MACHINE LEARNING TO PREDICT RECOMMENDATION BY TOURISTS IN A SPANISH PROVINCE

Santiago Aparicio Castillo[1], Nuño Basurto Hornillos[2], Pablo Arranz Val[1,¶],
Paula Antón Maraña[1] y Alvaro Herrero Cosío[2]
[1] Department of Applied Economics
Faculty of Economics and Business Studies
University of Burgos, Pza. de la Infanta Dª. Elena s/n, 09001 Burgos, Spain
[2] Department of Computer Engineering, Polytechnic School
University of Burgos, Avda. Cantabria s/n, 09006 Burgos, Spain
Received 30 July 2020
Revised 9 March 2022
Accepted 13 March 2022

The analysis of the opinions and experiences of tourists is a key issue in tourist promotion. More precisely, forecasting whether a tourist will or will not recommend a given destination, based on his/her profile, is of utmost importance in order to optimize management actions. According to this idea, present research proposes the application of cutting-edge Machine Learning techniques in order to predict tourist recommendation of rural destinations. More precisely, classifiers based on supervised learning (namely Support Vector Machine, Decision Trees, and k-Nearest Neighbor) are applied to survey data collected in the province of Burgos (Spain). Available data suffer from a common problem in real-life datasets (data unbalance) as there are very few negative recommendations. In order to address such problem, that penalizes learning, data balancing techniques have been also applied. The satisfactory results validate the proposed application, being a useful tool for tourist managers.

Keywords: Artificial Intelligence, Supervised Learning, Classification, Tourism Management, Recommendation.

1. Introduction & Previous Work

Today, tourism is one of the most important and fastest growing sectors in the world; it plays a very significant role in the economy's growth.[1] The tourism sector accounted for 10.4% of the world's GDP and generates one in eleven jobs in 2019.[2]

Focusing on Spain, the Spanish GDP is almost 155,000 million euros in 2019 in terms of final tourist demand, 12.4% of GDP, three tenth more than in 2018. Since 2015, the weight of tourism in GDP has grown 1.3 points, from 11.1% to 12.4%. Meanwhile, employment in the characteristic economic branches of tourism reached 2.68 million jobs. This accounted for 12.7% of total employment in the economy in 2019, one tenth less than in 2018. The weight of tourism-related employment has grown 0.6 points since 2015, from 12.1% to 12.7% of total employment in the economy. The largest component of domestic tourism consumption in 2019 was domestic tourism expenditure, with 27.4% of the total. Final demand associated with tourism increased by 3.4%, in terms of volume, in 2019. Since 2015 the evolution in real terms of the tourism economy has been greater than that of the economy as a whole and, despite the crisis suffered by the Covid-19 pandemic, the weight of tourism in the global and national economy is expected to recover pre-crisis levels, demonstrating the strength and resilience of the sector.[3,4]

In the province of Burgos (Spain), tourism is a sector that fosters both economic and cultural activities. This is reflected in the contribution of the hospitality sector to employment in Burgos, which has increased from 6.82% in December 2019 to 7.38% in December 2021, despite the fall caused by the pandemic crisis.[5,6]

In 2019 the province of Burgos was the province in the Spanish region of Castilla y León that received the highest number of tourists (1,503,199) which meant a total of 2,329,692 overnight stays. These data have declined due to the pandemic effect and have not yet recovered to previous levels since in 2021, the number of tourists was 847,594 and the overnight stays were 1,341,116.[6,7] If we compare with other destinations in Spain, Burgos is in fifth position in the ranking of interior tourism cities according to EXCELTUR.[8]

The importance of tourism in Burgos is also evidenced by the confluence of three World Heritage Sites (Burgos Cathedral -244,311 visitors in 2021-, the Pilgrims' Route of the Camino de Santiago and Archaeological Site of Atapuerca -350,644 visitors in 2021 which has resulted in the Museum of Human

Evolution, one of the most visited museums nationally with 91,599 visitors in 2021-) and an Intangible Heritage (Burgos Creative City of Gastronomy) recognised by UNESCO. [7,9,10]
This work arises as a result of the collection and analysis of tourism data to comply with the agreement signed between the University of Burgos and the Society for the Development of the Province of Burgos for the development of actions included in the strategic plan PEBUR 15-20.
For all these reasons, the analysis of tourism in Burgos requires attention, since the importance of this sector suggests special interest for tourism decision makers and researchers.
As is well known, tourism is an economic power and therefore has the capacity to contribute to all the UNWTO Sustainable Development Goals (SDGs). Particularly in the province of Burgos, tourism is an important factor in economic and social development that brings benefits to the local communities. Tourism generates income by providing jobs at the local and community level (SDG 1 - No to poverty) and tax revenues generated by tourism may be reinvested in health care and services (SDG 3 - Good health and well-being). Likewise, tourism in Burgos has the potential to be a powerful key for community development and the reduction of inequalities (SDG 10 – Reduced inequalities) as regards the rural destination development of Burgos. All this with an integrated vision that promotes the evolution of tourism through the joint development of new information technologies, sustainability, innovation and social cohesion. Indeed, tourism can foster urban infrastructures and accessibility and promote the regeneration and preservation of cultural and natural heritage, on which tourism depends (SDG 11 - Sustainable Cities and Communities), as the rich biodiversity and natural heritage are often the major reasons why tourists visit a destination (SDG 15 - Life on Land). The achievement of all these objectives will help to relieve a major current issue that is the depopulation suffered by many regions in the Spanish hinterland. The adoption of measures to preserve natural and heritage resources is a crucial factor in enhancing the value of tourist destinations from a sustainable territorial and tourist attraction point of view.[11]
In line with Spain's Sustainable Tourism Strategy 2030,[12] the Spanish Tourism Plan Horizon 2020[13] and the Strategic Tourism Plans of the region of Castilla y León,[14] the Strategic Plan of Burgos states,[15] it is essential to enhance the province's competitive advantages by developing a more economically efficient and sustainable tourist destination model, in order to position Burgos as a benchmark for economic activity, quality of life and smart destination. In order to create this model, the present paper clearly identifies whether a tourist will make positive or negative recommendations taking into account their profile, their behaviour during the trip and their level of satisfaction.
In this context, it is important to value the emergence of concepts such as smart tourism,[16-22] and the implementation of advanced analysis tools linked to Big Data,[23-30] Artificial Intelligence [31-36] and Machine Learning[37], applied to tourism and which allow us a better and greater analysis of the behaviour of tourists who arrive at a destination, their reasons for choosing it, the resources used, their itinerary, their expenses, among others. Although supervised ML has been widely researched[38] and applied to a wide variety of problems, there is still a long way to go for studying its contribution to smart tourism.
One of the fundamental information pool when choosing a destination are the opinions and experiences of other tourists beyond marketing campaigns, guides and brochures, websites and tourism fairs. Thus, the likes or comments in different social networks -Facebook, Twitter, Instagram... -, the reviews in TripAdvisor of: hotels, restaurants, means of transport, shops..., or the recommendations of friends, family, colleagues are conditioning significantly the behavioural intention of tourists to visit a certain destination.[39]
This shows the importance of the recommendations of others on the decision to choose a particular destination. According to psychologists, behavioural intentions have a substantial positive effect on one's decision that will eventually translate into behaviour. Therefore, having the intention of traveling to Burgos explains adequately the behaviour of visiting Burgos.[40]
As for the factors that lead tourists to recommend their experience, many authors have shown that overall tourist satisfaction and/or satisfaction with different destination attributes have a positive effect on the probability of recommend it to others.[41] Thus, positive experiences during the trip give rise to positive Worth of Mouth (WOM) recommendations, which are considered the most valuable and reliable information source for potential tourists.[42]

The concept of WOM refers to the exchange of non-commercial, interpersonal information about products or services.[43] Its non-commercial and interpersonal nature explains why consumers trust WOM more than other sources of information.[44] Due to the expansion of the Internet and Web 2.0 and the widespread use of smart technologies and social media in our daily lives, traditional WOM has become electronic word-of-mouth (e-WOM).[45] Litvin et al., defined e-WOM as all informal communications to consumers via the Internet related to the use or characteristics of goods and services or their sellers.[46] According to Pourfakhimi et al., e-WOM plays a decisive role in the choice of hospitality and tourism products by tourists[47] as more and more of them use online social platforms to exchange their opinions on products, services and to learn about other travel experiences before making a decision.[48,49] Today, e-WOM has become one of the main sources of information that allows tourists to share and view comments and reviews on many Internet applications, such as online platforms, blogs, review sites and Social Networking Sites (SNSs).[45] Tourism recommendations that are now present in these media have the speed, scale and reach of mass communication channels as well as an incomparable level of persuasiveness, trust and availability.[47]

For these reasons, we consider it important to analyse the recommendations of tourists who arrive in Burgos and to identify the elements that influence their satisfaction and the possibility of tourists recommending Burgos as a tourist destination to other people in their atmosphere.

Accordingly, the findings of this study provide solid evidence and so, have important theoretical and practical implications.

From the theoretical perspective, most previous work aims to understand the antecedents that influence tourists' behavioural intentions toward a destination, i.e. the likelihood of recommendation and revisiting. In this line, Kozak [50] suggests that tourist harassment has an impact on tourist behaviour since considerable differences in overall holiday satisfaction, word-of-mouth recommendation and repeat visit intentions were observed between those reporting the absence and presence of harassment. Other study[51] focuses on the direct and indirect relationships among online destination brand experience, perceived online destination brand credibility and users' behavioral intentions toward the destination. Regarding heritage tourism, the interplay of visitor engagement, authenticity and destination image in driving revisit and electronic word of mouth (eWOM) intentions is investigated[52]. In terms of domestic tourism, the relationships among motivations, satisfaction and behavioral intentions are analysed[53]. In a similar vein, a model linking involvement, experience quality, satisfaction and recommendation intention has been empirically tested in a cultural tourism destination context[54]. Also, for a specific tourism activity, another study tries to examine how stargazing positively affects tourists' revisit and recommend intentions, based on the theories of peak experiences and conservation of resources.[55]

Unlike these studies, which consider different areas within the tourism business, the current study aims to predict the tourists' behavioural intentions in the context of a rural tourism destination by focusing on willingness to recommend the visited rural destination.

Furthermore, no studies have been found that predict the behaviour of tourists towards rural destinations using Machine Learning techniques, as they generally aim to model this behaviour with structural equation models. [56-57]

The use of ML algorithms and the application of cutting-edge balancing techniques improves the accuracy of tourist recommendation forecasts and so, allows tourism decision makers to use a robust tool. In this sense, policy makers focus on improving the public facilities and services offered to tourists, and managers provide quality tourism experiences that more than meet the needs and wants of visitors.

Definitely, this study aims to guide tourism stakeholders towards the formulation of more effective destination management policies in order to get tourists to recommend the destination. As a result, an increase in tourist demand and tourist satisfaction will be achieved, which will substantially boost the socio-economic effects derived from tourist activity in Burgos in a sustainable way.

In order to forecast tourist recommendations, many different methods can be used. According to previous knowledge[58], Machine Learning (ML) techniques, based on abductive and inductive research, offer a complementary perspective to deductive statistical estimation techniques as the former permit the observation and identification of data patterns that other techniques, such as the classic deductive

regressions, can overlook due to their inherent constraints. Such perspective is adopted in the present study in order to identify and learn the patterns that drive tourist's recommendations.
From this standpoint, Artificial Intelligence in general and different ML alternatives in particular have been previously applied to a wide range of management problems[59-64], including Tourism Management[65-68]. Within this field, it can be observed that in general terms, most published papers using ML techniques focus on predicting tourism demand.[69-73] Other studies have used different ML techniques to create a Tourist information system to intelligently provide personalized information to each user,[74] identify tourism attractiveness in tourist destinations,[75,76] develop a Named Entity Recognition for tourism, travel, hotel and point of interest domain[77] or analyse Korean Medical Tourism from surveys on foreign tourists.[78]
The examination of the tourist's behaviour after the visit is a topic widely covered in the literature. In fact, it is often studied as a construct formed both by the intention of revisiting the destination and by the willingness to recommend it to others through Structural Equation Modelling.[79-83] However, other authors have used ML techniques such as regression analysis,[84,85] confirmatory factor analysis (CFA),[86] decision trees,[87] Naïve Bayes,[88] K-Nearest Neighbours,[89] Support Vector Machine (SVM)[90] or K-Means clustering algorithm[91,92] to specifically analyze tourist recommendations.
Differentiating from previous work, in present research up-to-date ML techniques are applied in order to predict tourist recommendation in Burgos as an appealing destination. More precisely, three different supervised learning models (classifiers) have been applied and are comprehensively compared. Up to the author's knowledge, this is the first time that such problem of predicting recommendations by tourist is addressed by applying the selected classifiers. This paper also proposes the pioneer application of cutting-edge balancing techniques to improve the prediction results. Such advanced ML techniques have been never benchmarked with the target classifiers and data. Trustworthy results are obtained, being an advantageous tool for tourism managers.
Therefore, given the importance of tourists' opinions and experiences for tourism stakeholders, this study pursues two main research objectives:

1. To predict whether or not a tourist will recommend a rural destination visited in the province of Burgos based on its profile.
2. To design a methodology that uses up-to-date Machine Learning techniques together with cutting-edge balancing techniques to improve the prediction.

In this way, this empirical research aims to help the agents involved in the planning, marketing and management of a tourist destination in Burgos to make decisions focused on encouraging tourists to have a favourable behaviour after their experience, recommending the destination to others through WOM comments or social networks.
The remaining sections of this study will be structured as follows: the applied ML techniques are described in section 2. The section 3 will introduce the dataset for analysis, the experiments and the results that have been obtained. Finally, both conclusions and future work proposals are discussed in section 4.

2. Applied Machine Learning

As previously stated, some different ML algorithms are applied in the present work and are described in the following subsections. Experiments have been run on supervised learning algorithms to carry out a data classification. This means that the class information to be predicted (the positive or negative recommendation) is available in the dataset, as described below.

2.1. *Support Vector Machine*

The Support Vector Machine (SVM)[93] is a supervised learning method whose purpose of is to find the hyperplane that maximizes the separation margin of the data, according to the classes defined in the training dataset. It's a widely-applied algorithm implementing Statistical Learning Theory.[94] The aim is to achieve an archetype that helps you to classify those data coming from unlabelled datasets in the different classes. To carry out a one-class classification it uses a loss function called Hinge defined as:

$$L[y, f(x)] = max[0, 1 - y\,f\,(x)] \quad (2.1)$$

In present research, the sigmoidal kernel has been used, whose activation function can be seen below:

$$k(x, y) = tanh(ax^T y + c) \tag{2.2}$$

The variables tuned for the SVM have been gamma and cost. The first one defines the influence of a single example; the lower values indicates far while a high value means closeness. The cost parameter indicates the bend of the model with the data, with a low value it takes smooth decisions. Oppositely, a high value leads to maximize the number of hits at the cost of misclassification.

2.2. *Random Forest*

A Random Forest (RF)[95] can be defined as a combination of decision trees, where each of these classifiers is generated from an input data sample. Based on some input values, each one of the classifier trees gives a vote for the classification of that data sample. The vote is given according to a margin function that is defined[95] as:

$$mr(\mathbf{X}, Y) = P_\Theta(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(\mathbf{X}, \Theta) = j) \tag{2.3}$$

Being the training set drawn at random from the distribution of the random vector *Y*, $\mathbf{X}$, Θ
is the random split selection, and *h()* is the classifier itself.
Each decision tree generates its own prediction from a subset called the bootstrapped[96] dataset. Those instances that have not been selected are in the Out-Of-Bag (OOB) dataset. The characteristics of the data in the bootstrapped dataset are selected randomly, which generates a large variance between the decision trees. Finally, a majority voting scheme is applied. That is, each one of the decision trees votes to see which class corresponds to an instance and the class with the highest number of votes is the one assigned to the instance as the output of the RF.

2.3. *k-Nearest Neighbour*

In the well-known k-Nearest Neighbour (k-NN) algorithm, the examples are classified depending on their nearest neighbours, where k indicates the number of close neighbours that are taken into account.
The distance between two points can be measured based on different metrics, such as the Euclidean, Manhattan or Minkowski distance among many others. In present work, Euclidean distance has been used, defined as:

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2.4}$$

Being p the n-dimensional data sample to be classified and q each one of the nearest neighbours.

3. Analyzed Dataset

The dataset that is analysed in present research is described in subsection 3.1, while the data balancing methods applied to it are discussed in the subsection 3.2.

3.1. *Dataset Features*

The results of this work are based on the 5,294 valid surveys, once the corresponding cleaning process has been carried out on the 5,896 surveys collected by the Province of Burgos Tourism Observatory. These surveys were collected from June 2013 to May 2015 and in 6 rural zones in which the province was segmented. The sample obtained consists of 5,294 valid questionnaires which determine the profile and perception of tourists -people over 16 years old, not resident in nearby municipalities, who visit the province of Burgos. The data collection was carried out by collaborators from tourist establishments and tourist offices and by professors and students from the University of Burgos by conducting face-to-face surveys. Fieldwork in the street has been carried out mainly taking into account the seasonal component of tourism in order to increase the efficiency of data collection. One of the main targets is to predict recommendation of both the majority and minority class.

The purpose of these surveys is to gather representative information from tourists in Burgos on the profile of the tourist, the preparation of their visit, the development of their visit, i.e. the duration of the visit, whether the tourist spends the night or not, the type of accommodation where the tourist is staying, the itinerary that the tourist takes, the total cost of the visit and the activities carried out by the tourist, and finally, the degree of satisfaction, both general and specific, in different aspects, as well as their loyalty.

The variables included in the questionnaire were chosen according to other scientific research[97,98] and to studies carried out by other Spanish Observatories such as Seville, Alicante and Turespaña. However, the questionnaire was gradually adjusted over successive survey periods to the objectives established for this work and to the specific characteristics of tourists in the province of Burgos, and it was finally consolidated in January 2014. The Likert scale is used for questions related to travel satisfaction and experience assessment (assessment of different aspects on a numerical scale ranging from 1 (minimum) to 5 (maximum). The type of sampling is random, stratified by region/district, and the localities are sampled according to the Tourist Potential Index (TPI)[*] calculated for each of the localities of Burgos. For this TPI we have taken into account the equipments - number of lodging places and number of catering establishments - the tourist resources - cultural heritage, natural heritage and festivals of interest - and the accessibility of the destination - roads, railways -.[100]

The gathered data provide a confidence level higher than 98% for global data and a sampling error lower than 2.486%. In summary, the information obtained through the surveys of tourists in the province of Burgos allows us to know their socio-demographic profile, their way of organizing the trip, the use of new technologies throughout the process, the reasons for their visit, the activities carried out or the places visited, their degree of satisfaction and their intention to return or recommend their visit, in other words, the degree of their loyalty.

Once missing values were removed, the following features have been kept from each survey in the 2014 period:

- Companionship: it comprises 6 binary features stating whether the person travelled alone, with his/her partner, on a package tour, with friends, family or other possibilities.
- Occupation: it consists of 6 binary features that indicate which is the current occupation of the interviewee, which are self-employed, employed, housework, retired, unemployed or student.
- Motivation: it consists of 24 binary features that indicate which is or are the main reasons to make the trip among which are: to enjoy the natural environment, for professional reasons, to practice sports activities, to visit archeological sites, to enjoy free time, to visit monuments and heritage, to visit family and friends, for hiking or climbing, to visit new places, to taste the local gastronomy, to attend to events, to walk the Pilgrim's Route to Santiago or the Route of El Cid, for wine tourism and for celebrations reasons.
- Knowledge: it consists of 10 binary features that indicate the means by which you have known the destination such as through guide or brochures friends or family, fairs, Internet, tourist information office, being local, social networks, mass media or other means, and even if the interviewee did not know the destination.
- Type of Transport: it consists of 7 binary features that indicate the means of transport used to get to the destination such as car, bus, train, bike, foot, plane or other means of transport
- Organization: it consists of 5 binary features that indicate the means that the person has used to organize his trip and that can be through the Internet, a travel agency, by phone, or other means, including the option that he has not used any means.
- Main Accommodation: it consists of 5 binary features that indicate which is the main accommodation used to stay overnight during the stay and that can be a hotel, a hostel or pension, rural accommodation, relatives or friends house, camping or other type of establishment.
- Itinerary: it consists of 5 binary features that indicate which is the itinerary made during the visit which can be only Burgos City, Province of Burgos and/or other provinces.

[*] The concept of Tourist Potential Index refers to "the sum of chances of the natural and social environment offers tourist activities, where the main work is focused on the installation or function activation of them to achieve the maximum potential of an area". [99]

- Main Expenses: it consists of 7 binary functions that indicate the main expenses that the visitor has made during the visit, which can be accommodation, bar/café, restaurants, culture/leisure, transportations, shopping or other expenses.
- Sex: it comprises 2 binary features stating whether the person is a man or a female.
- Age: it comprises 5 binary features stating the age range of the visitor which can be less than 25 years, between 25 and 39 years, between 40 and 55 years, between 55 and 65 years and more than 65 years.
- Education: it comprises 2 binary features stating whether the visitor has pre-university or university studies.
- Internet Info: it comprises 2 binary features stating whether or not the visitor used the Internet to search information about the visit and if so, indicate on which web pages.
- Internet to Pay: it comprises 2 binary features stating whether or not the visitor used the Internet to contract or pat any aspect of the visit and if so, indicate on which web pages.
- Knowing Web: it comprises 2 binary features stating whether or not the visitor knows "www.turismoburgos.com" and if so, indicate an assessment between very bad, bad, regular, good and very good.
- Total Expense (€) during the visit per person and day: it comprises 6 binary features stating the expenditure range per person and per day which can be less than 30 euros, between 31 and 60 euros, between 61 and 100 euros, between 101-150 euros, between 151-250 euros and more than 250 euros.
- Before Deciding: it comprises 2 binary features stating whether or not the visitor compared the destination with other options.
- Incidents: it comprises 9 binary features stating whether or not the visitor had any problem or incident during the visit and if so, indicate if the incidence was related to any of these aspects accommodation, museum and monuments, restaurants and bars, information services, security, transportation, hospitality and attention, infrastructures and/or other reasons.
- Region: the place where the tourist come from.
- Areas: district of the province that the tourist is visiting.

For a thorough definition of the analysed features, Table 1 shows the different values taken by the above-described features.

Table 1: Values taken by the dataset features.

Feature	Values					
Companionship	Alone	Partner	Package Tour	Friends	Family	Others
Occupation	Self-employed	Employed	Housework	Retired	Unemployed	Student
Motivation	Nature	Professional	Sports	Archeology	Free Time	Monuments
	Family	Hiking/Climbing	New Place	Gastronomy	Espectacles	Route Santiago
	Route Cid	Wine Turism	Celebration	Other		
Knowledge	Bochures	Turism Office	Media	Friends	Close Place	Party
	RRSS	No Knowledge	Internet	Other		
Type of Transport	Car	Bus	Train	Bike	Foot	Plane
	Others					
Organization	Internet	Telephone	Travel Agency	Others	No planned	
Main Accomodation	Hotel	Hostel	Rural	Family / Friends	Camping	Other
Itinerary	City	Province	Other Provinces			
Main Expenses	Accommodattion	Bar	Restaurants	Culture	Transporting	Shopping
	Others					
Sex	Male	Female				
Age	<25	25<39	39<55	55<65	>65	

Education	Pre-Universitaries	Universitaries				
Internet Info	Yes	No				
Internet to Pay	Yes	No				
Knowing Web	Yes	No				
Total Expense (€)	<30	30<60	60<100	100<150	150<250	>250
Before deciding	Yes	No				
Incidents	Accommodation	Information services	Restaurants	Transports	Attention	Monuments
	Security	Infraestructurtes	Others			
Region	Andalucia	Aragon	Asturias	Baleares	Canarias	Cantabria
	CastillaLaMancha	CastillaLeon	Cataluña	C.Valenciana	Extremadura	Galicia
	La Rioja	Madrid	Murcia	Navarra	PaisVasco	
Areas	1	2	3	4	5	6

The Table 2 shows the demographic data of the sample, including those questions that were not answered (Not Available – NA).

Table 2: Demographic information about survey respondents.

Feature	Values	No. of respondents	(%)
Sex	Male	2751	46.66%
	Female	2960	50.20%
	NA’s	185	3.14%
Age	<25	533	9.04%
	25<39	1791	30.38%
	39<55	1888	32.02%
	55<65	911	15.45%
	>65	453	7.68%
	NA’s	320	5.43%
Education	Undergraduate	1682	28.53%
	Graduate	3622	61.43%
	NA’s	592	10.04%
Occupation	Self-employed	671	11.38%
	Employed	314	5.33%
	Housework	3036	51.49%
	Retired	553	9.38%
	Unemployed	184	3.12%
	Student	98	1.66%
	Other	659	11.18%
	NA’s	381	6.46%

As previously explained, the main target of this work is to predict the tourist recommendations. Consequently, the recommendation (whether positive or negative) information, that is used as the class information, has also been included in the dataset. The recommendation of a tourist destination can be measured through the Net Promoter Score (NPS),[101,102] which seeks to analyze the tourists' experience in the destination in order to group them into different categories.[103-105] According to this popular industry standard, customers are split up into those who can be considered as promoters or prescribers (continue returning to the destination and will recommend it to others), thus boosting the growth linked to the tourist resource or destination that is considered, in our case Burgos. Two types of customers can be distinguished: on the one hand, passive customers who are satisfied but not very enthusiastic customers and can therefore be vulnerable to other alternative tourist offers and, on the other hand, detractors who are unhappy customers who can harm our destination and prevent or reduce its growth by negative word of mouth. [106-108]

3.2. *Data Balancing*

It is well known the negative effect that data imbalance has in the performance of classifiers.[109] However, this characteristic is present in many real-life datasets as the one previously introduced. In the dataset analyzed under study, 98.6% of the data instances belong to the majority class (positive recommendation) while only 1.4% belong to the minority one (negative recommendation). It means that this dataset is strongly unbalanced. In order to address such issue and to reduce classification errors, some balancing algorithms have been used. These balancing algorithms can be classified into three groups: undersampling, oversampling and hybrid ones. They are defined in the following subsections.

When analysing the classification results on imbalance datasets, traditional metrics such as accuracy, are not advisable as they distort the real values. Under such circumstances, there are other metrics whose objective is to value the capability of the model to distinguish between classes. In present experimentation, the metric used to compare results is the Area Under Curve (AUC)[109], defined as the area under the Receiver Operating Characteristic (ROC) curve. It confronts the True Positive Rate with the False Positive Rate. The higher value obtained by AUC (between 0 and 1), the greater capability of the model to distinguish between classes.

3.2.1. *Undersampling Methods*

This type of algorithms has as a strategy to generate a new subset of the data by removing some instances. Usually this is done by reducing the number of instances of the majority class, thus achieving a balance of the data with respect to the target class. Random Under Sampling (RUS) has been applied in this research, which randomly removes instances of the majority class until the data is balanced.

3.2.2. *Oversampling Methods*

These methods generate a new and bigger dataset by generating new data instances "artificially". In this case, the instances that are generated are usually from the minority class. The best known method of this kind, which has been used, is called Random Oversampling (ROS). It randomly selects instances of the minority class and duplicates them, generating new ones to be added to the set. Additionally, other advance oversampling methods have been used:

- **Synthetic Minority Over-sampling Technique (SMOTE)[110]:** It generates new synthetic instances by interpolating different instances of the minority class, using the k-NN algorithm to carry it out.
- **Borderline - SMOTE (BLSMOTE)[111]:** Based on SMOTE, it tries to locate those instances of the minority class located at the borderline and generate new synthetic examples from them. It is based on the fact that most of the times these instances are misclassified and thus reinforced.
- **Density-Based SMOTE (DBSMOTE)[112]:** This variant combines SMOTE with DBSCAN. It is a density clustering algorithm, that recognizes those regions that have a higher density of instances of different classes. DBSMOTE generates new instances by taking this idea as a basis. It could be said that it oversamples data in an opposite way to BLSMOTE because while the latter tries to reinforce those more isolated instances, DBSMOTE works on the more compact ones.

3.2.3. *Hybrid Methods*

These types of techniques combine the views above, eliminating some instances from the majority class and generating new instances from the minority class. To carry out this, the different oversampling and undersampling algorithms can be combined. In present research, combinations of ROS plus RUS and RUS plus SMOTE have been used.

4. Results

In this section, the results of the different classification algorithms combined with the balancing techniques are presented and discussed. For comparison purposes, the results obtained by the classification algorithms directly applied to the unbalanced data are also shown (labelled as "None").

4.1. *Support Vector Machine*

Results obtained by SVM after applying the data balancing algorithms are shown in Table 3. The well-known method called Cross Validation (CV) with 10 folds has been applied to validate these results.

Table 3: AUC values obtained by applying SVM and the data balancing techniques for different vaules of cost and gamma parameters. In bold: best AUC for each Cost value.

Cost	Method/Gamma	0.005	0.01	0.05	0.1	0.15
1	None	0.5000	0.5000	0.5000	0.5000	0.5000
	SMOTE	0.7013	0.6338	0.6220	0.5000	0.5000
	SMOTE + RUS	**0.7175**	0.5988	0.6108	0.5973	0.6111
	ROS	0.6531	0.6759	0.6740	0.6573	0.6422
	RUS	0.6822	0.5765	0.6856	0.6032	0.6385
	ROS + RUS	0.6751	0.4250	0.6435	0.5000	0.5000
	DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000
	BLSMOTE	0.5555	0.5193	0.5876	0.5000	0.5000
3	None	0.5000	0.5000	0.5000	0.5000	0.5000
	SMOTE	0.6978	0.6138	0.6836	0.5000	0.5000
	SMOTE + RUS	0.6720	0.6438	0.6346	0.5483	0.5610
	ROS	**0.7302**	0.6704	0.7187	0.6241	0.5345
	RUS	0.6107	0.5847	0.5946	0.6294	0.6563
	ROS + RUS	0.6331	0.6277	0.6532	0.5000	0.5000
	DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000
	BLSMOTE	0.5377	0.5946	0.5911	0.4941	0.5000
5	None	0.5000	0.5000	0.5000	0.5000	0.5000
	SMOTE	0.6523	0.6508	**0.7059**	0.5111	0.5000
	SMOTE + RUS	0.6616	0.6689	0.7052	0.6272	0.5880
	ROS	0.6169	0.6619	0.6578	0.6183	0.6056
	RUS	0.6594	0.6535	0.6637	0.6508	0.6489
	Over + RUS	0.6134	0.6234	0.6755	0.5000	0.5000
	DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000
	BLSMOTE	0.5412	0.5416	0.6098	0.5684	0.5000
7	None	0.5000	0.5000	0.5000	0.5000	0.5000
	SMOTE	0.6862	0.6693	0.7009	0.6792	0.5000
	SMOTE + RUS	0.7451	0.6562	0.6697	0.5586	0.5079

	ROS	0.6773	0.6331	0.6501	0.6122	0.5834
	RUS	0.5743	0.5300	0.7071	0.6281	0.5505
	ROS + RUS	0.6196	0.7147	**0.7710**	0.5000	0.5000
	DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000
	BLSMOTE	0.5674	0.5524	0.5582	0.6023	0.5191
10	None	0.5000	0.5000	0.5000	0.5000	0.5000
	SMOTE	0.6138	0.6319	0.6439	0.6702	0.5000
	SMOTE + RUS	0.6762	0.6500	0.6805	0.6002	0.6230
	ROS	0.6280	0.6381	0.5915	**0.6892**	0.6233
	RUS	0.6309	0.6798	0.6316	0.6456	0.6538
	ROS + RUS	0.6334	0.6696	0.6543	0.4853	0.5000
	DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000
	BLSMOTE	0.5454	0.5439	0.5856	0.6088	0.5000

From Table 3, a general trend of improvement can be observed when increasing the value of the cost parameter up to 7. For the highest value that has been included in the experiments (10), the obtained AUC values are smaller. Similarly, as far as the Gamma parameter is concerned, it can be seen how the increase in the value means an improvement in the AUC values, until reaching the 0.05 value. For the two highest values of such parameter (0.1 and 0.15), obtained results are worse.

When comparing the results by the balancing algorithms, the worst results are obtained by None (raw imbalanced data) and DBSMOTE. In addition to the AUC values that are close to those that would be obtained by a random classification (0.5), the SVM classifier has assigned all data instances to the majority class. That is, as a result of the data unbalance, the minority class is ignored by the SVM classifier. The DBSMOTE algorithm is not able to overcome such weakness in the data. When considering the results obtained when applying all the other algorithms, none of the methods is the best one in all cases. Nevertheless, it is observed that in general terms the oversampling techniques works better than the undersampling or hybrid ones. However, it is a hybrid algorithm (the combination of ROS and RUS) the one that has enabled the SVM to obtain the best result (0.7710 AUC value), with a cost of 7 and a gamma value of 0.05. It should be noted that the use of improved SMOTE algorithms, (DBSMOTE or BLSMOTE) do not means an improvement with regard to the use of the original method.

4.2. *Random Forest*

Results obtained by RF after applying the data balancing algorithms are shown in Table 4. CV is not required for this method as the error is calculated on the OOB data, that is not used to build the model.

Table 4: AUC values obtained by applying RF and the data balancing techniques for different numbers of trees (ntrees). In bold: best AUC.

Method/ntrees	100	250	500	750	1000	1250
None	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
SMOTE	0.5223	0.5227	0.5000	0.5227	0.5227	0.5227
SMOTE + RUS	0.5223	0.5223	0.5232	0.5223	0.5219	0.5223
ROS	0.5654	0.5651	0.5658	0.5651	0.5654	0.5651
RUS	0.5846	0.6327	0.6286	0.6457	**0.6765**	0.6424
ROS + RUS	0.4398	0.4148	0.4360	0.4375	0.4383	0.4375
DBSMOTE	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000
BLSMOTE	0.5143	0.505	0.5097	0.5166	0.5073	0.5027

In general terms, it can be said that results by RF are worse than those obtained by the SVM algorithm. The only parameter that has been tuned for this classification algorithm is the number of trees; its variation (from 100 to 1250) has a small impact in results and no clear conclusions can be obtained aimed at selecting the best value.
When analysing the results by taking into account the balancing algorithm, it can be said that the pure undersampling method (RUS) has obtained the highest AUC values in all cases. The highest one (0.6765) has been obtained with a large number of trees (1000) while an even larger number (1250) does not imply a best result. In is worth mentioning that this same method, when combined in a hybrid formulation (ROS + RUS), has obtained the lowest AUC values (worse than random) for all the different numbers of trees. Once again, the use of improved SMOTE algorithms, (DBSMOTE or BLSMOTE) has led to deterioration of the AUC values with regard to the use of the original method.

4.3. *k-Nearest Neighbour*

Results obtained by k-NN after applying the data balancing algorithms are shown in Table 5.

Table 5: AUC values obtained by applying k-NN and the data balancing techniques for different values of the k parameter. In bold: best AUC.

Method/k	3	5	7	9
None	0.5000	0.5000	0.5000	0.5000
SMOTE	0.5873	0.6201	0.5090	0.5853
SMOTE + RUS	0.5552	0.6037	0.5322	0.5341
ROS	0.5610	0.5787	0.6041	0.6095
RUS	0.5649	0.5751	0.5269	0.6075
ROS + RUS	0.3959	0.5088	**0.6558**	0.5877
DBSMOTE	0.5020	0.5000	0.5000	0.5000
BLSMOTE	0.5105	0.5392	0.5235	0.5109

4.4. *Discussion*

From a general perspective, it can be said that the results obtained by k-NN, although similar to those obtained by RF, are worse. As it happened in the case of RF, only one parameter has been tuned for this classification algorithm: the number of neighbors. Its variation (from 3 to 9) has a small impact in results and no clear conclusions can be obtained aimed at selecting the best value. Similarly, unsatisfactory results have been obtained by DBSMOTE and BLSMOTE.
The best result (0.6558 AUC value) has been obtained with 7 nearest neighbors and the hybrid balancing algorithm (ROS+RUS). However, this very same algorithm, when using 3 neighbours, has obtained the worst result in present study (0.3959).
It can be said that none of the data balancing methods is the best one in all cases. As it is widely known, a thorough experimental study is required on the dataset under analysis in order to know which method leads to best prediction results. From the experiments that have been carried out, the best result is obtained when applying the SVM (with a cost of 7 and a gamma value of 0.05) to the dataset improved by the ROS + RUS hybrid algorithm. RUS leads to best results for the 3 classifiers that are compared; in isolation for RF and combined with ROS for SVM and k-NN. Improved versions of SMOTE (DBSMOTE and BLSMOTE) imply worse predictions than the seminal SMOTE algorithm.
When comparing the 3 classifiers, best results are obtained by SVM. However, crystal-clear conclusions about the parameter tuning of these models can not be obtained and their values must be adjusted case by case.

5. Conclusions

The main conclusion that arises from present study is that ML can be successfully applied to predict tourist recommendations of rural destinations. Experimental results obtained from real-life data collected in different rural zones of the province of Burgos validate such proposal. The class unbalance in this dataset leads to a poor forecast that can be greatly improved by applying data balancing algorithms.

Theoretically, this study focuses on a specific field (rural destinations), differentiating from most previous studies. Additionally, achieving the two research objectives set, it proposes the use of three ML algorithms in order to classify or predict the positive or negative tourist recommendation from real data obtained from surveys. More important is the application of cutting-edge balancing techniques, a technique not previously used and which allows for more accurate results that do not distort the real data. Indeed, data balancing significantly improves the forecasting accuracy of tourist recommendations, allowing tourism decision makers to use a more trustworthy tool. Furthermore, evidence is provided supporting the idea that the proposed ML techniques can successfully model the idiosyncrasy or tourist visiting rural destinations. As a result, this study significantly contributes to the knowledge in this field.

In terms of practical implications, this study allows policymakers to clearly identify whether a tourist will make positive or negative recommendations taking into account their profile, their behaviour during the trip and their level of satisfaction regarding a rural destination. In this way, and according to the guidelines set by the national, regional and provincial Tourism Strategic Plans, institutions and policy makers can use the findings of this study to develop a more sustainable and competitive tourism model.

On the other hand, there are also managerial implications, referring to travel agents and tour operator managers. To encourage positive recommendations from tourists, managers should improve tourist satisfaction, taking into account their suggestions for improvement as well as their complaints or incidents. Also, anticipating the recommendation of the tourist can help tourism professionals optimize resource allocation, meet the needs of tourists in a valuable way and rationally formulate pricing strategies. Definitely, offering customized services that are the most appropriate for each tourist using a personalized experience strategy allows for a strong competitive advantage.

All in all, this paper contributes to improve the management of tourism in rural areas, being a lifeline for rural communities, that urgently need support not only in Europe but worldwide.

Future work will focus on analysing additional data to forecast tourist recommendations in wider scopes and the application of hybrid intelligent systems to improve the classification results. Furthermore, additional ML models will be investigated in order to provide managers with explanations about the predicted values. Thanks to these outcomes, in subsequent research, the most relevant features and its impact on the recommendations themselves will be studied in order to provide managers and policy makers will useful information.

Acknowledgments

This work was financially supported by the Society for the Development of the Province of Burgos thanks to the collaboration agreement signed between the University of Burgos and this institution for the development of actions included in PEBUR 15-20, in the fields of tourism, economic environment and human capital.

References

1. Z. Osman, Mediating effect of customer satisfaction on service quality and customer loyalty relationship in Malaysian rural tourism*, SSRG Int. J. Economics Manag. Stud.* **2**(1) (2013) 25-37, e-ISSN: 2304-6945
2. World Travel & Tourism Council, Global Economic Impact & Trends (2021), https://wttc.org/Research/Economic-Impact
3. Statista, Statistics and market data on tourism and hospitality (2020), https://es.statista.com/sectores/1186/turismo-y-hosteleria/
4. INE, Cuenta Satélite del Turismo de España (CSTE). Revisión estadística 2019 (2021), https://www.ine.es/prensa/cst_2020.pdf

5. JCyL, Boletines de Coyuntura (2021), https://www.turismocastillayleon.com/es/espacio-profesionales/boletines-coyuntura.
6. INE, Operaciones Estadísticas de Hostelería y Turismo (2021), https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735576863
7. OTB, Publicaciones oficiales del Observatorio de Turismo de Burgos (2021), https://www.ubu.es/observatorio-de-turismo-de-burgos/publicaciones
8. EXCELTUR, Principales Indicadores del Monitor de Competitividad Turística de los Destinos Urbanos Españoles - 2016 Urbantur (2017), https://www.exceltur.org/urbantur/
9. UNESCO, World Heritage List (2021), https://whc.unesco.org/en/list/
10. UNESCO, Creative Cities Network (2021), https://en.unesco.org/creative-cities/burgos
11. UNWTO, Sustainable Development Goals (2019), https://www.un.org/sustainabledevelopment.
12. Ministry of Industry, Trade and Tourism. Government of Spain, Estrategia de Turismo Sostenible de España 2030 (2019), https://turismo.gob.es/es-es/estrategia-turismo-sostenible/Paginas/Index.aspx
13. Consejo Español de Turismo, Plan del Turismo Español Horizonte 2020 (2007), https://www.tourspain.es/es-es/Conozcanos/Documents/HistoricoPoliticaTuristica/Horizonte%202020%20-%20Plan%20Turismo%20Espa%C3%B1ol%200812.pdf
14. Junta de Castilla y León, Plan Estratégico de Turismo de Castilla y León 2019 – 2023 (2019), https://www.turismocastillayleon.com/es/espacio-profesionales/planes-estrategicos/plan-estrategico-turismo-castilla-leon-2019-2023.ficheros/172680-Plan%20Estrat%C3%A9gico%20de%20Turismo%20de%20Castilla%20y%20Le%C3%B3n%202019-2023.pdf
15. SODEBUR, Plan Estratégico Burgos Rural 2015-2020 (2015), http://pebur1520.org/
16. Asociación Española de Normalización, UNE 178501:2018 Management system of smart tourist destinations. Requirements. (2018), https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060239
17. A. Matos, B. Pinto, F. Barros, S. Martins, J. Martins and M. Au-Yong-Oliveira, Smart Cities and Smart Tourism: What Future Do They Bring?, *World Conference on Information Systems and Technologies*. (2019) 358-370, doi:10.1007/978-3-030-16187-3_35
18. D.Z. Jovicic, From the traditional understanding of tourism destination to the smart tourism destination, *Curr. Issues Tour*. **2**(3) (2019) 276-282, doi:10.1080/13683500.2017.1313203
19. P. Del Vecchio, G. Mele, V. Ndou and G. Secundo, Creating value from social big data: implications for smart tourism destinations, *Inform. Process. Manag*. **54**(5) (2018) 847-860, doi:10.1016/j.ipm.2017.10.006
20. C. Koo, L. Mendes Filho and D. Buhalis, Smart tourism and competitive advantage for stakeholders, *Tour. Rev*. **74**(1) (2019) 1-128, doi:10.1108/TR-02-2019-208
21. F. Mehraliyev, Y. Choi and M.A. Köseoglu, Progress on smart tourism research, *J. Hosp. Tour. Technol*. **10**(4) (2019) 522-538, doi:10.1108/JHTT-08-2018-0076
22. D. Buhalis, Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article, *Tour. Rev.* (2019), doi:10.1108/TR-06-2019-0258/
23. M. Inanc–Demir and K. Metin, *Big Data and Innovation in Tourism, Travel, and Hospitality* (Springer, Singapore, 2019), pp. 213-223, doi:10.1007/978-981-13-6339-9_13
24. M. Mariani, Big Data and analytics in tourism and hospitality: a perspective article, *Tour. Rev*. (2019), doi:10.1108/TR-06-2019-0259
25. A.R. Alaei, S. Becken and B. Stantic, Sentiment analysis in tourism: capitalizing on big data, *J. Travel Res.* **58**(2) (2019) 175-191, doi:10.1177/0047287517747753
26. J. Li, L. Xu, L. Tang, S. Wang and L. Li, Big data in tourism research: A literature review, *Tour. Manag*. **68**(2018) 301-323, doi:10.1016/j.tourman.2018.03.009
27. M.H. Salas-Olmedo, B. Moya-Gómez, J.C. García-Palomares and J. Gutiérrez, Tourists' digital footprint in cities: Comparing Big Data sources, *Tour. Manag*. **66**(2018) 13-25, doi:10.1016/j.tourman.2017.11.001
28. P. Centobelli, and V. Ndou, Managing customer knowledge through the use of big data analytics in tourism research, *Curr. Issues Tour*. (2019) 1-22, doi:10.1080/13683500.2018.1564739
29. W. Höpken, M. Fuchs, and M. Lexhagen, Big Data Analytics for Tourism Destinations, *Encyclopedia of Information Science and Technology*, (2018) 349-363.
30. T. Gajdošík, Big Data Analytics in Smart Tourism Destinations. A New Tool for Destination Management Organizations?, *Smart Tourism as a Driver for Culture and Sustainability*, (2019) 15-33, doi:10.1007/978-3-030-03910-3_2
31. L.N. Cain, J.H. Thomas and M. Alonso Jr, From sci-fi to sci-fact: the state of robotics and AI in the hospitality industry, *J. Hosp. Tour. Technol*. **10**(4) (2019) 624-650, doi:10.1108/JHTT-07-2018-0066
32. A. N. Kazak, P. V. Chetyrbok and N. N. Oleinikov, Artificial intelligence in the tourism sphere, *IOP Conference Series: Earth and Environmental Science*. **421**(4) (2020), doi:10.1088/1755-1315/421/4/042020
33. G. Telonis and P.P. Groumpos, The Artificial Intelligence in the Tourism Sector: Issues, Challenges and Opportunities, *Proc. UBT Int. Conf.*, 2018, doi:10.33107/ubt-ic.2018.208
34. S. Ivanov and C. Webster, *Robots, Artificial Intelligence and Service Automation in Travel, Tourism and Hospitality* (Emerald Publishing Limited, 2019), pp. 7-37, doi:10.1108/978-1-78756-687-320191001

35. R. Baggio, Measuring tourism: methods, indicators, and needs, *The Future of Tourism*, (2019) 255-269, doi:10.1007/978-3-319-89941-1_13
36. C. Webster, and S. Ivanov, Robotics, artificial intelligence, and the evolving nature of work, Digit. Transform. Bus. Soc. (2020) 127-143, doi:10.1007/978-3-030-08277-2_8
37. A. Aluri, B.S. Price, and N.H. McIntyre, Using machine learning to cocreate value through dynamic customer engagement in a brand loyalty program, *J. Hosp. Tour. Res.* **43**(1) (2019) 78-100, doi:10.1177/1096348017753521
38. G. Kou, Y. Lu, Y. Peng, and Y. Shi, Evaluation of Classification Algorithms Using MCDM and Rank Correlation, *Int. J. Inf. Technol. Decis. Mak.* **11**(1) (2012) 197-225, doi: 10.1142/S0219622012500095
39. A. Almeida-Santana and S. Moreno-Gil, Understanding tourism loyalty: Horizontal vs. destination loyalty, *Tour. Manag*. **65**(2018) 245-255.
40. M.S. Soliman and M.A. Abou-Shouk, Predicting Behavioural Intention of International Tourists Towards Geotours, *Geoheritage*. **9**(2017) 505–517, doi:10.1007/s12371-016-0200-5
41. J. Alegre and M. Cladera, Analysing the effect of satisfaction and previous visits on tourist intentions to return, *Eur. J. Mark*. **43**(5/6) (2009) 670-685, doi:10.1108/03090560910946990
42. H. C. Wu and T. Li, A study of experiential quality, perceived value, heritage image, experiential satisfaction, and behavioral intentions for heritage tourists, *J. Hosp. Tour. Res*. **41**(8) (2017) 904-944, doi:10.1177/1096348014525638
43. J. Arndt, *Word of mouth advertising and information communication*, in Cox, D.F. (Ed.), Risk Taking and Information Handling in Consumer Behavior, (Harvard University, Boston, 1967).
44. J. Brown, A.J. Broderick and N. Lee, Word of mouth communication within online communities: conceptualizing the online social network, *J. Interact. Mark.* **21**(3) (2007) 2-20.
45. A. Reyes-Menendez, J. R. Saura and J. G. Martinez-Navalon, The Impact of e-WOM on Hotels Management Reputation: Exploring TripAdvisor Review Credibility With the ELM Model, *IEEE*, **7** (2019) 68868-68877, doi:10.1109/ACCESS.2019.2919030.
46. S. W. Litvin, R. E. Goldsmith and B. Pan, Electronic word-of-mouth in hospitality and tourism management, *Tour. Manage.* **29**(3) (2008) 458-468, doi: 10.1016/j.tourman.2007.05.011
47. S. Pourfakhimi, T. Duncan and W. J. Coetzee, Electronic word of mouth in tourism and hospitality consumer behaviour: state of the art, *Tour. Rev.* (2020), ISSN:1660-5373
48. X. Cheng and M. Zhou, Study on effect of eWOM: A literature review and suggestions for future research, *Proc. Int. Conf. Manage. Service Sci.* (2010), 1-4.
49. T. Daugherty and E. Hoffman, eWOM and the importance of capturing consumer attention within social media, *J. Marketing Commun.*, 20(1) (2014) 82-102, doi: 10.1080/13527266.2013.797764
50. M. Kozak, Tourist harassment: A marketing perspective, *Ann. Tour. Res.* **34**(2) (2007) 384-399, doi: 10.1016/j.annals.2006.10.003
51. T. Kee Hui, D. Wan, and A. Ho, Tourists' satisfaction, recommendation and revisiting Singapore, *Tour. Manag*. **28**(4) (2007) 965-975. doi: 10.1016/j.tourman.2006.08.008
52. S. Mostafa Rasoolimanesh, S. Seyfi, C. Michael Halland and P. Hatamifar, Understanding memorable tourism experiences and behavioural intentions of heritage tourists, *J. Dest. Mark. Manag.* **21**(2021) 100621, doi: 10.1016/j.jdmm.2021.100621
53. B. Esubalew Bayih and A. Singh, Modeling domestic tourism: motivations, satisfaction and tourist behavioral intentions, *Heliyon,* **6**(9) (2020) e04839, doi:10.1016/j.heliyon.2020.e04839
54. M. Cevdet Altunel and B. Erkurt, Cultural tourism in Istanbul: The mediation effect of tourist experience and satisfaction on the relationship between involvement and recommendation intention, *J. Dest. Mark. Manag.* **4**(4) (2015) 213-221, doi: 10.1016/j.jdmm.2015.06.003
55. T. Li, Universal therapy: A two-stage mediation model of the effects of stargazing tourism on tourists' behavioral intentions, *J. Dest. Mark. Manag.* **20**(2021) 100572, doi: 10.1016/j.jdmm.2021.100572
56. Z. Osman, Mediating effect of customer satisfaction on service quality and customer loyalty relationship in Malaysian rural tourism, *Int. J. Economics Manag. Stud.* **2**(1) (2013) 25-37, e-ISSN: 2304-6945
57. S. M. Correia-Loureiro and F. Miranda-González, The importance of quality, satisfaction, trust, and image in relation to rural tourist loyalty, *J. Travel Tour. Mark*. **25**(2) (2008) 117-136, doi: 10.1080/10548400802402321
58. P. Choudhury, R.T. Allen, and M.G. Endres, Machine Learning for Pattern Discovery in Management Research, *Strategic Manag. J.* **42**(1) (2021) 30-57, doi:10.1002/smj.3215
59. A. Trunk, H. Birkel, and E. Hartmann, On the current state of combining human and artificial intelligence for strategic organizational decision making. *Bus. Res.* **13**(2020) 875–919, doi:10.1007/s40685-020-00133-x
60. B. van Liebergen, Machine learning: A revolution in risk management and compliance?, *J. Financ. Transform.*, Capco Institute, **45**(2017) 60-67.
61. J. Ifft, R. Kuhns, and K. Patrick, Can machine learning improve prediction–an application with farm survey data, *Int. Food Agribusiness. Manag. Rev*. **21**(8) (2018) 1083-1098, doi: 10.22004/ag.econ.284919
62. L. Sáiz, A. Pérez, Á. Herrero, and E. Corchado, Analyzing key factors of human resources management. In Int. Conf. on Intelligent Data Eng. and Automated Learn. (2011) 463-473, doi: 10.1007/978-3-642-23878-9_55

63. Á. Herrero and A. Jimenez, Improving the management of industrial and environmental enterprises by means of soft computing, *Cybern. and Syst.* **50**(1) (2019) 1-2, doi: 10.1080/01969722.2019.1560961
64. Á. Herrero, A. Jimenez, and S. Bayraktar, Hybrid unsupervised exploratory plots: a case study of analysing foreign direct investment, *Complexity*, **2019**(2019), doi: 10.1155/2019/6271017
65. C.H. Ku, Y.C. Chang, Y. Wang, C.H. Chen, and S.H. Hsiao, Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews and Responses. Proc. of the 52nd Hawaii Int. Conf. on System Sciences, 2019, doi:10.24251/HICSS.2019.634
66. K.A. Astuti and G.S. Darma, Community-based tourism: measuring readiness of artificial intelligence on traditional village, *Int. J. Soc. Sci. Humanity Stud*. **3**(3) (2019) 81-89, doi:10.29332/ijssh.v3n3.352
67. O. Rodriguez, Artificial Intelligence in the Business of Tourism: A Market Strategy in the UK Travel Distribution (Doctoral dissertation, University of East London), 2019. doi:10.15123/PUB.8036
68. O. Panasiuk, Z. Akbar, U. Şimşek and D. Fensel, Enabling conversational tourism assistants through schema.org mapping, *European Semantic Web Conference*. (2018) 137-141, doi:10.1007/978-3-319-98192-5_26
69. S. Cankurt and A. Subasi, Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components, *Balkan J. Electr. Comput. Eng*. **3**(1) (2015) 42-49.
70. O. Claveria, E. Monte and S. Torra, Combination forecasts of tourism demand with machine learning models, *Appl. Econ. Lett*. **23**(6) (2015) 428-431, doi:10.1080/13504851.2015.1078441
71. S. Sun, Y. Wei, K. L. Tsui, and S.Wang, Forecasting tourist arrivals with machine learning and internet search index, *Tour. Manag*. **70**(2019) 1-10, doi:10.1016/j.tourman.2018.07.010
72. Y.Y. Liu, F.M. Tseng and Y.H. Tseng, Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model, *Technol. Forecast. Soc. Change.* **130**(2018) 123-134, doi:10.1016/j.techfore.2018.01.018
73. O. Claveria, E. Monte, and S. Torra, Modelling tourism demand to Spain with machine learning techniques. The impact of forecast horizon on model selection*, Revista de Economía Aplicada de Cornell University*. **72**(24) (2016) 109-132.
74. R.S. Ganga, P.C.P. Reddy and B.C. Mohan, System for Intelligent Tourist Information using Machine Learning Techniques*, Int. J. Appl. Eng. Res*. **13**(7) (2018) 5321-5327.
75. S. Giglio, F. Bertacchini, E. Bilotta, and P. Pantano, Using social media to identify tourism attractiveness in six Italian cities, *Tour. Manag*. **72**(2019) 306-312, doi:10.1016/j.tourman.2018.12.007
76. S. Giglio, F. Bertacchini, E. Bilotta and P.Pantano, Machine learning and points of interest: typical tourist Italian cities, *Curr. Issues Tour*. (2019) 1-13, doi:10.1080/13683500.2019.1637827
77. J. Vijay and R. Sridhar, A Machine Learning Approach to Named Entity Recognition for the Travel and Tourist Domain, *Asian J. Inf. Technol*. **15**(21) (2016) 4309-4317.
78. J. H., Huh, H. B. Kim and J. Kim, A method of modeling of basic big data analysis for Korean medical tourism: A machine learning approach using apriori algorithm, *Int. Conf. on Information Science and Applications*, 2017, pp. 784-790, doi:10.1007/978-981-10-4154-9_91
79. C. Eusébio and A. L. Vieira, Destination attributes' evaluation, satisfaction and behavioural intentions: A structural modelling approach, *Int. J. Tour. Res*. **15**(1) (2013) 66-80.
80. H. C. Wu and T. Li, A study of experiential quality, perceived value, heritage image, experiential satisfaction, and behavioral intentions for heritage tourists, *J. Hosp. Tour. Res*. **41**(8) (2017) 904-944.
81. M. S. Soliman, and M. A. Abou-Shouk, Predicting behavioural intention of international tourists towards geotours, *Geoheritage*. **9**(4) (2017) 505-517, doi:10.1007/s12371-016-0200-5
82. M. Bigovic and J. Prašnikar, Predicting tourists' behavioural intentions at the destination level, *Curr. Issues Tour*. **18**(8) (2015) 744-764.
83. N. Chen, L. Dwyer and T. Firth, Factors influencing Chinese students' behavior in promoting Australia as a destination for Chinese outbound travel, *J. Travel Tour. Mark*. **32**(4) (2015) 366-381.
84. M. F. A. A. Muhamed, F. A. Jabar, S. N. S. Wahid, H. Paino and M. R. M. Dangi, Predicting Customer Recommendation Towards Homestay at West Pahang Region, *Adv. Sci. Lett*. **23**(4) (2017) 2978-2982.
85. Z., Zhang and Y. Yang, The Power of Expert Identity: How Website-Recognized Expert Reviews Influence Travelers' Online Rating Behavior, *Tour. Manag.* **55**(2016) 15-24.
86. J. F. Petrick, The roles of quality, value, and satisfaction in predicting cruise passengers' behavioral intentions, *J. Travel Res*. **42**(4) (2004) 397-407.
87. P. Sun, D.A. Cárdenas and R. Harrill, Chinese Customers' Evaluation of Travel Website Quality: A Decision-Tree Analysis, *J. Hosp. Mark. Manag*. **25**(4) (2016) 476–497.
88. H. Kang, S.J. Yoo and D. Han, Senti-Lexicon and Improved Naïve Bayes Algorithms for Sentiment Analysis of Restaurant Reviews, *Expert. Syst. Appl*. **39**(5) (2012) 6000–6010.
89. L. Dey, S. Chakraborty, A. Biswas, B. Bose and S. Tiwari, Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier, *Int. J. Inf. Eng. Electr. Bus*. **8**(4) (2016) 54–62.
90. Y.-C. Chang, C.-H. Ku and C.-H. Chen, Social Media Analytics: Extracting and Visualizing Hilton Hotel Ratings and Reviews from Tripadvisor, *Int. J. Inf. Manage*. (2017).

91. T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, Travel Route Recommendation Using Geotagged Photos, *Knowl. Inf. Syst*. **37**(1) (2013) 37–60.
92. X. Zhang and Q. Yu, Hotel Reviews Sentiment Analysis Based on Word Vector Clustering, 2nd IEEE Int. Conf. on Computational Intelligence and Applications (ICCIA), (2017) 260– 264.
93. C. Cortes and V. Vapnik, Support-Vector Networks, *Mach. Learn*. **20**(3) (1995) 273-297, doi:10.1023/a:1022627411411
94. V. Vapnik, *The nature of statistical learning theory* (Springer-Verlag, New York, 1995).
95. L. Breiman, Random Forests, *Mach. Learn*. **45**(1) (2001) 5-32, doi:10.1023/A:1010933404324
96. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap* (CRC press, 1994).
97. T. K. Hui, D. Wan, and A. Ho, Tourists' satisfaction, recommendation and revisiting Singapore, *Tour. Manag.* **28**(4) (2007) 965-975, doi: 10.1016/j.tourman.2006.08.008
98. B. Ozdemir, A. Aksu, R. Ehtiyar, B. Çizel, R. B. Çizel, and E. T. İçigen, Relationships among tourist profile, satisfaction and destination loyalty: Examining empirical evidences in Antalya region of Turkey, *J. Hospit. Market. Manag.* ***21***(5) (2012) 506-540, doi:10.1080/19368623.2012.626749
99. M.J. Mikery Gutiérrez and A. Pérez-Vázquez, Métodos para el análisis del potencial turístico del territorio rural, *Revista mexicana de ciencias agrícolas*, **5**(SPE9) (2014) 1729-1740, doi: 10.29312/remexca.v0i9.1060
100. A. Alvear-González and S. Aparicio-Castillo, Índice de Potencialidad Turística en Burgos: Aplicación en el diseño y planificación del trabajo de campo, (2014) Universidad de La Laguna.
101. R. Owen, Net Promoter Score and Its Successful Application, *Marketing Wisdom*. (2019) 17-29, doi:10.1007/978-981-10-7724-1_2
102. N.I. Fisher, and R.E. Kordupleski, Good and bad market research: A critical review of Net Promoter Score, *Appl. Stoch. Models Bus. Ind.* **35**(1) (2019) 138-151, doi:10.1002/asmb.2417
103. R. Haarhoff, Tourist perceptions of factors influencing destination image: a case study of selected Kimberley resorts, *African J. Hosp. Tour. Leis.* **7**(4) (2018).
104. H., Hanin, H. Hosni, H. Kievit, P. D. Drakos, G. Angelakis and G. van Dijk, Measuring Tourist Attitudes and Consumer Behavior, *Int. J. Food Beverage Manufacturing Bus. Models*. **3**(2) (2018) 29-41, doi:10.4018/IJFBMBM.2018070103
105. C. Lewis and D. Donnelly, Towards challenging the use of the NPS in practice as a measure of visitor satisfaction. CAUTHE 2019*:* Sustainability of Tourism, Hospitality & Events in a Disruptive Digital Age: Proc. of the 29th Annual Conf., (2019).
106. A.-G. Hoarau, L. Gendre and V. Ricard, Effective or offensive advertising?: An exploratory study on negative Word-of-Mouth and consumers' perception, *Bachelor Thesis on Business Administration* **62**(2018).
107. T. Ma and F. Ding, Research on the dynamic effect of the intelligent urban experience to the tourists' two-way internet word-of-mouth, *Int. J. Commun. Syst*. **31**(16) (2018), doi: 10.1002/dac.3467
108. C. Wu and S. Zhang, Impact of Tourists' Negative Word of Mouth on the Tourism Image of Coastal City of Dalian: Text Analysis Based on Comments from Tourists Network, *Ocean Development and Management*, **7**(2018) 17.
109. V. López, A. Fernández, S. García, V. Palade and F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, *Inf. Sci*. **250**(2013) 113 – 141, doi:10.1016/j.ins.2013.07.007.
110. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *J. Artif. Intell. Res*. **16**(2002) 321–357, doi:10.1613/jair.953.
111. H. Han, W.Y. Wang and B.H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *Lect. Notes Comput. Sci*. **3644** (2005) 878-887, doi:10.1007/11538059_9
112. C. Bunkhumpornpat, K. Sinapiromsaran and C. Lursinsap, DBSMOTE: Density-Based Synthetic Minority Over-sampling Technique, *Appl. Intell*. **36** (2012) 664–684, doi:10.1007/s10489-011-0287-y