Aprendizaje supervisado para clasificación de experiencias de viajes turísticos en Colombia

Gonzales Matinez Edwin Fernando[1], Ortegon Palomino Frank Sebastian[2], and Salamanca Bernal Julian Andres[2]

[2]Fundación Universitaria Los Libertadores

Resumen

En Colombia, según la Asociación Colombiana de Agencias de Viajes y Turismo (ANATO), cerca de 2,2 millones de personas se desplazaron en el país como visitantes internos durante el primer semestre del 2022, cifra que representa un 10,4 % de la población total; En consecuencia las compañías que ofrecen viajes turísticos a nivel nacional se enfrentan a un entorno altamente competitivo. Para abordar este tema se realiza el análisis de los resultados, se encontró que el modelo de Regresión Logística es más efectivo en clasificar correctamente las experiencias de los usuarios. En conclusión, esta investigación brinda valiosa información para mejorar la toma de decisiones en las empresas de viajes turísticos y promoviendo la rentabilidad al ofrecer experiencias más satisfactorias a los usuarios.

Palabras clave— regresion logistica, naive bayes, machine learning, smote

Abstract

In Colombia, according to the Colombian Association of Travel and Tourism Agencies (ANATO), about 2.2 million people traveled in the country as domestic visitors during the first half of 2022, a figure that represents 10.4% of the total population; consequently, companies that offer tourist trips at the national level face a highly competitive environment. To address this issue, the analysis of the results is carried out, it was found that the Logistic Regression model is more effective in correctly classifying the experiences of users. In conclusion, this research provides valuable information to improve decision making in tourist travel companies and promote profitability by offering more satisfactory experiences to users.

Keywords— logistic regression, naive bayes, machine learning, smote

1. Introducción

El turismo es una industria importante en Colombia y ha experimentado un crecimiento constante en los últimos años. Con el objetivo de mejorar la calidad de los servicios turísticos y proporcionar una experiencia más satisfactoria a los turistas, muchas agencias de viajes están buscando soluciones tecnológicas para mejorar sus procesos.

En este contexto, se ha llevado a cabo una investigación para desarrollar un modelo de machine learning utilizando una base de datos de una agencia de turismo en Colombia. El objetivo de este modelo es clasificar la experiencia de los usuarios en servicios turísticos en función de variables como la temporada, la ubicación, el tipo de viaje, entre otras.

La importancia de esta investigación radica en la posibilidad de mejorar la planificación de los servicios turísticos y realizar ofertas a los usuarios con malas experiencias en sus viajes, lo que podría generar mayores ingresos para las agencias de viajes y mejorar la satisfacción de los turistas, así como mejorar la planificación de los servicios turísticos y ajustar la oferta a la demanda real.

El objetivo general de esta investigación es desarrollar un modelo de machine learning que permita clasificar la experiencia de los usuarios en servicios turísticos en Colombia. Para ello, se utilizará una base de datos proporcionada por una agencia de viajes y se aplicarán técnicas de aprendizaje automático para analizar los datos y construir el modelo.

En cuanto a los antecedentes, existen estudios previos que han utilizado técnicas de machine learning en la industria turística, pero pocos se han enfocado en la realidad colombiana. Por lo tanto, esta investigación busca llenar ese vacío y proporcionar información relevante para la industria turística en Colombia.

Teniendo en cuenta lo anterior, se realiza un estudio de antecedentes relacionados donde Se han llevado a cabo numerosos estudios en el campo de la minería de datos y el aprendizaje supervisado para predecir la experiencia en viajes turísticos. Estos estudios buscan comprender y anticipar los diferentes aspectos que influyen en la satisfacción y preferencias de los turistas, con el objetivo de proporcionar recomendaciones más precisas y personalizadas, mejorar la toma de decisiones en la planificación de viajes y optimizar la calidad de los servicios turísticos. A continuación, se presenta algunos estudios.

H. Zhang, Liu, y Li, utiliza técnicas de aprendizaje supervisado, específicamente soporte de maquina vectorial, para clasificar la experiencia de los viajeros. Para ello, recopilaron datos relacionados con las

preferencias, satisfacción y opiniones de los turistas en diferentes destinos turísticos. Por otro lado Buhalis y Michopoulou se centra en investigar las fuentes de información utilizadas por los turistas europeos al planificar y seleccionar destinos turísticos. El objetivo es comprender las necesidades de información de los turistas e identificaron diferencias significativas en el uso de fuentes de información según las características demográficas de los turistas, como la edad, el nivel educativo y la experiencia previa de viaje.

Por otro lado X. Zhang y Song se recopilaron grandes cantidades de reseñas de hoteles chinos de diversas fuentes en línea. Se utilizó una técnica de modelado de temas llamada Latent Dirichlet Allocation para identificar los temas subyacentes en las reseñas y asignar las palabras relevantes a cada tema y una vez que se identificaron los temas, se realizó un análisis de sentimiento para determinar el sentimiento general asociado con cada tema. Esto se logró mediante la clasificación de las palabras en positivas, negativas o neutrales

En resumen, estos estudios abordan diferentes aspectos relacionados con el turismo y la experiencia de los viajeros por medio de técnicas asociadas a la estadística.

2. Marco teórico

En esta sección se investiga la teoría sobre las técnicas de Machine Learning aplicadas en problemas con grandes volúmenes de información al clasificar los comentarios del comprador. La finalidad de esta primera parte del trabajo es describir los métodos de clasificación de la experiencia de los clientes con algoritmos de Machine Learning como herramienta metodológica para descubrir patrones en bases de datos de paquetes de viajes turísticos que a simple vista no son posibles de encontrar.

2.1. Machine Learning

Machine learning es una disciplina de las ciencias informáticas relacionados en el campo de la inteligencia artificial, así como también con el análisis predictivo, básicamente le permite a un algoritmo la capacidad de identificar y aprender de los datos para que de esta manera convierte la información en datos masivos y pueda elaborar predicciones, a medida que el algoritmo adquiere datos de entrenamiento, es posible ir generando modelos más precisos basados en datos, en otras palabras un modelo de marchine learning es la salida de información que se obtiene de entrenar un algoritmo con datos. IBM.

Los métodos de Machine Learning presentan diferentes categorías las cuales se presentan a continuación.

2.2. Aprendizaje no supervisado

Este algoritmo es entrenado utilizando un conjunto de datos que no tiene etiquetas, es decir, a este aprendizaje no se le dice lo que representan los datos, ya que la idea es que el algoritmo encuentre los patrones para que pueda entender el conjunto de datos. López.

2.3. Aprendizaje supervisado

El algoritmo aprende a partir de datos donde ya se encuentran etiquetados con la respuesta correcta, entre mayor es el conjunto de datos, el algoritmo puede aprender más sobre el tema relacionado con los datos. Una vez finalizado el entrenamiento, se le proporcionan nuevos datos, estos ya sin etiquetas de las respuestas correctas y el algoritmo de aprendizaje aprovecha la experiencia pasada que esté adquirió durante la etapa de entrenamiento para predecir un resultado. KeepCoding.

En este proyecto se va a utilizar modelos de aprendizaje supervisado, considerando que la base de datos empleada tiene una variable de respuestas donde se clasifica la experiencia del usuario como buena o mala. Teniendo en cuenta lo anterior, se va emplear los siguientes modelos:

2.3.1. Regresión logística

Método de análisis estadístico para predecir algún resultado binario, como por ejemplo un sí o un no, apoyado en observaciones de un conjunto de datos, este modelo predice una variable de datos dependiente, donde se analiza la relación entre una o más variables independientes existentes, adicional puede tener en cuenta múltiples criterios de entrada. Posibilita que los algoritmos empleados en las aplicaciones de aprendizaje automático clasifique los datos entrantes basado en datos históricos, de igual manera cada vez que llegan datos relevantes adicionales, el algoritmo mejora la predicción de clasificaciones dentro de los conjuntos de datos. La regresión logística ha sido una herramienta importante en la disciplina del Machine Learning. Universidad Piloto de Colombia.

2.3.2. Naive Bayes

Se asume que las variables predictoras son independientes entre sí, es decir, la presencia de una característica en un conjunto de datos es completamente independiente de la presencia de cualquier otra característica. estos algoritmos se hacen proporcionando una manera de calcular la probabilidad “posterior”del evento A, dada la probabilidad de eventos “anteriores”. Roman.

3. Metodología

Se ha optado por utilizar para el desarrollo del proyecto la metodología SEMMA debido a sus beneficios clave para abordar problemas complejos en el análisis de datos. Dado que el turismo en Colombia implica una amplia variedad de factores y variables interrelacionadas, es fundamental tener un enfoque estructurado que nos guíe a lo largo de todo el proceso. SEMMA nos brinda un marco claro y

sistemático, que abarca desde la comprensión del dominio turístico en Colombia y la identificación de los objetivos específicos del proyecto, hasta la implementación y evaluación de los modelos de aprendizaje supervisado.

Al seguir la metodología, podemos garantizar que estamos cubriendo todos los aspectos importantes del proyecto, desde la recolección y preparación de los datos turísticos relevantes, hasta la construcción y evaluación de modelos precisos y efectivos. En última instancia, la utilización de la metodología nos ayudará a maximizar el valor de los datos turísticos en Colombia y a obtener resultados confiables y accionables para impulsar la toma de decisiones y mejorar la industria turística en el país.

3.1. Datos

Los datos que se van a trabajar para el proyecto se obtienen de una base de datos de una empresa Colombiana que se dedica a la venta de paquetes turísticos para empresas a nivel nacional. Los datos están compuestos por 7731 registros de clientes del periodo de enero a diciembre del 2022, además de tener 16 variables que se explican a continuación :

Nombre Variable	Descripción Variable	Tipo De Variable
Ciudad Origen	La Ciudad De La Cual Proviene El Cliente.	Categórica
Edad Del Cliente	La Edad Del Cliente Que Adquiere El Paquete Turístico.	Entero
Género Del Cliente	El Género Del Cliente (masculino, Femenino U Otro).	Categórica
Número De Personas	El Número De Personas Que Viajan En El Paquete Turístico	Entero
Ciudad Destino	La Ciudad Hacia La Cual Se Dirige Para El Viaje Turístico.	Categórica
Duración Del Viaje	La Cantidad De Días Que Dura El Viaje Turístico	Entero
Tipo De Alojamiento	El Tipo De Alojamiento Proporcionado Durante El Viaje (hotel, Apartamento, Casa	Categórica
Transporte	El Medio De Transporte Utilizado Durante El Viaje (avión, Autobús, Carro Particular,etc.).	Categórica
Actividades	Las Actividades Incluidas En El Paquete Turístico (visitas A Lugares De Interés, Excursiones, Actividades Deportivas, Etc.).	Categórica
Mes Del Viaje	El Mes En El Cual Se Realiza El Viaje Turístico.	Entero
Costo Total	El Costo Total Del Paquete Turístico (Está En Pesos Colombianos).	Entero
Tipo De Paquete Turístico	El Tipo De Paquete Turístico Adquirido (todo Incluido, Solo Alojamiento, Etc.).	Categórica
Profesión Del Cliente	La Profesión U Ocupación Del Cliente.	Categórica
Estado Civil Del Cliente	El Estado Civil Del Cliente(soltero, Casado, Divorciado, Etc.).	Categórica
Motivación	La Motivación O Razón Principal Por La Cual Se Realiza El Viaje Turístico.	Categórica
Comentarios	Comentarios Adicionales O Sugerencias Proporcionadas Por Los Clientes.	Categórica

Tabla 1: Tabla descripción de los datos

3.2. Exploración de datos

En esta fase, se llevan a cabo las actividades para comprender los datos y la relación entre ellos. Un proceso clave es el análisis de tablas de contingencia o correlación tanto de las variables cualitativas como de las cuantitativas. En este proceso se observa la relación de las variables entre si para comprender su distribución. La visualización de datos se usa mucho para ayudar a discernir mejor la información.

3.2.1. Variables Cuantitativas

En la figura 1 se realiza las correlaciones entre las variables $edad$, $num\_personas$, $duracion$ y $costo\_total$. Estas correlaciones sugieren que la cantidad de personas tiene cierta influencia en la duración y el costo total, mientras que la edad no parece estar fuertemente relacionada con las otras variables. Sin embargo, es importante considerar que estas correlaciones son lineales y no reflejan necesariamente relaciones causales entre las variables.

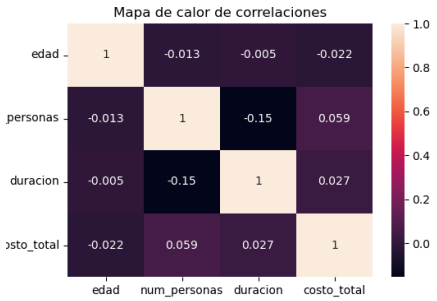

Figura 1: Correlación de las variables cuantitativas.

3.2.2. Variables Cualitativas

Para las variables cualitativas se realiza tablas de contingencia para analizar la frecuencia y de ese modo observar la distribución conjunta de las categorías de ambas variables, lo que permite examinar la asociación o dependencia entre ellas. Además se realiza el *p-valor* y el estadístico *chi-cuadrado* para saber si hay una asociación significativa entre las variables. Con lo dicho anteriormente a continuación se muestran las variables con mayor relación entre si :

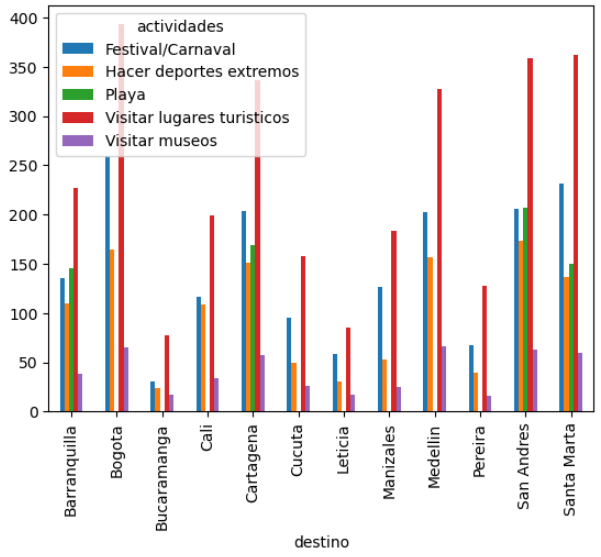

Figura 2: Correlación variables destino y actividades

La figura 2 muestra la participación en actividades turísticas en diferentes destinos. Bogotá destaca como el destino con mayor participación en festivales, visitas a lugares turísticos y museos, mientras que San Andrés lidera en actividades de playa y deportes extremos. Otros destinos como Cartagena y Santa Marta también muestran una buena participación en varias categorías. Por otro lado, destinos como Bucaramanga, Cali, Cúcuta, Leticia, Manizales y Pereira tienen una participación más limitada en estas actividades. En general, Bogotá se destaca como un destino turístico versátil, ofreciendo una amplia gama de opciones para los visitantes. Además se realizó una prueba de chi-cuadrado y los resultados mostraron una fuerte evidencia en contra de la hipótesis nula. El estadístico chi-cuadrado obtenido fue de 788.95, lo que indica una significativa discrepancia entre los datos observados y esperados. Además, el p-valor obtenido fue extremadamente bajo (3.26e-137), lo que refuerza la idea de que la discrepancia no se debe al azar. Estos resultados sugieren una relación significativa entre las variables analizadas en la prueba de chi-cuadrado.

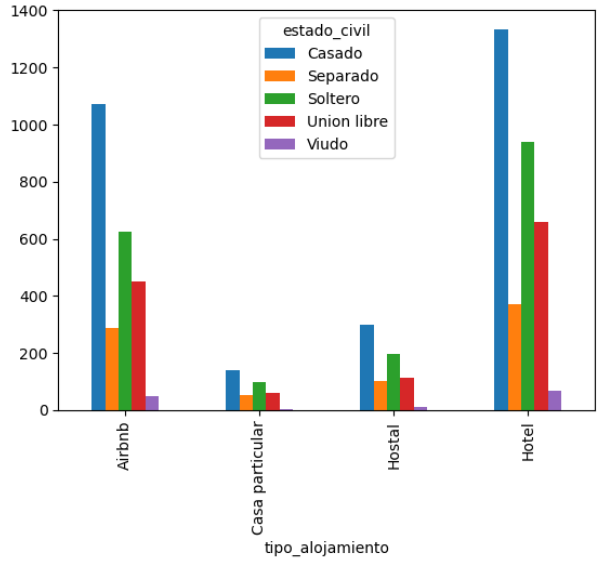

Figura 3: Correlación variables destino y actividades

La figura 3 presenta la distribución de personas en diferentes categorías de estado civil y tipos de alojamiento. Se observa que la mayoría de las personas son solteras y se hospedan en hoteles, seguidas de las casadas y las que utilizan Airbnb. El número de personas separadas, en unión libre y viudas es menor en comparación. La prueba de chi-cuadrado arrojó resultados significativos con un estadístico chi-cuadrado de 23.70 y un p-valor de 0.022. Estos resultados indican que hay una asociación significativa entre las variables evaluadas en el estudio.

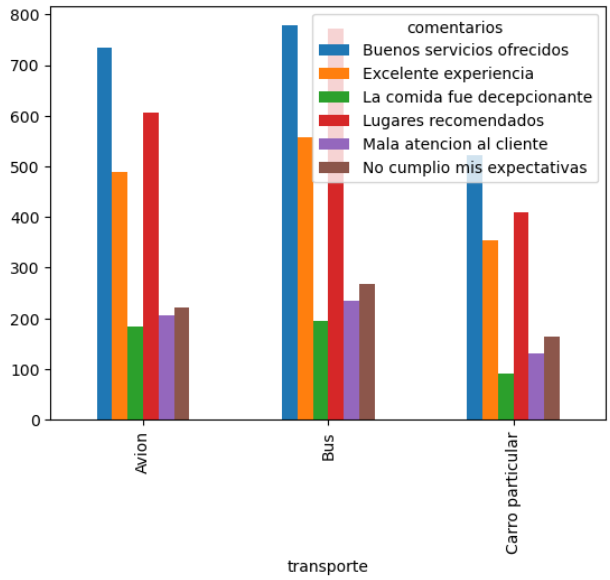

Figura 4: Correlación variables destino y actividades

La figura 4 presenta la distribución de comentarios relacionados con la experiencia del cliente en diferentes

medios de transporte. Se observa que la mayoría de los comentarios provienen de usuarios que utilizaron el bus, seguido por el avión y el carro particular. En términos de aspectos de la experiencia, la mayoría de los comentarios son positivos, destacando los buenos servicios ofrecidos y las excelentes experiencias. Sin embargo, también se mencionan aspectos negativos como la comida decepcionante, la mala atención al cliente y las expectativas incumplidas. Los resultados muestra que hay una asociación significativa entre las variables analizadas, con un estadístico chi-cuadrado de 18.894 y un p-valor de 0.042. Esto indica que la discrepancia observada entre los datos no es aleatoria y es estadísticamente significativa.

3.3. Modificación de datos

La fase de modificación de datos desempeña un papel fundamental en la preparación de la información para el análisis. Durante esta etapa, se llevan a cabo tareas como la limpieza, la transformación y la creación de nuevas variables. Estas acciones buscan mejorar la calidad de los datos, eliminar ruido y establecer relaciones significativas entre las variables. En este caso se eliminan los registros que tengan datos nulos los cuales son el **10.49 %** de la data, esta decisión se toma a partir del conocimiento del negocio y descartando los otros métodos que se pueden implementar para datos nulos.

Las variables cuantitativas se transforman a categóricas y se crea la variable de respuesta a partir de los comentarios de los usuarios para poder implementar los modelos de aprendizaje supervisado, en la figura 5 se muestra el cambio de las variables numéricas.

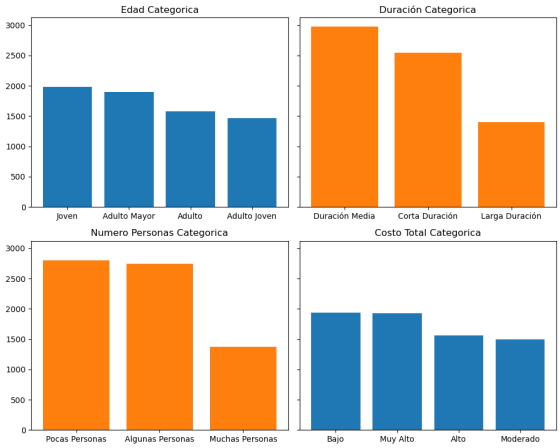

Figura 5: Trasformación variables cuantitativas

Al analizar la figura 6 que muestra el diagrama de frecuencia de la variable $experiencia$ indica si a los clientes les gustó o no sus viajes, se encontró que aproximadamente el 75.4 % de los clientes reportaron una buena experiencia, mientras que alrededor del 24.6 % tuvo una experiencia negativa.

Estos resultados sugieren que la mayoría de los clientes se sienten satisfechos con sus viajes. Además se muestra que hay una clase minoritaria y mayoritaria donde se analizará si hay desbalanceo de clases.

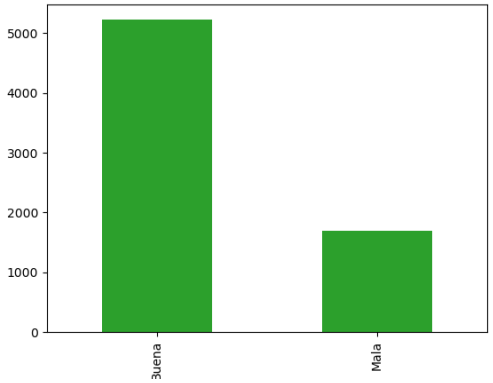

Figura 6: Trasformación variables cuantitativas

3.4. Modelo

En esta fase, se realiza la partición de la data para testeo que son 2076 registros y 4843 registros de entrenamiento, se considera un des-balanceo, por lo tanto, se realiza un tratamiento bajo dos posibles soluciones entre las cuales está SMOTE que según Cruz, Reyes, Rendón, y Alejo es una técnica de sobre-muestreo que aborda el des-balanceo de clases que al aumentar la presencia de la clase minoritaria, mejora el rendimiento de los modelos de aprendizaje automático al evitar el sesgo hacia la clase mayoritaria y submuestreo que según Cruz et al. es una técnica de aprendizaje automático que reduce la cantidad de muestras de la clase mayoritaria para equilibrar el des-balanceo de clases. Se elimina aleatoriamente instancias de la clase mayoritaria para lograr una proporción deseada entre las clases.

En la tabla 2 se evidencia la cantidad de registros en cada uno de los escenarios.

	Entrenamiento	**Testeo**
class weight	4843	2076
smote	7284	7284
submuestreo	2402	2402

Tabla 2: Tabla de registros des-balanceo

A continuación se presenta los resultados de los modelos aplicados de acuerdo al remuestreo.

3.4.1. Regresión logística

En la figura 7 se presentan las métricas de evaluación para un modelo de clasificación. La precisión, el recall y el F1-Score son relativamente bajos para ambas clases, lo que indica dificultades en la clasificación precisa de instancias positivas y negativas. La precisión global del modelo es moderada, pero aún muestra una capacidad limitada para predecir correctamente las clases en general.

	precision	recall	f1-score	support
0	0.25	0.62	0.36	495
1	0.78	0.42	0.54	1581
accuracy			0.46	2076

Figura 7: Métricas Regresión Logística

3.4.2. Naive Bayes

La figura 8 presenta dificultades en la clasificación precisa de ambas clases. Muestra un rendimiento relativamente bajo en términos de precisión y recall para la clase 0, lo que indica problemas para identificar correctamente los casos negativos. Aunque la precisión para la clase 1 es alta, el recall sugiere que no se han identificado todos los casos positivos.

	precision	recall	f1-score	support
0	0.25	0.35	0.29	495
1	0.77	0.67	0.72	1581
accuracy			0.59	2076

Figura 8: Métricas Naive Bayes

3.5. Evaluación

En esta ultima fase se evalúan los modelos para decidir por el más adecuado para la clasificación de experiencia de los usuarios que adquieren paquetes de viaje. A continuación se muestran los gráficos de barras comparando los modelos con sus métricas.

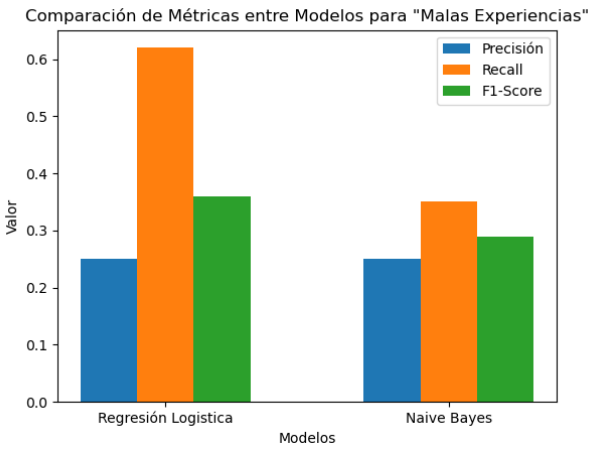

Figura 9: Comparación modelos malas experiencias

Como se observa en la figura 9 al comparar los resultados de los dos modelos de machine learning para la clasificación de malas experiencias, se analiza lo siguiente:

El modelo de Regresión Logística muestra una precisión del 0.25, lo que significa que de todas las predicciones clasificadas como malas experiencias, solo el 25 % son acertadas. Además, el modelo tiene un recall del 0.62, lo que indica que es capaz de identificar el 62 % de todas las malas experiencias reales.

Por otro lado, el modelo de Naive Bayes también tiene una precisión del 0.25, similar al modelo de Regresión Logística. Sin embargo, su recall es del 0.35, lo que indica que identifica un porcentaje menor de malas experiencias reales en comparación con el modelo anterior.

En resumen, ambos modelos presentan resultados modestos en términos de precisión, recall y F1-score. Sería necesario realizar un análisis más detallado para identificar las posibles causas de estos resultados y considerar ajustes en los algoritmos o la exploración de otros enfoques para mejorar el rendimiento de los modelos.

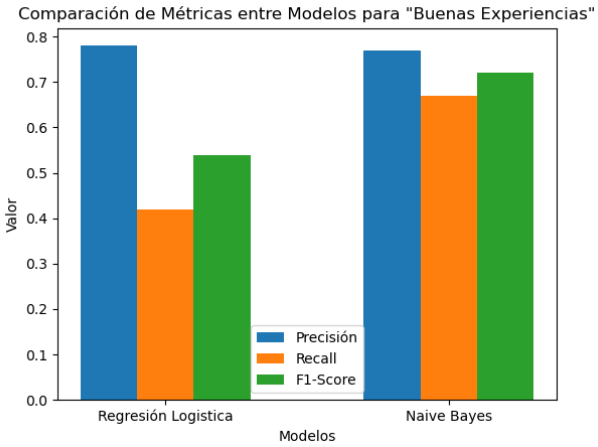

Figura 10: Comparación modelos buenas experiencias

Como se puede apreciar en la figura 10 al comparar los resultados de los dos modelos de machine learning utilizados para clasificar las buenas experiencias de los usuarios, se observa que ambos modelos presentan una alta precisión, con valores cercanos al 78 % y 77 % respectivamente. Sin embargo, el modelo de Naive Bayes muestra un mejor rendimiento en términos de recall y F1-score, con valores de 0.67 y 0.72 respectivamente, superando ligeramente al modelo de Regresión Logística que obtiene un recall de 0.42 y un F1-score de 0.54.

Lo mencionado en el párrafo anterior indica que el modelo de Naive Bayes es más eficiente para identificar correctamente las buenas experiencias de los usuarios en comparación con el modelo de Regresión Logística. Estos resultados son relevantes para la selección del modelo más adecuado en la industria turística, priorizando el equilibrio entre la precisión y la capacidad de detectar las buenas experiencias.

3.5.1. Curva de aprendizaje de los modelos

La curva de aprendizaje se refiere a la representación gráfica de cómo mejora el rendimiento de un modelo a medida que se aumenta la cantidad de datos de entrenamiento disponibles. A continuación se evidencia las curvas de aprendizaje de los modelos.

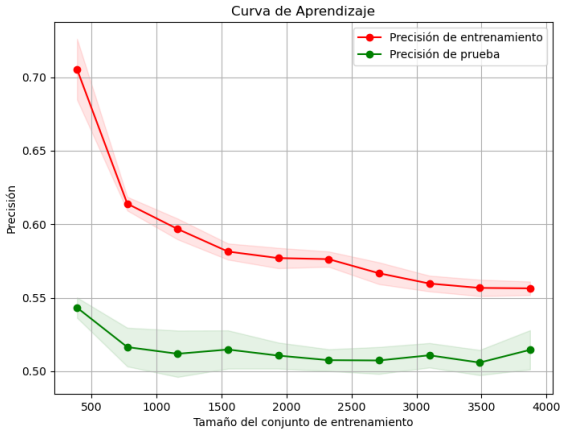

Figura 11: Curva aprendizaje regresión logística

En la curva de aprendizaje del modelo de regresión logística se identifica que a medida que aumenta el tamaño de muestra el modelo pierde precisión, sin embargo se estabiliza en el tamaño de muestra de 3500 registros y se observa que la curva de aprendizaje de la prueba tiende a mejorar a partir de los 3500.

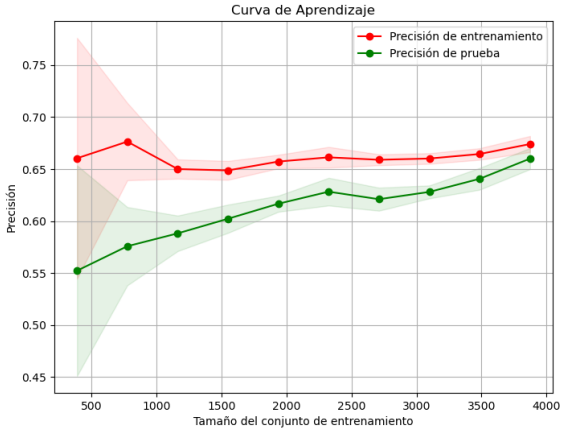

Figura 12: Curva aprendizaje naive bayes

En el caso del modelo naive bayes se identifica que la curva de aprendizaje es mucho mejor a diferencia del modelo de regresión logística porque a medida que aumenta el tamaño de muestra el este mejora su precisión.

4. Resultados y discusión

Al analizar las métricas de los dos modelos, se observa que el modelo de Regresión Logística presenta una precisión ligeramente superior en comparación con el modelo de Naive Bayes, con un valor de 0.78 frente al 0.77 respectivamente. Esto significa que el modelo de Regresión Logística tiene una mayor capacidad para clasificar correctamente las buenas experiencias de los usuarios en la compra de paquetes turísticos.

Aunque el modelo de Naive Bayes muestra un mejor rendimiento en términos de recall y F1-score, con valores de 0.67 y 0.72 respectivamente, es importante tener en cuenta que la precisión es una métrica fundamental en este contexto. La precisión nos indica la proporción de casos positivos clasificados correctamente, es decir, la capacidad del modelo para identificar las buenas experiencias de los usuarios sin clasificar erróneamente otras experiencias.

En el contexto de la industria turística, es crucial maximizar la precisión, ya que la empresa especializada en paquetes turísticos busca brindar una experiencia satisfactoria a sus clientes. Al seleccionar el modelo de Regresión Logística, que tiene una precisión ligeramente más alta, se asegura una mayor probabilidad de identificar correctamente las buenas experiencias de los usuarios, lo que puede contribuir a mejorar la calidad del servicio y la satisfacción del cliente.

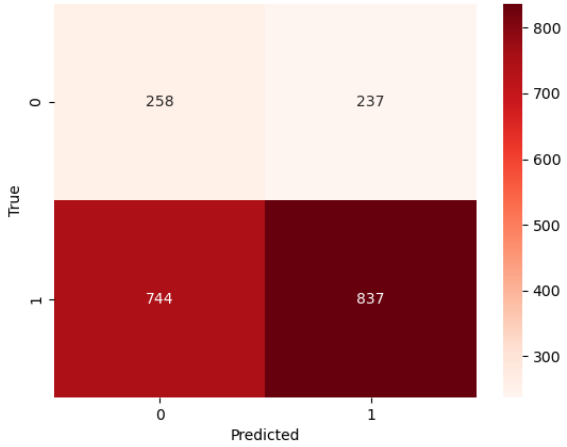

Figura 13: Matriz de confusión

Los resultados de la matriz de confusión en la figura 13 muestran que el modelo ha clasificado correctamente 258 casos como malas experiencias (0) y 837 casos como buenas experiencias (1). Sin embargo, se observa que ha habido cierta confusión en la clasificación de los casos.

Se identificaron 744 casos de buenas experiencias que fueron clasificados erróneamente como malas experiencias (falsos negativos), lo que indica que el modelo no logró reconocer estas experiencias positivas. Por otro lado, se clasificaron incorrectamente 237 casos como buenas experiencias cuando en realidad eran malas experiencias (falsos positivos).

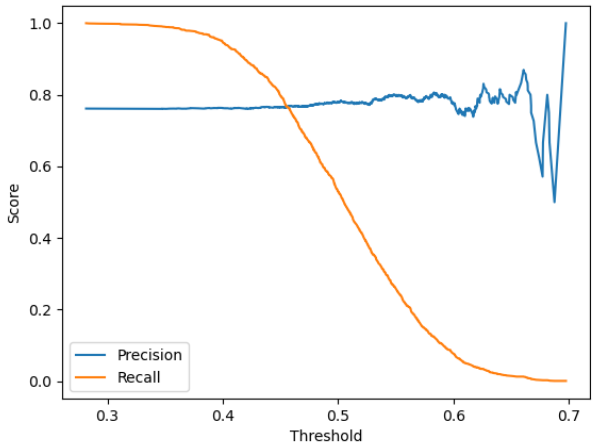

Figura 14: Gráfica de curva de precisión y recall

Como se observa en la figura 14 se ajusta el umbral para poder optimizar las métricas del modelo pero esto implica cambiar el valor mínimo requerido de probabilidad para clasificar una instancia como positiva o negativa. En este caso, el umbral se ha movido a 0.52, lo que significa que una instancia se clasificará como positiva si la probabilidad estimada por el modelo es mayor a 0.52, y como negativa en caso contrario.

Este ajuste del umbral puede tener un impacto en el equilibrio entre la precisión y el recall del modelo. Un umbral más alto, como en este caso, puede aumentar la precisión al clasificar menos instancias como positivas, pero también puede disminuir el recall al perder algunas instancias que realmente son positivas.

	precisión	recall	f1-score
Modelo "1"	0.78	0.53	0.63
Modelo "0"	0.26	0.52	0.34
Umbral "1"	0.78	0.42	0.54
Umbral "0"	0.25	0.62	0.36

Tabla 3: Tabla comparación diferentes escenarios

En resumen, en la tabla 3 se compararon dos enfoques diferentes para abordar el desequilibrio de clases en la clasificación de las experiencias de los usuarios: el "Modelo Balanceadoz el Ümbral".

En comparación, el "Modelo Balanceado"muestra un mejor desempeño general en términos de precisión y recall en comparación con el enfoque del Ümbral". Aunque el modelo balanceado tiene un recall ligeramente más bajo, logra una mejor capacidad de clasificación global, lo que lo convierte en una opción más confiable para identificar tanto las buenas como las malas experiencias de los usuarios. Sin embargo, la elección entre ambos enfoques dependerá de las necesidades y prioridades específicas del problema en cuestión.

5. Conclusiones

La regresión logística mostró un mejor rendimiento en términos de precisión, recall, F1-score, en comparación con el modelo Naive Bayes. Estos resultados muestran que la regresión logística capturó patrones significativos y más complejos en los datos de entrada.

La elección de la regresión logística también se respalda debido a la interpretación de los coeficientes, siendo estos los que señalan la importancia de cada característica en su clasificación, ayudando de esta manera a comprender los factores que impactan la experiencia del usuario.

Los modelos supervisados pueden beneficiarse de técnicas de mejora y optimización, como la selección de características, la ingeniería de características o la búsqueda de hiper-parámetros. Estas técnicas permiten mejorar la precisión y el rendimiento del modelo, así como reducir el sobre-ajuste y aumentar la capacidad de generalización.

La implementación de un modelo para la toma de decisiones en estrategias de marketing se vuelve indispensable en este contexto. Un modelo sólido y bien diseñado permite a las agencias de viaje aprovechar los datos recopilados y utilizarlos para identificar los planes más populares entre los usuarios, comprender los factores que influyen en las decisiones de compra y determinar qué segmentos de la población son más propensos a elegir planes de viaje específicos.

Referencias

Buhalis, D., y Michopoulou, E. (2011). Information sources for tourist destinations: A study of the information needs of european tourists. journal of hospitality marketing management. , 587 - 607.

Cruz, H., Reyes, A., Rendón, y Alejo, R. (2018). Information sources for tourist destinations: A study of the information needs of european tourists. journal of hospitality marketing management. , 198 - 200.

IBM. (s.f.). *Ibm.* (Sitio Web. Recuperado de `https://www.ibm.com/mx-es/analytics/machine-learning`)

KeepCoding. (2022). *Keepcoding.* (Sitio Web. Recuperado de `https://relopezbriega.github.io/blog/2015/10/10/machine-learning-con-python/`)

López, R. (2015). *Matemáticas, análisis de datos y python.* (Sitio Web. Recuperado de `https://relopezbriega.github.io/blog/2015/10/10/machine-learning-con-python/`)

Roman, V. (2019). *Medium.* (Sitio Web. Recuperado de `https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaciÃşn-4bcb24b307f`)

Universidad Piloto de Colombia. (2022). *Universidad piloto de colombia.* (Sitio Web. Recuperado de `https://estudiarvirtual.unipiloto.edu.co/blog/evaluar-modelo-de-regresion-logistico/`)

Zhang, H., Liu, C., y Li, G. (2018). A classification model for tourist experience based on support vector machine. in 2018 ieee international conference on cybernetics and intelligent systems (cis) and ieee conference on robotics, automation and mechatronics (ram). *IEEE*, 10-14.

Zhang, X., y Song, H. (2015). Information sources for tourist destinations: A study of the information needs of european tourists. journal of hospitality marketing management. , 458 - 475.