EasyChair Preprint

№ 8800

An Application of Machine Learning Techniques to the Prediction of Purchase in the Tourism Sector

Nuria Gómez Cuenca and Mónica Carmona Arango

September 6, 2022

An application of Machine Learning techniques to the prediction of purchase in the tourism sector

Nuria Gómez Cuenca
Business Management and Marketing
University of Huelva
Huelva, Spain
nuria.gomez@dem.uhu.es

Mónica Carmona Arango
Business Management and Marketing
University of Huelva
Huelva, Spain
monica@dem.uhu.es

***Abstract*— The concept of smart tourism has gained more and more importance over the years in the tourism sector due to the evolution of information and communication technologies. This evolution has contributed to the development of intelligent systems that achieve an increasingly personalized tourist experience every day. This article presents the application of various machine learning techniques for the prediction of purchasing behavior in the tourism sector.**

Keywords*— *turismo inteligente, experiencia personalizada, tecnologías, aprendizaje automático.

I. INTRODUCCIÓN

El nacimiento del concepto *smart* podemos situarlo a finales del siglo XX, cuando "se detecta la poderosa influencia de las TIC en el turismo de la mano de los sistemas globales de distribución (GDS) y las centrales de reserva (CRS)" [1].

Estas novedades han originado una revolución en las estrategias de mercado ligada con tecnologías de Big Data e Inteligencia Artificial, debido a la necesidad de toma de decisiones basadas en datos.

Los datos se hacen esenciales hoy en día en cualquier empresa. Con ellos, van a poder entender a sus clientes y poder ofrecerles lo más adecuado a sus necesidades, lo que les permitirá ahorrar tiempo y dinero.

Además, los datos también permiten conocer y estudiar los movimientos de los usuarios, y tanto en lugares tangibles (ciudades, monumentos, restaurantes, etc), como intangibles (como páginas web). Con el registro de estos movimientos podremos obtener mapas de calor para estudiar qué llama más la atención, y con ello, poder hacer recomendaciones relacionadas con sus preferencias.

También, gracias a estas tecnologías, las reservas *online* o los *chatbots*, así como la obtención de información en tiempo real, han facilitado bastante las decisiones de los usuarios, agilizando de este modo cualquier proceso y eliminando barreras.

Sin embargo, es a partir del año 2019 cuando da comienzo una pandemia mundial ocasionada por la COVID-19 que revolucionaría al planeta en todos sus aspectos, entre ellos el turismo.

El sector turístico podría considerarse uno de los más afectados durante la pandemia, puesto que se restringió la movilidad ocasionando grandes pérdidas a las empresas enfocadas en este sector y una baja demanda debido a la precaución que tomaron los viajeros para evitar contagios.

Este sector resulta estratégico para nuestro país, España, puesto que antes del impacto de la pandemia este aportaba el 12.3% del PIB y el 12.7% del empleo [2].

Sin embargo, se estimó en marzo de 2020 que las llegadas de turistas se reducirían en un 1-3%, siendo las regiones de Asia y el Pacífico las más afectadas, con una caída del 9-12% de las previsiones de llegadas. Esto supondrá una pérdida estimada en torno a los 30 000-50 000 millones de dólares estadounidenses; cuando, antes de la pandemia se preveía un crecimiento del 3-4% [3].

Además, Zurab Pololikashvili, Secretario General de la OMT, subraya que "las pequeñas y medianas empresas constituyen alrededor del 80% del sector turístico y están particularmente expuestas, siendo millones las personas, muchas de ellas en comunidades vulnerables, para las que el turismo es su medio de vida" [3].

El uso de las nuevas tecnologías de la comunicación y la aplicación de las nuevas técnicas de recopilación, almacenamiento y análisis de datos deben suponer una oportunidad para la recuperación de los mercados turísticos [4].

En este artículo nos planteamos el uso de las técnicas de *machine learning*, concretamente el uso de algoritmos de aprendizaje supervisado para construir un modelo que ayude a predecir la decisión de los turistas cuando se enfrentan a un proceso de decisión de compra.

II. REVISIÓN DE LA LITERATURA

En los últimos años se han desarrollado numerosos estudios acerca del impacto que genera el uso de grandes datos en la economía del sector turístico. Las investigaciones llevadas a cabo han tratado de generar métodos que logren precisos valores de predicción.

Uno de los últimos trabajos es el de predecir modelos de comportamiento de los turistas en Java con el objetivo de incrementar el PIB del sector turístico en el país [5]. Para ello, emplean el algoritmo de máquinas de vector soporte (SVM), prediciendo la llegada de turistas con alta precisión, siendo de apoyo esta investigación para las políticas turísticas del país.

Otro trabajo reciente de la aplicación del aprendizaje automático es el de la predicción de las visitas turísticas a

[1] Celdrán-Bernabéu, M. A., Mazón, J.-N., Ivars-Baidal, J. A., & Vera-Rebollo, J. F. (2018). "Smart tourism. Un estudio de mapeo sistemático". Cuadernos de Turismo, 41.

[2] del Castillo, J. M. D. (2020). "Turismo y medio ambiente en un escenario post COVID-19." En Turismo post COVID-19: el turismo después de la pandemia global, análisis, perspectivas y vías de recuperación (pp. 363-385). Ediciones Universidad de Salamanca.

[3] Organización Mundial del Turismo (6 de marzo de 2020). "COVID-19: La OMT pide que el turismo se incluya en los planes de recuperación." https://www.unwto.org/es/news/covid-19-la-omt-pide-que-el-turismo-se-incluya-en-los-planes-de-recuperacion.

[4] Martínez, S. M. (2020). "Redes sociales y big data: Una oportunidad para la recuperación de los mercados turísticos." En Turismo post COVID-19: el turismo después de la pandemia global, análisis, perspectivas y vías de recuperación (pp. 147-165). Ediciones Universidad de Salamanca.

[5] Purnaningrum, E., & Athoillah, M. (marzo de 2021). "SVM Approach for Forecasting International Tourism Arrival In East Java." En Journal of Physics: Conference Series (Vol. 1863, No. 1, p. 012060). IOP Publishing.

China, utiliza el método *Kernel Extreme Learning Machine* (KELM) [6].

En la referencia [7], se utilizan tanto técnicas de aprendizaje automático como de aprendizaje profundo, en la predicción de visitas a Europa.

También resulta relevante el estudio llevado a cabo por [8] que emplea un método llamado Red Neuronal de Memoria a Corto Plazo (LSTM) que se basa en una comparación del método ARIMA y el aprendizaje automático para predecir las tendencias del turismo.

Por último, también destacamos el trabajo desarrollado para realizar un pronóstico de la demanda turística en España combinando métodos KNN y SVR [9].

III. OBJETIVOS Y METODOLOGÍA

A. Objetivos

Como hemos visto, se hace necesario el uso de tecnologías Big Data para el desarrollo eficiente de las empresas englobadas en el sector turístico.

Por ello, nuestro principal objetivo será estudiar la viabilidad de un nuevo paquete turístico que contemple todas las medidas necesarias ante la COVID-19, mediante la implementación de diferentes técnicas de aprendizaje automático supervisado. Para ello comprobaremos la precisión de cada una de ellas, concluyendo finalmente si este nuevo paquete generará o no beneficios a la empresa turística.

B. Datos

Para generar un modelo donde podamos aplicar el concepto del turismo inteligente, hemos utilizado una base de datos, obtenida a través de *Kaggle* (https://www.kaggle.com/), de la agencia turística www.trips-travel.com, en la que encontramos 4.888 observaciones y 20 variables relativas a clientes potenciales del último año, que son las representadas en la [Tabla I].

Hasta ahora, la compañía disponía de 5 tipos de paquetes turísticos: Básico, Estándar, Deluxe, Super Deluxe y King. Sin embargo, recientemente ha presentado un nuevo paquete turístico que se ajusta a las medidas de seguridad necesarias ante la COVID-19, llamado Bienestar, el cual será objeto de nuestro estudio.

C. Metodología

En primer lugar, realizaremos un análisis de componentes principales con el objetivo de reducir la dimensionalidad del conjunto de datos, evitando así problemas de escala, y que ciertas variables ejerzan influencia sobre el resto. Posteriormente, se irán aplicando las siguientes técnicas:

- Bosques aleatorios (*Random forest*)
- K vecinos más cercanos (KNN)
- Máquinas de vector soporte (SVM)

TABLA I. VARIABLES DE ESTUDIO

Variable	Descripción	Valores válidos
CustomerID	ID de cliente único	De 200 000 a 204 887
ProdTaken	Los clientes contrataron el producto o no	0: No, 1: Sí
Age	Edad del cliente	De 10 a 70 años
TypeofContact	Cómo se contactó con el cliente	Self Enquiry: Autoconsulta, Company Invited: Invitación de la empresa
CityTier	El nivel de ciudad depende del desarrollo de la ciudad, la población, las instalaciones y los niveles de vida	1: Bajo, 2: Medio, 3: Alto
DurationOfPitch	Duración del discurso de un vendedor al cliente	De 0 a 120 minutos
Occupation	Ocupación del cliente	Salaried: Asalariado, Free Lancer: Autónomo, Small Business: Pequeña empresa, Large Business: Gran empresa
Gender	Género del cliente	Female: Femenino, Male: Masculino
NumberOfPersonVisiting	Número total de personas que planean hacer el viaje con el cliente	De 1 a 5 personas
NumberOfFollowups	El vendedor ha realizado el número total de seguimientos después del argumento de venta	De 0 a 7 seguimientos
ProductPitched	Producto presentado por el vendedor	Deluxe, Basic, Standard, Super Deluxe, King
PreferredPropertyStar	Calificación de propiedad hotelera preferida por cliente	De 3 a 5 estrellas
MaritalStatus	Estado civil del cliente	Single: Soltero, Divorced: Divorciado, Married: Casado, Unmarried: Separado
NumberOfTrips	Número medio de viajes en un año por cliente	De 0 a 20 viajes
Passport	El cliente tiene pasaporte o no	0: No, 1: Sí
PitchSatisfactionScore	Puntaje de satisfacción del argumento de venta	De 1 (Muy bajo) a 5 (Muy alto)
OwnCar	Si los clientes poseen un automóvil o no	0: No, 1: Sí
NumberOfChildrenVisiting	Número total de niños menores de 5 años que planean hacer el viaje con el cliente	De 0 a 3 niños
Designation	Designación del cliente en la organización actual	Manager: Gerente, Executive: Ejecutivo, Senior Manager: Gerente Senior, AVP: Vicepresidente adjunto, VP: Vicepresidente
MonthlyIncome	Ingreso mensual bruto del cliente	De 0 a 100 000 dólares

[6] Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). "Forecasting tourist arrivals with machine learning and internet search index." Tourism Management, 70, 1-10.

[7] Chen, N. C., Xie, W., Welsch, R. E., Larson, K., & Xie, J. (junio de 2017). "Comprehensive predictions of tourists' next visit location based on call detail records using machine learning and deep learning methods." En 2017 IEEE International Congress on Big Data (BigData Congress) (pp. 1-6). IEEE.

[8] Li, Y., & Cao, H. (2018). "Prediction for tourism flow based on LSTM neural network." Procedia Computer Science, 129, 277-283.

[9] Claveria, O., Monte, E. & Torra, S. (2016). "Modelling tourism demand to Spain with machine learning techniques." The impact of forecast horizon on model selection. Revista de Economia Aplicada. 24. 109-132.

Para la aplicación de estos diferentes métodos vamos a tomar como posibles clientes potenciales aquellos que viajan por motivos de trabajo, ya que en tiempos de pandemia existen diferentes restricciones de movilidad en bastantes países, lo que podría ocasionar una menor movilidad de aquellos que viajen por motivos de ocio.

De igual modo, también vamos a prestar una mayor consideración a aquellos clientes que provienen de una ciudad de nivel 3, puesto que entendemos que ante la crisis económica que ha generado la COVID-19 serán aquellos los que dispongan de un presupuesto más elevado para poder realizar viajes.

Tras una selección de variables más relevantes, dispondremos de un conjunto más reducido, con 1840 observaciones, siendo más fácil ahora la aplicación de algoritmos para estudiar la predicción.

El motivo que nos ha llevado a elegir estos tres algoritmos es que son los principales clasificadores que producen las precisiones más altas. Según [10], cuando se utilizan dichos algoritmos, los parámetros juegan un importante papel en la producción de resultados de alta precisión.

Sin embargo, antes de aplicar estas técnicas necesitaremos realizar un breve análisis de los datos con los que trabajaremos:

Para empezar, resulta esencial realizar un estudio de la correlación de nuestras variables. Para ello, lo realizaremos a través del método de Pearson, y comprobaremos si existe alguna fuerza o dirección de la relación entre dichos elementos, a través de coeficientes de correlación por cada par de variables [Fig. 1].

Tras la representación gráfica y de los coeficientes, observamos que la mayor correlación positiva se da entre las variables *NumberOfPersonVisiting* y *NumberOfChildrenVisiting*, con un 0.61; la mayor correlación

negativa es entre *CityTier* y *PitchSatisfactionScore*, con -0.04;

y la menor correlación la tenemos entre *CityTier* y *NumberOfChildrenVisiting*, con un valor de 0.

No observamos grandes rasgos de correlación, ya que el dato que representa la mayor correlación no es muy alto, exceptuando aquellos coeficientes que se dan en la diagonal principal.

Seguidamente hemos realizado un análisis de variables categóricas y otro de variables continuas, para estudiar cómo se encuentran distribuidos los datos en función de la variable *ProdTaken* que nos indica si el cliente contrató o no el último producto que se le ofreció.

Los datos obtenidos para las variables categóricas son los representados en la [Fig. 2].

Claramente observamos una cierta tendencia de los clientes a no contratar los productos que se le ofrecen, siendo la variable más significativa *Passport*, que nos demuestra que una gran parte de usuarios que no disponen de pasaporte no contratan ningún paquete. Esto, consecuentemente, ocasiona pérdidas a la agencia de viajes; y en lo que refiere a nuestro modelo, se nos presenta datos no balanceados que dificultará su correcto estudio.

A continuación, generamos la representación gráfica para las variables continuas [Fig. 3].

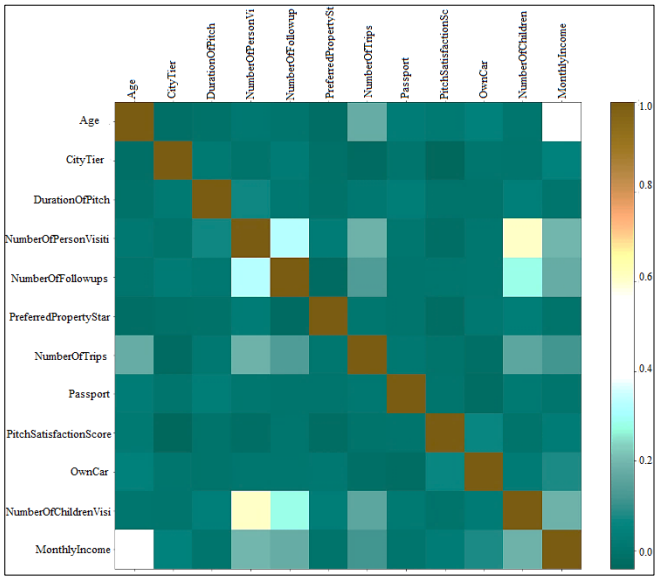

Fig. 1. Matriz de correlación

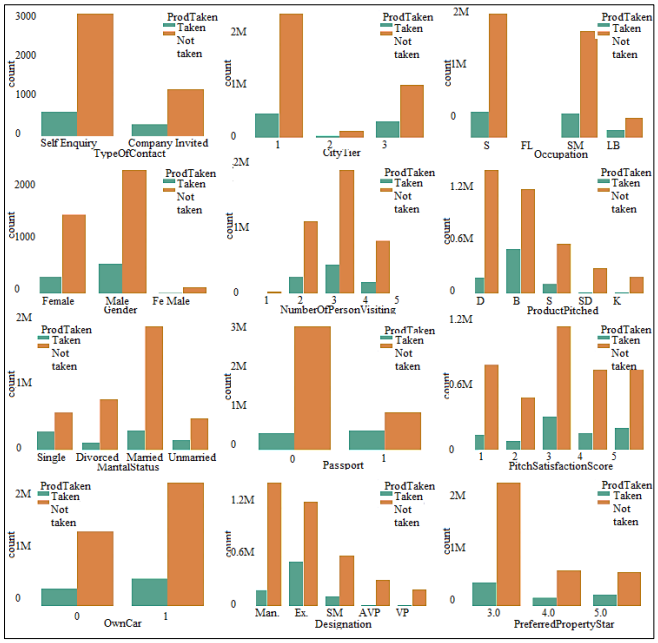

Fig. 2. Análisis de variables categóricas

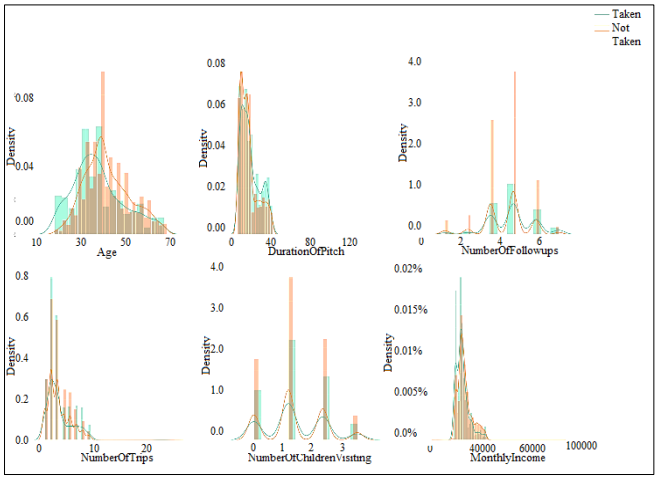

Fig. 3. Análisis de variables continuas

[10] Thanh Noi, P., & Kappas, M. (2017). "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery." Sensors (Basel, Switzerland), 18(1), 18.

En este caso, sí que es posible observar cierta tendencia a contratar el producto ofrecido, especialmente cuanto más largo es el discurso del vendedor (*DurationOfPitch*) y cuanto más jóvenes son los clientes (*Age*). De igual modo, pero de un modo más sutil, encontramos alguna tendencia en función del número de viajes (*NumberOfTrips*) y con ingresos algo más reducidos (*MonthlyIncome*).

A modo resumen, tenemos que las 3 variables más significativas de nuestro modelo serán *MonthlyIncome*, *Passport* y *Age*, en dicho orden de más a menos relevancia; siendo las 3 menos significativas *NumberOfPersonVisiting*, *NumberOfChildrenVisiting* y *OwnCar*, en dicho orden de menos a más relevancia.

De este modo, hemos calculado que en nuestra base de estudio tan solo el 0.18% contrataron algún producto, por lo que nos encontramos ante un problema de datos no balanceados, que tendremos que resolver.

El hecho de tener que realizar un remuestreo de los datos para conseguir el deseado balanceo de datos, podrá conllevar sobreajuste o pérdida de información; sin embargo, y debido a su sencillez, optamos por el método del submuestreo, que nos permitirá obtener muestras aleatorias de cada clase y removerá aquellos valores de la clase mayoritaria.

Una vez que hemos conseguido datos balanceados, correspondiendo 920 observaciones a cada valor de la variable *ProdTaken*, podemos proceder a la aplicación de técnicas que desarrollaremos a continuación.

IV. Resultados

Como comentábamos, el análisis de datos lo realizaremos utilizando técnicas de aprendizaje automático. Concretamente, nos basaremos en el uso de los algoritmos *Random Forest*, KNN y SVM.

Para ello, probaremos diferentes valores para cada algoritmo utilizando los parámetros más adecuados basados en la clasificación general. Con estos algoritmos generaremos los siguientes resultados:

- El valor de exactitud
- La matriz de confusión
- El informe de clasificación
- La curva de aprendizaje

A partir de estos datos, podremos concluir si los algoritmos resultan eficaces o no, en función de los valores obtenidos. Esto nos ayudará a decidir si resultará conveniente invertir en la supuesta campaña de marketing que quiere llevar la agencia de viajes para promocionar su nuevo paquete turístico con medidas COVID-19.

Pero, antes de la aplicación de estos algoritmos necesitaremos realizar un par de pasos más: uno de ellos es dividir nuestro conjunto en datos de entrenamiento y evaluación, reservando el 80% de los datos para el primer conjunto, al que corresponderán 1472 observaciones; y el 20% para el último, con 368 observaciones. Con los datos de entrenamiento podremos entrenar el modelo y con los de evaluación comprobaremos que el modelo generado funciona correctamente. El otro de ellos será realizar un análisis de los componentes principales que veremos a continuación.

A. *Análisis de componentes principales (PCA)*

Se hace totalmente indispensable el análisis de componentes principales para nuestro estudio, ya que permite reducir la dimensión del número de variables originales que se han considerado en el análisis [11].

En este caso, hemos decidido trabajar con una escala robusta para normalizar los datos, ya que al basarse en percentiles, evitará que nuestro modelo se vea influenciado por *outliers*.

Seguidamente, necesitaremos aplicar una función que permita transformar los datos que tenemos en componentes no correlacionados y tipificados. De este modo, estaremos evitando que se produzcan problemas de escala, puesto que siendo la varianza de una variable tipificada igual a 1, la suma de las varianzas será *p* [11].

Por último, nos quedaremos con aquellas variables que representen hasta un 85% de variabilidad explicada acumulada, que serán los primeros 11 componentes.

Si ahora representamos los valores, obtenemos la gráfica [Fig. 4].

Una vez reducidas las variables de nuestra base de datos que logran explicar hasta el 85% de las observaciones en un espacio bidimensional definido por las dos primeras componentes principales, es posible advertir una distribución de los clientes en seis grupos diferentes.

Además, gracias a esta agrupación, es posible detectar la presencia de *outliers*, como es la observación que se encuentra en el centro del lado derecho de la gráfica, que serán los que provoquen mayor varianza en nuestro modelo.

Podemos ver los grupos de clientes que contratan algún paquete (1) y los que no (0) con los puntos de la predicción del conjunto de entrenamiento; de este modo, podemos comprender la relación que guardan las decisiones de los clientes en busca de similitudes.

A la izquierda observamos dos grupos, uno de clientes que contrataron algún producto y otro de clientes que no lo hicieron. Entre estos tipos de clientes es posible que exista una

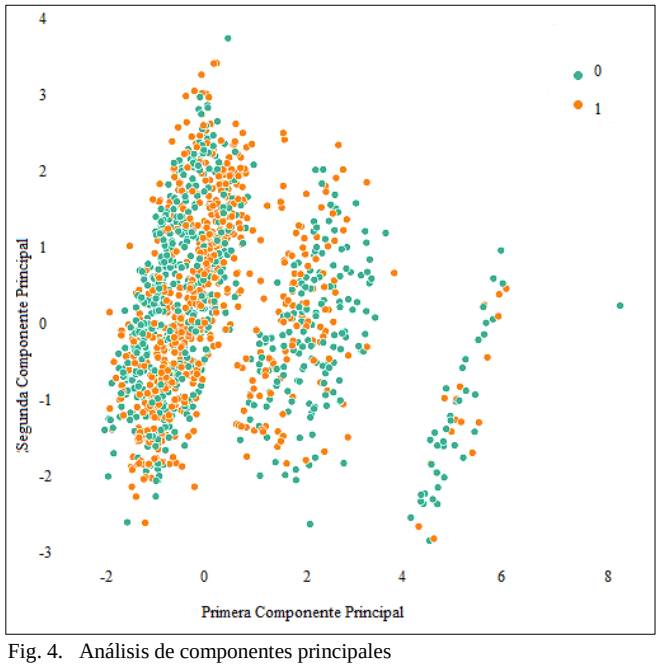

Fig. 4. Análisis de componentes principales

[11] De la Fuente, S. (2011). "Componentes principales." Madrid: Facultad de Ciencias económicas y empresariales. UAM.

transferencia de preferencias entre ellos. Lo mismo ocurre con los dos grupos que encontramos en el medio y en la derecha.

Existe cierta evidencia de que la gran mayoría de clientes, tanto si contratan algún paquete como si no, tienden a ubicarse unas cercas de otras. Estas se encuentran en su mayoría en los cuadrantes izquierdos al centro de gravedad.

Cabe destacar que es posible que se haya producido cierta pérdida de información al realizar esta reducción, pero como comentamos, hemos conseguido datos incorrelados y ordenados de acuerdo con la información que llevan incorporados [12]; y, a continuación, podremos trabajar con ellos aplicando los algoritmos descritos con la seguridad de que no se darán problemas de escala.

B. Bosques aleatorios (Random Forest)

El presente algoritmo se basa en la creación de árboles de decisión seleccionando muestras al azar, y predice clasificando como positivo si la mayoría de árboles predicen la observación como positiva, o viceversa [13].

Son varios los motivos que nos han llevado a emplear este algoritmo en nuestro estudio, principalmente, por destacar en el manejo de variables categóricas. Además, es útil en el manejo de espacios con grandes dimensiones, y por su rápido funcionamiento.

En la siguiente [Tabla II] se presentan las principales fortalezas y debilidades de este modelo [13][14].

Aplicado a nuestro caso de estudio, hemos obtenido un valor de exactitud del 86%, y la siguiente matriz de confusión [Tabla III].

donde (a) son los verdaderos positivos, (b) los falsos negativos, (c) los falsos positivos y (d) los verdaderos negativos.

Los valores (a) y (d) corresponden con los valores estimados de forma correcta por nuestro modelo, mientras que los valores (b) y (c) corresponde a los valores en los que nuestro modelo ha cometido error.

Como observamos, el porcentaje de predicciones correctas frente al total es bastante alto, por lo que nuestro algoritmo clasifica correctamente.

TABLA II. FORTALEZAS Y DEBILIDADES DE LOS MODELOS BASADOS EN *RANDOM FOREST*

Fortalezas	Funcionan bien en problemas de clasificación. Pueden utilizar características numéricas o categóricas, y datos faltantes. Excluye características con baja o sin importancia. Se puede utilizar para grandes cantidades de características.
Debilidades	No son fáciles de interpretar. Pueden requerir un trabajo adicional para ajustar el modelo a los datos.

TABLA III. MATRIZ DE CONFUSIÓN *RANDOM FOREST*

	Valores estimados	
Valores reales	160 (a)	27 (b)
	24 (c)	157 (d)

Seguidamente obtenemos la curva de aprendizaje, donde observamos las puntuaciones del conjunto de entrenamiento y test para diferentes tamaños de conjuntos de entrenamiento, y comprobaremos cómo varía el error en función del tamaño del conjunto de entrenamiento [Fig. 5].

Como vemos, el hecho de agregar más datos de entrenamiento al conjunto de entrenamiento no provoca que cambie mucho, pero al hacerlo en el conjunto de validación, hasta llegar a las 1000 observaciones, sí que nos reporta beneficio, puesto que hará que esta curva ascienda y que nuestro modelo alcance una puntuación más alta en el aprendizaje, disminuyendo así el sesgo.

También obtenemos la curva característica operativa del receptor (ROC), que nos va a ayudar a encontrar el umbral donde la tasa de verdaderos positivos es alta y la tasa de falsos positivos es baja [Fig. 6].

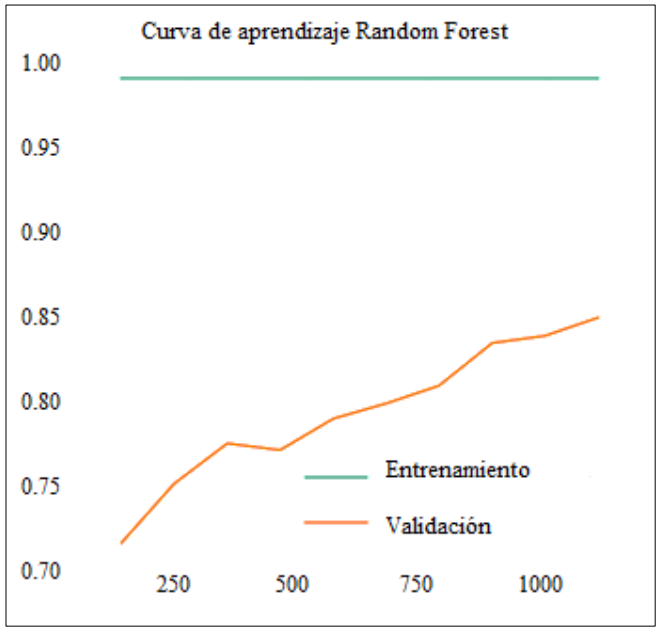

Fig. 5. Curva de aprendizaje *Random Forest*

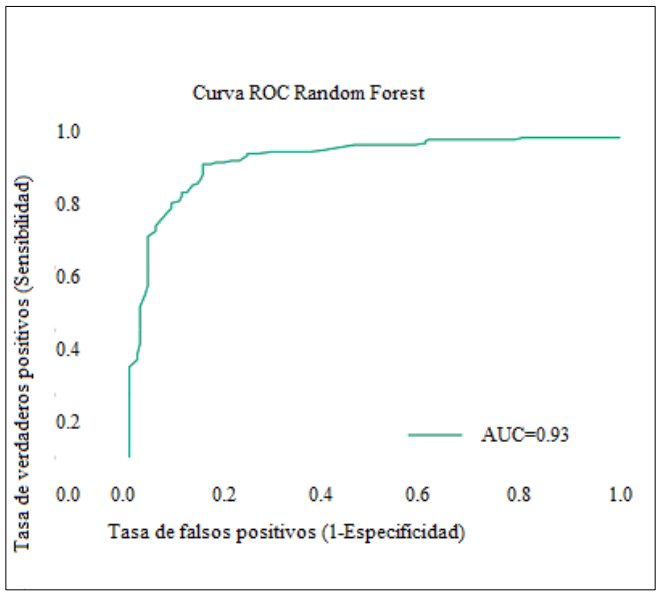

Fig. 6. Curva ROC *Random Forest*

[12] De la Fuente, S. (2011). "Componentes principales." Madrid: Facultad de Ciencias económicas y empresariales. UAM.

[13] Martínez Rodríguez, J. (2018). "Estudio comparativo de modelos de machine learning para la detección de dianas microARN."

[14] Lantz, B. (2019). "Machine learning with R: expert techniques for predictive modeling." Packt publishing ltd.

Tenemos que con bajos valores de la tasa de falsos positivos, ya obtenemos altos valores en la tasa de verdaderos positivos; por lo que, para mejorar la eficacia de nuestro modelo, resultaría conveniente reducir mínimamente la especificidad, al menos a 0.2, puesto que así obtendremos una sensibilidad mucho más alta, en torno al 0.9. De este modo, obtenemos un área bajo la curva de 0.93 puntos, por lo que identificaremos fácilmente aquellos verdaderos negativos, es decir, aquellos a los que no deberíamos dedicar tiempo y dinero en promocionar nuestro nuevo paquete turístico.

C. K vecinos más cercanos (KNN)

Este algoritmo realiza una clasificación de un nuevo ejemplo basándose en los de la muestra de entrenamiento más cercanos, es decir, se asigna el nuevo ejemplo a la clase más frecuente entre los k vecinos más próximos.

El principal motivo por el que usamos este algoritmo es por su resistencia a datos de entrenamiento ruidosos, aunque este se trata de un algoritmo que necesita mayor tiempo de computación. Para ello necesitamos determinar el valor de K, que será el número de vecinos más cercanos y en nuestro caso hemos tomado aleatoriamente el valor de 40.

Podemos encontrar varias fortalezas y debilidades que las resumimos a continuación [Tabla IV][15][16].

Si empleamos dicho algoritmo a nuestro modelo, obtenemos una exactitud del 74%, y la matriz de confusión es la siguiente [Tabla V].

donde (a) son los verdaderos positivos, (b) los falsos negativos, (c) los falsos positivos y (d) los verdaderos negativos.

Los valores (a) y (d) corresponden con los valores estimados de forma correcta por nuestro modelo, mientras que los valores (b) y (c) corresponde a los valores en los que nuestro modelo ha cometido error.

Observamos que el porcentaje de predicciones correctas frente al total es bastante alto, por lo que nuestro algoritmo clasifica correctamente. No obstante, sí es posible observar un ligero aumento respecto al algoritmo de bosques aleatorios de los valores con error, y como consecuencia, un ligero descenso en los valores estimados de forma correcta.

TABLA IV. FORTALEZAS Y DEBILIDADES DE LOS MODELOS BASADOS EN KNN

Fortalezas	Simple y efectivo. No hace suposiciones sobre la distribución de datos subyacentes. Fase de entrenamiento rápida.
Debilidades	No produce un modelo, limitando la capacidad de encontrar nuevos conocimientos en las relaciones entre las características. Fase de clasificación lenta. Requiere gran cantidad de memoria. Las variables nominales y los datos faltantes requieren un proceso adicional.

TABLA V. MATRIZ DE CONFUSIÓN KNN

	Valores estimados	
Valores reales	135 (a)	52 (b)
	42 (c)	139 (d)

A continuación, obtenemos la curva de aprendizaje, donde observamos las puntuaciones del conjunto de entrenamiento y test para diferentes tamaños de conjuntos de entrenamiento, y comprobaremos cómo varía el error en función del tamaño del conjunto de entrenamiento [Fig. 7].

Tenemos que, el hecho de agregar más datos de entrenamiento tanto al conjunto de entrenamiento como al de validación no reportará gran beneficio, puesto que las dos curvas se mantienen prácticamente en paralelo, siendo casi el mismo sesgo tanto para un tamaño del conjunto de 250 como de 1000 observaciones.

También obtenemos la curva característica operativa del receptor (ROC), que nos va a ayudar a encontrar el umbral donde la tasa de verdaderos positivos es alta y la tasa de falsos positivos es baja [Fig. 8].

Podemos ver que, al contrario del algoritmo de bosques aleatorios, aquí sí que necesitaremos disminuir mucho más la

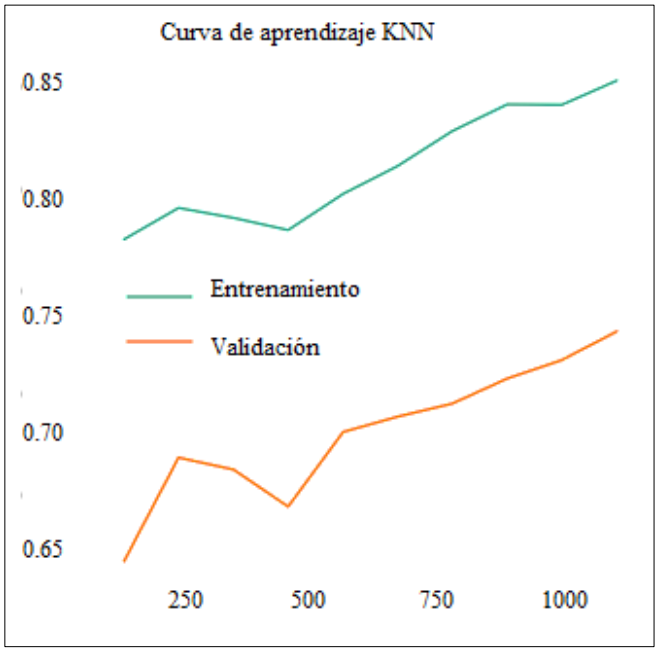

Fig. 7. Curva de aprendizaje KNN

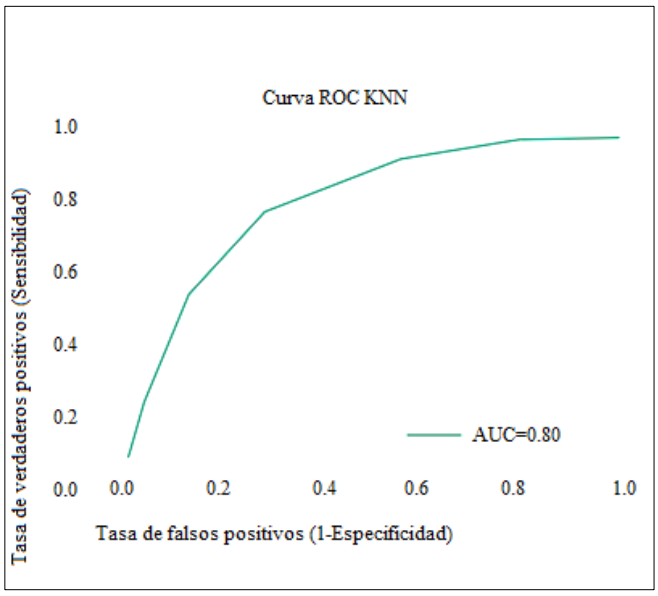

Fig. 8. Curva ROC KNN

[15] Martínez Rodríguez, J. (2018). "Estudio comparativo de modelos de machine learning para la detección de dianas microARN."

[16] Lantz, B. (2019). "Machine learning with R: expert techniques for predictive modeling." Packt publishing ltd.

especificidad para obtener una mayor sensibilidad, siendo el área bajo la curva de 0.8 puntos. Esto tampoco resultaría muy conveniente puesto que al disminuir la especificidad no podremos identificar fácilmente los verdaderos negativos, es decir, aquellos clientes con los que no resultaría conveniente emplear tiempo en promocionar la campaña de marketing.

D. Máquinas de vector soporte (SVM)

El presente algoritmo se basa en la separación lineal entre datos que representan ejemplos y sus características, dividiendo el espacio en particiones homogéneas (Martínez Rodríguez, 2018).

La principal razón del uso de este algoritmo es debido a que nuestro conjunto de datos no puede ser separado linealmente; sin embargo, presenta como el anterior algoritmo un tiempo de computación más elevado que el primero.

A continuación, se proponen varias fortalezas y debilidades resumidas en la siguiente tabla [Tabla VI][17][18].

Cuando aplicamos este algoritmo a nuestro modelo, obtenemos una exactitud del 75%, y la siguiente matriz de confusión [Tabla VII].

donde (a) son los verdaderos positivos, (b) los falsos negativos, (c) los falsos positivos y (d) los verdaderos negativos.

Los valores (a) y (d) corresponden con los valores estimados de forma correcta por nuestro modelo, mientras que los valores (b) y (c) corresponde a los valores en los que nuestro modelo ha cometido error.

Tenemos de nuevo que el porcentaje de predicciones correctas frente al total es bastante alto, por lo que nuestro algoritmo clasifica correctamente. Sin embargo, como ocurre también con el algoritmo KNN, también es posible observar un ligero aumento respecto al algoritmo de bosques aleatorios de los valores con error, y como consecuencia, un ligero descenso en los valores estimados de forma correcta; aunque, el algoritmo SVM destaca ligeramente sobre el KNN en cuanto a mejores resultados.

A continuación, representamos la curva de aprendizaje, donde observamos las puntuaciones del conjunto de entrenamiento y test para diferentes tamaños de conjuntos de entrenamiento, y comprobaremos cómo varía el error en función del tamaño del conjunto de entrenamiento [Fig. 9].

TABLA VI. FORTALEZAS Y DEBILIDADES DE LOS MODELOS BASADOS EN SVM

Fortalezas	Para problemas de clasificación o predicción numérica. Buen funcionamiento con datos ruidosos y no es propenso al sobreajuste. Más fácil que las redes neuronales, debido a la existencia de varios algoritmos SVM bien soportados.
Debilidades	Requiere probar diferentes kernels y parámetros de prueba y error para encontrar el mejor modelo. Lento de entrenar, sobre todo a medida que aumenta el número de características. Los resultados son difíciles de interpretar.

TABLA VII. MATRIZ DE CONFUSIÓN SVM

	Valores estimados	
Valores reales	144 (a)	43 (b)
	48 (c)	133 (d)

Observamos que, al agregar más datos de entrenamiento, tanto al conjunto de entrenamiento como al de validación, nos reportará beneficio, ya que las líneas van convergiendo a medida que aumenta el número de observaciones, siendo en 1000 observaciones donde observamos el menor sesgo.

A continuación, generamos la curva característica operativa del receptor (ROC), que nos va a ayudar a encontrar el umbral donde la tasa de verdaderos positivos es alta y la tasa de falsos positivos es baja [Fig. 10].

Tenemos esta vez un caso similar al de k vecinos más cercanos. De nuevo, necesitaremos reducir bastante más la especificidad, en comparación con el algoritmo de Random Forest, para poder conseguir una sensibilidad más alta, ya que

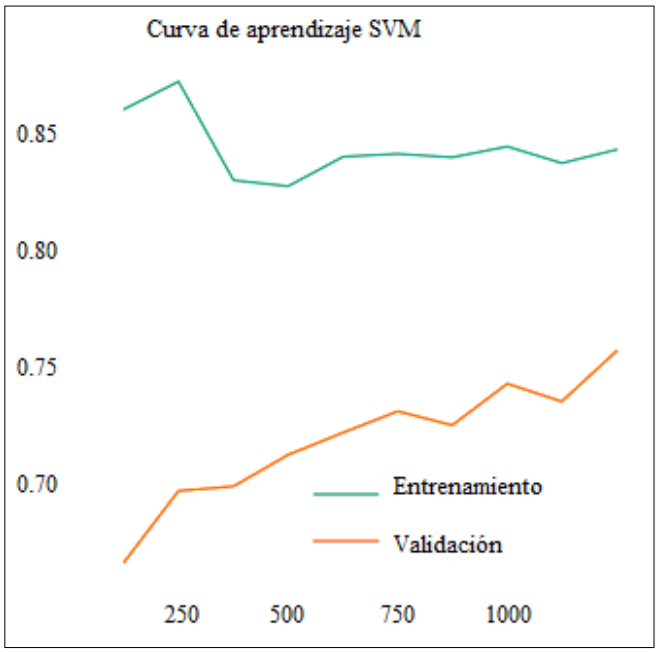

Fig. 9. Curva de aprendizaje SVM

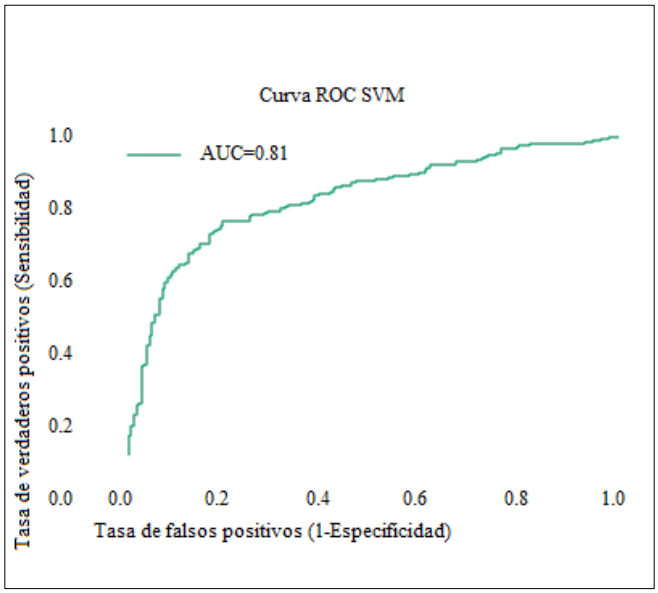

Fig. 10. Curva ROC SVM

[17] Martínez Rodríguez, J. (2018). "Estudio comparativo de modelos de machine learning para la detección de dianas microARN."

[18] Lantz, B. (2019). "Machine learning with R: expert techniques for predictive modeling." Packt publishing ltd.

obtenemos una puntuación de 0.81 en el área bajo la curva. Como comentamos anteriormente, no resultaría conveniente, puesto que aumentaría la tasa de falsos positivos, provocando posibles pérdidas en la campaña de marketing a realizar.

V. Análisis de los resultados

Tras el correspondiente balanceo de datos y análisis de componentes principales, en el estudio de los diferentes algoritmos de aprendizaje automático empleados, hemos obtenido resultados bastante significativos que mostramos en la siguiente tabla [Tabla VIII].

El algoritmo *Random Forest* es el que mejor resultado nos proporciona, con un 86% de exactitud en la predicción. No tan lejos, encontramos el algoritmo SVM con un 75%, y seguidamente, el algoritmo KNN con un 74%.

Además, el algoritmo de Bosques aleatorios es el que presenta mayor sensibilidad y precisión, lo que a su vez provoca que área bajo la curva ROC sea mayor; de este modo, el modelo tendrá mayor habilidad para distinguir verdaderos positivos (aquellos clientes que sí contratarán el producto a ofrecer). También es el que mayor valor-F presenta, es decir, asumiendo que damos el mismo valor tanto a la exactitud como a la sensibilidad.

Sin duda, y principalmente gracias al algoritmo *Random Forest*, nuestro modelo nos ofrece estimaciones con valores bastante altos. Así, de este modo, nos proporciona bastante fiabilidad a la hora de estimar si a nuestra agencia de viajes le resultará conveniente realizar campañas de marketing para promocionar su nuevo paquete turístico con todas las medidas COVID-19 necesarias.

Los resultados que obtendrá la agencia turística resultarán bastante beneficiosos, porque concluimos que esta nueva oferta planteada resultará de gran interés a los clientes recogidos en nuestra base de datos, pudiendo así llamar la atención también de nuevos clientes potenciales.

Sin embargo, cabe decir que estos valores siempre podrán mejorarse bien con cambios de escalas o bien con una base de datos más actualizada, donde poder observar los efectos que va teniendo la campaña de marketing, así como la evolución de la pandemia, y a partir de ahí estimar si seguiría siendo conveniente invertir en ella o dejar de promocionar el paquete puesto que ya no interese a los viajeros o no resultase rentable.

VI. Conclusiones

Como mencionábamos anteriormente, hemos obtenido unos resultados bastante fiables, pero es sobre todo gracias al estudio del algoritmo de bosques aleatorios y su alto valor de predicción lo que nos hará confirmar que a nuestra agencia de viajes www.trips-travel.com le reportará grandes beneficios invertir en una interesante campaña de marketing para promocionar el nuevo paquete turístico que garantice las medidas contra la COVID-19.

Este nuevo paquete, llamará la atención tanto de nuevos clientes como de aquellos que ya contrataron algún paquete anteriormente con la agencia, puesto que los viajeros podrán sentirse más seguros a la hora de viajar.

Resaltaríamos de nuevo la importancia que presenta el Big Data en el sector turístico, sobre todo ahora que vivimos una situación más complicada y donde las decisiones que tomemos presentarán un mayor riesgo que en tiempos anteriores a la pandemia. Ya comentamos que el uso de estas tecnologías, han llegado para quedarse, y es que, hoy más que nunca, nos podrán ofrecer lo que más necesitamos, que se trata de la seguridad.

Seguridad tanto a nivel empresarial como a nivel personal: las empresas podrán tomar decisiones con datos que aporten mayor precisión para no incurrir en pérdidas, y los turistas, en este caso, viajarán con tranquilidad sabiendo que podrán desde casa realizar cualquier reserva y conocer en cualquier momento las aglomeraciones en los sitios que pretendan visitar.

TABLA VIII. Informe de clasificación

Variable	RF	KNN	SVM
Exactitud=(VP+VN)/Total	0.86	0.74	0.75
Sensibilidad=VP/(VP+FN)	0.86	0.72	0.77
Precisión=VP/(VP+FP)	0.87	0.76	0.75
Valor-F	0.86	0.74	0.76
Área bajo la curva (ROC)	0.93	0.80	0.81