Received September 29th, 2024; accepted December 23th, 2024. Date of publication December 30th, 2024
Digital Object Identifier: https://doi.org/10/25047/jtit.v11i2.5440

Development of Machine Learning Model to Predict Hotel Room Reservation Cancellations

EKA RAHMAWATI[1], GALIH SETIAWAN NUROHIM[2], CANDRA AGUSTINA[3], DENNY IRAWAN[4], ZAINAL MUTTAQIN[5]

[1,2,3,4,5]Universitas Bina Sarana Informatika, Jakarta, Indonesia

CORESPONDING AUTHOR: EKA RAHMAWATI (email:eka.eat@bsi.ac.id)

ABSTRACT The frequent cancellations of hotel room reservations have become a pressing issue for the hospitality industry, especially in high-tourism areas such as Borobudur, Indonesia. This research develops a predictive machine learning (ML) model to identify cancellation probabilities to support proactive decision-making for hotel management. Using datasets from Borobudur-based hotels, key variables such as booking lead time, arrival month, and reservation outcomes were analyzed. Random Forest demonstrated the best performance, achieving an accuracy of 86.36% with a precision of 88.06%, recall of 93.65%, and F1-score of 90.77%. Logistic Regression demonstrated moderate effectiveness, while Bayesian Networks underperformed, highlighting the importance of robust algorithms for such tasks. The findings underscore the potential of ML models, particularly Random Forest, to reduce financial losses and enhance operational efficiency in the hospitality sector by anticipating cancellations and facilitating better resource allocation.

KEYWORDS: Hotel reservation cancellations; machine learning; Random Forest; Logistic Regression; Bayesian Networks

I. INTRODUCTION

The hospitality industry is one of the sectors in tourism that have a crucial role in supporting the economy, particularly in popular tourist destinations such as the Borobudur area. However, one of the biggest challenges faced by hotel management is room reservation cancellations [1]. Sudden cancellations, especially those made close to the check-in date, can lead to significant financial losses, reduce operational efficiency, and impact hotel occupancy rates. In many cases, canceled rooms are difficult to resell on short notice, ultimately leading to underutilization of hotel resources.

To mitigate the negative impact of reservation cancellations, predicting cancellations becomes an important solution for hotel management [2]. By having early information on the likelihood of cancellations, hotel managers can implement more effective strategies such as overbooking, dynamic pricing, and offering additional services to reduce the risk of cancellations.

Machine learning (ML)-based prediction techniques have been used in recent years to improve the accuracy of cancellation predictions in the hospitality industry[3][4][5]. ML models can analyze various factors influencing cancellation decisions, such as stay duration, channel distribution, room rate average, reservation day, day of arrival, lead time, and payment conditions[6]. One of the most promising approaches is integrating Bayesian Networks (BN) with machine learning methods, allowing probabilistic analysis and more transparent interaction between predictor variables. This approach not only improves prediction accuracy but also provides clearer interpretation of the factors contributing to cancellation decisions. Moreover, research involving the implementation of Logistic Regression and K-Nearest Neighbors has also been conducted, demonstrating that such studies can assist hotels in increasing their revenue [7].

This research focuses on developing a hotel room cancellation prediction model using data from hotels around the Borobudur area, Indonesia. By utilizing ML techniques integrated with interpretable feature interactions, this study aims to help hotel managers predict cancellations more accurately, thus reducing potential losses and improving operational efficiency. Several popular machine learning models can be used for prediction. This study will implement Random Forest, Logistic Regression and Bayesian Network models. Existing studies have predominantly applied general-purpose machine learning methods without tailoring them to the specific conditions of regional tourism hubs like Borobudur. This research contributes by optimizing models for local hotel data, incorporating unique variables such as arrival month and booking lead time.

Random Forest is a machine learning approach that utilizes ensemble methods, making it highly effective for tasks involving classification and regression[8]. This algorithm operates by constructing a series of decision trees during the learning phase. For classification problems, it determines the output by selecting the class with the highest frequency among the trees, while for regression tasks, it computes the mean of the predictions generated by the trees [9]. Each decision tree is trained on a randomly selected subset of the data and a randomly chosen set of features, enhancing the model's resilience and minimizing the likelihood of overfitting[10].

Random Forest is highly valued to handle large datasets with higher dimensionality, its resilience to overfitting, especially when working with noisy data, and its ability to maintain accuracy even with missing data[11]. Additionally, it provides feature importance rankings, offering insights into which features contribute most to the predictive power of the model. This algorithm is computationally efficient in parallel processing environments, making it scalable to larger datasets, and it consistently delivers strong predictive performance across a wide range of applications.

The model that designed for binary classification task is Logistic Regression. The model can have one of two possible outcomes as the target of classification[12]. The model estimates the likelihood of a binary outcome by employing the logistic function, widely known as the sigmoid function[13]. This approach ensures that the predicted probabilities lie between 0 and 1, making it ideal for classification tasks. The key concept of logistic regression is to find a linear combination of the input features to predict the probability of an event occurring (such as classifying a label as 0 or 1).

Bayesian networks are graphical probabilistic models employed in machine learning to represent relationships and dependencies among a set of variables[14][15]. They use Bayes' Theorem to compute conditional probabilities, which is essential in dealing with uncertainty in a structured way.

The structure of Bayesian Networks consists of two key components: a Directed Acyclic Graph (DAG) and Conditional Probability Distributions (CPDs) [16]. In the DAG, nodes represent random variables, which can be either discrete or continuous, while edges indicate conditional dependencies between these variables. An edge from node X to node Y signifies that Y is dependent on X, establishing X as a parent of Y. Each node in the network is further defined by its CPD, which quantifies the influence of its parent nodes on the variable it represents. For instance, for a node Y with parents X1, X2, ..., Xn, the CPD is expressed as P(Y | X1, X2, ..., Xn), capturing the conditional relationships within the network.

The primary strengths of Bayesian Networks lie in their capacity to systematically manage uncertainty through probabilistic reasoning [17]. The modular nature of these networks allows for independent treatment of different components, which significantly enhances computational efficiency [18]. Furthermore, the graphical representation of variable dependencies improves interpretability, a feature that is particularly valuable in complex domains such as medicine, genetics, and diagnostics. Additionally, Bayesian Networks exhibit versatility in learning tasks, being applicable in both supervised and unsupervised contexts [19]. They possess the ability to infer both the network structure and the corresponding parameters from data, or alternatively, they can be constructed based on expert knowledge.

II. METHOD

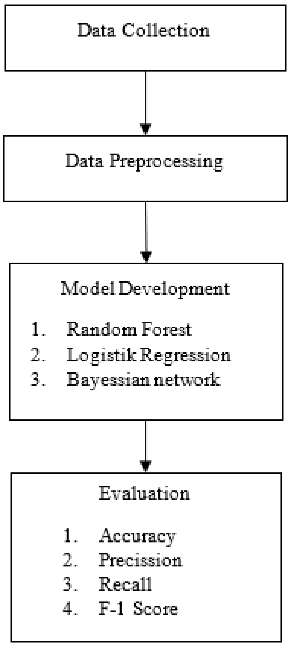

FIGURE 1. Research Methodology.

The methodological steps in this study are described in Figure 1: Stage 1 (Data Collection), Stage 2 (Data Preprocessing), Stage 3 (Model Development) and Stage 4 (Model Evaluation). These stages are designed to systematically address the research objectives and ensure robust and interpretable results.

1. Data Collection

The dataset used in this research contains reservation records from hotels in the Borobudur area. The data consists of 6 columns and 293 samples, representing various attributes related to bookings and cancellations. Table 1 show the data atrributes.

TABLE 1. Key Attributes

Attribute	Description
no_of_week_nights	Number of weeknights booked by the customer.
required_car_parking_space	Indicates whether a parking space was requested (0: No, 1: Yes).
lead_time	The duration in days between the date a booking is made and the scheduled arrival date.
arrival_month	The month when the customer is scheduled to arrive, represented numerically (e.g., 1 for January, 2 for February, and so on up to 12 for December).
market_segment_type	Source of the booking (e.g., Online, Offline, Corporate, Complementary).
booking_status	Indicates the reservation status (Canceled or Not_Canceled).

Then, the visualizations of key features are shown in Figures 2 and 3.

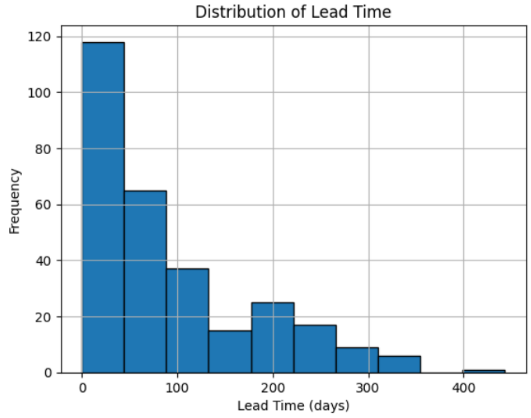

FIGURE 2. Histogram and Leadtime Distribution.

Figure 2 depicts the distribution of lead_time, which indicates the number of days between the booking date and the scheduled arrival date. The data shows a high frequency of bookings with short lead times, particularly within the first 50 days, indicating that most customers make reservations closer to their check-in date. As lead time increases, the frequency of bookings decreases with very few instances exceeding 300 days. This pattern highlights the tendency of customers to book reservations on relatively short notice. Understanding this distribution is crucial for hotel management to optimize marketing strategies, such as offering early booking discounts or managing inventory for last-minute bookings effectively. This insight supports the development of predictive models by emphasizing the importance of lead time as a key feature influencing booking behaviors.

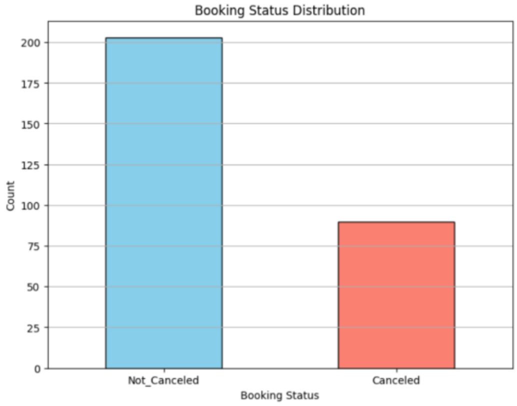

FIGURE 3. Distribution of Booking Status.

The chart presented in Figure 3 illustrates the distribution of booking statuses within the dataset, distinctly categorized into two groups: Not_Canceled and Canceled. This visualization highlights a notable imbalance, with the Not_Canceled category representing the majority of reservations, exceeding 200 instances. In contrast, the Canceled category is significantly smaller, amounting to less than half the count of the Not_Canceled reservations. This disparity provides essential insight into the dataset's characteristics, underscoring the predominance of successful bookings over cancellations.

Understanding such an uneven distribution is crucial for developing predictive models capable of addressing the needs of hotel management effectively. Models trained on this dataset must account for the imbalance to ensure accurate predictions and minimize biases that could favor the majority class. By tailoring predictive approaches to handle this discrepancy, hotels can better anticipate booking behaviors, improve resource allocation, and enhance decision-making processes to optimize operational efficiency..

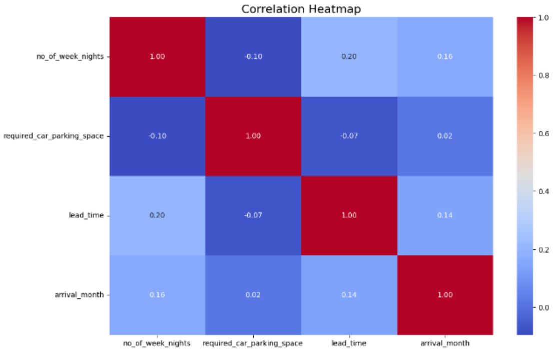

FIGURE 4. Correlation Heatmap.

Figure 4 resents the Pearson correlation coefficients between the numerical variables in the hotel reservation cancellation dataset. The diagonal values (1.00) indicate the perfect correlation of each variable with itself. Observing the relationships, lead_time shows a moderate positive correlation of 0.20 with no_of_week_nights, suggesting that longer booking lead times may influence longer stays. Meanwhile, required_car_parking_space exhibits very low correlations with other variables (close to 0), indicating negligible or insignificant relationships. Similarly, arrival_month does not show a strong correlation with any variables, with the highest value being 0.16.

The heatmap uses a color gradient where red signifies stronger positive correlations and blue indicates weak or negative correlations. This visualization helps identify the degree of linear relationships between variables, which is particularly useful for feature selection in machine learning models. By highlighting the most and least correlated variables, the heatmap provides insights into which features might contribute meaningfully to predictive tasks.

2. Data Preprocessing

The collected data must be cleaned and processed before being input into the models. The objective of data preprocessing is to ensure that the data is ready for modeling by addressing issues such as imbalance or noise[20]. This step includes:

a. Addressing Missing Values

Missing values were handled using mean and median imputation based on the distribution of the data. This method was chosen to retain the dataset's integrity without introducing significant bias, as the proportion of missing values was less than 5% for most features.

b. Encoding Categorical Variables

Converting variables such as room type, payment method, and guest country of origin into numerical form using techniques like One-Hot Encoding.

c. Splitting Data

The dataset was divided into testing (20%) and training (80%) subsets. Cross-validation was not performed due to computational constraints and the availability of a large testing set, which provides sufficient evaluation robustness.

3. Model Development

The selected machine learning models are then developed to predict reservation cancellations. The goal of this stage is to develop models capable of accurately predicting cancellations while offering insights into the factors that influence these outcomes. The methods include Random Forest, Logistic Regression and Bayesian Networks. Random Forest was selected because it effectively handles datasets with non-linear relationships and imbalanced data. This ensemble learning method is particularly robust in scenarios where datasets exhibit complex patterns, making it an ideal choice for predicting hotel reservation cancellations. Logistic Regression, on the other hand, serves as a simple yet effective model often employed for preliminary analysis due to its straightforward implementation and interpretability. Lastly, Bayesian Networks were incorporated to understand probabilistic relationships between variables, offering clear and interpretable insights into decision-making processes in the hotel industry.

4. Evaluation

After developing the models, they are trained with training data and tested with testing data. The objective of evaluation is to assess the model's performance based on its prediction on unseen data. The models are evaluated using several performance metrics:

a. Accuracy (ACC)

Accuracy (ACC) is a key performance metric used to evaluate the overall effectiveness of a predictive model by determining the proportion of correct predictions it makes out of all the predictions. It reflects the model's ability to classify instances correctly, whether they belong to the positive or negative class. Accuracy is calculated as the ratio of the total number of correctly predicted instances to the total number of instances in the dataset. While it provides a straightforward measure of a model's performance, accuracy alone may not always be sufficient in cases where the dataset is imbalanced, as it can lead to misleading conclusions by overemphasizing the majority class.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

ACC	= Accuracy
TP	= True Positives
FN	= False Negatives
FP	= False Positives
FP	= False Positives

b. Precision (P)

Precision (P) is a critical metric that measures the model's ability to accurately identify true positives—in this case, correctly predicting cancellations—relative to the total number of instances it classified as cancellations. Precision is calculated as the ratio of true positive predictions to the sum of true positives and false positives, reflecting the model's reliability in making positive predictions. A high precision score indicates that the model minimizes false alarms, ensuring that the cancellations it predicts are mostly accurate. This metric is particularly important in scenarios where the cost of false positives, such as wrongly classifying a non-canceled booking as canceled, can lead to inefficiencies or missed opportunities.

$$P = \frac{TP}{TP + FP}$$

Where:
P = Precision
TP = True Positives
FP = False Positives
FP = False Positives

c. Recall (R)

Recall (R), also known as sensitivity or the true positive rate, is a critical performance metric used to evaluate a model's ability to detect all actual occurrences of a desired category, in this case, booking cancellations. Mathematically, recall is defined as the ratio of true positives (TP) to the total sum of true positives (TP) and false negatives (FN). A high recall value indicates that the model effectively identifies the majority of cancellation cases while minimizing the number of false negatives. This metric is particularly significant in contexts where accurately detecting all relevant instances, such as cancellations, has substantial implications for decision-making processes. For example, in hotel management, high recall ensures efficient room reallocation and better resource planning. However, recall must be interpreted alongside other metrics, such as precision, to avoid a disproportionate number of false positives, which could undermine the reliability of predictions. Consequently, recall is often combined with precision and summarized using the F1-score to provide a more comprehensive evaluation of a model's overall performance. The model's ability to detect actual cancellations.

$$R = \frac{TP}{TP + FN}$$

Where:
R = Recall
TP = True Positives
FN = False Negatives
FP = False Positives
FP = False Positives

d. F1-Score (F1)

The F1-Score (F1) is a crucial metric that represents a harmonic mean between precision and recall, providing a balanced measure of a model's performance in scenarios where both metrics are equally important. This metric is particularly valuable in situations where there is an inherent trade-off between precision and recall, ensuring that neither is disproportionately favored. A high F1-Score indicates that the model achieves a good balance, excelling in both accurately predicting true positives and minimizing false positives or false negatives. By incorporating both metrics, the F1-Score offers a comprehensive evaluation of the model's ability to handle imbalanced datasets or critical decision-making tasks, making it a standard choice for assessing classification performance in many real-world applications.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

Where:
F1 = F1-Score
P = Precision
TP = True Positives
FN = False Negatives
FP = False Positives
FP = False Positives

Default parameters were used during the initial model training to establish baseline performance. For Random Forest, key parameters such as the number of estimators and maximum depth were tuned iteratively to optimize performance. Logistic Regression was adjusted for regularization strength and the Naive Bayes model was evaluated with its default settings as it has fewer tunable parameters.

This setup ensures that the methods and parameter choices are well-documented, providing a clear pathway for replication and further optimization by readers. Default parameters were used during initial model training, followed by hyperparameter tuning for optimization. Hyperparameter settings included the number of estimators and maximum depth for Random Forest, regularization strength for Logistic Regression, and prior probabilities for Bayesian Networks. These settings ensured reproducibility and maximized model performance.

III. RESULT AND DISCUSSION

Predictive analysis for hotel reservation cancellations using Random Forest, Logistic Regression, and Bayesian Network was conducted. The evaluation metrics used to compare the models with Acc, P, R, and F1-Score.

TABLE 2. Performance Metrics of Random Forest

Metric	Value
ACC (Accuracy)	0.863636
P (Precision)	0.880597
R (Recall)	0.936508
F1-Score	0.907692

From the results presented in Table 2, the Random Forest model demonstrated the highest performance across most metrics, with an Accuracy of 86.36%, a Precision of 88.06%, a Recall of 93.65%, and an F1 Score of 90.77%. This indicates that Random Forest is the best-performing model for this prediction task.

TABLE 3. Performance Metrics of Logistic Regression

Metric	Value
ACC (Accuracy)	0.795455
P (Precision)	0.816901
R (Recall)	0.920635
F1-Score	0.865672

Table 3 shows that the Logistic Regression model also have good performance with an Accuracy of 79.55%, a Precision of 81.69%, and a Recall of 92.06%, leading to an F1 Score of 86.57%. Despite a slightly lower performance than Random Forest, Logistic Regression still showed strong predictive capabilities, particularly in recall. This suggests that it effectively identified the most positive instances of reservation cancellations.

TABLE 4. Performance Metrics of Bayessian Network

Metric	Value
ACC (Accuracy)	0.602273
P (Precision)	1.000000
R (Recall)	0.444444
F1-Score	0.615385

The results of the Bayesian Network model at Table 4 performed significantly worse, with an Accuracy of only 60.23%. While the model achieved perfect Precision (1.000), its Recall was extremely low (44.44%), indicating that it failed to correctly identify the majority of cancellation instances. This resulted in a lower F1 Score of 61.54%. This suggests that while the Bayesian Network model was highly conservative in its predictions (i.e., it produced few false positives), it could not generalize to a larger set of cancellation cases, making it unsuitable for this task.

The Random Forest model outperformed the other two algorithms in this study. Random Forest is known for its robustness in handling classification tasks, especially in datasets with complex and non-linear relationships. The high recall and precision values indicate that this model was able to effectively identify both positive and negative cases of hotel reservation cancellations, making it the most reliable model for this specific dataset.

Random Forest's capacity to construct multiple decision trees and combine their outputs minimizes the risk of overfitting while improving generalization. Its balance between *P* and *R*, reflected in the high F1 score, underscores its effectiveness for predictive tasks with critical implications, such as hotel reservation cancellations. In such cases, both *FP* (predicting a cancellation that does not occur) and *FN* (failing to predict an actual cancellation) carry operational consequences for hotels, highlighting the model's practical relevance.

The Logistic Regression model showed good predictive power, although it was outperformed by Random Forest in every metric except for recall, where it achieved a slightly higher score of 92.06%. Logistic Regression is a simpler, linear model compared to Random Forest, and it may not capture the complex relationships in the data as effectively as Random Forest does. However, its relatively high recall indicates that it is capable of identifying a majority of the cancellations, though perhaps at the expense of introducing more false positives, as suggested by its lower precision.

Given the simplicity of Logistic Regression, its performance here is still commendable and may be preferred in situations where computational efficiency or interpretability is more critical than maximizing prediction accuracy.

The Bayesian Network model, while achieving perfect precision, suffered from a notably low recall. This suggests that while the model was highly conservative in predicting cancellations (leading to no false positives), it missed a substantial number of actual cancellations. This trade-off between precision and recall is problematic in contexts like hotel reservations, where missing a true cancellation could lead to operational inefficiencies, such as rooms being blocked unnecessarily or potential revenue loss.

The low accuracy and F1 Score further emphasize that the Bayesian Network model was not able to capture the complexities of the dataset. One possible explanation for this underperformance could be the inherent assumptions of conditional independence in Bayesian Networks, which might not hold in a real-world dataset like this one, where many factors affecting hotel cancellations (e.g., customer behaviour, external events) are interdependent.

Based on the evaluation results, the Random Forest model was selected as the optimal model for predicting hotel reservation cancellations in the area around Borobudur. The high performance across all metrics suggests that this model is both accurate and reliable for use in practical applications. In contrast, while Logistic Regression also showed potential, particularly in recall, the Bayesian Network model was found to be unsuitable for this task due to its poor generalization performance.

IV. CONCLUSSION

In this study, the Random Forest model demonstrated superior performance in predicting hotel reservation cancellations, achieving the highest scores across all evaluation metrics, including accuracy (86.36%), precision (88.06%), recall (93.65%), and F1 score (90.77%). This makes it the most reliable model for this task compared to Logistic Regression and Bayesian Networks. While Logistic Regression also showed good performance, particularly in recall, its overall effectiveness was slightly lower than Random Forest. On the other hand, the Bayesian Network underperformed, primarily due to its inability to generalize well to the data, as evidenced by its low recall and F1 score. Consequently, the Random Forest model has been selected as the best predictive model and has been saved for future use. This study highlights the importance of model selection in operational tasks like hotel reservation management, where accurate predictions can mitigate potential revenue losses and improve efficiency. Further research may explore additional advanced models and feature engineering to improve prediction outcomes.

ACKNOWLEDGMENT

We extend our heartfelt gratitude to the Indonesian Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) for their generous financial support, which has been pivotal in enabling the successful completion of our research on hotel room reservation cancellations using machine learning models. This funding has provided the necessary resources to advance our study and explore innovative solutions to challenges faced by the hospitality industry. The unwavering trust and support from Kemendikbudristek have not only facilitated the progression of this research but also fostered a spirit of innovation and academic excellence. We deeply appreciate their commitment to empowering researchers and driving impactful advancements in the field.

REFERENCE

[1] N. K. Hikmawati, Y. Ramdhani, and Wartika, “Exploring ADR Trends: A Data Mining Approach to Hotel Room Pricing, Cancellations, and EDA,” *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 189–202, Jan. 2024, doi: 10.47738/jads.v5i1.165.

[2] M. Yoo, A. K. Singh, and N. Loewy, “Predicting hotel booking cancelation with machine learning techniques,” *Journal of Hospitality and Tourism Technology*, vol. 15, no. 1, pp. 54–69, Jan. 2024, doi: 10.1108/JHTT-07-2022-0227.

[3] X. Tian, B. Pan, L. Bai, and D. Mo, “Md-Pred: A Multidimensional Hybrid Prediction Model Based on Machine Learning for Hotel Booking Cancellation Prediction,” *Intern J Pattern Recognit Artif Intell*, vol. 37, no. 5, Apr. 2023, doi: 10.1142/S0218001423510096.

[4] S. Chen, E. W. T. Ngai, Y. Ku, Z. Xu, X. Gou, and C. Zhang, “Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction,” *Decis Support Syst*, vol. 170, Jul. 2023, doi: 10.1016/j.dss.2023.113959.

[5] A. Herrera, Á. Arroyo, A. Jiménez, and Á. Herrero, “Forecasting hotel cancellations through machine learning,” *Expert Syst*, Sep. 2024, doi: 10.1111/exsy.13608.

[6] S. Chalupa and M. Petricek, “Understanding customer’s online booking intentions using hotel big data analysis,” *Journal of Vacation Marketing*, vol. 30, no. 1, pp. 110–122, Accessed: Sep. 14, 2024. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/13567667221122107

[7] M. F. Sholahuddin *et al.*, “Perbandingan Model Logistic Regression dan K-Nearest Neighbors Dalam Prediksi Pembatalan Hotel,” *SNESTIK Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika* , pp. 137–142, 2023, doi: 10.31284/p.snestik.2023.4040.

[8] Y. Zhang, J. Liu, and W. Shen, “A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications,” Sep. 01, 2022, *MDPI.* doi: 10.3390/app12178654.

[9] Z. Liu, K. W. De Bock, and L. Zhang, “Explainable profit-driven hotel booking cancellation prediction based on heterogeneous stacking-based ensemble classification,” *Eur J Oper Res*, Aug. 2024, doi: 10.1016/j.ejor.2024.08.026.

[10] E. Y. Boateng, J. Otoo, and D. A. Abaye, “Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review,” *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.

[11] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[12] A. L. Lynam *et al.*, “Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults,” *Diagn Progn Res*, vol. 4, no. 1, Dec. 2020, doi: 10.1186/s41512-020-00075-2.

[13] A. Zaidi and A. S. M. Al Luhayb, “Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression,” *Math Probl Eng*, vol. 2023, pp. 1–11, Apr. 2023, doi: 10.1155/2023/5525675.

[14] G. Briganti, M. Scutari, and R. J. McNally, “A Tutorial on Bayesian Networks for Psychopathology Researchers,” *Psychol Methods*, vol. 28, no. 4, pp. 947–961, Feb. 2022, doi: 10.1037/met0000479.

[15] B. Mihaljević, C. Bielza, and P. Larrañaga, “Bayesian networks for interpretable machine learning and optimization,” *Neurocomputing*, vol. 456, pp. 648–665, Oct. 2021, doi: 10.1016/j.neucom.2021.01.138.

[16] N. K. Kitson, A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, “A survey of Bayesian Network structure learning,” *Artif Intell Rev*, vol. 56, no. 8, pp. 8721–8814, Aug. 2023, doi: 10.1007/s10462-022-10351-w.

[17] J. Bharadiya and J. P. Bharadiya, “A Review of Bayesian Machine Learning Principles, Methods, and Applications,” *Article in International Journal of Innovative Research in Science Engineering and Technology*, vol. 8, no. 5, 2023, doi: 10.5281/zenodo.8002438.

[18] P. Santos *et al.*, “Recommending Words Using a Bayesian Network,” *Electronics (Switzerland)*, vol. 12, no. 10, May 2023, doi: 10.3390/electronics12102218.

[19] F. Castelletti, F. Niro, M. Denti, D. Tessera, and A. Pozzi, “Bayesian Learning of Causal Networks for Unsupervised Fault Diagnosis in Distributed Energy Systems,” *IEEE*

Access, vol. 12, pp. 61185–61197, 2024, doi: 10.1109/ACCESS.2024.3394046.

[20] S. Albahra *et al.*, "Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts," Mar. 01, 2023, *W.B. Saunders*. doi: 10.1053/j.semdp.2023.02.002.

EKA RAHMAWATI received her Bachelor's degree in Computer Science from Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Jakarta, in 2017, and her Master's degree in Computer Science from the same institution in 2019. She is currently pursuing her Doctorate degree in Information Systems at Universitas Diponegoro, Semarang, Indonesia. She is a lecturer in the Information Systems Department at Universitas Bina Sarana Informatika, Solo, Indonesia. Her research interests include information systems development, machine learning applications, and digital transformation in various industries. She can be contacted at email: eka.eat@bsi.ac.id.

GALIH SETIAWAN NUROHIM earned his Bachelor's degree in Computer Science from Universitas Amikom Yogyakarta in 2013 and his Master's degree in Computer Science from the same institution in 2017. He is currently a lecturer in the Information Systems Department at Universitas Bina Sarana Informatika, Solo, Indonesia, with the functional position of Lektor. His research interests focus on topics such as information systems development, data analytics, and the application of technology in education and industry. His commitment to advancing knowledge in these areas is reflected in his active participation in academic initiatives and contributions to the academic community. Galih remains dedicated to fostering innovation and excellence in the field of information systems.

CANDRA AGUSTINA received her Bachelor's degree in Computer Science from Universitas Dian Nuswantoro, Semarang, in 2004, and her Master's degree in Computer Science from STMIK Nusa Mandiri, Jakarta, in 2012. Currently, she is pursuing her Doctorate degree in Information Systems at Universitas Diponegoro (UNDIP), Semarang, Indonesia. She is a lecturer in the Information Systems Department at Universitas Bina Sarana Informatika, Indonesia. Her research interests include e-tourism, machine learning applications, digital transformation, and information systems development. She can be contacted at email: candra.caa@bsi.ac.id.

DENNY IRAWAN, photograph and biography not available at the time of publication.

ZAINAL MUTTAQIN, photograph and biography not available at the time of publication.