fue lo que ayudó a mejorar el poder predictivo del modelo. Esto significa un avance para el campo económico en relación con la oportunidad en la toma de decisiones informadas, pues se puede hacer uso de las predicciones del modelo sin que se tenga que esperar o usar información rezagada no tan precisa a la realidad.

La novedad del modelo se encuentra en la identificación de las búsquedas de Google Trends que complementaron el poder predictivo con las variables cuantitativas convencionales, pues son estos términos y su actualización constante los que permiten obtener información precisa para realizar el *nowcasting.* Lo anterior se corroboró a través del contraste hecho con las métricas de evaluación del MAE y RMSE, donde se demostró la superioridad predictiva sobre procesos autorregresivos y otras aproximaciones de modelamiento explicadas anteriormente.

Finalmente, este trabajo puede usarse en el futuro como ejemplo para la predicción de otros indicadores económicos con la inclusión de la información suministrada por Google Trends. No obstante, siempre debe verificarse la relación entre estos datos y la posibilidad de predicción, pues puede no ser generalizable para otro tipo de indicadores. En relación con el modelamiento del ICC, se pueden usar aproximaciones más complejas y herramientas de Aprendizaje Automático para refinar aún más la predicción, para lo cual este trabajo puede funcionar como base y adelanto en identificación de términos relevantes.

7. BIBLIOGRAFÍA

Bánbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). *Now-Casting and the Real-Time Data Flow*. Obtenido de European Central Bank - Eurosystem: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf

Carrière-Swallow, Y., & Labbé, F. (2010). *Nowcasting with Google Trends in an emerging market.* Banco Central de Chile.

Edward, G., & Box, G. M. P. (s/f). SERIES TEMPORALES, MODELO ARIMA METODOLOGÍA DE BOX - JENKINS. Estadistica.net. Recuperado el 27 de agosto de 2023, de https://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf

Hayes, A. (2010, abril 11). *Autoregressive integrated moving average (ARIMA) prediction model*. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Hoyos, C. (2021). *Nowcasting de la tasa de desempleo en Colombia, usando información de Google Trends.* Cali.

Jayathilake, Wattegama, Gamage, & Dassanayak. (2017). Incorporating Google Trends Data in Predicting Consumer Confidence in Sri Lanka. Sri Lanka.

Keynes, J. (1936). *Teoría General del Interés y el Empleo.*

Kumar, D. (05 de 2023). *A Complete understanding of LASSO Regression.* Recuperado el 08 de 2023, de Great Learning: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20uses%20shrinkage%2C%20where,zero%20or%20near%2Dzero%20coefficient.

Manwani, R. (15 de 09 de 2021). *Lasso and Ridge Regularization – A Rescuer From Overfitting.* Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/09/lasso-and-ridge-regularization-a-rescuer-from-overfitting/

Masui, T. (10 de 01 de 2022). *All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression.* Recuperado el 08 de 2023, de Toward Data Science: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Pickering, J. F., Harrison, J. A., & Cohen, C. D. (1973). Identification and Measurement of Consumer Confidence: Methodology and Some Preliminary Results. *Journal of the Royal Society, 136*(1), 43-63.

Rodrigo, J. A. (12 de 2016). *elección de predictores, regularización ridge, lasso, elastic net y reducción de dimensionalidad.* Recuperado el 08 de 2023, de Ciencia de Datos: https://cienciadedatos.net/documentos/31_seleccion_de_predictores_subset_selection_ridge_lasso_dimension_reduction#Regularización_(Shrinkage)

Southern, M. (Marzo de 2022). *Google Search Trends Show How Customer Expectations Are Changing.* Obtenido de Search Engine Journal: https://www.searchenginejournal.com/google-search-trends-show-how-customer-expectations-are-changing/440522/#close

Tjandrasa, & Budiawan, B. (2022). Determinants of Consumer Confidence Index to Predict the Economy in Indonesia. Indonesia.

Varian, H. (2009). *New York Times.* Obtenido de How Google Flattens Information?

Woo, J., & Owen, A. (2019). *Forecasting private consumption with GT data.* Obtenido de Wiley Online Library: https://doi.org/10.1002/for.2559