

Multi-dimensional Analysis and Prediction Model for Tourist Satisfaction

Deepanjal Shrestha¹, Tan Wenan^{1,2}, Bijay Gaudel¹, Neesha Rajkarnikar³
and Seung Ryul Jeong^{4*}

¹School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Nanjing, China

²School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, China
[e-mail: deepanjal@hotmail.com, wtan@foxmail.com, gaudelbijay10@gmail.com]

³School of Energy Science and Engineering, Nanjing Tech University
Nanjing, China

[e-mail: neesha_rk@hotmail.com]

⁴Graduate School of Business IT, Kookmin University
Seoul, South Korea

[E-mail: srjeong@kookmin.ac.kr]

*Corresponding author: Seung Ryul Jeong

*Received November 28, 2021; revised January 2, 2022; accepted February 2, 2022;
published February 28, 2022*

Abstract

This work assesses the degree of satisfaction tourists receive as final recipients in a tourism destination based on the fact that satisfied tourists can make a significant contribution to the growth and continuous improvement of a tourism business. The work considers Pokhara, the tourism capital of Nepal as a prefecture of study. A stratified sampling methodology with open-ended survey questions is used as a primary source of data for a sample size of 1019 for both international and domestic tourists. The data collected through a survey is processed using a data mining tool to perform multi-dimensional analysis to discover information patterns and visualize clusters. Further, supervised machine learning algorithms, kNN, Decision tree, Support vector machine, Random forest, Neural network, Naïve Bayes, and Gradient boost are used to develop models for training and prediction purposes for the survey data. To find the best model for prediction purposes, different performance matrices are used to evaluate a model for performance, accuracy, and robustness. The best model is used in constructing a learning-enabled model for predicting tourists as satisfied, neutral, and unsatisfied visitors. This work is very important for tourism business personnel, government agencies, and tourism stakeholders to find information on tourist satisfaction and factors that influence it. Though this work was carried out for Pokhara city of Nepal, the study is equally relevant to any other tourism destination of similar nature.

Keywords: Tourism satisfaction, multi-dimensional analysis, machine learning, Pokhara, Nepal.

The paper is supported in part by the National Natural Science Foundation of China under Grant (No. 61672022 and No. U1904186), and the Collaborative Innovation Platform of Electronic Information Master under Grant No. A10GY21F015 of Shanghai Polytechnic University

1. Introduction

Tourism is the second largest industry of Nepal and one of the biggest contributors to the GDP and creation of jobs in the country. In 2019, travel and tourism had a contribution of USD 9170 billion with 34 million jobs (1 in 4 net new jobs) [1]. The report published by World Travel and Tourism Council has forecasted that tourism will be the fastest growing industry after the post COVID period [1]. Therefore, Nepal has an active and important role to play to put itself as a top tourism destination on the tourism world map. Tourism is a competitive business industry that has competition at the country level and as well as destination level. In this context, governments, public organizations, and private organizations work hand in hand to make the tourism industry more pleasing, attractive, and satisfactory. Satisfaction is a very important aspect of the tourism business as it measures the quality of tourism products and services. Satisfaction in the tourism industry is a high priority area and a lot of research is conducted time and again around it [2]. Satisfaction and tourism go hand in hand, as a satisfied tourist is a viable marketing agent who contributes positively to the construction of business value and brand image of a destination. The economies that rely heavily on the tourism industry take a lot of care to see that tourist visiting their places are returned happy and satisfied. Measurement of tourist satisfaction is a complex job and depends on many aspects of the tourism business [3]. Reviews, feedbacks, ratings, closed surveys, and suggestion forms are some of the strategies that are used by business houses to measure and strengthen satisfaction. Tourism satisfaction is multidimensional data and its analysis requires a complete understanding and multidimensional view of it [3]. Normal statistical models may not fully comply with the analysis of such data and may not fully predict the satisfaction value completely. A data mining or machine learning model can make better analysis, visualization, and prediction of such data. This work attempts to study tourist survey data of Pokhara city of Nepal with a multi-dimension perspective using data mining and machine learning models to discover information patterns, visualize them and predict the satisfaction value of tourists.

2. Literature Review

The hospitality industry takes customer satisfaction as an absolutely important aspect and prioritizes it as the highest attribute of service quality. This aspect is responsible to establish a strong link between customers' loyalties and improve the financial performance of a business [4]. Satisfaction can be understood as a combination of qualities of tourism products and services that are combined with positive experiences of the tourist [5]. Satisfaction is a function of user expectations before the purchase and the anticipation after using it [6]. A positive feeling of using a product or service increases the satisfaction quotient of a tourist which in return helps them to form a positive outlook for it and become loyal customers [5]. The measure of satisfaction of tourism products and services has become an important parameter with the advent of ICT and internet technologies. Reviews, feedback, ratings, and recommendations of a product and service are directly connected to tourist level of satisfaction [7]. A satisfied tourist will provide positive reviews, feedback, and ratings for a product and service and vice versa in case they are not satisfied [7]. Research around tourism and satisfaction has always attracted scholars and they have studied different aspects of tourism products and services about satisfaction [6]. The earlier tourism studies were focused on benchmarking of micro-medium tourism sector like hotels, destinations, transportation access

in combination with tourist needs and expectations [5][6]. Later, the works related to tourist satisfaction gained importance, and studies were conducted in the area of the image of tourist destinations [5][6], tourists destination choice attributes, expectations and reality [6], tourists expectation in urban destinations [5] tourist attractions and events [6]. Research on tourism satisfaction gained more importance with the development of the tourism industry and internet technologies. Authors carried out research using mixed dimensions like information and communication technology, machine learning systems, e-commerce, and other similar artifacts [9]. The study of multi-objective tourism product optimization design model and tourist satisfaction [7], the relationship between user satisfaction and tourism e-commerce function [8], The role of big data technology architecture for tourist management and tourist satisfaction [9] elaborated the role of technology in tourism. Scholars have also studied tourist satisfaction in tourism destinations based on variable precision, geotagged data, web, and big data mining of online text as other important aspects [10]. Some other notable works on tourist satisfaction include a factor-cluster segmentation approach [11], multi-criteria-based studies, destinations as a function of attribute importance, performance, and travel motivation [12], and constructs of satisfaction and dissatisfaction of tourist satisfaction attributes [13]. Similarly, work is also seen in the area of evaluation of destination satisfaction based on fuzzy multi-data decision models [14], the relationship between facilities in tourism destinations, positive impact, satisfaction, and re-visit intention [15]. Tourism satisfaction gained a primary place in the later research with the application of the Internet and WWW [16]. Scholars used technological dimensions to study tourism and various aspects related to it [16]. Some recent studies in the context of ICT and tourism include big data visualization in tourism [17], the relationship between reviews and tourist satisfaction [18], social data analysis for tourism destinations and activities, tourist demands, and the role of technology in tourism [16]. Studies are seen in the area of online tourism information and tourist behavior [19], Tourism 4.0 technologies and tourist experiences [20], Influence of IoT in Tourism Industry [21], measuring satisfaction through factor analysis and multiple regression [22], studies using multi-criteria and data mining, big data analysis in tourism [23], new model developments for user satisfaction, and recommendations [24] [25]. Studies related to tourism in Nepal were also prevalent and research shows that tourist satisfaction was a common subject with many scholars working on this aspect. Satisfaction studies have covered topics related to hotels [26], tourist behavior and satisfaction [27], tourist satisfaction and revisit intention [28], visitor perceptions of the world heritage value of Mt. Everest [29], community tourism, cultural and religious tourism [29]. Besides numerous studies on tourism and satisfaction, the Nepalese literature lacked studies in the area of application of technology and tourism satisfaction. The studies that were traced mainly talked about technology applications in tourism with digital tourism security systems [30], Religious recommender systems [31], Sustainable tourism [32] Tourism ecosystem, and Tourism and e-commerce [33]. Studies were also discovered related to ICT implications in tourism, geotags and tourism points of interest, e-tourism, Nepal digital tourism framework, and mobile technology in tourism [34]. The literature review provided a good understanding of tourist satisfaction and its related aspects. It was seen that satisfaction has become an important area of research and studies are carried out for existing tourism business systems through the use of technology [7]. Business houses nowadays are applying technology to develop systems that can capture data on a real-time basis, work with reviews, and post of users and improve their business to provide better service and increase tourist satisfaction [15].

3. Research Framework and Methodology

Data collection and developing a good data source is a challenging task as a poor model can yield bad results. In this study, survey questionnaires were designed carefully based on the literature. The designed questionnaire consists of 96 variables which included 95 features revolved around tourist satisfaction attributes. The processed data set was analyzed using the statistical and data mining tools and visualized as Scatter plots, PCA, and linear projections. The data set was further analyzed with seven machine learning models to validate the best prediction model. The research framework of the study is depicted in Fig. 1.

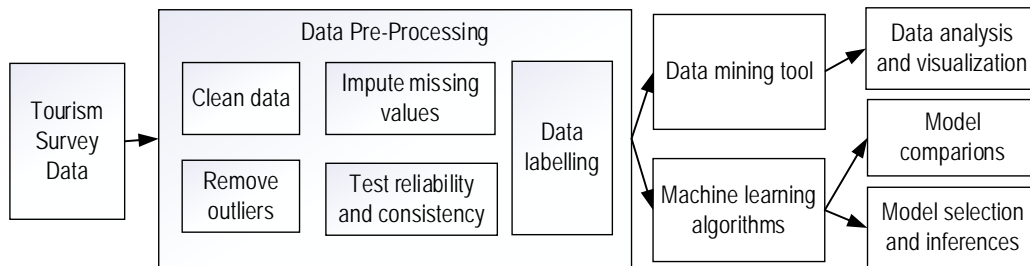


Fig. 1. Research Framework of the Study

3.1 Research Domain

The study is carried out in Pokhara city of Nepal (Tourism capital) and both international and domestic tourists, were chosen as respondents. The study did not discriminate based on sex, marital status, and income and treated all equally. The objectives of the study include:

1. Analyze the survey data based on a multi-dimensional approach to understand the tourist pattern for tourism destinations.
2. Visualize data to discover information clusters for understanding tourist satisfaction.
3. Apply machine learning models to survey data for learning and prediction purposes.
4. Evaluate and identify the best machine learning algorithms for the classification and prediction of tourist satisfaction for the tourist survey data and draw a conclusion.

3.2 Sampling

Sampling is very important to acquire proper data and in this work, a stratified random sampling method was used as it represents the entire population in the best way and yields good statistical test results. This method is good for medium to smaller data sample sizes that provide better accuracy results and is convenient to visualize data and train models. This methodology is also best known for model selection in the case of machine learning and data mining of real-world data sets that are imbalanced and has many attributes to consider.

3.3 Data Acquisition

The data was acquired through a survey questionnaire that was distributed to 1500 respondents through the different mediums in Pokhara city of Nepal. 1019 responses were finally considered after preprocessing the data. The survey question was designed to capture the satisfaction parameters of the tourist combined with other parameters like demography, product choices, travel information, travel motives and behaviors, specific activities, personal travel satisfaction parameters, and destination satisfaction parameters. These seven broad categories of data were further subdivided into specific categories to capture the specific

details of tourists to finally come up with satisfaction value. The tourist survey data consisted of 1019 instances, 96 variables which included 95 features (52 categorical, 43 numeric) and had 0.0% missing values.

3.4 Data Pre-processing

The real-world data are always faced with some intrinsic issues like noise, incomplete values, inconsistency, and missing values. Besides these problems, there are problems from the respondent side, where he/she can intentionally fill in wrong data or may provide unrealistic estimates. Mistakes in data can also occur during data entry and data merging from various sources. The quality of data is a very important aspect for any data-dependent study to come up with good results and inferences. The role of good quality of data becomes mandatory for machine learning and data mining systems, also. In this study, the survey data was pre-processed using data integration, cleansing, transformation, and reduction techniques. In the initial selection process, the data that has no significant contribution to the study was dropped. The data relating to email-id, timestamp, agreement statement, and user-id were dropped. In the next step, data cleaning was applied, by removing data with offending cases and variables with excessive levels of mismatch. The set of data that were discovered to miss accidentally or at random were corrected using the imputation method. The data were cleaned for outlier and extreme values using an outlier algorithm as shown in [Algorithm 1](#).

Algorithm 1. Outlier detection

```

1   Input: dataset
2   Output: dataset after removal of outlier
3   For i = 1 to len(dataset):
4       For j = 1 to number of variables
5           If (is CategoricalVariable)
6               X = range of Variable
7               If dataset [i] [j] not in X :
8                   dataset = dataset – dataset [i][j]
9           End
10      End
11  end
12  end
13  return dataset

```

Besides this, manual inspection, the data was checked for unmatched, extreme values, and ambiguous data. The data is prepared for two purposes, one is to do a multi-dimensional analysis of data to discover tourist patterns and the second is to apply machine learning models to train and test data for predictions. To fulfill the machine learning objective supervised machine learning algorithms are applied for three classes of satisfied, unsatisfied, and neutral user's labels. The class labels were decided based on the user's answer's to the overall satisfaction value in terms of satisfied, unsatisfied, or neutral users.

3.5 Statistical Description and Reliability Testing

A Cronbach's alpha (α) test is carried out to test the internal consistency of the data under consideration. The analytical result in [Table 1](#), indicates that Cronbach's alpha (α) value is greater than the threshold of 0.7. Thus, the internal consistency and reliability of the constructs are confirmed. Further, the statistical analysis of the demography data shows that standard

deviation values are less than 1 for tourist type, gender, and age group while it is greater than 1 for marital status, monthly income, academic qualification, and profession. Similarly, the variance is also seen as less than 1 for tourist type, gender, and age group and greater than 1 for marital status, monthly income, academic qualification, and profession as shown in **Table 2**. These values confirm that data is scattered closely in the above data set.

Table 1. Representing internal consistency of the data

Cronbach's Alpha	N of Items
.834	110

Table 2. Demographic data of the respondents

	Tourist Type	Gender	Marital Status	Age Group	Monthly Income USD.	Academic Qualification	Profession
N Valid	1019	1019	1019	1019	1019	1019	1019
Missing	0	0	0	0	0	0	0
Std. Deviation	.482	.490	1.023	.940	2.803	2.497	2.121
Variance	.233	.240	1.046	.883	7.859	6.236	4.499
Range	1	2	3	5	9	7	6
Minimum	1	1	1	1	1	1	1
Maximum	2	3	4	6	10	8	7

4. Multi-Dimensional Data Analysis and Visualization

The multidimensional analysis of the tourism survey data was carried out using data mining and the visual programming tool Orange 3.30. A combination of data attributes was considered to visualize and interpret data and discover information clusters, using scatter plots and PCA. **Fig. 2** depicts a scatter plot for demographic data which shows four distinct and three nominal clusters based on tourist type, gender, and marital status. International male tourists with married status were more frequent visitors than female international tourists whereas, for domestic tourists, male and female tourists number was equal for married and unmarried status.

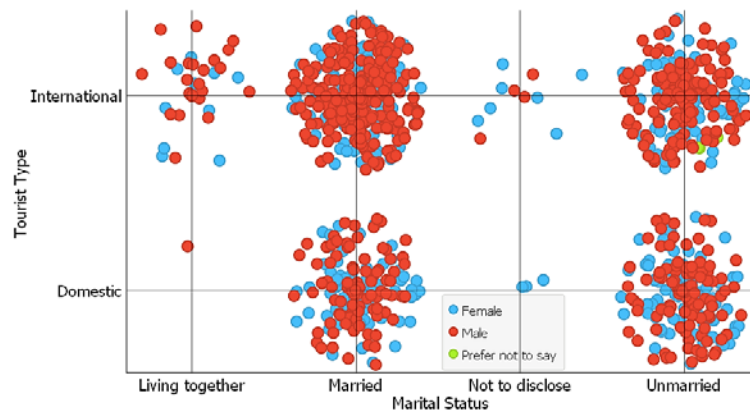


Fig. 2. Visualizing data cluster of a tourist type, gender, and marital status

The scatter plot for age group and the number of people accompanying together for both tourists type generates 11 distinct and dense clusters with 5 nominal clusters and some sparse data with one or two data instances scattered in the plot. Fig. 3 depicts that domestic tourists have dense clusters marked in 20-30 and 30-40 age groups and people accompanying are distinctly in 3 to 5 group, undecided and more than 6 people category. Similarly, for international tourists 20-30, 30-40 and 40-50 age groups make dense clusters with 1 to 2, 3, undecided, and more than 6 people category.

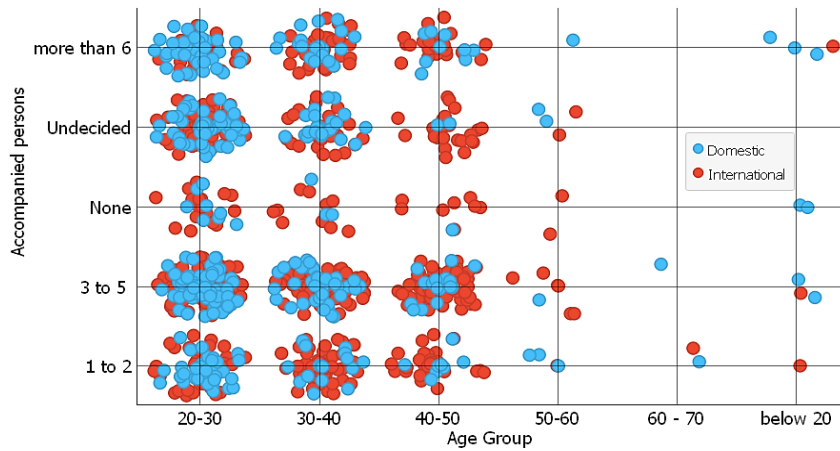


Fig. 3. Visualizing data cluster for an accompanied person, tourist type, and age group

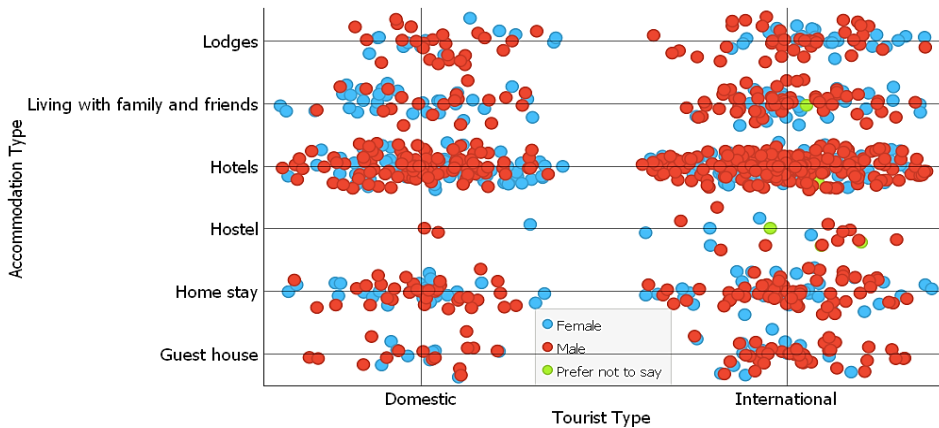


Fig. 4. Visualizing data cluster for accommodation preference, tourist type, and gender

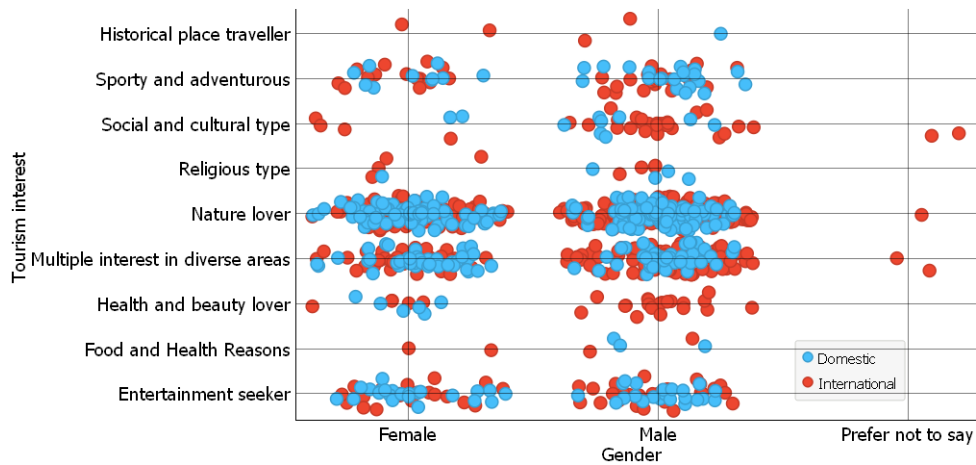


Fig. 5. Visualizing data cluster for tourism interest, tourist type, and gender

The analysis of tourist type for accommodation choice and gender shows that for domestic tourists (male and female) hotel is the first preferred category followed by living with family and friends (for females) followed by homestay as third (males and females). The plot for international tourist shows that hotels are the first preferred accommodation (male and female) followed by homestay (male and female) and living with family and friends (mostly males) as the third choice as shown in Fig. 4. Tourism interest, tourism motive, and tourist planning factors are important attributes to check tourist satisfaction, besides the demographical dimensions. It is seen from Fig. 5 that domestic tourists have distinct tourism interest clusters for nature, multiple interests in diverse areas, and entertainment (both males and females). The international tourist interest clusters are also seen distinctively for nature, multiple interests in diverse areas, and entertainment (both males and females).

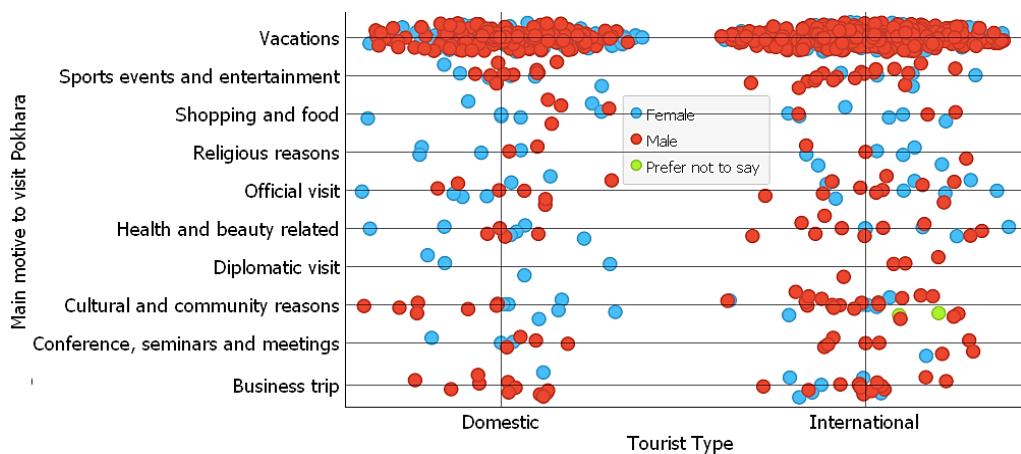


Fig. 6. Visualizing data cluster for motive to visit, tourist type, and gender

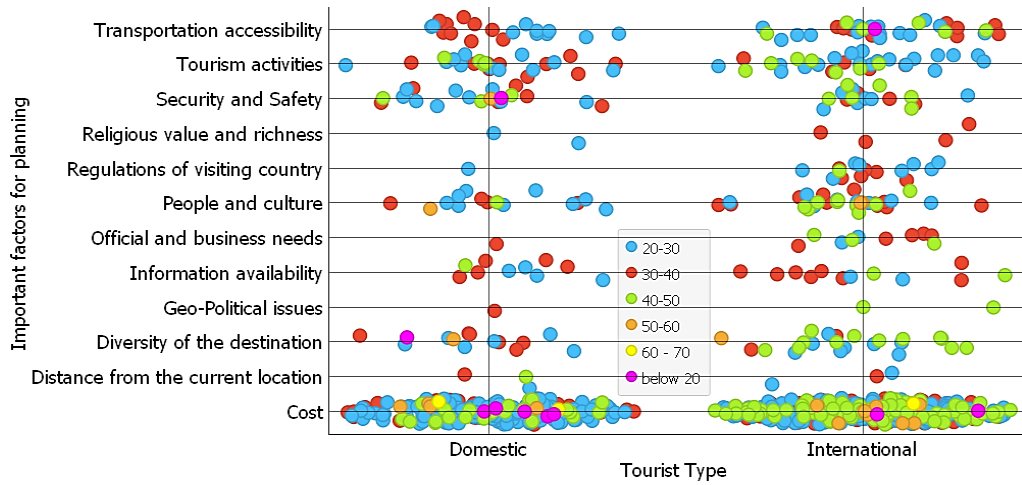


Fig. 7. Visualizing data cluster for planning factors, tourist type, and age group

The analysis of motive to visit Pokhara depicts that vacation is the most prominent cluster (both males and females), followed by sports events and entertainment, and cultural and community as small clusters for both domestic and international tourists as shown in Fig. 6. Planning factors were analyzed concerning age group and it is seen that cost is a very important factor for both genders, all age groups, and tourist types. The international tourist was also concerned about the diversity of destination, people and culture, transportation, and tourism activities. The domestic tourist was also concerned about transportation, tourism activities along with safety and security issues as seen in Fig. 7. The satisfaction attributes were finally visualized using linear projection through circular placement. It is seen that the cluster is dense at the center for satisfied tourists depicting that all 18 attributes measured for satisfaction met the desired expectation. The unsatisfied and neutral tourist clusters were seen scattered for safety, health, and hygiene, tourism service, QoS, and people attitude as shown in Fig. 8.

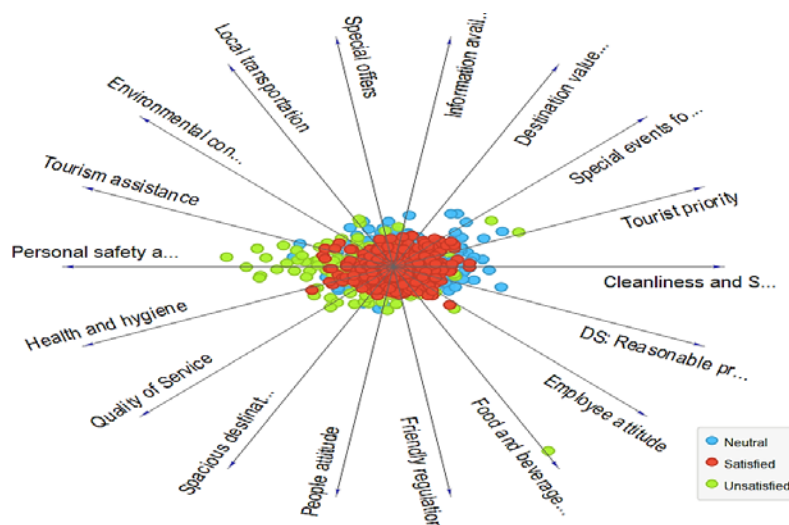


Fig. 8. Visualizing data clusters and satisfaction attributes through circular placement

5. Model Design and Experiment

5.1 The Classification and Prediction Algorithms

The work considers seven algorithms, kNN, Decision tree, Support vector machine, Random forest, Neural network, Naïve Bayes, and Gradient boost for the classification and prediction purpose. The basic theory and initial environment for the experiment are discussed below.

5.1.1 kNN

The first algorithm considered is KNN, as it is a simple, efficient, and supervised machine learning algorithm that works by calculating the distance between the data points and finding the nearest data points (neighbor) [35]. This model is non-parametric and considers entire instances of the training data set to predict the output for the test data or unseen data. kNN model relies on the selection of the value of K for training and test data, as a low value of K in training leads to overfitting and high value leads to an underfitting problem. The generalized equation for calculating the distance is shown in equation (1).

$$D = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \Rightarrow D = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (1)$$

where a, b are the data points and D is the distance between them. The first part of the equation provides an expanded form with $a_1, b_1, a_2, b_2, \dots, a_n, b_n$ representing different dimensions of the data. The second part provides a generalized equation for the distance between (a, b) with different dimensions from i to n . For this study, kNN was set up with the number of neighbors as 6, Euclidean metric, and distance as weights. Distance is chosen as a parameter because the closer the neighbors of a query point are and the greater is the influence of that neighbor.

5.1.2 Decision Tree

A Decision tree is a simple, yet powerful supervised machine learning algorithm used in many applications for classification and prediction problems. It is a tree-like structure that uses simple flowchart notations to depict predictions that are a result of splits based on some features [35]. The decision tree depends on three important parts to decide a split which includes, Information gain, Entropy, and Gain as shown in the equation (2), (3), and (4) below.

$$\text{Information Gain: } I(P, n) = \frac{-P}{P+n} \log_2 \left(\frac{P}{P+n} \right) - \frac{-n}{P+n} \log_2 \left(\frac{n}{P+n} \right) \quad (2)$$

$$\text{Entropy: } E(A) = \sum_{i=1}^v \frac{P_i + n_i}{P+n} (I(P_i, n_i)) \quad (3)$$

$$\text{Gain: } \text{Gain}(A) = I(P, n) - E(A) \quad (4)$$

The experiment carried out with DT was initially set up for 3 as the minimum number of instances in leaves, with split subsets as 5, maximal tree depth to 100, and making the classification stop when the majority reaches a threshold of 95%.

5.1.3 Support Vector Machine (SVM)

SVM is one of the most popular supervised machine learning algorithms that work by calculating a hyperplane that best divides a dataset into classes [35]. It is considered a good choice for medium and small data sizes with n number of dimensions and works well for both classification and regression problems. The equation (5) represents hypothesis function h that divides the class into $+1$ and -1 , where the point above hyperplane is categorized as $+1$ class and the point below hyperplane is classified as -1 class. A generalized equation of the model is shown in equation (6).

$$h(x_i) \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases} \quad (5)$$

$$\left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(w \cdot x_i - b)) \right] + \lambda \|w\|_2 \quad (6)$$

SVM is a powerful model which was initially set with Cost (C) as 1.0, regression loss epsilon (ϵ) 0.10, with the Radial Basis Function (RBF) kernel, where two parameters C and gamma are implemented from GridSearchCV that uses a “fit” and a “score” method for the experiment.

5.1.4 Random Forest

Random forest is a powerful supervised machine learning algorithm that is considered for this work. In the classification setting, the prediction of the random forest is the most dominant class among predictions by individual decision trees [35]. If there are T trees in the forest, then the number of votes received by a class m is calculated based on equation (7).

$$v_m \sum_{t=1}^T I(\hat{y}_t == m) \quad (7)$$

where \hat{y}_t is the prediction of the *t*-th tree on a particular instance. The indicator function $I(\hat{y}_t == m)$ takes on the value 1 if the condition is met, else it is zero. Given these votes, the final prediction of the algorithm is the class with the most votes. In the regression setting, the prediction of the random forest is the average of the predictions made by the individual trees. If there are T trees in the forest, each making a prediction \hat{y}_t , the final prediction \hat{y} as in equation (8)

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T \hat{y}_t \text{ (Regression)} \quad (8)$$

Considering the basic properties of the random forest algorithm, the basic setup for the experiment is initialized with the number of trees as 7, the number of attributes at each split as 5, with replicable training and limiting the depth of individual trees up to 25.

5.1.5 Neural Network

A neural network is a powerful supervised algorithm that is inspired by human neurons and can learn by examples [35]. The mathematical equation of the model can be represented as (Y) the summation of inputs multiplied with weights and a bias value that is added to the total value as shown in equation (9). Inputs in this case are the representation of neurons.

$$Y = \sum(\text{Inputs} * \text{Weights}) + \text{bias} \quad (9)$$

Assuming the basic working mechanism of the neural network, the experiment environment for this study is set up with neurons per hidden layer as 100 having rectified linear unit function (ReLU) and Adam as the stochastic gradient-based optimizer with max iterations to 200. The other parameters for this model are set to sklearn’s default.

5.1.6 Naive Bayes

Naive Bayes is a powerful extensible algorithm that is very fast and works for multiclass classification data. Naive Bayes is a powerful predictive algorithm that works on the principle of the Bayesian theorem [35]. Due to its powerful features, this model is considered for this work. The basic mathematical model for this algorithm is explained in equation (10).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (10)$$

The Naive Bayes model parameters are initially set to sklearn’s defaults for the tourism survey data set and the reading of the experiment is noted for this work.

5.1.7 Gradient Boost

Gradient boosting is a powerful supervised machine learning algorithm that uses the ensemble technique of weak prediction models as the basic principle for classification and regression. The algorithms can be customized to meet the needs of a particular problem and are hence suitable for this work. In equation (11) we can see that the final output of the algorithm is the aggregation of the output of the base model with the learning rate and residual model until minimum residual error is achieved [35].

$$\text{Final Output} = \text{O/P of Base model} + \eta\text{RM1} + \eta\text{RM2} + \eta\text{RM3} + \dots + \eta\text{RMn} \quad (11)$$

Considering the customization principle of gradient boost the initial environment was set up to include the number of trees initialized to 100, learning rate for each tree was specified as 0.100, with replicable training to replicate test results, the depth of individual tree was limited to 3, with subset splitting limited to 2 and fraction of training instances were defined as 100.

5.2 Measurement

5.2.1 Accuracy, Precision, and Recall

The classifier accuracy is very important to evaluate a particular algorithm as it represents the overall correctness of the model [35]. Accuracy is calculated based on the sum of true positive and true negative cases divided by the total number of cases as shown in equation (12) below:

$$\text{Accuracy} = \frac{|TP|+|TN|}{|FN|+|FP|+|TN|+|TP|} \quad (12)$$

Also, the performance of a classifier can be expressed as a misclassification error rate using the equation (13) below:

$$\text{Error Rate} = \frac{|FN|+|FP|}{|FN|+|FP|+|TN|+|TP|} \quad (13)$$

Precision and recall measure the performance of the classifier. The precision measures how many selected items are relevant (True Positive divided by False Positive and True Positive values), and recall measures how many relevant items are selected (True Positive divided by False Negative and True Positive values). The equation (14) and (15) represent both as:

$$\text{Precision} = \frac{|TP|}{|FP|+|TP|} \quad (14)$$

$$\text{Recall} = \frac{|TP|}{|FN|+|TP|} \quad (15)$$

5.2.2 F-score

F-Score is another measure used in this study which is the test of accuracy and is calculated based on Precision and Recall. F-Score is also known as F-Measure and is an improvement in accuracy as it takes class discrimination into account. 1 represents the highest value of F-Score and 0 represents the lowest value. It can be calculated as shown in equation (16).

$$Fscore = 2 \times \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \quad (16)$$

5.3 Experimental Design

The work is carried out using basic tools that include Python 3.6 with scikit-learn libraries, MS-Excel for preprocessing initial data, and orange as the data mining tool. The environment consisted of Windows 10 home user edition with a 64-bit operating system and had all the required software libraries installed. The hardware consists of 11th Gen Intel(R) Core(TM) i5-1135G7, 2.42 GHz processor with 8.00 GB of RAM. The environment for the model experiment was set up for the best performance of the machine learning algorithms. In some cases where the choice for the variable assignment was difficult to choose, scikit-learn defaults were considered. Data for the experiment was divided into a 70:30 ratio and a cross-validation method with stratified sampling was used for 10, and 20 folds. The target variable is chosen as average over classes, which returns scores that are weighted averages of the overall classes.

6. Result Analysis and Performance Evaluation

6.1 Learning Model Analysis

Table 3. Testing result of classification algorithms for training data using stratified sampling with a 10-fold cross-validation method

Model	AUC	CA	F1-Score	Precision	Recall
kNN	0.97190	0.87675	0.87474	0.87793	0.87675
Tree	0.89384	0.84874	0.84720	0.84641	0.84874
SVM	0.97309	0.88235	0.88264	0.88516	0.88235
Random Forest	0.94543	0.85294	0.85367	0.85678	0.85294
Neural Network	0.96878	0.87115	0.87019	0.86974	0.87115
Naive Bayes	0.98106	0.89636	0.89846	0.90616	0.89636
Gradient Boosting	0.99291	0.93697	0.93726	0.93776	0.93697

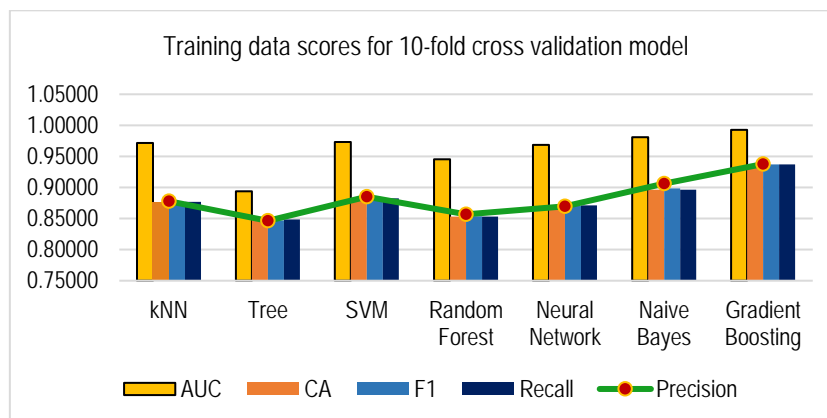


Fig. 9. Representing training data performance for 10 cross-validation model

The learning algorithms are executed to test the models with the highest accuracy and lowest error rate by assessing their performance and selecting the best model. The different measures like AUC, Classification accuracy, F1-Score, Precision, Recall, and Specificity are used to evaluate the performance of the model. To make an extensive comparison of the models, they are executed with different parameters, and results are analyzed in extensive detail. **Table 3** shows the result of 7 models for a 10 fold cross-validation method using stratified sampling. It can be seen from **Table 3** that Gradient boost has the best performance as a training model with AUC .9934035, CA .952382, F1-Score 0.9525812 Precision 0.953042, and recall 0.952381 followed by Naïve Bayes as second best model and kNN as the third-best model. A graphical representation of all seven models is also shown in **Fig. 9**.

Table 4. Testing result of classification algorithms for training data using stratified sampling with a 20-fold cross-validation method

Model	AUC	CA	F1-Score	Precision	Recall
kNN	0.9734135	0.8753501	0.8732781	0.8764873	0.8753501
Tree	0.9080911	0.8613445	0.859638	0.8592149	0.8613445
SVM	0.9728764	0.8809524	0.8811621	0.8831505	0.8809524
Random Forest	0.9408542	0.8305322	0.82976	0.8324706	0.8305322
Neural Network	0.9673604	0.8683473	0.8675698	0.8671152	0.8683473
Naive Bayes	0.9807848	0.894958	0.8970731	0.9051694	0.894958
Gradient Boosting	0.9934035	0.952381	0.9525812	0.953042	0.952381

To validate the performance of the model further, all seven models were executed with 20 fold cross-validation. It was seen that as the K-fold parameter was increased, models performed better. The Gradient boost improved its performance with an increase in AUC by 0.0011135, CA by 0.012601, F1-Score by 0.0126612, Precision by 0.012692, and Recall by 0.012601 and performed the best among all the seven algorithms, followed by Naïve Bayes and SVM as the third-best model as shown in **Table 4**. In the 20 fold cross-validation setup, SVM performed better than kNN as the third-best model. A visual representation of the performance of 7 models using 20-fold cross-validation is shown in **Fig. 10**.

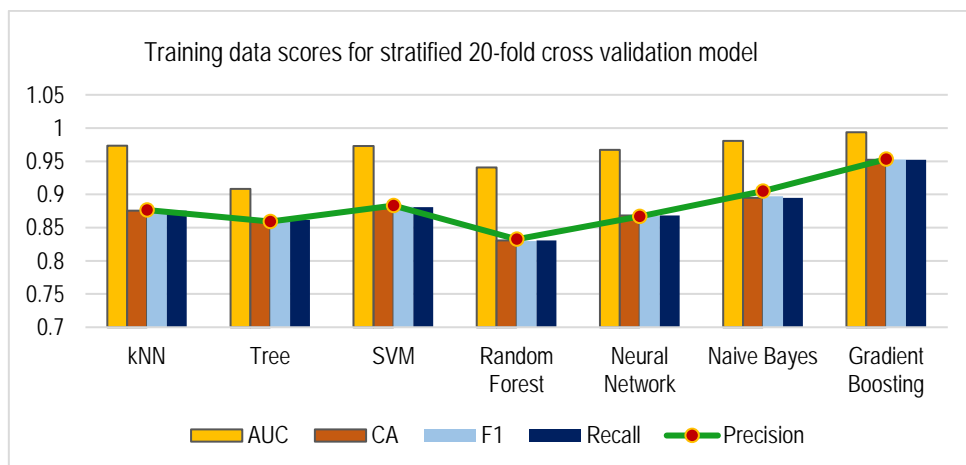


Fig. 10. Representing training data performance for 20 cross-validation model

6.2 Prediction Model Analysis

The prediction analysis of the seven models was done with the remaining 30 percent of the data that consisted of 305 instances, 95 variables, and 94 features. AUC, CA, F1-Score, Precision, Recall, and Specificity was used to analyze the best prediction algorithm.

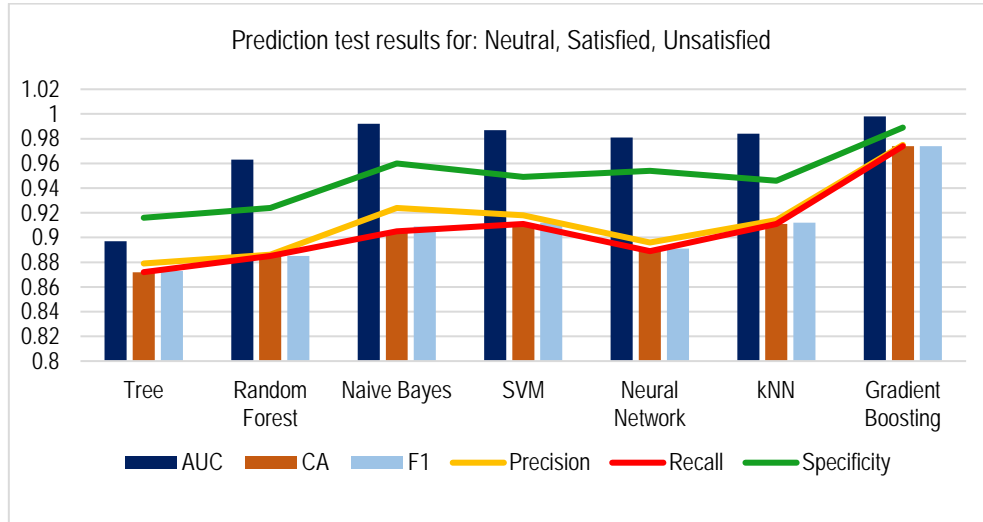


Fig. 11. Representing prediction test results for Neutral, Satisfied, Unsatisfied classes

The test results show that six models achieve an accuracy of more than 90% while only one model had an accuracy below 90%. It can be seen from [Table 5](#), that Gradient boost has the best prediction accuracy with AUC = 0.998, CA = 0.974, F1-Score = 0.974, Precision = 0.975, Recall = 0.974, Specificity = 0.989, followed by Naïve Bayes and SVM. kNN also shows good prediction results and performs better than Naïve Bayes for classification accuracy and F1-Score whereas Naïve Bayes has better overall performance. The SVM and kNN perform equally, except for AUC, where SVM is better. [Fig. 11](#) shows the performance of all seven algorithms graphically with precision, recall, and specificity shown as lines over the bars.

Table 5. Testing result of prediction algorithms for 30% testing data set

Model	AUC	CA	F1-Score	Precision	Recall	Specificity
Tree	0.897	0.872	0.873	0.879	0.872	0.916
Random Forest	0.963	0.885	0.885	0.886	0.885	0.924
Naive Bayes	0.992	0.905	0.909	0.924	0.905	0.960
SVM	0.987	0.911	0.912	0.918	0.911	0.949
Neural Network	0.981	0.889	0.891	0.896	0.889	0.954
kNN	0.984	0.911	0.912	0.914	0.911	0.946
Gradient Boosting	0.998	0.974	0.974	0.975	0.974	0.989

6.2.1 ROC Analysis

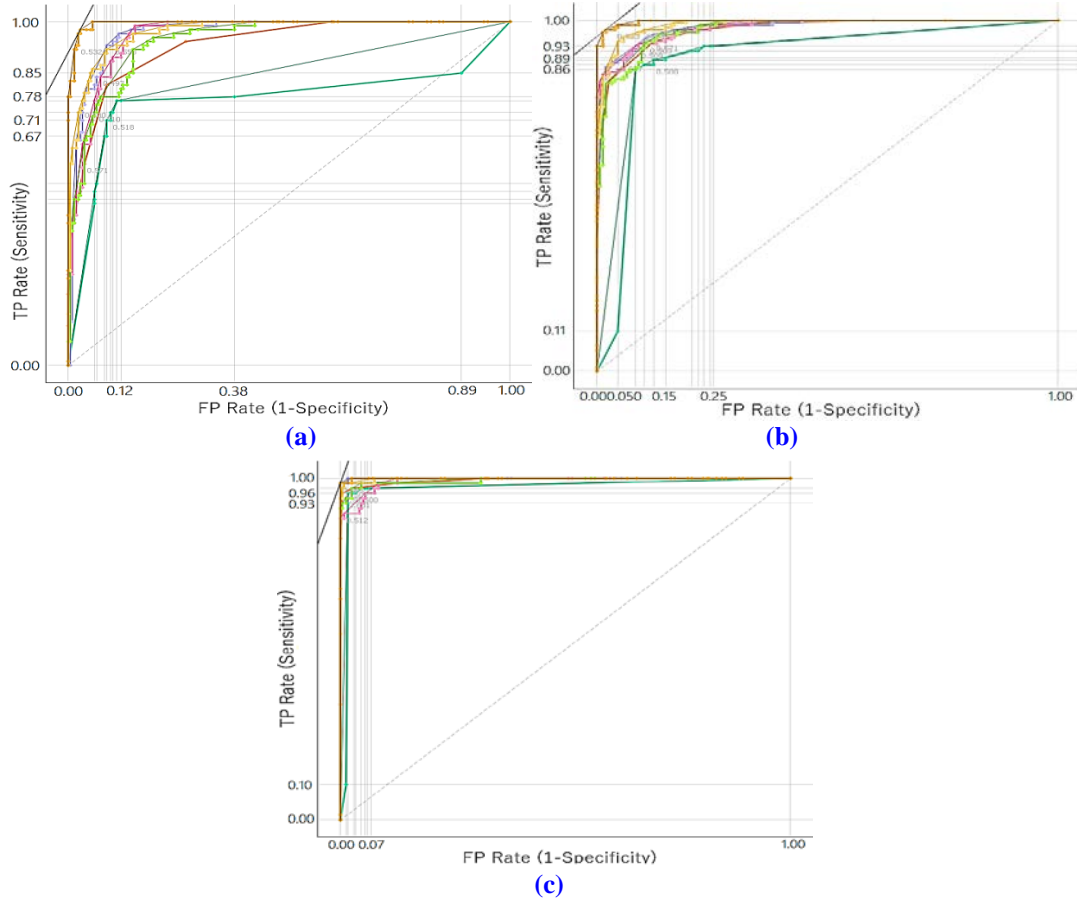


Fig. 12 (a) Representing ROC analysis for target class (Neutral) (b) Target class (Satisfied) (c) Target class (Unsatisfied)

The prediction models were further analyzed with ROC and it was observed that Gradient boost has the best test results as the curve of it is closer to the top (nearly 1 in the y-axis) and left-hand border of the ROC space with a testing environment having default threshold at 0.5 and performance line having FP cost = 500 and FN cost = 500 as seen in Fig. 12 (a), (b) and (c). Naïve Bayes and SVM performed better in predicting neutral classes, whereas in the case of satisfied and unsatisfied classes, kNN performed better than the two as seen in the figures.

6.2.2 Lift Curve

Further analysis of test results of the prediction model was done using a lift curve. It was seen that for the neutral class the highest lift is obtained in the first 20% of the data with 3.5 times more positive instances compared to a random model for Gradient boost. Similarly, for SVM the highest lift of 3.5 is obtained for the first 10% of the data followed by kNN. In the case of satisfying class prediction, the highest lift of 2.1 is obtained at first 50% of data for Gradient boost followed by Naïve Bayes that has the highest lift of 2.1 for the first 30% of data and then SVM with 2.1 for 38% of data compared to the random model. For unsatisfied class Gradient boost and Naïve Bayes has the highest curve with a value of 4.5 for the first 22% of data

followed by SVM and kNN with 4.5 lift for a data representation of 20% compared to a random model. The analysis of figures shows that Gradient boost is more stable and performs the best followed by Naïve Bayes and SVM on average as seen in **Fig. 13 (a), (b) and (c)**.

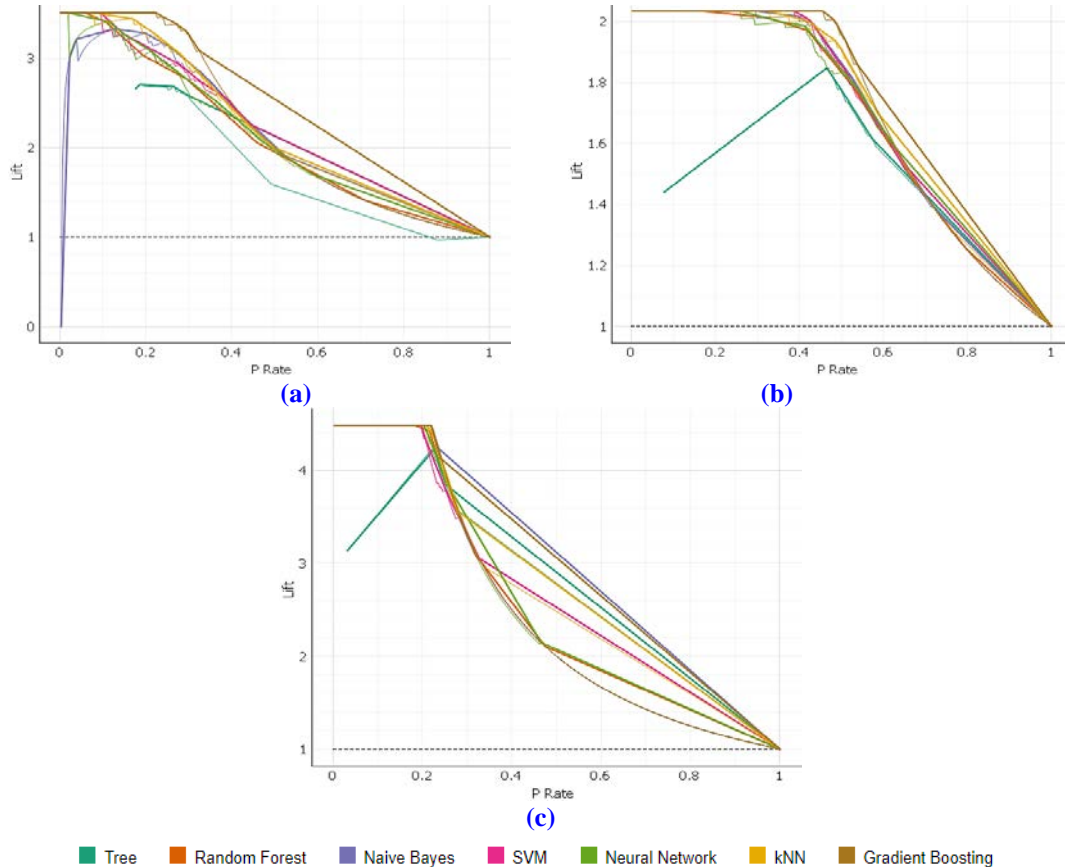


Fig. 13 (a) Representing Lift Curve analysis for target class (Neutral) **(b)** Target class (Satisfied) **(c)** Target class (Unsatisfied)

6.3 Confusion Matrix Evaluation for Gradient Boost.

A confusion matrix was constructed to further understand the prediction test result for the Gradient boost and Naïve Bayes algorithm. It can be seen in **Table 4** that the confusion matrix shows that predicted and actual results of the neutral class is 94.2% (81 instances), 96.1% (146 instances) for satisfied class, and 100% (67 instances) for unsatisfied class for Gradient boost. Further, it can be seen that 4.7% (4 instances) of satisfied instances are miss-classified as neutral, 1.2% (1 instance) of unsatisfied classes are miss-classified as unsatisfied and 3.9% (6 instances) of neutral classes are miss-classified as satisfied. The overall classification accuracy of Gradient boost is better than the accuracy of Naïve Bayes as seen in **Table 6** and **Table 7**.

Table 6. Confusion matrix for Gradient boost

	Predicted				Total
	Neutral	Satisfied	Unsatisfied		
Actual	Neutral	94.20%	3.90%	0.00%	87
	Satisfied	4.70%	96.10%	0.00%	150
	Unsatisfied	1.20%	0.00%	100.00%	68
	Total	86	152	67	305

Table 7. Confusion matrix for Naïve Bayes

	Predicted				Total
	Neutral	Satisfied	Unsatisfied		
Actual	Neutral	78.60%	4.30%	0.00%	87
	Satisfied	17.50%	95.70%	0.00%	150
	Unsatisfied	3.90%	0.00%	100.00%	68
	Total	103	138	64	305

6.4 Feature Contribution

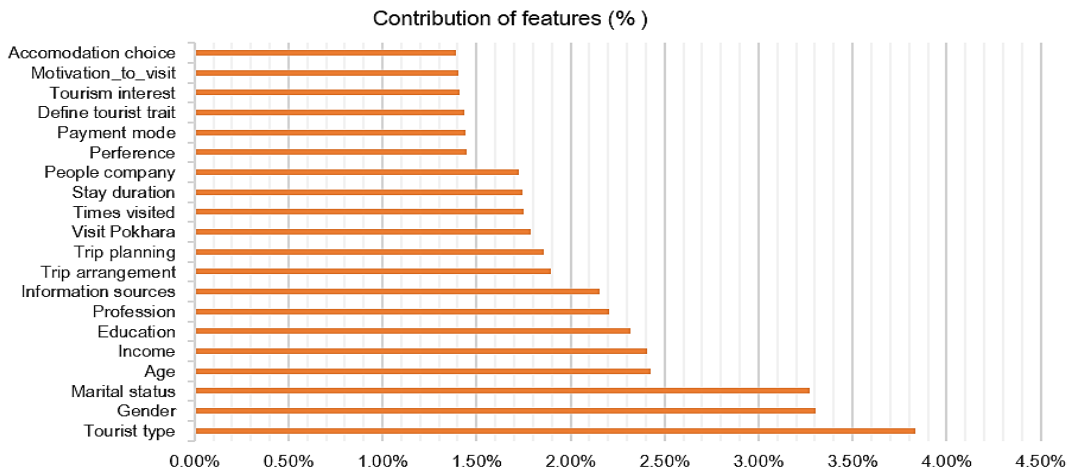


Fig. 14. Visualization of features that contribute to the prediction model

The role of contributing features in tourist satisfaction is an important aspect to understand. It is seen that 77 features contributed to the satisfaction value of tourists. The top twenty features with their contribution in percentage are plotted as a bar graph as shown in Fig. 14. It can be seen that tourist type (3.8%) is the most contributing feature followed by gender (3.25%), marital status (3.22%), age (2.4%), income (2.4%), profession (2.2%), followed by other features contributing less than 2%. These features help a tourism business to understand what features are most contributing and which features need improvement for the tourism business.

7. Discussions

This study analyzes tourist survey data of Pokhara city of Nepal for satisfaction value using multidimensional data analysis and machine learning models. The literature survey shows that it is the first study of its kind in Nepal that uses machine learning models for predicting satisfaction values. It is observed that for this study data preparation is a major activity and a lot of care is taken in pre-processing data as a bad data set would yield bad results and may not be appropriate for machine learning. This process removes 181 data sets after pre-processing, which accounts for 15.08% of data. This indicates that data obtained through a fixed type of questionnaire can have a lot of problems and appropriate methods must be applied to pre-process data before using it further. Multidimensional analysis and visualization performed on data depict many informative and interesting information clusters. The study shows that tourist type is an important component which in combination with gender, marital status, and age group gives a lot of information on tourist aspects and satisfaction. The international male tourists with married status are more frequent to the city compared to

female tourists. In the case of domestic tourist, gender and marital status has no significant contribution and are seen in equal numbers. In terms of age group 20-30 and 30-40 is the popular age group for domestic tourists and they are seen visiting in groups for both genders, whereas for international tourists an added group of 40-50 is also seen popular besides 20-30 and 30-40 age group. The above data indicate that Pokhara city is a favorite destination for people aged between 20 - 50 years having good potential. In terms of demography and accommodation, it is seen that irrespective of the gender and age group, hotels are the first choice, a homestay is second, and living with family and friends is the third important category. This data indicates that in terms of accommodation, Pokhara is a rich place. Understanding the visiting motives it is analyzed that vacations are the most important motive followed by cultural and community, and sports and entertainment. This visualization depicts that Pokhara is a popular vocational destination with rich culture, and is popular for entertainment and sporting events. Cost, people and culture, and regulations of the country are important factors that the tourist visiting Pokhara have considered. This indicates that Pokhara is competitive in terms of cost, is rich in people and culture with friendly regulations for tourists. The analysis of the overall satisfaction indicates that tourists irrespective of demography are seen as satisfied neutral for Pokhara destination. This inference of the study is also supported by studies [6] [26] [27] [28] [29] of Nepalese scholars.

The application of machine learning models to predict satisfaction class for this study provided very good results with an accuracy above 90% for all the models, excluding the Decision tree. This result is seen for both the training phase and prediction phase. This indicates that all the models under consideration have performed well and gradient boost is the best performing model. To study the model performances in more detail, matrices like AUC, CA, F1-Score, Precision, and Recall were used. All these matrices provided high output indicating that all models have performed well in identifying the class labels. Further, the test results of ROC analysis, Lift curve analysis and confusion matrix also confirm the high accuracy of the models. The study of machine learning models states that gradient boost is the best model for satisfaction prediction and it can be used in the future for other similar data sets.

The current study deals with tourist satisfaction data that keeps on growing in volume and has both categorical and numeric types (in this study there are 52 categorical and 43 numeric data). The analysis of such data type is best performed by machine learning models as they are not constrained by volume or data type [35]. Machine learning systems are capable of analyzing such data beyond the boundaries of linearity or continuity and are not constrained by dependent or independent variables [35]. Tourist satisfaction data is dependent on many variables that need a deep observation, analysis, and prediction of results in real-time [23] [24]. Machine learning models are capable of discovering hidden patterns and information by performing a deep analysis of such data. The learning and predictive power of these systems are generally very strong and can outperform any statistical analysis [35]. The satisfaction models of this study (based on machine learning) can be further used to build recommender systems, tourist satisfaction models, and intelligent tourist information systems. Machine learning models can help in the regular and continuous analysis of tourist satisfaction data to enforce the process of improvement of tourism products and services. These models are the ultimate solution to make the satisfaction model dynamic and use it for any future implementations including system development or tourist satisfaction analysis.

This study serves as a satisfaction barometer for the tourism stakeholders who are responsible for devising the tourism business policy. Regular measurement of satisfaction can lead towards the betterment of the tourism business and enforce improvements in the weaker areas of this business, which is also supported by study references [3] [6]. The study depicts

that tourist demography has the highest contribution to satisfaction (as per machine learning analysis). The tourism business stakeholders should devise packages considering demographic data (age, income, gender, etc.) that help them in the growth of the overall tourism business. The study shows the dissatisfaction of tourists in the area of personal safety, tourism assistance, health and hygiene, and local transport. The improvement in this sector will bring relatively high efficiency and increase the satisfaction of tourists, increasing the overall business. This study can also be applied to other major tourist destinations of Nepal and measure the satisfaction level. An overall tourist satisfaction model for the whole country can be developed, which will help to gain insight into the tourism business and impose improvements at the national level. The inferences of this study serve as a knowledge base for tourism governing bodies and business houses in understanding the tourist expectation and availability of tourist services in reality. Policy formations, quality improvement processes, the design of attractive tourism packages, and plans to improve the deficient sector of tourism can be implemented more precisely. The tourist on the other hand can use this study for assessing the satisfaction level of tourism products and services of Nepal.

8. Limitations of the Study

This study excludes tourist behavioral data and has 1019 samples limited to Pokhara city only. A more extensive survey with behavioral aspects combined with current attributes in other cities can bring a generalized model for tourist satisfaction in Nepal. The addition of other attributes can create specialized groups, which can be used to target future tourists through policies and plans. The overall study and a common model can be developed considering other geographical areas and using the machine-learning system in tourism.

9. Conclusion

The study, multi-dimensional analysis, and prediction model for tourist satisfaction brought many interesting facts into the light through analysis and visualization of the survey data. The data analysis depicts that Pokhara city is a favorite tourist destination for both international and domestic tourists. The demographical data shows that gender, marital status, age group are important attributes that play a vital role in the consumption of tourism products and services in Pokhara. The scattered tourist clusters for tourism activities especially in sports, entertainment, culture, community, nature, entertainment, and health display the tourism vitality of the destination. Pokhara is considered a vacation destination for tourists, followed by cultural and community purposes. Tourist meets their expectations in terms of cost, information access, service quality, priority, safety, health, hygiene, etc. The overall satisfaction level is high for tourists in Pokhara compared to neutral and dissatisfied tourists. The use of state of art models for training and prediction purposes for the tourism data set yields a high accuracy above 90% for all the seven models except the decision tree. The performance metrics, test results, and other measures depict that the supervised machine algorithms have performed well in both the training and prediction phase by identifying the class labels with an accuracy of close to 99%. The use of confusion matrix, ROC analysis, and lift curve further confirm this output. The visualization of attributes identifies 77 attributes that contribute as a whole to tourist satisfaction that include tourist type, gender, marital status, age, income, information sources, trip arrangement factors, motivations, etc. as the most important attributes.

It can be finally concluded that tourists are satisfied in Pokhara as a tourism destination and it holds a good tourism destination value. The use of machine learning further confirms that the gradient boost model can be used as a base model to further build the satisfaction model with high accuracy. The application of machine learning with multidimensional analysis of the survey data infers that Pokhara city provides satisfaction to all types of tourists irrespective of their demographics and is a popular tourist destination of Nepal.

References

- [1] World Travel & Tourism Council, "Nepal 2021 Annual Research: Key Highlights," WTTC Tourism report, London, SE1 0HR, UK, March 2021. [Online]. Available: <https://wttc.org/Research/Economic-Impact>
- [2] D. Foris, M. Popescu, and T. Foris, "A Comprehensive Review of the Quality Approach in Tourism, Mobilities," *Tourism and Travel Behavior*, IntechOpen, 2017. [Article \(CrossRef Link\)](#).
- [3] R. Castellano, F.M. Chelli, M. Ciommi, G. Musella, L. Salvati, "Trahit sua quemque voluptas. The multidimensional satisfaction of foreign tourists visiting Italy," *Socio-Economic Planning Sciences*, Vol. 70, 2020. [Article \(CrossRef Link\)](#)
- [4] G. Dominici, F. Palumbo, "The drivers of customer satisfaction in the hospitality industry: applying the Kano model to Sicilian hotels," *International Journal of Leisure and Tourism Marketing*, Vol.3, No.3, pp. 215-236, 2013. [Article \(CrossRef Link\)](#)
- [5] P. Ba I., "Tourist Perceived Value, Relationship to Satisfaction, and Behavioral Intentions: The Example of the Croatian Tourist Destination Dubrovnik," *Journal of Travel Research*, Vol. 54. No. 1, pp 122-134, 2015. [Article \(CrossRef Link\)](#)
- [6] N. Matsatsinis, E. Krassadaki, and P. Delias, "A tourists' satisfaction analysis using multiple criteria analysis and machine learning techniques. The case of Chania as destination place," *Modern Methodological Trends in Tourism Management*, Kleidarithmos, 2006.
- [7] D. Buhalis, R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research," *Tourism Management*, Vol. 29, no. 4, pp. 609-623, 2008. [Article \(CrossRef Link\)](#)
- [8] G. Zang, "Cultural Tourism Products Design Optimization for User Satisfaction," in *Proc. of International Conference on Virtual Reality and Intelligent Systems*, 2018. [Article \(CrossRef Link\)](#).
- [9] X. Yang, "Satisfaction Evaluation and Optimization of Tourism E-Commerce Users Based on Artificial Intelligence Technology," in *Proc. of 2019 International Conference on Robots & Intelligent System*, pp. 373-375, 2019. [Article \(CrossRef Link\)](#).
- [10] H. Su, X. Lin, Q. Xie, W. Chen and Y. Tang, "Research on the Construction of Tourism Information Sharing Service Platform and the Collection of Tourist Satisfaction," in *Proc. of 3rd International Conference on Smart City*, pp. 640-643, 2018. [Article \(CrossRef Link\)](#).
- [11] L. Lin, "Study on the evaluation of tourist satisfaction in tourism destination based on a variable precision rough set," in *Proc. of International Conference On Computer Design and Applications*, pp. V1-124-V1-128, 2010. [Article \(CrossRef Link\)](#).
- [12] A. K. Agiomirgianakis G, Mihiotis A, "Measuring tourist satisfaction: A factor-cluster segmentation approach," *Journal of Vacation Marketing*, Vol. 14, pp. 221-235, 2008. [Article \(CrossRef Link\)](#)
- [13] F. Meng, Y. Tepanon, M. Uysal, "Measuring tourist satisfaction by attribute and motivation: The case of a nature-based resort," *Journal of Vacation Marketing*, vol. 14, pp. 41-56, 2008. [Article \(CrossRef Link\)](#)
- [14] A.M Oliveri, G.Polizzi, & A.M Parroco, "Measuring Tourist Satisfaction Through a Dual Approach: 4Q Methodology," *Soc Indic Res*, Vol. 146, pp 361–382, 2019. [Article \(CrossRef Link\)](#)

- [15] J. C. Martin, M. Saayman, E. du Plessis, "Determining satisfaction of international tourist: A different approach," *Journal of Hospitality and Tourism Management*, Vol. 40, pp. 1-10, 2019. [Article \(CrossRef Link\)](#)
- [16] J. C. Castro, M. Quisimalin, C. de Pablos, V. Gancino, J. Jerez, "Tourism Marketing: Measuring Tourist Satisfaction," *Journal of Service Science and Management*, Vol.10, No. 03, pp. 280-308, 2017. [Article \(CrossRef Link\)](#)
- [17] S. Gössling, "Technology, ICT, and tourism: from big data to the big picture," *Journal of Sustainable Tourism*, Vol 29, No. 5, pp 849-858, 2021. [Article \(CrossRef Link\)](#)
- [18] P. Beqiraj, E. Gjermëni, "Tourist's Satisfaction in Terms of Accommodation: A Case Study in Albania," *Business Perspectives and Research*, Vol 8, pp 67-80, 2020. [Article \(CrossRef Link\)](#)
- [19] S. Majeed, Z. Zhou, C. Lu, and H. Ramkissoon, "Online Tourism Information and Behavior: A Structural Equation Model Analysis," *Front. Psychol.*, Vol. 11, 2020. [Article \(CrossRef Link\)](#)
- [20] U., Stankov, Gretzel, U, "Tourism 4.0 technologies, and tourist experiences: a human-centered design perspective," *Inf Technol Tourism*, Vol. 22, pp 477-488, 2020. [Article \(CrossRef Link\)](#)
- [21] A. Verma, and V. Shukla, "Analyzing the Influence of IoT in Tourism Industry," in *Proc. of International Conference on Sustainable Comp. Sc. Tech. and Mgmt*, 2019. [Article \(CrossRef Link\)](#)
- [22] S. Suthathip, "Factors Affecting Tourist Satisfaction: An Empirical Study in the Northern Part of Thailand," in *Proc. of SHS Web of Conferences*, 2014. [Article \(CrossRef Link\)](#)
- [23] H. Su, Q. Xie, X. Lin, W. Chen, D. Gao, and Y. Tang, "Analysis of Tourist Satisfaction Based on Internet Public Opinion and Big Data Collection," in *Proc. of 3rd International Conference on Smart City and Systems Engineering*, pp. 721-724, 2018. [Article \(CrossRef Link\)](#)
- [24] F. Huang and L. Su, "A study on the relationships of service fairness, quality, value, satisfaction, and loyalty among rural tourists," in *Proc. of 7th International Conference on Service Systems and Service Management*, pp. 1-6, 2010. [Article \(CrossRef Link\)](#)
- [25] B. Wang and C. Wu, "A Systematic Comparison of First-Time and Repeat Visitors' Satisfaction with a Destination," in *Proc. of 4th International Conference on Wireless Communications, Networking, and Mobile Computing*, pp. 1-4, 2008. [Article \(CrossRef Link\)](#)
- [26] R. Baniya, P. Thapa, "Hotel Attributes Influencing International Tourists' Satisfaction and Loyalty," *Journal of Tourism and Hospitality Education*, Vol.7, pp. 44-61, 2017. [Article \(CrossRef Link\)](#)
- [27] A. Poudel, R. K. Phuyal, "An analysis of foreign tourist behavior and their satisfaction in Nepal," *International Journal of Applied Business and Economic Research*, Vol. 14(3), pp. 1955-1974, 2016.
- [28] N. Bam, A. Kunwar, "Tourist Satisfaction: Relationship Analysis among its Antecedents," *Advances in hospitality and tourism research*, Vol. 8, no. 1, pp. 30-47, 2020. [Article \(CrossRef Link\)](#)
- [29] N. Baral, H. Hazen & B. Thapa, "Visitor perceptions of World Heritage value at Sagarmatha National Park," *Journal of Sustainable Tourism*, Vol 25, No 10, pp. 1494-1512, 2017. [Article \(CrossRef Link\)](#)
- [30] D. Shrestha, T. Wenan, A. Khadka, and S. R. Jeong, "Digital Tourism Security System for Nepal," *KSII Trans. Internet Inf. Syst.*, Vol 14, No. 11, pp. 4331-4354, 2020. [Article \(CrossRef Link\)](#)
- [31] T. Wenan, D. Shrestha, N. Rajkarnikar, B. Adhikari and S. R. Jeong, "Digital reference model system for religious tourism & its safeties," in *Proc. of International Conference on Engineering Technologies and Applied Sciences*, pp. 1-6, 2020. [Article \(CrossRef Link\)](#)
- [32] T. Wenan, D. Shrestha, D. Shrestha, B. Gaudel and S. R. Jeong, "Analysis and Design of Tourism Recommender System for Religious Destinations of Nepal," in *Proc. of International Conference on Sustainable Engg. & Creative Computing*, 2020. [Article \(CrossRef Link\)](#)
- [33] D. Shrestha, T. Wenan, B. Gaudel, N. Rajkarnikar, and S. R. Jeong, "Digital Tourism Business Ecosystem: Artifacts, Taxonomy and Implementation Aspects," *International Journal of Innovative Research in Computer Science & Technology*, Vol. 9, no. 5, September 2021. [Article \(CrossRef Link\)](#)

- [33] D. Shrestha, T. Wenan, D. Shrestha, N. Rajkarnikar, S. Niroula, “Analysis of ICT Infrastructure and Tourism Informational Needs: A Case Study of Nepal,” *International Journal of Innovative Research in CS & Tech.*, Vol 9, Issue 6, pp 1-10, December 2021. [Article\(CrossRefLink\)](#)
- [34] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, & TensorFlow, Concepts, Tools, and Techniques*, Second Edition, O'Reilly Media, Inc., ISBN: 9781492032649, 2019.



Deepanjali Shrestha is a Ph.D. research candidate at Nanjing University of Aeronautics and Astronautics, School of Computer Science and Technology, Nanjing, China. He is also an Asst Prof. at the School of Business, Pokhara University, Nepal. His research interest includes Management information systems, Software engineering, Digital ecosystems, Service engineering and Machine learning.



Tan Wenan is a Professor of Computer Science at the Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is also a Professor at Shanghai Polytechnic University, Shanghai, China. He has attained an Engineering degree in Computer Software, a Ph.D. degree in Computer Software and Theory from Beihang University, Beijing, China. His main research interests include Software service engineering, Service Computing, and Enterprise intelligent information systems.



Bijay Gaudel holds a Master of Engineering Degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, and a BE in Electrical Engineering from the IOE, Nepal. His master's degree research is in Graph representation learning using deep learning and his research interest includes Computer vision, graph representation learning, and complex network analysis.



Neesha Rajkarnikar is a Ph.D. Scholar at Nanjing Tech University, Department of New Energy Science & Engineering, Nanjing, China. She has a Master's degree in Computer Applications, Masters in Gender & Development. Her area of interest includes Computer Applications, Energy Science & Social Computing.



Seung Ryul Jeong is a Professor and Vice President of Kookmin University, South Korea. He has attained a Bachelor's degree in Economics from Sogang University, Seoul, Korea, an M.S. in Management Information Systems from the University of Wisconsin - Milwaukee, U.S.A and a Ph.D. in MIS from the University of South Carolina, SC, USA. He has published more than 200 research articles and papers in various refereed journals and conferences. His current research includes System implementation, Process Innovation, Resource Optimization, and Machine learning.