International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

To Classify the Tourist Behavior through Non Verbal Communication Using Machine Learning

Deepak[1], Dr. Anupam Bhatia, Dr. Naveen kundu
[1]MCA Student, [3]Asst. Prof., [2]Associate Professor
Deptt. of Comp. Sc. and Application, Chaudhary Ranbir Singh University, Jind, Haryana, India

Abstract

Tourist behavior analysis is essential for improving customer satisfaction and offering personalized tourism services. Traditional methods for understanding tourist preferences often depend on surveys and questionnaires, which may not accurately reflect tourists' true feelings and reactions. This research suggests a machine learning approach to analyze tourist behavior by using nonverbal communication features. The dataset for this study was collected from Kaggle and consists of 73 records with 23 attributes. These attributes include demographic information, gestures, postures, audio preferences, emotional traits, and proxemics. We applied data preprocessing techniques to address missing values, convert categorical data to numerical form, and improve data quality. To identify the most suitable variables for model training, we use feature selection. To tackle class imbalance, we used the SMOTE technique to increase the dataset from 73 records to 300. We implemented basic Ml methods like decision tree, logistic regression, XGboost, and random forest. The dataset was divided into 20% for testing and 80% for training. Accuracy, recall, precision, F1-score, and confusion matrix were used to evaluate the model performance. According to the results, XGBoost outperformed the other models with the highest testing accuracy of 95.00%. These findings indicate that nonverbal behavioral traits can effectively classify different types of tourist customers and support smart decision-making in the tourism industry. The system we proposed offers an effective way to understand tourist behavior and improve personalized tourism services.

Keywords: Tourist Behavior, Nonverbal Communication, Customer Relationship management (CRM), Machine Learning Algorithm

Introduction

Tourism plays an important role in the global economy. It contributes to job creation, economic development, cultural exchange, and international relations. Our success does not depend on how many visitors come, but on whether tourists are happy [1].

Tourist Satisfaction- Tourist satisfaction depends on many factors, including the quality of service, available facilities, and employee customer relationships. If a customer is satisfied with the service once, they are more likely to visit the place again and recommend it to others. CRM Customer Relationship Management is a technology that helps in gathering and understanding customer behavior. It also helps to identify the customers, attract them, and maintain relationships with them. The main goal is to identify customer needs and, based on those needs, provide good service and improve our business [1].

Nonverbal Communication: In verbal communication, people primarily express their thoughts by speaking or writing, whereas in nonverbal communication, people are understood through facial expressions, body gestures, tone of voice, and body language. That's why body language can be trusted more than words.

People's emotions were assessed solely through surveys or responses, which did not always reflect their true feelings. Therefore, to understand people's true emotions, it is necessary to connect with them directly and interpret their facial expressions, hand gestures, tone of voice, and body language rather than relying only on their responses or surveys. Even though Machine Learning is widely used in tourism, very few studies focus on understanding tourist behavior using nonverbal communication like facial expressions, gestures, and body language. Most research relies on surveys and written feedback, which may not reflect tourists' true feelings. Also, there is a lack of adequate models that can integrate multiple nonverbal signals to produce accurate results. Collecting good-quality data and handling privacy issues are also major challenges. So, there is a research gap in developing simple and effective ML-based systems to study tourist behavior using nonverbal communication.

Literature Review

Sana Abyari et al. (2025) presented an advanced machine learning technique and a framework for applying AI approaches to analyze the nonverbal behavior of tourists. To improve model reliability and handle class imbalance, Smote technique was applied. They used a machine learning algorithm Extreme Gradient Boosting, and Random Forest for classification and found that the Optimized XGBoost model performed best, achieving an overall accuracy of 98.3% [1].

Tusell-Rey et al. (2021) collected data via a questionnaire at the hotel. They used clustering & classification algorithms to study tourist behavior and group similar tourists based on their nonverbal communication. The data is obtained from a mixed, incomplete questionnaire; therefore, to handle this type of data, we use a clustering algorithm [2].

Melike Sak et al. (2024) examined the relationship between communication styles and non-verbal communication of tourist guides. The study highlighted that facial expressions, hand gestures, body posture, and eye contact play an important role in communication with tourists. Good nonverbal communication can improve the quality of interaction and lead to a better tourist experience [3].

Kun Zhang et al. (2019) examined tourist photos using a deep learning model to understand more about travelers' attitudes and actions. The study showed that AI can identify preferences, customer behavior, and destination experiences using visual content. This approach helps researchers better understand tourist behavior and enhances tourism management decisions [5].

Marij A. Hillen et al. (2015) studied how the trust of patients is affected by doctors' nonverbal communication in medical appointments. This research paper study found that making eye contact, smiling, and proper posture all increase communication and patients' improved satisfaction. Good nonverbal communication builds trust and helps better communication between patients and doctors [6].

Dataset description

This study's datasets were gathered from Kaggle [4]. It contains 73 records and 23 attributes. It includes features like facial expressions, hand gestures, tone of voice, body postures, proxemics, and customer type classification. The target variable in the dataset is **Type of Client**, which classifies tourists into different customer groups.

Descriptive Statistics of Qualitative Variables

Qualitative variables in this research represent the categorical information related to tourists' nonverbal preferences. It includes features such as gender, country, customer return status, reactions to images (group and individual), reactions to sounds (tone and quality), proxemics, and customer type [1]. These variables help to understand customer non verbal behavior and improve satisfaction.

Descriptive Statistics of Quantitative Variables

Quantitative variables in this research represent the numerical information related to tourists' nonverbal preferences. It includes features such as age, tense-relaxed, authoritative-anarchic, and hostile-friendly attitude [1]. These variables provide measurable information that can be used in statistical analyses and ML models to predict tourist behaviour and improve the quality of service.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

Table.1 Dataset Description [2]

Sr. No.	Name	Feature Description	Admissible Values
1	Sex	Sex of client	Female, male, ?
2	Country	Country of client	United nation admitted country, ?
3	Age	Age of client	0- 100, ?
4	Returning	Returning status	Yes, no, ?
5	PImg 1	Consent posture	Like, dislike, indifferent, ?
6	PImg 2	Interest posture	Dislike, like, indifferent, ?
7	PImg 4	Reflexive posture	Like, dislike, indifferent, ?
8	PImg 3	Neutral posture	Indifferent, like, dislike, ?
9	PImg 5	Negative posture	Indifferent, dislike, like, ?
10	GImg 1	handshake	Dislike, like, indifferent, ?
11	GImg 3	kiss	Like, dislike, indifferent, ?
12	GImg 2	hug	Dislike, like, indifferent, ?
13	TAudio 2	Sarcastic tone	Like, dislike, indifferent, ?
14	TAudio 1	Authoritative tone	Dislike, indifferent, like, ?
15	TAudio 3	Friendly tone	Dislike, like, indifferent, ?
16	Hostile-friendly	Emotional clime observed	10: Too friendly, 1: Too hostile 1-10, ?
17	Authoritative-anarchic	Emotional clime observed	10: Too anarchic, 1: Too authoritative 1-10, ?
18	QAudio 1	Spitting sound	Like, dislike, indifferent, ?
19	QAudio 3	Sigh sound	Dislike, like, indifferent, ?
20	QAudio 2	Hum sound	Like, indifferent, dislike, ?
21	Tense-relaxed	Observed emotional climate	1-10, ? 1= Too tense, 10= Too relaxed
22	Proximics	Physical distance	A, B, C, D, ?(intimate:15-45 cm, personal:46-122 cm, social:123-360 cm, public:>360 cm)
23	Type of Client	Customer Classification	0= Low-Interaction, 1= Family-oriented, 2= Cost-sensitive, 3= loyal, 4= Regular, 5= Special

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

Methodology

The flowchart represents the complete methodology followed in this research. It describes the flow of present study from data collection to the final prediction. Figure 1 show the present study workflow:

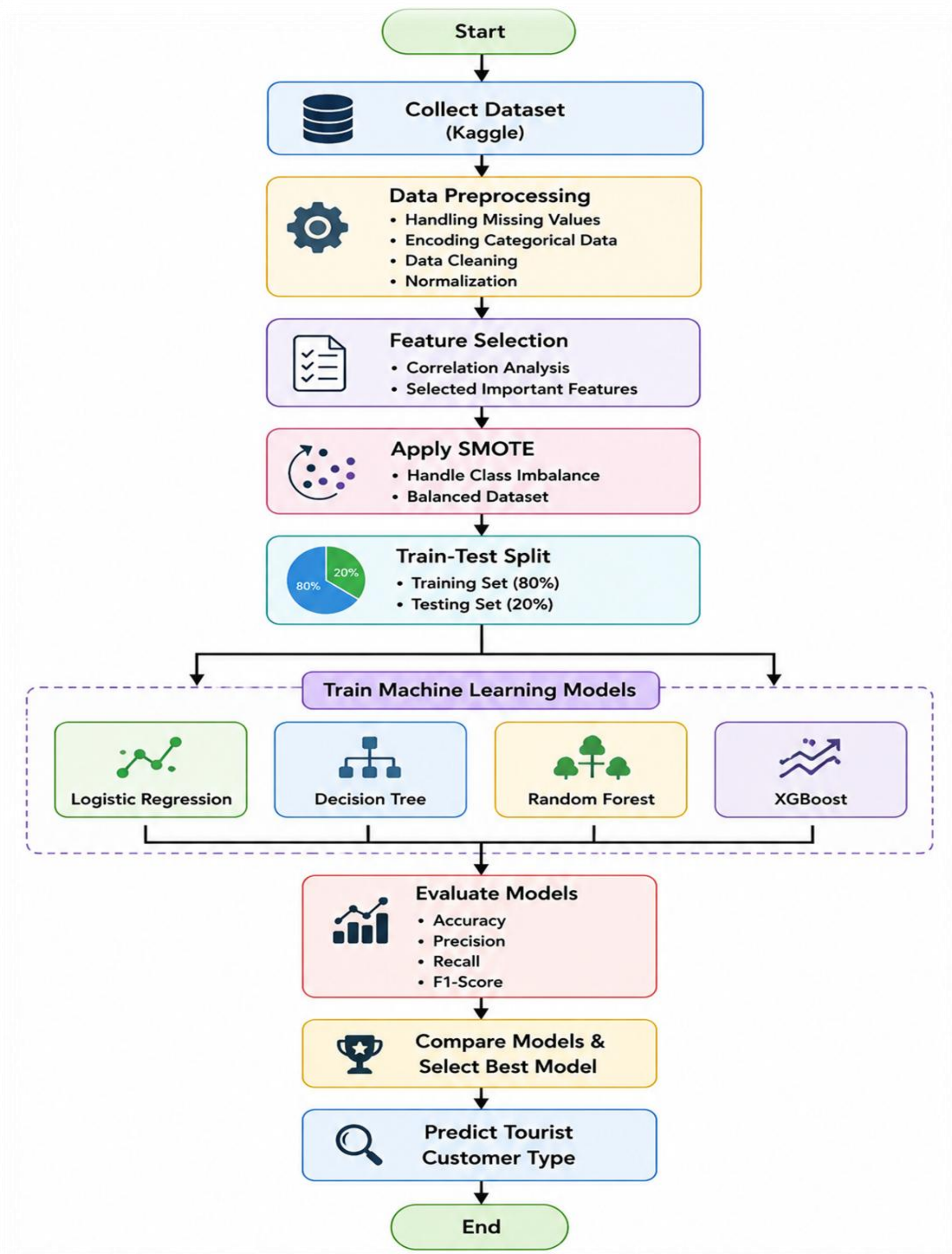

Figure.1: The Flowchart of present study

1. **Data Collection:** The datasets was downloaded from Kaggle [4]. It contains 73 records and 23 attributes.
2. **Data Preprocessing:** Many data processing methods were applied to improve the quality of the datasets.
    - **Handling Missing Values:** In this we identify missing values and fill them by mode, mean, and median.
    - **Removing Duplicate Records:** Duplicate entries were checked and removed to avoid redundancy in the dataset.
    - **Data Cleaning:** Inconsistent and irrelevant data were corrected or removed from the dataset.
    - **Feature Engineering:** Feature Engineering was performed to convert the raw data into a suitable format for Ml algorithms. **Label encoding and one-hot encoding** were used to translate categorical data into numerical values.
3. **Feature & Target Selection:** Feature selection performed to select the important features & remove unnecessary features from the datasets, improving the accuracy of the machine learning model. The datasets was divided into input feature(X) and output feature(y).

   Input Features (X):
    - Age
    - Sex
    - Returning
    - Country
    - Tense-relaxed
    - TAudio1 – TAudio3
    - PImg1 – PImg5
    - Hostile-friendly
    - GImg1 – GImg3
    - Proxemics
    - QAudio1 – QAudio3
    - Authoritative-Anarchic

   Important Input Features Used for Model Training (X):
    - PImg3, PImg1
    - GImg2
    - Tense - relaxed
    - QAudio1, TAudio1

   Output feature(y):
    - Type of Client

4. **SMOTE Technique:** SMOTE is used to handle class imbalance in a datasets. Such imbalance can decrease the accuracy of machine learning algorithms, as model usually learns dominant classes more efficiently and, at the same time, ignores minority classes. SMOTE generates new synthetic data points by forming linear combinations between existing samples and their nearest neighbors, rather than simply duplicating minority class samples. The original datasets are 73 records. Then the smote technique was applied to balance the distributions of classes and the dataset was enhanced to 300 records. This improved the efficacy and accuracy of the classification models [1].
5. **Train-test-split data:** The dataset was divided into 20% testing data and 80% training data.
6. **Model Selection:** Four classification algorithms were applied to predict the tourist behavior. **Logistic Regression, XGboost Classifier, Decision Tree Classifier, and Random Forest Classifier.**
7. **Model Evaluation:** Model evaluation was used to evaluate the performance of the models. These metrics are precision, recall, accuracy, and F1-score.

   Metrics defines as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

TN(true negative), TP(true positive), FN(false negative), and FP(false positive) are the correct negative predictions, correct positive predictions, false negative predictions, and false positive predictions respectively [1].

Result & Discussion

The outcomes of the machine learning models for classifying tourist behaviour are shown in this section. The result findings main three areas: correlation matrix analysis, confusion matrix analysis and model performance evaluation.

Correlation Matrix Analysis

The correlation matrix was used to show the relationship between each feature in the dataset. It helps to identify whether two features are correlated with each other. The correlation value lies between +1 and -1, where 0 denotes no correlation, +1 denotes strongly positive correlation and -1 denotes strongly negative correlation. A correlation matrix is displayed as a heatmap, making it

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

easier to identify strong and weak relationships between variables. It is used during data analysis to select important features, reduce redundancy, and improve the accuracy of machine learning models. Based on the correlation matrix analysis, the important features such as PImg3, PImg1, GImg2, Tense–Relaxed, QAudio1, and TAudio1 were selected and used to train the machine learning model. Figure 2 show correlation matrix analysis:

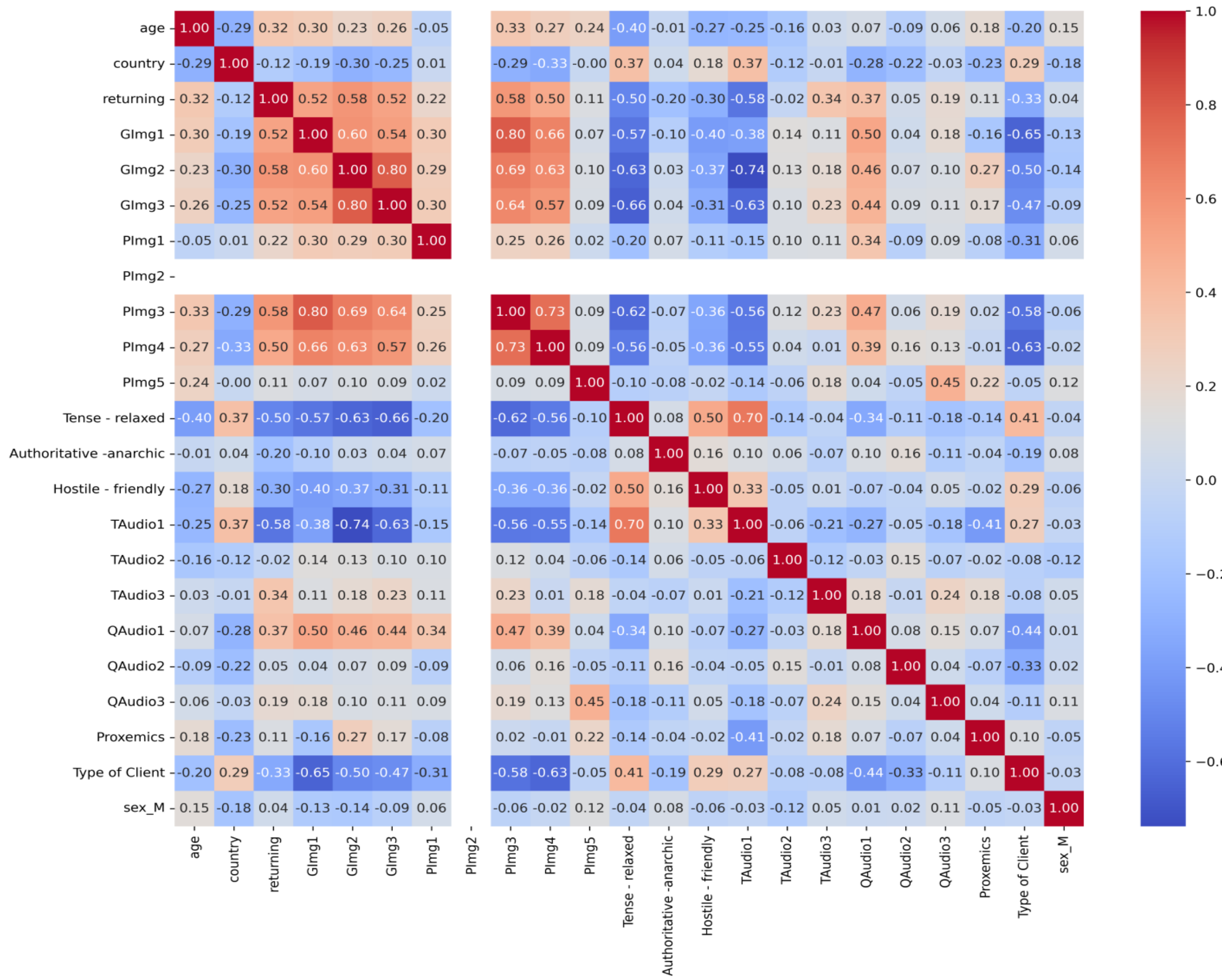

Figure.2: Correlation Matrix Analysis

Confusion Matrix Analysis

The confusion matrices are used to compare the performance between actual and the predicted class. The off-diagonal represents incorrect guesses, and the diagonal represents correct guesses. Figure 3 present the confusion matrices for all four classification models **Logistic Regression, Decision Tree Classifier, XGBoost Classifier, and Random Forest Classifier**. These matrices provide a detailed class-wise view of prediction accuracy. According to the findings, the XGBoost (XGB) model produced the best classification performance across all classes, with the most accurate predictions and the lowest number of incorrect classifications. The Random Forest

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

Classifier (RFC) also performed well but gives a few more incorrect predictions as compared to XGBoost. The decision tree also performed moderately but gives more incorrect predictions as compared to XGBoost and Random Forest. The logistic regression model gives a high number of incorrect predictions and performs poorly. Overall, in four classification models, XGBoost model is the best for predicting tourist behavior.

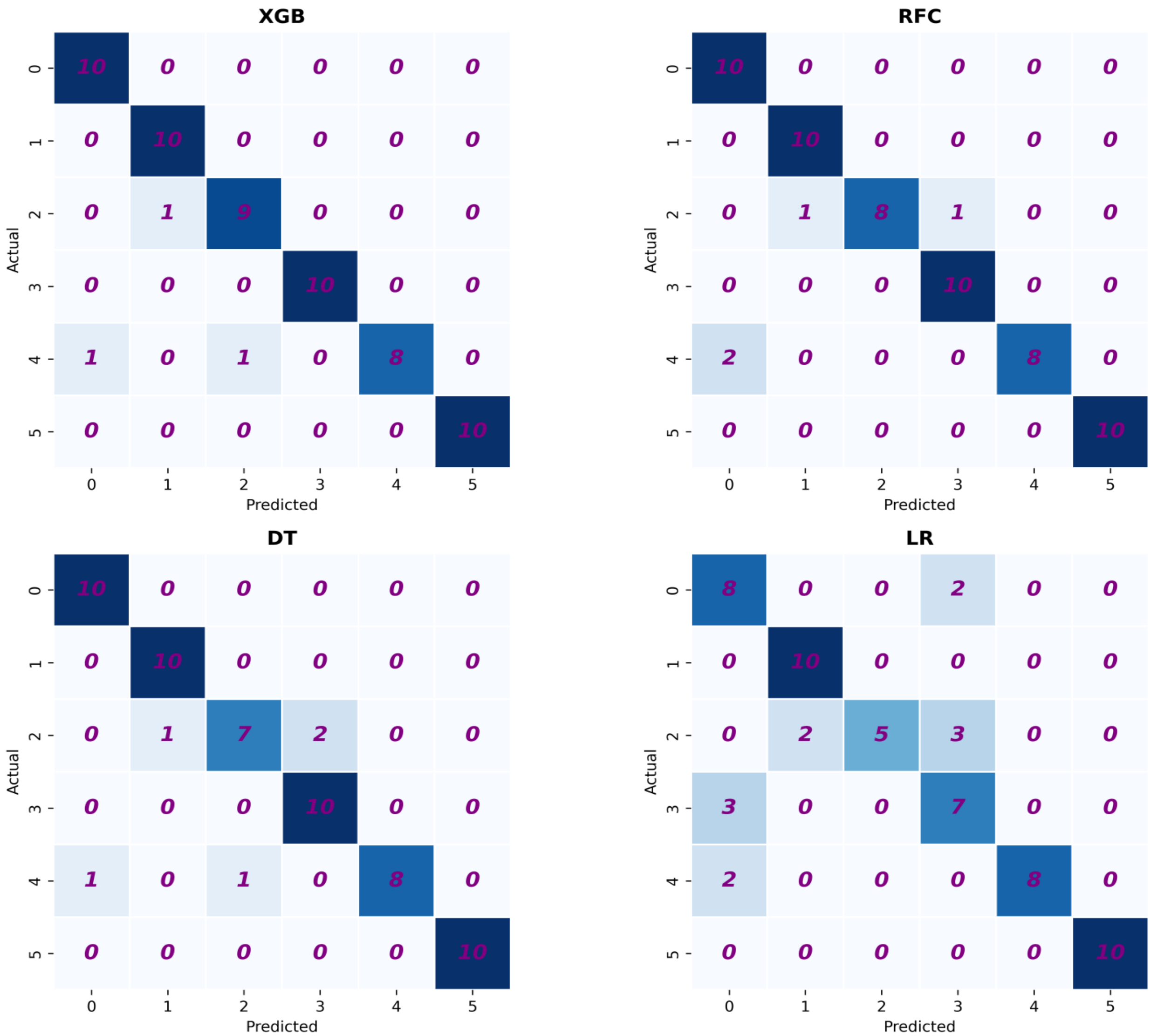

Figure.3: Confusion Matrix Analysis

The performance comparison of four machine learning models using training accuracy, testing accuracy, precision, recall, and F1-Score is shown in Table 2.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 242-252 March 2026-Sept 2026 www.csjournals.com

Table2. Model performance comparison

Model	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
XGB	95.42	95.00	95.30	95.00	94.89
RFC	95.83	93.33	94.19	93.33	93.19
DT	95.83	91.67	92.11	91.67	91.34
LR	84.58	80.00	83.87	80.00	79.94

Among all the models, XGBoost obtained the highest accuracy of 95.00% in testing, followed by Random Forest (93.33%), Decision Tree (91.67%) and Logistic Regression (80.00%). A similar pattern is observed for recall, f1-score, and precision. These results show that XGboost was the best model for predicting tourist customer types.

Conclusion

In this research, a machine learning model was built to classify tourist behavior based on features related to nonverbal communication including hand gestures, body postures, voice preferences, emotion properties, facial expression and proxemics. This study uses a dataset of 73 records with 23 features that was obtained from Kaggle. Some preprocessing methods, which include handling missing value, encode categorical values, select some important features and SMOTE method, were performed to enhance the dataset quality and balance the dataset. There are four ML models in this research, namely decision tree, logistic regression, random forest, and XGboost. The result shows that the highest testing accuracy is achieved by XGBoost model which is 95.00%, followed by Random Forest which is 93.33%, Decision Tree which is 91.67%, and Logistic Regression which is 80.00%. From the result above, it can be seen that nonverbal behavioral properties are helpful in classifying tourist customers. Hence, XGBoost model is chosen as the best model in this research.

Future Scope

In future, this research can be extended in the following ways:

- Use a big dataset for better model training.
- Collect real-time tourist behavioral data.
- Use video-based behavior analysis.
- Apply deep learning algorithms such as CNN.
- Develop a real-time tourist monitoring system.
- Improve prediction accuracy using advanced feature engineering.
- Use IoT devices for behavior data collection.

References

[1] Sana Abyari, Vali Nemati, Rasoul Abbasi Taghidizaj, Sadegh Afzal, "Optimizing Tourist Satisfaction through Nonverbal Preference Analysis Using Advanced Artificial Intelligence Techniques," Journal of Artificial Intelligence and System Modelling, vol. 3, no. 4, 2025, https://doi.org/10.22034/jaism.2025.553283.1166

[2] C. C. Tusell-Rey, R. Tejeida-Padilla, O. Camacho-Nieto, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Improvement of Tourists Satisfaction According to Their Non-Verbal Preferences Using Computational Intelligence," Applied Sciences, vol. 11, no. 6, p. 2491, 2021, doi: 10.3390/app11062491.

[3] Melike Sak, Ali Turan Bayram, Gül Erkol Bayram, Anuj Kumar and Shakeel Basheer, "The Relationship Between Communication Styles And Nonverbal Communication Level: A Study On Tourist Guides," Journal of Mediterranean Tourism Research, vol. 3, no. 2, 2024, https://www.doi.org/10.5038/2770-7555.3.2.1029

[4] Kaggle, "Tourist Behavior Dataset," [Online]. Available: https://www.kaggle.com/. Accessed: Jun. 24, 2026.

[5] Ye Chen, Kun Zhang, Chunlin Li, "Discovering the Tourist Behaviors and Perception in a Tourism Destination by Analyzing Photos Visual Content with a Computer Deep Learning Model: The Case of Beijing," Tourism Management, vol. 75, pp. 595-608, 2019, doi:10.1016/j.tourman.2019.07.002.

[6] Hanneke C. J. M. de Haes, Marij A. Hillen, Hanneke W. M. van Laarhoven, Danielle M. Vermeulen, Nina Bijker, Geertjan van Tienhoven, Ellen M. A. Smets, "All eye on the patient: The influence of oncologists nonverbal communication on breast cancer patients trust," Breast cancer research and treatment, vol. 153, no. 1, pp. 161-171, 2015, doi: 10.1007/s10549-015-3486-0.