Nowcasting del Turismo y la Ocupación Hotelera en Colombia Usando Tendencias de Google

Juan Felipe Gómez Borja

Valeria Ospina Del Castillo

Universidad Icesi

Economía y Negocios Internacionales

Julio Cesar Alonso Cifuentes

Santiago de Cali

Noviembre de 2023

Nowcasting del Turismo y la Ocupación Hotelera en Colombia Usando Tendencias de Google

Juan Felipe Gómez Borja

Valeria Ospina Del Castillo

Julio Cesar Alonso Cifuentes

Facultad de Ciencias Administrativas y Económicas

Economía y Negocios Internacionales

Santiago de Cali

2023

Tabla de Contenido

Pág.

LISTA DE TABLAS

LISTA DE FIGURAS

Resumen

La información económica es la base para la elaboración y seguimiento de los objetivos de numerosas empresas que buscan mejorar su estructura interna con el propósito de atraer a un público más amplio. Con dichos fines, las compañías se apoyan en reportes mensuales que les ayuden a anticiparse al futuro, presentados por las instituciones económicas y financieras. No obstante, esta información se encuentra desactualizada frente a su fecha de publicación, lo que es un problema para los agentes del mercado al ser información obsoleta.

Este documento tiene el propósito encontrar una alternativa a la búsqueda de información sobre el comportamiento de un indicador en tiempo real, centrándose en el sector turismo. De esta manera y mediante el uso de Google Trends, se pretende encontrar resultados que ayuden a las empresas turísticas a tomar decisiones estratégicas con información reciente y futura del mercado.

1.1. Palabras Claves

Google Trends, Nowcasting, Ocupación Hotelera, Turismo

Abstract

Economic information is the basis for the development and monitoring of the objectives of many companies seeking to improve their internal structure to attract a wider public. To this end, companies rely on monthly reports presented by economic and financial institutions to help them anticipate the future. However, this information is outdated with respect to its publication date, which is a problem for market agents as it is obsolete information.

The purpose of this document is to find an alternative to the search for information on the behavior of an indicator in real time, focusing on the tourism sector. In this way, using Google

Trends, it is intended to find results that help tourism companies to make strategic decisions with recent and future market information.

1.2. Key Words

Google Trends, Nowcasting, Hotel Occupancy, Tourism

2. Introducción

2.1. Justificación

Durante los últimos años el turismo se ha convertido en un pilar importante para la reactivación de la economía colombiana frente a la problemática del Covid-19. Según el Ministerio de Comercio, Industria y Turismo, en 2021, este sector generó 3.101 millones de dólares en divisas, esto es un aumento de 59.5% en comparación al 2020. En el segundo trimestre de 2022 el producto interno bruto (PIB) de la rama de alojamiento y servicios de comida se destacó por su comportamiento durante este periodo, representando el 4.67% del PIB en el siguiente trimestre según el DANE. De esta forma, las empresas encargadas de diseñar políticas públicas de turismo requieren de documentos guía que les permitan crear estrategias orientadas a fomentar las diferentes actividades turísticas del país.

En Colombia, las empresas del sector turismo emplean los reportes del DANE, específicamente la Encuesta Mensual de Alojamiento (EMA) con el objetivo de crear estrategias de promoción que atraigan a los clientes a consumir sus productos. Sin embargo, esta información presenta un rezago de un mes aproximadamente, lo que significa que a agosto de 2023 sólo se cuenta con información desfasada de junio de 2023, de manera que no se pueden tomar decisiones inmediatas con la información que se encuentra disponible. Esto puede generar que las organizaciones escojan un curso de acción erróneo basándose en información desactualizada.

Por esta razón, es necesario el desarrollo de indicadores adelantados que permitan a las empresas predecir el comportamiento de las personas en el futuro cercano. Estos resultados permiten reducir la incertidumbre sobre escenarios en diferentes industrias. Durante los últimos años, las compañías han implementado la herramienta del nowcasting, modelo que ha adquirido

gran relevancia debido a que permite realizar diagnósticos de distintas variables económicas combinando el presente y el futuro. Gracias a este modelo, las empresas trabajan con un margen de error menor, que permite garantizar una probabilidad de éxito más alta en sus nuevas estrategias de mercado, contrario a lo que obtendrían si desarrollaran estrategias a ciegas.

Adicionalmente, nuevas fuentes de información permiten alimentar el desarrollo del nowcasting. En este caso, los datos de búsqueda obtenidos de Google Trends han demostrado contribuir de manera positiva en la predicción del nowcast de distintos indicadores económicos. Se observan casos como: el desempleo en Estados Unidos (Nagao et al., 2019), el sector automotriz en economías emergentes (Carrière-Swallow et al., 2013), y el PIB en Alemania y Finlandia (Heikkinen, 2019).

2.2. Planteamiento del Problema

Debido a los acontecimientos de los últimos años, se ha convertido en una necesidad para todos los agentes económicos contar con información actualizada acerca de variables que brinden información relevante para sus procesos internos. A raíz de esto, se hace indispensable el desarrollo de una herramienta que, contrario a los indicadores existentes —los cuales requieren de un tiempo específico para presentar al público— proporcione la información necesaria para desarrollar estrategias que respondan a los nuevos fenómenos.

En este caso, la publicación de la variable del EMA, ocupación hotelera mensual, tiene un rezago de aproximadamente un mes, lo que significa que las empresas que requieran realizar cambios en sus procedimientos deberán esperar aproximadamente un mes para llevar a cabo dichas estrategias. Por ende, las decisiones tomadas por compañías del sector turismo nacen bajo incertidumbre, lo que puede causar que se creen expectativas que no se puedan cumplir con los datos disponibles.

2.3. Objetivo General

Realizar un nowcasting de la ocupación hotelera usando información de las tendencias de Google.

2.4. Objetivos Específicos

- Identificar la capacidad de Google Trends como insumo para el pronóstico de indicadores económicos.
- Evaluar diferentes aproximaciones para pronosticar la ocupación hotelera mensual en Colombia.
- Validar la calidad de la estimación usando modelos de regresión no convencionales.

3. Antecedentes

3.1. Marco Teórico

En este apartado se definen los conceptos económicos en los que se apoya el proyecto. De manera que supone un soporte técnico para lo abarcado/desarrollado a lo largo del documento.

3.1.1. Conceptos económicos

Nowcasting

"El nowcasting consiste en llevar a cabo una serie de acciones para predecir el presente, el pasado reciente y el futuro cercano de una serie temporal" (Cabria, 2023, p.13). Este modelo permite estimar variables económicas a partir de variables con menor periodicidad a la que se busca evaluar.

Indicador económico

Es un dato económico estadístico que permite analizar el entorno económico de un territorio. Spinak (2001) lo define como medidas que permiten capturar la evolución y coyuntura económica, con base en un conjunto de datos que se obtienen en un tiempo determinado. Gracias a este, se pueden realizar predicciones sobre posibles fenómenos que puede experimentar la economía.

Ocupación hotelera

El grado de ocupación hotelera es un indicador del sector turismo que brinda información a las empresas a partir del porcentaje de habitaciones ocupadas con relación al total de habitaciones disponibles en un alojamiento. De esta variable dependen las estrategias de precio y promoción de las organizaciones.

3.1.2. Concepto de datos y modelos machine learning

Google Trends

Google Trends es una herramienta desarrollada por Google con el objetivo de brindar información al usuario acerca de sus temas de interés. Para esto, realiza seguimiento de cualquier temática a partir de una muestra de las búsquedas realizadas en la página de Google.

Machine Learning

Los modelos de machine learning son algoritmos que permiten capturar automáticamente relaciones complejas en los datos que los modelos econométricos tradicionales no logran representar. A diferencia de los métodos econométricos tradicionales que se basan en especificar relaciones estructurales entre variables, el machine learning busca encontrar patrones predictivos en los datos sin imponer restricciones a priori sobre la forma funcional.

Overfitting

El overfitting o sobreajuste es un problema que puede ocurrir durante el entrenamiento de modelos de machine learning. Se da cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando tanto el patrón real como el ruido aleatorio.

Esto puede suceder cuando el modelo es muy complejo (muchos parámetros) con relación al tamaño de datos disponibles. Así, el modelo termina modelando perfectamente el conjunto de entrenamiento, pero no es capaz de generalizar bien a nuevos datos.

Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Vectores de Soporte (SVM por sus siglas en inglés), son un modelo de regresión que busca encontrar el hiperplano óptimo que separe los datos en categorías, de manera

que actúa como un clasificador discriminativo de los datos. Utiliza kernels —funciones que permiten transformar los datos de entrada a un espacio dimensional diferente— para transformar los datos a un espacio de dimensión mayor donde sea más fácil separarlos linealmente. Luego, minimiza el error permitiendo algunos puntos fuera del margen para evitar overfitting. Es útil cuando hay muchas variables predictoras y se quiere identificar las más relevantes.

Regresión Lasso

La regresión Lasso penaliza la suma de los valores absolutos de los coeficientes, reduciendo algunos de estos coeficientes a cero, eliminándolos efectivamente del modelo. De este modo, la regresión Lasso es una medida práctica para la selección de variables, debido a que elimina predictores irrelevantes, y penaliza la magnitud de los coeficientes con el fin de reducir el overfitting.

Boosting Regressor

Este modelo combina regresores débiles (como árboles de decisión simples) para producir un regresor fuerte. Cada regresor se entrena en los residuos del anterior para concentrarse en los casos difíciles, reduciendo el sesgo y overfitting. Así, al combinar modelos simples, Boosting es rápido y fácil de entrenar incluso con muchas variables.

Random Forest

El Random Forest construye diversos árboles de decisión en bootstrapped samples de los datos; para después promediar las predicciones que permiten reducir la varianza y overfitting. Al usar subconjuntos aleatorios de variables en cada árbol, identifica variables importantes. Por esta razón, el Random Forest es práctico cuando se trabaja con muchas variables, ya que no se sobreajusta, es fácil de entrenar y paralelizar.

Auto-Regressive Integrated Moving Average (ARIMA)

El modelo ARIMA es una clase popular de modelos para series de tiempo que captura tres tipos de efectos: autorregresivos (AR), integrativos (I) y de medias móviles (MA). Los componentes autorregresivos modelan la dependencia de una observación en observaciones previas de la serie. Los componentes integrativos eliminan tendencias no estacionarias mediante differencing (diferenciación que se aplica a la serie de tiempo para hacerla estacionaria). Los componentes de medias móviles modelan la dependencia entre una observación y los residuos de observaciones previas.

El modelo lleva a cabo un proceso de ajuste, donde se identifica y estima el orden óptimo de los componentes AR, I y MA. Una vez ajustado, el modelo puede usarse para predecir valores futuros de la serie de tiempo. Así, al combinar estas piezas, el modelo ARIMA puede modelar series temporales no estacionarias que exhiben patrones como tendencia, estacionalidad, autocorrelación, entre otros.

3.2. Revisiones

3.2.1. Revisión internacional

Una de las investigaciones de nowcasting llevadas a cabo para el sector turismo y usando la herramienta de Google Trends, fue desarrollada por Rivera (2016), donde el autor busca la posibilidad de explotar los sistemas de búsqueda como Google Trends para modelar la demanda de turismo. Esta investigación fue llevada a cabo para Puerto Rico y se concluyó causalidad entre las variables estimadas y las del motor de búsqueda de Google.

Otra investigación pertinente se trata de un estudio de caso propuesto por Blanco (2020), aplicado a la Costa del Sol en el contexto de la crisis global provocada por la pandemia de

coronavirus en la primera mitad de 2020. El autor propone la utilización de herramientas tecnológicas, en concreto Google Trends, para predecir la predisposición a viajar de los mercados emisores hacia la Costa del Sol. Se concluye que Google Trends es válido para realizar aproximaciones generales de la demanda turística, aunque su fiabilidad aumenta si se complementa con otras herramientas y fuentes de información. El estudio indica también que la demanda turística se desplomó a niveles mínimos históricos debido a la pandemia, pero que la Costa del Sol ha mantenido un interés suficiente por parte de sus principales mercados emisores, demostrando así una alta fidelidad.

3.2.2. Revisión nacional

Para el caso colombiano, existen estudios de herramientas como Google Trends usadas para el sector turismo a través del nowcasting, uno de ellos es el artículo de Correa, por medio del cual se determinó que la viabilidad del motor de búsqueda es un camino para pronosticar un indicador de turismo en el país. “Los resultados evidencian que la industria del turismo y las autoridades encargadas de la política pública de turismo se pueden beneficiar en utilizar los datos de Google Trends” (Correa, 2021).

Otra propuesta para la región nacional es un estudio llevado por Ospina A. C. & Caballero L. (2019), que analiza la capacidad de los datos de Google Trends para predecir la llegada de turistas a Colombia. Utilizando la metodología MIDAS, que permite estimar modelos con variables de diferente frecuencia, se construye un indicador a partir de búsquedas en Google para Estados Unidos, Canadá y Reino Unido. Los resultados sugieren que la información de Google Trends ofrece beneficios significativos para pronosticar la llegada de turistas, superando los modelos sin esta información. Se concluye que Google Trends tiene potencial como herramienta para mejorar la evaluación y pronóstico del turismo en Colombia.

4. Metodología

4.1. Selección del indicador

Para este trabajo, se utilizaron los reportes expedidos por el Departamento Administrativo Nacional de Estadística (DANE) sobre la encuesta mensual de alojamiento. El DANE es la entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia. Esta institución se encarga mensualmente de realizar la encuesta EMA, donde se obtiene información de los establecimientos que prestan servicio de alojamiento a nivel nacional y regional a través de índices y variaciones e indicadores del sector turismo. Anteriormente, este indicador se denominaba muestra mensual de hoteles (MMH), pero a partir de los resultados de junio de 2020, debido a un cambio en su estructura, metodología, y mejora en la muestra, pasa a nombrase como actualmente se conoce, EMA. Este cambio corresponde a una nueva unidad estadística de operación, donde en lugar de recopilar información de las empresas que ofrecen servicios de alojamiento, se tienen en cuenta todos los establecimientos que cumplen con este mismo objetivo.

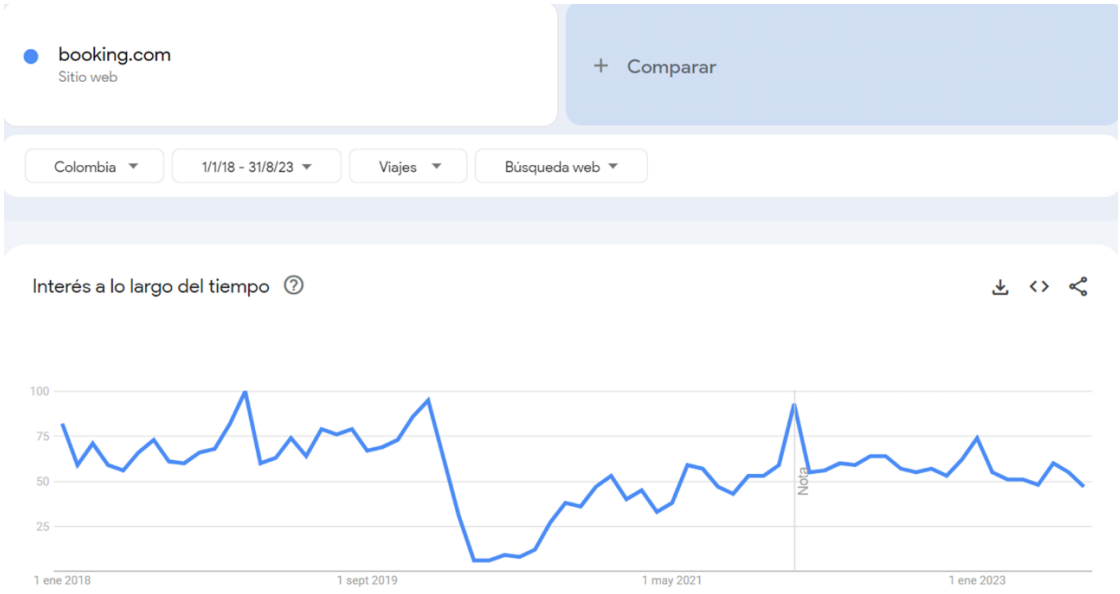
Por ello, y considerando hacer un pronóstico para el turismo en Colombia, se decide usar este indicador por su disponibilidad y competencia actual, debido a la necesidad que surge de encontrar una variable que se actualizara constantemente en las estadísticas del DANE, de modo que fuera aplicable para el desarrollo de este ejercicio, y tuviera relevancia en el sector real del turismo colombiano. Adicionalmente, se eligió dentro del EMA, la variable “ocupación hotelera mensual” que, por criterio del investigador, es la variable que más significancia mantiene con el indicador económico.

4.2. Recolección de datos y criterios

Según StatCounter (2020), Google ostenta una cuota de mercado global del 92.71% en el ámbito de los motores de búsqueda. El término "Googlear" ha sido añadido al diccionario coloquial del ciudadano mundial, resaltando la importancia diaria que las personas le otorgan a Google como plataforma de información. Referente al sector turismo, se habla de una industria masiva en información, por lo cual Google alberga una gran cantidad de datos relacionados con este sector; particularmente, Google Trends que permite analizar cómo la fama de diferentes términos de búsqueda ha variado a lo largo del tiempo.

Figura 1

Ejemplo de criterio de búsqueda de Google Trends



Fuente: Google Trends.

Como se puede observar en la **Figura 1**, Google Trends construye sus datos de acuerdo con una categoría de búsqueda (por ejemplo, "Booking.com"), una región geográfica o país específico (por ejemplo, "Colombia"), una categoría de interés (por ejemplo, "Viajes") y un

periodo de tiempo determinado. Después de ello, los valores arrojados por la plataforma se elaboran al calcular la proporción de búsquedas —es decir, se toma en cuenta el número de búsquedas relacionadas con "GoogleVuelos" en comparación con todas las demás búsquedas realizadas en la región geográfica específica en la categoría objetivo— y el mayor grado de búsquedas diarias en un período de tiempo determinado se estandariza en 100 puntos.

Por ello, y considerando a Rivera (2016) y Cabria (2023), junto con la plataforma turística TripAdvisor, se decidieron los criterios de búsqueda en Google Trends más pertinentes para la construcción de la base de datos. A su vez, se evaluaron cuáles indicadores macroeconómicos tenían un efecto directo en el sector turismo. Así, se encontró que la tasa representativa del mercado (TRM), la inflación y el desempleo son los factores macroeconómicos con mayor impacto en el turismo (Velázquez et al., 2022). Respecto a la tasa representativa del mercado, Muhammad et al. (2018), afirma que un incremento en la TRM tiene una connotación positiva entre en la depreciación de la moneda del país anfitrión y la demanda turística. Mientras que la inflación, según Tang y Lean (2007), un incremento en este indicador trae como consecuencia un aumento en el costo de vida, que conlleva a una disminución en el poder de compra y, a su vez, afecta negativamente el flujo de turistas. Asimismo, el desempleo y su estrecha relación con el bienestar de los individuos, ocasiona que altos niveles del indicador limiten la capacidad de las personas para generar un ingreso que permita financiar su consumo y generar ahorro (Eslava y Fernández, 2022). Por ende, los niveles de desempleo se encuentran directamente relacionados con la toma de decisión de las personas acerca del gasto y tiempo que dedican a sus actividades de ocio, entre las que se pueden encontrar viajes por dentro o fuera de su país. Estos tres factores influyen directamente en el comportamiento del consumidor, debido a que determinan el atractivo de un país, y el poder de compra de los individuos.

4.3. Base de datos

La base de datos está conformada por 20 variables, entre las cuales se encuentran: tres variables macroeconómicas, dieciséis variables de Google trends, y una variable objetivo. Dentro del primer grupo se encuentran: la tasa de desempleo (%), la inflación (%), y la tasa representativa del mercado (USD/COP). El segundo grupo está compuesto por: agencias de viajes (google vuelos, despegar, vuelos baratos, Skyscanner, tiquetes baratos, Avianca, Latam), alojamiento en Colombia (hoteles, Booking, Airbnb), blog de recomendaciones turísticas (TripAdvisor), y destinos vacacionales (Colombia, Bogotá, Cartagena de Indias, Medellín, Santa Marta). Por último, el tercer grupo hace referencia a la ocupación hotelera mensual (%).

Para efectos del ejercicio, se buscó información mensual de los últimos seis años, desde enero de 2018 hasta agosto de 2023; de manera que se disponen de 68 datos para cada variable. Por otra parte, se debe tener en cuenta que para que la información sea coherente con el objetivo de este proyecto, se rezagaron las variables explicativas. En otras palabras, para predecir la tasa de ocupación hotelera para las festividades de diciembre, los índices de los términos de búsqueda a emplear deberán ser los del mes inmediatamente anterior, es decir, noviembre.

4.4. Trato de datos y modelamiento

Para efectos del ejercicio, se utiliza el software estadístico R, con el objetivo de corroborar la hipótesis realizada en este trabajo y con la ayuda de Excel y su herramienta PowerPivot, se hace el trato de datos como se presenta en la **Figura 2**.

Figura 2

Fragmento de la base de datos

Año	Meses	OcupHot	OcupHot1	OcupHot2	OcupHot3	TD	Inflacion	TRM	Hoteles	Hoteles2
2018(p)	Enero	55,6								
	Febrero	53,9	55,6				0,63	2867,68	97	
	Marzo	56,3	53,9	55,6		12,1	0,71	2860,00	51	97
	Abril	52,8	56,3	53,9	55,6	11,2	0,24	2852,46	63	51
	Mayo	52,6	52,8	56,3	53,9	9,8	0,46	2765,96	51	63
	Junio	54,0	52,6	52,8	56,3	9,7	0,25	2862,95	52	51
	Julio	58,7	54,0	52,6	52,8	10,0	0,15	2893,22	66	52
	Agosto	60,0	58,7	54,0	52,6	9,3	-0,13	2885,55	72	66
	Septiembre	56,9	60,0	58,7	54,0	10,0	0,12	2959,57	60	72
	Octubre	58,8	56,9	60,0	58,7	9,4	0,16	3037,80	53	60
	Noviembre	61,0	58,8	56,9	60,0	9,7	0,12	3080,48	65	53
	Diciembre	54,7	61,0	58,8	56,9	9,4	0,12	3198,13	61	65
2019(p)	Enero	47,2	54,7	61,0	58,8	9,0	0,3	3212,48	80	61
	Febrero	48,1	47,2	54,7	61,0	10,0	0,6	3161,91	100	80
	Marzo	47,3	48,1	47,2	54,7	13,1	0,57	3115,15	53	100
	Abril	45,7	47,3	48,1	47,2	12,1	0,43	3125,34	58	53
	Mayo	45,6	45,7	47,3	48,1	11,2	0,5	3155,22	68	58
	Junio	48,3	45,6	45,7	47,3	10,7	0,31	3310,49	56	68
	Julio	49,5	48,3	45,6	45,7	11,0	0,27	3256,02	80	56
	Agosto	51,8	49,5	48,3	45,6	9,7	0,22	3208,11	76	80
	Septiembre	50,1	51,8	49,5	48,3	11,3	0,09	3412,65	71	76
	Octubre	49,7	50,1	51,8	49,5	11,2	0,23	3399,62	59	71
	Noviembre	53,3	49,7	50,1	51,8	10,8	0,16	3437,73	66	59
	Diciembre	49,8	53,3	49,7	50,1	10,1	0,1	3411,42	63	66
2020(p)	Enero	50,3	49,8	53,3	49,7	9,6	0,26	3383,00	86	63

Fuente: Elaboración propia.

Primero, a excepción de la variable dependiente, y buscando la aplicabilidad del modelo, se rezagan las variables para simular datos de series de tiempo, y encontrar un sentido práctico del ejercicio. De este modo, exceptuando la tasa de desempleo que, por su publicación bimensual se rezaga dos meses, se rezagan todas las variables explicativas un mes. Después, se crean otras 19 variables, de las cuales tres son nuestra variable de turismo rezagada uno, dos y tres meses respectivamente, esto porque, debido a la percepción de las empresas turísticas, puede que la ocupación hotelera de diciembre dependa también de sus valores en los meses anteriores. Las otras 16 variables se crean de las existentes de Google Trends y son las mismas variables rezagadas un mes adicional, o sea, dos meses. Así, ahora se tiene una base de datos de 39 variables y 68 observaciones, que se dividirá en dos diferenciadas: la primera, se obtiene al sacar

los últimos meses del año 2023 de la base de datos original, esta nueva base emplea información de enero a agosto de 2023, y se denomina “Base Testing (evaluación)” y la segunda, se compone de los 60 datos restantes después de extraer la base de evaluación de la base original, y utiliza datos de enero de 2018 a diciembre de 2022, y es designada como “Base Training (entrenamiento)”.

Segundo, se aplican los cuatro modelos de predicción explicados anteriormente: SVM, Regresión Lasso, Boosting Regressor, y Random Forest. Estos modelos facilitan desarrollar la regresión analítica, no obstante, toman los datos como si fueran de corte transversal, y no de series de tiempo. Por ello, los rezagos efectuados.

Por último, con estas dos bases de datos, se corren los modelos de regresión, y se aplica un modelo de validación ARIMA.

5. Presentación de la Propuesta

5.1. Modelos

Para usar los modelos, primero se aplica una base de datos de entrenamiento para "entrenar" al modelo sobre lo que debe hacer con las bases de datos originales. Después, se corren los modelos con la muestra de entrenamiento y luego se hace la evaluación de cada uno con las dos bases de datos existentes, buscando discriminar al mejor de ellos por el que arroje una menor media de la suma de los errores (MAE); y a su vez, se asegura que las dos evaluaciones de cada una de las bases de datos, sea parecida dentro de un mismo modelo. Por último, escogido el mejor modelo, se corre la regresión que arroja las predicciones de nuestra variable turismo.

5.2. Evaluación MAE

Para elegir el mejor modelo, se calculó el MAE de los pronósticos en la muestra de prueba para cada uno de los modelos estimados. Formalmente, el MAE se denomina como:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$$

Que describe una sumatoria de 1 hasta n, donde y_i son la serie de datos reales y $\widehat{y_i}$ las predicciones.

El modelo con menor MAE entre las predicciones y los valores reales observados se considera el que mejor se ajusta a los patrones en los datos. Por lo tanto, el modelo con menor MAE en la evaluación se selecciona como el modelo ganador para el pronóstico de la ocupación hotelera. De este modo, el MAE provee una métrica objetiva y cuantitativa para comparar los modelos y determinar cuál tiene el mejor desempeño predictivo.

5.3. Validación ARIMA

Para validar la propuesta de pronóstico con modelos de machine learning, se comparó el desempeño contra aproximaciones tradicionales de series de tiempo como ARIMA. Como se mencionó anteriormente, el modelo ARIMA permite modelar series temporales no estacionarias a partir de un proceso de ajuste que involucra identificar los órdenes óptimos para tres componentes: autorregresivo (AR), integrativo (I) y de medias móviles (MA).

Así, se estimó un modelo ARIMA para la ocupación hotelera mensual. Esta captura la estacionalidad anual de los datos. El modelo se ajustó usando los últimos 5 años de datos mensuales previos al periodo de prueba. Luego se calculó el error MAE del ARIMA para el periodo de ocho meses de prueba. Se comparó este error con el obtenido por el mejor modelo de machine learning seleccionado previamente.

6. Discusión de Resultados

A continuación, se presentan los resultados de la evaluación por el parámetro MAE de los modelos y la validación del uso del mejor modelo seleccionado.

Tabla 1

Resultados de la prueba de evaluación MAE y prueba de validación ARIMA

Aproximación	**MAE test set**
Random Forest	2.10
Regresión Lasso	3.12
SVM	3.60
Boosting Regressor	2.35
ARIMA	4.70

Fuente: Elaboración propia.

Como se puede observar, la **Tabla 1** presenta los resultados de la evaluación de los modelos, donde se advierte que Random Forest cuenta con el menor MAE traducido en mejores resultados; de manera que se ha seleccionado una mejor aproximación para anticiparse a la ocupación hotelera mensual, usando como insumo los datos de Google Trends.

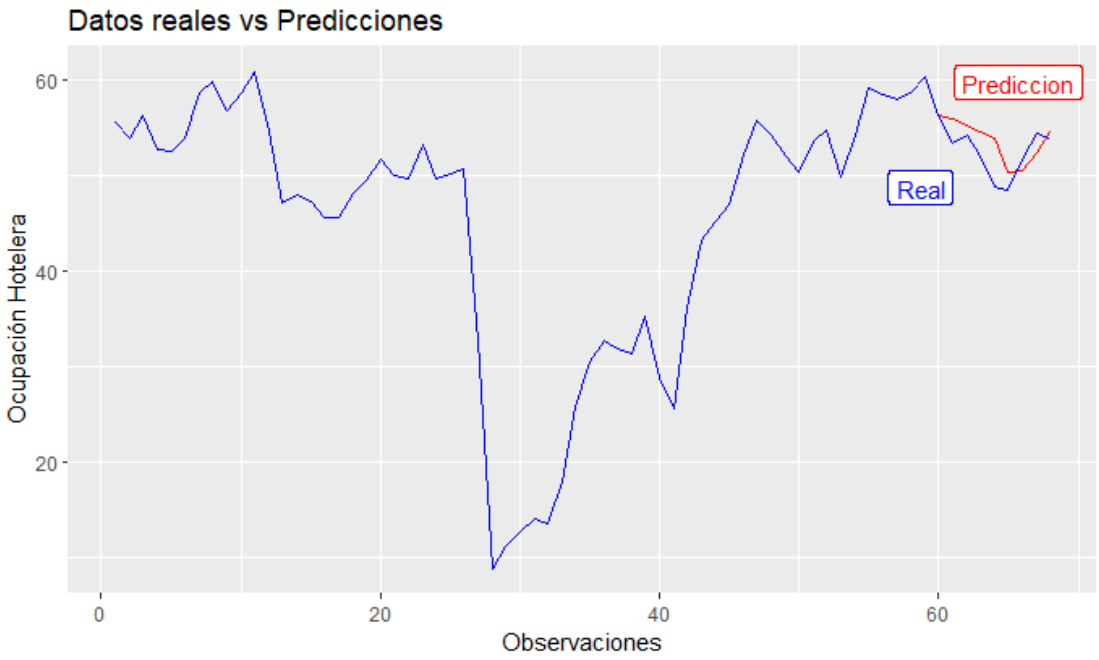
A su vez, el modelo ARIMA presentó un error de predicción mayor que el modelo de Random Forest seleccionado. Esto sugiere que la incorporación de información de búsquedas en Google mediante el enfoque de machine learning, logra mejoras en la precisión del pronóstico para la ocupación hotelera mensual en Colombia.

Por lo tanto, la propuesta basada en modelos de aprendizaje automático supera un enfoque tradicional de series de tiempo según la métrica de desempeño MAE. Esta validación refuerza la utilidad de la metodología planteada en el documento para pronosticar el indicador de interés.

Por último, se puede ver reflejado en la **Figura 3** el histórico de los datos reales y los pronosticados por el modelo Random Forest de la variable de ocupación hotelera donde, como lo explica su nombre, cada observación se da mes a mes.

Figura 3

Ocupación hotelera pronosticada y real, serie histórica



Fuente: Elaboración propia.

7. Conclusiones

El modelo de nowcasting para la ocupación hotelera en Colombia usando datos de Google Trends es una herramienta útil para obtener información adelantada sobre este indicador económico. Al comparar el desempeño contra un modelo tradicional de series de tiempo como ARIMA, el enfoque propuesto logra mejores resultados en términos del error MAE. Esto valida la metodología planteada y su capacidad para pronosticar la ocupación hotelera mensual.

Los resultados indican que la incorporación de información de búsquedas en Google Trends mediante modelos de machine learning como Random Forest permite capturar cambios en la demanda turística de forma oportuna. De esta manera, se supera la limitación de información desactualizada que enfrentan actualmente las empresas del sector.

El uso de Google Trends para nowcasting en turismo tiene gran potencial en Colombia, debido a que se pueden construir indicadores adelantados para diferentes destinos y segmentos, enfocándose en los términos de búsqueda más relevantes. También, es factible extender el enfoque a otras variables de interés en el sector, como: reservas aéreas, recaudo hotelero, entre otros.

En conclusión, este trabajo demuestra el potencial de las técnicas de nowcasting para suministrar información adelantada valiosa sobre el sector turismo en Colombia. Los resultados obtenidos motivan la adopción de estas metodologías para apoyar la toma de decisiones gerenciales y de política pública.

Anexos

En este apartado, se especifican características de los modelos empleados en este trabajo que, a pesar de no haber mencionado durante el desarrollo de este proyecto, fueron relevantes para entender su aplicabilidad, y el valor agregado que le añadían a las predicciones.

En un primer momento, haciendo uso del software estadístico R, se hace una prueba de correlaciones y se elabora una gráfica y una tabla de correlaciones, se identifican las variables explicativas que tienen más peso estadísticamente hablando y se corre una regresión múltiple de la variable dependiente.

Luego de este preprocesamiento de la base de datos, se aplican 4 modelos de predicción: regresión Lasso, regresión Ridge, Máquinas de Vectores de Soporte (SVM) y redes neuronales.

Las regresiones Ridge y Lasso son métodos de regresión que penalizan la magnitud de los coeficientes, minimizan el error de predicción y disminuyen la complejidad del modelo. Por su parte, el SVM permite la búsqueda de hiperplanos no lineales que ayuden en la clasificación de los datos para disminuir la complejidad del cálculo.

Se incorpora el concepto de *Redes neuronales* —basado en inteligencia artificial, denominado Deep Learning— que al tomar las variables como entrada, construye una función no lineal y devuelve una salida, con una pequeña diferencia en la estructura del modelo en comparación a los algoritmos generales.

Agradecimientos

Queremos agradecer a la Universidad Icesi por brindar las herramientas necesarias para la construcción de este trabajo de grado. En el marco de los cinco años que componen la carrera, tuvimos la oportunidad de asistir a distintas clases que enriquecieron nuestro conocimiento y nos motivaron a informarnos acerca de la situación económica mundial, en especial la colombiana. Por otro lado, los espacios de diálogo ofrecidos por la universidad permitieron que nos interesamos en problemáticas a las cuales es necesario dar solución, y de esa manera, surge este tema de grado en compás con nuestras áreas de interés. Asimismo, queremos agradecer a los profesores Julio Cesar Alonso Cifuentes y Cristian Camilo Hoyos del área económica de la Facultad de Ciencias Administrativas y Económicas, que desempeñaron el papel de tutor y cotutor respectivamente de este trabajo. Gracias a ellos tuvimos una idea clara de lo que se queríamos desarrollar en esta tesis; su guía, apoyo, y supervisión, permitieron que existiera un orden en el desarrollo de este escrito, así como una estructura de investigación coherente. También, queremos reconocer el apoyo emocional que representaron tanto nuestra familia como amigos, puesto que sus palabras de ánimo siempre fueron de gran ayuda en momentos de estrés y frustración.

Referencias

(S/f). Tripadvisor.co. Recuperado el 13 de octubre de 2023, de https://www.tripadvisor.co/Tourism-g294073-Colombia-Vacations.html

Blanco Vílchez, M. (2022). *Turismo y Coronavirus: análisis de la predisposición de la demanda de viajar a la Costa del Sol con una herramienta tecnológica de nowcasting, Google Trends*. Universidad de Málaga.

Cabria, M. (2023). Nowcasting de indicadores económicos combinando series de Google Trends. Tesis de economía. Universitat Politècnica de València. https://riunet.upv.es/bitstream/handle/10251/194367/Cabria%20%20Nowcasting%20de%20indicadores%20economicos%20combinando%20series%20de%20Google%20Trends.pdf?sequence=1&isAllowed=y

Correa, A. (2021). Prediciendo la llegada de turistas a Colombia a partir de los criterios de Google Trends. *Lecturas de economía*, *95*, 105–134. https://doi.org/10.17533/udea.le.n95a343462

DANE - Empleo y desempleo. (s/f). Gov.co. Recuperado el 15 de octubre de 2023, de https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo

DANE - Encuesta mensual de alojamiento (EMA). (s/f). Gov.co. Recuperado el 16 de octubre de 2023, de https://www.dane.gov.co/index.php/estadisticas-por-tema/servicios/encuesta-mensual-de-alojamiento-ema

Eslava M. & Fernández M. (2022). *Problemática del mercado laboral en Colombia*. Universidad de los Andes. https://uniandes.edu.co/es/noticias/economia-y-negocios/problematica-del-mercado-laboral-en-colombia

Inflación total y meta. (s/f). Gov.co. Recuperado el 15 de octubre de 2023, de https://www.banrep.gov.co/es/estadisticas/inflacion-total-y-meta

Meo, M. S., Chowdhury, M. A. F., Shaikh, G. M., Ali, M., & Masood Sheikh, S. (2018). Asymmetric impact of oil prices, exchange rate, and inflation on tourism demand in Pakistan: new evidence from nonlinear ARDL. *Asia Pacific Journal of Tourism Research*, *23*(4), 408–422. https://doi.org/10.1080/10941665.2018.1445652

Muestra mensual de hoteles (MMH). (s/f). Gov.co. Recuperado el 16 de octubre de 2023, de https://www.dane.gov.co/index.php/estadisticas-por-tema/servicios/muestra-mensual-de-hoteles-mmh

Ospina, A. C., & Caballero, L. (2019). *MODELO DE PRONÓSTICO PARALA DEMANDA DE TURISTAS EN COLOMBIA A PARTIR DE CRITERIOS DE BÚSQUEDA EN GOOGLE, UNA APROXIMACIÓN UTILIZANDO LA METODOLOGÍA MIDAS*. Unpublished. https://doi.org/10.13140/RG.2.2.31350.42567

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, *57*, 12–20. https://doi.org/10.1016/j.tourman.2016.04.008

Spinak E. (2001). Indicadores cienciométricos. *Acimed*, *9*, 16–18. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352001000400007

StatCounter. (2020). Search Engine Market Share Worldwide. https://gs.statcounter.com/search-engine-market-share

Tang, C. F., & Lean, H. H. (2007). Will inflation increase crime rate? New evidence from bounds and modified Wald tests. *Global Crime*, *8*(4), 311–323. https://doi.org/10.1080/17440570701739694

Tasa Representativa del Mercado (TRM - Peso por dólar). (s/f). Gov.co. Recuperado el 15 de octubre de 2023, de https://www.banrep.gov.co/es/estadisticas/trm

Velázquez López, O. N., Saavedra García, M. L., & Saavedra García, M. E. (2022). Las variables macroeconómicas y la demanda de la industria hotelera mexicana, periodo 2010-2016. *RAN*, *8*(1), 97–110. https://doi.org/10.29393/ran8-8vmom30008