Predicting and determining antecedent factors of tourist village development using naive bayes and tree algorithm

Nafiah Ariyani [1]*, Akhmad Fauzi [2], Farhat Umar [3]

[1]*Sahid University, Indonesia*
[2]*IPB University, Indonesia*
[3]*Sahid University, Indonesia*

*Corresponding Author: *nafiah_ariyani@usahid.ac.id*

Abstract: This study aims to predict the progress status of tourism villages in the Kedung Ombo area, Java, Indonesia, and find the antecedent factors of the progress of tourism villages in Indonesia. This study uses a modern approach, namely data mining. Data sources for tourist villages use the data available on the Google link and the observation method. The prediction technique uses the Naïve Bayes machine learning algorithm and Tree Decision on Orange 3.3.0 software. The number of tourist villages analyzed was 126. The results showed that all tourist villages in the Kedung Ombo area were at the development level of the four tourist village classifications of the Ministry of Tourism and Creative Economy. The antecedent factors for the progress of tourism villages are the completeness of ICT facilities, multi-stakeholder partnerships, strong government support, community involvement, and various attractions. Another finding is that the Tree Decision algorithm provides better predictions than the Naïve Bayes method. The results of this study can be used to design policies for developing tourist villages throughout Indonesia.

Keywords: data mining, decision tree algorithm, naïve bayes, tourist village development

Introduction

Rural tourism is becoming a significant trend, especially in developing countries, as a manifestation of the concept of community-based tourism (CBT) which is deemed to be able to counteract the negative impacts of mass tourism related to social equality, environmental degradation, and preservation of community culture (Khalid et al., 2019; Muganda et al., 2013). Rural tourism could be a vehicle of sustainable development that could generate employment and income creation, prevent rural exodus, encourage socio-economic networks, save and enhance cultural and natural heritage, and improve the quality of life for residents (Rodrigues et al., 2021; Powell et al., 2018). Gohori & van der Merwe (2020) propose a reciprocal relationship between tourism, poverty alleviation, and community development. In the context of sustainability, rural tourism is synonymous with sustainable tourism development in nature, scale, character, and development process (Sharpley & Roberts, 2004).

Rural development through tourism has become an essential concept for both developed and developing countries as it represents a process of mobilizing innovation and aligning change, focusing on increasing opportunities for the population, economic growth, protection of natural resources, and social equality. Rural tourism is considered capable of supporting development in rural areas that are structurally weak (Neumeier & Pollermann, 2014). In rural areas, especially in developing countries, tourism is enthusiastically accepted as a panacea for revitalizing the rural economy: thus prompting many government agencies, particularly those related to tourism, to invest in promoting more sustainable forms of community-based rural tourism (Kamarudin et al., 2020). The tourist village program is a priority rural development program (Ariyani et al., 2022).

In Indonesia, rural tourism is manifested in the form of tourism village development (TVD), which since 2021 has been set by the Coordinating Ministry for Economic Affairs. TVD is directed

toward increasing economic growth for people's welfare, eradicating poverty, overcoming unemployment, preserving nature, the environment, and resources, and promoting culture. The development of tourist villages is expected to be one form of accelerated village development in an integrated manner to encourage the village's social, cultural, and economic transformation. The success of the tourist village could become leverage for the village and regional economy, which will ultimately encourage national economic growth. In Indonesia, tourist villages are categorized as pilot, developing, developed, and independent villages (Ariyani & Fauzi, 2023).

Along with these provisions, various rural areas develop tourist villages. No less than thousand eight hundred and seventy tourist villages are spread throughout Indonesia. One area designated as a tourist village is the Kedung Ombo reservoir in Central Java. There are eight tourist destinations with the concept of a tourist village. Against the background of the limited benefits of reservoirs for communities in the upstream area, several community groups take advantage of the reservoir panorama as a tourist attraction. The existence of tourist villages in this area was expected to solve this area's socio-economic problems and infrastructure limitations.

However, to date, these efforts have not shown significant progress. Instead of creating alternative jobs for the local community and reducing poverty, the tourist villages have not been able to bring in enough visitors even though they have been developed over the years. As an integral part of the national tourism development agenda, this community initiative in the Kedung Ombo area needs support.

One of the reasons for the condition of tourist villages in the Kedung Ombo area is the development approach, which is based more on conventional methods that focus on in situ characteristics of the villages. Although this approach has advantages related to the ability to identify local needs, it has disadvantages due to the lack of understanding of hidden factors that may determine the development of the tourist village. For example, several tourist villages have the same characteristics, yet they may produce different outcomes related to the performance of the tourist village.

This shortcoming can be bridged with a modern approach based on data mining through machine learning such as Naïve Bayes and Tree. Machine learning is an analytical method using past information, usually in the form of available electronic data, to make accurate predictions. In principle, this approach is carried out by harvesting big data. Then, machine learning will produce forecasts regarding whether the tourist villages developed successfully or not. In addition, machine learning algorithms such as decision trees can be used to determine what factors (the antecedent factors) produce a pattern of success for tourist villages to develop.

The use of machine learning has been widely used in various studies, both economic, social, environmental, technological, and political (Ariyani et al., 2022). Naïve Bayes (NB) algorithm is trained and is used to classify or determine the gage length of the wheat straw based on target mechanical properties (Naik & Kiran, 2018). In the field of tourism, machine learning, for instance, has been used to forecast demand for tourism (Ahmed et al., 2007; Li, 2022; Yu & Chen, 2022), marketing strategies for rural tourism (Xie & He, 2022), and recommendations for smart tourism strategies (Ho, 2022). The use of machine learning for tourism cases in Indonesia is limited to several aspects, such as predictions of international tourist arrivals during the Covid-19 period (Andariesta & Wasesa, 2022) or estimation of international tourists (Purnaningrum & Athoillah, 2021). There remains a lot of room for tourism analysis with machine learning that can be done for the current case, such as predicting how the tourist village will develop or not and what antecedent factors determine the development of the tourist area. This study aims to predict the progress of tourist villages in Kedung Ombo area and determine the antecedent factors of successful development.

Rural tourism has various definitions with a very broad scope. Researchers from different countries develop their definitions based on the unique experiences or contexts they encounter (Nair et al., 2015). According to Tang (2022), rural tourism has no single definition. This statement implies that determinants of the success of tourist villages are difficult to identify (Ayazlar & Ayazlar, 2015).

In their research, Rodrigues et al. (2021) found that community involvement is a critical factor in developing tourist villages. Community participation is vital in creating sustainable community-based tourism (Amin & Ibrahim, 2015). In line with that, Bajrami et al. (2020) stated that

one of the critical factors for the success of sustainable tourism in rural areas is the support from community members. Community participation can empower communities and significantly contribute to rural tourism development, which helps eradicate poverty, depopulation, hydro-geological instability, and degradation of cultural heritage and landscapes (Basile et al., 2021). Local people play a crucial role in developing sustainable tourism in rural areas (Yu et al., 2018). They must be involved in decisions that will influence themselves, their families, and their communities (Powell et al., 2018).

In addition to requiring the participation of all stakeholders, sustainable tourism development requires involvement from all relevant stakeholders and strong political leadership to ensure broad participation and build consensus (Kantsperger et al., 2019). The government is a leading actor with a paramount role in tourism development (Firdaus et al., 2021; McLennan et al., 2014). A more participatory rural development involving horizontal and vertical coordination places the government as the driver of the partnership process between stakeholders in helping to develop and oversee the strategic direction of rural tourism development (Koopmans et al., 2018). This interventionist approach adopting a more active government role in tourism development has become widespread worldwide, even in countries with different ideologies (Liu et al., 2020).

In its development, CBT requires public partnerships both in local and global contexts (Purbasari & Manaf, 2018). Regardless of the type of CBT services, these ventures should remain wholly owned, managed, and controlled by community members (or groups of independent micro and small experiences under the same CBT management organization). Meanwhile, external partners should provide facilitative and other support services instead of being a partner in the CBT venture itself (Mtapuri & Giampiccoli, 2013).

Research conducted by Kristianto et al. (2019) in the tourist village of Pahawang, Lampung, Indonesia, found the antecedents of the success of rural tourism development include attractions, amenities, accessibility, image, human resources, and tourism prices. The determination of ticket prices in rural tourism is a factor directly related to the interests of each stakeholder (Wu et al., 2017). Another factor is infrastructure, including homestays (Bhalla et al., 2016). Information and communication technology (ICT) plays a role in supporting the development of rural tourism as media promotion, booking, and payment facilities for transactions (Hidayatullah et al., 2018; Waghmode, M. L., & Jamsandekar, 2013). ICT-based management can be formulated into a tourist village development strategy (Pantiyasa et al., 2019).

León-Gómez et al. (2021) found that the criteria for the success of CBT from evaluations in several countries in Asia: involving the masses of people; benefits being distributed equitably to all communities; good tourism management; strong partnerships both inside and outside; unique attractions; environmental preservation; the uniqueness of the location; the facilitation of existing embryo activities; involvement of the broader community as tourism actors; and partnerships. In a different context, Yang et al. (2016) stated that tourism resources, tourist traffic, and social and economic factors drive island tourism. Government policies, tourism companies, and the tourist market are external drivers.

The diversity of findings on the determinants of the success of the development of tourist villages can be used as material for analysis in developing tourist villages in Indonesia. The in-situ characteristics of tourist villages in Indonesia are closely related to the community's geographical and socio-economic conditions and require guidance to be developed successfully in the future time spectrum.

This research has two objectives. First, it is to predict village tourist development in the Kedung Ombo area. Second, it is to analyze the antecedent factors that determine the success of tourist village development.

Methodology

This study uses a modern approach to achieve research objectives, namely data mining. Data mining is a data acquisition method and information collection methodology that can guide decision-making efficiently by extracting and analyzing accumulated datasets (big data) to obtain helpful knowledge (Adekitan et al., 2019). To achieve the first objective, Naïve Bayes and Trees algorithm are used, and for the second goal, the Decision Tree algorithm is used. All analyzes were performed using the Orange 3.3.0 software.

To analyze the tourist village development profile, use nine (9) attributes: (1) the level of community participation, (2) the variety of attractions, (3) the level of government support; (4) the intensity of partnership, (5) the completeness of infrastructure, (6) the distance of object location to the main road, (7) the ticket price, (8) management, and (9) the application of information and communication technology (Table 1). For analysis progress status of tourist villages, the data of tourist villages are grouped into developed and progressive tourism villages based on the modified criteria of the Ministry of Tourism and Creative Economy: (1) the number of tourist visits both from within and outside the region, (2) the community can manage tourism businesses, and (3) the creation of employment from the tourism.

Table 1. Attributes of Tourism Village Development and Measurement Scale

Attribute	Measurement
Community Participation	Active – Passive
Attraction	Single – Diversity
Government Support	Strong – Weak
Partnership	Mono - Multi (government, private, academics)
Infrastructure facilities (meeting hall, prayer room, toilets, gift stalls, food stalls, selfie spots, and homestays)	Complete – Partial
Location Distance	0-10 km; 10-30 km; ≥ 30 km
Management	Community; Government; Collaboration (Community, Village-owned Enterprises, Youth Organization)
Ticket Price	0-30 (IDR 000); 30 -75 (IDR 000); ≥ 75(IDR 000)
ICT	Good-Limited
Target Variable	Tourism village development status: developed-progress

According to the machine learning approach, the data in this study consists of two categories, namely the training data and testing data. The training data represent features from 126 tourism villages and is used to build a prediction basis model. The training data were obtained from various news and reviews about tourist villages in Indonesia on Google Search. At the same time, data testing is test data that will be tested based on the predicted results of training data. Data testing is the profile of eight tourist villages in the Kedung Ombo area (Table 2). The data in the Kedung Ombo area was obtained using the observation method.

Table 2. Profile of Tourist Village in the Kedung Ombo Area

No	Tourist Village	Partici-pation	Attraction	Government Support	Part-nership	Infra-struc-ture	Dis-tance	Manage-ment	Ticket Price	ICT
1	Bo-yolayar	Active	Mono	Weak	Mono	Partial	0-10	Commu-nity	0-30	Limited
2	Wono-sari	Active	Mono	Strong	Multi	Com-plete	0-10	Govern-ment	0-30	Limited
3	Wono-harjo	Passive	Mono	Strong	Multi	Com-plete	0-10	Govern-ment	0-30	Good
4	Kedung Grujug	Passive	Mono	Strong	Mono	Partial	0-10	Commu-nity	0-30	Limited
5	Soko	Active	Mono	Strong	Mono	Com-plete	0-10	Collabo-ration	0-30	Limited
6	Pen-dem	Passive	Mono	Weak	Mono	Partial	0-10	Commu-nity	0-30	Limited
7	Bagor	Active	Mono	Weak	Mono	Partial	10-30	Commu-nity	0-30	Limited
8	Bulu Serang	Passive	Mono	Strong	Mono	Partial	0-10	Collabo-ration	0-30	Limited

The Naïve Bayes algorithm is a probabilistic classification algorithm regularly used to handle big data (Panawong et al., 2014). Naïve Bayes is a widely used method because it is simple, measurable, and efficient in classifying (Ramoni & Sebastiani, 2001; Naik & Kiran, 2018). Naïve Bayes uses probability and statistics and the basic Bayes theorem. Probability is the chance that an event will occur randomly. Bayes' theorem was discovered by Thomas Bayes (1701-1761), who introduced that the conditional probability of a non-single event, i.e., the probability that an event will occur, is affected by the previous event. In this research, the naïve Bayes method is used as the first approach to prediction. The equation of Bayes' theorem is as follows:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$ (1)

P(A|B) is the posterior probability which indicates how often event A occurs under condition B. P(B|A) is the prior probability which shows how often B occurs given A occurs. P(A) and P(B) are the probabilities of events A and B, respectively.

The decision tree is a classification method applying a tree structure or decision hierarchy. According to Anggarwal (2015), it is a classification method whose model uses a set of decisions in a ranking, forming a tree structure with feature variables. Decision trees are easy-to-understand and often accurate decision-making applications (Witten et al., 2017). The decision tree algorithm criteria commonly used are ID3, C4.5, and CART. Iterative dichotomized 3 (ID3) is an algorithm with a basic iterative structure, and its features are divided into two classes at each step. This method produces a classification in the form of a decision tree that starts from the root of the tree of possible decisions—following the explanation of Quinlan (1992)(Quinlan, 1992), who developed an improvement on the previous method and called it the C4.5 algorithm. Breiman (2001) also developed another decision tree algorithm called the classification and regression tree (CART). This classification divides a binary data set into two sets separately. The CART method calculation process has several stages (Anggarwal, 2015).

A collection of points on the data S. Suppose that p is included in the dominant class. The error rate is calculated as 1-p. For the Split r-way from the set S to the set S1….Sr, the error rate of the split can be qualified as a weighted average of the error rates of the individual sets of Si, where Si is |Si|. The separation with the lowest error rate is selected from the alternatives.

- The Gini index G(S) is the training data for S in class p1..pk distribution from the training data points in S.

$$G(S) = 1 - \sum_{j=i}^{k} p_j^2$$ (2)

The overall Gini Index for the r-way split from the set S to the set S1…Sr can be quantified as a weighted average of the Gini Index values G(Si) of each Si, where the weight of Si is |Si|.

$$Gini - Split\ (S \rightarrow S_1 \ldots S_r = \sum_{i=1}^{r} [s_i]\ GS_i$$ (3)

The split with the lowest Gini Index is selected from the alternatives. The CART algorithm uses the Gini Index as the split criterion.

The illustration of data mining using machine learning Orange in this research is shown in Figure 1.
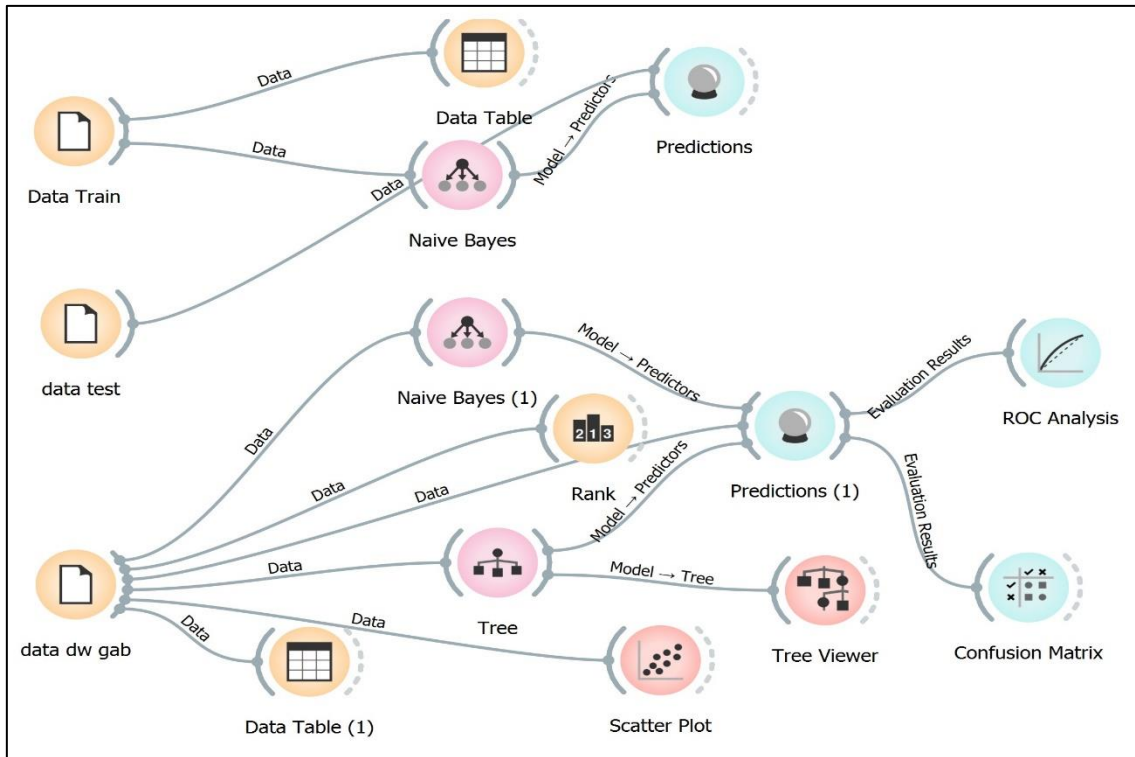
Figure 1. Image Data Mining to Determine the Status of Tourism Villages

Performance Evaluation

Performance Test

A performance test is used to get the best validation and learning model through cross-validation. This process tests which algorithm gives the best classification probability and is suitable for use as a prediction. These results can be seen in the prediction scores and test scores. The performance test was carried out using the cross-validation method as a sampling method because this method was effective in avoiding unintentional effects, primarily due to data limitations. This method was also suggested by Witten et al. (2017).

The learning technique separates the data into two categories: training data (training data) to form the model and test data (testing data) to test the model's performance. The classification results are likely to be accurate or rarely incorrect. The data will then be divided into several parts, symbolized by k in n data, known as k-fold cross-validation. Each iteration has a representation so that all data elements are met and data strata are used. The average result of each iteration obtained is used as the validation value.

The measurement of performance values is based on the confusion matrix value, which represents the prediction compared to the actual condition of the data generated by the machine learning algorithm (Table 3). True Positive (TP) is a positive and correct prediction; True Negative (TN) is a negative and true prediction; False Positive (FP) is a positive and false prediction result, while False Negative (FN) is a negative and false prediction result.

Table 2. Confusion Matrix

Actual		Predicted	
		-	+
	-	True Positive (TP) = A	False Positive (FP) = B
	+	False Negative (FN) = C	True Negative (TN) = D

Based on the Confusion Matrix in the naïve Bayes method, prediction performance is measured using the following values:

(1) The Area Under Curve (AUC) describes how accurately the model can classify correctly visually. The accuracy of the ROC classification is done by visually calculating the area under the Receiver Operating Characteristics Curve (ROC) curve. An excellent model has an AUC value close to 1. The following is the formula for the area Under the ROC Curve:

$$\text{AUC} = \frac{1}{2}\sum_{i=1}^{n}(x_{i+1} - x_i)(y_{i+1} - y_i) \tag{4}$$

The accuracy of the predicted values was confirmed using the criteria developed by Gorunescu (2011) in Table 3.

Table 3. Prediction Value Classification

AUC Score	Classification
$0.90 - 1.00$	Excellent
$0.80 - 0.90$	Good
$0.70 - 0.80$	Fair
$0.60 - 0.70$	Poor
$0.50 - 0.60$	Failure

(Source: Gorunescu, 2011)

(2) Classification Accuracy (CA) shows the accuracy of the predictions generated from the predicted and actual values divided by the total results. As with AUC, the higher the CA value, the closer to 1, the more accurate the model prediction.

$$CA = (TP + TN)/(Total\ A+B+C+D) \tag{5}$$

(3) Precision. Precision is the ratio of the true positive predicted value to the overall positive predicted result. The value of precision in validation is more practical and provides an accurate picture.

$$Precisison = TP/(TP + FP) \tag{6}$$

(4) Recall. The Recall is a positive true value comparison with all true values. The Recall compares the predicted positive true value with the overall positive true value. F1 is a weighted comparison of the average precision and Recall.

$$Recall = TP/(TP+FN) \tag{7}$$

(5) F1. the F1 score combines Recall and Precision (described below) into one performance metric. The F1 score is a weighted average of Precision and Recall. Therefore, this score takes into account false positives and false negatives.

$$F1 = 2FP / (FP + FN + 2TP) \tag{8}$$

Results and Discussion
Results

From the machine learning process on train data and test data using the Naïve Bayes and Tree methods, predictions regarding the status of villages in the Kedung Ombo area are obtained, as shown in Table 4. The prediction results show that the eight villages in the Kedung Ombo area, which are the basis of this study, are predicted to be developing tourism villages.

Table 4. Prediction Results of Tourism Village Status Based on Naïve Bayes and Tree Methods

No.	Naïve Bayes	Tree	Status	Desa
1	1.00: 0.00 developing	1.00: 0.00 developing	Developing	Boyolayar
2	0.37:0.63 developing	1.00;0.00 developing	Developing	Wonosari
3	0.91:0.09 developing	0.92: 0.08 developing	Developing	Wonoharjo
4	1.00:0.00 developing	1.00:0.00 developing	Developing	Kedung Grujug
5	0.92:008 developing	1.00:0.00 developing	Developing	Soko
6	1.00: 0.00 developing	1.00: 0.00 developing	Developing	Pendem
7	1.00:0.00 developing	1.00:0.00 developing	Developing	Bagor
8	1.00:0.00 developing	1.00:0.00 developing	Developing	Bulu Serang

(Source: Data processed with Orange Software 3.3.0)

The results of performance predictions were tested using AUC, CA, F1, Precision, and Recall values, as shown in Table 5.

Table 5. Performance Test Results Through Cross-Validation of Naïve Bayes and Tree Method

Model	AUC	CA	F1	Precision	Recall
Naïve Bayes	0.925	0.858	0.822	0.721	0.957
Tree	0.966	0.888	0.828	0.878	0.783

(Source: Data processed with Orange Software 3.3.0)

Table 5 shows that the performance results of the AUC validation value of the naïve Bayes model are 0.925. Following the classification of prediction accuracy values developed by Gorunescu (2011), this prediction falls into the excellent category. While the AUC value of the decision tree is 0.966, so it is also included in the excellent classification. The AUC value shows that the neural network is the best method for predicting the status of tourist villages in the Kedung Ombo area.

Moreover, according to the results of the prediction classification test on the accuracy value (CA), the highest value is obtained from the Naïve Bayes gets a value of 0.858. and the decision tree has 0.888, which is in the good category.

Furthermore, from the prediction test results on the accuracy value (CA), it is known that the Naïve Bayes method gets a value of 0.858, while the Tree method gets a value of 0.888. These results include the good category. Predictive performance assessment uses the F1 value. The Naïve Bayes method gets a value of 0.822, while the Tree method gets a value of 0.828. Thus, the predictive performance of both falls into the fair category.

Performance assessment using the precision value is known to be the best found in the Naïve Bayes model, having a value of 0.933, and the decision tree has a value of 0.878. The latter methods are good classification. The opposite result is obtained from the Recall test, where the Naïve Bayes method has an enormous value of 0.957, so it is classified as an excellent category. While the value of the Tree method of 0.783 falls into the good classification.

The prediction accuracy value can also be known through the ROC graph in Figure 3. Testing the model by looking at the ROC curve is the easiest way to graphically compare the accuracy values of each classification model. Figure 2 shows that all models are good at predicting because they are close to point 1.0, and the neural network model has the best accuracy value because the ROC curve is most intimate to point 1.0. The results of this analysis confirm the previous test.
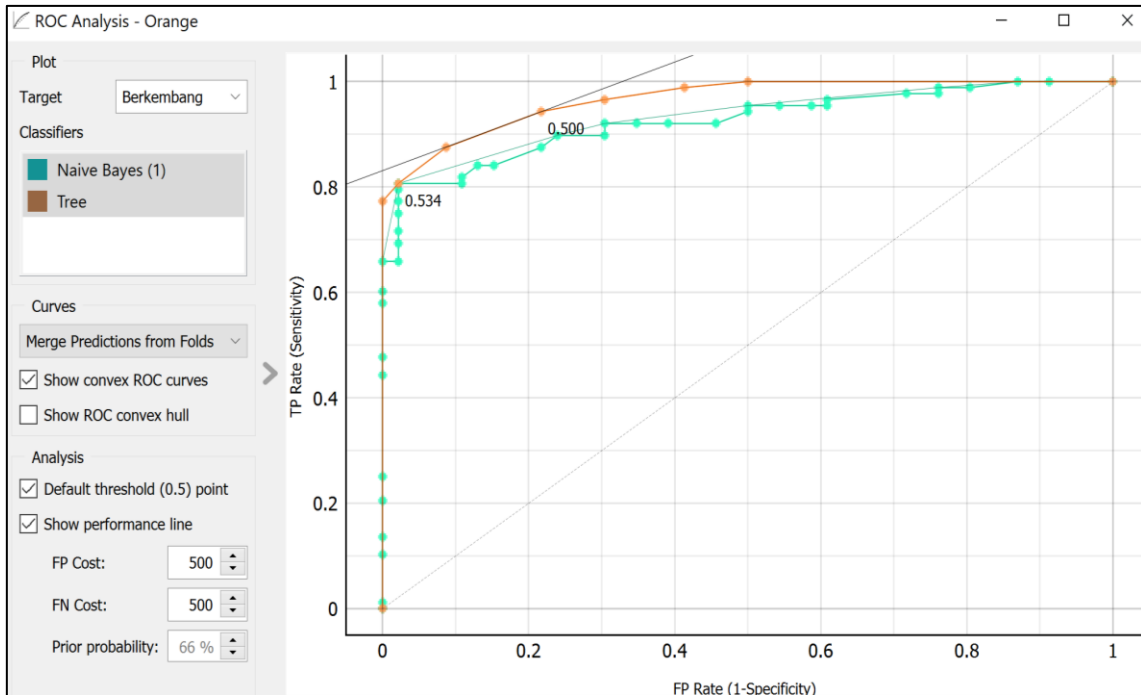
Figure 2. ROC Analysis

The results of this prediction were then tested for accuracy using the Confusion Matrix (Figure 3).
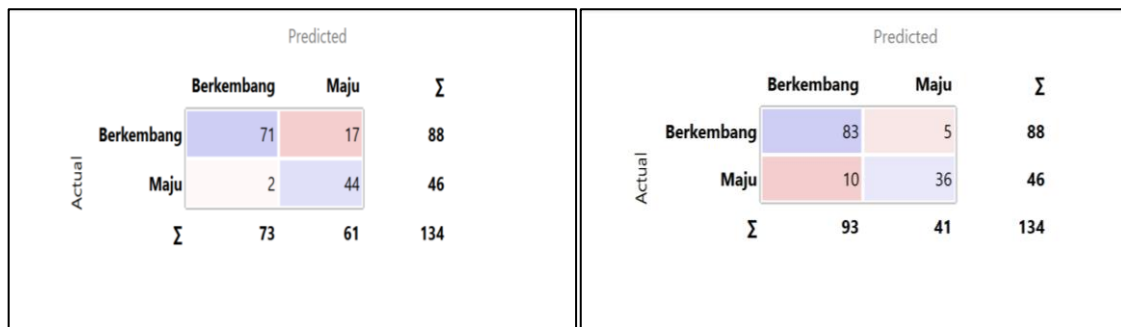

Figure 3. Confusion Matrix on Naïve Bayes and Tree Method

Based on the formula given previously, it is known that the results of the model performance test are based on the Confusion Matrix (Table 6).

Table 6. Prediction Test Results through Confusion Matrix from Naïve Bayes and Tree Method

Model	CA	F1	Precision	Recall
Naïve Bayes	86%	88%	82%	97%
Tree	88%	91%	94%	89%

(Source: Data processed with Orange Software 3.3.0)

The Confusion Matrix test results show that the predictive values for all the methods used are classified as excellent and good. The Naïve Bayes algorithm provides more consistent valida-

tion than other methods in the three parameters of the five methods. Thus, this method is considered the most accurate in providing views on development policies in tourist villages in the Kedung Ombo area.

The overall distribution of tourist villages based on cross-correlation between several tested attributes is shown in Figure 4.
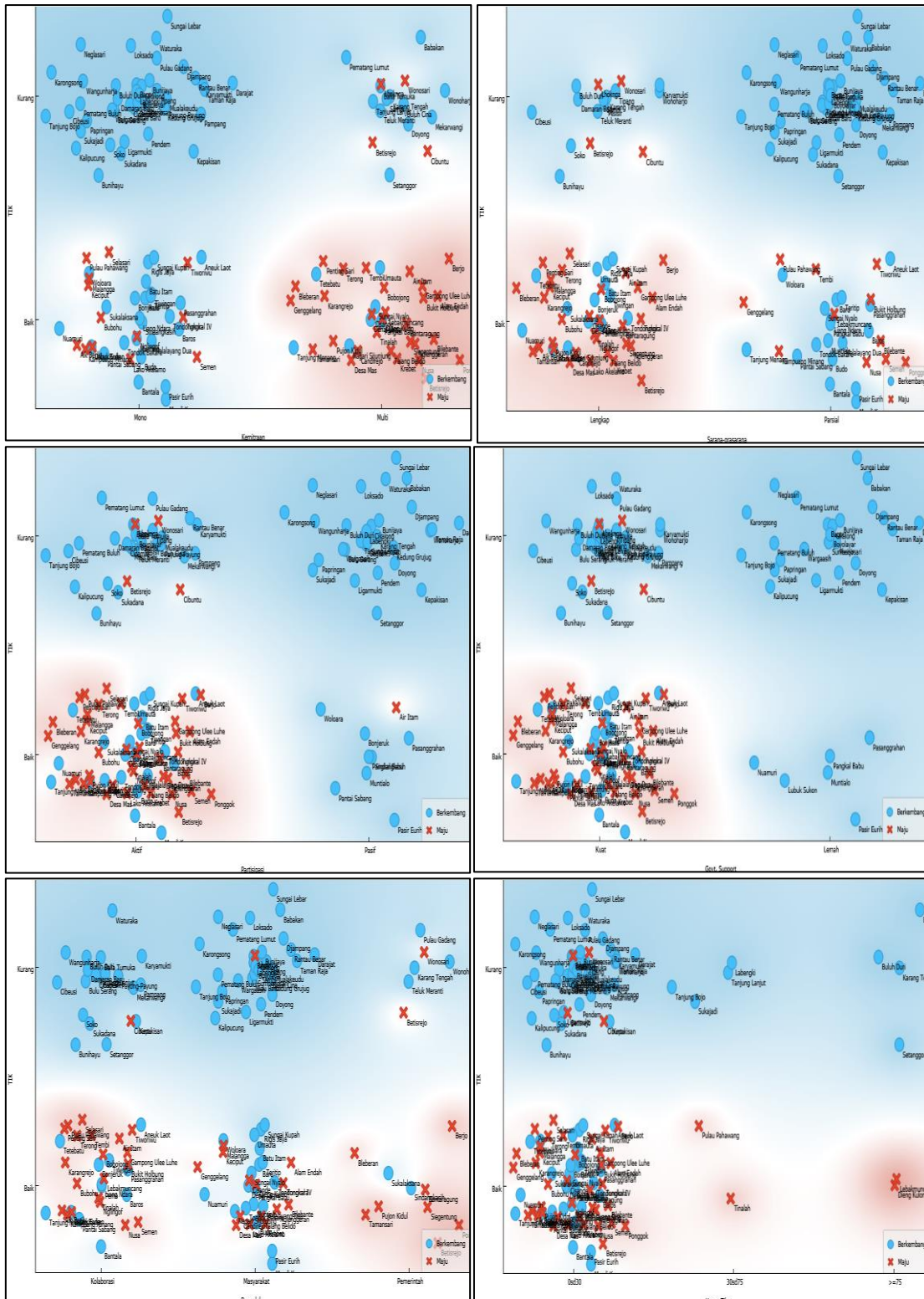

Figure 4. Scatter Plot between Target and Feature Variables

Discussions

From this point, to find out what attributes most determine the progress of a tourist village, it can be seen from the decision tree model. The results of the decision tree classification provide a good analysis of the predictions generated. The gain ratio value from the decision tree determines which variable is the split classification. The decision tree begins with the formation of roots (located at the top). Then the existing data is divided based on attributes suitable to be made into leaves connected and developed through branches.

Decision rules are formed from the Tee method that has been created and then derived by tracing from roots to leaves. Based on the dataset processing using the Tree method, the accuracy of the classification process obtained an AUC value of 96%, which means that predictions through the decision tree method are excellent. The decision tree image in Figure 5 provides information about the attributes/conditions that determine the progress of a tourist village.

Figure 5. Decision Tree Prediction Results

Base on Decision Tree (Figure 5) shows the factors determining the progress of a tourist village are the type of collaborative management, the location distance from 0 km to 30 km or 10 km to 30 km, various attractions, and a maximum ticket price of 30,000 rupiahs, ongoing partnerships with multiple partners (government, academia, and the private sector), as well as good ICT facilities and applications.

Furthermore, from the results of the decision tree prediction, the order of attributes that most influence the progress of the tourist village can be seen. The Orange operation shows the ranking results for the antecedent factors of the progress of tourism villages (Table 7).

Table 7. Attribute Rank Determining Tourism Village Progress

			#	Info. gain	Gain ratio	Gini	χ²
1	C	TIK	2	0.217	0.219	0.119	19.862
2	C	Kemitraan	2	0.166	0.171	0.102	18.347
3	C	Govt. Support	2	0.166	0.212	0.071	16.205
4	C	Partisipasi	2	0.164	0.190	0.080	16.935
5	C	Sarana-prasarana	2	0.139	0.140	0.084	11.591
6	C	Atraksi	2	0.101	0.102	0.060	9.518
7	C	Pengelola	3	0.053	0.038	0.033	0.132
8	C	Jarak	3	0.010	0.007	0.006	0.060
9	C	Harga Tiket	3	0.000	0.000	0.000	0.008

(Source: Data processed with Orange Software 3.3.0)

At Table 7 ICT is the attribute that has the most significant influence on the progress of tourist villages from various measurement bases (Ingo. gain, Gain Ratio, Gini, and X^2). At the next level is the attribute of partnership, followed by government support, community participation, and infrastructure. Next are the tourist attractions. The management and distance attributes have little influence, and the ticket price attribute has almost no effect.

These findings can be explained by the presence of ICT (especially the internet) that has changed tourists' behavior, especially in seeking information. Therefore, ICT should be utilized in creating ideas and information about rural tourism in the minds of tourists and enhancing seek information, such as tourism services, locality accessibility, ticket booking, and payment methods, as well as other tourist facilities. In the case of rural tourism, ICTs can facilitate the establishment of relationships between tourists and managers, even after the visit is over. Managers can build a visitor database to develop these relationships to contribute to building visitor loyalty. Friendly service during and after travel activities which ICT facilities support, can determine this loyalty.

The next determinant factor most significantly influencing tourist village progress is multi-stakeholder partnerships. Furthermore, the following factors are strong government support, community involvement, completeness of infrastructure facilities, and various attractions. However, the characteristic of each village are factors that must be developed in an integrated manner to accelerate the progress of tourist villages.

The results of this study support research conducted by Hidayatullah et al. (2018) and Waghmode & Jamsandekar (2013) that ICT is a factor that plays a role in supporting the development of rural tourism and tourist villages as a promotional medium and booking and payment facilities transaction. Regarding partnerships, the results of this study are in accordance with research conducted by Purbasari & Manaf (2018) and Mtapuri & Giampiccoli (2013), which state that rural tourism development requires public partnerships in both local and global contexts. While related to community participation, the results of this study follow the results of research by Rodrigues et al. (2021); Bajrami et al. (2020); Basile et al. (2021); Yu et al. (2018); (Powell et al., 2018); and Amin & Ibrahim (2015) who in their research found that community involvement is a critical factor in the development of tourist villages. While the findings regarding the completeness of tourism infrastructure, the results of this study are in accordance with research conducted by Bhalla et al. (2016).

Conclusions

Tourism management, especially rural-based tourism, faces complex and challenging issues. The dynamic interaction between various components affects the tourism sector's performance. Hence, a paradigm shift is needed in determining the right rural tourism policy. The use of machine learning will be critical to help develop science-based and data-driven policies in the future. Besides being used to predict how a tourist village will grow or if it will not develop, machine learning will also reduce bias in determining tourism development policies, which tend to be more subjective.

In this study, the Tree Decision algorithm is the best approach to predict the progress status and determine antecedent factors of the progress of tourist villages. Likewise, the variables analyzed, namely community participation, diversity of attractions, government support, partnerships, infrastructure, ticket prices, distance to locations, management, and ICT, are the right factors to predict village status.

Among the antecedents of these variables, this study finds ICT is the main factor predicted to determine the progress of tourist villages. However, government support, completeness of infrastructure, community participation, multi-stakeholder partnership, and various tourist attractions are other essential factors that must be developed in integrated ways for tourist village development.

References

Adekitan, A. I., Abolade, J., & Shobayo, O. (2019). Data mining approach for predicting the daily Internet data traffic of a smart university. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0176-5

Ahmed, N., Atiya, A., El Gayar, N., & El-Shishiny, H. (2007). Tourism Demand Forecasting Using Machine Learning Methods. *AIML*, *January*. http://infos2007.fci.cu.edu.eg/tourism/07184.pdf

Amin, A., & Ibrahim, Y. (2015). Model of Sustainable Community Participation in Homestay Program. *Mediterranean Journal of Social Sciences*, *6*(3), 539–545. https://doi.org/10.5901/mjss.2015.v6n3s2p539

Andariesta, D. T., & Wasesa, M. (2022). Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach. *Journal of Tourism Futures*, 1–17. https://doi.org/10.1108/JTF-10-2021-0239

Anggarwal, C. . (2015). *Data Mining: The Textbook*. Springern _c Springer International Publishing Switzerland. https://doi.org/10.1016/0304-3835(81)90152-X

Ariyani, N., & Fauzi, A. (2023). Pathways toward the Transformation of Sustainable Rural Tourism Management in Central Java, Indonesia. *Sustainability*, *15*(3), 2592. https://www.mdpi.com/2071-1050/15/3/2592

Ariyani, N., Fauzi, A., & Umar, F. (2022). Predicting determinant factors and development strategy for tourist villages. *Decision Science Letters*, *12*, 137–148. https://doi.org/10.5267/dsl.2022.9.003

Ayazlar, G., & Ayazlar, R. (2015). Rural Tourism: A Conceptual Approach. In A. Chevdet, M. Dinu, N. Hacioglu, R. Efe, & A. Spykan (Eds.), *Tourism, Environment and Sustainability* (Issue 14, pp. 167–184). St. Kliment Ohridski University Press.

Bajrami, D. D., Radosavac, A., Cimbaljević, M., Tretiakova, T. N., & Syromiatnikova, Y. A. (2020). Determinants of residents' support for sustainable tourism development: Implications for rural communities. *Sustainability (Switzerland)*, *12*(22), 1–16. https://doi.org/10.3390/su12229438

Basile, G., Tani, M., Sciarelli, M., & Ferri, M. A. (2021). Community participation as a driver of sustainable tourism. The case of an Italian village: Marettimo Island. *Sinergie*, *39*(1), 81–102. https://doi.org/10.7433/s114.2021.06

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

Firdaus, Hardjosoekarto, S., & Lawang, R. M. Z. (2021). The Role of Local Government on Rural Tourism Development: Case Study of Desa Wisata Pujonkidul, Indonesia. *International Journal of Sustainable Development and Planning*, *16*(7), 1299–1307. https://doi.org/10.18280/ijsdp.160710

Gohori, O., & van der Merwe, P. (2020). Towards a tourism and community-development framework: An African perspective. *Sustainability (Switzerland)*, *12*(13). https://doi.org/10.3390/su12135305

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques. In Intelligent Systems Reference Library (Vol. 12)*. Springer Link. https://doi.org/https://doi.org/10.1007/978-3-642-19721-5

Hidayatullah, S., Malang, U. M., & Rachmawati, I. K. (2018). Development of Tourist Village Model Through "Pokdarwis" Empowerment and Information Technology Utilization. *European Journal of Business and Management*, *January*. https://doi.org/10.7176/ejbm-10-23-12

Ho, P. T. (2022). Smart Tourism Recommendation Method in Southeast Asia under Big Data and Artificial Intelligence Algorithms. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/4047501

Kamarudin, K. H., Wahid, S. N. A., & Chong, N. O. (2020). Challenges for Community Based Rural Tourism Continuity and Resilience in Disaster Prone Area: The Case of Mesilou, Sabah. *IOP Conference Series: Earth and Environmental Science*, *409*(1). https://doi.org/10.1088/1755-1315/409/1/012003

Kantsperger, M., Thees, H., & Eckert, C. (2019). Local participation in tourism development-roles of non-tourism related residents of the Alpine Destination Bad Reichenhall. *Sustainability (Switzerland)*, *11*(24). https://doi.org/10.3390/SU11246947.

Khalid, S., Ahmad, M. S., Ramayah, T., Hwang, J., & Kim, I. (2019). Community empowerment and sustainable tourism development: The mediating role of community support for tourism. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226248.

Koopmans, M. E., Rogge, E., Mettepenningen, E., Knickel, K., & Šūmane, S. (2018). The role of multi-actor governance in aligning farm modernization and sustainable rural development. *Journal of Rural Studies*, *59*, 252–262. https://doi.org/10.1016/j.jrurstud.2017.03.012.

Kristianto, B., Fahrudin, A., & Salim, Z. (2019). Determinants of sustainable small islands tourism management: Case study of Pahawang Island, Indonesia. *IOP Conference Series: Earth and Environmental Science*, *241*(1). https://doi.org/10.1088/1755-1315/241/1/012015.

León-Gómez, A., Ruiz-Palomo, D., Fernández-Gámez, M. A., & García-Revilla, M. R. (2021). Sustainable tourism development and economic growth: Bibliometric review and analysis. *Sustainability (Switzerland)*, *13*(4), 1–20. https://doi.org/10.3390/su13042270.

Li, W. (2022). Prediction of Tourism Demand in Liuzhou Region Based on Machine Learning. *Mobile Information Systems*, 1–9. https://doi.org/10.1155/2022/9362562.

Liu, C., Dou, X., Li, J., & Cai, L. A. (2020). Analyzing government role in rural tourism development: An empirical investigation from China. *Journal of Rural Studies*, *79*(October 2019), 177–188. https://doi.org/10.1016/j.jrurstud.2020.08.046.

McLennan, C. lee J., Ritchie, B. W., Ruhanen, L. M., & Moyle, B. D. (2014). An institutional assessment of three local government-level tourism destinations at different stages of the transformation process. *Tourism Management*, *41*, 107–118. https://doi.org/10.1016/j.tourman.2013.09.007.

Mtapuri, O., & Giampiccoli, A. (2013). Interrogating the role of the state and nonstate actors in community-based tourism ventures: Toward a model for spreading the benefits to the wider community. *South African Geographical Journal*, *95*(1), 1–15. https://doi.org/10.1080/03736245.2013.805078.

Muganda, M., Sirima, A., & Ezra, P. M. (2013). The Role of Local Communities in Tourism Development: Grassroots Perspectives from Tanzania. *Journal of Human Ecology*, *41*(1), 53–66. https://doi.org/10.1080/09709274.2013.11906553.

Naik, D. L., & Kiran, R. (2018). Naïve Bayes classifier, multivariate linear regression and experimental testing for classification and characterization of wheat straw based on mechanical properties. *Industrial Crops and Products*, *112*(December 2017), 434–448. https://doi.org/10.1016/j.indcrop.2017.12.034.

Nair, V., Munikrishnan, U. T., Rajaratnam, S. D., & King, N. (2015). Redefining Rural Tourism in Malaysia: A Conceptual Perspective. *Asia Pacific Journal of Tourism Research*, *20*(3), 314–337. https://doi.org/10.1080/10941665.2014.889026.

Neumeier, S., & Pollermann, K. (2014). Rural Tourism as Promoter of Rural Development - Prospects and Limitations: Case Study Findings from a Pilot Project Promoting Village Tourism. *European Countryside*, *6*(4), 270–296. https://doi.org/10.2478/euco-2014-0015.

Panawong, N., Snae Namahoot, C., & Brückner, M. (2014). Classification of tourism web with modified Naïve Bayes algorithm. *Advanced Materials Research*, *931–932*(March 2015), 1360–1364. https://doi.org/10.4028/www.scientific.net/AMR.931-932.1360.

Pantiyasa, I. W., Sulasmini, N. M. A., & Rosalina, P. D. (2019). Construction Model of Tourist Village Development Towards Smart Eco-Tourism Village Destination in Paksebali To Become Tourism Icon in Klungkung Regency, Bali. *Journal of Business on Hospitality and Tourism*, *5*(1), 44. https://doi.org/10.22334/jbhost.v5i1.149

Powell, R. B., Green, T. F., Holladay, P. J., Krafte, K. E., Duda, M., Nguyen, M. T., Spencer, J. H., & Das, P. (2018). Examining Community Resilience to Assist in Sustainable Tourism Development Planning in Dong Van Karst Plateau Geopark, Vietnam. *Tourism Planning and Development*, *15*(4), 436–457. https://doi.org/10.1080/21568316.2017.1338202

Purbasari, N., & Manaf, A. (2018). Comparative Study on the Characteristics of Community-Based Tourism between Pentingsari and Nglanggeran Tourism Village, Special Region Yogyakarta. *E3S Web of Conferences*, *31*. https://doi.org/10.1051/e3sconf/20183109007

Purnaningrum, E., & Athoillah, M. (2021). SVM Approach for Forecasting International Tourism Arrival in East Java. *Journal of Physics: Conference Series*, *1863*(1). https://doi.org/10.1088/1742-6596/1863/1/012060

Quinlan, J. R. (1992). Learning with continuous classes. *Australian Joint Conference on Artificial Intelligence*, *92*, 343–348.

Ramoni, M., & Sebastiani, P. (2001). Robust Bayes classifiers. *Artificial Intelligence*, *125*(1–2), 209–226. https://doi.org/10.1016/S0004-3702(00)00085-0

Rodrigues, C., Liberato, D., & Melo, C. (2021). Tourism sustainable practices in rural territories: The case of Caretos de Podence. *Journal of Tourism and Development*, *36*, 205–220. https://doi.org/10.34624/rtd.v1i36.23736

Sharpley, R., & Roberts, L. (2004). Rural tourism — 10 years on. *International Journal of Tourism Research*, *6*(3), 119–124. https://doi.org/10.1002/jtr.478

Tang, Y. (2022). Discrete Dynamic Modeling Analysis of Rural Revitalization and Ecotourism Sustainable Prediction Based on Big Data. *Discrete Dynamics in Nature and Society*, *2022*. https://doi.org/10.1155/2022/9158905

Waghmode, M. L., & Jamsandekar, P. P. (2013). Role of ict in tourism. *ASM's International E-Journal of Ongoing Research in Management and IT*, *5*(8), 1–10.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques.* (4 th). Burlington, MA: In Morgan Kaufmann.

Wu, W., Zhang, L., & Qiu, F. (2017). Determinants of tourism ticket pricing for ancient villages and towns: Case studies from Jiangsu, Zhejiang, Shanghai and Anhui provinces. *Tourism Management*, *58*, 270–275. https://doi.org/10.1016/j.tourman.2016.03.011

Xie, D., & He, Y. (2022). Marketing Strategy of Rural Tourism Based on Big Data and Artificial Intelligence. *Hindawi, Mobile Information System*, *2022*, 7. https://doi.org/https://doi.org/10.1155/2022/9154351

Yang, J., Ge, Y., Ge, Q., Xi, J., & Li, X. (2016). Determinants of island tourism development: The example of Dachangshan Island. *Tourism Management*, *55*, 261–271. https://doi.org/10.1016/j.tourman.2016.03.001

Yu, C. P., Cole, S. T., & Chancellor, C. (2018). Resident support for tourism development in rural midwestern (USA) communities: Perceived tourism impacts and community quality of life perspective. *Sustainability (Switzerland)*, *10*(3). https://doi.org/10.3390/su10030802

Yu, N., & Chen, J. (2022). Design of Machine Learning Algorithm for Tourism Demand Prediction. *Computational and Mathematical Methods in Medicine*, *2022*. https://doi.org/10.1155/2022/6352381