STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–20.
Published online in International Academic Press (www.IAPress.org)

Enhancing Hotel Rating Predictions Through Machine Learning: Data Analytics Applications in Indian Hospitality and Digital Marketing

Sapna Kumari [1], M.T. Beig [2*], Mohammad Anas [1], Haresh Kumar Sharma [3]

[1] *Department of Mathematics, Faculty of Applied and Basic Sciences, Shree Guru Gobind Singh Tricentenary University, Gurugram 122505, India*
[2] *Department of Physics, Faculty of Applied and Basic Sciences, Shree Guru Gobind Singh Tricentenary University, Gurugram 122505, India*
[3] *Area of Operations and Decision Sciences, Birla Institute of Management Technology, Greater Noida - 201306, India*

Abstract With more consumers relying on online reviews, predicting hotel ratings accurately has become very important. This study investigates the use of machine learning models to predict overall hotel ratings based on key service-related features, including location, hospitality, cleanliness, facilities, food, value for money, and price. Using a real-world dataset of Indian hotels, we evaluate and compare the performance of six supervised learning models: Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, K-nearest neighbours, and PCA-based Linear Regression. The models were evaluated using Mean Squared Error (MSE) and R-squared (R^2) as performance metrics. Gradient Boosting demonstrated the highest predictive accuracy, closely followed by Random Forest. Feature importance analysis identified hospitality, cleanliness, and location as the most significant predictors of customer satisfaction. Principal Component Analysis (PCA) further reduced dimensionality while retaining over 90% of the dataset's variance within the first four components. These findings demonstrate the effectiveness of ensemble learning methods for hotel rating prediction and offer actionable insights for service improvement in the Indian hospitality sector. Furthermore, the results underscore the role of data-driven analytics in shaping effective digital marketing and promotional strategies tailored to diverse customer preferences.

Keywords: Machine Learning, Hotel Ratings, Linear Regression, Random Forest, K-Nearest Neighbors (KNN), Principal Component Analysis, Support Vector Machine, Gradient Boosting
AMS 2010 subject classifications 62J05, 62H25

DOI: 10.19139/soic-2310-5070-2848

1. Introduction

The rise of online communication has profoundly transformed the hospitality industry, particularly through travel websites where users routinely post reviews and rate their hotel experiences. These user-generated reviews serve as critical data points for travellers, offering authentic perspectives beyond traditional marketing materials. Ratings and reviews influence consumer decision-making and play a vital role in enhancing brand reputation, attracting new customers, and driving business profitability. For tourism businesses, effectively analysing this vast customer feedback is essential for maintaining competitive service standards. To this end, researchers have increasingly employed machine learning (ML) techniques to process and interpret hotel reviews. These models can identify key service quality indicators and predict overall hotel ratings, offering valuable insights for travellers and industry stakeholders. In a related vein, Benitez et al. [1] introduced a fuzzy multi-attribute decision-making approach to dynamically evaluate service quality across three Gran Canaria hotels. Their study contributes to

*Correspondence to: M.T. Beig (Email:mirzatanweer@gmail.com). Department of Physics, Faculty of Applied and Basic Sciences, Shree Guru Gobind Singh Tricentenary University, Gurgaon 122505, India.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

the hospitality and tourism marketing literature by applying fuzzy number theory to assess service performance and its impact on the competitive positioning of hotels in the vacation market. Chou et al. [2] developed a fuzzy multi-criteria decision-making (FMCDM) model to aid in selecting the optimal location for tourist hotels. Their framework offers a systematic and objective methodology for evaluating various location-based factors, thereby supporting more informed site selection decisions in the hospitality sector. Extending this decision-support approach, Tsai et al. [3] proposed an integrated model combining DEMATEL, Analytic Network Process (ANP), Zero-One Goal Programming (ZOGP), and Activity-Based Costing (ABC) to guide the selection of Corporate Social Responsibility (CSR) programs in international tourist hotels. Their findings emphasise the pivotal role of organisational image in achieving CSR objectives. Akincilar and Dagdeviren [4] introduced a hybrid evaluation model that combines the Analytic Hierarchy Process (AHP) with the Preference Ranking Organisation Method for Enrichment of Evaluations (PROMETHEE) to assess the quality of hotel and hospitality websites. This approach offers a structured assessment mechanism for enhancing online service delivery. Similarly, Gil-Lafuente et al. [5] employed the Fuzzy Delphi Method (FDM) alongside the Fuzzy Analytic Hierarchy Process (FAHP) to establish evaluation criteria for luxury resort hotels in Taiwan and Macao. Their comparative analysis revealed regional differences in strategic priorities- Taiwan emphasized consumer orientation and operational management, whereas Macao placed a greater focus solely on operational management. Masiero et al. [6] used a discrete choice modelling approach to investigate how much hotel guests are willing to pay for specific room features in a study centered on customer preferences. The results indicated significant valuation differences between leisure and business travellers and between first-time and repeat visitors. These insights can support targeted marketing strategies and revenue optimisation in the hotel industry. Chen [7] employed quantile regression to examine how inbound tourism growth influences Taiwanese hotel firms' sales performance and financial outcomes. The analysis revealed an asymmetric impact on hotel equity returns, with smaller hotels being more sensitive to tourism fluctuations, highlighting the need for tailored financial strategies within the sector. In a related area of sustainability, Mardani et al. [8] developed a hierarchical evaluation framework to assess and prioritise energy-saving technologies in large Iranian hotels. By integrating fuzzy Delphi, Fuzzy Analytic Hierarchy Process (FAHP), and fuzzy-based ranking techniques, the study offers a comprehensive decision-support tool for promoting energy efficiency in hospitality operations. Rianthong et al. [9] proposed a two-stage stochastic programming model that optimises hotel sequencing based on customer-generated reviews from online travel agencies (OTAs). The model aims to streamline the search process for users by intelligently ranking hotels, and its practical utility was validated using real OTA data. Focusing on service quality, Akbaba [10] investigated customer expectations in business hotels using the SERVQUAL framework. The study identified five core dimensions-tangibles, adequacy in service supply, understanding and caring, assurance, and convenience-and emphasised the importance of contextual adaptation for different service environments and cultural settings. These findings highlight the ongoing relevance of SERVQUAL while acknowledging its limitations in diverse hospitality contexts. Wang et al. [11] found that compatibility, firm size, technological competence, and the presence of a critical mass were positively associated with hotel adoption of Mobile Hotel Reservation Systems (MHRS). At the same time, system complexity had a significant adverse effect. These results highlight the importance of both organisational readiness and external market factors in driving technological adoption in the hospitality sector. Yu et al. [12] proposed a novel Multi-Attributive Border Approximation Area Comparison (MABAC) approach incorporating interval type-2 fuzzy sets to support multi-criteria decision-making. Their algorithm was validated through a hotel selection case study, demonstrating its utility in ranking alternatives under uncertainty. Lai et al. [13] utilised the Partial Least Squares (PLS) method to analyse satisfaction patterns among traveller segments in Macau luxury hotels. The study revealed that satisfaction drivers vary across new, repeat, and frequent guests, providing nuanced insights into evolving customer expectations and segment-specific service strategies. In a survey related to pandemic responses, Lee et al. [14] applied Semantic Network Analysis (SNA) to online reviews of quarantine hotels, identifying patrons' service quality perceptions during crises. Their work offers guidance on how hotels can adapt communication and service delivery under emergency conditions. Wu et al. [15] explored hotel service robots' psychological and experiential dimensions, focusing on attributes such as anthropomorphism and perceived intelligence. Using a hybrid Structural Equation Modelling-Artificial Neural Network (SEM-ANN) approach, the study found that service and brand authenticity are critical in fostering customer affection, termed "brand love", for robotic service offerings in hospitality. Viglia et al. [16] examined

the effects of price adjustments by competing hotels on consumers' reference prices through laboratory and field experiments. Their findings indicate that synchronised pricing changes among hotels can lower consumers' perceived price baselines, influencing booking behaviour-a key consideration for online travel agencies that display dynamic rate comparisons. Finally, Boo et al. [17] investigated the influence of Customer Social Marketing (CSM) initiatives on consumer behaviour during the COVID-19 pandemic. The study concluded that well-executed CSM campaigns effectively altered customer attitudes and behaviours, underlining the strategic value of social engagement during periods of uncertainty.

Cruz et al. [18] investigated the relationship between hotel management structures and guest satisfaction. Their findings revealed that while chain-operated hotels typically achieve the highest satisfaction levels, owner-operated hotels perform better in the economy and midscale segments, suggesting different operational strengths based on hotel type. Fang et al. [19] developed the Hotel-Guest-Robot Interaction Experience (HGRIE) scale to quantify guest interactions with service robots. This scale provides a framework for improving robot management and enhancing service experiences in technologically augmented hospitality environments. Lee et al. [20] proposed a method for analysing energy consumption patterns in hotel guestrooms to promote occupant-specific appliance control. Their work supports initiatives for cost reduction and green certification within the hotel industry. Lim et al. [21] assessed the impact of augmented and virtual reality (AR/VR) on tourist satisfaction. They found that perceived ease of use, innovativeness, and overall usefulness significantly enhance guest satisfaction and increase the likelihood of bookings. Nakamura et al. [22] explored the influence of Airbnb listings on hotel occupancy rates in Japan. Contrary to common assumptions, their results showed that fluctuations in Airbnb supply do not significantly disrupt traditional hotel occupancy, suggesting market segmentation between accommodation types. Oukil et al. [23] introduced a two-stage hybrid Data Envelopment Analysis (DEA) model for identifying optimal hotel pairings to enhance sector-wide performance and benchmarking.

Saez et al. [24] emphasised the growing importance of digital innovation in the hotel industry. Analysing 322 Spanish hotels using Structural Equation Modelling (SEM), they demonstrated how digitalisation positively affects competitiveness and operational efficiency. Shehawy et al. [25] conducted a cross-national study examining consumers' willingness to pay more for green hotels. Using the Theory of Planned Behaviour (TPB), they identified notable variations in eco-conscious behaviour across seven countries, reinforcing the global relevance of sustainable hotel practices.

Yu et al. [26] explored the emerging concept of vegan hotels, identifying six key guest-attracting attributes: health, guilt, social ethics, environmental concern, religion, and curiosity. These factors significantly enhanced perceived well-being, enjoyment, and behavioural intentions such as word-of-mouth and repeat bookings. Building on this diverse body of research, our study applies machine learning techniques for predicting hotel ratings based on key service attributes. While prior literature has concentrated mainly on qualitative assessments and structured decision models, there is a clear need for data-driven predictive frameworks that can harness numerical review data to guide service optimization. In this context, a robust predictive framework based on machine learning is essential for accurately analysing location, hospitality, cleanliness, facilities, value for money, food, and price-the core dimensions that shape customer satisfaction and influence rating behaviour. To address this gap, we evaluate the predictive power of several machine learning algorithms-including Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, K-Nearest Neighbours, and PCA-based Linear Regression-in modelling overall hotel ratings. By comparing model performance and identifying the most influential predictors, this study contributes actionable insights for hospitality professionals and underscores the value of machine learning in customer experience management. This paper explores using various machine learning models – Linear Regression, Random Forest, Gradient Boosting, Support Vector Machines, K-nearest Neighbour, and PCA-based Linear Regression to predict overall hotel ratings based on a feature set. The study assesses the accuracy of these methods to evaluate their performance. The rest of the paper is organized as follows. Section 2 describes the dataset. Section 3 presents the dataset's basic statistics and proposed system is introduced. Section 4 details the machine learning techniques used. Section 5 evaluates the performance of various machine learning models. Section 6 presents and compares the outcomes of the study. Finally, Section 7 presents the conclusions of the study and outlines potential directions for future work.

2. Dataset Description

The dataset used in this study was collected from MakeMyTrip [27], an online travel agency platform. It contains 609 entries and eight columns representing various hotel features and ratings. These include location, hospitality, facilities, cleanliness, value for money, food, price, and overall rating. The data types within these columns are predominantly 'float64', indicating decimal or continuous values, except for the 'price' column, which is of type 'int64' and signifies integer values. Our dataset contains no missing values, encompasses eight variables, and the final column represents the target variable. The first seven variables serve as independent features reflecting customer perceptions and hotel characteristics, while 'overall rating' is the target variable for predictive modelling. All entries are complete, with no missing values, ensuring the dataset's suitability for supervised machine learning applications.

3. Basic Statistics

We will utilise the hotel rating dataset, which comprises eight attributes: 'location' , 'hospitality' , 'facilities' , 'cleanliness', 'value for money' , 'food' , 'price' and 'overall rating'. Our objective is to extract meaningful information on hotel ratings. Understanding the dataset characteristics and the variability in ratings is essential for businesses and researchers in the hospitality industry, as it aids in comprehending customer preferences and facilitates data-driven decision-making.

3.1. Sample of the Subset Data Frame

Table 1 shows a snapshot of the data. We show the first five entries for columns like 'overall rating' , 'value for money' , 'hospitality' , 'cleanliness' and 'location'. This sample provides an overview of typical rating distributions. For example, the first row shows a hotel with an 'overall rating' of 4.3, 'value for money' at 4.2, 'hospitality' at 4.2, 'cleanliness' at 3.7, and 'location' at 4.7. These results give us a basic idea of the dataset and how hotel ratings are spread out. They help us see what the data is like and how different aspects are rated. This info is a good starting point for digging deeper and discovering more about the data. It helps users make better decisions and understand the data better.

Table 1. Sample of the subset data frame (data snippet)

Overall rating	Value for money	Hospitality	Cleanliness	Location
4.3	4.2	4.2	3.7	4.7
4.3	4.1	4.2	4.1	4.6
4.7	4.7	4.7	3.9	4.7
4.9	4.9	4.9	4.5	4.9
4.8	4.8	4.8	4.6	4.8

3.2. Descriptive Statistics

To understand the dataset better, we looked at hotel reviews. Table 2 shows the explanation of statistical analysis. We focused on five important columns: 'overall rating,' 'value for money,' 'hospitality,' 'cleanliness,' and 'location.' This analysis showed that there are 609 hotel ratings in the dataset, which you can see from the count in the 'overall rating' column. The mean is about 4.01, so most hotels fall around that score. The standard deviation was 0.47, so there is definitely some variance in the scores. Ratings range from a low of 1.70 to a high of 5.00 We observe that there's quite a bit of spread there. We extended this analysis to other key variables such as 'value for money', 'hospitality', 'cleanliness' and 'location. This analysis helps us comprehend which different areas the customers are really paying attention to when rating those areas.'

Table 2. Statistical Analysis

	Overall rating	Value for money	Hospitality	Cleanliness	Location
Count	609	609	609	609	609
Mean	4.01	3.93	3.86	3.93	4.19
Standard Deviation	0.47	0.53	0.58	0.53	0.49
Minimum	1.70	1.50	1.50	1.50	1.40
25%	3.80	3.70	3.60	3.70	4.00
50%	4.00	4.00	3.90	4.00	4.30
75%	4.30	4.30	4.20	4.30	4.50
Maximum	5.00	5.00	5.00	5.00	5.00

The ‘mean’ (average) is a central measure that provides insight into the typical or expected value. It is a reference point for hotel ratings in the aspects analysed. The ‘standard deviation’ quantifies the level of variability or dispersion in the data. A higher standard deviation suggests that ratings in a particular aspect are more spread out from the mean. We can gain a more nuanced understanding of the hotel ratings dataset by performing a statistical analysis. For instance, the average overall rating is approximately 4.01, indicating a generally positive sentiment among reviewers. The standard deviation of approximately 0.47 suggests moderate variability, implying that hotel ratings vary around the mean. While the minimum and maximum ratings indicate the full spectrum of experiences, the percentiles reveal the distribution of ratings.

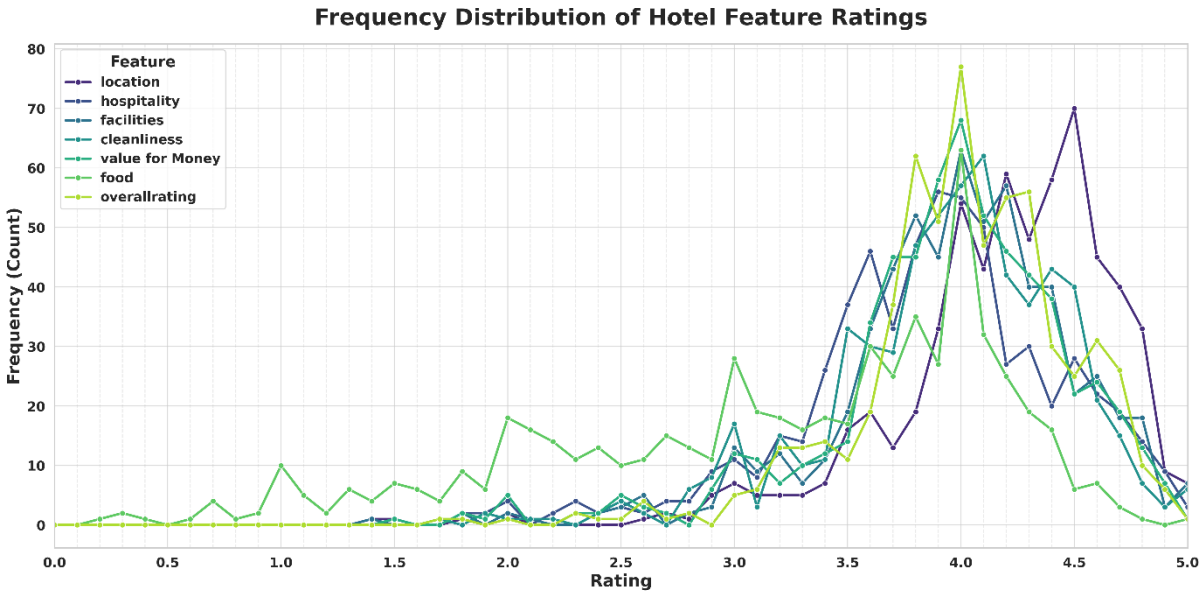

Figure 1. Frequency Distribution of Hotel Feature Ratings.

The line plot Figure 1 visualises the count of user ratings for each hotel feature (e.g., location, food, cleanliness) across the full rating scale. Most features cluster around the 3.5-4.5 range, indicating generally favourable reviews. Notably, location and hospitality exhibit tighter, right-skewed peaks, suggesting more consistent positive experiences. Food shows a wider spread, indicating varied guest perceptions.

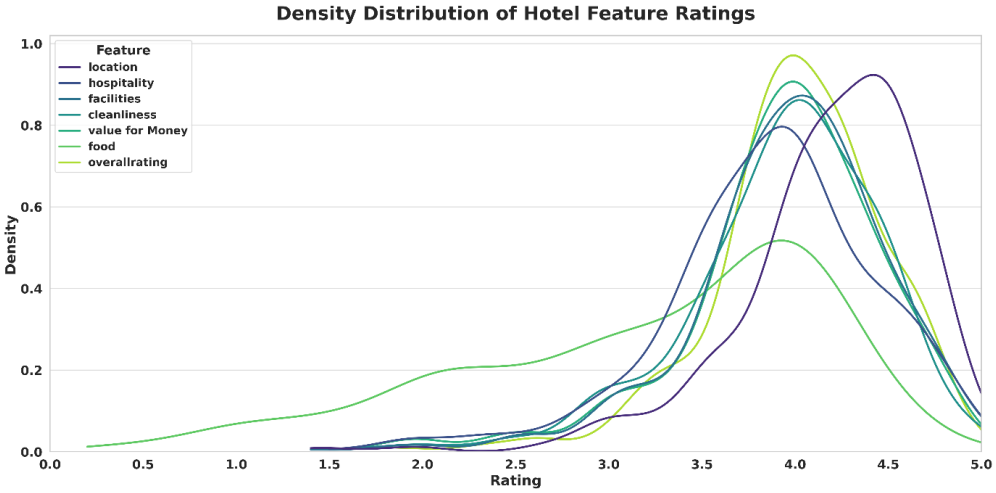

Figure 2. Density Distribution of Hotel Feature Ratings.

The kernel density plot displays the normalized distribution of ratings for various hotel features. The sharper peaks near 4.0 highlight consistent satisfaction with services like location and hospitality. Broader or flatter curves (e.g., food) indicate more diverse opinions. These distributions help assess variability and central tendency in guest experiences.

3.3. Correlation Heatmap

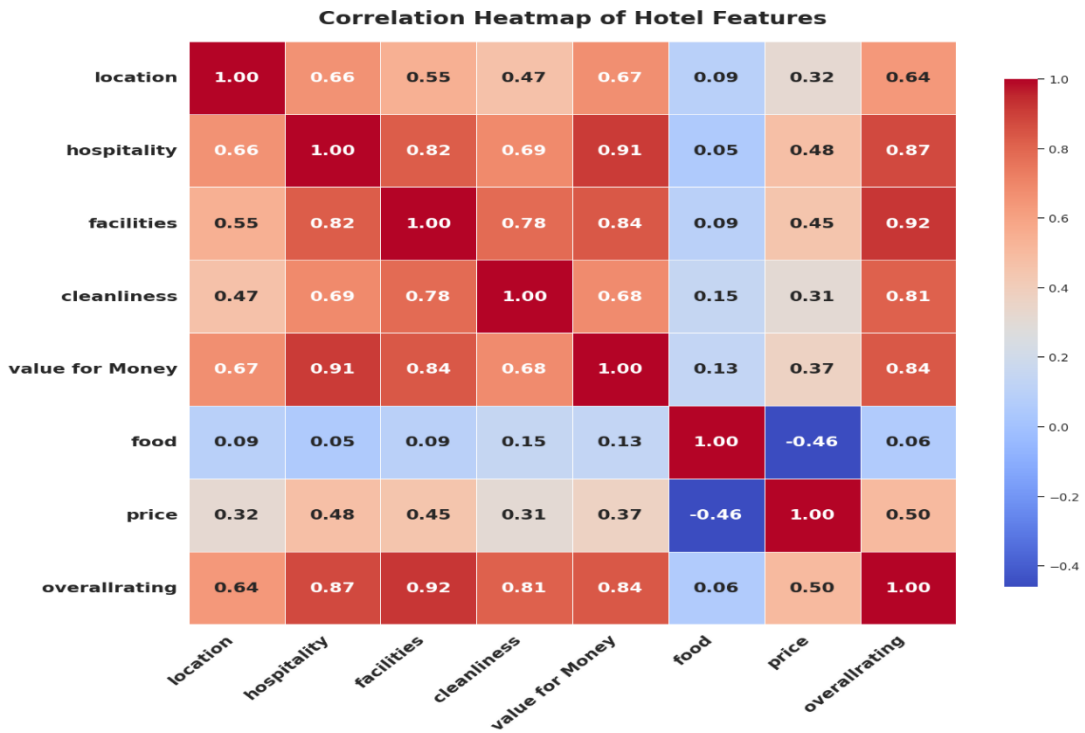

Figure 3. Correlation Heatmap.

To understand the interrelationships among hotel attributes, a Pearson correlation heatmap was generated. The matrix displays coefficients ranging from -1 to +1, where values closer to +1 indicate strong positive linear relationships, values near -1 suggest strong inverse relationships, and values around 0 imply negligible correlation. Figure 3 illustrates the correlation heatmap, capturing the linear relationships among various hotel features and the overall rating. Interestingly, 'food' has a weak correlation with the overall rating (r = 0.06), indicating it is not a primary driver of customer satisfaction in this dataset. This suggests food is either subjective or consistently rated, contributing little to rating variance. Among inter-feature relationships, 'hospitality' is highly correlated with 'value for money' (r = 0.91) and 'facilities' (r = 0.82), indicating these aspects often co-occur in positive guest experiences. Similarly, 'cleanliness' and 'facilities' (r = 0.78) are strongly related, suggesting that well-equipped hotels also tend to maintain better hygiene. A moderate negative correlation between 'food' and 'price' (r = -0.46) is observed, suggesting a perception gap: customers paying higher prices may not perceive corresponding improvements in food quality. These insights emphasize that facilities, hospitality, value for money, and cleanliness are the most predictive features of a hotel's overall rating. In contrast, food appears to have minimal influence and could be deprioritized in predictive modelling. Price and location, while not dominant, still offer complementary value and should be retained as supporting features in regression and classification models.

3.4. Top 3 most highly correlated feature pairs plot

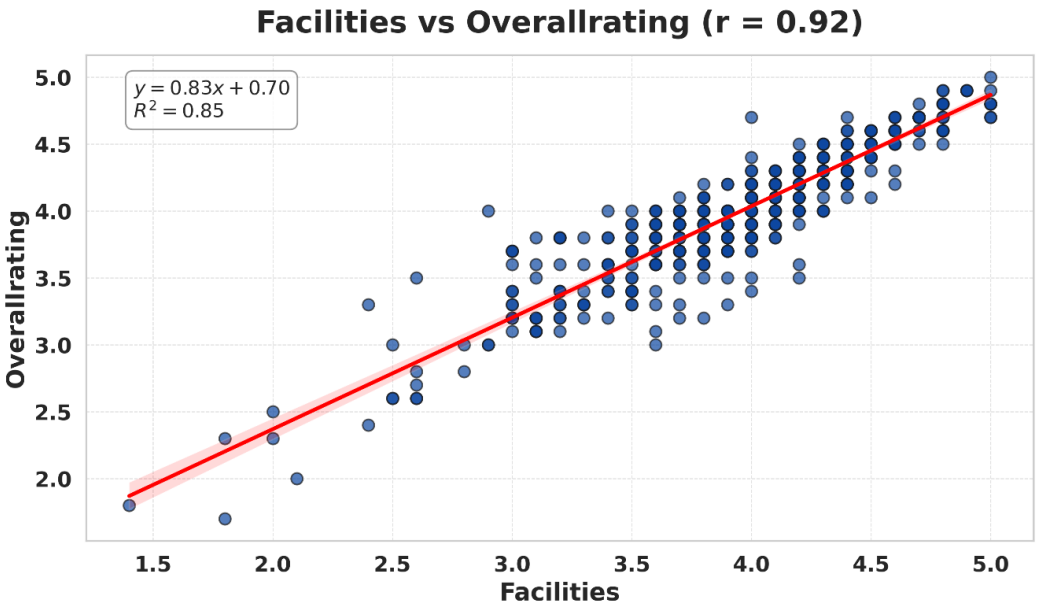

Figure 4. Facilities vs Overall Rating (r = 0.92)

This scatter plot shows a strong positive linear relationship between hotel facilities and the overall rating. It implies that better facilities significantly contribute to higher guest satisfaction scores.

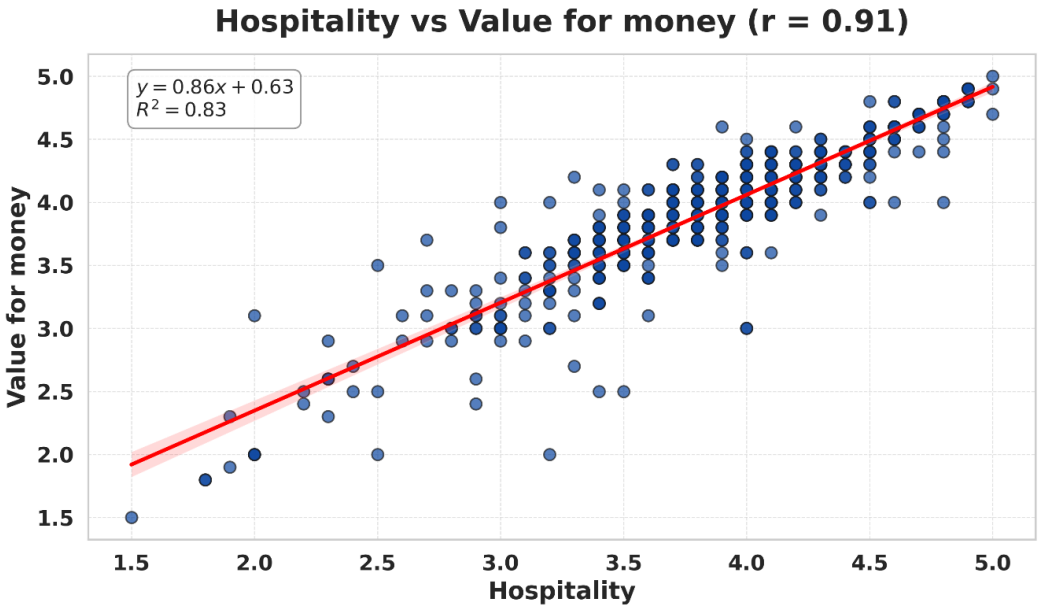

Figure 5. Hospitality vs Value for Money (r = 0.91)

This plot illustrates a high correlation indicating that hotels with better hospitality are also perceived as offering greater value for money, reinforcing the interconnectedness of service quality and perceived cost-effectiveness.

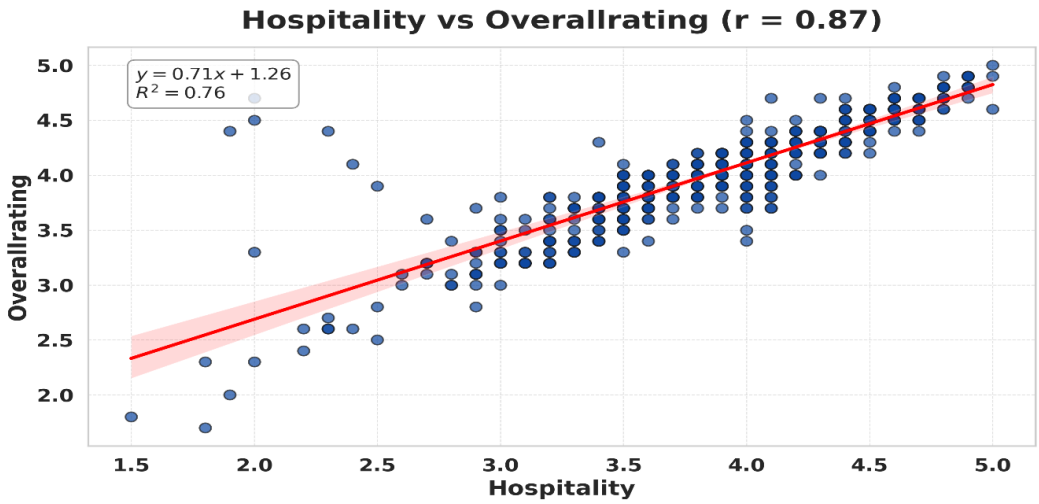

Figure 6. Hospitality vs Overall Rating (r = 0.87)

The graph depicts a strong linear trend, suggesting that improved hospitality is directly linked to higher overall ratings.

3.5. Proposed Methodology

The dataset attributes and samples have been pre-processed to eliminate null values, preparing them for the application of various machine learning algorithms for predictive purposes.

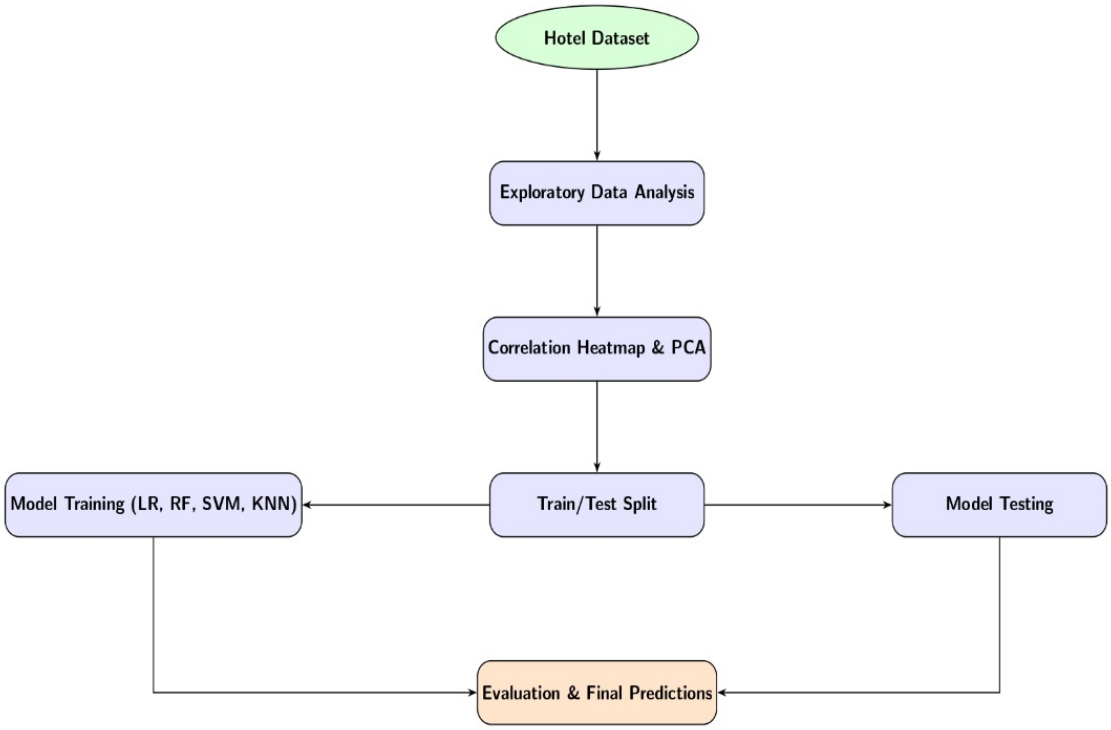

Figure 7. A model flow chart for hotel rating predictions

4. Classification Using Various Algorithms

In this study, a range of machine learning techniques were employed, including Linear Regression, Random Forest, Gradient Boosting, Support Vector Machines, K-nearest neighbour, and PCA-based Linear Regression for classification models. These models received input attributes from a dataset divided into two segments: 70% of the data sample was used for training the model, while the remaining 30% served as a test dataset. The training dataset refers to the portion of data specifically used to train the machine learning model. Principal Component Analysis (PCA) was used for dimensionality reduction [28]. The k-Nearest Neighbors (kNN) algorithm is widely used for clustering and classification tasks. In particular, the paper by Begum et al. [29] demonstrates how kNN, combined with feature selection methods, can improve classification performance on benchmark datasets. The hybrid SVM-SVR model for build-up rate prediction, as proposed in the paper by Wang et al. [30], demonstrates significant improvement in prediction accuracy compared to traditional models.

4.1. Linear Regression

Linear regression is a fundamental technique in both statistics and machine learning, primarily used for modelling the linear relationship between a dependent (target) variable and one or more independent (predictor) variables. As a type of supervised machine learning algorithm, its main objective is to find the best-fit line that minimizes the error between the predicted and actual values. From the independent characteristics provided, the outcome of the dependent variable is ascertained. Consequently, it is crucial to underline that linear regression basically focuses

on determining and quantifying the direct proportionality between variables, so as to produce correct predictions and relevant data interpretations.

4.2. Random Forest

Random Forest is a widely used machine learning algorithm known for its effectiveness in both regression and classification tasks. It operates by constructing an ensemble of decision trees, each trained independently on different randomly selected subsets of the data and features. This approach enhances predictive accuracy and reduces the risk of overfitting. Random Forest is also recognized as a powerful feature selection tool, capable of identifying the most important variables that contribute to predictions. Additionally, it supports dimensionality reduction by evaluating feature importance across multiple trees. The implementation of Random Forest involves two key steps: first, building the ensemble of decision trees; and second, using the aggregated output of these trees to make accurate predictions.

4.3. Gradient Boosting

Gradient boosting is a well-known machine learning method applied successfully for both classification and regression problems. Unlike Random Forest's concurrent tree-building method, Gradient Boosting creates its trees one after the other. Its great precision and deft handling of difficult datasets depend much on this methodological difference. However, maximising Gradient Boosting's performance and accuracy depends on rigorous evaluation of the necessary computational resources and meticulous optimisation of its parameters, both of which are absolutely necessary for proper application.

4.4. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm primarily used for classification tasks, though it is also well-suited for regression problems (referred to as Support Vector Regression, or SVR). It is particularly effective in high-dimensional spaces and can handle both linear and non-linear data efficiently. The core idea behind SVM is to identify the optimal hyperplane that best separates data points of different classes in the feature space. This hyperplane maximizes the margin between the closest data points of the classes, ensuring better generalization on unseen data. In regression tasks, SVM aims to find a function that deviates from the actual target values by a margin no greater than a specified threshold, while remaining as flat as possible.

4.5. K-nearest neighbour

This algorithm is widely applied for both classification and regression. K-nearest neighbour is a flexible and basic machine learning method. It uses a distance metric, either Euclidean or Manhattan distance, to find the 'k' closest data points or neighbours near a query point. While for regression, KNN forecasts a value based on their average, in classification problems, it finds the most common class among these neighbours. The performance of the method depends much on the distance metric and the choice of "k". KNN is sensitive to irrelevant features, particularly in high-dimensional spaces. Its performance also declines with big datasets, although it is simple to apply and understand.

4.6. Principal Component Analysis

A sophisticated technique meant to lower the dimensionality of difficult datasets while keeping important information is PCA. This method uses a mathematical construction of Principal Components by means of a smaller selection of uncorrelated variables from the decomposition of a large set of correlated variables. The main aim of PCA is to find a lower-dimensional representation of the original dataset. The first main component is created to capture the maximum variation feasible; each succeeding component is made to have the same variance, provided it is orthogonal to the previous component. This methodical technique for dimension reduction is quite helpful for simplifying difficult data and guarantees that important features are maintained for efficient study.

5. Performance Evaluation of Machine Learning Models

5.1. Mean Squared Error (MSE)

MSE is a metric used to gauge the accuracy of a model in predicting quantitative data. It represents the average of the squares of the errors, i.e., the difference between the actual and predicted values

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Where n is the number of observations, Y_i is the actual value for the i^{th} observation, and $\hat{Y}_i$ is the predicted value for the i^{th} observation.

5.2. Coefficient of Determination (R^2) for Model Evaluation

The R^2 statistic is a key metric for evaluating the goodness of fit of a regression model. It quantifies the proportion of variance in the dependent variable that can be predictable from the independent variables. This measure helps in understanding the effectiveness of the model in explaining the variability of the data.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Where, SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

5.3. Confusion Matrix

A confusion matrix in Table 3 has explained machine learning and classification tasks, helping to assess how well a model performs. It organizes predictions and actual outcomes into a structured matrix, making understanding the model's behaviour easy.

The elements of a confusion matrix are as:

1. **True Positives (TP):** These are the cases where the model correctly predicts a positive data point.
2. **True Negatives (TN):** These are the cases where the model correctly predicts a negative data point.
3. **False Positives (FP):** These are the cases where the model incorrectly predicts a positive data point.
4. **False Negatives (FN):** These are the cases where the model incorrectly predicts a negative data point.

Table 3. Confusion Matrix

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

6. Results

6.1. Distribution of hotel prices within the dataset

Table 4. Summary Statistics of Hotel Prices

Count	Mean	Std	Min	25%	50%	75%	Max
609	4294.84	4612.41	500.00	1250.00	2000.00	6138.00	29296.00

The summary statistics for the 'price' feature in the dataset provide a comprehensive overview of the distribution of hotel prices. Figure 8 illustrates this distribution. With a total count of 609 entries, the dataset contains 609 distinct hotel prices available for analysis. The mean (average) hotel price is approximately ₹4,294.85, indicating that, on average, hotel accommodations in the dataset are priced around this amount. The standard deviation is about ₹4,612.41, reflecting significant variability in hotel prices-suggesting a broad spectrum of options ranging from budget to luxury hotels. The minimum price is ₹500, representing the most affordable hotel option in the dataset. At the lower end, the 25th percentile (Q1) is ₹1,250, meaning that 25% of hotels are priced below this value. The median (50th percentile) is ₹2,000, indicating that half of the hotels cost less than this amount and the other half more, serving as a central reference point. At ₹6,136.8, the 75th percentile (Q3) indicates that 75% of hotels are priced below this level, with the remaining 25% being higher-priced properties. The maximum price, recorded at ₹29,296, represents the most expensive hotel in the dataset, highlighting the availability of high-end luxury options. Overall, these statistics reveal a wide and diverse pricing distribution, ranging from economical stays to premium accommodations. While the mean price is ₹4,294.85, the spread from ₹500 to ₹29,296, along with the quartile values, offers deeper insights into the segmentation of hotel prices and underscores the variability present in the hospitality market.

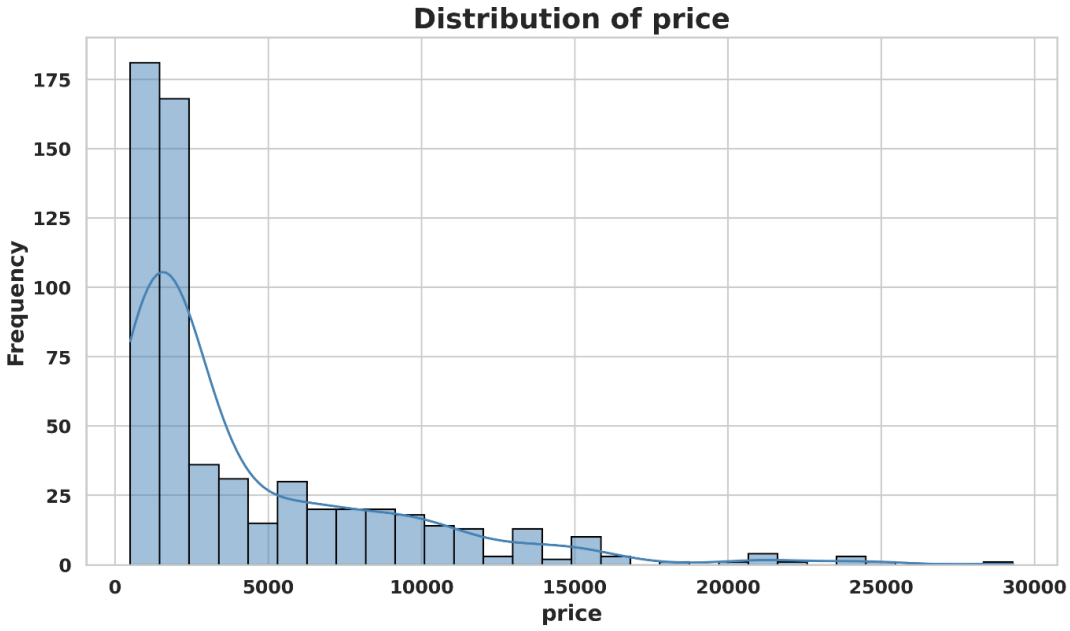

Figure 8. Distribution of Hotel Price

6.2. Analysis of the relationship between hotel prices and hospitality levels

The box plot in Figure 9 provides a detailed visualization of hotel price distributions across different levels of hospitality ratings. It effectively highlights how hotel pricing correlates with varying levels of perceived hospitality, offering insights into market segmentation and pricing strategies. From the plot, a clear upward trend is observed: as the hospitality rating increases, hotel prices generally rise. Hotels with lower hospitality scores (e.g., 1.5 to 3.0) tend to have lower and more tightly clustered prices, often ranging between ₹500 and ₹5,000. In contrast, higher-rated hotels (4.5 to 5.0) exhibit both higher median prices and greater variability, with prices extending up to ₹24,395, as marked on the chart. For instance, the median price for hotels rated around 5.0 is approximately ₹21,490, while their 25th percentile (Q1) lies at ₹17,521, suggesting that 25% of these top-rated hotels charge at or below this threshold. This reflects a more premium pricing strategy and potentially greater service offerings. In contrast, hotels with ratings around 2.0 to 3.0 show lower medians and a narrower interquartile range, reflecting standardized,

economy-class pricing. Additionally, the presence of numerous outliers, particularly at higher ratings, indicates occasional high-priced luxury properties within otherwise similarly rated groups. The variability (as shown by the height of the boxes and whiskers) also increases with rating, suggesting that as hospitality improves, the range of pricing options broadens. In summary, this box plot demonstrates a strong positive relationship between hospitality rating and hotel price. It underscores how service quality influences pricing, and provides valuable information for targeted pricing strategies, consumer segmentation, and competitive market positioning. Understanding these patterns allows hotel businesses to align their pricing with perceived service value, optimize revenue, and cater to varied customer preferences in a competitive market landscape.

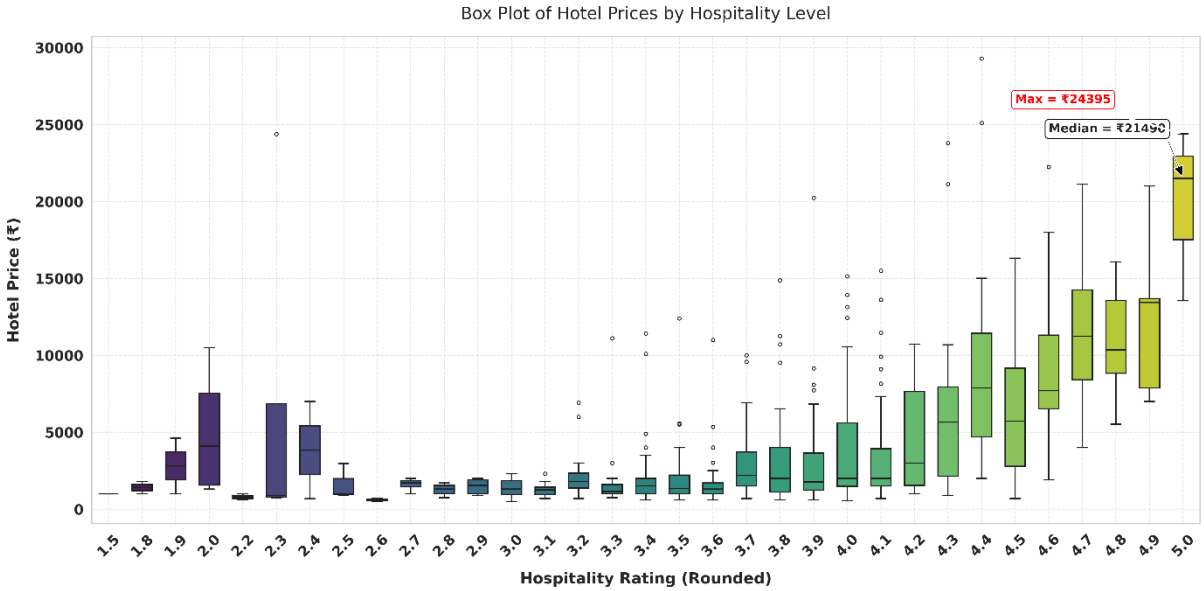

Figure 9. Box plot for all the variables

6.3. Preprocessing and Feature Extraction

We used PCA on this research paper in order to analyse how various aspects of satisfaction and general rating combine in the field of hotels. The dataset we utilized was from different hotels that encompassed key categories such as “location,” “hospitality,” “facilities,” “cleanliness,” “value for money,” “food,” and “price.” Each of these attributes contributes toward determining the final score assigned to any given hotel; however, we aimed at differentiating their significance. Following preprocessing of the data and standardising of the features to make sure each attribute was on the same scale, we then used PCA in order to uncover the hidden structure and dependencies. Our analysis revealed that the five most important principal components combined for most of the variation in the dataset. Notably, the results indicated that ‘value for money’, ‘cleanliness’, ‘food’, ‘hospitality’ and ‘location’ consistently appeared as significant variables within all principal components. Moreover, the table below shows the explained variance ratios for all principal constituents. As predicted, the first principal component has the greatest contribution to the total variance of the dataset, with the second principal component following it, then the third and so on. With the fourth principal component, more than 92% of the total variance in a data set has been retained according to the cumulative explained variance ratio. This measurable insight allows data scientists and analysts to make educated choices about how many principal components (Table 5) they should keep, depending on how much variance is explained.

Table 5. Explained Variance by Principal Components

	Principal Component	Explained Variance Ratio	Cumulative Explained Variance Ratio
0	1	0.626	0.626
1	2	0.176	0.802
2	3	0.075	0.877
3	4	0.045	0.923
4	5	0.040	0.963
5	6	0.020	0.983
6	7	0.012	0.994

These findings illustrate how important these five qualities are when it comes to evaluating how satisfied customers feel with specific hotels. 'Value for Money' deals with economic issues relating to staying at a hotel where cost is essential, while also considering what one perceives as valuable. 'Cleanliness' and 'food' touch upon hygiene levels and food quality, respectively, both leading contributors to guest satisfaction rates. 'Hospitality' entails various forms of guest service delivery and staff friendliness that imply a sense of being home away from home; hence, this ensures unforgettable experiences. Ultimately, location stands out as a key factor, reminding us of a desirable locality, which translates into overall scores.

Through identifying these crucial elements, hotels can enhance their strategies and pay attention to vital areas that significantly influence the guest experience. By stressing these features in their provisions and promotional activities hotels will improve customer satisfaction as well as overall ratings, thus acquiring an advantage over others in the ever-changing hospitality industry.

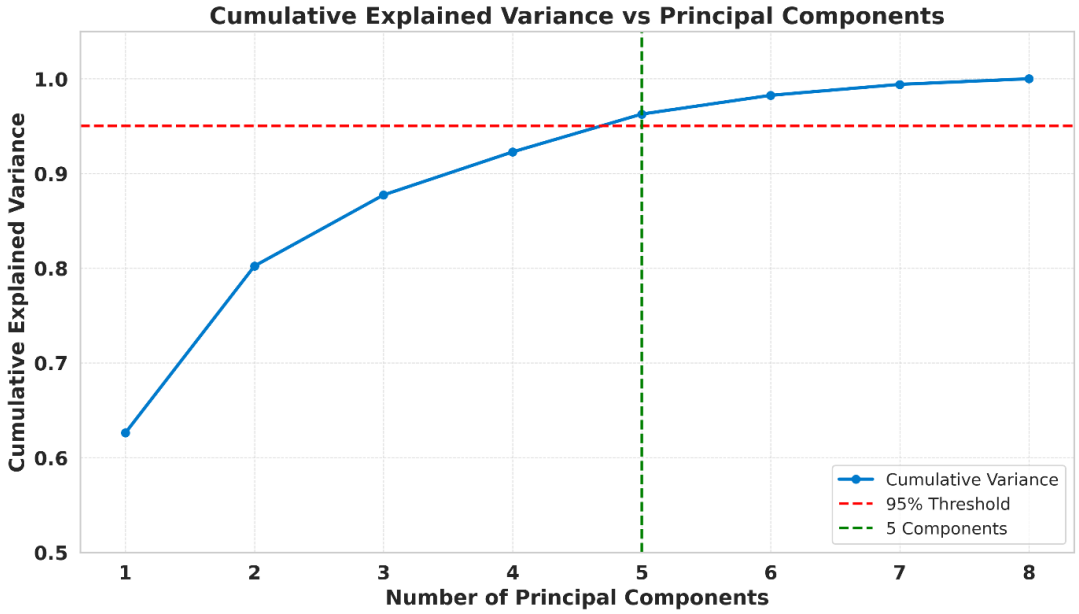

Figure 10. Cumulative Explained Variance Ratio

The most important part of this PCA evaluation is the cumulative explained variance ratio that is shown in a figure. This graph shows a good aspect of the bargaining between dimensionality and proportion of variance explained by data. The 'x' axis depicts principal components while the 'y' axis indicates the cumulative explained variance ratio. Cumulative explained variances are determined progressively where each successive component is added to it. The illustration gives visual guidance about how many principal components are needed for a given

proportion of the dataset's total variance to be kept constant. An in-depth analysis in Figure 10 shows that the first principal components account for a large share of. In addition, after the first few components, the slope of the graph starts becoming less steep, which signifies that the explanatory power for more variance decreases in value. For instance, for this particular data set, it has been noted that about 95% of the total variance is captured by the first four principal components taken together. Therefore, we could retain only these first four principal components and preserve most of their information, thereby possibly reducing their dimensionality. In conclusion, the PCA analysis is a necessary advance in data processing and dimensionality reduction. It allows the researcher to choose which principal components wisely. The components are simplified yet retain their most important aspects by knowing how much variance is left. Such a reduction in dimensionality becomes particularly valuable when a dataset has many features or when one needs to compress data for various statistics and machine learning applications.

6.4. Performance Comparison

For example, our first model, Linear Regression, had a MSE of approximately 0.015 and a reasonably high R^2 value of 0.930, indicating a good linear fit to the data. In contrast, Random Forest and Gradient Boosting models performed better than Linear Regression, with lower MSE values of 0.010 and 0.009, respectively. Their R^2 scores, at 0.950 each, highlight their superior predictive performance. This suggests that both ensemble methods provide stronger predictive power, with Gradient Boosting slightly edging out as the most accurate model overall. On the other hand, Support Vector Regression (SVR) and K-Nearest Neighbours (KNN) delivered comparatively poor performance. SVR had the highest MSE at 0.150 and the lowest R^2 at 0.290, indicating it failed to capture key data relationships. KNN followed a similar trend, with an MSE of 0.110 and an R^2 of 0.480, reflecting weak generalization ability. Additionally, we explored the impact of dimensionality reduction by applying Principal Component Analysis (PCA) followed by Linear Regression. However, the PCA-based Linear Regression model recorded a higher MSE of 0.020 and a lower R^2 of 0.890 compared to the ensemble models. This suggests that while PCA helped reduce input complexity, it may not be the best approach for preserving predictive accuracy in this dataset.

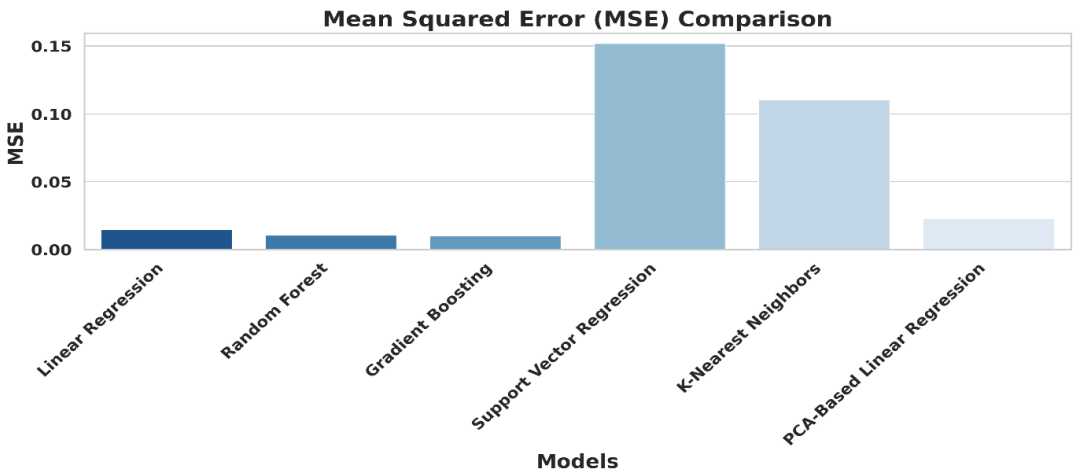

Figure 11. MSE correction of different ML models

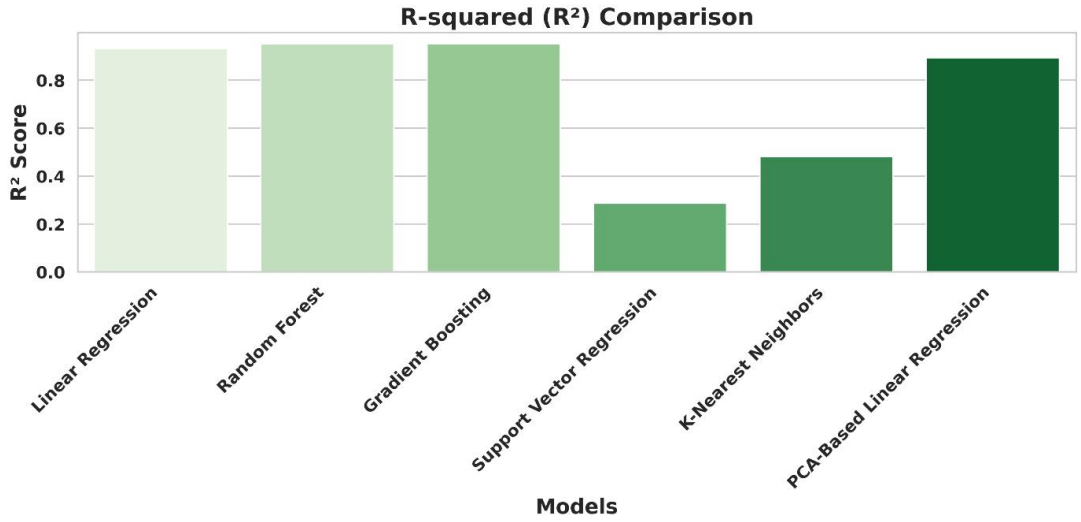

Figure 12. R^2 error correction of different ML models

Table 6. Performance Comparison of Regression Models Using MSE and R^2 Metrics

Models	MSE	R^2
Linear Regression	0.015	0.930
Random Forest	0.010	0.950
Gradient Boosting	0.009	0.950
Support Vector Regression	0.150	0.290
K-nearest Neighbours	0.110	0.500
PCA-based Linear Regression	0.020	0.890

Figures 11 and 12 compares the performance of six regression models using both MSE and R^2 metrics. The values are listed in Table 6. The light blue bars represent MSE-where lower values indicate better predictive accuracy-and green bars show R^2 scores-where higher values indicate a better model fit.

Among the models, Random Forest and Gradient Boosting clearly outperform the others, achieving the lowest MSE (0.01 and 0.009) and highest R^2 scores (0.950). Linear Regression also performs well, with an R^2 of 0.930 and low MSE of 0.015. In contrast, Support Vector Regression and K-Nearest Neighbors exhibit higher MSE values (0.150 and 0.110, respectively) and notably lower R^2 scores (0.290 and 0.480), suggesting weaker predictive capabilities. These comparisons highlight the superior performance of ensemble methods-especially Random Forest and Gradient Boosting-for predicting hotel ratings. The results also demonstrate that model selection plays a crucial role in accuracy, and not all algorithms perform equally well for the same dataset.

6.5. Model Performance Evaluation using MSE

The Random Forest Regressor stands out with an MSE of 0.010 and an R^2 of 0.95, indicating that it accurately captures the relationship between hotel features and overall ratings. It explains 95% of the variance in the ratings using factors like location, hospitality, and cleanliness. This high explanatory power confirms the model's robustness and predictive reliability.

6.6. Prediction Error Plot for Random Forest Regressor with $R^2 = 0.929$ Showing High Model Accuracy

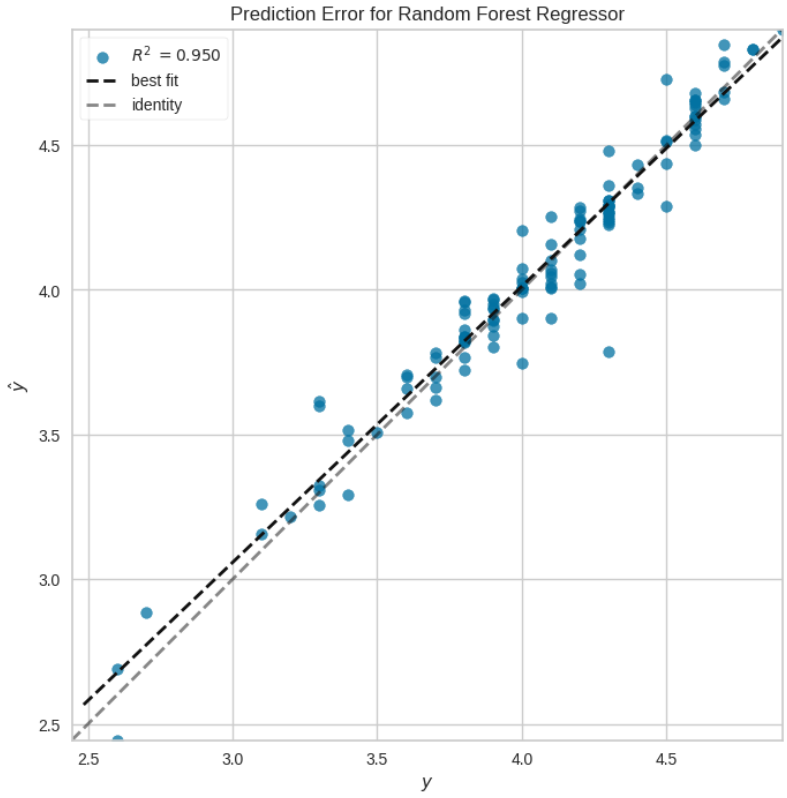

Figure 13. Prediction Error Plot for Random Forest Regressor

Figure 13 gives the prediction error plot for the Random Forest Regressor shows actual hotel ratings (x-axis) versus predicted ratings (y-axis), with most points closely aligned along the ideal diagonal. The high R^2 value of 0.929 indicates that 92.9% of the variation in ratings is well captured by the model. The proximity of the best-fit line to the identity line confirms the model's strong predictive accuracy.

6.7. Decision Tree Values

This information can be found in the "Decision Tree Values" part and highlights values associated with leaf nodes of the first decision tree in Random Forest. Every individual value corresponds to the mean overall rate for a certain subset of data that went through that specific leaf node. Such principles provide insight into how the data was divided by decision tree according to feature values; In contrast, when predictions are made with a Random Forest model then predictions coming from different decision trees are summed up together where its first decision tree structure and parameters offer an idea about contribution of each single tree to the final prediction made by the ensemble. Therefore, given the suggested characteristics, the Random Forest Regressor seems to be an excellent model for predicting general hotel scores as it has low MSE and high R^2. Moreover, by interpreting Decision Tree Values, also shown in Figure 14 and Figure 15, it is possible to see what this model does with data segments, thus facilitating the identification of predictors that are crucial when determining overall hotel scores.

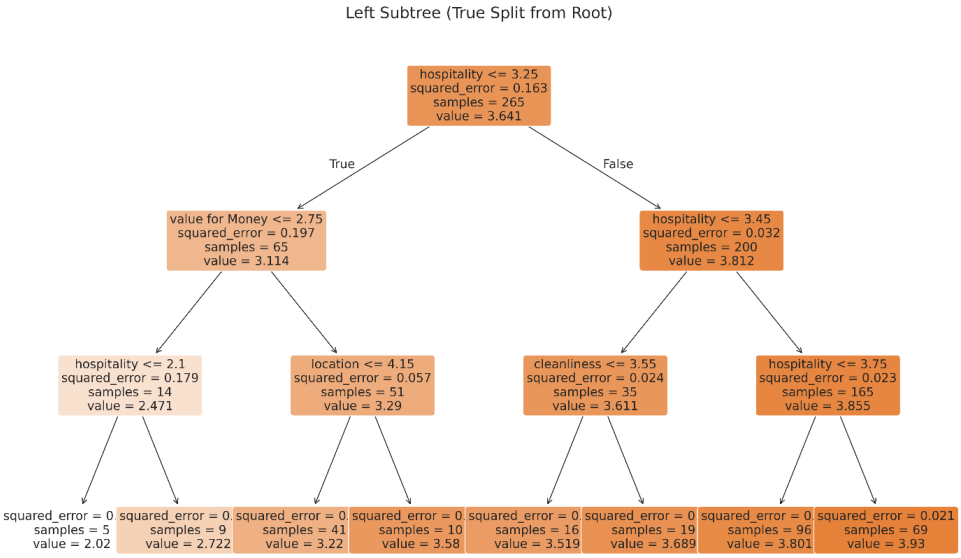

Figure 14. Left subtree for Decision tree for hotel rating

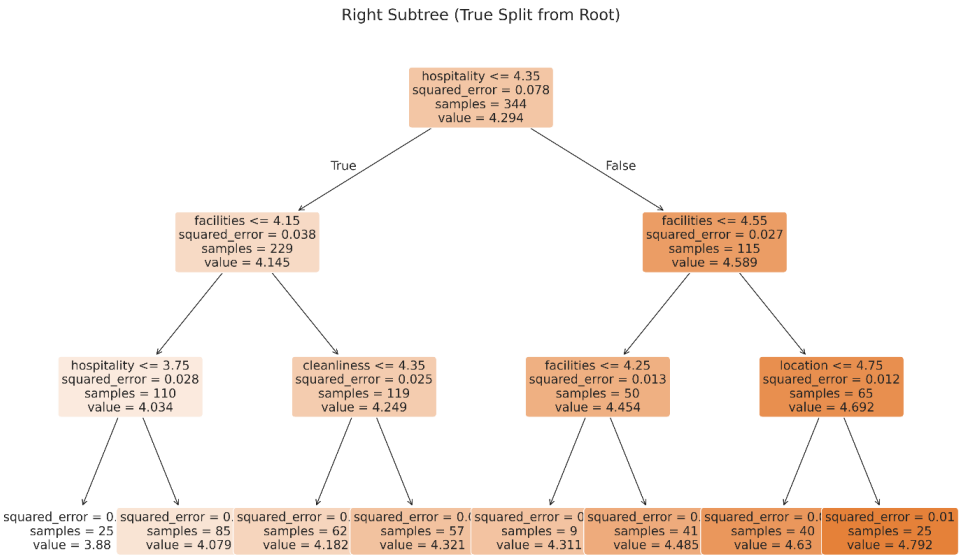

Figure 15. Right subtree for Decision tree for hotel rating

6.8. Understanding the Confusion Matrix in Evaluating Predictive Models

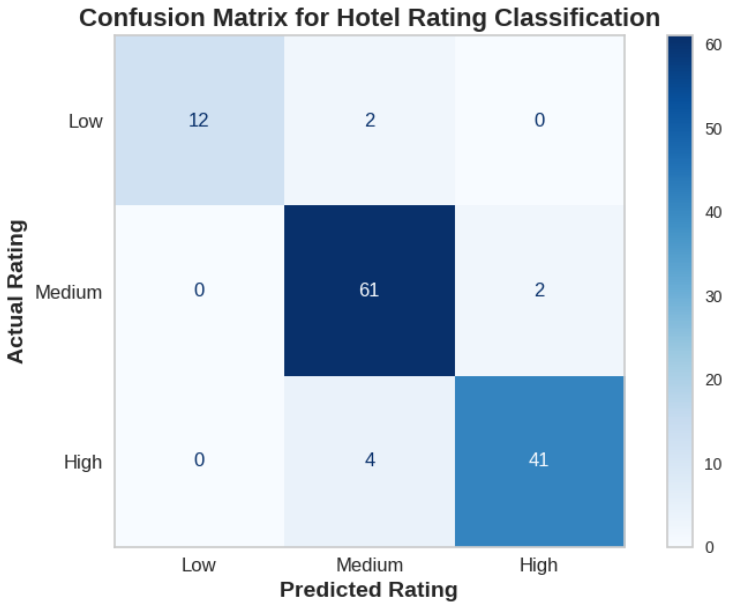

Figure 16. Confusion Matrix for Multiclass Hotel Rating Classification

In addition to the regression-based analysis of hotel ratings, a classification model was employed to predict categorical versions of the ”overall rating” by grouping them into three classes: Low, Medium, and High. This classification task is distinct from the earlier regression models, as it focuses on predicting discrete categories rather than continuous values. The model was trained using the same set of hotel features-location, hospitality, facilities, cleanliness, value for money, food, and price. To evaluate the model's performance in this classification task, a confusion matrix was generated, as shown in Figure 16. The confusion matrix compares actual hotel rating categories against the predicted ones and reveals that the model performs strongly in identifying medium and high-rated hotels. It correctly classified 61 medium-rated and 41 high-rated hotels. Additionally, 12 low-rated hotels were accurately identified, though two were misclassified as medium. For medium-rated hotels, only two instances were misclassified as high, and among the high-rated hotels, four were predicted as medium. Notably, there were no misclassifications between the extreme categories (e.g., no low-rated hotels predicted as high), suggesting that the model learned to distinguish boundary cases effectively. Overall, this classification model demonstrates high predictive reliability, particularly in adjacent class distinctions, and complements the regression analysis by providing a categorical perspective on hotel quality segmentation. The results affirm the usefulness of machine learning in capturing nuanced patterns in hotel ratings for both continuous and categorical interpretations.

7. Conclusion

This study conducted an in-depth investigation into the application of various machine learning techniques for predicting hotel ratings based on features such as location, hospitality, facilities, cleanliness, value for money, food, and price. Multiple regression and classification models were evaluated, including Linear Regression, Gradient Boosting Regressor, k-nearest neighbours (KNN), Support Vector Regression (SVR), PCA-based Linear Regression, and Random Forest Regressor. Among these, Random Forest and Gradient Boosting achieved the best predictive performance, with Random Forest ultimately selected as the most suitable model due to its comparable accuracy, greater interpretability, reduced sensitivity to hyperparameter tuning, and consistent stability across

runs. The ensemble nature of Random Forest enables it to effectively handle non-linearity, multicollinearity, and interactions among features, outperforming simpler linear models or distance-based methods such as KNN. In addition to regression tasks, we also explored a classification approach by discretizing continuous hotel ratings into categorical classes: Low, Medium, and High. The confusion matrix results demonstrated high predictive reliability, particularly for medium- and high-rated hotels, with minimal misclassifications and no extreme-category errors (e.g., low classified as high). These findings highlight the value of machine learning in the hospitality industry, particularly for platforms aiming to provide personalized recommendations, automated rating predictions, and quality assessments for users. For consumers, accurate rating predictions support better decision-making tailored to their expectations and budgets. For hotel operators and aggregators, the models serve as strategic decision-making tools for quality control, market segmentation, and competitive pricing strategies.

Future Work: Future studies could explore more advanced machine learning methods such as XGBoost, LightGBM, and deep neural networks to further improve predictive accuracy and generalization. Incorporating unstructured textual review data through natural language processing (NLP) would provide deeper insights into customer sentiment, complementing numerical ratings and enabling more nuanced predictions. Multi-modal approaches that combine images, textual reviews, and structured hotel features could create more comprehensive models capturing various aspects of hotel quality. Longitudinal studies could track changes in hotel ratings over time to capture evolving customer preferences and market trends. Expanding the dataset to include international hotels or different market segments would test the model's generalizability across diverse contexts. Finally, employing interpretability methods such as SHAP or LIME could help identify key features driving predictions, offering actionable insights to hotel operators and platform designers for decision-making and service improvements.

Conflict of Interest Statement

The authors certify that they have no affiliations with any organization or entity with financial or non-financial interests. No conflicts of interest are declared related to this work.

REFERENCES

1. J. M. Benitez, J. C. Martín, and C. Román, *Using fuzzy number for measuring quality of service in the hotel industry*, Tourism Management, vol. 28, no. 2, pp. 544–555, 2007.
2. T. Y. Chou, C. L. Hsu, and M. C. Chen, *A fuzzy multi-criteria decision model for international tourist hotel location selection*, International Journal of Hospitality Management, vol. 27, no. 2, pp. 293–301, 2008.
3. W. H. Tsai, J. L. Hsu, C. H. Chen, W. R. Lin, and S. P. Chen, *An integrated approach for selecting corporate social responsibility programs and costs evaluation in the international tourist hotel*, International Journal of Hospitality Management, vol. 29, no. 3, pp. 385–396, 2010.
4. A. Akincilar, and M. Dagdeviren, *A hybrid multi-criteria decision-making model to evaluate hotel websites*, International Journal of Hospitality Management, vol. 36, pp. 263–271, 2014.
5. A. M. Gil-Lafuente, J. M. Merigó, and E. Vizuete, *Analysis of luxury resort hotels by using the fuzzy analytic hierarchy process and the fuzzy Delphi method*, Economic Research-Ekonomska Istraživanja, vol. 27, no. 1, pp. 244–266, 2014.
6. L. Masiero, C. Y. Heo, and B. Pan, *Determining guests' willingness to pay for hotel room attributes with a discrete choice model*, International Journal of Hospitality Management, vol. 49, pp. 117–124, 2015.
7. M. H. Chen, *A quantile regression analysis of tourism market growth effect on the hotel industry*, International Journal of Hospitality Management, vol. 52, pp. 117–120, 2016.
8. A. Mardani, E. K. Zavadskas, D. Streimikiene, A. Jusoh, K. M. Nor, and M. Khoshnoudi, *Using fuzzy multiple criteria decision-making approaches for evaluating energy saving technologies and solutions in five-star hotels: A new hierarchical framework*, Energy, vol. 117, pp. 131–148, 2016.
9. N. Rianthong, A. Dumrongsiri, and Y. Kohda, *Optimizing customer searching experience of online hotel booking by sequencing hotel choices and selecting online reviews: A mathematical model approach*, Tourism Management Perspectives, vol. 20, pp. 55–65, 2016.
10. A. Akbaba, *Measuring service quality in the hotel industry: A study in a business hotel in Turkey*, International Journal of Hospitality Management, vol. 25, no. 2, pp. 170–192, 2006.
11. Y. S. Wang, H. T. Li, C. R. Li, and D. Z. Zhang, *Factors affecting hotels' adoption of mobile reservation systems: A technology-organization-environment framework*, Tourism Management, vol. 53, pp. 163–172, 2016.
12. S. M. Yu, J. Wang, and J. Q. Wang, *An interval type-2 fuzzy likelihood-based MABAC approach and its application in selecting hotels on a tourism website*, International Journal of Fuzzy Systems, vol. 19, pp. 47–61, 2017.

13. I. K. W. Lai, and M. Hitchcock, *Sources of satisfaction with luxury hotels for new, repeat, and frequent travelers: A PLS impact-asymmetry analysis*, Tourism Management, vol. 60, pp. 107–129, 2017.
14. M. J. Leutwiler-Lee, S. S. Kim, F. Badu-Baiden, and B. King, *Dimensionality in the service quality perceptions of quarantine hotel guests*, Tourism Management Perspectives, vol. 47, 101124, 2023.
15. M. Wu, G. W. H. Tan, E. C. X. Aw, and K. B. Ooi, *Unlocking my heart: Fostering hotel brand love with service robots*, Journal of Hospitality and Tourism Management, vol. 57, pp. 339–348, 2023.
16. G. Viglia, A. Mauri, and M. Carricano, *The exploration of hotel reference prices under dynamic pricing scenarios and different forms of competition*, International Journal of Hospitality Management, vol. 52, pp. 46–55, 2016.
17. S. Boo, M. Kim, and T. J. Kim, *Effectiveness of corporate social marketing on prosocial behavior and hotel loyalty in a time of pandemic*, International Journal of Hospitality Management, vol. 117, 103635, 2024.
18. M. Cruz, D. Hodari, and S. Raub, *The impact of management structure on guest satisfaction in chain-affiliated hotels and the moderating influence of chain scale*, International Journal of Hospitality Management, vol. 117, 103651, 2024.
19. S. Fang, X. Han, and S. Chen, *Hotel guest-robot interaction experience: A scale development and validation*, Journal of Hospitality and Tourism Management, vol. 58, pp. 1–10, 2024.
20. J. Lee, J. Kim, T. Hong, S. G. Mun, K. Koh, and C. Koo, *Scalable investigation of energy usage patterns and saving potential in hotel guestrooms: Focused on occupancy states and electrical installations*, Energy and Buildings, vol. 302, 113735, 2024.
21. W. M. Lim, K. M. Jasim, and M. Das, *Augmented and virtual reality in hotels: Impact on tourist satisfaction and intention to stay and return*, International Journal of Hospitality Management, vol. 116, 103631, 2024.
22. S. Nakamura, A. Baskaran, and S. K. Selvarajan, *Impact of Airbnb on the hotel industry in Japan*, Journal of Destination Marketing & Management, vol. 31, 100841, 2024.
23. A. Oukil, R. E. Kennedy, A. Al-Hajri, and A. A. Soltani, *Unveiling the potential of hotel mergers: A hybrid DEA approach for optimizing sector-wide performance in the hospitality industry*, International Journal of Hospitality Management, vol. 116, 103620, 2024.
24. P. Zaragoza-Sáez, B. Marco-Lajara, M. Úbeda-García, and E. Manresa-Marhuenda, *Exploratory and co-exploratory innovation. The mediating role of digitalization on competitiveness in the hotel industry*, Technological Forecasting and Social Change, vol. 199, 123069, 2024.
25. Y. M. Shehawy, G. Agag, H. O. Alamoudi, M. D. Alharthi, A. Brown, T. G. Labben, and Z. H. Abdelmoety, *Cross-national differences in consumers' willingness to pay (WTP) more for green hotels*, Journal of Retailing and Consumer Services, vol. 77, 103665, 2024.
26. J. Yu, S. S. Kim, N. G. BAAH, and H. Han, *Veganism, a new hotel paradigm: Exploring the attributes of vegan-friendly hotels and guest approach behaviors*, International Journal of Hospitality Management, vol. 117, 103639, 2024.
27. *MakeMyTrip: Online travel agency platform*, https://www.makemytrip.com/.
28. I. Jolliffe, *Principal Component Analysis*, In Wiley StatsRef: Statistics Reference Online, Wiley, 2014.
29. S. Begum, D. Chakraborty, and R. Sarkar, "Data Classification Using Feature Selection and kNN Machine Learning Approach," in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, India, pp. 811–814, 2015.
30. H. Wang, Y. Geng, M. Zhang, W. Wang, Y. Yang, H. Qian, and C. Xi, "A hybrid model based on novel SVM-SVR and weighted combination strategy for build-up rate prediction," *Measurement Science and Technology*, IOP Publishing, vol. 36, no. 1, p. 016012, Oct. 2024.