sustainability

MDPI

Article

Application of Machine Learning Techniques to Predict Visitors to the Tourist Attractions of the Moche Route in Peru

Jessie Bravo *, Roger Alarcón, Carlos Valdivia and Oscar Serquén

Professional School of Computer Engineering and Informatics, Pedro Ruiz Gallo National University, Lambayeque 14013, Peru; ralarcong@unprg.edu.pe (R.A.); cvaldivias@unprg.edu.pe (C.V.); aserquen@unprg.edu.pe (O.S.)

* Correspondence: jbravo@unprg.edu.pe

Abstract: Due to the COVID-19 pandemic, the tourism sector has been one of the most affected sectors and requires management entities to develop urgent measures to reactivate and achieve digital transformation using emerging disruptive technologies. The objective of this research is to apply machine learning techniques to predict visitors to tourist attractions on the Moche Route in northern Peru, for which a methodology based on four main stages was applied: (1) data collection, (2) model analysis, (3) model development, and (4) model evaluation. Public data from official sources and internet data (TripAdvisor and Google Trends) during the period from January 2011 to May 2022 are used. Four algorithms are evaluated: linear regression, KNN regression, decision tree, and random forest. In conclusion, for both the prediction of national and foreign tourists, the best algorithm is linear regression, and the results allow for taking the necessary actions to achieve the digital transformation to promote the Moche Route and, thus, reactivate tourism and the economy in the north of Peru.

Keywords: predictive model; tourist arrival; Moche Route; machine learning; online forums; open data; digital transformation; tourism management

Citation: Bravo, J.; Alarcón, R.; Valdivia, C.; Serquén, O. Application of Machine Learning Techniques to Predict Visitors to the Tourist Attractions of the Moche Route in Peru. *Sustainability* **2023**, *15*, 8967. https://doi.org/10.3390/su15118967

Academic Editors: Hector Cardona-Reyes, Albert Barreda, Rosse Marie Esparza Huamanchumo, Sandra Zubieta Zamudio and Jun (Justin) Li

Received: 31 March 2023
Revised: 26 May 2023
Accepted: 29 May 2023
Published: 1 June 2023

1. Introduction

The tourism sector after the COVID-19 pandemic has been one of the most affected, but since 2022 its reactivation has been in effect with more force and the application of information technologies has been considered to generate better results and achieve the digital transformation that the sector needs.

There is a type of tourism called heritage tourism, which according to [1] is usually based on living and built elements of culture and refers to the use of the tangible and intangible past as a tourism resource. It encompasses current cultures and customs, since they are also legacies of the past. There is a socioeconomic perspective, divided into developed and developing countries, sometimes referred to as "rich" and "poor", "North" and "South" (due to the high concentration of poor countries in the southern hemisphere).

From a tourism point of view, less developed countries are extremely important as destinations and players in the global industry. Travel to and within developing countries is growing at a faster rate than in more developed regions.

In this sense, Peru is an ideal destination for this type of tourism because it has many places that meet the aforementioned characteristics and provides sustainability for the actors involved. In this regard, in the analysis of tourism in [2], it states that sustainability, in its most basic form, summarizes the growing concern for the environment and natural resources, although sustainability has also had a growing resonance in social and economic issues.

Socioeconomic sustainability is a factor related to this research, which is understood as an improvement in the quality of life of the local population, of the people who live and

work in tourist destinations through the generation of multiple sources of employment, as well as an improvement in the infrastructure, communications, and local transport.

The result of this research will allow for the planning and generation of strategies developed by the managing entities of the tourism sector based on information and the application of disruptive technologies.

Therefore, the digital transformation of the tourism sector is essential for the development of sustainable tourism, this being a comprehensive plan led by managing entities but involving people and organizations who are from the local community and make their livelihoods from tourism, with the aim of maintaining the operation of the industry, guaranteeing social, economic, and environmental sustainability, also evidenced in [3–5], which indicate that digitalization, data analytics, and artificial intelligence are great facilitators of sustainable solutions of tourist destinations and that managing entities must define policies that motivate their application and achieve the greater satisfaction of tourists.

According to [6], in recent years there has been a revolution in both the content and methodology of work in the field of artificial intelligence. It is now more usual to build upon existing theories than to propose totally novel theories, to take as basis rigorous theorems or solid experimental evidence rather than intuition, and to demonstrate the usefulness of applications in the real world rather than to create prototypes.

1.1. The Moche Route in Peru

One of the archaeological destinations visited by tourists to Peru is the Moche Route, which has great cultural value due to the archaeological, natural, cultural, and landscape attractions that define it; this includes the regions of La Libertad and Lambayeque located on the north coast of Peru [7,8].

The authorities of the sector have the objective of working together so that both regions can improve the flow of national and foreign visitors who visit, thus making the Moche Route the main tourist destination in northern Peru, allowing tourism to be reactivated and improving the local and national economies.

According to [9], the Moche Route is a destination where Moche's old traditions still continue to this day, which is evident in its gastronomy, the work of its people, and the beauty of its beaches, hence the phrase, "Moche Route, live the experience", as part of its positioning strategy, which also focuses on the values of friendship, kindness, and the unique experiences that both regions, La Libertad and Lambayeque, can offer.

The Moche Route has two of the most visited destinations by tourists in Peru. One is the Lord of Sipan, located at the archaeological site of Huaca Rajada de Sipan in the Lambayeque region, and the other is the Lady of Cao, located in the El Brujo archaeological complex in La Libertad. The creation of a branding strategy that allowed for the promotion and disseminating of tourism in these destinations was therefore necessary [10].

While it is true that the southern part of Peru is very attractive for national and international tourists because of the Inca culture (i.e., Cuzco and Machu Picchu); however, the northern part of Peru allows for learning of Moche culture, one that predates the Inca and one of the most important in Peru, through archaeological sites (huacas, pyramids, and a mud city), museums (collections of gold, ceramics, and textiles are exhibited), and cultural manifestations (dances and gastronomy) existing throughout more than 200 km of the northern region of Peruvian territory.

In addition to the above, its importance lies in two important discoveries: The Lord of Sipan, an ancient priest ruler in the third century, whose discovery in 1987 was a major milestone in the world of archeology, as it was the first royal burial found intact in South America and one of the most lavish burials ever discovered in Peru, as the priest was found covered with gold, silver, copper, and semiprecious stones. The second great discovery was the Lady of Cao, a young ruler, who revealed the role of women in Moche culture, with magical–religious tattoos of snakes and spiders, as well as necklaces and earrings on her hands, legs, and face.

The Moche Route, as an official tourist destination, dates back to 2009, as a result of the Action Plan of the Moche Route management entity [11], developed by the Ministry of Foreign Trade and Tourism (MINCETUR), as a tourism alternative to the southern area, currently joined in its promotion institutions such as PROMPERU, who this year, 2023, aim to promote this tourist destination with the support of the World Bank, being the first project of the tourism sector scheduled for this year.

In 2011, the World Tourism Organization (UNWTO) awarded the Ulysses Award for Innovation in Tourism to MINCETUR with the purpose of improving the infrastructure of this tourist destination, with the goal that by 2016 the Moche Route will receive 650,000 visitors per year, which could generate revenues in excess of USD 140 million, but according to [12], it is still far from achieving this goal.

In addition, according to the National Strategic Tourism Plan 2017–2025, Peru's goal is to consolidate itself as a competitive, sustainable, quality, and safe tourist destination, as is also evidenced in the National Strategy for the Reactivation of the Tourism Sector 2021–2023 [13,14], where they indicate that the monthly arrival of international tourists was reduced by the pandemic, as it was 364,000 in 2019, dropping to 75,000 in 2020, and reaching 19,000 in 2021.

Therefore, the reactivation of tourism to Moche Route is one of the objectives programmed in the plans of the Peruvian government, where the efforts to promote tourism should focus on the revitalization of internal tourism, maintenance of receptive tourism, and articulation and advice of competencies in the quality of service and technological innovation to service providers, as well as investment in the improvement of infrastructure of tourist sites.

1.2. Digital Transformation Applied to Tourism

Technological innovation that allows for the digital transformation of this sector focuses on artificial intelligence, metaverse applications, big data, and IoT, among others. Artificial intelligence, specifically machine learning, has contributed to the management of efficient tourism processes based on data, the reactivation of tourism, and the improvement of the tourist experience.

Studies have contributed through the use of regression algorithms, such as observed in [15–19], with the main algorithms being used for both prediction and classification, including linear regression, k-nearest neighbors, support vector regression (SVR), random forest regression (RFR), decision tree, and multiple linear regression (MLR), among others, applied both in the tourism and other sectors.

In other works, deep learning algorithms and neural networks have been applied, such as in [20], where a machine learning model was developed to estimate the number of foreign visitors leaving Turkey for certain reasons. In addition, algorithms have been applied to batch type genetics to learn unknown model parameters when considering disruptions, for example, because of COVID-19.

Additionally, in [21] machine learning techniques were applied that integrated different hybrid models based on principal component analysis (PCA) and autoencoders for the prediction of tourism demand based on data from Google Trends in Morocco. Added to this is the work in [22], which applied an interlinked neural network model to predict short- and long-term tourist arrivals in the United States.

1.3. Data Sources for Machine Learning

Regarding the data sources used in this investigation, they were enriched thanks to information that exists in open databases and from different sources on the Internet, as demonstrated in [23], in which it was concluded that models using a combination of predictors from an online travel forum, such as TripAdvisor and Google Trends, plus a search engine, have better accuracy than those using predictors from a single internet data source.

In [24], the study used data from social networking services on hotels in Saudi Arabia obtained from TripAdvisor.

Additionally, ref. [25] used data from two search engines, Baidu and Google, to predict the arrival of tourists to a city based on seven influence factors. In [26], it was demonstrated that online platform data can be used as profitable and precise alternative data sources for statistics on sustainable tourism.

Finally, it is worth mentioning a case study applied in Peru, as developed in [27], for sun and beach destinations in the District of Ancon in which a qualitative study presented a diagnostic model of the tourism system. Peru is an important tourism destination worldwide, but the impact of the pandemic has been high, requiring an improvement in the tourist experience.

In this research, based on a review of the literature, it is possible to observe that there are very few studies that have applied machine learning to the tourism sector in Peru. Using various data sources and machine learning allows for digital transformation, and above all, it contributes to the better management of regulatory bodies based on data and the sustainability of companies and communities that make a living from tourism.

In addition, this document contributes, with a first approximation, using regression algorithms, to the understanding of the data and sources used in prediction analysis in the tourism sector in Peru, identifying the models that have the best performance according to the metrics, facilitating information-based decision making.

Therefore, the main objective of this research was to apply machine learning techniques to predict visitors to tourist attractions to Moche Route in northern Peru, thus contributing to the digital transformation of tourism.

2. Literature Review

From a literature review on the subject under study, the different prediction models that have been applied can be observed in addition to the various data sources used for analysis, requiring greater in-depth research into this topic to improve the accuracy of the results obtained. Table 1 summarizes the most important research articles found.

Regarding linear regression algorithms, ref. [28] states that linear predictors are intuitive, easy to interpret, and fit the data reasonably well for many natural learning problems.

In terms of nearest neighbors, ref. [29] defines it as one of the simplest prediction models available. It makes no assumptions and does not need robust computational equipment. It requires some concept of distance and assumes that points close to each other are similar. In addition, it indicates that a decision tree uses a tree structure to represent a series of possible decision paths and an outcome for each path.

Regarding random forest, ref. [28] states that it is a classifier formed by a collection of decision trees, and a prediction is obtained by a majority vote on the predictions of the individual trees.

2.1. Machine Learning Model and Deep Learning

Learning models are widely used tools in forecasting research. As can be seen in [20], a machine learning model was developed to estimate the number of foreign visitors leaving Turkey, and it also identified 10 reasons for foreign visitor departures over the next 10 years in order to gain a deeper understanding of their future behaviors.

In [21], a new hybrid deep learning framework was proposed combining data from search queries, autoencoders, and stacked long short-term memory (LSTM), which improved the accuracy of tourism demand forecasting. Similarly, in [25] a tourism demand forecasting model based on existing machine learning forecasting model research was built. This article used search engine data on the monthly volume of tourists from city A and its related influence factors as the dataset. The dataset was processed to make the model fit the input data. The mean absolute error (MAE), root mean square error (RMSE), MAPE, and other model evaluation indicators were applied to it. Then, the LSTM was used, and

the SAE-LSTM model was built to perform comparative experiments to predict the number of tourist arrivals over four years.

Regarding the data used for learning models, in [30] they used various data sources, such as the Baidu search engine and online review platforms, including Ctrip and Qunar, to carry out their forecast study, finding that the integration of several platforms is significant for this type of model.

2.2. Artificial Neural Networks

In [22], a model of interlinked neural networks was proposed, using data from tourist arrivals, which were broken down by two low-pass filters into long-term trend components and short-term seasonal components and then modeled by a pair of autoregressive neural network models as a parallel structure.

Similarly, ref. [31] proposed a new tourist arrival forecasting model based on multiscale learning to explore different data characteristics. Two popular models were introduced: modal decomposition (MD) and convolutional neural network (CNN).

It should be noted that, in [32], interval prediction models addressing two significant issues were developed. A simple mean with an additive property to derive pooled forecasts and time series often does not conform to any statistical assumptions. The genetic algorithm optimally determines all parameters needed to build an interval model. The empirical results for tourism demand showed that the proposed nonadditive interval model outperformed the other interval prediction models considered.

2.3. Prediction Models

The literature shows the utility of predictive methods [33]. It also uses techniques to treat data anomalies and recommends data processing based on decomposition to obtain reliable forecasts by detecting change points and use of data characteristics, pandemic characteristics, and payback periods. To calibrate the predictive performance, results are compared to a sequence of imputation techniques and forecasts derived from autoregressive models, machine learning, and deep learning models.

In [17], prediction models for visit time were used, which included linear regression, decision tree, and K-nearest neighbors, among others. In [8], a performance prediction study of students was conducted that used artificial neural networks, naive Bayes, decision tree, and logistic regression.

In [16], multilinear regression models were built for independent variables, and linear regression models were used to analyze the relationship between them and energy consumption to identify critical design variables for the energy performance of a building.

This study in [18] estimated the number of visitors from five tourism agencies using a machine learning method. The number of cases and deaths in Europe during the COVID-19 pandemic were considered using an artificial neural network (ANN), and regression of support vector (SVR) and multiple linear regression (MLR) were used as machine learning models.

Additionally, in [34] user activities based on the nature of various locations were predicted, and four models based on known machine learning techniques were proposed, including the generalized linear, logistic regression, deep learning, and gradient-boosted trees.

Table 1. Summary of the literature review.

No.	Title	Author(s)	Variables/Characteristics	Source of Data	Machine Learning Algorithm/Method	Validation Method
1	A Machine Learning-Based 10 Years Ahead Prediction of Departing Foreign Visitors by Reasons: A Case of Turkiye	Tutsoy, Onder, Tanrikulu, Ceyda [20]	Foreign citizens over 14 years of age for education, health, religion, shopping, transit, business, companions, travel	Face-to-face questionnaires 4 times a year between 2003 and 2022	Fractional-order polynomial prediction model and genetic algorithm optimization	Accuracy
2	A Novel Hybrid Deep Learning Approach for Tourism Demand Forecasting	Houria Laaroussi, Fatima Guerouate, Mohamed Sbihi [21]	Data from the Moroccan tourist office and Google Trends	Tourism Observatory in Morocco, search engines, Google Trends	Stacked LSTM	MAE, MSE, R2
3	A Paired Neural Network Model for Tourist Arrival Forecasting	Yao, Y.; Cao, Y.; Ding, X.; Zhai, J.; Liu, J.; Luo, Y.; Ma, S.; Zou, K. [22]	Tourism data: long-term trend component, seasonal component	Claveria and Torra (2014); Hassani, Silva, Antonakakis, Filis, and Gupta (2017)	Interlinked neural networks	MAPE, RMSE
4	Constructing Interval Models Using Neural Networks with Non-Additive Combinations of Grey Prediction Models in Tourism Demand	Jiang, Penga;Hu, Yi-Chung [32]	Visitor arrivals	Historical visitors data	Fuzzy integralgenetic algorithms	N/A
5	Design of Machine Learning Algorithms for Tourism Demand Prediction	Nan Yu, Jiaping Chen [25]	Monthly tourist volume search engine strength monthly data	Search engines	SAE-LSTM	MAE, RMSE, MAPE
6	Tourist Arrival Forecasting Using Multiscale Mode Learning Models	He, K.J. (He, Kaijian); Wu, D. (Wu, Don); Zou, Y.C. (Zou, Yingchao) [31]	Daily tourist arrival data in Macau from five major countries or regions	Extracted from the Macao Government Tourism Office (MGTO)	MD-CNN	MSE
7	Impact of COVID-19 on Demand Planning: Building Resilient Forecasting Models	Sreeja Ashok, Kanu Aravind [33]	Historical data	Real-word datasets derived from property group bookings	ARIMAXGBoostProphet LSTMCNN-ID	MAPE and RMSE
8	Visiting Time Prediction Using Machine Learning Regression Algorithms	Indri Hapsari, Isti Surjandari, Komarudin [17]	Information on the destination, transportation, hotel, weather, and exchange rate	Google data for both dependent variables and independent variables	Linear regression, KNN, decision tree, support vector machine, multilayer perceptron	Correlation coefficient, RMSE
9	Predicting Students' Performance Using Machine Learning Techniques	Hussein Altabrawee Osama Abdul Jaleel Ali, Samir Qaisar Ajmi [15]	Department, gender, studying style group, internet for study, extra learning resources, interest in studying computer, has computer experience, etc.	Faculty of Humanities: 2015 and 2016	ANN, DT, logistic regression, naive Bayes	ROC, F measure precision, precision, recall
10	Using Regression Models to Develop Green Building Energy Simulation by BIM Tools	Faham Tahmasebinial, Ruifeng Jiang, Samad Sepásgozar, Jinlin weil, Yilin Dingy, Hongyi ma [16]	Twelve variables for each model, including WWR, wall construction, roof construction, infiltration, lighting efficiency, plug load, HVAC, interior loads, envelope, building form, daylighting, occupancy, and building orientation	Variables determined automatically by the building performance simulation system	multilinear regression	Significances F and R2
11	Estimating the Changes in the Number of Visitors on the Websites of the Tourism Agencies in the COVID-19 Process by Machine Learning Methods	Mehmet Kayakus [18]	Daily numbers of cases and deaths in 54 European countries obtained from a European Union agency	Daily website visitor information was obtained from websiteiq.com	Red neuronal artificial (ANN), support vector regression (SVR), multiple linear regression (MLR)	R2, MSE, RMSE MAE, MAPE
12	Prediction and Classification of User Activities Using Machine Learning Models from Location-Based Social Network Data	Naimat Ullah Khan, Wanggen Wan, Rabia Riaz, Shuitao Jiang, and Xuzhi Wang [34]	Dataset to be applied to classify location-based social media data	Weibo data as the main source of research	Four models based on well-known machine learning techniques, including generalized linear model, logistic regression, deep learning, and gradient-boosted trees	Confusion matrices for classification, area under the curve (AUC) or receiver operating characteristic (ROC), accuracy, precision, recall, F-score, sensitivity

3. Materials and Methods

As detailed in Figure 1, the applied research methodology consisted of four main stages: (1) data collection, (2) model analysis, (3) model development, and (4) model evaluation, in addition to the technological support required for data processing.

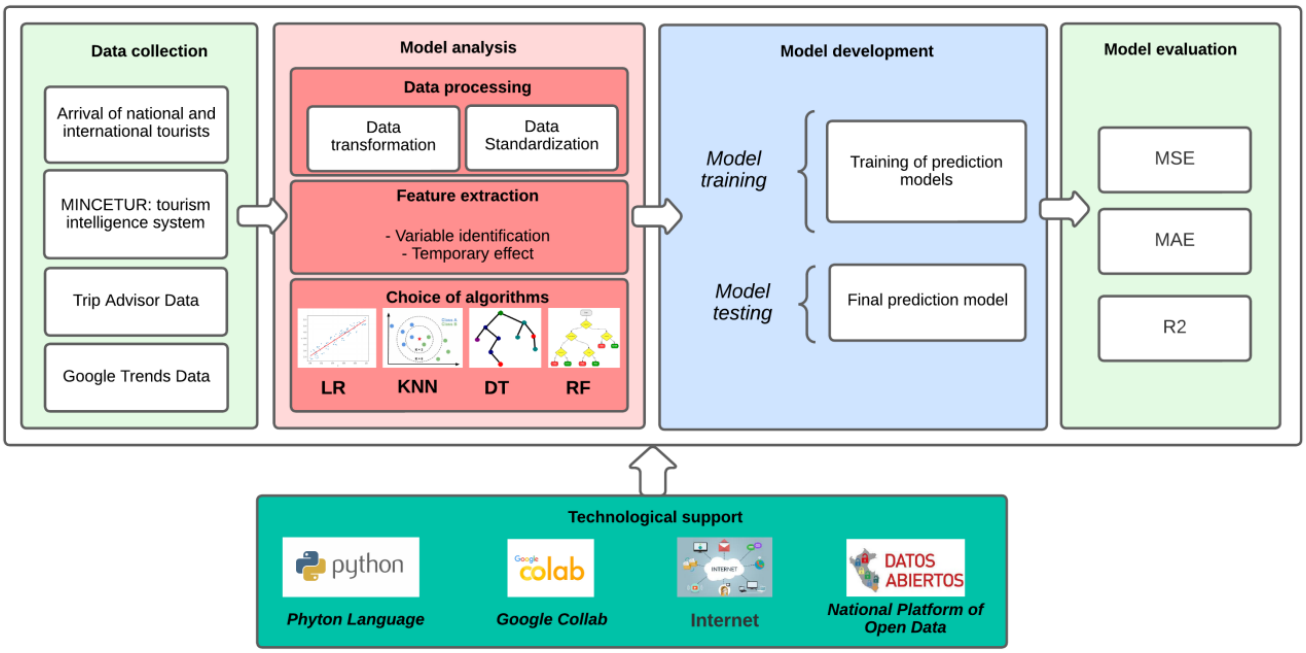

Figure 1. Investigation methodology.

3.1. Data Collection

In the first stage, we collected the data used in the proposed model, whose data dictionary was divided into five dimensions, as described in Table 2, over the period from January 2011 to May 2022.

The temporal dimension represents the months and years within the period evaluated.

The dimension MINCETUR Tourism Resources Data was extracted from the MINCETUR Tourism Intelligence System [35], representing data corresponding to the tourist attractions of the Moche Route, such as description, type, location, area, distance, nearby hotels, types of access (on foot, own mobility, on tour), and number of festivities.

For the Google Trends dimension, the Google Search tool was used, which allowed searching for keywords related to the Moche Route and determining trends of interest of potential tourists, such as access to airports and tourist interest, as well as their interest in Moche culture, as detailed in the results section, during the period covered by the research.

In the TripAdvisor dimension, which is the world's leading tourism platform, the opinions and comments of tourists who visited tourist sites included in the Moche Route were analyzed.

Finally, the last dimension, Tourist Arrivals—Open Data, are data extracted from the Internet [12], particularly the Open Data Platform of the Government of Peru, which represents the number of domestic and foreign tourists who visited the tourist attractions on a monthly basis.

It is worth mentioning that the tourist attractions lack qualitative information regarding the rating of the service by tourists (i.e., satisfaction surveys are not applied), whose information would be valuable to include in this research in order to obtain a much more accurate model.

Regarding the Moche Route, the scope of this research includes various types of tourist attractions: archaeological sites, natural sites, and museums, as listed in Table 3, since only for these attractions are data available for the arrival of national and foreign visitors.

Table 2. Variables used in the prediction model.

Dimension	Variable	Description	Type
Temporary	MONTH	Month of visit	Numeric
	YEAR	Year of visit	Numeric
MINCETUR Tourist Resources Data	PLACE	Description of the tourist attraction	String
	RESOURCE_TYPE	Type of tourist resource (1: Museum; 2: Archaeological site; 3: Natural site)	Categorical
	PROVINCE	Province where it is located	String
	DEPARTAMENT	Department where it is located	String
	DISTRICT	District where it is located	String
	DISTANCE_KM	Distance to the nearest city (Km)	Numeric
	HOTELS	Number of hotels close to the attraction	Numeric
	ZONE	Zone type (1: Rural; 2: City)	Categorical
	ACCESS_FOOT	Sf there is access by foot to the tourist attraction (1: Yes; 0: No)	Categorical
	ACCESO_MOV	Is there is mobile access to the tourist attraction (1: Yes; 0: No)	Categorical
	ACCESS_TOUR	Is there access through a tour to the tourist attraction (1: Yes; 0: No)	Categorical
	AMOUNT_FEST	Number of festivities near the tourist place	Numeric
Google Trends	ACCESS_AIRPORT	Index of trends to the main entry points to the Moche Route (Chiclayo and Trujillo airports)	Numeric
	INTEREST_TUR	Index of trends to each tourist attraction	Numeric
	INT_CULMOCHE	Moche Culture Trends Index	Numeric
TripAdvisor	COMMENTS	Number of comments	Numeric
Tourist Arrivals—Open Data	NATIONAL_TOTAL	Number of national tourists who visit the tourist attractions monthly	Numeric
	FOREIGN_TOTAL	Number of foreign tourists who visit tourist attractions monthly	Numeric
	AVERAGE_NATIONAL	Average number of national visitors according to the month of visit by tourist attraction	Numeric
	AVERAGE_FOREIGN	Average number of foreign visitors according to the month of visit by tourist attraction	Numeric

Table 3. Tourist attractions of the Moche Route.

Tourist Attraction	Type	Location
Huaca Arco Iris Archaeological Complex	Archaeological site	La Libertad
Huaca del Sol y la Luna Archaeological Complex	Archaeological site	La Libertad
Huaca el Brujo Archaeological Complex	Archaeological site	La Libertad
Chan Chan Site Museum	Museum	La Libertad
Brüning National Archaeological	Museum	Lambayeque
Huaca Chotuna Site Museum—Chornancap	Museum	Lambayeque
Huaca Rajada Site Museum—Sipan	Museum	Lambayeque
Tucume Site Museum	Museum	Lambayeque
Sican National Museum	Museum	Lambayeque
Royal Tombs of Sipan Museum	Museum	Lambayeque
Pomac Forest Historical Sanctuary	Natural site	Lambayeque

3.2. Model Analysis

In this stage, data preprocessing was carried out followed by feature extraction to obtain valuable and representative information from the dataset, culminating in the choice of the prediction algorithms that were used in this research.

Within the prediction algorithms, it was determined to use regression algorithms, such as linear regression (LR), KNN, decision tree (DT), and random forest (RF).

Linear Regression (LR)

The linear regression model tries to explain the relationship between a dependent variable (i.e., response variable) and a set of independent variables (i.e., explanatory variables), $X_1, \ldots, X_n$, which are mainly used when the dependent variable is a continuous variable.

The linear regression model fits a straight line or a surface that minimizes discrepancies between predicted and actual output values [36].

The general formula that represents this algorithm is observed in Equation (1).

$$Y = \beta_0 + \beta_1 X + \in \quad (1)$$

where Y is the predictable variable; X is the variable(s) used to make a forecast; and $\in$ is the error.

For this algorithm, the following parameter values were used. For the *fit_intercept* parameter, the β_0 intercept was calculated and used with a default value of true. For the *normalize* parameter, which normalizes the input variables, a parameter value of false was considered.

Finally, the *n_jobs* parameter, which represents the number of jobs used in the parallel computation, was defaulted to none.

K Nearest Neighbor (KNN)

The KNN method is widely used in data mining and machine learning applications because of its simple implementation and high performance [37].

It is a nonparametric supervised learning classifier that uses proximity to make classifications or predictions about the clustering of an individual data point. It can be used for both regression and classification problems.

For regression, it takes the average of the k nearest neighbors to make a prediction about a classification. The main difference here is that classification is used for discrete values, while regression is for continuous values. However, before a classification can be made, the distance must be defined. The Euclidean distance is the most used.

For the prediction of national and foreign tourists, the "minkowski" metric was used to calculate the *Euclidean distance*, adding a power to this metric ($p = 2$).

The *n_neighbours* parameter represents the number of close neighbors, and the value of five was used.

The *weights* parameter was the weight function for the prediction, and the value used was "uniform", where all points of each neighbor were weighted equally.

The *algorithm* parameter was used to calculate the nearest neighbor, and "auto" was used, where the algorithm determined the function for the past values that gave the best results for the training.

The *leaf_size* had a value of 30.

Decision Tree (DT)

The decision tree (DT) is the simplest inductive learning method [31]. It belongs to the data mining tool and can handle continuous and noncontinuous variables. It establishes the tree structure diagram mainly by the given classification fact and induces some principles. The principles are mutually exclusive, and the generated DT can also make an out-of-sample prediction.

Regarding the parameters, the value "squared_error" was used for the *criterion* parameter, which expresses the quality of the division of the nodes based on the reduction of the variance.

The strategy for choosing the split in each node was the *splitter* parameter, and it was set to "best".

The *max_depth* parameter represents the nodes to expand until they reach the configured minimum, which was set to none. The rest of the parameters worked with their default values.

Random Forest (RF)

RF has grown in popularity because of its high reliability and practical application in various fields of study [31,38].

RF can be used for a categorical response variable, referred to in [39] as "classification", or a continuous response, referred to as "regression". Similarly, predictor variables can be categorical or continuous, as expressed by in [40].

The random forest (RF) classifier is a meta-estimator that fits a series of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. The subsample size is always the same as the original input sample size.

Regarding the main parameters used (i.e., the *n_estimators*), it operated with a value of 100, which refers to the number of trees in the forest. The true value of the *bootstrap* parameter was configured, which means that it used the bootstrap samples when building the trees.

The rest of the algorithm's parameters used the default values.

3.3. Model Development

In the third stage, the forecast models were developed, which included two steps, and then the models were trained, culminating in the testing of the proposed models.

3.4. Model Evaluation

The model evaluation, based on mean absolute error (MAE), mean square error (MSE), and R2, was conducted to examine the accuracy of the out-of-sample prediction model.

3.5. Technological Support

Regarding technological support for the processing, the Python programming language version 3 and storage in Google Collab were used.

In addition, the Internet was used to search for information and access to the national open data platform published on Peruvian portals.

4. Results

The results provided below are based on the developed methodology and applied to the prediction of national and foreign tourists to the tourist attractions of the Moche Route in Peru.

4.1. Data Collection

For the present study, data collected on arrivals of national and international tourists to the Moche Route were analyzed, obtaining official data from the Open Data Platform of the Government of Peru in the .csv format, as well as from the Information System Tourist Intelligence of MINCETUR in the .xlsx format, from January 2011 to May 2022.

Table 4 shows the keywords used in Google Trends that allowed us to analyze search trends and user preferences regarding access to airports and interests of tourists, as well as their interests related to Moche culture, as the objects of the investigation.

With the complete data, the monthly arrivals of national and foreign tourists, keywords in Google Trends, and comments received for each tourist attraction of the Moche Route, a descriptive statistical analysis was carried out to obtain summarized and quantitative information on the different variables.

Table 5 summarizes the descriptive statistics, finding important data. For example, regarding the distance in km from the tourist destination to the nearest city, there is an

average of 21.97 km and a standard deviation of 17.73, revealing considerable variability between distances. In the case of hotels around the tourist place, an average of 1.73 and a standard deviation of 0.44 were obtained, which indicates that the number of hotels tends to be similar or constant for most tourist attractions.

Table 4. Keywords in Google Trends.

Variable *	Keyword
ACCESS_AIRPORT	• Chiclayo Airport • Trujillo Airport
INTEREST_TUR	• Huaca Arco Iris Archaeological Complex • Huaca del Sol y la Luna Archaeological Complex • Huaca el Brujo Archaeological Complex • Chan Chan Site Museum • Brüning National Archaeological Museum • Huaca Chotuna Site Museum—Chornancap • Huaca Rajada Site Museum—Sipan • Tucume Site Museum • Sican National Museum • Royal Tombs of Sipan Museum • Pomac Forest Historical Sanctuary
INT_CULMOCHE	• Moche culture

* ACCESS_AIRPORT: main entry points to the Moche Route (Chiclayo and Trujillo Airports); INTEREST_TUR: index of trends to each tourist attraction; INT_CULMOCHE: index of trends on the Moche culture.

Table 5. Descriptive analysis for the data under study.

Variable *	Count	Mean	Std.	Min.	25%	50%	75%	Max.
DISTANCE_KM	1534	21.97275	17.73202	3.8	5.4	17.1	32.6	60
HOTELS	1534	1.732073	0.443024	1	1	2	2	2
ZONE	1534	1.267927	0.443024	1	1	1	2	2
AMOUNT_FEST	1534	3.228814	2.70255	0	0	3	6	6
TYPE_RESOURCE	1534	1.469361	0.661128	1	1	1	2	3
ACCESS_FOOT	1534	0.732073	0.443024	0	0	1	1	1
ACCESS_MOV	1534	1	0	1	1	1	1	1
ACCESS_TOUR	1534	1	0	1	1	1	1	1
INT_CULMOCHE	1534	57.50261	19.32601	0	43	60	70	98
ACCESS_AIRPORT	1534	28.99804	14.26231	0	21	29	39	67
INTEREST_TUR	1534	24.59518	23.76916	0	0	21.5	38	100
COMMENTS	1534	2.624511	5.04408	0	0	0	3	40
NATIONAL_AVERAGE	1534	3862.555	3701.999	132	1693	2891	4430	19,697
FOREIGN_AVERAGE	1534	599.9296	628.14	5	167.75	374	780	2709
TOTAL_NATIONAL	1534	3862.527	4015.052	0	1403.25	2701	4619	27,506
TOTAL_FOREIGN	1534	599.9035	714.6349	0	115.25	330.5	786	6365

* DISTANCE_KM: distance to the nearest town; HOTELS: number of hotels nearby; ZONE: rural or city type; AMOUNT_FEST: number of festivities; TYPE_RESOURCE: Museum, Archaeological Site, or Natural Site; ACCESS_FOOT: if there is access by foot or not; ACCESO_MOV: if there is access with mobility or not; ACCESS_TOUR: if there is access through a TOURS or not; INT_CULMOCHE: index of trends on Moche Culture; ACCESS_ AIRPORT: corresponds to entry points to the airports of Chiclayo and Trujillo; INTEREST_TUR: index of trends to each tourist attraction; COMMENTS: number of comments made by tourists extracted from Google Trends and TripAdvisor; NATIONAL_AVERAGE: average number of national visitors according to the month of visit; FOREIGN_AVERAGE: average number of foreign visitors according to the month of visit; TOTAL_NATIONAL: total monthly number of national tourists visiting tourist attractions; TOTAL_FOREIGN: total monthly number of foreign tourists visiting tourist attractions.

On TripAdvisor, the opinion of tourists when visiting tourist places was analyzed. Table 6 shows an example of this analysis for one tourist place of the Moche Route, where 696 comments were observed for the Royal Tombs of Sipan Museum, representing lived tourist experiences.

Similarly, the trend in the number of visitors over time was analyzed using a line graph, seeking to identify the patterns and seasonality in tourist flows, as well as to identify the most visited tourist destinations that comprise the Moche Route.

Table 6. Data on comments for the Royal Tombs of Sipan Museum on TripAdvisor.

Variable	Type	Example
Forum	String	Royal Tombs of Sipan Museum
Comment	String	This museum is an absolute must see as the treasures are phenomenal
Comment link	String	https://www.tripadvisor.com/ShowUserReviews-g1926372-d1951110-r735889734-Royal_Tombs_of_Sipan_Museum-Lambayeque_Lambayeque_Region.html (accessed on 10 March 2023)
Author of the comment	String	Evelynne G.
Author profile	String	https://www.tripadvisor.com/Profile/EvelynneG (accessed on 10 March 2023)
Comment date	Date	31 December 2019
Number of comments	Numeric	696

Figure 2 shows the arrival of national visitors to Moche Route—La Libertad, where it can be seen that the Huaca del Sol y de la Luna Archaeological Complex had the highest number of visitors. On the other hand, Figure 3 shows the arrival of national visitors to Moche Route—Lambayeque, where the Royal Tombs of Sipan Museum had the highest number of visitors.

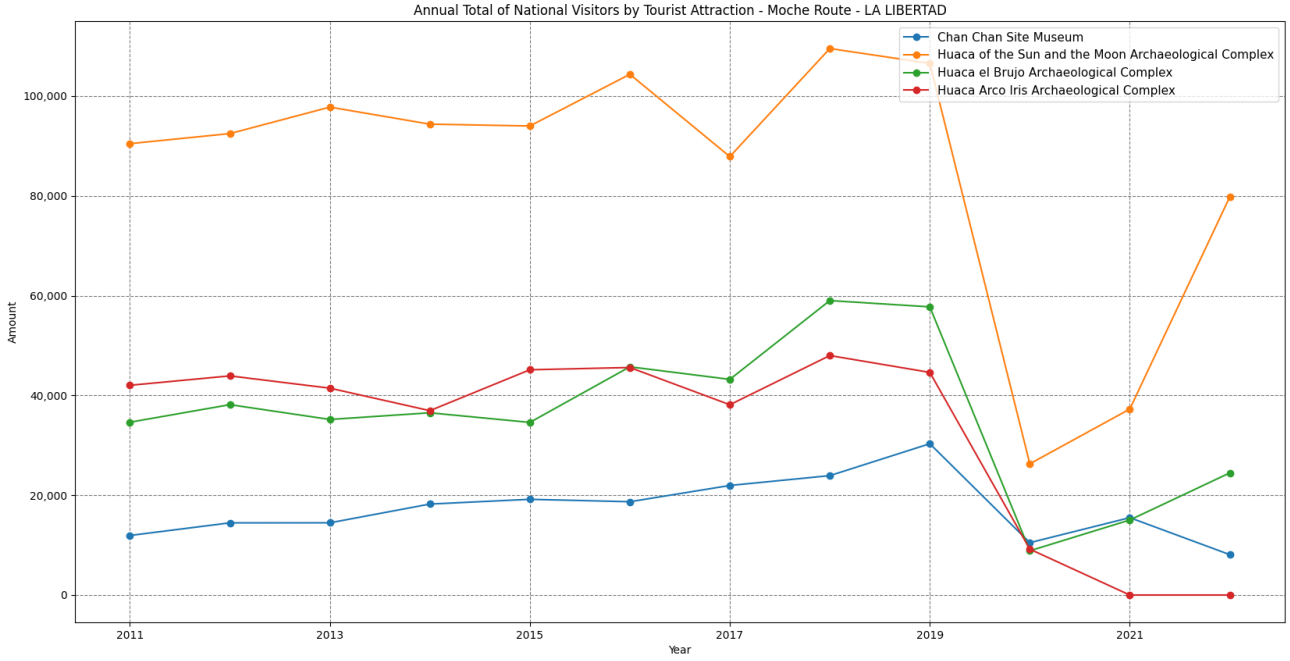

Figure 2. Arrivals of national visitors to the Moche—La Libertad Route between 2011 and 2022.

Figure 4 shows the arrival of foreign visitors to Moche Route—La Libertad, where it can be seen that the Huaca del Sol y de la Luna Archaeological Complex had the highest number of visitors. On the other hand, Figure 5 shows the arrival of foreign visitors to Moche Route—Lambayeque, where the Royal Tombs of Sipan Museum had the highest number of visitors.

Finishing this analysis of the data, as shown in Figure 6, where it is observed that national tourists represented the largest number of visitors to the tourist resources that make up the Moche Route.

4.2. Model Analysis

Data preprocessing allowed for the records that were obtained in the .csv and .xlsx files to be unified into a single data collection format, consolidating a total of 1534 records

that comprised the tourists who arrived at the eleven tourist resources that make up the Moche Route. Some records were found that did not have data in some months, which were normalized using the average of that month for all years of the study period. None of the data obtained were deleted.

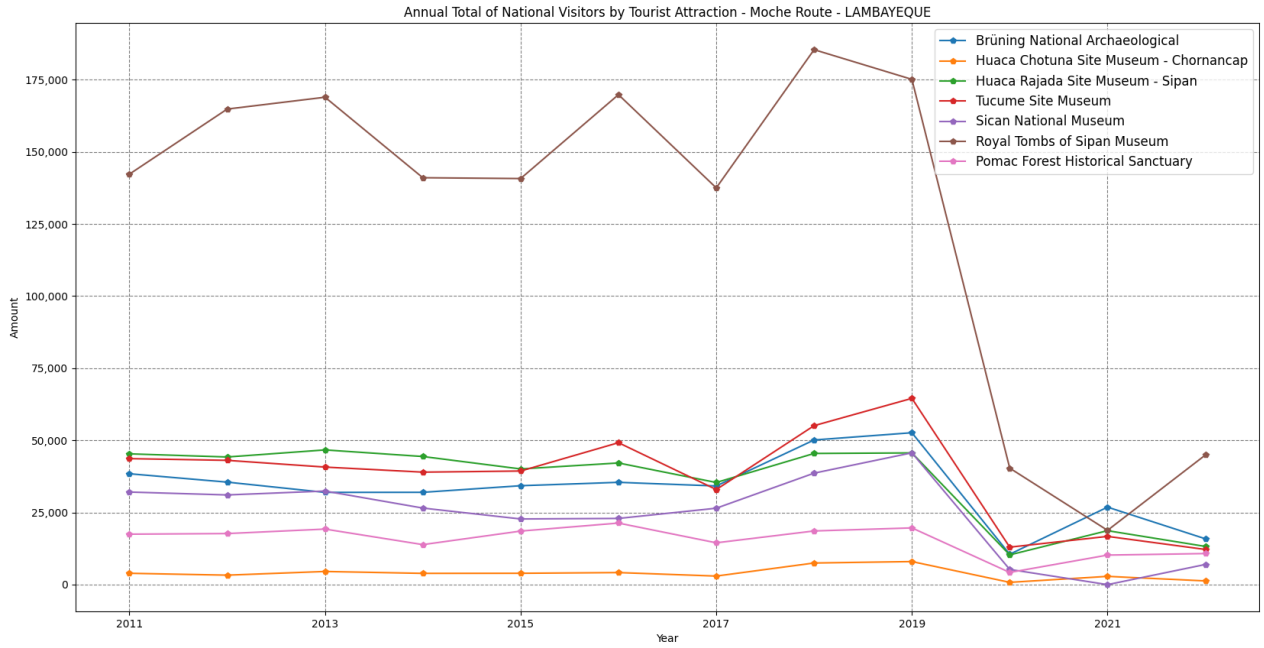

Figure 3. Arrivals of national visitors to the Moche—Lambayeque Route between 2011 and 2022.

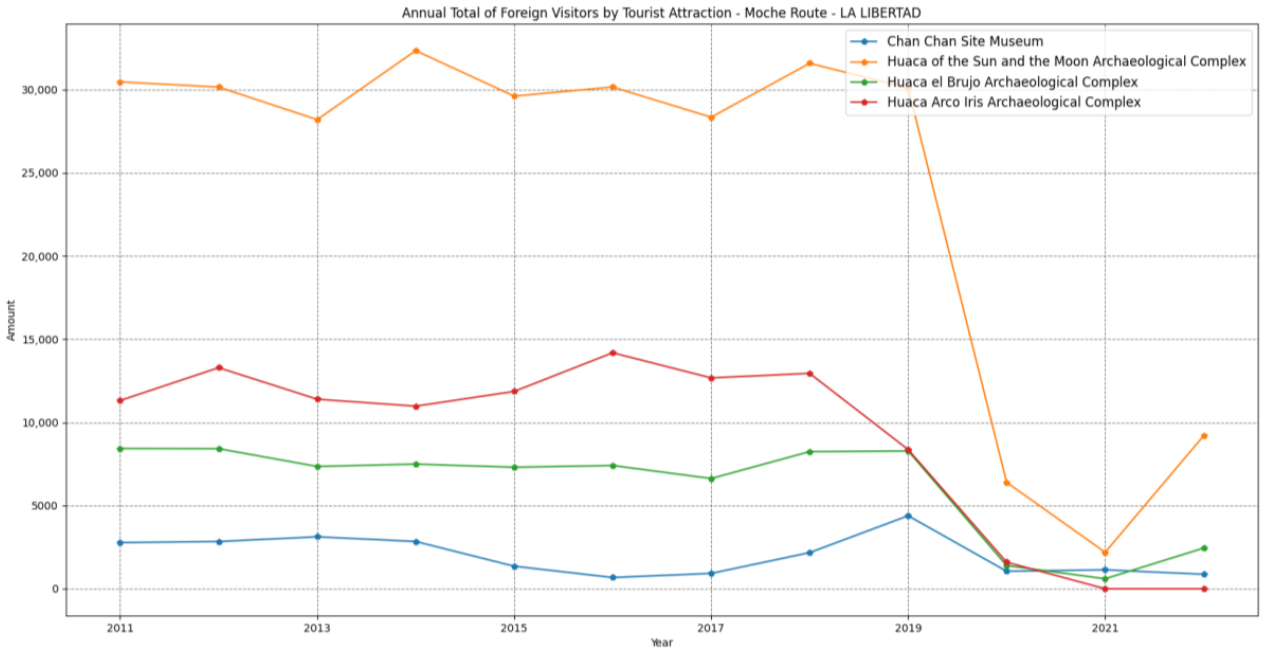

Figure 4. Foreign visitor arrivals to the Moche—La Libertad Route between 2011 and 2022.

The extraction of the characteristics made it possible to define the variables that complemented the data under study, where a total variable was built for each month as a variable of inertia or delay characteristic, represented by the average of all tourist visits in the 12 months of each year. The initial characteristics were 26, finally leaving 11 relevant characteristics in the data.

According to the correlation of the variables using the Pearson method, the degree of the linear relationship between each pair of variables was measured, while the correlation

value was closer to the value of one, which indicates that the variables can increase or decrease at the same time.

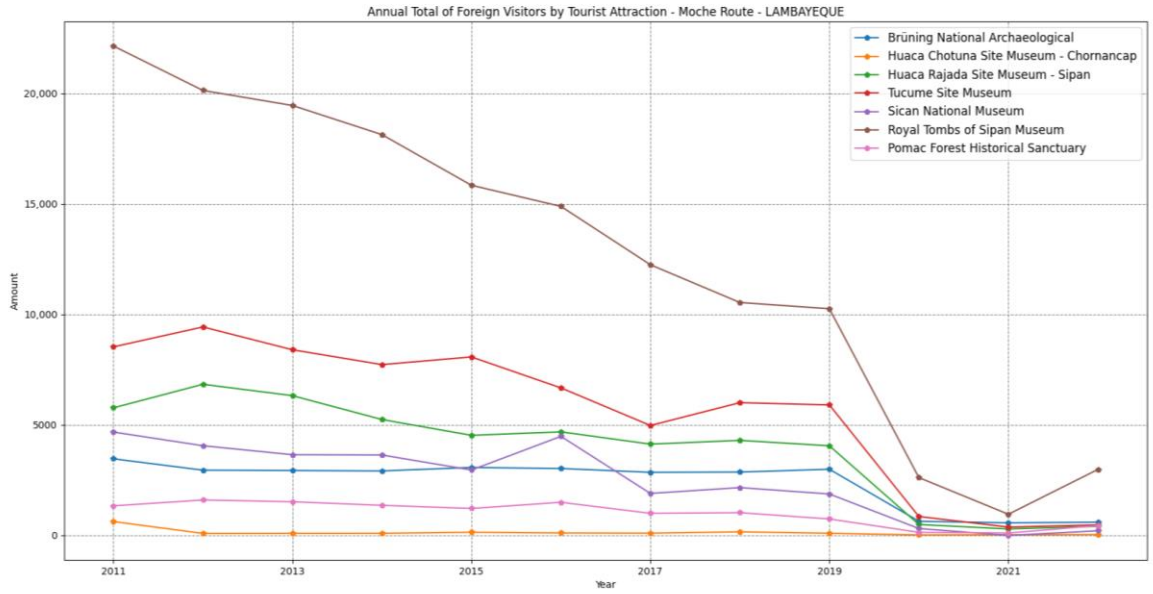

Figure 5. Foreign visitor arrivals in the Moche—Lambayeque Route between 2011 and 2022.

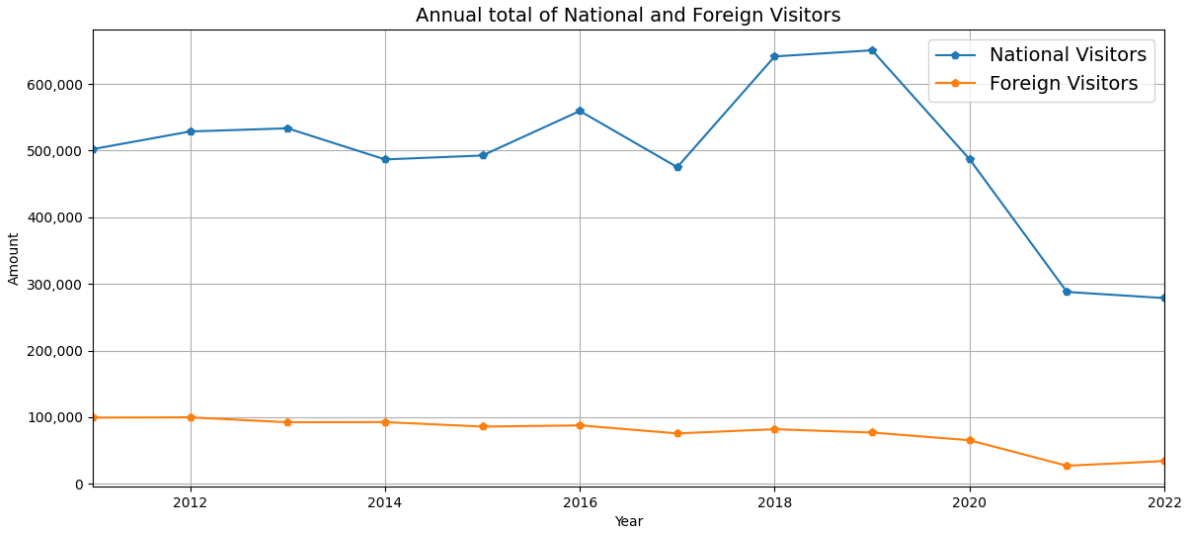

Figure 6. Arrival of national and foreign visitors between 2011 and 2022.

Figure 7 shows the correlations among the variables under study; for example, a high correlation was identified between the variables ZONE and HOTELS, which indicates the presence of a greater number of hotels in the city areas; also, for ZONE and ACCESS_FOOT, it is understood that there is ease of access on foot in the areas where they are located; in addition, a medium correlation was detected between variables such as AMOUNT_FEST and HOTELS and AMOUNT_FEST and TYPE_RESOURCE, among other existing variables.

Figure 8, for the model of foreign visitors, shows the correlation among the variables under study. For example, a high correlation is identified between ZONE and ACCESS_FOOT, understood as the ease of access on foot in the areas where they are located; there is also a medium correlation between AMOUNT_FEST and HOTELS and AMOUNT_FEST and ZONE, among others.

Taking into account that it seeks to apply machine learning techniques to predict visitors, in addition to the fact that large volumes of data have not been found and there is few and incomplete data, in this first scope of research on the application of machine

learning, the use of main regression algorithms, such as linear regression, KNN, random forest, and decision tree, was applied to understand their behavior and interpretation of their results, which will allow the future application of more advanced methods.

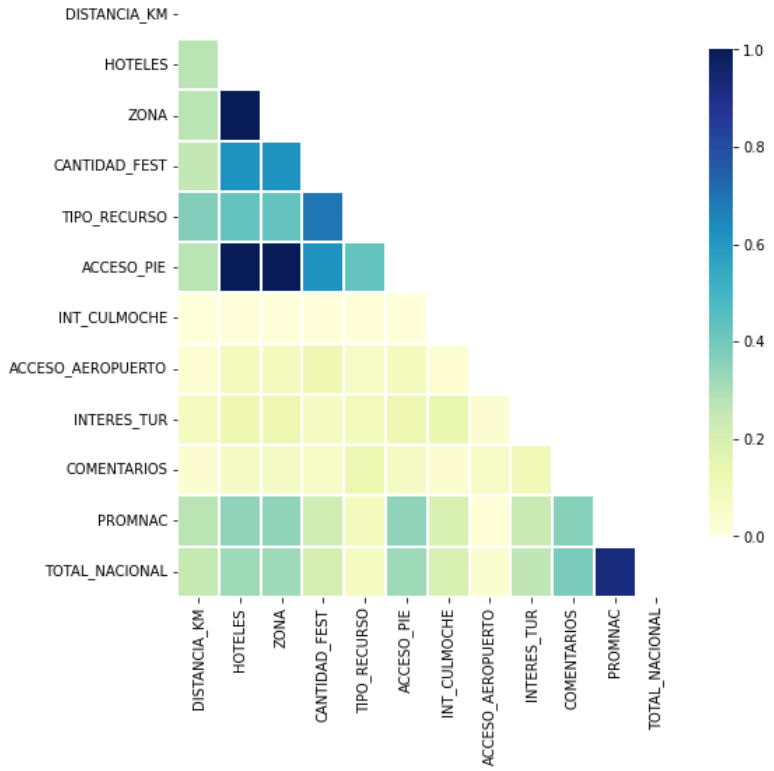

Figure 7. Correlation among variables for the model of national visitors.

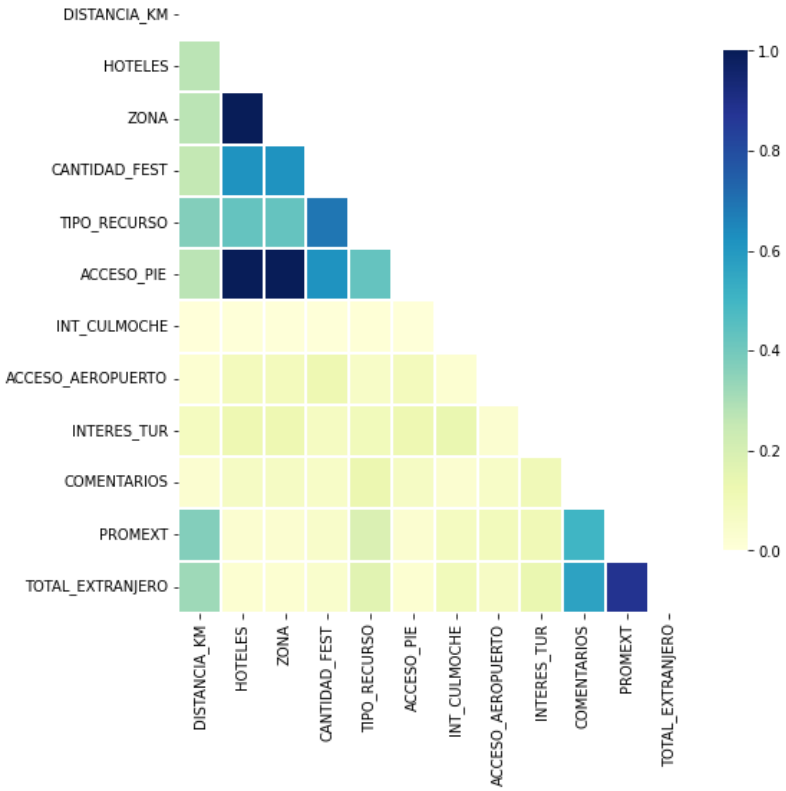

Figure 8. Correlation among variables for the model of foreign visitors.

4.3. Development of Models

The training and testing of the machine learning models (national and foreign) allowed for the evaluation and improvement of their predictive capacity. The entire dataset was divided into two parts: 80% data (1227 instances) for the model training and 20% for the testing (307 instances). The models were implemented with Python and the Google Collab tool, using the default settings for the algorithms used.

The training allowed the models to learn the relationships among the provided variables, such as the distance to a nearest city and the number of visitors, with which the models adjusted its parameters to minimize the error between the predictions and the actual values of the training set. The testing of the models and the evaluation of their performance allowed the models to generalize and make accurate predictions.

The algorithms were evaluated with the training data and using the cross-validation strategy with a value of 10 folds, as well as using the negative mean square error (neg_mean_squared_error) as scoring.

In Figure 9, for the National Tourist Model, a box diagram was built where the evaluation of the regression algorithms and the metric "Negative Mean Quadratic Error" was appreciated, which is commonly used in automatic learning to evaluate the performance of the models. The best result is when its value is closest to zero. In this case, the linear regression algorithm was better than the rest of the models, which indicates that it is more precise in the prediction of the data. In each box, the central brand (orange line) is the median, and the edges of the boxes are the 25 and 75 percentiles. The small circles represent atypical values.

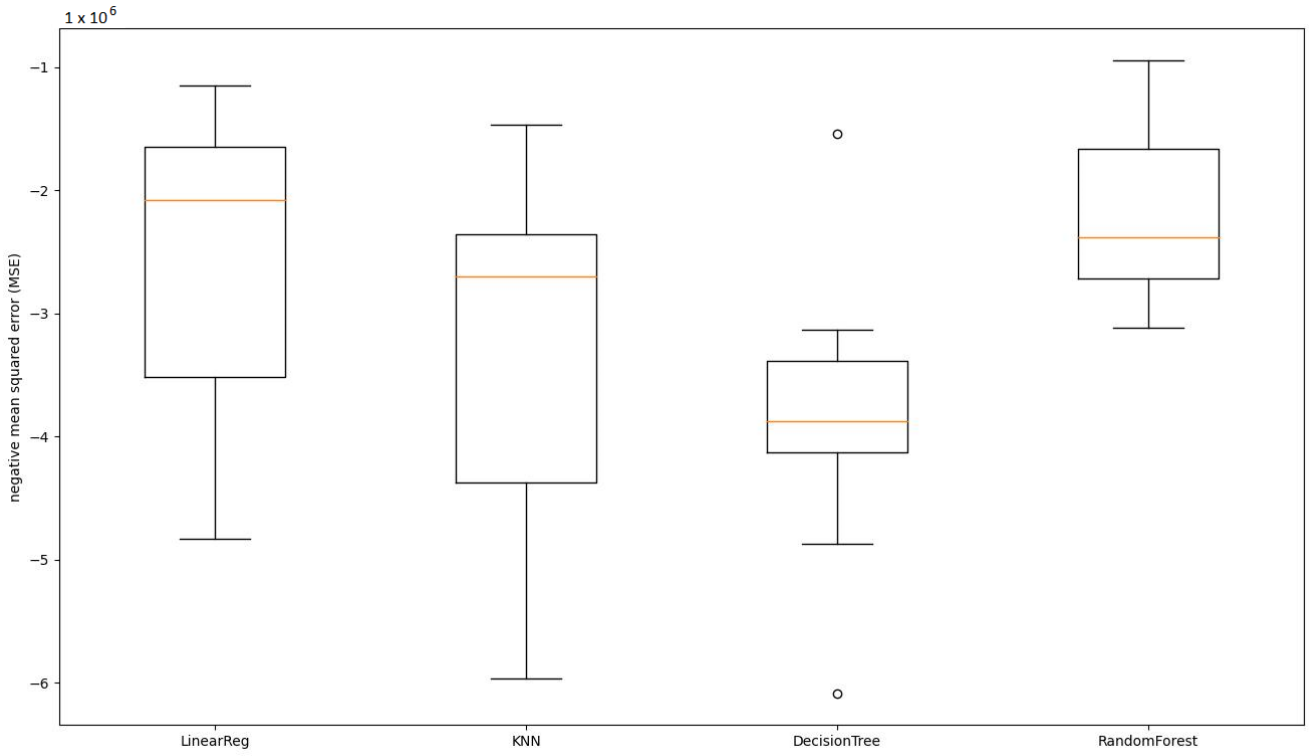

Figure 9. Results of the evaluation of the algorithms for the National Tourist Model.

In the analysis of the same regression algorithms used in the Foreign Tourist Model, in the analysis of the box diagram, it is observed in Figure 10 that the best result is the linear regression algorithm, which means that it is the most precise in the prediction of the data. In each box, the central brand (orange line) is the median, and the edges of the boxes are the 25 and 75 percentiles. The small circle represents an atypical value.

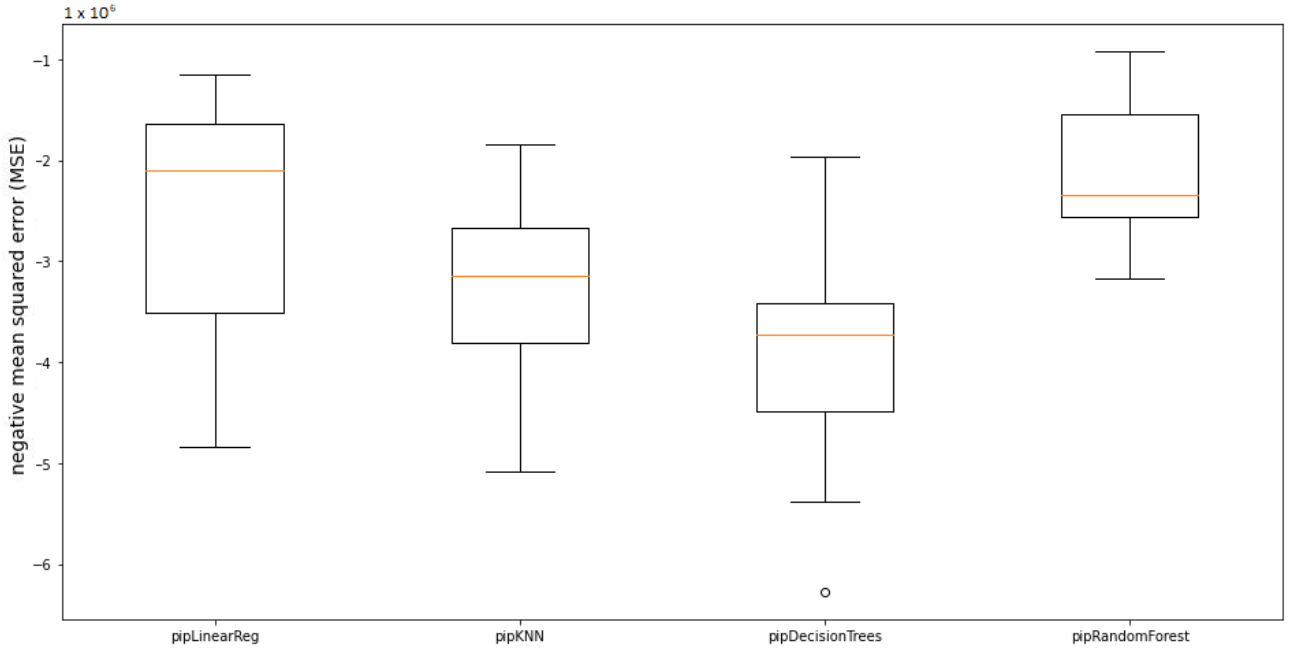

Figure 10. Results of the evaluation of the algorithms for the Foreign Tourist Model.

Regarding the testing and validation, in Figure 11 it can be seen that between the real data (blue) and predicted data (orange) of the National Tourist Model, for the first 50 instances of the 307 instances for the test, it had moderately high peaks in prediction, but not in the instances between 150 and 200, where the peaks had low prediction. Regarding the Foreign Tourist Model, according to Figure 12, for the instances between 0 and 100, there were moderately high prediction peaks, which fell for the instances between 250 and 300.

In Figures 11 and 12, the instances representing the test dataset are visualized and compared with the real data in order to observe the prediction level of the applied models.

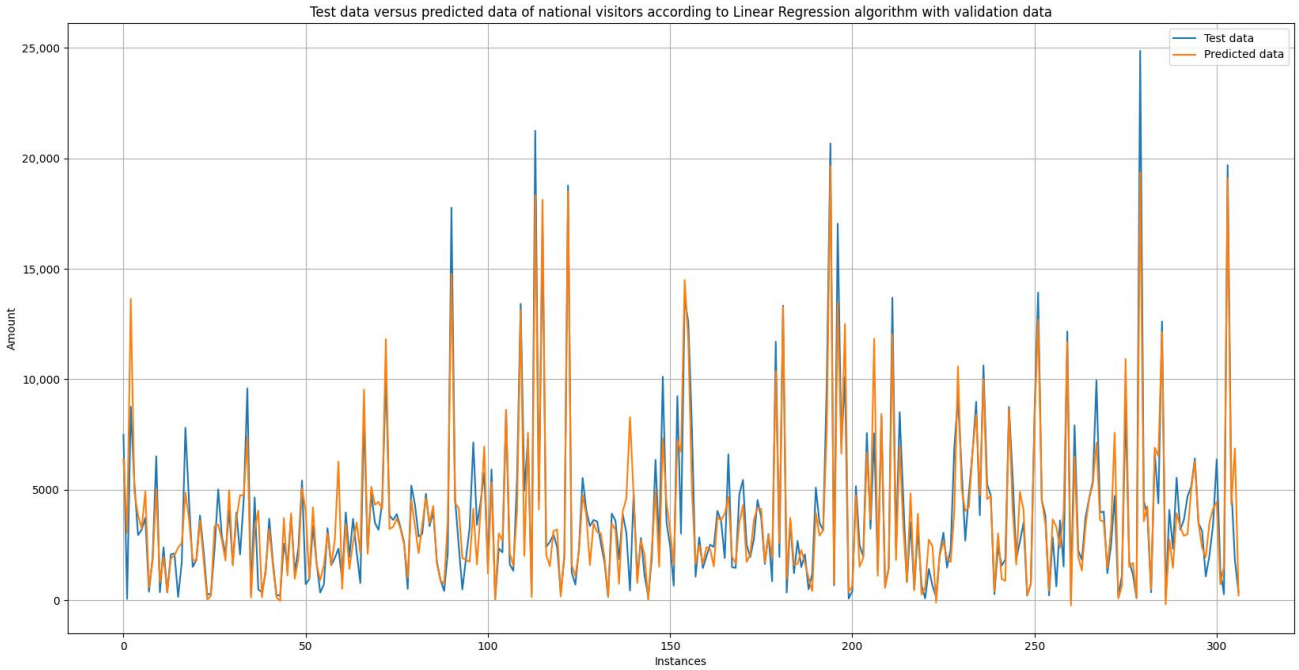

Figure 11. Testing and validation of the National Tourist Model.

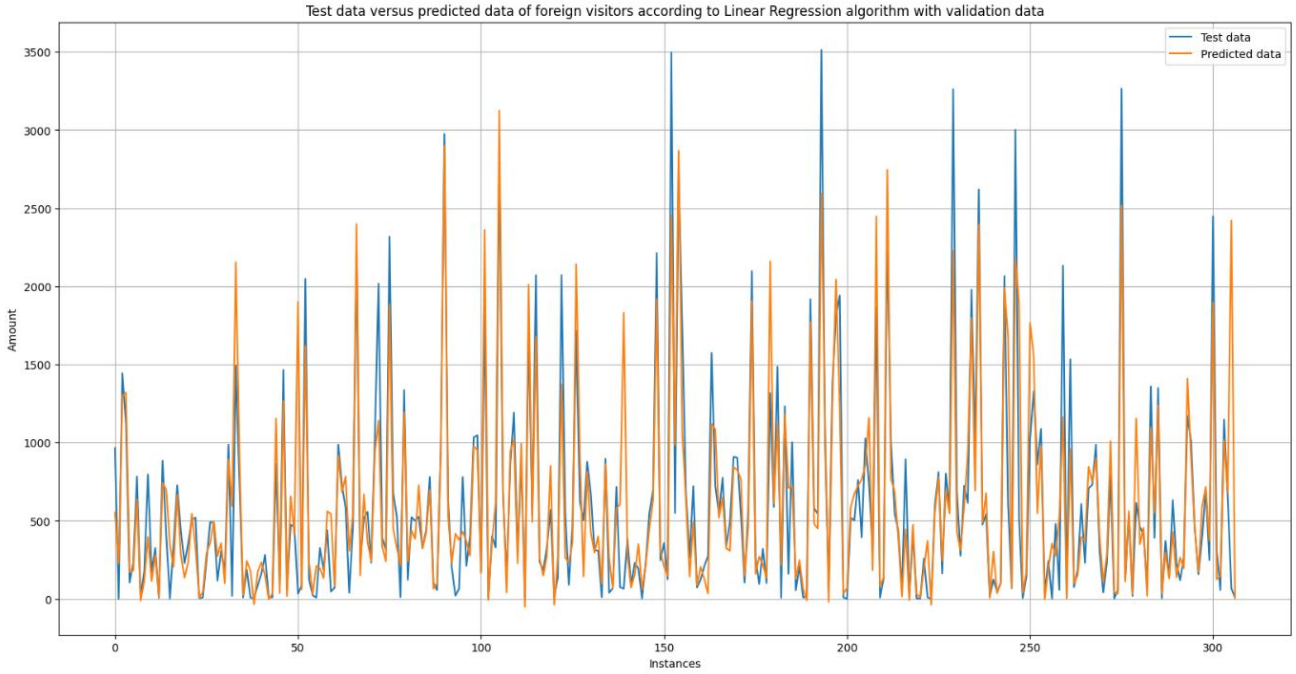

Figure 12. Testing and validation of the Foreign Tourist Model.

4.4. Model Evaluation

Since the algorithms used belong to regression models, their performances were evaluated using evaluation techniques or metrics: MSE, MAE, and R2. Precision and accuracy metrics, which are applied to classification models, were not considered.

- Prediction models for national visitors

Based on the results in Table 7, which determine the quality indicator models developed for the prediction of national visitors, it can be observed that the linear regression algorithm had the lowest values of MSE and MAE, indicating that it was the most accurate model. In addition, it had the highest value of R2, which suggests that the model can better explain the variability of the observed data compared to the other models, achieving more accurate predictions.

Table 7. Evaluation metrics of the models for the prediction of national visitors.

Regression Algorithm	MSE *	MAE *	R2 *
Linear regression	1,733,504.069299	849.511049	0.884650
KNN	2,161,560.164951	951.232573	0.856167
Random forest	1,854,965.651957	894.229218	0.876568
Decision tree	2,212,562.291144	975.122146	0.852773

* MSE: mean squared error; MAE: mean absolute error; R2: coefficient of determination or R-squared.

On the other hand, the decision tree algorithm had the lowest value of R2 and the highest values of MSE and MAE, indicating that it was the least accurate model.

The prediction model of national visitors with the linear regression algorithm had the highest precision, and it can be seen in Figure 13 that, for the relationship between the real data and the prediction data, there were many points near the line, so they followed the same relationship as the other data and can be used to predict values.

The linear regression equation used to calculate the prediction of the arrival of national visitors is expressed in Equation (2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \in \quad (2)$$

where Y is the predicted variable; X_i is the variable(s) used to make the forecast; β_i is the coefficient for each variable used in the prediction; and $\in$ is the error.

Applied to the data used in the model, Equation (3) is observed.

$$Y = 3856.44 + 2.58 \times 10^{+01}X_1 - 3.37 \times 10^{+16}X_2 - 1.98 \times 10^{+16}X_3 + 1.80 \times 10^{+01}X_4 - 3.80 \times 10^{+01}X_5 + 1.39 \times 10^{+16}X_6 + 6.05 \times 10^{+01}X_7 + 6.25 \times 10^{+01}X_8 + 2.00 \times 10^{+02}X_9 + 2.62 \times 10^{+02}X_{10} + 3.55 \times 10^{+03}X_{11} + \in \quad (3)$$

where: X_1: DISTANCIA_KM; X_2: HOTELES; X_3: ZONA; X_4: CANTIDAD_FEST; X_5: TIPO_RECURSO; X_6: ACCESO_PIE; X_7: INT_CULMOCHE; X_8: ACCESO_AEROPUERTO; X_9: INTERES_TUR; X_{10}: COMENTARIOS; and X_{11}: PROMNAC.

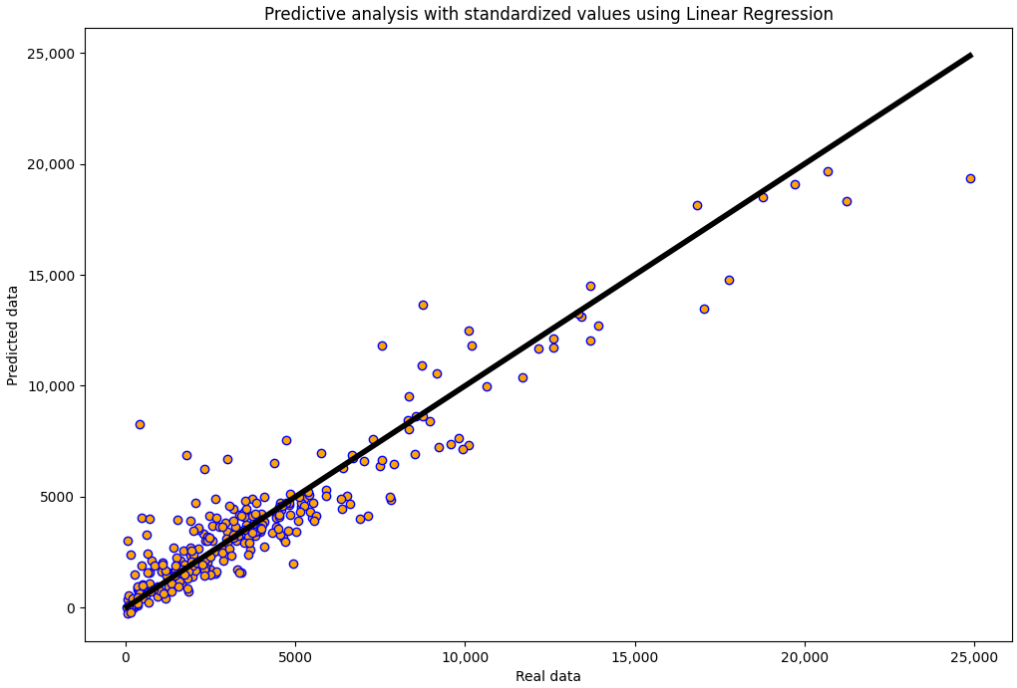

Figure 13. Predictive analysis of the model with the linear regression algorithm for national visitors.

- Prediction models for foreign visitors

Regarding the models developed for the prediction of foreign visitors, according to Table 8, it can be observed that the four algorithms used present a moderate performance in terms of prediction accuracy. The linear regression algorithm obtained the lowest MSE, which indicates that it had the fewest mean square errors in the prediction; in addition, it also obtained the highest value of R2, which suggests that it adequately explained the variation in the data.

Table 8. Evaluation metrics of the models in the prediction of foreign visitors.

Regression Algorithm	MSE *	MAE *	R2 *
Linear regression	113,269.739472	192.293311	0.765636
KNN	142,328.774201	216.495114	0.706867
Random forest	121,397.996075	189.389087	0.749273
Decision tree	178,482.557003	239.671009	0.630705

* MSE: mean squared error; MAE: mean absolute error; R2: coefficient of determination or R-squared.

The random forest algorithm had a similar MSE and MAE as the linear regression, with a slightly lower R2 value, suggesting that it did not explain the variation in the data as well as the linear regression. The KNN and decision tree algorithms had the highest MSE and MAE, which indicates that they had a lower prediction accuracy; likewise, they had lower R2 values, which is why they insufficiently explained the variation in the data.

Figure 14 shows the relationship between the real data and the predicted data for foreign visitors with the prediction model and the linear regression algorithm, and it can be seen that it presents a greater dispersion of some of the points towards the central line, for which reason it can be inferred that they are outliers that do not have the same relationship with those that are closer to the line, and, therefore, have moderate prediction performance.

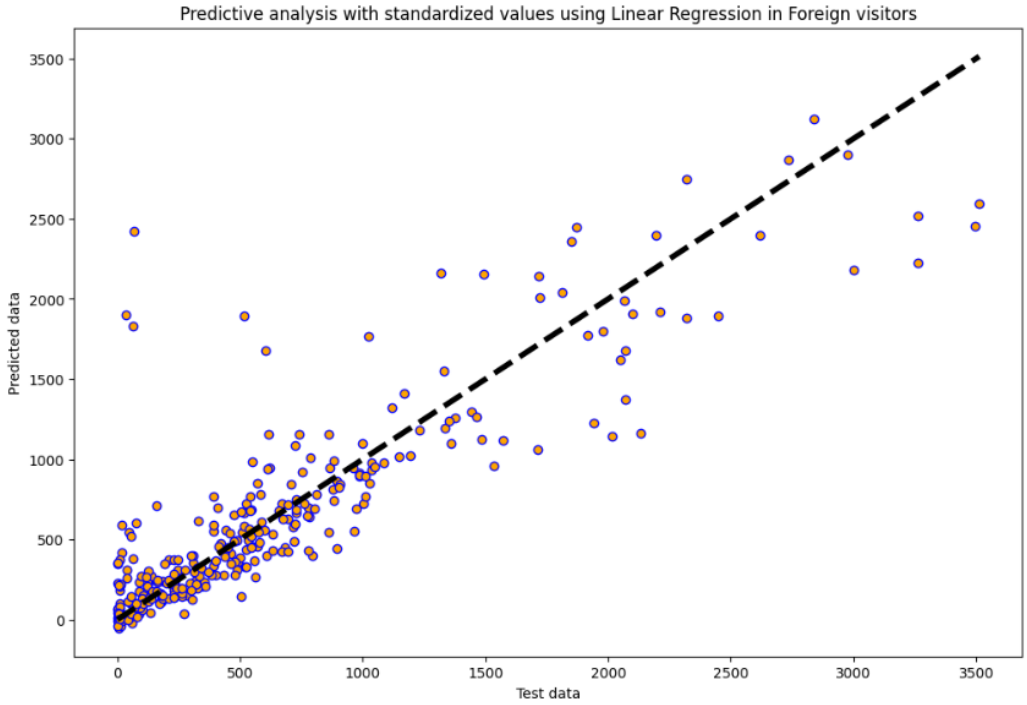

Figure 14. Predictive analysis of the model with the linear regression algorithm for foreign visitors.

The linear regression equation used to calculate the prediction of the arrival of foreign visitors is shown in Equation (4).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \in \quad (4)$$

where Y is the predicted variable; X_i is the variable(s) used to make a forecast; β_i is the coefficient for each variable used in the prediction; and $\in$ is the error.

Applied to the data used in the model, the formula is expressed in Equation (5).

$$Y = 596.79 - 2.6 \times 10^{+01} X_1 - 1.24 \times 10^{+15} X_2 - 6.12 \times 10^{+14} X_3 + 4.13 \times 10^{+00} X_4 + 1.24 \times 10^{+01} X_5 + 6.30 \times 10^{+14} X_6 + 1.99 \times 10^{+01} X_7 - 2.09 \times 10^{+01} X_8 + 1.21 \times 10^{+01} X_9 + 1.28 \times 10^{+02} X_{10} + 5.56 \times 10^{+02} X_{11} + \in \quad (5)$$

where X_1: DISTANCIA_KM; X_2: HOTELES; X_3: ZONA; X_4: CANTIDAD_FEST; X_5: TIPO_RECURSO; X_6: ACCESO_PIE; X_7: INT_CULMOCHE; X_8: ACCESO_AEROPUERTO; X_9: INTERES_TUR; X_{10}: COMENTARIOS; and X_{11}: PROMEXT.

Figures 13 and 14 show how the real and predicted data are close to the straight line; when the model is well trained, the predicted values should approach the real values. The results obtained allow for the confirmation that the developed prediction models for national and foreign visitors to tourist attractions on the Moche Route are acceptable.

In summary, integrated data from various sources were generated using the main regression algorithms (linear regression, KNN, random forest, and decision tree), finding that, due to its performance, the linear regression algorithm was the one that best predicted the number of both national and foreigners visitors, with this finding being important for the objective set out in this research, in the context of the data and variables used.

In addition, it was possible to use model evaluation metrics, such as MSE, MAE, and R2, to quantify the prediction performance of the models with promising results,

which has significant implications for decision-making in the management and planning of tourism in Peru.

5. Discussion

The present study included the development of predictive models for several tourist places that comprise the Moche Route; however, the reviewed studies differentiated in terms of the results, because the focus was on a single place of analysis or the one with the greatest influx, as in [20], which focused on Turkey. This difference in the analysis in the present investigation and the previous studies that focused on a single place could have contributed to the models not having the expected performance with high predictions.

In the study in [25], data and influence factors were used that coincided with that in the present research in addition to normalization according to their correlations, as well as the use of analysis techniques such as MAE and RMSE, which allowed for the determine of the predictive scope of the models.

Taking [27] as a reference, where a predictive model with a qualitative scope was built to identify the potential of its tourist destinations, this investigation was similar because it intended to analyze, based on official data, the various places that make up the Moche Route, managing to identify which tourist attractions are already recognized and the others that are still in the stages of improvement and enhancement.

The study in [33] applied imputation techniques to manipulate data anomalies and their processing, specifically for the pandemic stage; similarly, in this research, mean values were used to fill in the missing data.

The models developed in the present study correspond to a set of regression algorithms (linear regression, KNN, random forest, and decision tree) that are still valid today and are used as starting points in prediction studies with reduced amounts of data. They have also been used in [17] linear regression models, decision tree for the prediction of visit times and in [15] logistic regression and decision tree for the prediction of student performance, as well as in [18], where the use of linear regression models were applied to the energy performance of a building.

As for the data used for the models, both for training and testing, in [30] they were obtained from online platforms, such as search engines and tourist experience reviews. In the present research, the data obtained from official tourism channels in Peru, the data from search trends and user preferences in Google Trends, and the data from tourist experiences on TripAdvisor with respect to the Moche Route were integrated to obtain models with significant results for the prediction of visitors.

The scarcity of data continues to be a significant limitation for the accurate prediction of the number of tourists arriving at tourist sites, which is why the models used yielded moderate precision values. It should be considered that the data used were obtained from official sources; however, due to the lack of or ignorance of a standard structure for data collection at the tourist sites, only monthly consolidated visitor data are available, when it would have been more important to have slightly more data: country of origin, age, economic level, and gender, among others. Similarly, some data obtained from the Internet represent the perception of visitors to the various sites, which does not necessarily represent the same for all. All details influence the findings found in this research, so it may not be generalizable to other tourism environments.

The Moche Route represents one of the most important routes in the north of the country, which can be considered a flagship route for tourist destinations in Peru, for which it is not only necessary to have tools and technological products, such as predictive models of visitors applying the machine learning of this study, but it is also important that the decision makers in the tourism field invest in communication technologies, such as the internet service in the country, both in rural and urban areas, and to train the actors involved in the tourism sector, including travel agencies, hotel service managers, and airline crew members, to promote internal and external tourism.

The Peruvian government, through its tourism promotion agency, PROMPERU, is in charge of promoting the country's image and strategies to boost the arrival of visitors; however, it is known that the routes in the southern part of the country, where the city of Cuzco and the archaeological site of Machu Picchu are located, have always been the most promoted and interesting; however, the Moche Route is located in the north of the country, and tourism there would allow for the exploration of the Moche culture, a civilization that predates the Inca; therefore, marketing strategies and promotion of tourism for this route should be enhanced.

This study presents significant implications related to sustainable tourism in the socioeconomic context, considering that data are fundamental to modern economies and facilitate the generation of value in a more efficient, sustainable, and transparent way. The precision of the influx of visitors will provide institutions involved in the tourism sector, such as local authorities, tourism agencies, and accommodation operators, with a tool to make decision on the distribution of resources and planning of activities, thereby ensuring the efficient and sustainable management of tourism which, in turn, can generate economic and social benefits for local communities. For example, by anticipating visitor demand, tourism companies can hire additional staff from local communities during peak seasons, thereby contributing to job creation and economic development in the region.

In addition, the proper management of the influx of tourists can help minimize the negative impact on the environment and local culture, thus ensuring sustainable and equitable tourism that benefits both visitors and host communities.

6. Conclusions

This is a pioneering study on the prediction analysis of a complete route and its tourist attractions in Peru, the Moche Route, which previously did have information related to the analysis of other routes that converge in the country, contributing to the reactivation of tourism that was affected by the pandemic.

This study presents predictive models based on regression algorithms, using public data from official sources and the Internet (TripAdvisor and Google Trends), with a combination of sources being a positive effect for the proposed models. The predictive model based on the linear regression algorithm proved to be the best performing in terms of predicting visitors to the Moche Route in the coming years.

Regarding national visitors who arrive to tourist places that make up the Moche Route, the largest influx was found for the Royal Tombs of Sipan Museum in Lambayeque and the Huaca del Sol y de la Luna Archaeological Complex in La Libertad. Foreign visitors choose to visit the Huaca of the Sun and the Moon Archaeological Complex in La Libertad and the Royal Tombs of Sipan and Bruning National Archaeological Museums in Lambayeque.

The Moche Route is still not well known or has not been adequately promoted, so it is necessary that institutions related to the management of the tourism sector carry out an analysis of the tourist sites that comprise it, identifying what needs to be improved to enhance its value and potential, developing strategies and actions to promote and highlight the characteristics and attractions of the tourist site. Among these actions, the creation of tourist packages, development of activities and events, improvement of infrastructure and services, promotion via digital and traditional media, contributing to the digital transformation of this sector, mainly in the North Peru, can be considered.

It should be taken into account that the southern zone of Peru is currently more attractive for national and international tourists, because it presents tourist expressions of the Inca period; however, the Moche Route, distributed with its tourist sites in the north of the country, represents an opportunity to explore cultures that precede the Incas, the Moches, who developed a vast culture that today can be appreciated in the museums and archaeological sites of this route.

Taking into account that the Moche Route is enhanced and already has several integrated tourist sites, it represent opportunities to visit, such as the Huaca del Sol y de la Luna Archaeological Complex and the Royal Tombs of Sipan museum, which is why it

is of utmost importance that the first project of the tourism sector for the year 2023 is to promote this tourist destination with the support of the World Bank.

This research, due to its scope, covered eleven tourist sites that make up the Moche Route, finding little official and incomplete information, allowing through machine learning and the main regression algorithms to understand their behavior and interpretation of their results in predictive models, to at a later stage, look for the application of advanced methods, and even the algorithms used could serve as preprocessing for deep learning.

The data collected from official public sources do not present data collection standards or other indicators necessary for a predictive analysis; likewise internet data may not accurately represent the perception of a population of tourist visitors, which is a limitation of the study, and the findings and analysis made with such data may not be generalizable beyond the study area.

It is also important to highlight that the prediction models for the arrival of visitors to the Moche Route, built in this research, would be of direct benefit to the tourism promotion agency PROMPERU at the government level, which is responsible for promoting the country's image in the field of tourism.

The information could be used to generate the necessary strategies with the knowledge obtained from the developed tools, the whole planning of tourism promotion strategies, and the necessary budget for this type of promotion, as well as the tourism infrastructure to receive visitors; likewise, transportation companies with this information could adjust their capacity and schedules according to the expected arrival of visitors, as well as the hotel sector and travel agencies.

The data obtained from the prediction models will make it possible to know in advance the influx of visitors and carry out the necessary planning to maintain the operation of the tourism sector in the north of the country, which includes directly participating institutions, as well as the generation of employment for local communities where tourist resources are located, contributing to sustainable tourism in a socioeconomic context, helping to minimize the negative impact on the environment and local culture, benefiting visitors and communities that surround the tourist attractions.

This study creates new research opportunities, where other data sources can be incorporated, such as social networks that are currently widely used to generate interest from tourists, which enrich the training data by developing predictive models, and the analysis could be complemented with neural networks and deep learning models.

Author Contributions: Conceptualization and methodology, J.B. and C.V.; software, validation, R.A., C.V. and O.S.; writing—original draft preparation, J.B.; writing—review and editing, J.B., C.V., R.A. and O.S.; project administration, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mowforth, M.; Munt, I. *Tourism and Sustainability: Development, Globalisation and New Tourism in the Third World*, 4th ed.; Routledge: London, UK, 2016.
2. Dallen, J.T.; Gyan, P.N. *Cultural Heritage and Tourism in the Developing World A Regional Perspective*, 1st ed.; Routledge: London, UK, 2009.
3. Schönherr, S.; Eller, R.; Kallmuenzer, A.; Peters, M. Organisational learning and sustainable tourism: The enabling role of digital transformation. *J. Knowl. Manag.* **2023**, *27*, 82–100. [CrossRef]
4. Agrawal, R.; Wankhede, V.A.; Kumar, A.; Luthra, S.; Huisingh, D. Big data analytics and sustainable tourism: A comprehensive review and network based analysis for potential future research. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100122. [CrossRef]

5. Alsahafi, R.; Alzahrani, A.; Mehmood, R. Smarter Sustainable Tourism: Data-Driven Multi-Perspective Parameter Discovery for Autonomous Design and Operations. *Sustainability* **2023**, *15*, 4166. [CrossRef]
6. Russel, S.; Norving, P. *Artificial Intelligence A Modern Approach*, 3rd ed.; Pearson Educational: London, UK, 2010.
7. Asmat Uceda, D.; Vargas Yong, J.; Cortina Mendoza, R.R.; Pinillos Romero, F.; Vallejos Mendoza, A. Plan Estratégico de Marketing de Turismo Cultural Arqueológico Ruta Moche. 2017. Available online: http://hdl.handle.net/20.500.12404/7988 (accessed on 11 March 2023).
8. PERÚ EP de SESAE Ruta Moche: Conoce este Notable Circuito Turístico y sus Milenarios Tesoros Arqueológicos. Available online: https://andina.pe/agencia/noticia-ruta-moche-conoce-este-notable-circuito-turistico-y-sus-milenarios-tesoros-arqueologicos-863036.aspx (accessed on 11 March 2023).
9. Turismo. Available online: https://www.gob.pe/institucion/mincetur/tema/turismo (accessed on 13 March 2023).
10. Fernández Aguilar, L.C. Estrategias de Marketing para promover internacionalmente la "Ruta Moche" de la Región Norte del Perú como destino turístico. 2011. Available online: http://sedici.unlp.edu.ar/handle/10915/18158 (accessed on 10 March 2023).
11. Lanzan Plan de Acción de Ente Gestor de Ruta Moche. Available online: https://www.regionlambayeque.pe/web/noticia/detalle/1271?pass=Mg== (accessed on 13 May 2023).
12. DatosTurismo. Available online: http://datosturismo.mincetur.gob.pe/appdatosTurismo/Content1.html (accessed on 13 March 2023).
13. Estrategia Nacional de Reactivación del Sector Turismo 2021–2023. 2022. Available online: https://cdn.www.gob.pe/uploads/document/file/1737796/Reactivaci%C3%B3n%20del%20Turismo%202021-2023%20-%20Presentaci%C3%B3n.pdf (accessed on 10 March 2023).
14. Resolución Viceministerial N.° 004-2021-MINCETUR/VMT. Available online: https://www.gob.pe/institucion/mincetur/normas-legales/1782386-004-2021-mincetur-vmt (accessed on 13 March 2023).
15. Altabrawee, H.; Ali, O.A.J.; Ajmi, S.Q. Predicting Students' Performance Using Machine Learning Techniques. *J. Univ. Babylon Pure Appl. Sci.* **2019**, *27*, 194–205. [CrossRef]
16. Tahmasebinia, F.; Jiang, R.; Sepasgozar, S.; Wei, J.; Ding, Y.; Ma, H. Using Regression Model to Develop Green Building Energy Simulation by BIM Tools. *Sustainability* **2022**, *14*, 6262. [CrossRef]
17. Hapsari, I.; Surjandari, I.; Komarudin. Visiting time prediction using machine learning regression algorithm. In Proceedings of the 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 3–5 May 2018; pp. 495–500.
18. Kayakus, M. Estimating the Changes in the Number of Visitors on the Websites of the Tourism Agencies in the COVID-19 Process by Machine Learning Methods. *Sosyoekonom* **2022**, *30*, 11–26. [CrossRef]
19. Birim, S.; Kazancoglu, I.; Mangla, S.K.; Kahraman, A.; Kazancoglu, Y. The derived demand for advertising expenses and implications on sustainability: A comparative study using deep learning and traditional machine learning methods. *Ann. Oper. Res.* **2022**, *2022*, 1–31. [CrossRef] [PubMed]
20. Tutsoy, O.; Tanrikulu, C. A Machine Learning-Based 10 Years Ahead Prediction of Departing Foreign Visitors by Reasons: A Case on Turkiye. *Appl. Sci.* **2022**, *12*, 11163. [CrossRef]
21. Laaroussi, H.; Guerouate, F.; Sbihi, M. A novel hybrid deep learning approach for tourism demand forecasting. *Int. J. Electr. Comput. Eng.* **2023**, *13*, 1989–1996. [CrossRef]
22. Yao, Y.; Cao, Y.; Ding, X.; Zhai, J.; Liu, J.; Luo, Y.; Ma, S.; Zou, K. A paired neural network model for tourist arrival forecasting. *Expert Syst. Appl.* **2018**, *114*, 588–614. [CrossRef]
23. Andariesta, D.T.; Wasesa, M. Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: A multisource Internet data approach. *J. Tour. Futures* **2022**, *2022*, 1–17. [CrossRef]
24. Alghamdi, A. A Hybrid Method for Big Data Analysis Using Fuzzy Clustering, Feature Selection and Adaptive Neuro-Fuzzy Inferences System Techniques: Case of Mecca and Medina Hotels in Saudi Arabia. *Arab. J. Sci. Eng.* **2023**, *48*, 1693–1714. [CrossRef]
25. Yu, N.; Chen, J. Design of Machine Learning Algorithm for Tourism Demand Prediction. *Comput. Math Method Med.* **2022**, *2022*, 6352381. [CrossRef] [PubMed]
26. Hoffmann, F.J.; Braesemann, F.; Teubner, T. Measuring sustainable tourism with online platform data. *EPJ Data Sci.* **2022**, *11*, 41. [CrossRef]
27. Ruiz Palacios, M.A.; Pereira Texeira de Oliveira, C.; Serrano Gonzalez, J.; Saenz Flores, S. Analysis of Tourist Systems Predictive Models Applied to Growing Sun and Beach Tourist Destination. *Sustainability* **2021**, *13*, 785. [CrossRef]
28. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*, 1st ed.; Cambridge University Press: Cambridge, UK, 2014.
29. Grus, J. *Data Science from Scratch*, 2nd ed.; O'Reilly: Springfield, MO, USA, 2019.
30. Li, H.; Hu, M.; Li, G. Forecasting tourism demand with multisource big data. *Ann. Tour. Res.* **2020**, *83*, 102912. [CrossRef]
31. He, K.; Wu, D.; Zou, Y. Tourist Arrival Forecasting Using Multiscale Mode Learning Model. *Mathematics* **2022**, *10*, 2999. [CrossRef]
32. Jiang, P.; Hu, Y.-C. Constructing interval models using neural networks with non-additive combinations of grey prediction models in tourism demand. *Grey Syst.* **2023**, *13*, 58–77. [CrossRef]
33. Ashok, S.; Aravind, K. Impact of Covid-19 on Demand Planning: Building Resilient Forecasting Models. In Proceedings of the 2021 The 5th International Conference on Compute and Data Analysis, Sanya, China, 2–4 February 2021; pp. 59–66.

34. Khan, N.U.; Wan, W.; Riaz, R.; Jiang, S.; Wang, X. Prediction and Classification of User Activities Using Machine Learning Models from Location-Based Social Network Data. *Appl. Sci.* **2023**, *13*, 3517. [CrossRef]
35. Sistema de Inteligencia Turística. Available online: https://www.mincetur.gob.pe/centro_de_Informacion/mapa_interactivo/index.html (accessed on 19 March 2023).
36. Fy, H.; Da, B.; Md, L. A simple method of sample size calculation for linear and logistic regression. *Stat. Med.* **1998**, *17*, 1623–1634.
37. Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Cheng, D. Learning k for kNN Classification. *ACM Trans. Intell. Syst. Technol.* **2017**, *8*, 119. Available online: https://dl.acm.org/doi/abs/10.1145/2990508 (accessed on 17 March 2023).
38. Rigatti, S.J. Random Forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning: Methods and Applications*; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, USA, 2012; pp. 157–175.