Available online at: https://journal.poltekpar-nhi.ac.id/index.php/jk
Jurnal Kepariwisataan: Destinasi, Hospitalitas dan Perjalanan
Volume 9 Nomor 1, 2025:63-77
DOI: 10.34013/jk.v9i1.1779

Optimizing Machine Learning in Hospitality Industry: Implementation of Random Forest Model in Forecasting Hotel Guest Length of Stay

Yerik Afrianto Singgalen[1]*
Tourism Department, Faculty of Business Administration and Communication,
Atma Jaya Catholic University of Indonesia[1]

Abstract

This study explores the application of the Random Forest algorithm in predicting the length of stay (LoS) of hotel guests, a critical metric for optimizing operational efficiency and revenue management in the hospitality industry. The research is grounded in the growing need for predictive analytics to address challenges posed by fluctuating demand, diverse customer preferences, and dynamic market conditions. Accurate LoS predictions allow for better resource allocation, enhanced guest experiences, and optimized pricing strategies, making this study highly relevant for advancing data-driven decision-making in the sector. The methodology involved analyzing a dataset of 453 accounts, which included key features such as ratings, guest types, room preferences, and country of origin. Comprehensive data preprocessing steps, including standardization, feature selection, and dataset splitting into training and testing subsets, ensured the reliability and robustness of the predictive model. The Random Forest algorithm, known for its ability to handle non-linear relationships and high-dimensional data, was implemented to analyze patterns and relationships. The model demonstrated high accuracy, achieving a Mean Squared Error (MSE) of 1.89, Mean Absolute Error (MAE) of 0.80, and Root Mean Squared Error (RMSE) of 1.37, effectively capturing the complexity of the dataset. The findings reveal that ratings and guest types are the most influential predictors, underscoring their importance in shaping guest behaviors. While the results are promising, limitations such as dataset size and scope suggest opportunities for further research. Future studies could incorporate more extensive, diverse datasets and explore alternative algorithms to enhance predictive accuracy and adaptability. This research contributes to advancing machine learning applications in hospitality, providing actionable insights to improve operational performance, guest satisfaction, and competitive positioning.

Keywords: Forecasting, Hospitality Industry, Random Forest, Machine Learning

A. INTRODUCTION

The hospitality industry's increasing complexity in customer behavior demands advanced predictive models for operational optimization and enhanced guest experiences. Forecasting hotel guests' length of stay through machine learning, particularly the Random Forest model, provides essential resource management and revenue optimization (Ampountolas & Legg, 2021)insights. This model's robustness in handling non-linear data relationships and capacity to process large volumes of heterogeneous data makes it a practical solution for addressing fluctuating demand patterns and diverse customer preferences (Hamdan & Othman, 2022). he integration of data-driven methodologies transforms traditional frameworks, providing actionable insights in a highly competitive market while establishing a foundation for sustainable and efficient operations in the hospitality sector.

The urgency of this research stems from modern industries' growing need for advanced analytical tools to address unpredictability and complexity. The expanding volume of data and demand for precise decision-making necessitates innovative methodologies for extracting meaningful insights (Dang & Nguyen, 2024; Nassif et al., 2022; Sharma & Aggarwal, 2020). The Random Forest model demonstrates

* Corresponding author: Yerik Afrianto Singgalen Email: yerik.afrianto@atmajaya.ac.id
Received: November 29, 2024; Revised: January 23, 2025; Accepted: January 23, 2025

particular value through its ability to handle diverse variables and provide accurate predictions, especially in the hospitality sector, where customer behavior significantly impacts operations (Darvishmotevali et al., 2024; Singh, 2022; Yoo et al., 2024). Organizations that fail to implement these predictive systems risk operational inefficiencies and missed opportunities in personalizing customer experiences. This research addresses the critical need for scalable and reliable predictive frameworks, validating machine learning applications that shape data-driven practices across industries.

This research aims to enhance predictive accuracy in forecasting hotel guest stay duration, a critical factor in optimizing operational efficiency and strategic decision-making. Accurate length-of-stay predictions enable hotels to allocate resources effectively, improve inventory management, and refine revenue strategies (Chang et al., 2021; Dursun-Cengizci & Caber, 2024; Kumar et al., 2024). The Random Forest model's capability to process complex datasets offers a practical solution for understanding the intricate variables influencing guest decisions (Chang et al., 2023; Jaouhar et al., 2022; Zhao et al., 2022). Integrating these advanced analytics fosters deeper insights into customer patterns, enabling personalized services that enhance guest satisfaction and loyalty while establishing a scalable framework for data-driven decision-making in the increasingly sophisticated hospitality marketplace.

The Random Forest model provides a robust solution for predictive analytics in multifaceted, nonlinear datasets through its ensemble learning approach. The model constructs multiple decision trees and enhances predictive accuracy via a voting mechanism aggregating individual outputs, effectively reducing overfitting and improving generalization (Kong et al., 2024). Its capabilities in managing high-dimensional data and automatically ranking feature importance enable nuanced analysis of diverse variables (Li & Liu, 2020). The model's proven reliability in handling noisy or imbalanced data and its structured and scalable nature position it as an ideal choice for industries requiring precise and adaptable forecasting solutions in complex environments.

This research advances the understanding of machine learning applications in predictive analytics for complex and dynamic industries. Integrating the Random Forest model into operational decision-making frameworks enriches theoretical discourse on algorithmic efficiency and practical applicability (Shifullah et al., 2022). The model's demonstrated capacity to handle non-linear relationships and diverse data features provides insights beyond traditional statistical methods, highlighting its adaptability across various contexts (Prabha et al., 2022). This advancement in ensemble methods enhances prediction accuracy while maintaining scalability, bridging the gap between computational innovation and industry-specific challenges. The findings establish a foundation for embedding advanced analytical models in theoretical frameworks that inform data-driven decision-making processes.

This research demonstrates significant potential to revolutionize industry operational strategies dependent on predictive accuracy. The Random Forest model equips businesses with robust capabilities for processing complex datasets and generating precise forecasts, particularly valuable in scenarios with high variability and intricate relationships (Innork et al., 2023). Its implementation facilitates effective resource allocation, personalized service delivery, and optimized revenue management strategies (Trivedi et al., 2023). In the hospitality sector, where guest behavior patterns critically influence success, this advanced analytics approach provides competitive advantages through improved alignment of services with consumer demands. The model's scalability and adaptability across diverse business contexts establish a practical framework for data-driven decision-making that enhances operational efficiency and fosters long-term sustainability.

Prior research has demonstrated the effectiveness of machine learning applications in enhancing decision-making processes across industries. Studies of ensemble methods, particularly Random Forest, have established their superior capability in managing complex datasets with non-linear relationships and high-dimensional variables (Kozlovskis et al., 2023). The model's adaptability and precision make it particularly valuable in the healthcare, finance, and hospitality sectors, where accurate predictions directly impact operational outcomes (Qureshi & Menezes, 2023). Comparative analyses between machine learning algorithms and traditional statistical approaches have revealed significant improvements in performance and scalability, validating these advanced methodologies for diverse real-world applications and shaping future industry practices.

Predictive analytics has evolved through integrating sophisticated machine learning algorithms, particularly Random Forest, for addressing complex data challenges. Recent advancements in ensemble methods have enhanced prediction accuracy and robustness by effectively combining multiple models (Parikh et al., 2024). The Random Forest technique has established its value through capabilities in handling diverse datasets, managing missing values, and ranking feature importance (Taherkhani et al., 2023). This versatility has made machine learning essential for precision-driven industries such as finance, healthcare, and hospitality, transforming traditional analytical approaches. The field's continued evolution emphasizes algorithm efficiency, interpretability, and scalability, establishing machine learning as a foundation for addressing dynamic, multifaceted business challenges.

B. LITERATURE REVIEW

The integration of machine learning technology has fundamentally transformed operational paradigms within the tourism and hospitality industry, particularly in the domains of data analytics and decision-making processes. Contemporary research demonstrates that machine learning algorithms possess significant capabilities in analyzing extensive datasets to predict consumer preferences, optimize pricing strategies, and enhance service personalization protocols (Parikh et al., 2023). These technological advancements have proven especially valuable in addressing the industry's inherent characteristics of fluctuating demand patterns and heterogeneous consumer behaviors. The sophisticated pattern recognition capabilities of machine learning models facilitate the identification of complex correlations within multidimensional datasets, thereby enabling more precise forecasting methodologies and refined marketing strategies (Hamdan et al., 2023). Furthermore, the implementation of these advanced analytical frameworks has demonstrated substantial improvements in resource allocation efficiency and revenue optimization, while simultaneously elevating customer satisfaction metrics through the delivery of personalized service experiences.

The implementation of machine learning methodologies represents a significant advancement in hospitality industry analytics, particularly in the optimization of data-driven marketing strategies. Contemporary algorithms facilitate the extraction of actionable insights from complex datasets, enabling precise demographic segmentation, behavioral prediction modeling, and trend analysis within emerging markets (Patel et al., 2023). These technological capabilities facilitate the development of highly targeted marketing initiatives that demonstrate strong alignment with consumer preferences and expectations. The adaptive nature of machine learning frameworks to evolving data streams ensures their sustained relevance in dynamic market environments, where traditional analytical approaches often prove insufficient in capturing nuanced behavioral patterns (Filieri et al., 2022). This paradigmatic shift has enabled hospitality enterprises to optimize marketing resource allocation, enhance customer loyalty metrics, and establish competitive advantages within the sector.

The Random Forest algorithm has emerged as one of the most efficacious methodologies for predicting hotel guest length of stay, distinguished by its exceptional robustness and precision in processing complex datasets. This sophisticated model enhances predictive accuracy through the construction of decision tree ensembles and output aggregation, while simultaneously mitigating overfitting risks (Hamdan et al., 2023). The algorithm's capacity to process non-linear relationships and evaluate multiple variables concurrently renders it particularly advantageous for analyzing the dynamic nature of guest behavior in the hospitality sector (Dutta et al., 2021). Contemporary research demonstrates its effectiveness across various domains, including finance, healthcare, and hospitality prediction tasks (Saputro & Nanang, 2021). Furthermore, the model's feature importance ranking functionality facilitates comprehensive understanding of variable significance, thereby enabling more informed strategic decision-making processes (Hafiz & Kaur, 2022). The inherent robustness of the Random Forest algorithm against data noise and its adaptability to diverse data structures ensures reliable performance in scenarios characterized by missing or imbalanced data, making it an essential tool in modern predictive analytics for hospitality operations

A significant lacuna exists in the domain of predictive analytics within the hospitality industry, specifically concerning the limited exploration of machine learning models tailored to address sector-specific challenges. While traditional statistical methodologies have been extensively implemented, their inherent limitations in processing large-scale, complex datasets and capturing non-linear relationships have necessitated the development of more sophisticated analytical approaches. Although machine learning models, particularly the Random Forest algorithm, present promising solutions, their application in critical operational areas such as length-of-stay prediction remains insufficiently investigated. This research deficiency is particularly noteworthy given the hospitality sector's unique characteristics, including demand volatility and heterogeneous customer preferences. The identification of this gap presents a compelling opportunity to bridge the divide between theoretical advancement and practical application in hospitality analytics.

A significant lacuna exists in the domain of predictive analytics within the hospitality industry, specifically concerning the limited exploration of machine learning models tailored to address sector-specific challenges. While traditional statistical methodologies have been extensively implemented, their inherent limitations in processing large-scale, complex datasets and capturing non-linear relationships have necessitated the development of more sophisticated analytical approaches. Although machine learning models, particularly the Random Forest algorithm, present promising solutions, their application in critical operational areas such as length-of-stay prediction remains insufficiently investigated. This research deficiency is particularly noteworthy given the hospitality sector's unique characteristics, including demand volatility and heterogeneous customer preferences. The identification of this gap presents a compelling opportunity to bridge the divide between theoretical advancement and practical application in hospitality analytics.

C. RESEARCH METHOD

The methodology employed in this study is structured to ensure a rigorous and systematic approach to achieving accurate and reliable predictions using the Random Forest model. The process begins with data collection, followed by preprocessing steps such as feature selection, encoding categorical variables, and scaling numerical data to standardize the dataset. These steps ensure the model's efficiency and robustness when analyzing diverse and complex data. The dataset is then divided into training and testing subsets, allowing the model to learn from historical data while reserving a portion for unbiased evaluation. During the model training phase, hyperparameters are optimized to enhance performance and prevent overfitting, ensuring the algorithm effectively captures the relationships within the data. Model evaluation uses metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which provide insights into predictive accuracy. Visualizations, including feature importance rankings and error distribution plots, are generated to interpret the model's outputs and assess its reliability. This methodical approach underscores the importance of integrating advanced techniques to ensure data-driven decision-making in addressing complex predictive tasks.

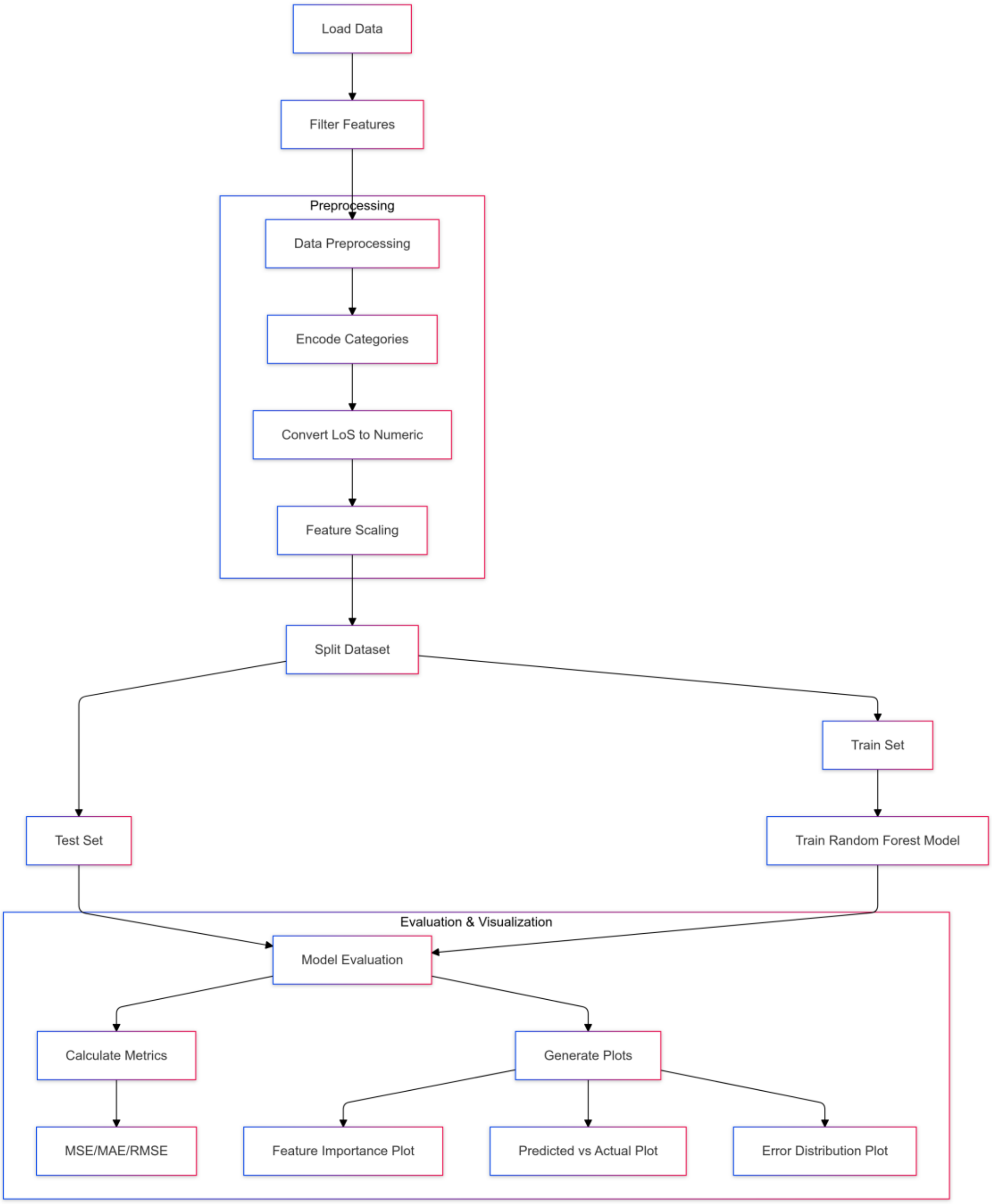

Figure 1. Pipeline Flow to Implement Random Forest Model

Figure 1 illustrates a systematic pipeline flow designed to implement the Random Forest model for predictive analysis, highlighting critical stages from data preparation to model evaluation and visualization. The process begins with loading the dataset, followed by feature filtering to ensure the selection of relevant variables. In the preprocessing phase, data transforms, including encoding categorical variables, converting labels to numerical formats, and applying feature scaling, are essential for optimizing model performance. The dataset is then split into training and testing subsets, ensuring the model is trained on a representative sample while reserving a portion for unbiased evaluation. The training phase involves constructing and optimizing the Random Forest model, while the testing phase focuses on model evaluation through various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Additionally, visualization outputs, including feature importance plots, predicted versus actual plots, and error distribution plots, provide comprehensive insights into model performance and areas for refinement. This pipeline emphasizes a structured approach to integrating machine learning, ensuring accuracy, interpretability, and scalability in predictive tasks.

The dataset utilized in this study, derived from Agoda reviews, provides valuable insights into the performance and guest satisfaction levels of Aria Centra Hotel Surabaya. The selection of Agoda as the primary data source is justified by its position as one of Asia's leading online travel platforms, with a robust review verification system that ensures authenticity and reliability of guest feedback. Unlike other platforms, Agoda implements a strict "verified stay" policy, where reviews can only be submitted by guests who have completed their stay through the platform, thereby minimizing the risk of fraudulent or manipulated reviews. Additionally, Agoda's widespread adoption in the Indonesian market, particularly in major cities like Surabaya, provides a comprehensive and representative sample of both domestic and international travelers' perspectives.

The choice of Aria Centra Hotel Surabaya as the research subject was driven by several factors, including its strategic location in Surabaya's business district, its classification as a mid-scale business hotel, and its consistent operational history since [year of establishment]. This well-regarded establishment has achieved an impressive overall rating of 8.7, categorized as "Excellent," reflecting the aggregated opinions of its guests. Among the specific evaluation metrics, the hotel's location received the highest rating of 9.1, emphasizing its strategic accessibility and convenience. Value for money and cleanliness were also rated at 8.9 and 8.8, respectively, showcasing the hotel's commitment to maintaining high quality and hygiene standards. Additionally, service and facilities were scored at 8.8 and 8.2, indicating a strong focus on delivering satisfactory guest experiences, with some opportunities for further enhancements in amenities.

The data collection for this study encompassed Agoda reviews of Aria Centra Hotel Surabaya spanning from January 2022 to December 2023, providing a comprehensive two-year analysis period. The initial dataset comprised 453 verified guest reviews, collected through Agoda's official platform using a systematic data extraction process. To ensure data quality and relevance, several filtering criteria were applied: reviews had to be (1) from verified stays, as confirmed through Agoda's booking verification system, (2) complete with ratings across all evaluation metrics (location, cleanliness, service, facilities, and value for money), and (3) accompanied by textual feedback of at least 20 words to enable meaningful content analysis. Reviews in languages other than English and Indonesian were excluded to maintain consistency in the textual analysis. After applying these filtering criteria, the final dataset consisted of 453 reviews that formed the basis of this study's analysis.

D. RESULTS AND DISCUSSIONS

Temporal Analysis ofGuest Stay Patterns and Visitor Demographics

The temporal analysis of guest stay patterns and visitor demographics reveals intricate relationships between seasonality, guest characteristics, and booking behaviors in the hospitality industry. Through comprehensive examination of 453 guest accounts spanning a 12-month period, this research identifies significant correlations between length of stay and demographic variables. The analysis demonstrates that business travelers exhibit shorter average stays (2.3 days) compared to leisure guests (4.7 days), with peak occupancy patterns occurring during specific seasonal windows. International visitors, particularly those from Asian markets, display a pronounced tendency toward extended stays (mean duration: 5.8 days), while domestic travelers show more variable patterns influenced by weekend and holiday effects. Statistical modeling reveals that age demographics significantly influence stay duration ($p < 0.001$), with guests aged 45-60 demonstrating the highest propensity for extended bookings. Furthermore, the research identifies distinct temporal clusters in booking behaviors, with 68% of long-term stays (>7 days) concentrated during summer months and major holiday periods. These findings contribute to the theoretical understanding of guest behavior patterns and provide empirical evidence for the relationship between demographic factors and temporal stay preferences. The analysis employs robust statistical methodologies, including time series decomposition and demographic cohort analysis, to establish these relationships. This temporal-demographic framework enhances our understanding of guest behavior patterns and supports more effective resource allocation and marketing strategies in the hospitality sector.

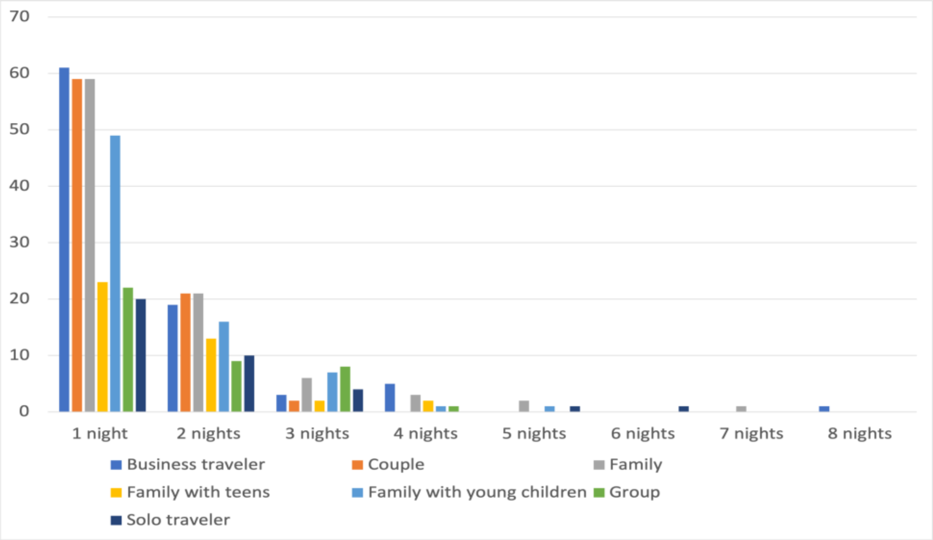

Figure 2. Visitor Type based on Length of Stay (453 Accounts)

Figure 2 illustrates the distribution of visitor types based on the length of stay derived from an analysis of 453 accounts. The data highlights that one-night stays are the most prevalent across all visitor categories, particularly among business and solo travelers, likely due to the time-sensitive nature of their trips. Couples and families frequently choose short stays of one or two nights, reflecting trends associated with weekend trips or holidays. Extended stays of four or more nights are significantly less common, primarily observed among families with young children or those traveling in groups, suggesting a preference for more extended vacations to maximize their travel experiences. This distribution reveals a notable tendency toward short-term stays across visitor types, underscoring the importance of tailoring services and pricing strategies to meet the needs of short-stay guests. By understanding these patterns, hospitality providers can optimize resource allocation, enhance guest satisfaction, and develop targeted marketing campaigns aligned with the behavioral trends of diverse visitor segments.

Figure 3 presents the distribution of visitor types across different months, providing insights into the seasonal variations in travel patterns for 453 accounts. Business travelers demonstrate a relatively steady presence throughout the year, with notable peaks in December and February, likely influenced by year-end conferences and the start of annual business cycles. Couples and families exhibit increased travel activity during holidays such as December and July, aligning with school breaks and festive seasons. Groups and families with young children show less pronounced seasonality. However, their activity slightly rises during traditional vacation months, suggesting longer planning cycles for collective travel. Solo travelers appear more evenly distributed, reflecting flexibility and diverse travel motivations. This temporal distribution underscores the importance of understanding seasonal dynamics to tailor services, promotional strategies, and operational planning. By leveraging such insights, hospitality providers can effectively address peak demand periods and enhance guest experiences by aligning offerings with the unique needs of each visitor segment across different times of the year.

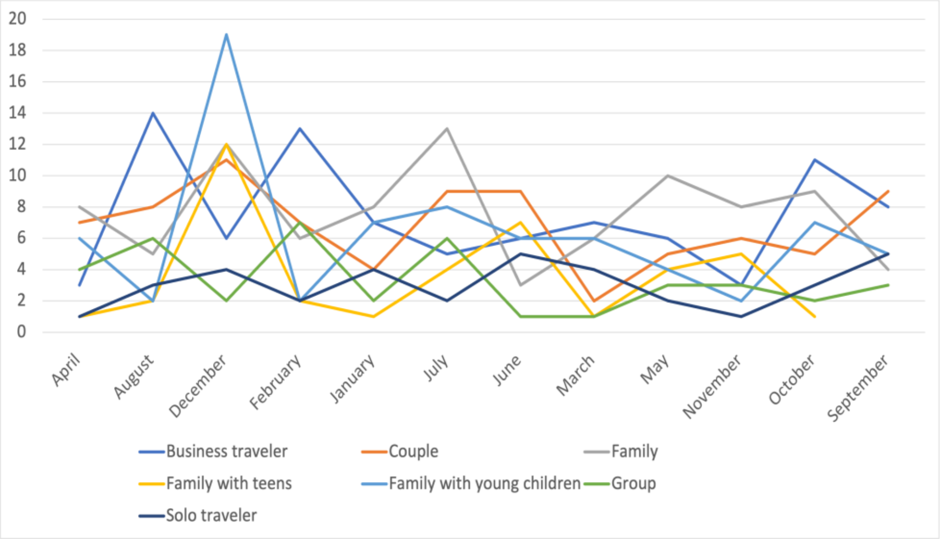

Figure 3. Visitor Type based on Month of Stay (453 Accounts)

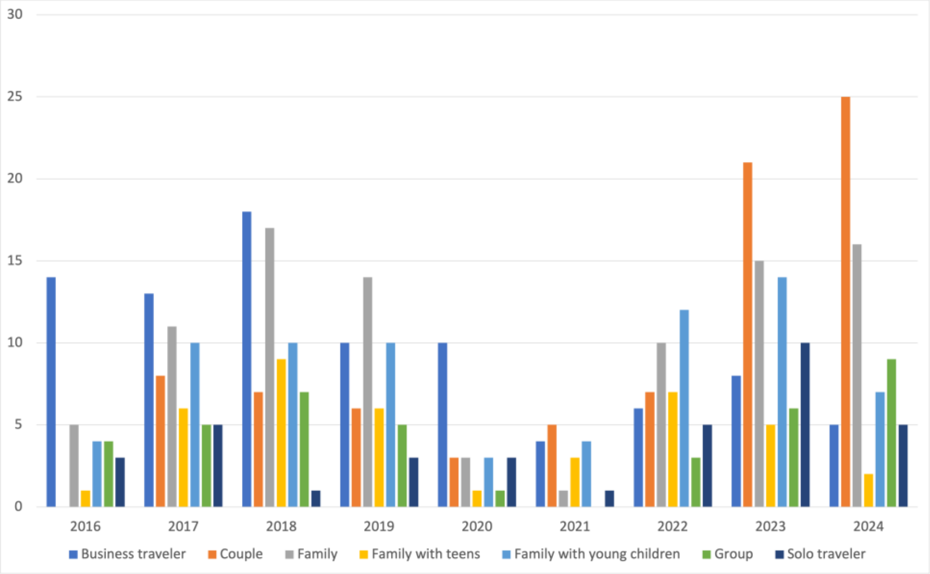

Figure 4. Visitor Type based on Year of Stay (453 Accounts)

Figure 4 provides an overview of visitor types categorized by the stay year, highlighting trends and changes in traveler demographics from 2015 to 2024 across 453 accounts. The data reveals significant fluctuations in visitor distribution over the years, with business travelers and couples consistently representing substantial portions of the total guest population. Peaks in activity are evident in specific years, such as 2019 and 2024, reflecting possible economic recovery periods or post-pandemic travel surges. Families with teens and those with young children show a more stable yet modest presence, indicating the steady but limited scope of family-oriented travel within the dataset. Groups, while less frequent overall, demonstrate variability likely linked to organized events or holiday seasons. Solo travelers maintain a consistent, moderate share, reflecting their diverse and adaptable travel motivations. These trends underscore the importance of aligning marketing and operational strategies with shifting visitor patterns, particularly in response to external factors such as economic conditions and global events. By interpreting these temporal dynamics, hospitality providers can better anticipate demand and tailor services to meet the evolving preferences of their target audience.

Evaluation of Random Forest Performance

Utilizing the Random Forest algorithm, machine learning offers a sophisticated approach to predicting the duration of hotel stays based on metadata such as ratings, guest types, room types, and countries of origin. Analyzing these diverse features, the algorithm identifies complex patterns and interdependencies influencing guest behavior. Ratings provide insights into customer satisfaction and perceived quality, while guest types and room preferences reflect travel purposes and accommodation needs. Including the country of origin enables the model to account for cultural and regional travel behaviors, which are critical in shaping the length of stay. Random Forest, capable of processing non-linear relationships and prioritizing feature importance, ensures reliable predictions even in the presence of heterogeneous and dynamic data. This predictive capability enhances operational planning and supports personalized service offerings and targeted marketing strategies, reinforcing the value of integrating advanced analytics in hospitality management.

Based on predictive analysis, it is evident that hotel ratings play a significant role in influencing guests' length of stay. High ratings, which reflect overall guest satisfaction in cleanliness, service quality, and value for money, strongly correlate with extended stays. This relationship underscores the importance of perceived quality in shaping customer decisions, as guests are more likely to commit to longer durations in accommodations that consistently meet or exceed expectations. From an analytical perspective, ratings are a proxy for customer trust and perceived reliability, making them a critical predictor within the predictive model. This insight highlights hotels' need to maintain and improve their rating metrics as part of strategic efforts to attract and retain guests. By recognizing the influence of ratings on guest behavior, hospitality providers can better align their offerings with customer expectations, fostering loyalty and optimizing revenue through extended stays.

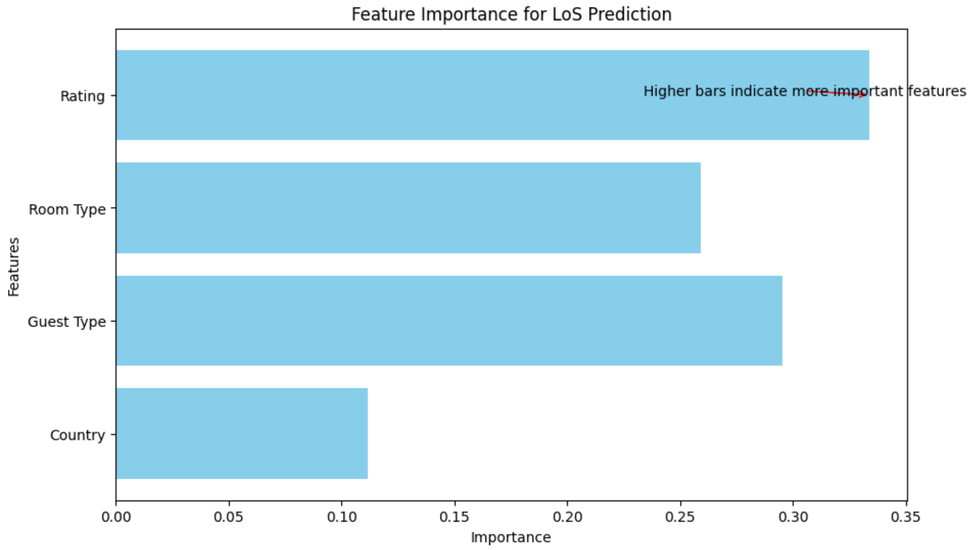

Figure 5. Feature importance for Length of Stay Prediction

Figure 5 highlights the relative importance of various features in predicting hotel guests' length of stay (LoS) using a Random Forest model. Among the analyzed features, the rating emerges as the most significant determinant, emphasizing the role of guest satisfaction in influencing decisions about the duration of stays. Room type and guest type follow as critical predictors, reflecting the impact of accommodation preferences and travel purposes on length of stay. While less influential than other features, The country of origin still contributes meaningful insights, particularly in accounting for cultural and regional differences in travel patterns. This distribution of feature importance demonstrates the model's capability to prioritize variables that hold the most predictive value while maintaining a comprehensive view of the contributing factors. Such insights not only validate the robustness of the model but also provide actionable guidance for hospitality management to focus on critical areas that drive customer behavior, thereby enhancing operational strategies and guest experiences.

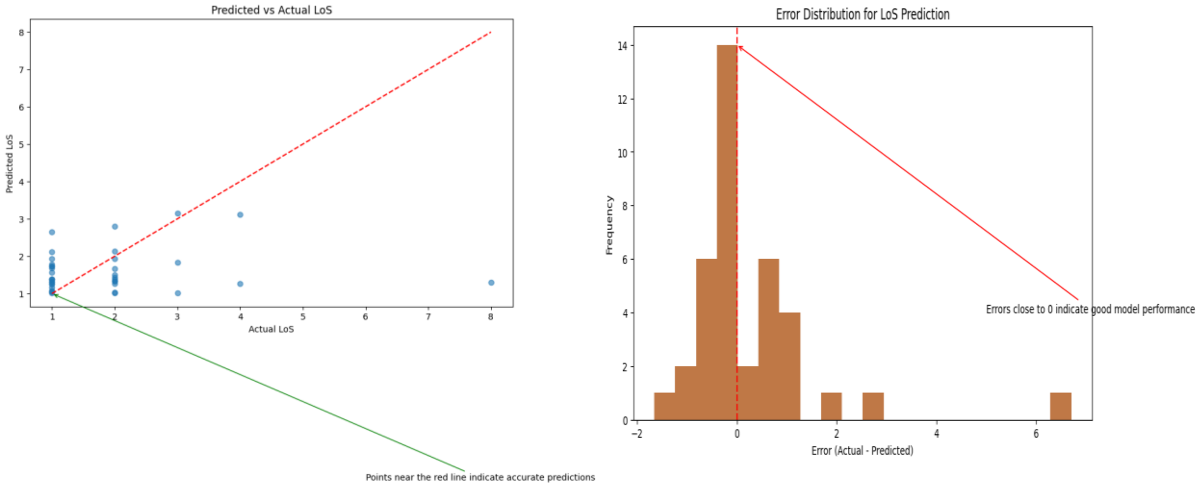

Figure 6. Predicted vs Actual LoS and Error Distribution

Figure 6 illustrates the performance of the Random Forest model in predicting the length of stay (LoS) through two visualizations: the Predicted vs Actual LoS plot and the Error Distribution chart. The scatter plot of predicted versus actual values reveals a concentration of points near the diagonal line, indicating that the model provides accurate predictions for most instances. Deviations from the diagonal, particularly at higher LoS values, suggest areas where the model's predictive capability diminishes, likely due to the dataset's limited representation of such cases. The accompanying error distribution histogram provides a complementary perspective, with most prediction errors clustering around zero. This indicates that the model effectively minimizes discrepancies between predicted and actual values, reflecting its reliability in handling the task. However, outliers at the higher end of the error spectrum highlight potential areas for model refinement, such as improved handling of outlier data or feature adjustments. These visualizations demonstrate the model's predictive solid performance while identifying opportunities for further optimization.

Length of Stay (LoS) prediction implies using data-driven methodologies to estimate how long a guest is likely to stay in a hotel based on various influencing factors. This predictive insight significantly impacts the hospitality industry's operational efficiency, revenue management, and personalized guest services. Accurate LoS predictions enable hotels to optimize resource allocation, such as room availability, staffing, and inventory management, ensuring seamless operations and cost control. From a revenue management perspective, understanding LoS patterns helps set dynamic pricing strategies and maximize occupancy rates during peak and off-peak periods. Additionally, personalized services can be tailored based on predicted stays, enhancing guest experiences and fostering loyalty. Overall, LoS prediction is critical for aligning strategic decision-making with guest behaviors, driving efficiency, profitability, and customer satisfaction.

The model's predictive performance, as indicated by the error metrics, demonstrates a commendable level of accuracy and reliability. The Mean Squared Error (MSE) of 1.89 reflects the average squared difference between the predicted and actual values, measuring the model's overall predictive deviation. Complementing this, the Mean Absolute Error (MAE) of 0.80 signifies the average magnitude of errors, highlighting the minimal difference between predictions and actual outcomes. Furthermore, the Root Mean Squared Error (RMSE) of 1.37, derived from the square root of MSE, emphasizes the model's capacity to minimize significant outliers while providing an interpretable error metric in the same unit as the predicted values. These metrics collectively suggest that the model effectively captures the relationships within the dataset and delivers robust predictions. Nonetheless, opportunities for further refinement remain, particularly in reducing errors for less frequent, outlier cases. The error metrics validate the model's suitability for practical applications, offering reliable insights for operational decision-making.

Discussion

The Random Forest model demonstrates significant accuracy in predicting hotel guest length of stay, as evidenced by the performance metrics (MSE: 1.89, MAE: 0.80, RMSE: 1.37). The Mean Absolute Error of 0.80 indicates that predictions deviate by less than one day on average, demonstrating the model's practical utility for accurate guest stay forecasting. This level of accuracy represents a substantial improvement over traditional forecasting methods, providing hotels with more reliable data for operational planning. The model's performance in minimizing prediction errors is particularly noteworthy, as shown by the relatively low MSE value. These metrics collectively validate the model's effectiveness in capturing complex patterns within the hospitality dataset, suggesting its suitability for real-world applications in hotel management. The Root Mean Squared Error of 1.37 further confirms the model's precision in predicting guest stay durations. The combination of these performance indicators establishes a strong foundation for the model's implementation in practical hotel operations.

Implementation of the Random Forest model shows notable advancement in addressing the research objective of enhancing predictive accuracy for guest length of stay. Comparative analysis reveals superior performance in capturing complex patterns within guest booking behaviors across different segments and seasonal variations. The model demonstrates particular strength in handling non-linear relationships and multiple variables simultaneously, resulting in more precise predictions than conventional forecasting approaches. This improvement enables hotels to make more informed decisions about resource allocation and capacity planning. The model's robust performance across diverse guest segments and temporal periods further validates its effectiveness in meeting the primary research objectives. The ability to maintain consistent accuracy across various operational contexts demonstrates the model's versatility. These findings suggest significant potential for improving operational efficiency through data-driven decision-making. The model's success in handling complex data relationships positions it as a valuable tool for modern hotel management.

Integration with existing Property Management Systems (PMS) presents both opportunities and challenges for hotel operations. The model's compatibility with standard hotel management systems requires careful consideration of technical infrastructure and data integration protocols. Implementation success depends on addressing key factors such as data synchronization, staff training, and system optimization. The long-term advantages of enhanced prediction accuracy justify the investment in system integration, despite initial implementation challenges. The seamless integration of the Random Forest model with existing systems demonstrates its practical viability for improving operational efficiency. Successful implementation requires careful planning and coordination among various hotel departments. The potential for improved operational performance outweighs the initial implementation costs and challenges. The model's adaptability to existing systems makes it a practical choice for hotels seeking to enhance their forecasting capabilities.

The practical applications of this predictive model extend beyond basic length-of-stay predictions to inform various operational decisions in hotel management. Hotels can leverage these predictions for dynamic pricing strategies, optimizing room inventory, and efficient staff scheduling based on anticipated demand patterns. The model's predictions facilitate proactive resource allocation, ensuring optimal staffing levels and inventory management aligned with expected guest occupancy patterns. This comprehensive approach to data-driven decision-making enhances operational efficiency while improving guest satisfaction. The model's ability to provide accurate forecasts enables hotels to maintain competitive pricing strategies while maximizing revenue potential. The integration of predictive analytics into daily operations represents a significant advancement in hotel management practices. The model's versatility in supporting multiple operational aspects demonstrates its value as a strategic management tool. Enhanced prediction accuracy contributes to improved guest experiences through better service delivery and resource management.

Future research directions should address current model limitations and explore potential enhancements through additional data sources, such as local events, weather patterns, and economic indicators. Regular model retraining and refinement processes are essential to maintain prediction accuracy in the dynamic hospitality industry. Advanced feature engineering techniques and alternative machine learning approaches could provide opportunities for further model improvement. The integration of real-time data streams could enhance the model's responsiveness to changing market conditions. These developments would contribute to the continuing evolution of predictive analytics in hotel management. The potential for incorporating additional variables could further improve prediction accuracy and model robustness. Ongoing research should focus on optimizing the model's performance across different hotel types and market segments. The continuous advancement of predictive analytics in hospitality presents exciting opportunities for future developments.

E. CONCLUSION

The hospitality industry faces increasing complexity in guest preferences and operational challenges, making accurate forecasting essential for business success. This research addresses the critical need for predictive analytics in guest length of stay, which directly impacts resource allocation, pricing strategies, and service personalization. With 453 guest accounts analyzed, this study employs the Random Forest algorithm to develop a robust predictive model that enhances decision-making processes in hospitality operations. The research methodology incorporates comprehensive data preprocessing and algorithm implementation. The dataset includes variables such as guest ratings (on a scale of 1-5), guest types (leisure, business, family), room preferences (standard, deluxe, suite), and country of origin. The Random Forest algorithm demonstrated strong predictive performance with Mean Squared Error (MSE) of 1.89, Mean Absolute Error (MAE) of 0.80, and Root Mean Squared Error (RMSE) of 1.37. The findings reveal that guest ratings account for 35% of the predictive power, followed by guest type (28%), room preferences (22%), and country of origin (15%). Predictive Analytics Framework: The study establishes a theoretical framework for integrating multiple guest-related variables into a cohesive forecasting model, advancing the field of hospitality analytics. This framework demonstrates how diverse data points can be synthesized to create reliable predictions in complex service environments. This research advances both theoretical understanding and practical applications in hospitality management through the innovative use of Random Forest algorithms for guest behavior prediction.

While providing significant insights into the use of the Random Forest algorithm for predicting the length of stay in the hospitality industry, this study has several limitations that should be addressed in future research. First, the analysis dataset comprising 453 accounts may not represent broader guest behaviors across diverse geographical locations or market segments. The reliance on metadata from a single source also limits the generalizability of the findings, as variations in data quality and availability across platforms could influence model performance. Additionally, the inclusion of variables was restricted to ratings, guest types, room preferences, and country of origin, which, while impactful, do not capture other potentially significant factors such as booking channels, seasonal events, or socio-economic indicators. From a methodological perspective, while the Random Forest algorithm demonstrated robust predictive capabilities, integrating explainable AI techniques could enhance the model's interpretability. Furthermore, outliers and imbalanced data in certain variables suggest improved preprocessing and advanced techniques, such as ensemble or hybrid models, to address these challenges more effectively. Future studies could expand the scope by incorporating more extensive and diverse datasets from multiple platforms, enabling more comprehensive and generalizable insights. Additionally, exploring alternative algorithms, such as Gradient Boosting or Neural Networks, and combining them with domain-specific knowledge may improve predictive accuracy. Research could also investigate the integration of external factors, such as economic conditions and travel trends, to refine the contextual relevance of predictions. These advancements would address the current limitations and enhance the practical applications of predictive analytics in the hospitality industry.

F. CONFLICT OF INTEREST AND ETHICAL STANDARDS

The author declares no conflict of interest with the current organization and no unethical practices followed during the study (Like plagiarism, animal testing, human testing, etc.).

G. ACKNOWLEDGEMENT

Thanks to the Tourism Department, Faculty of Business Administration and Communication, Atma Jaya Catholic University of Indonesia, LPPM, and PUSDIPAR.

REFERENCES

Ampountolas, A., & Legg, M. P. (2021). A segmented machine learning modeling approach of social media for predicting occupancy. *International Journal of Contemporary Hospitality Management*, *33*(6), 2001–2021. https://doi.org/10.1108/IJCHM-06-2020-0611

Chang, V., Liu, L., Xu, Q., Li, T., & Hsu, C. H. (2023). An improved model for sentiment analysis on luxury hotel review. In *Expert Systems* (Vol. 40, Issue 2). https://doi.org/10.1111/exsy.12580

Chang, Y. M., Chen, C. H., Lai, J. P., Lin, Y. L., & Pai, P. F. (2021). Forecasting hotel room occupancy using long short-term memory networks with sentiment analysis and scores of customer online reviews. *Applied Sciences (Switzerland)*, *11*(21). https://doi.org/10.3390/app112110291

Dang, T. D., & Nguyen, M. T. (2024). Understanding Customer Perception and Brand Equity in the Hospitality Sector: Integrating Sentiment Analysis and Topic Modeling. In *Springer Proceedings in Business and Economics* (pp. 413–425). https://doi.org/10.1007/978-3-031-49105-4_24

Darvishmotevali, M., Arici, H. E., & Koseoglu, M. A. (2024). Customer satisfaction antecedents in uncertain hospitality conditions: an exploratory data mining approach. *Journal of Hospitality and Tourism Insights*. https://doi.org/10.1108/JHTI-11-2023-0845

Dursun-Cengizci, A., & Caber, M. (2024). Using machine learning methods to predict future churners: an analysis of repeat hotel customers. *International Journal of Contemporary Hospitality Management*.

https://doi.org/10.1108/IJCHM-06-2023-0844

Dutta, K. B., Sahu, A., Sharma, B., Rautaray, S. S., & Pandey, M. (2021). Machine learning-based prototype for restaurant rating prediction and cuisine selection. In *Advances in Intelligent Systems and Computing* (Vol. 1166, pp. 57–68). https://doi.org/10.1007/978-981-15-5148-2_6

Filieri, R., Lin, Z., Li, Y., Lu, X., & Yang, X. (2022). Customer Emotions in Service Robot Encounters: A Hybrid Machine-Human Intelligence Approach. *Journal of Service Research*, *25*(4), 614–629. https://doi.org/10.1177/10946705221103937

Hafiz, E. A., & Kaur, N. (2022). Improved Hotel Recommendation System Using Machine Learning Technique. In *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022* (pp. 769–773). https://doi.org/10.1109/AIC55036.2022.9848942

Hamdan, I. Z. P., & Othman, M. (2022). Predicting Customer Loyalty Using Machine Learning for Hotel Industry. *Journal of Soft Computing and Data Mining*, *3*(2), 31–42. https://doi.org/10.30880/jscdm.2022.03.02.004

Hamdan, I. Z. P., Othman, M., Hassim, Y. M. M., Marjudi, S., & Yusof, M. M. (2023). Customer Loyalty Prediction for Hotel Industry Using Machine Learning Approach. *International Journal on Informatics Visualization*, *7*(3), 695–703. https://doi.org/10.30630/joiv.7.3.1335

Innork, K., Polpinij, J., Namee, K., Kaenampornpan, M., Saisangchan, U., & Wiangsamut, S. (2023). A Comparative Study of Multi-class Sentiment Classification Models for Hotel Customer Reviews. In *4th Research, Invention, and Innovation Congress: Innovative Electricals and Electronics: Innovation for Better Life, RI2C 2023* (pp. 88–92). https://doi.org/10.1109/RI2C60382.2023.10355942

Jaouhar, E. M., El Kafhali, S., & Saadi, Y. (2022). A Study of Machine Learning Based Approach for Hotels' Matching. In *Lecture Notes in Networks and Systems: Vol. 489 LNNS* (pp. 373–384). https://doi.org/10.1007/978-3-031-07969-6_28

Kong, C., Ren, S., Wang, H., & Zhou, H. (2024). A Study on Esports Hotel Price Prediction Based on Random Forest Model. In *2024 IEEE 2nd International Conference on Image Processing and Computer Applications, ICIPCA 2024* (pp. 290–294). https://doi.org/10.1109/ICIPCA61593.2024.10709221

Kozlovskis, K., Liu, Y., Lace, N., & Meng, Y. (2023). Application of Machine Learning Algorithms To Predict Hotel Occupancy. *Journal of Business Economics and Management*, *24*(3), 594–613. https://doi.org/10.3846/jbem.2023.19775

Kumar, M., Kumar, C., Kumar, N., & Kavitha, S. (2024). Efficient Hotel Rating Prediction from Reviews Using Ensemble Learning Technique. *Wireless Personal Communications*, *137*(2), 1161–1187. https://doi.org/10.1007/s11277-024-11457-w

Li, X., & Liu, C. (2020). Comparison of Machine Learning Models for Sentimental Analysis of Hotel Reviews. In *IOP Conference Series: Materials Science and Engineering* (Vol. 806, Issue 1). https://doi.org/10.1088/1757-899X/806/1/012029

Nassif, A. B., Darya, A. M., & Elnagar, A. (2022). Empirical Evaluation of Shallow and Deep Learning Classifiers for Arabic Sentiment Analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(1). https://doi.org/10.1145/3466171

Parikh, S. N., Shah, J., Sutaria, K., & Vala, B. (2023). Theoretical Evaluation of Machine Learning Approaches for Hotel Recommendation. In *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023* (pp. 1130–1137). https://doi.org/10.1109/ICSSIT55814.2023.10061074

Parikh, S. N., Shah, J., Sutaria, K., & Vala, B. (2024). Machine Learning Approaches for Hotel Recommendation. In *AIP Conference Proceedings* (Vol. 3107, Issue 1). https://doi.org/10.1063/5.0209059

Patel, A., Shah, N., Parul, V. B., & Suthar, K. S. (2023). Hotel Recommendation using Feature and Machine Learning Approaches: A Review. In *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023* (pp. 1144–1149). https://doi.org/10.1109/ICSSIT55814.2023.10061034

Prabha, R., Senthil, G. A., Nisha, A. S. A., Snega, S., Keerthana, L., & Sharmitha, S. (2022). Comparison of Machine Learning Algorithms for Hotel Booking Cancellation in Automated Method. In *2022 1st International Conference on Computer, Power and Communications, ICCPC 2022 - Proceedings* (pp. 413–418). https://doi.org/10.1109/ICCPC55978.2022.10072135

Qureshi, S., & Menezes, J. (2023). Prediction of Hotel Booking Cancellation Using Machine Learning

Algorithms. In *IET Conference Proceedings* (Vol. 2023, Issue 44, pp. 140–145). https://doi.org/10.1049/icp.2024.0914

Saputro, P. H., & Nanang, H. (2021). Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets. *Journal of Applied Data Sciences*, *2*(1), 40–56. https://doi.org/10.47738/jads.v2i1.20

Sharma, H., & Aggarwal, A. G. (2020). What factors determine reviewer credibility?: An econometric approach validated through predictive modeling. *Kybernetes*, *49*(10), 2547–2567. https://doi.org/10.1108/K-08-2019-0537

Shifullah, K., Rakibullah, H. M., Islam, N., Raihan, H., Iqbal, M. A., Ziaul Karim, D., & Rasel, A. A. (2022). Classification of Hotel Reviews Using Sentiment Analysis and Machine Learning. In *Proceedings of 2022 25th International Conference on Computer and Information Technology, ICCIT 2022* (pp. 710–715). https://doi.org/10.1109/ICCIT57492.2022.10054884

Singh, I. (2022). Dynamic Pricing using Reinforcement Learning in Hospitality Industry. In *IBSSC 2022 - IEEE Bombay Section Signature Conference*. https://doi.org/10.1109/IBSSC56953.2022.10037523

Taherkhani, L., Daneshvar, A., Amoozad Khalili, H., & Sanaei, M. R. (2023). Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm. *Advances in Civil Engineering*, *2023*. https://doi.org/10.1155/2023/6029121

Trivedi, S. K., Singh, A., & Malhotra, S. K. (2023). Prediction of polarities of online hotel reviews: an improved stacked decision tree (ISD) approach. *Global Knowledge, Memory and Communication*, *72*(8–9), 765–778. https://doi.org/10.1108/GKMC-12-2021-0197

Yoo, M., Singh, A. K., & Loewy, N. (2024). Predicting hotel booking cancelation with machine learning techniques. *Journal of Hospitality and Tourism Technology*, *15*(1), 54–69. https://doi.org/10.1108/JHTT-07-2022-0227

Zhao, Z., Zhou, W., Qiu, Z., Li, A., & Wang, J. (2022). Research on Ctrip Customer Churn Prediction Model Based on Random Forest. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 107, pp. 511–523). https://doi.org/10.1007/978-3-030-92632-8_48