MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi   ISSN : 1978-161X(p); 2477-2550(e)
Volume 16, No. 2(2024), pp 122-127   DOI : 10.18860/mat.v16i2.28862
Received : August 16th 2024; **Accepted** : August 27th 2024 ; **Avalaible Online** : September 20th 2024

Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree for Hotel Reservation Cancellation Prediction Using Random Forest

Hamida Maulana Lailatal Baroah, Lukman Hakim

***Abstract*—This study focuses on predicting hotel booking cancellations using machine learning to improve accuracy and operational efficiency. The methods used include Random Forest for modeling and Artificial Bee Colony (ABC) and Gradient Boosting Decision Tree (GBDT) for feature selection. ABC, which excels in optimization but is prone to local optima, is combined with GBDT for feature selection. The dataset used is Hotel_Bookings from Kaggle, containing 119.390 entries and 28 features. The data was processed through cleansing, normalization, and split into 75% for training and 25% for testing. Model evaluation using a confusion matrix and metrics like precision, recall, f1-score, and accuracy shows that combining ABC and GBDT achieved an accuracy of 86.81%. Increasing the number of trees and selected features generally improved model performance, with feature selection showing significant improvements over models without feature selection.**

***Index Terms*— Artificial Bee Colony (ABC), Feature Selection, Gradient Boosting Decision Tree (GBDT), Machine Learning, Model Evaluation, Random Forest, Hotel Booking Cancellation Prediction.**

I. INTRODUCTION

The hospitality industry is one of the business sectors with great potential for growth and development. Competition in this industry has become increasingly fierce, influenced by factors such as service quality, room rates, and hotel amenities [1]. Customers play a critical role in this business, as they often make room reservations before their stay. However, one of the main challenges for hotel managers is the high cancellation rate[2].

Reservation cancellations can be triggered by various factors, such as changes in travel plans or better offers from other hotels. This not only causes financial losses but also disrupts hotel operational planning [3]. Therefore, developing a predictive model that canaccurately forecast cancellation probabilities is crucial for hotel management.

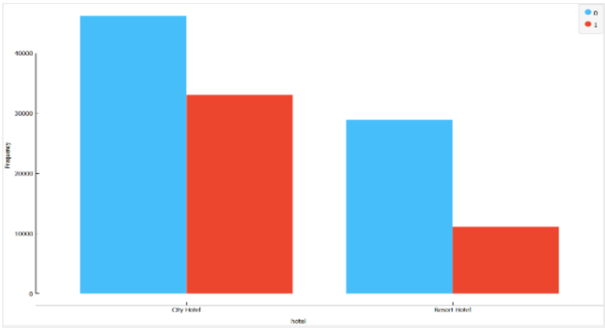

Fig 1 Amount Cancellation Hotel Orders

Figure 1 shows that city hotels experience higher cancellation rates compared to resort hotels[4]. Accurate predictions of hotel cancellations are essential for current revenue management systems. Accurate predictions allow hotels to plan operations more efficiently, take steps to reduce cancellation losses, and better understand consumer behavior for improved business strategies[5].

Machine learning has become an effective method for classification and is widely applied in various fields [6]. Previous studies have used different algorithms such as SVM, K-NN, and Logistic Regression to predict hotel cancellations with varying results[7][8]. However, there is still room for improvement in model accuracy and efficiency.

II. THEORETICAL FRAMEWORK

A. Prediction

Prediction is a process of forecasting or estimating what may happen in the future based on past and present information, with the aim of minimizing errors [11].

B. Hotel Reservation Cancellation

Hotel reservation cancellations pose a major challenge in the hospitality industry. Factors such as changes in travel plans and offers from

Manuscript received March 22, 2007. This work was supported in part by Department of Informatics Engineering at Yudharta University Pasuruan.
Hamida Maulana Lailatal Baroah is with the Department of Informatics Engineering at Yudharta University Pasuruan
Lukman Hakim is with Department of Informatics Engineering at Yudharta University Pasuruan.

competitors are common causes[3]. Understanding cancellation patterns helps hotels plan and strategize to mitigate risks.

C. Feature Selection

Feature selection is a preprocessing technique used to choose the most relevant and significant features, reducing irrelevant or redundant features. This helps improve model performance and prevents overfitting [12].

D. Artificial Bee Colony (ABC)

The Artificial Bee Colony algorithm is an optimization method inspired by the behavior of honey bee colonies searching for food sources. Artificial Bee Colony has a simple structure and few control parameters but is prone to local optima [13]. To address this weakness, Artificial Bee Colony is often combined with other algorithms.

E. Gradient Boosting Decision Tree (GBDT)

GBDT is an ensemble method that builds a predictive model by combining a series of weak decision tree classifiers into a strong one. GBDT effectively handles complex feature interactions and improves accuracy [14].

F. Random Forest

Random Forest is a machine learning algorithm that uses bagging to create multiple decision trees from different data subsets and combines the results. This reduces variance and prevents overfitting [15].

III. RESEARCH METHODOLOGY

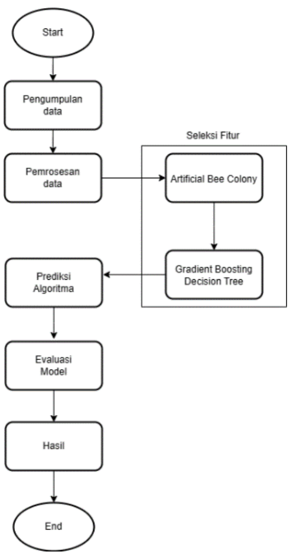

Fig 2 Research Flow Diagram

Figure 2 explains stages channel research, ie as following :

A. Data Collection

The dataset used is ***Hotel_Bookings*** from Kaggle, containing 119,390 data points with 28 features.

B. Data Prossesing

1. **Data Cleansing**: Columns such as 'country', 'agent', 'company', and 'reservation_status_date' were removed to reduce complexity and ensure data privacy.
2. **Data Normalization**: Categorical data was transformed into numerical values using one-hot encoding, and all features were scaled to ensure uniformity.
3. **Data Splitting**: The dataset was divided into 75% for training and 25% for testing.

C. Feature Selection

1. **Using ABC**: The ABC algorithm was used to search for the optimal subset of features through exploration and exploitation of the search space.
2. **Using GBDT**: GBDT was used to calculate feature importance and select the most impactful features for the model.

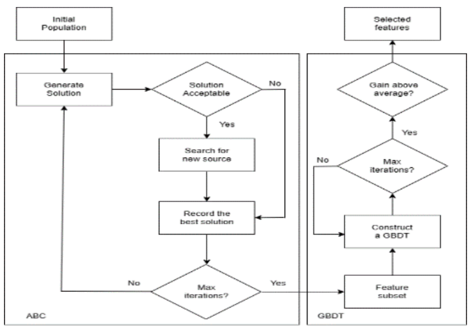

Fig 3 Flowchart Algorithm Selection feature

Figure 3 explains the diagram illustrates a combined process of the Artificial Bee Colony (ABC) and Gradient Boosting Decision Tree (GBDT) algorithms for feature selection and optimization. ABC generates solutions and iterates through a process of improving and recording the best solution until maximum iterations are reached. GBDT then refines the selected features by checking their performance and constructing the model, iterating until optimal feature subsets are achieved.

D. Prediction Algorithm

A ***Random Forest*** algorithm was used to build the predictive model based on the selected features, with tuned parameters to improve accuracy.

E. Evaluation

The model was evaluated using a confusion matrix and metrics such as accuracy, precision, recall, and f1-score.

a.) Confusion Matrix

To obtain the proportion of data predicted by the model compared to the actual labels from the existing data, you can refer to the values in the confusion matrix.

	Positive	Negative
Positive	True Positive	False Negatives
Negative	False Positive	True Negative

Fig 4 *Confusion Matrix*

In figure 4 it is explained that *confusion matrix*, generates mark *True Positive (TP)*, if the model outputs mark positive and the target dataset also outputs mark positive. Whereas produce mark *True Negative (TN)*, if the model outputs mark negative and the target dataset emit mark negative. If the model outputs mark positive but the target dataset emit mark negative, then mark the named *False Positive (FP)* (also known with *type I error*). And, if the model issues mark negative but the target dataset emit mark positive, then mark the named *False Negative (FN) (type II error)* [11].

b.) Precision
Precision state ratio predictions Correct compared to with whole predicted results cancelled. Following formula For measure Precision value :

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

c.) Recall
Percentage of classified data with Correct indicated by *recall*. Following formula For measure mark *Recall* :

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

d.) F1-Score
F1-score show average comparison between *precision* and *recall*. Following formula For measure mark *F1-Score* :

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

e.) Accuracy
Accuracy state ratio of total classified data Correct in test data. Following formula For measure mark *Accuracy* :

$$Accuracy = \frac{TP+TN}{P+N}$$

IV. Results and Discussion

A. Data Collection

On research This using the taken *Hotel_bookings* dataset from *Kaggle.com*, the amount of data from the dataset namely 119,390 data, has 28 features and 2 classes.
Hotel bookings (kaggle.com)

B. Data Processing

a.) Cleansing Data
Delete columns This Can So step in *the* data cleaning process for objective analysis or more modeling specific. This can also be done done for avoid redundancy or for reduce data dimensions so more easy For analyzed and processed by machine learning models.

Table 4.1 Cleansing Data

Features	*Type*
country	*meta*
agent	*meta*
company	*meta*
reservation_status_date	*meta*
meals	*categorical*
reserved_room_type	*categorical*
assigned_room_type	*categorical*
deposit_type	*categorical*
reservation_status	*categorical*

Removal of metadata such as *'country'*, *'agent'*, *'company'*, and *'reservation_status_date'* is possible considered for a number of reason important. First, from corner look privacy and security, this metadata possible containing information sensitive as can be endanger privacy user or company If misused. As for example, the *'agent'* and *'company'* metadata can be disclose identity the parties involved in something transaction or reservation, which is possible become a target for activity wicked or spam. Second, from perspective efficiency and performance, metadata does not relevant or worn can add burden storage and slows down the data management process. With remove metadata that does not Again required, system can work more faster and more efficient. Additionally, metadata is not relevant can cause chaos in search and analysis of data, creating it more difficult for find and use really information important. Therefore the, removal of this metadata can help guard data integrity, improve performance system, and protect privacy as well as security all parties involved .
Columns *'meal'*, *'reserved_room_type'*, *'assigned_room_type'*, *'deposit_type'*, *and 'reservation_status'* deleted direct without changed become dummy variable or No participate in the data normalization process because columns This considered no relevant or no enough informative, redundant or contain possible information obtained from another column, or own problem with data quality. Delete columns this help reduce complexity and dimensionality, so keep the model constant simple and more effective[3]

b.) Normalization Data
Normalization data aim for change data to in scale or the same distribution, so makes it easier algorithm for process and analyze data. Aspect important from *normalization data* is change existing data types to in an appropriate and consistent format. For example, numeric data possible need changed to in form scale certain, like range 0 to 1, or customized with normal distribution. Categorical data Possible need changed become numeric data type use technique like *one-hot encoding* or *label encoding*[12].

This process help in ensure that all feature in the dataset has balanced and equal contribution in the model, so avoid possible bias happen Because difference scale or data type. With change data types and commits normalization, we can also increase efficiency computational and model performance, as well make it easier interpretation results analysis.

Table 4.2 Initial Data

hotel	*arrival_date_month*	*market_segment*	*distribution_channel*	*customer_type*
Resort Hotel	January	Online	Direct	Transient
City Hotel	February	Offline	Corporate	Contract

Table 4.3 After one-hot encoding

Resort Hotel	*February*	*Online*	*Corporate*	*Contract*
0	0	1	0	0
1	1	0	1	1

Table 4.2 and Table 4.3 show that columns original has deleted, and now the data is there in form ready *numeric* used for analysis.

C. Split Data

Split data used for divide the dataset into two part, namely the training data *(training data) and* testing data *(testing data)*. Variable X containing features from the data, meanwhile variable y contains the desired label or target predicted.

On research this will be 25% of the total data used as test data *(testing data)*, while the remaining 75% will used as training data *(training data)*. So, from 119.390 data, 29.848 data will be available used as *test data* and 89.542 data will used as *data train.*

Table 4.4 Split Data

Data Testing	25 %	0.25 x 119.390 = 29.848
Training Data	75 %	0.75 x 119.390 = 89.542

D. Feature Selection

Selection feature aim for select a subset of the most relevant and significant features for used in machine learning model training. The main purpose from *selection feature* is for increase model performance with remove features that are not important or *redundant*, which can be cause *overfitting* and extending time computing. With only use important features, the model can become more simple, more fast, and more easy interpreted. With do selection feature the right one, yes confirmed that algorithm more efficient, have more accuracy high, and more generalist in facing new data [13].

E. Feature Selection Artificial Bee Colony Algorithm

Algorithm Artificial Bee Colony modeling three type bee in colony bee: bee worker, bee observers, and bee explorer. Optimization process started with choose bunch solution initial, which represents position source possible food in context algorithm this. Then, bees worker visit solutions thi , fix or improve it If possible. Bee observer Then choose solutions discovered by bees worker based on the quality. Temporary that, bee explorer responsible answer for find solutions new with do exploration random in room search. Through repeated iteration, algorithm Artificial Bee Colony in a way gradually increase quality the resulting solutions. This process continues until criteria stop certain fulfilled, like reach amount iteration maximum or reach satisfactory solution [14].

F. Feature Selection Gradient Boosting Decision Tree Algorithm

Gradient Boosting Decision Tree is combine series classifier base weak become classifier base strong. Different from traditional boosting methods that give weight on the sample positive and negative, *Gradient Boosting Decision Tree* make algorithm global convergence with follow direction negative [15].

G. Model Performance Without Feature Selection

Table 4.5 *Model Performance without Feature Selection*

N_estimators	*Max_depth*	*Acc*	*Pr*	*Rc*	*F1-score*
50	10	80.69%	82%	76%	78%

Table 4.5 shows results evaluation from algorithm *Random Forest* with variation *n_estimators* and *max_depth* parameters. The baseline model using all features showed an accuracy of 80.69%. This serves as a baseline for comparing the improvements after feature selection.

H. Feature Selection with Algorithm Artificial Bee Colony

Table 4.6 Feature Selection with Algorithm *Artificial Bee Colony*

N_estimators	*Num_features*	*Acc*	*Pr*	*Rc*	*F1-score*
50	40	86.17%	86%	84%	85%

Table 4.8 presents results evaluation selection feature use algorithm *Artificial Bee Colony* with variation of parameters *n_estimators* and *num_features*. Using ABC for feature selection increased accuracy to 86.17%, indicating that feature selection improved model performance.

I. Feature Selection with Algorithm Gradient Boosting Decision Tree

Table 4.7 Feature Selection with Algorithm *Gradient Boosting Decision Tree*

N_estimators	*Num_features*	*Acc*	*Pr*	*Rc*	*F1-score*
50	40	86.65%	87%	85%	85%

Table 4.7 shows results evaluation selection feature use algorithm *Gradient Boosting Decision Tree* with variation of parameters *n_estimators* and *num_features*. With GBDT, model accuracy increased to 86.65%, slightly higher than with ABC.

J. Combination of ABC and GBDT

Table 4.8 *Combination of ABC and GBDT*

Num_features ABC	Num_features GBDT	Acc	Pr	Rc	F1-score
30	25	86.81%	87%	85%	86%

Table 4.8 shows results evaluate with The combination of both feature selection methods yielded the highest accuracy at 86.81%, showing that combining ABC and GBDT positively contributed to model performance.

K. Results Interpretation

Appropriate feature selection significantly improved the accuracy of hotel reservation cancellation predictions. Features like 'lead_time', 'adr', and 'total_of_special_requests' had a significant impact on predictions. With a more accurate model, hotel management can make more informed decisions, such as overbooking strategies or offering promotions to reduce cancellation rates.

L. Comparison with Other Studies

Previous studies used methods like SVM and K-NN with lower accuracy [7][8]. The use of Random Forest with ABC and GBDT feature selection in this study showed a significant improvement in accuracy, making it a more effective approach in this context.

V. Conclussion and Recommendations

This study demonstrated that feature selection using the ABC and GBDT algorithms can improve the performance of hotel reservation cancellation prediction models. With an accuracy of 86.81%, the model can be a valuable tool for hotel management to anticipate cancellations. Recommendations for Future Research:

1. **Test on Different Datasets**: Test the model on other datasets to validate its generalizability.
2. **Integrate Additional Data**: Add additional features such as customer reviews or weather data to improve accuracy.
3. **Explore Other Algorithms**: Use other machine learning techniques such as Neural Networks or XGBoost to compare performance.
4. **Hyperparameter Tuning**: Further optimization of model hyperparameters to improve performance.

VI. References

[1] H. Annur, "Penerapan Algoritma Naïve Bayes Berbasis Backward Elimination Untuk Prediksi Pemesanan Kamar Hotel," *J. Ilm. Ilmu Komput. Banthayo Lo Komput.*, vol. 1, no. 1, pp. 1–5, 2022, doi: 10.37195/balok.v1i1.99.

[2] F. H. Qani'ah, R. Ramadhan, and ..., "Prediksi Pembatalan Reservasi Hotel Menggunakan Algoritma Naive Bayes," ... *Informatics* ..., vol. 4, no. 1, pp. 76–80, 2023, [Online]. Available: https://journal.univpancasila.ac.id/index.php/jiac/article/view/5499%0Ahttps://journal.univpancasila.ac.id/index.php/jiac/article/download/5499/2514

[3] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, 2023, doi: 10.52158/jacost.v4i1.491.

[4] I. Gusti Naufhal Daffa Adnyana, R. Mufli Arjuna, A. Nur Indraini, and D. Sandya Pasvita, "Pengaruh Seleksi Fitur pada Algoritma Machine Learning untuk Memprediksi Pembatalan Pesanan Hotel," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, no. April, pp. 551–558, 2021.

[5] F. Sholahuddin, Mohammad, A. Holik, C. Suprapto, I. Mahendra, Iqbal, S. Wibawanto, and M. Kurniawan, "Perbandingan Model Logistic Regression dan K-Nearest Neighbors Dalam Prediksi Pembatalan Hotel," *Semin. Nas. Tek. Elektro, Sist. Informasi, dan Tek. Inform.*, pp. 137–143, 2023.

[6] R. Rosaly and A. Prasetyo, "Pengertian Flowchart Beserta Fungsi dan Simbol-simbol Flowchart yang Paling Umum Digunakan," *Https://Www.Nesabamedia.Com*, vol. 2, p. 2, 2019, [Online]. Available: https://www.nesabamedia.com/pengertian-flowchart/https://www.nesabamedia.com/pengertian-flowchart/

[7] E. Rahmawati, A. B. Nando, C. Agustina, and F. C. Kusumarini, "Perbandingan Teknik Resample pada Algoritma K-NN dan SVM untuk Prediksi Pembatalan Pemesanan Kamar Hotel," *J. Teknol. Inf. dan Terap. (J-TIT*, vol. 10, no. 2, pp. 2580–2291, 2023, [Online]. Available: https://doi.org/10/25047/jtit.v10i2.333

[8] Y. Azhar, G. A. Mahesa, and M. C. Mustaqim, "Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, pp. 15–21, 2021, doi: 10.14710/jtsiskom.2020.13790.

[9] I. S. Manuel and I. Ernawati, "Implementasi GLCM dan Algoritma Naive Bayes Dalam Klasifikasi Jenis Bunga Anggrek," *Senamika*, vol. 1, no. 2, pp. 99–109, 2020, [Online]. Available: https://conference.upnvj.ac.id/index.php/senamika/article/download/638/427

[10] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *J. Krisnadana*, vol. 3, no. 1, pp. 25–33, 2023, doi: 10.58982/krisnadana.v3i1.470.

[11] M. S. T. Putra and Y. Azhar, "Perbandingan Model Logistic Regression dan Artificial Neural Network pada Prediksi Pembatalan Hotel," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 1, pp. 29–37, 2021, doi: 10.14421/jiska.2021.61-04.

[12] P. H. Saputro and H. Nanang, "Exploratory Data Analysis & Booking Cancelation Prediction on Hotel Booking Demands Datasets," *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 40–56, 2021, doi: 10.47738/jads.v2i1.20.

[13] J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *J. Rekayasa Elektr.*, vol. 18, no. 2,

pp. 102–108, 2022, doi: 10.17529/jre.v18i2.24047.

[14] A. Nurdiansyah, M. T. Furqon, and B. Rahayudi, "Prediksi Harga Bitcoin Menggunakan Metode Extreme Learning Machine (ELM) dengan Optimasi Artificial Bee Colony (ABC)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 6, pp. 5531–5539, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5507

[15] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput. J.*, vol. 74, pp. 634–642, 2019, doi: 10.1016/j.asoc.2018.10.036.

VII. AUTHORS PROFILE

Hamida Maulana Lailatal Baroah is a graduate with a Bachelor of Computer Science (S.Kom) from the Department of Informatics Engineering at Yudharta University Pasuruan. Focused on technology, data mining, and information systems. Passionate about learning the latest IT developments and mastering data analysis and data mining algorithms. Committed to continuously improving skills in extracting valuable insights from complex data and contributing to innovative technological solutions.

Lukman Hakim is a researcher with expertise in Image Processing, Computer Vision, and Deep Learning. He completed his Ph.D. in Information Engineering at Hiroshima University, Japan, where he focused on developing advanced deep learning techniques for medical image segmentation. His research interests include deep learning applications in medical imaging, agriculture, chemistry, and materials science. Dr. Hakim has published extensively in international journals and conferences, and he is currently affiliated with Yudharta University, Pasuruan, Indonesia, and can be reached at lukman@yudharta.ac.id.