Juan Gabriel Vanegas
Magíster en Economía de la Universidad de Antioquia
Institución Universitaria Visión de las Américas
Colombia
[juan.vanegas@uam.edu.co]

Guberney Muñetón Santa
Magíster en Estudios Socioespaciales de la Universidad de Antioquia
Instituto de Estudios Regionales, Universidad de Antioquia
Colombia
[guberney.muneton@udea.edu.co]

REVISTA TURISMO y SOCIEDAD

SATISFACCIÓN DEL TURISTA USANDO FACTORES MOTIVACIONALES: COMPARACIÓN DE MODELOS DE APRENDIZAJE ESTADÍSTICO

TOURIST SATISFACTION USING MOTIVATIONAL FACTORS: COMPARISON OF STATISTICAL LEARNING MODELS

Para citar el artículo: Vanegas, J. & Muñetón, G. (2024). Satisfacción del turista usando factores motivacionales: comparación de modelos de aprendizaje estadístico. *Turismo y Sociedad*, XXXIV, 149-178. DOI: https://doi.org/10.18601/01207555.n34.06

Fecha de recepción: 26 de agosto de 2022
Fecha de modificación: 14 de septiembre de 2022
Fecha de aceptación: 14 de octubre de 2022

Resumen

El nivel de satisfacción de un turista con el destino visitado y su intención de volver a visitarlo se asumen como dependientes de su experiencia previa con el lugar. Para observar esta perspectiva relacional, se utilizó un conjunto de datos de 386 turistas que visitaron la ciudad de Medellín (Colombia) durante el año 2018. Para predecir la variable de volver a visitar la ciudad y la satisfacción con el destino, se usaron las variables consideradas de empuje (*push*) y aquellas que halan (*pull*) al turista. Se estimaron cuatro modelos de aprendizaje estadístico para la clasificación de los turistas: regresión logística, árboles aleatorios, máquinas de soporte vectorial y el conjunto de aumento de gradiente extremo. Las variables más importantes en las estimaciones de la satisfacción fueron 'hablar sobre una experiencia de viaje en el futuro' e 'ir a lugares que mis amigos no han visitado'; y para volver a visitar la ciudad fueron 'visitar lugares históricos' y 'viajar a bajos precios'.

Palabras clave: satisfacción del turista, motivaciones del turista, aprendizaje supervisado, algoritmos de aprendizaje estadístico, máquinas de soporte vectorial, Medellín

Abstract

The level of satisfaction of a tourist with the destination visited, as well as his or her intention to revisit the destination, is assumed to be dependent on his or her previous experience with the place. To observe this relational perspective, a dataset of 386 tourists who visited the city of Medellin (Colombia) in 2018 was used. To predict the variables of revisiting the city and satisfaction with the destination, we consider push and pull variables. Four statistical learning

models were estimated to classify tourists: Logistic Regression (LR), Random Forests (RF), Support Vector Machines (SVM), and the Extreme Gradient Boosting algorithm. The most important variables in the satisfaction estimation were: 'talk about future travel experiences' and 'go to places my friends have not visited', while for revisiting the city the variables were: 'visit historical places' and 'travel at low prices'.

Keywords: tourist satisfaction, tourist motivations, supervised learning, machine learning algorithms, Support Vector Machines, Medellín

1. Introducción

Las formas en que los turistas expresan sus motivaciones para visitar un destino y los cambios que se desarrollan en el tiempo en tales motivaciones influyen directamente en la demanda y la oferta turísticas (Mansfeld, 1992). Así, en el proceso de elección de un destino, la comprensión de la motivación de los turistas aparece como un elemento clave no solo para lograr una adecuada gestión del turismo (Fodness, 1994), sino también para realizar mejoras en la competitividad del destino (Gil-León et al., 2021). Las dinámicas cambiantes y la continua evolución hacen que no solo los tomadores de decisiones requieran de información fundamentada para mantener los destinos atractivos, también los proveedores de productos y servicios turísticos necesitan conocer y anticiparse a los cambios en las motivaciones que determinan que los turistas los elijan (Bloom, 2004; Oh et al., 2004).

De esta forma, las motivaciones del turista ocupan un papel preponderante en los modelos de relacionamiento con la imagen, la satisfacción y la lealtad con el destino visitado (Chi & Qu, 2008; Correia et al., 2013; Yoon & Uysal, 2005). No obstante, la evidencia empírica que corrobora la importancia relativa de las motivaciones proviene estrictamente de modelos relacionales basados en ecuaciones estructurales (Albayrak & Caber, 2018; Chi & Qu, 2008; Correia et al., 2013; Do Valle et al., 2006; Luna-Cortés, 2020; Prebensen et al., 2010; Yoon & Uysal, 2005) y de modelos logísticos o multinomiales (Huang et al., 2018; Jang & Cai, 2002; Lam-González et al., 2019; Lee et al., 2002; Yoo et al., 2018).

Varios estudios de segmentación en el campo turístico han aplicado modelos de aprendizaje estadístico como los que se proponen en este trabajo, es decir, aplicaciones para la predicción de la satisfacción del turista (Li et al., 2009) y de los asistentes a eventos (Oh & Lee, 2021). Ahora bien, en el campo turístico, un libro reciente repasa varios algoritmos de aprendizaje estadístico aplicados en diversos escenarios predictivos (Egger, 2022). En los últimos años, la implementación de estas técnicas ha mostrado no solo una tendencia creciente de su uso, sino también un diverso tipo de áreas temáticas en las que se han publicado los trabajos, que principalmente se recogen en el campo de la inteligencia artificial en las ciencias de la computación, el turismo deportivo de ocio y hospitalidad, y métodos teóricos en las ciencias de la computación (Egger, 2022, p. 88). Cabe asimismo señalar que este tipo de técnicas son una oportunidad para reaprender de los ejercicios empíricos que hasta ahora han dominado la temática de estudio, con el fin de contrastar las posturas teóricas que se derivan de este tipo de análisis (Salganik et al., 2020). Además, desde el punto de vista de los indicadores de confiabilidad de los modelos, es importante destacar

que, en el campo del conocimiento de las ciencias administrativas, esta clase de ejercicios muestran resultados de desempeño conjunto modestos (Żbikowski & Antosiuk, 2021).

En este sentido, el objetivo principal de este trabajo se centra en predecir la satisfacción del visitante mediado por las variables de motivación. De esta manera, si el conjunto de factores motivacionales explica o predice el nivel de satisfacción del turista, es necesario implementar modelos basados en aprendizaje estadístico para demostrar tanto la estimación como la predicción de estos factores sobre los cambios en la satisfacción. Considerando el contexto anterior, en el presente trabajo se aborda esta brecha investigativa desde la perspectiva metodológica. Así, se utilizan tres de los principales algoritmos por su desempeño y rendimiento en las competencias de clasificación tipo *Kaggle*: máquinas de soporte vectorial, árboles aleatorios y el incremento extremo del gradiente (XGBoost). Además, se utiliza la regresión logística como algoritmo de base para la comparación de los rendimientos por ser uno de los más usados en el área de investigación de las motivaciones del turista. Los algoritmos aquí implementados se usan para predecir la satisfacción de los turistas con base en un conjunto de variables motivacionales (*push* y *pull*).

Este trabajo se estructura en cinco secciones, comenzando por la introducción de los factores motivaciones y las metodologías empleadas en su estudio. Luego, en la segunda sección se aborda la revisión de la literatura sobre los trabajos aplicados en esta temática. En la tercera parte se describe en detalle la metodología utilizada para realizar el estudio. Posteriormente, en el cuarto apartado se destacan los principales resultados de los modelos propuestos y se contrastan entre sí empleando un diverso tipo de métricas. Por último, se exponen las principales aportaciones del estudio, las recomendaciones y las futuras líneas de actuación en la materia.

2. Trabajos relacionados

Existen varios trabajos aplicados en la explicación de los determinantes de la satisfacción de un turista desde diversas posturas conceptuales y temáticas. Desde la perspectiva de Kozak (2001), existen cuatro modelos evaluativos de esta dimensión: modelo de expectativa-desempeño, modelo de importancia- desempeño, modelo de desconfirmación de la expectativa y modelo de solo rendimiento. Ahora bien, tomando como referencia lo planteado por Chen et al. (2013), la teoría de la satisfacción del turista adopta ampliamente posturas derivadas de la satisfacción del cliente provenientes de estudios de la administración. Este mismo autor aduce que cuestiones como la amplitud de la satisfacción del turista en los destinos y la singularidad del efecto de la interacción entre los turistas y los destinos requieren de mayor exploración y desarrollo conceptual.

De igual manera, se destaca que la literatura muestra cómo los enfoques transitan desde la satisfacción, pero enfocan otras dimensiones asociadas, como el papel de la cadena de valor (Ghaderi et al., 2018), las condiciones relacionadas con la prestación de los servicios turísticos (Yu & Goulden, 2006), así como el valor de la marca, que considera aspectos vinculados al conocimiento, la imagen, la calidad y la lealtad (San Martín et al., 2019). No obstante, la revisión adelantada en este trabajo prioriza la identificación y selección de variables desde la perspectiva motivacional (*push* y *pull*), así como desde el punto de vista aplicado de la modelación.

2.1 Estudios empíricos que consideran las motivaciones del turista

Un diverso tipo de factores determinan la elección de un destino por parte de un turista. La literatura académica ha abordado los determinantes críticos de las motivaciones, la satisfacción y la imagen que se forja un visitante al momento de disfrutar un destino (Chi & Qu, 2008; Correia et al., 2013; Do Valle et al., 2006; Jang & Cai, 2002; Lee, 2009; Olague de la Cruz, 2015; Yoon & Uysal, 2005).

Vale la pena reseñar algunos estudios que soportan este trabajo, la mayoría de los cuales opta por metodologías multivariadas como forma de explicar las variables independientes y su relacionamiento con la variable dependiente. Así, Yoon y Uysal (2005) en su estudio analizan un modelo de ecuaciones estructurales para comprender mejor las motivaciones que tienen los turistas para viajar a un destino, además de aportar al tema de la relación que tienen las motivaciones (empuje y atracción), la satisfacción y la lealtad para con el destino. Usando esta misma técnica, Do Valle et al. (2006) indagan por la relación que existe entre la satisfacción del viaje y la intención de lealtad al destino. Chi y Qu (2008) ofrecen un enfoque sobre la lealtad al destino mediante la examinación de la relación entre la imagen del destino, los atributos turísticos y la satisfacción. Lee (2009) analiza el comportamiento de los turistas a partir de cinco factores: imagen del destino, satisfacción, motivación, actitud y comportamiento futuro. Por su parte, Correia et al. (2013) muestran que la satisfacción surge de la oportunidad de experimentar las especificidades culturales y sociales de un destino. Autores como Villamediana-Pedrosa et al. (2020) y Dean y Suhartanto (2019) analizan las motivaciones usando modelos de regresión con escalamiento óptimo en el primer caso, mientras que en el segundo los autores utilizaron mínimos cuadrados parciales; la estrategia de emplear el escalonamiento óptimo permite convertir datos de tipo tabulación a una forma continua que hace posible un amplio uso de estadística paramétrica. Para el caso latinoamericano, más concretamente en Monterrey (México), Olague de la Cruz (2015) propone un modelo capaz de medir y explicar la satisfacción y la lealtad de los turistas.

2.2 Estudios aplicados que emplean aprendizaje estadístico en contextos turísticos

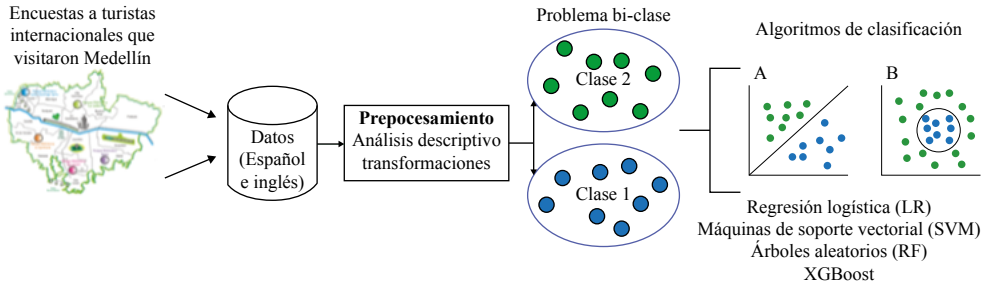
Muy pocos ejercicios usan técnicas de aprendizaje estadístico para la predicción del grado de satisfacción. Uno de estos presenta un modelo de máquinas de soporte vectorial para la evaluación del grado de satisfacción del turista (Li et al., 2009). Otro de los trabajos fue desarrollado por Ghaderi et al. (2018), quienes proponen un modelo de redes neuronales en la evaluación de la gestión de la cadena de suministro del turismo que influye en la satisfacción de los turistas. También se encuentra la propuesta de Oh y Lee (2021), quienes identifican las variables más significativas de la estrategia turística en la predicción de la satisfacción de asistentes a festivales. En otros temas específicos, las herramientas que aplican el aprendizaje estadístico van ganando terreno, aunque es de destacar la creciente utilización de este tipo de ayudas en los ejercicios de modelación que van desde redes neuronales artificiales, árboles de decisión, métodos probabilísticos y bayesianos, hasta aprendizaje basado en instancias, conjuntos (*ensembles*), agrupamiento y minería de reglas de asociación (Egger, 2022; Guerra-Montenegro et al., 2021).

Para el análisis de segmentación de turistas, se han usado técnicas de redes neuronales con algoritmos autoorganizados y de retropropagación (Bloom, 2004). También para el análisis de segmentación de mercado en hoteles y predicción de elecciones de viaje, se han combinado técnicas de agrupación y reducción de dimensiones (mapas autoorganizados y descomposición de valores singulares de orden superior) con técnicas de predicción usando árboles de decisión (Ahani et al., 2019). Para el sentido del lugar en un atractivo turístico, se han usado modelos de regresión logística y modelos de análisis no supervisado, como la asignación latente de Dirichlet (Song et al., 2021). Y para determinar los factores subyacentes en el valor del cliente en restaurantes, se han empleado redes neuronales artificiales para el modelamiento de tópicos, usando como variable de entrada la incrustación de palabras (*word embeddings*) (Kwon et al., 2020). Sin embargo, dada la cantidad de algoritmos disponibles con demostrados rendimientos en las competencias computacionales tipo *Kaggle*, se abre la discusión sobre el mejor modelo para los problemas de clasificación o identificación de patrones en el tema de satisfacción de los turistas con el destino turístico.

3. Materiales y métodos

La investigación tiene un flujo metodológico que se puede observar en la Figura 1. Los datos se recolectaron vía encuesta cerrada cara a cara en el año 2019. Se procedió luego a una consolidación de la base de datos y al procesamiento de limpieza de errores. Se identificaron dos variables de interés relacionadas con la evaluación del destino y la posibilidad de volver a visitarlo. Las variables se estructuraron en dos niveles al incluir las diversas opciones de respuesta con el fin de tener más muestras por clase; no se consideró la estimación con las variables originales porque eran muy pocos datos para estimar los algoritmos. Las dos variables se estimaron probando los principales modelos de clasificación, entre ellos, la regresión logística (LR), las máquinas de soporte vectorial (SVM, por su sigla en inglés), los árboles aleatorios (RF) y el incremento extremo del gradiente (XGBoost).

Figura 1. Metodología general



Nota. Elaboración propia.

3.1 Datos

La información fue obtenida por medio de un cuestionario administrado a 404 turistas que visitaron la ciudad de Medellín en diciembre de 2018. La recolección de información se hizo cara a cara en diferentes hoteles y lugares turísticos de la ciudad; la encuesta la

aplicaron profesionales en el área de aplicación de instrumentos en campo. El proceso de depuración consistió en eliminar registros sin un total de respuestas superior al 80 % y aquellos con inconsistencias en las respuestas. Luego del proceso de depuración quedaron 386 registros válidos con información completa.

El cuestionario aplicado estaba compuesto por características sociodemográficas (12 preguntas), motivaciones inherentes al viajero (22 preguntas), motivaciones inherentes al destino (24 preguntas) y la satisfacción con el destino visitado (7 preguntas). Las variables de motivaciones son usadas para analizar la percepción del destino en comparación con otras experiencias de los viajeros; en la literatura, estas variables se denominan endógenas (*push*) y exógenas (*pull*) al viajero. Las variables *push* —que se puede traducir como *empuje*— y *pull* —*atracción*— se midieron en una escala de 5 puntos: 1 es *nada importante*; 2, *poco importante*; 3, *neutral*; 4, *importante*; y 5, *muy importante*. Dichas variables fueron consideradas como las predictoras de los modelos (ver Tabla 1).

Tabla 1. Motivaciones específicas según categorías

Motivaciones de empuje		*Motivaciones de atracción*	
*push*1.1	Estar activo físicamente	*pull*1.1	Es una ciudad moderna
*push*1.2	Reunirme con o conocer a otras personas	*pull*1.2	Tiene una atmósfera exótica
*push*1.3	Encontrar algo nuevo y excitante	*pull*1.3	Museos y galerías de arte
*push*1.4	Experimentar cómo viven otras personas	*pull*1.4	Restaurantes de alta calidad
*push*2.1	Experimentar nuevos o diferentes estilos de vida	*pull*2.1	Variedad de actividades para realizar
*push*2.2	Probar nuevos alimentos o comidas	*pull*2.2	Alojamiento económico
*push*2.3	Visitar lugares históricos	*pull*2.3	Destino de bajo costo
*push*2.4	Conocer gente nueva	*pull*2.4	Restaurantes económicos
*push*2.5	Ser libre para actuar como quiera	*pull*3.1	Se puede recorrer fácilmente
*push*3.1	Cambiar de ambiente por tanto trabajo	*pull*3.2	Tiene un muy buen clima
*push*3.2	Ir a lugares que mis amigos no han visitado	*pull*3.3	Es segura
*push*3.3	Hablar sobre experiencias de viaje en el futuro	*pull*4.1	Paisaje excepcional
*push*3.4	Redescubrir experiencias pasadas	*pull*4.2	Variedad de patrimonio cultural diferente al mío
*push*3.5	Escapar del estrés diario	*pull*4.3	Gente local interesante y amable
*push*4.1	Visitar lugares donde mi familia ha estado	*pull*4.4	Diferentes culturas
*push*4.2	Visitar a amigos o parientes. Estar junto a familiares	*pull*4.5	Pueblos antiguos e históricos
*push*5.1	Alejarme de las demandas del hogar	*pull*5.1	Limpieza de sus calles
*push*5.2	Vivir un estilo de vida más simple	*pull*5.2	Realizar compras

Motivaciones de empuje		*Motivaciones de atracción*	
*push*6.1	Estar entretenido y pasarla bien	*pull*6.1	Vida nocturna y entretenimiento
*push*6.2	Viajar a bajos precios	*pull*6.2	Cocina local
*push*7.1	Sentirme como en casa lejos de ella	*pull*7.1	Actividades económicas
*push*7.2	Conocer tanto como sea posible	*pull*7.2	Actividades deportivas
		*pull*7.3	Turismo sexual
		*pull*7.4	Turismo médico

Nota. Elaboración propia tomando como referencia a Chi & Qu (2008) y Yoon & Uysal (2005).

Se estimaron dos variables diferentes relacionadas con la evaluación del destino y la posibilidad de volver a visitarlo. Una de las variables fue la satisfacción con el destino en comparación con otras ciudades visitadas; la pregunta precisa en el cuestionario fue la siguiente: "¿Cómo calificaría usted a Medellín como un destino para vacaciones comparado con otras ciudades similares que ha visitado?". Dicha variable se midió en una escala de 5 valores, donde 1 significaba *mucho peor*, y 5, *mucho mejor*. El valor de 1 no se presentó en las respuestas; los porcentajes de distribución de las demás respuestas fueron 2 (4,92 %), 3 (28,49 %), 4 (38,86 %) y 5 (27,72 %). Para la clasificación se colapsaron los valores 1, 2 y 3 en la clase *aceptable*, y los valores 4 y 5 en la clase *agradable*. Para efectos de la estimación, la clase *agradable* es la principal y se recodifica como 1, y la clase *aceptable*, como 0. La base de datos no presenta valores perdidos, todas las variables tienen datos completos para cada observación.

La segunda variable que se estimó tuvo en cuenta la recomendación de la ciudad. Al respecto, concretamente se preguntó: ¿Recomendaría visitar Medellín a familiares y amigos? La opción de respuesta tenía una escala de tres valores, donde 1 se asociaba a *poco probable*, 2 a *no lo sé* y 3 a *muy probable*. La proporción de respuestas fue 1 (16,58 %), 2 (35,75 %) y 3 (47,66 %). De esta forma, se asumió una posición conservadora para crear dos niveles de respuesta de clasificación: las opciones 1 y 2 se codificaron como *no*, y la opción 3, como *sí*. Al simplificar las opciones de manera conservadora, se pretende que los algoritmos utilizados encuentren más desafío en predecir la opción positiva.

3.2 Técnicas aplicadas

El problema abordado es de clasificación, entonces, los algoritmos estiman funciones para asignar las observaciones en una de las clases dadas en el problema. Para el presente trabajo se implementaron cuatro algoritmos: regresión logística (RL), *Support Vector Machines* (Vapnik, 2000) o máquinas de soporte vectorial (SVM), *Random Forest* (RF) (Breiman, 2001) o árboles aleatorios y *eXtreme Gradient Boosting* (XGBoost) (Chen & Guestrin, 2016) o incremento extremo del gradiente. La línea base de comparación será el resultado de la regresión logística, que se considera un método de estimación paramétrica, es decir, parte del conocimiento *a priori* de la distribución de los datos. La comparación se hará contra los demás algoritmos que son de estimación no paramétrica: SVM, RF y XGBoost,

este último considerado el 'gold standard' para los tipos de datos cuando las variables predictoras son nominales u ordinales.

3.2.1 Regresión logística

La regresión logística permite predecir una respuesta binaria al problema de clasificación, así: $p(x) = Pr(y = 1|x)$, donde la probabilidad de x sigue esta función logística:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

El ajuste se hace por el método de máxima verosimilitud. La regresión logística produce una curva en función de S. De acuerdo con el valor de x se obtiene la predicción. La regresión logística es un método de clasificación usado como línea base de comparación debido a que su desempeño ha demostrado ser menor que el de otro tipo de algoritmos, aunque es un algoritmo que permite comprender de forma directa las relaciones entre las variables.

3.2.2 Máquinas de soporte vectorial

Las máquinas de soporte vectorial (Vapnik, 2000) son una generalización de los clasificadores de margen máxima, pero con capacidad para acomodarse a problemas en los que las clases que se van a separar tienen fronteras no lineales (James et al., 2013). Para manejar el problema en el que la frontera de separación de clases es no lineal, se extienden las llamadas máquinas de soporte vectorial de margen dura. Entonces, el espacio original de medición de las características se cambia aplicando una serie de funciones sobre las variables predictoras y resolviendo un problema de optimizar el hiperplano separador. Al usar una función no lineal sobre las variables originales, el algoritmo mapea el vector de variables originales en un espacio de dimensión superior llamado espacio de características. Sin embargo, una ventaja del algoritmo es la posibilidad de usar el truco *Kernel* para evitar el manejo explícito del espacio de características, aunque se mapean las variables originales en el espacio de características (Abe, 2005). Por consiguiente, el *Kernel* es parte de la elección que debe realizarse al implementar el algoritmo, puesto que existen funciones *kernels* lineales, polinomiales, de base radial, basadas en la distancia de Mahalanobis, además de funciones *Kernel* desarrolladas para problemas específicos de procesamiento de imágenes, clasificación de texto o reconocimiento de voz; sin embargo, el *Kernel* de base radial ha sido el más usado (Abe, 2005).

En problema general de optimización en casos no separables se define como:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \xi_i^p$$

$$Sujeto\ a\ y_i\ (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [m]$$

Esta se interpreta como la maximización del margen de separación del hiperplano entre dos clases donde no hay separación lineal.

3.2.3 Árboles aleatorios

Se trata de uno de los algoritmos más usados para las tareas de clasificación por sus altos desempeños y la identificación de las variables explicativas (Schonlau & Zou, 2020). El algoritmo es un método de ensamblaje que utiliza diferentes árboles de decisión construidos sobre muestras de entrenamiento repetidas tomadas desde el mismo conjunto de datos de entrenamiento (método *boostrap*); ello permite reducir la varianza al calcular el valor promedio del conjunto de modelos estimados de forma independiente. Cada vez que se construye un árbol, se hace un muestreo aleatorio para elegir *m* variables predictoras que se usan para la partición de cada nodo del árbol; solo se emplea un muestreo aleatorio del conjunto de predictoras disponibles para el problema (James et al., 2013).

La elección típica de la cantidad de predictoras es $m \approx \sqrt{p}$, donde p es el número total de variables predictoras. La idea de usar diferentes variables predictoras para la construcción de los árboles mejora la aleatoriedad y reduce la varianza, en comparación con los métodos clásicos basados en árboles (James et al., 2013). El algoritmo es robusto ante valores atípicos (*outliers)* y el ruido, también es más rápido que el *Adaboost* y el *Bagging.* Además, con el algoritmo se pueden obtener una estimación interna del error, la correlación y las variables de importancia (Breiman, 2001). Los hiperparámetros que se pueden estimar son el número de variables *m*, el número de árboles que se van a considerar y el valor mínimo de observaciones para permitir que se active un nodo dentro de un árbol.

3.2.4 XGBoost

El XGBoost es un sistema escalable para el algoritmo que combina árboles en serie (*tree boosting*), por consiguiente, está basado en ensamblaje de árboles; la idea del ensamblaje es reducir el error de predicción o error de generalización. En el caso del XGBoost, se enfoca en reducir el sesgo al mejorar la velocidad de cálculo y la precisión (Chen & Guestrin, 2016). El XGBoost realiza la construcción de los árboles en paralelo, similar al modelo de árboles aleatorios; sin embargo, a diferencia de estos últimos, el XGBoost usa el algoritmo de gradiente descendente para la optimización (*Gradient Boosting*). La idea de la optimización es ir corrigiendo los errores del modelo y aprender en las siguientes iteraciones para mejorar el rendimiento hasta que no se logren mejoras. Es un algoritmo que ha conseguido ubicarse como uno de los más populares en las competencias de *machine learning*, al lado de los modelos de redes neuronales (Chen & Guestrin, 2016). Su principal factor de éxito es la escalabilidad para diferentes procesos.

3.3 Experimentos

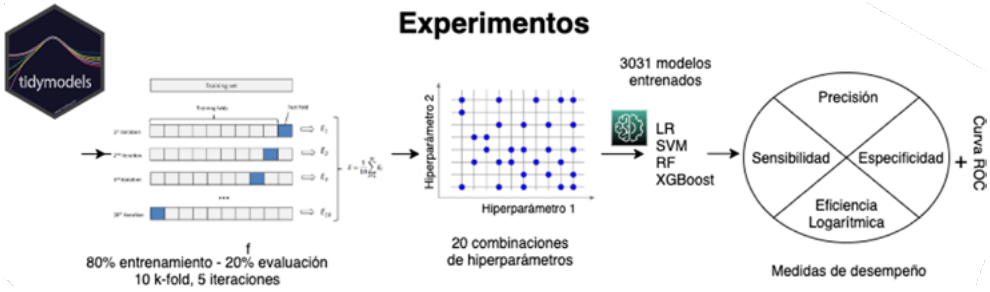
Se utilizó el *software R* para la estimación de los modelos, principalmente, el ecosistema de paquetes de *Tidyverse* (Wickham et al., 2019), que incluye *Tidymodels* (Kuhn & Wickham, 2020). Este último contiene una serie de paquetes para *machine learning* que permiten mantener un flujo de trabajo idéntico para correr en serie diferentes modelos usando los

mismos controles para la estimación de los hiperparámetros y, con ello, garantizar menos sesgos en la comparación de los resultados.

Para estimar los cuatro algoritmos de aprendizaje estadístico, se realizó una malla de búsqueda con 20 combinaciones diferentes de los hiperparámetros de los modelos ajustados. Para realizar una eficiente comparación de los modelos, se usó un flujo de trabajo común entre ellos, es decir, se empleó la misma validación cruzada de 10 particiones con 5 repeticiones en cada una; igualmente, se utilizaron las mismas métricas de desempeño, a saber: la precisión (*accuracy*), la curva *ROC*, la sensibilidad, la especificidad y la pérdida del logaritmo. En total, se entrenaron 3.031 modelos (Figura 2).

Una vez realizado el preprocesamiento de los datos, se balancearon las clases usando un método de elección por muestreo aleatorio de las observaciones con la clase predominante, así se obtuvo una base de datos balanceada para el entrenamiento. Para los procesos de optimización de los algoritmos, se empleó una matriz de *one-hot* en la que se dicotomizaron las variables para correr los algoritmos de SVM y XGBoost; para el caso de la regresión logística y los árboles aleatorios no se aplicó la transformación de las variables a variable *one-hot* o *dummy*.

Figura 2. Flujo de los experimentos realizados



Nota. Elaboración propia.

En el algoritmo de árboles aleatorios se optimizaron los siguientes hiperpárametros: el número de variables utilizadas en cada partición del nodo (*mtry*); la cantidad de árboles, que se fijó en 1.000; y el mínimo de observaciones para activar un nodo dentro de cada árbol (*min_n*). En las máquinas de soporte vectorial se estimaron el costo de la violación de la restricción —es decir, la constante del término de regularización en la formulación del lagrangiano (*cost*)— y el valor del sigma del *Kernel* gaussiano (*rbf_sigma*). En el XGBoost se estimaron el número de variables usadas en cada partición del nodo (*mtry*), el número de árboles del ensamblaje, el mínimo de observaciones para activar un nodo dentro de cada árbol (*min_n*), la profundidad máxima de cada árbol (*tree_depth*), la tasa de aprendizaje (*learn_rate*), la reducción en la función de pérdida (*loss_reduction*) para crear una nueva partición en el árbol y el número de observaciones para el ajuste en una estimación (*sample_size*).

Se utilizaron cinco medidas de desempeño para interpretar los resultados de cada modelo: el desempeño (*accuracy*, por su nombre en inglés), la sensibilidad, la especificidad, la pérdida del logaritmo (*Mn_log_loss*) y la curva *ROC*. El mejor modelo se eligió de acuerdo con el indicador de desempeño. Comparado con el *accuracy,* la pérdida del logaritmo toma en cuenta la incertidumbre en la predicción.

4. Resultados y discusión

4.1 Análisis descriptivo

Las características generales de los turistas encuestados se presentan en la Tabla 2. En cuanto al sexo, se encuentra un balance en la base de datos. La mayoría de las personas que respondieron la encuesta se encontraban entonces en el rango etario de 25 a 31 años (45,3 %), eran solteros (67,6 %), empleados (49,2 %), con nivel educativo universitario (60,1 %), viajaron principalmente con familiares o amigos (45,8 %) y visitaron la ciudad por razones de turismo (69,7 %).

Tabla 2. Características sociodemográficas de los encuestados

Característica	*Porcentaje (%)*
Sexo	
Femenino	49,7
Masculino	50,3
Edad	
18-24	24,1
25-31	45,3
Más de 32	30,6
Estado civil	
Casado	22,3
Soltero	67,6
Otro	10,1
Ocupación	
Desempleado	9,8
Empleado	49,2
Estudiante	25,4
Otro	15,5
Nivel educativo	
Posgrado	22,8
Primaria/Secundaria	17,1

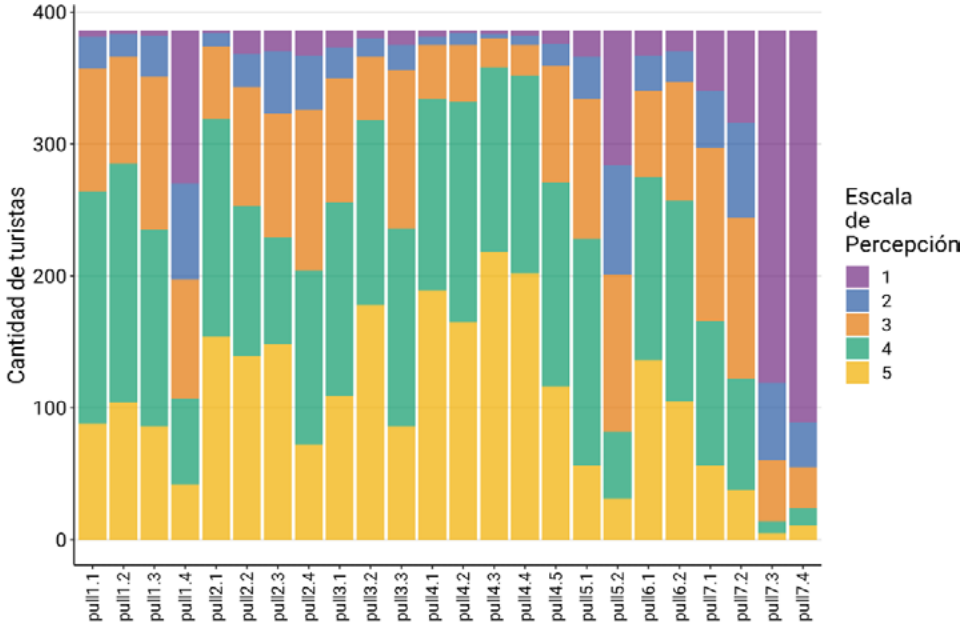
Característica	*Porcentaje (%)*
Universitario	60,1
¿Con quién viaja?	
Amigos o familiares	45,8
Esposo(a) e hijos	24,4
Solo	29,8
Tipo de turismo	
Negocios	13,7
Otro	16,6
Turismo	69,7
n = 386	

Nota. Elaboración propia.

Las variables independientes o predictoras del modelo, *pull* y *push*, se presentan en una escala de respuesta que va del 1 al 5, donde el primer valor representa la menor percepción, y la segunda, la mayor. En este sentido, las variables *pull* 7.3 y 7.4 son las que presentan mayor proporción de turistas con calificaciones negativas (para observar el nombre de la variable, ver la Tabla 1). A esas apreciaciones negativas también se pueden agregar las variables *pull* 1.4 y 5.2, las cuales, agrupando los valores de respuesta de 1, 2 y 3, logran más del 60 % de los turistas (Figura 3). Igualmente, de acuerdo con la Figura 4, las variables *push* que evidencian una tendencia a las respuestas negativas son la 4.1, la 4.2, la 3.2, la 3.4 y la 1.1. En cuanto a los factores positivos, en las variables *pull* se destacan las siguientes: 4.3, 4.4, 4.1, 4.2 y 2.1 (Figura 3). En el caso de las variables *push*, entre aquellas con mayor porcentaje de turistas que las evalúan de forma positiva se encuentran estas: 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4 y 6.1 (Figura 4).

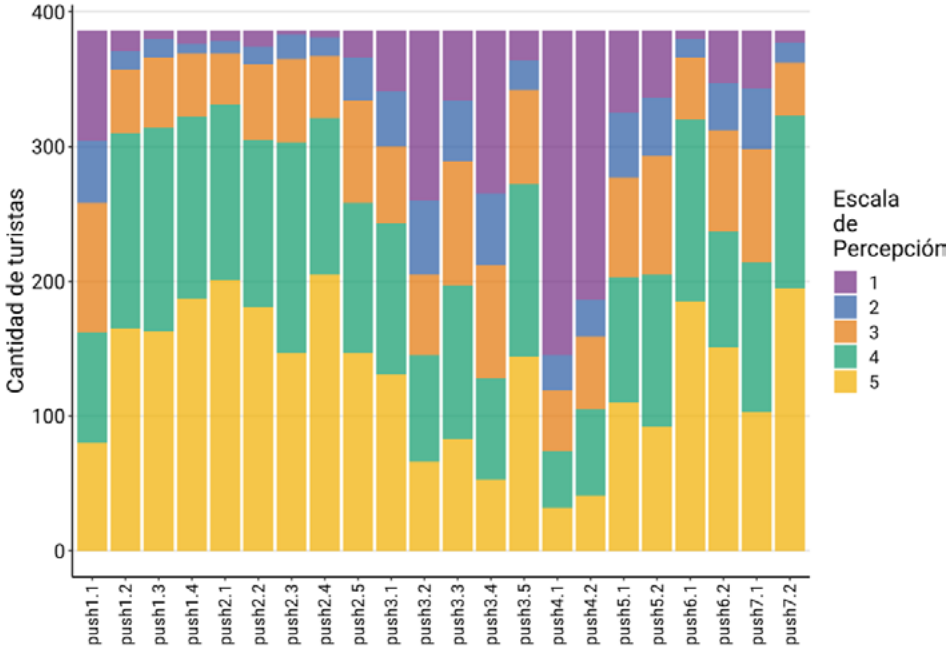
El resultado descriptivo muestra que las variables "*push* 5.1" y "*push* 5.2" son aquellas en las que más del 50 % de los turistas respondieron que son aspectos *nada importantes*, relacionados con visita a familiares o recomendaciones de ellos (Figura 3). Igualmente, se observa un porcentaje alto de respuestas (alrededor del 45 %) con valoraciones de *nada importante* y *poco importante* en los aspectos "*push* 3.2" y "*push* 3.4", que están relacionados con ir a los lugares que los amigos no han visitado y redescubrir experiencias pasadas. En cuanto a las variables con calificaciones altas de importancia, se observan la 1.2, asociada a la emoción y la pasión por viajar; la 2.4, relacionada con el conocimiento y el aprendizaje; la 7.1, sobre pasarla bien; y la 7.2, acerca de conocer tanto como sea posible.

Figura 3. Respuesta de las variables *pull*



Nota. Elaboración propia.

Figura 4. Respuesta de las variables *push*



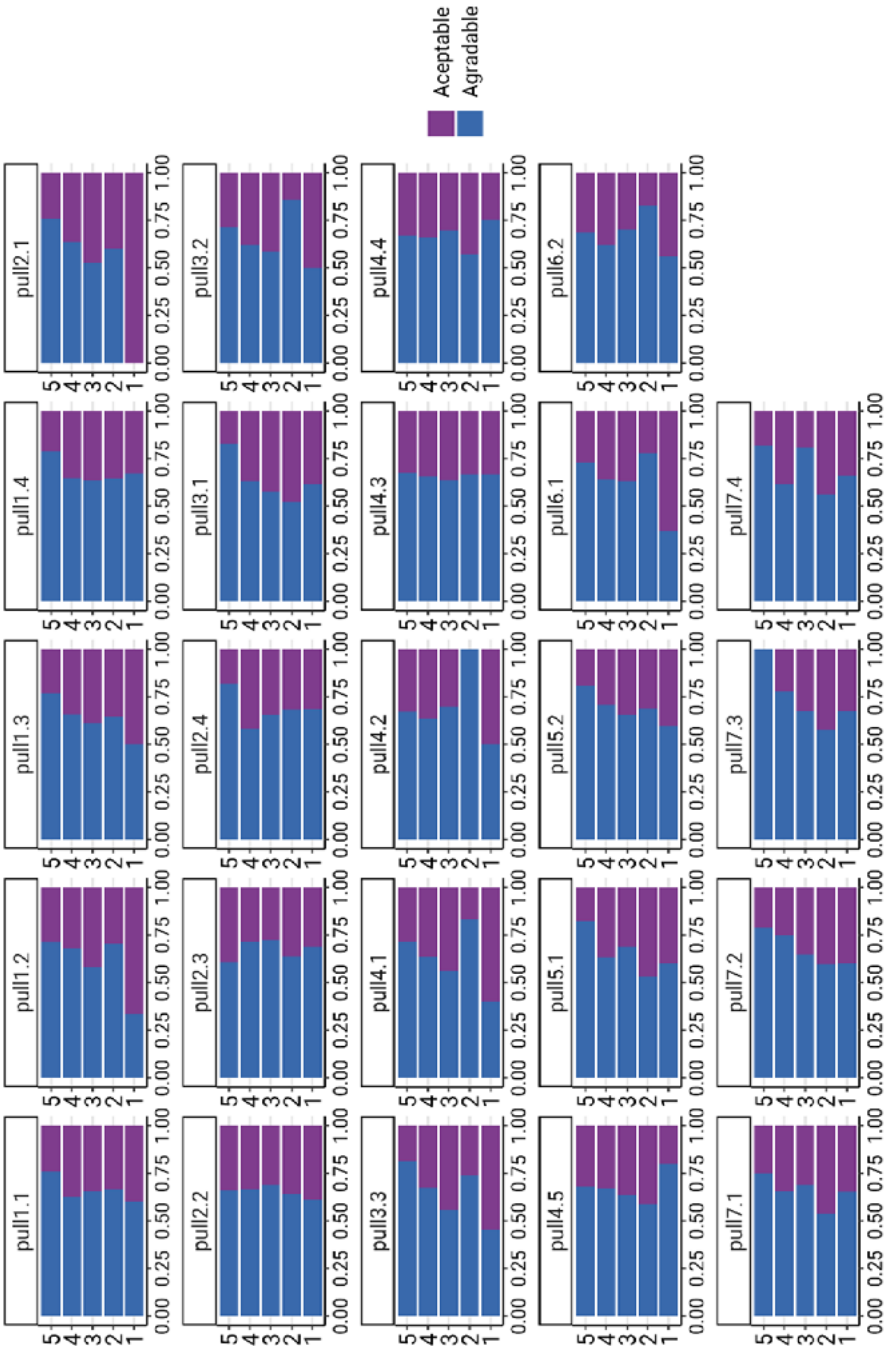
Nota. Elaboración propia.

Ahora bien, al cruzar las variables de los factores motivacionales (*push* y *pull*) con la variable de satisfacción con la ciudad que se busca explicar, codificada como *aceptable* y *agradable*, se pueden observar algunos rasgos distintivos (Figuras 5 y 6). Así, se puede esperar que motivaciones del grupo de la relajación (3.1 a 3.4), del grupo de lazos (5.1 y 5.2) y del grupo de comodidad (8.1 y 8.2) tiendan a asignar un nivel de satisfacción más alto, lo cual está relacionado con la valoración agradable de la ciudad.

Al observar las relaciones entre las motivaciones inherentes al destino y la satisfacción con la ciudad, no se evidencia una tendencia clara de separación entre las respuestas altas o bajas de las motivaciones y la variable de la satisfacción con la ciudad. Esta evidencia preliminar que se observa en la figura descriptiva sugiere la necesidad de algoritmos que logren estimar los patrones ocultos en el análisis descriptivo. El tener patrones difusos implica que se está ante un problema de investigación complejo para la predicción y, por consiguiente, se esperan valores de estimación moderados.

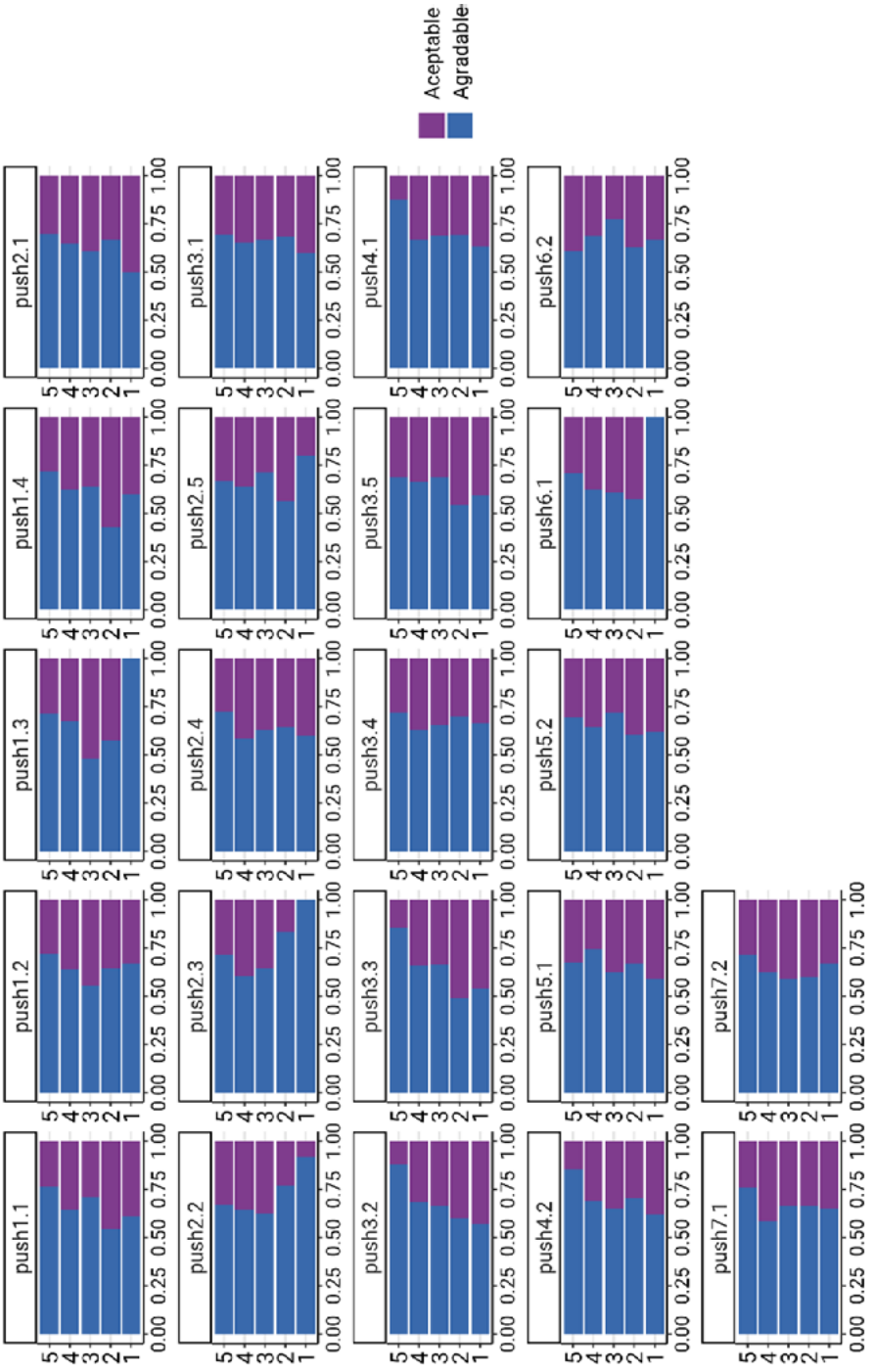
Al evaluar el cruce de la variable respuesta de volver a visitar la ciudad (*sí* o *no*) con los factores motivacionales (*push* y *pull*), se aprecia que algunas variables intentan mostrar respuestas hacia valores positivos de volver a visitar la ciudad, por ejemplo, los factores tipo *pull* 1.2, 4.2, 7.3 y 7.4 (Figura 7), sin embargo, los patrones no son tan claros, similar al caso anterior. Ahora bien, en el tema de las motivaciones *push* (Figura 8), la tendencia parece apuntar hacia la respuesta *no* en la mayoría de los casos; aquellas que se acercan más hacia la respuesta *sí* son la 1.4, la 4.1 y la 4.2.

Figura 5. Relación entre las variables *push* y la satisfacción con la ciudad



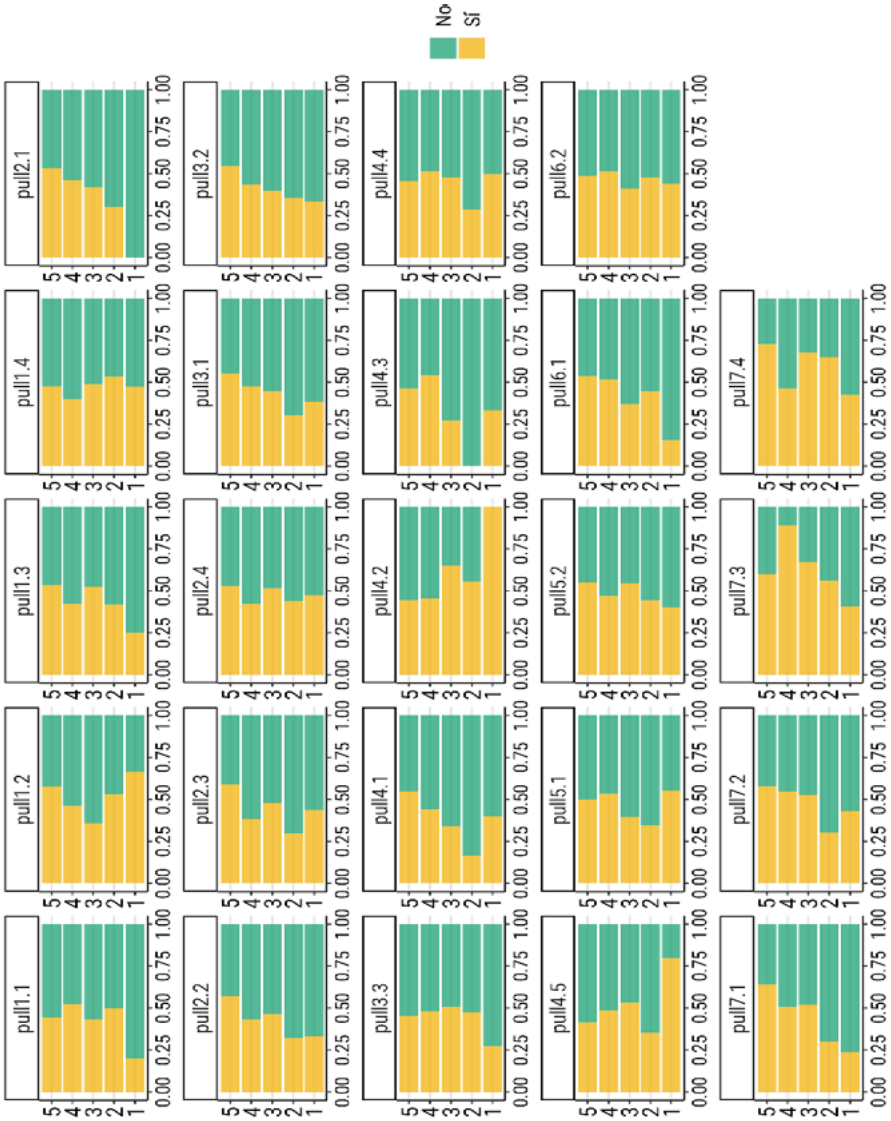
Nota. Elaboración propia.

Figura 6. Relación entre las variables *pull* y la satisfacción con la ciudad



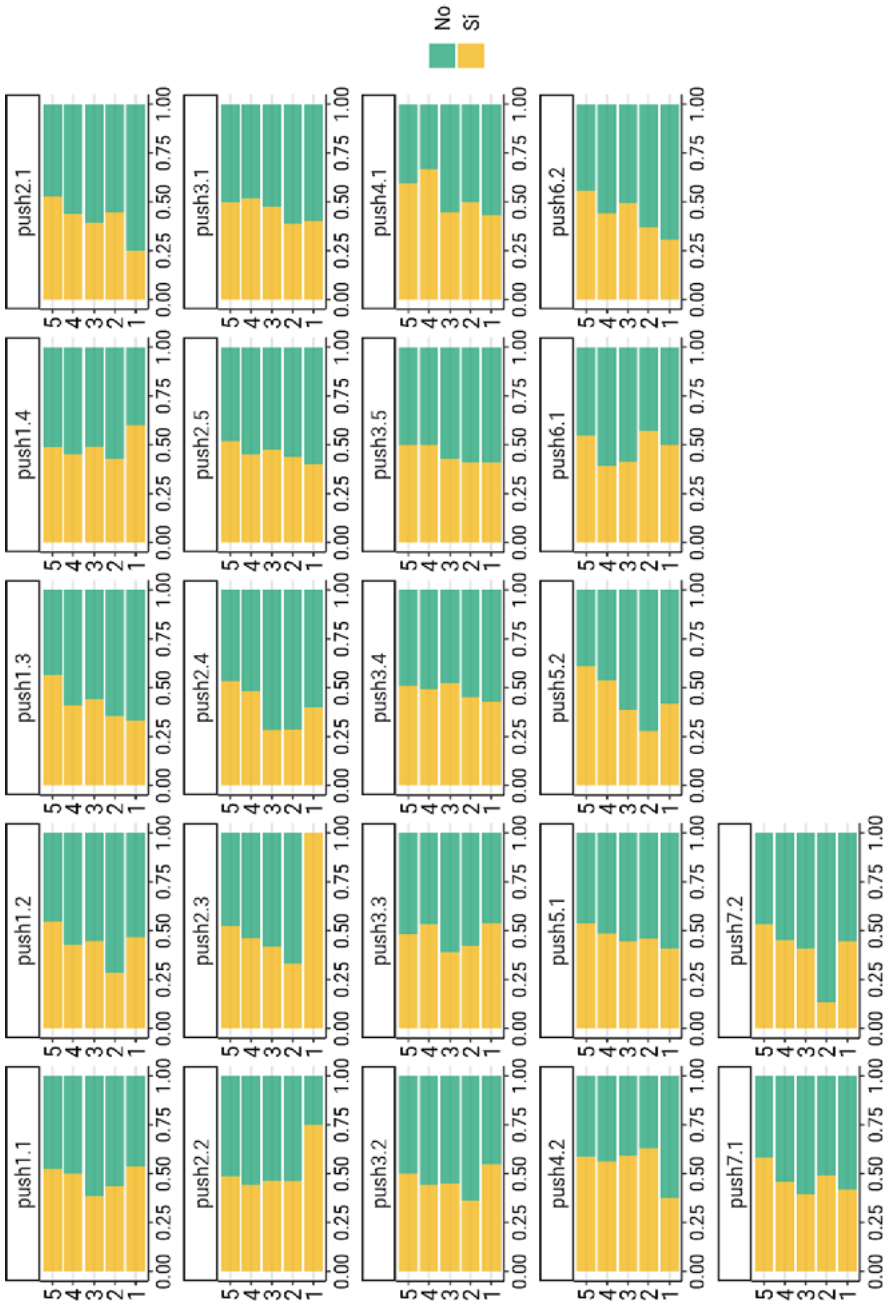
Nota. Elaboración propia.

Figura 7. Relación entre las variables *push* y volver a visitar la ciudad



Nota. Elaboración propia.

Figura 8. Relación entre las variables *pull* y volver a visitar la ciudad



Nota. Elaboración propia.

4.2 Desempeño de los modelos

Las Figuras 9 y 10 contienen las cuatro medidas de desempeño de los modelos para la predicción de las variables de satisfacción con la ciudad y volver a visitar la ciudad, respectivamente. La representación de los valores se ordena de acuerdo con el resultado del indicador de *accuracy*; en el eje de las ordenadas se grafica el valor obtenido con la métrica de desempeño específica. Cada color indica un modelo y en cada cuadrante se observa una métrica específica. Las gráficas de los modelos se apoyan en las Tablas 3 y 4 para cada variable que se va a explicar. En adelante, se describen los resultados de las predicciones. La representación gráfica y las tablas con el desempeño de los modelos son la forma estándar de reportar los resultados en aprendizaje estadístico cuando se comparan modelos predictivos con el fin de escoger el mejor.

Los modelos de *machine* evidencian una capacidad predictiva de la variable de satisfacción con la ciudad, que se encuentra entre 0,58 y 0,63, medida por la métrica de precisión (*accuracy* en inglés), con el modelo de árboles aleatorios como el mejor. Incluso evaluando la métrica de *Mn log loss* y la precisión en la curva *ROC*, los árboles aleatorios tienen mejor desempeño (ver Tabla 3 y Figura 9). Valores similares de capacidad predictiva se observan para la variable de volver a visitar la ciudad, cuyos valores se encuentran entre 0,58 para el peor desempeño y 0,62 para el mejor modelo, esto bajo la métrica de precisión; en este caso, el mejor modelo fue el algoritmo de las máquinas de soporte vectorial (ver Tabla 4 y Figura 10).

Si bien se pueden considerar los resultados como modestos, lo importante de la medición realizada es proponer una línea base para la comprensión de variables subjetivas, como la satisfacción con la ciudad y el volver a visitarla, puesto que son variables que presentan una compleja relación causal de los valores que adoptan, como se ha mostrado en recientes estudios en el campo de la administración y los negocios, donde la predicción para una de las clases se hace difusa (Żbikowski & Antosiuk, 2021). Las variables que se han considerado importantes para la atracción del turista en modelos clásicos y que han sido usadas en esta estimación presentan resultados moderados, incluso si se utilizan modelos robustos de aprendizaje estadístico, bajo optimización de hiperparámetros. Un asunto de interés debe orientarse a revisar las discusiones teóricas en asocio con las mediciones empíricas donde se incluya un mayor volumen de datos, como se ha resaltado en estudios de colaboración masiva sobre la modelación de aprendizaje estadístico e inteligencia artificial para predecir variables de corte subjetivo (Salganik et al., 2020).

Además, se debe considerar que las variables de corte subjetivo que representan la elección de los individuos tienen sesgos marcados (Kahneman et al., 2021) entre estos por la imposibilidad de materializar el constructo de la satisfacción de manera homogénea en cada persona que responde el cuestionario. En contraste, las decisiones se toman de manera intuitiva, muchas dependientes del estado de ánimo; las realiza lo que se ha llamado el sistema 1 para la toma de decisiones, que opera de manera rápida y automática y, en ocasiones, sin esfuerzo (Kahneman, 2011). Sin embargo, para la toma de decisiones, como lo ha expresado Angus Deaton, es mejor que una persona se encuentre feliz o satisfecha, no lo contrario (2013). En este sentido, los resultados de los modelos se pueden considerar una línea base de análisis para explicar los patrones similares en la satisfacción con el

destino turístico que declaran las personas y la posibilidad de volver a visitarlo, en función de las variables de *push y pull*.

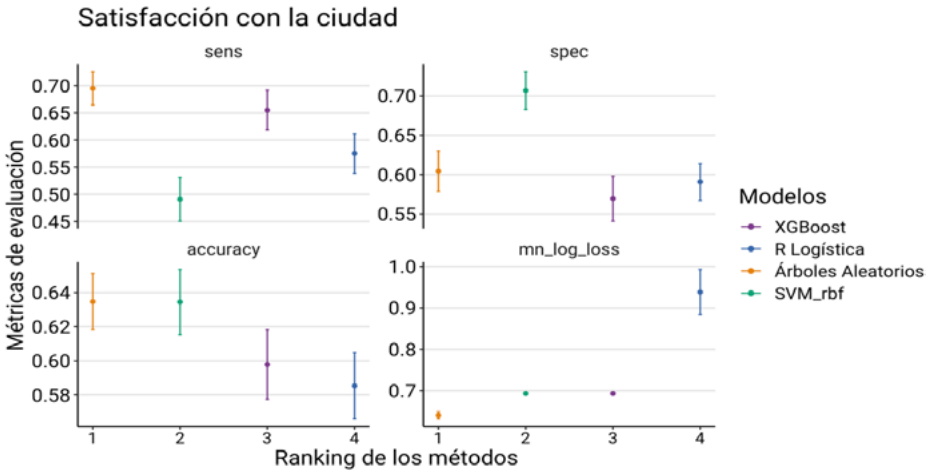
Cabe resaltar que, para la variable de satisfacción con la ciudad, los modelos entregan una información importante con respecto a la sensibilidad y la especificidad. El modelo de árboles aleatorios tiene una mayor capacidad para identificar a las personas que no están satisfechas con la ciudad (sensibilidad del 69 %), en cambio, el modelo de máquinas de soporte vectorial tiene más capacidad para identificar a quienes se encuentran satisfechos (especificidad del 70 %) (Tabla 3 y Figura 9). Con un método de ensamblaje de modelos (*model staking*) se podría aumentar la capacidad predictiva, en el caso particular, tomando los árboles aleatorios y las máquinas de soporte vectorial, un experimento para futuros trabajos. En los resultados, los modelos de XGBoost y LR presentan los peores desempeños; en este caso, los trabajos empíricos en el área del turismo que usan el modelo de regresión logística como único modelo de predicción (Huang et al., 2018; Jang & Cai, 2002; Lam-González et al., 2019) pueden migrar a nuevos modelos de aprendizaje de máquinas para mejorar la precisión de la estimación.

Tabla 3. Desempeño de los modelos para la variable "satisfacción con la ciudad".

Modelo	*Acc*	*Roc Auc*	*Sens*	*Spec*
RF	0,63	0,69	0,69	0,60
SVM	0,63	0,57	0,49	0,70
XGBoost	0,59	0,67	0,65	0,57
LR	0,58	0,63	0,57	0,59

Nota. Elaboración propia.

Figura 9. *Ranking* de los modelos estimados para la variable "satisfacción con la ciudad"



Nota. Elaboración propia.

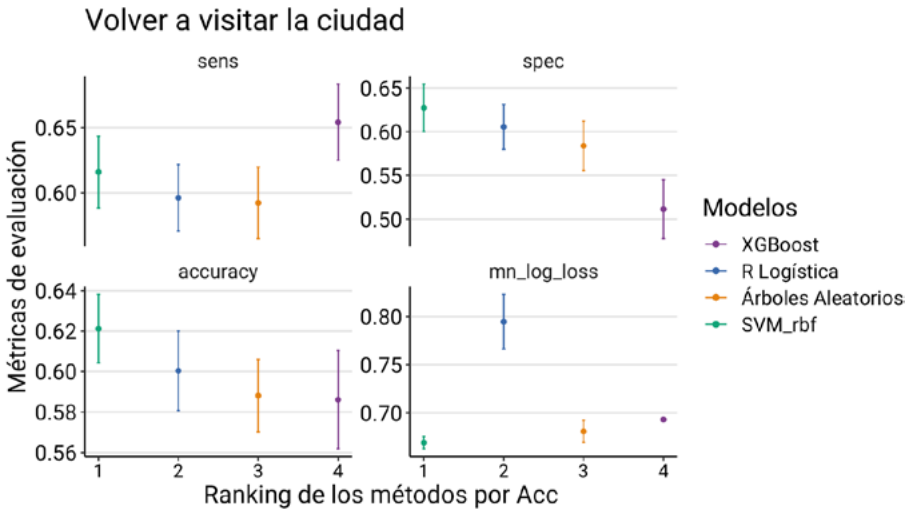
En el caso de la variable de volver a visitar la ciudad no se presentan ventajas de los modelos para predecir una clase u otra, es decir, el algoritmo de máquinas de soporte vectorial tiene los mejores resultados para predecir especificidad y sensibilidad. Los resultados son de 0,61 para sensibilidad y de 0,62 para especificidad (Tabla 4 y Figura 10).

Tabla 4. Desempeño de los modelos para la variable "volver a visitar la ciudad"

Modelo	*Acc*	*Roc Auc*	*Sens*	*Spec*
SVM	0,62	0,65	0,61	0,62
LR	0,60	0,62	0,59	0,60
RF	0,58	0,61	0,59	0,58
XGBoost	0,58	0,58	0,65	0,51

Nota. Elaboración propia.

Figura 10. *Ranking* de los modelos estimados para la variable "volver a visitar la ciudad"



Nota. Elaboración propia.

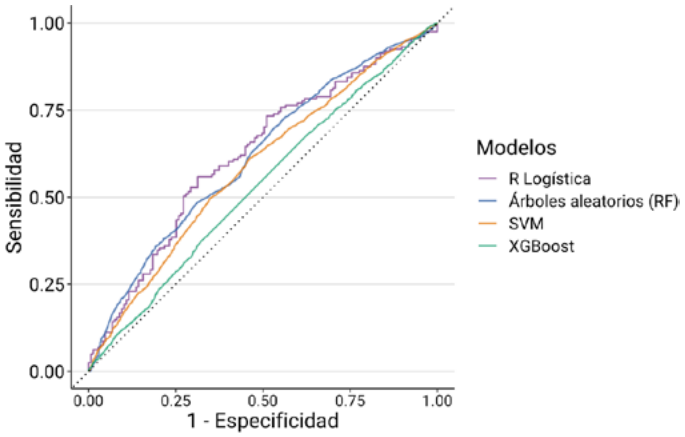
El desempeño de los modelos, medido con la curva *ROC*, se aprecia en la Figura 11. Se evidencian las curvas de los modelos entrenados tanto para la variable respuesta "volver a visitar la ciudad" (a) como para la variable "satisfacción con la ciudad" (b). Como se evidenció en las métricas de desempeño, los modelos con mejor capacidad para predecir el volver a visitar la ciudad son las máquinas de soporte vectorial, los árboles aleatorios y la regresión logística. Debido a que dichos modelos son similares en desempeño, para la evaluación de las variables de importancia se usa el modelo de regresión logística, que es menos complejo y permite evidenciar las variables que más inciden en la estimación. Cabe resaltar que el modelo de máquinas de soporte vectorial tiene un comportamiento

más satisfactorio para la predicción de datos no vistos en el entrenamiento, pues el valor de la precisión por curva *ROC* está por encima de 0,70.

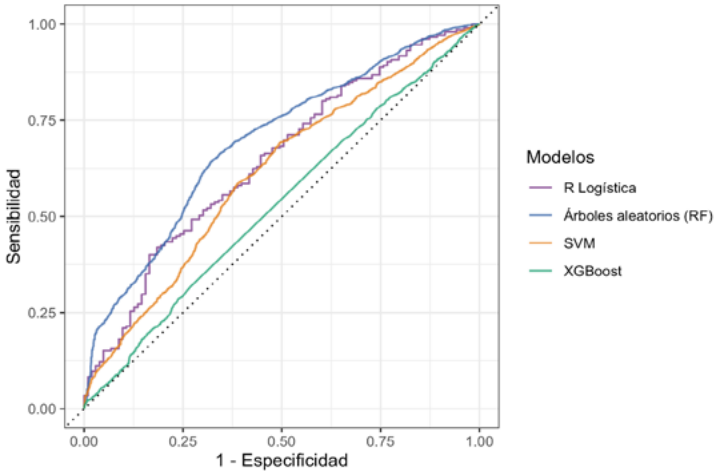
En cuanto a la variable de satisfacción con la ciudad, medido con curva *ROC*, se observa un desempeño mejor del modelo de árboles aleatorios, por consiguiente, es el usado para evaluar las variables de relevancia. Es importante anotar que los modelos presentan desempeños poco satisfactorios; aunque se usan algoritmos robustos, los resultados muestran que el problema de las motivaciones es complejo de modelar con las variables propuestas. En este sentido, los resultados se pueden tomar como un punto de partida para la discusión sobre las motivaciones de los turistas que los llevan a volver a visitar la ciudad y a quedar satisfechos con su viaje.

Figura 11. Curvas ROC en los modelos entrenados

(a) Volver a visitar la ciudad



(b) Satisfacción con la ciudad

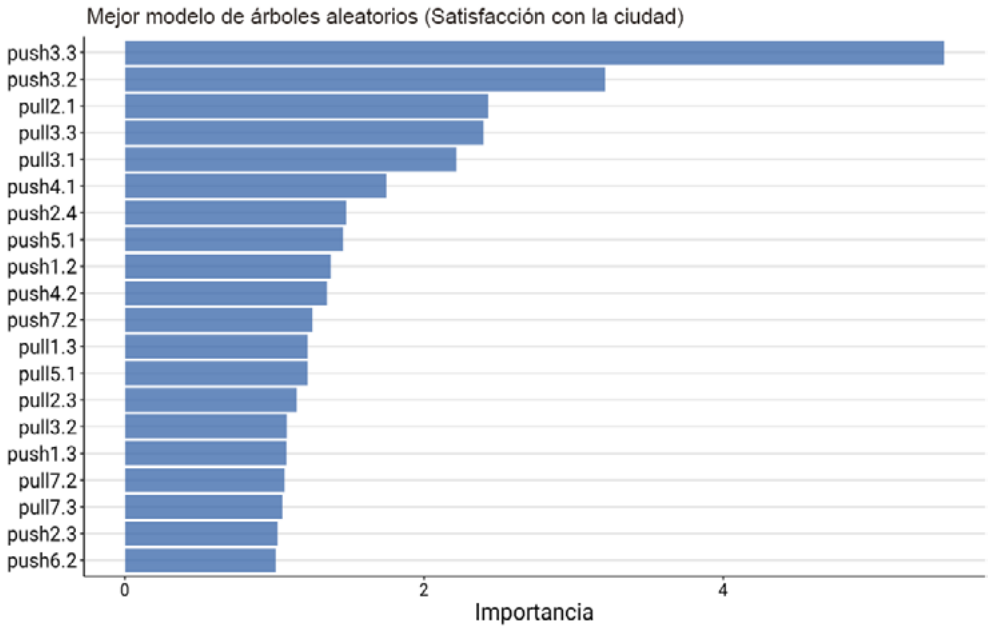


Nota. Elaboración propia.

4.3 Variables de importancia

En la modelación de la satisfacción con la ciudad, las cinco variables más importantes en el proceso de predicción incluyen dos de *push* (3.3 y 3.2) y tres de *pull* (2.1, 3.3 y 3.1) (Figura 12). Las variables están relacionadas con la dimensión latente de relajación y el logro en *push*, también con las actividades para realizar en el destino, la facilidad para recorrerlo y su clima. Al comparar los resultados del modelo con otros similares de la literatura sobre el tema en los que se explica la relación existente entre la satisfacción y la lealtad mediada por factores motivacionales, se encuentra que, tomando como referencia los factores de *pull,* el modelo de árboles aleatorios arrojó como los más relevantes la facilidad para recorrer la ciudad (*pull* 3.1) y el tema de la seguridad (*pull* 3.3), resultados que se pueden asociar a los exploratorios alcanzados en Chi y Qu (2008), donde el porcentaje más alto de la varianza explicada (14,74 %) se debe a estos factores.

Figura 12. Variables más importantes para explicar la satisfacción con la ciudad (árboles aleatorios)



Nota. Elaboración propia.

Entretanto, para el caso de las motivaciones de *push* se obtuvo que ir al lugar que no han visitado los amigos (*push* 3.2) y acumular experiencias de viaje para el futuro (*push* 3.3) se pueden relacionar con los resultados de Yoon y Uysal (2005), en los cuales dichos factores aparecen en el tercer escalón de importancia de factores explicativos, con una capacidad del 7,63 %; no obstante, es importante considerar que los factores de emocionante y conocimiento se ubicaron como los más importantes en el trabajo de Yoon y Uysal, pero en nuestro caso aparecen en los niveles medios de importancia (primeros 10 lugares). Posiblemente estas diferencias se deban a que aquí se estimaron los efectos

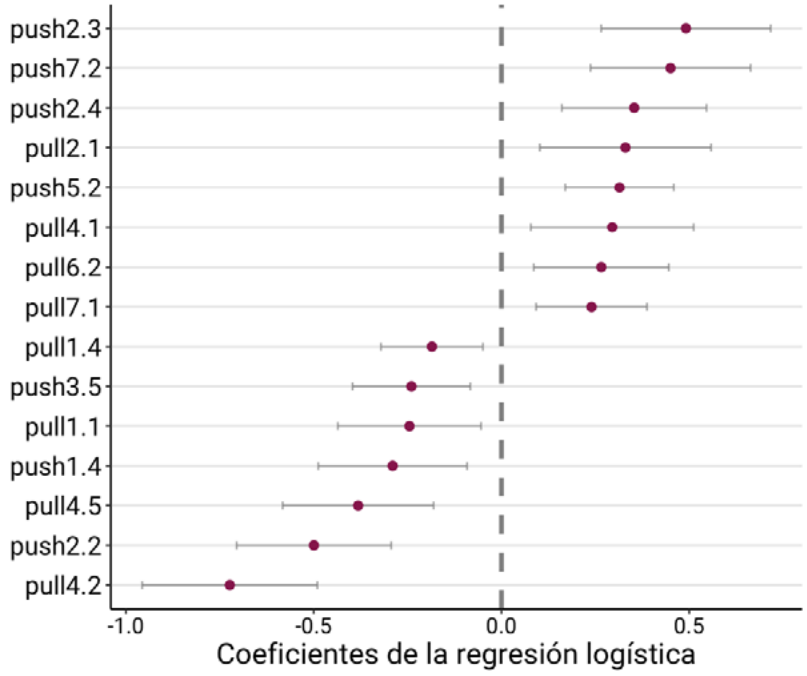
conjuntos de las dimensiones *push* y *pull,* además, se usaron métodos de estimación con algoritmos de aprendizaje estadístico; en cambio, en los artículos sobre la temática por lo general se encuentran las estimaciones de forma separada y usando el método de ecuaciones estructurales.

Aunque el modelo de máquinas de soporte vectorial presenta el mejor resultado de desempeño, medido por el *accuracy,* el modelo de regresión logística (*glm*) es el segundo mejor en desempeño, con una distancia de dos puntos. Por consiguiente, como el modelo logístico es de menos complejidad que el de máquinas de soporte vectorial y permite la interpretación directa de las variables por medio de su nivel de significancia para explicar la variable de volver a visitar la ciudad, se elige el modelo de regresión logística para este fin. Así, los coeficientes de la regresión logística con mayores valores se presentan en la Figura 13, los cuales dan una idea de las variables de importancia para la estimación del valor de volver a visitar la ciudad, que son significativas a un valor-*p* menor a 0,05.

En este sentido, las motivaciones que se relacionan en una mayor magnitud de forma positiva son visitar lugares históricos (*push* 2.3), viajar a precios económicos (*push* 7.2) y conocer gente nueva (*push* 2.4); estos resultados también se destacaron en el trabajo de Yoon y Uysal (2005) para la primera de estas. Ahora bien, aparecen asimismo algunas motivaciones que aportan de forma negativa: pueblos antiguos e históricos (*pull* 4.5), probar nuevos alimentos (*push* 2.2) y variedad de patrimonio cultural diferente del propio (*pull* 4.2), variables que mostraron un peso de importancia menor en los trabajos de Jang y Cai (2002) y Chi y Qu (2008).

El proceso metodológico propuesto en el presente artículo está en línea con las implementaciones que se vienen realizando en el área de turismo (Oh & Lee, 2021). En cada problema se requiere probar diferentes modelos con procesos de optimización de hiperparámetros que permitan identificar la capacidad que tienen para predecir y explicar comportamientos. En los resultados se observa que los algoritmos de los árboles aleatorios y las máquinas de soporte vectorial son los dos modelos con mejor desempeño, contrario a los resultados de Oh y Lee (2021) en el área del turismo, donde el mejor modelo es el *Light Gradient Boosting* para un problema de regresión. En este sentido, los nuevos métodos son una ventana de oportunidad para comprender los problemas de turismo. Los métodos de aprendizaje estadístico y la inteligencia artificial aplicados a los problemas de turismo (Egger, 2022) abren la posibilidad de diversificar las metodologías concentradas en los modelos de ecuaciones estructurales (Yoo et al., 2018; Yoon & Uysal, 2005); así, se incentiva el uso de nuevos algoritmos que posibiliten descubrir los patrones de los datos y aportar a la toma de decisiones informadas.

Figura 13. Variables más importantes para explicar el volver a visitar la ciudad (regresión logística)



Nota. Elaboración propia.

5. Conclusiones, recomendaciones y limitaciones

En este estudio se abordó un problema de clasificación desde la perspectiva del aprendizaje estadístico. Se realizó la predicción de dos variables: (1) la satisfacción con la ciudad de Medellín y (2) la posibilidad de volver a visitar la ciudad. Se estimaron cuatro modelos: regresión logística, máquinas de soporte vectorial, árboles aleatorios e incremento extremo del gradiente. La clasificación se realizó tomando como referencia un conjunto de factores motivacionales tanto internos, denominados *push*, como externos, llamados *pull*.

En la predicción de la satisfacción con la ciudad, los resultados en el conjunto de entrenamiento evidencian que el mejor modelo fue el de los árboles aleatorios, que logró un 0,63 de precisión, seguido por el de las máquinas de soporte vectorial con *Kernel* gaussiano, con un 0,62 de precisión. Al estimar los modelos en los datos de prueba, los desempeños fueron modestos y solo se obtuvieron valores de precisión del 0,53. Cabe anotar que en el presente estudio se evidenció mejor precisión de clasificación en los algoritmos de RF y SVM que en el de XGBoost, aunque este último ha sido uno de los principales para la precisión de las competencias de aprendizaje estadístico.

La ventaja de los árboles aleatorios permitió identificar las variables que más aportaron a la composición del modelo. Las motivaciones de sensación de seguridad, la posibilidad

de practicar actividades deportivas e ir a lugares que los amigos no han visitado explican la satisfacción del viajero, principalmente en la categoría *aceptable*. Del análisis se desprende que esos factores elevan la ponderación de la estadía comparada con otros viajes realizados por el turista.

Para el caso de la variable de volver a visitar la ciudad, el mejor modelo fue el de las máquinas de soporte vectorial, que logró un 62 % de capacidad para predecir el comportamiento de la decisión de las personas. Las variables que tuvieron más influencia para volver a visitar los lugares, dado el valor de desempeño del modelo de regresión logística (que fue el segundo mejor desempeño), fueron visitar lugares históricos, viajar a precios económicos y conocer gente nueva. Si se evalúa el aporte negativo a la variable respuesta de volver a visitar la ciudad, aparecen las siguientes variables: pueblos antiguos, probar nuevos alimentos y variedad de patrimonio cultural diferente del propio.

Los resultados obtenidos en el presente estudio se consideran como línea base para el desarrollo de modelos de aprendizaje estadístico para la explicación de la motivación de los turistas a visitar un destino, un área de estudio prometedora por los constantes avances en la modelación y los diferentes algoritmos que pueden ayudar a la toma de decisiones informada. La estimación de diversos modelos sirve como herramienta de planificación para desarrollar propuestas de atracción turística por medio de estrategias de *marketing* o de planificación de recursos del destino. Ahora bien, los modelos presentan desempeños diferentes, por ejemplo, los árboles aleatorios tienen mejor desempeño para predecir los turistas que califican como agradable el destino, por su parte, las máquinas de soporte vectorial tienen mejor capacidad explicativa para predecir la clase contraria. En el caso de la predicción de la variable de volver a visitar la ciudad, los modelos fueron equilibrados en su capacidad para predecir la clase positiva o negativa. Es de anotar que los mejores resultados se obtuvieron en el conjunto de datos de entrenamiento, puesto que las pruebas de generalización del modelo en los datos de prueba no logran resultados satisfactorios.

El estudio tiene limitación en el conjunto de datos, dado que se redujeron los grados de libertad en el proceso de entrenamiento; es limitado porque el algoritmo no logra aprender la variabilidad de los datos para obtener mejores resultados. Igualmente, las variables asociadas como relaciones causales de la elección del destino turístico requieren más discusión teórica y empírica para poder precisar las variables predictoras. En trabajos futuros se espera probar con redes neuronales para datos categóricos, aunque la gran limitación es el número de observaciones disponibles para resolver el problema; también se espera probar ensamblajes de modelos y modelos tipo *stack*, que aprovechan lo mejor de cada modelo en una agrupación de ellos.

Referencias

Abe, S. (2005). *Support vector machines for pattern classification*. Vol. 2. Springer. https://doi.org/10.1007/1-84628-219-5

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. *International Journal of Hospitality Management*, *80*, 52-77. https://doi.org/10.1016/j.ijhm.2019.01.003

Albayrak, T. & Caber, M. (2018). Examining the relationship between tourist motivation and satisfaction by two competing methods. *Tourism Management*, *69*, 201-13. https://doi.org/10.1016/j.tourman.2018.06.015

Bloom, J. (2004). Tourist market segmentation with linear and non-linear techniques. *Tourism Management*, *25*(6), 723-733. https://doi.org/10.1016/j.tourman.2003.07.004

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A*:1010933404324*

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Association for Computing Machinery (Ed), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen, Y., Zhang, H., & Qiu, L. (2013). Review on tourist satisfaction of tourism destinations. *Journal of System and Management Sciences*, *3*(1), 74-86. http://www.aasmr.org/jsms/Vol3/No1/JSMS _Vol3_No1_8.pdf

Chi, C. & Qu, H. (2008). Examining the structural relationships of destination image, tourist satisfaction and destination loyalty: An integrated approach. *Tourism Management*, *29*(4), 624-636. https://doi.org/10.1016/j.tourman.2007.06.007

Correia, A., Kozak, M., & Ferradeira, J. (2013). From tourist motivations to tourist satisfaction. *International Journal of Culture, Tourism and Hospitality Research*, *7*(4), 411-424. https://doi.org/10.1108/IJCTHR-05-2012-0022

Dean, D. & Suhartanto, D. (2019). The formation of visitor behavioral intention to creative tourism: The role of push–pull motivation. *Asia Pacific Journal of Tourism Research*, *24*(5), 393-403. https://doi.org/10.1080/10941665.2019.1572631

Deaton, A. (2013). *The great escape*. Princeton University Press.

Do Valle, P., Silva, J., Mendes, J., & Guerreiro, M. (2006). Tourist satisfaction and destination loyalty intention: A structural and categorical analysis. *International Journal of Business Science & Applied Management*, *1*(1), 25-44. https://acortar.link/C2k3Jh

Egger, R. (2022). Machine learning in tourism: A brief overview. En R. Egger (Ed.), *Applied data science in tourism: Interdisciplinary approaches, methodologies & applications* (pp. 85-107). Springer. https://doi.org/10.1007/978-3-030-88389-8

Fodness, D. (1994). Measuring tourist motivation. *Annals of Tourism Research*, *21*(3), 555-581. https://doi.org/10.1016/0160-7383(94)90120-1

Ghaderi, Z., Hatamifar, P., & Khalilzadeh, J. (2018). Analysis of tourist satisfaction in tourism supply chain management. *Anatolia: An International Journal of Tourism and Hospitality Research*, *29*(3), 433-444. https://doi.org/10.1080/13032917.2018.1439074

Gil-León, J., Gutiérrez-Ayala, J., & Ramírez-Hernández, E. (2021). El papel del turismo patrimonial en el índice de competitividad turística regional de Colombia: una evaluación de las relaciones mediante PLS-PM. *Revista Escuela de Administración de Negocios*, (90), 169-192. https://doi.org/10.21158/01208160.n90.2021.2973

Guerra-Montenegro, J., Sánchez-Medina, J., Laña, I., Sánchez-Rodríguez, D. Alonso-González, I., & Del Ser, J. (2021). Computational Intelligence in the hospitality industry: A systematic literature review and a prospect of challenges. *Applied Soft Computing*, *102*, 107082. https://doi.org/10.1016/j.asoc.2021.107082

Huang, Z., Kong, Y., & Zhou, C. (2018). A study on relationship between sports tourism motivation and tourists' re-visiting intention: Based on Logistic Model. *Advances in Social Science, Education and Humanities Research: Proceedings of the 2nd International Conference on Economics and Management, Education, Humanities and Social Sciences (EMEHSS 2018)*, *151*, 54-61. https://doi.org/10.2991/emehss-18.2018.13

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer. https://www.stat.berkeley.edu/users/rabbee/s154/ISLR _First_Printing.pdf

Jang, S. & Cai, L. (2002). Travel motivations and destination choice: A study of British outbound market. *Journal of Travel & Tourism Marketing*, *13*(3), 111-133. https://doi.org/10.1080/10548400209511570

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. Hachette Book Group.

Kozak, M. (2001). Comparative assessment of tourist satisfaction with destinations across two nationalities. *Tourism Management*, *22*(4), 391-401. https://doi.org/10.1016/S0261-5177(00)00064-9

Kuhn, M. & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org

Kwon, W., Lee, M., & Back, K-J. (2020). Exploring the underlying factors of customer value in restaurants: A machine learning approach. *International Journal of Hospitality Management*, *91*, 102643. https://doi.org/10.1016/j.ijhm.2020.102643

Lam-González, Y., León, C., & De León, J. (2019). Coopetition in maritime tourism: Assessing the effect of previous islands' choice and experience in tourist satisfaction. *Sustainability*, *11*(22), 6334. https://doi.org/10.3390/su11226334

Lee, T. (2009). A structural model to examine how destination image, attitude & motivation affect the future behavior of tourists. *Leisure Sciences*, *31*(3), 215-236. https://doi.org/10.1080/01490400902837787

Lee, G., O'Leary, J., Lee, S., & Morrison, A. (2002). Comparison and contrast of push and pull motivational effects on trip behavior: An application of a Multinomial Logistic Regression Model. *Tourism Analysis*, *7*(2), 89-104. https://doi.org/10.3727/108354202108749970

Li, W., Xu, S., & Meng, W. (2009). A support vector machines method for tourist satisfaction degree evaluation. En IEEE Computer Society (Ed.), *2009 6th International Conference on Service Systems and Service Management* (pp. 883-887). IEEE. https://doi.org/10.1109/ICSSSM.2009.5175007

Luna-Cortés, G. (2020). Análisis de la percepción de los estadounidenses que visitan Colombia. Un modelo de ecuaciones estructurales. *Estudios y Perspectivas en Turismo*, *29*(1), 51-71. https://acortar.link/WATPWR

Mansfeld, Y. (1992). From motivation to actual travel. *Annals of Tourism Research*, *19*(3), 399-419. https://doi.org/10.1016/0160-7383(92)90127-B

Oh, H., Kim, B. Y., & Shin, J. H. (2004). Hospitality and tourism marketing: Recent developments in research and future directions. *International Journal of Hospitality Management*, *23*(5), 425-447. https://doi.org/10.1016/j.ijhm.2004.10.004

Oh, H. & Lee, S. (2021). Evaluation and interpretation of tourist satisfaction for local Korean festivals using explainable AI. *Sustainability*, *13*(19), 10901. https://doi.org/10.3390/su131910901

Olague de la Cruz, J. (2015). *La imagen del destino y la motivación de viaje como determinantes de la satisfacción y lealtad del turismo urbano de ocio en Monterrey, México* (Tesis doctoral, Universidad Autónoma de Nuevo León). Repositorio Académico Digital UANL. http://eprints.uanl.mx/9248/

Prebensen, N., Skallerud, K., & Chen, J. (2010). Tourist motivation with sun and sand destinations: Satisfaction and the wom-effect. *Journal of Travel & Tourism Marketing*, *27*(8), 858-873. https://doi.org/10.1080/10548408.2010.527253

Salganik, M., Lundberg, I., Kindel, A., Ahearn, C., Al-Ghoneim, K., Almaatouq, A., Altschul, D., Brand, J., Bohme, N., Compton, R., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B., Jahani, E., Kashyap, R., Kirchner, A. ... & McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *PNAS: Proceedings of the National Academy of Sciences*, *117*(15), 8398-8403. https://doi.org/10.1073/pnas.2118703118

San Martín, H., Herrero, A., & García, M. (2019). An integrative model of destination brand equity and tourist satisfaction. *Current Issues in Tourism*, *22*(16), 1992-2013. https://doi.org/10.1080/13683500.2018.1428286

Schonlau, M. & Zou, R. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3-29. https://doi.org/10.1177%2F1536867X20909688

Song, Y., Wang, R., Fernández, J., & Li, D. (2021). Investigating sense of place of the Las Vegas Strip using online reviews and machine learning approaches. *Landscape and Urban Planning*, *205*, 103956. https://doi.org/10.1016/j.landurbplan.2020.103956

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.

Villamediana-Pedrosa, J, Vila-López, N., & Küster-Boluda, I. (2020). Predictors of tourist engagement: Travel motives and tourism destination profiles. *Journal of Destination Marketing & Management*, *16*, 100412. https://doi.org/10.1016/j.jdmm.2020.100412

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino, L., François R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin, T., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open-Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yoo, C., Yoon, D., & Park, E. (2018). Tourist motivation: An integral approach to destination choices. *Tourism Review*, *73*(2), 169-185. https://doi.org/10.1108/TR-04-2017-0085

Yoon, Y. & Uysal, M. (2005). An examination of the effects of motivation and satisfaction on destination loyalty: A structural model. *Tourism Management*, *26*(1), 45-56. https://doi.org/10.1016/j.tourman.2003.08.016

Yu, L. & Goulden, M. (2006). A comparative analysis of international tourists' satisfaction in Mongolia. *Tourism Management*, *27*(6), 1331-1342. http://dx.doi.org/10.1016/j.tourman.2005.06.003

Żbikowski, K. & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), 102555. https://doi.org/10.1016/j.ipm.2021.102555