Prediction of Hotel Booking Cancellation using CRISP-DM

Zharfan Akbar Andriawan
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
zharfan@student.undip.ac.id

Satriawan Rasyid Purnama
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
satriawanrasyid@students.undip.ac.id

Adam Sukma Darmawan
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
adamsukma@students.undip.ac.id

Ricko
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
ricko@students.undip.ac.id

Adi Wibowo
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
bowo.adi@live.undip.ac.id

Aris Sugiharto
Department of Computer Science/Informatics
Diponegoro University
Semarang, Indonesia
aris.sugiharto@live.undip.ac.id

Feri Wijayanto
Institute for Computing and Information Sciences
Radboud University
Nijmegen, The Netherlands
f.wijayanto@cs.ru.nl

***Abstract*—Online travel sales continue to increase every year. Recorded in 2019, digital transactions related to online travel reached USD 755.4 billion. One of the supports of the travel business is the tourism and hospitality industry. The online reservation system is one of the most attractive solutions in the hospitality industry. Cancellation of hotel bookings or reservations through the online system is currently one of the problems in the hotel management system. When the reservation has been canceled, the hotel will be harmed. Therefore, predicting whether a booking will be canceled or not using the help of data science is needed so that the hotel can minimize lost profits. Therefore, by using datasets related to hotel booking requests, a predictive analysis using the CRISP-DM framework is conducted. By first performing some data preparation processes, this study uses a tree-based algorithm to perform the prediction. The experiment produced that Random Forest model has the best value with an accuracy value of 0.8725 and it is found that the time difference between booking is made and arrival time is the most influential feature in predicting the level of cancellation.**

Keywords—Hotel Booking Cancellation, Hotel Booking System, Machine Learning, CRISP-DM

I. INTRODUCTION

Online travel sales continue to increase every year, recorded in 2019, digital transactions related to online travel reached USD 755,4 billion [1]. One of the supports of the travel business is the tourism and hospitality industry [2]. The hotel business operates 365 days a year, 7 days a week and 24 hours a day without exception. People travel all over the world for various reasons. Therefore, room reservations are needed for customers [2].

Management in the company use information technology to perform various tasks that increase the efficiency of employees at work, especially online reservations [3]. The online reservation system is one of the most attractive solutions in the hospitality industry. Their main feature is to fill accommodation capacity in order to increase sales and company profits.

Cancellation of bookings or room reservations through the online system is currently one of the problems in the hotel management system. When a reservation has been canceled, the hotel will be harmed, this is because the room that was previously usable can become unusable on the day the user ordered. Therefore, predicting whether a booking will be canceled or not using the help of data science is urgently needed so that the hotel can minimizes lost profits. Cancellations can reach 20% of the total bookings received by the hotel [4]. This value increases to 60% at airport hotels [5]. In an effort to mitigate losses, resort hotels employ strict cancellation policies as well as overbooking strategies [6].

Although according to Morales and Wang [4], it is hard to imagine that someone can predict whether a booking will be canceled or not with high accuracy. António et al have shown that the possibility of canceling a booking can be predicted with high accuracy. Theoretically, building a model to predict hotel booking cancellations is possible [7].

Therefore, using datasets related to hotel booking requests obtained from articles written by Antonio, 2019 [8]. We carry out predictive analysis using the CRISP-DM framework. CRISP-DM itself is one of the most widely used frameworks in data mining and data science. The contribution of this paper is that we use other tree-based algorithms in solving prediction problems in hotel booking cancellations apart from previous work on the same dataset which gets the best model namely XGBoost [9][10]. In this paper, we find the best accuracy by comparing XGBoost model with another decision-tree based models, namely Random Forest, Catboost and LightGBM also with several different data preprocessing methods. By using Random Forest that have the best accuracy score with value of 0.8725, we also get the most influential features when predicting the booking cancellation, we found that number of days between when a booking is made by the customer and the date when the customer arrives at the hotel is the most important variables.

978-1-7281-9526-1/20/$31.00 ©2020 IEEE

This paper will be described in four parts. The first part provides a brief introduction related to this research. The second part describes the methodology. The third part describes the results. The fourth part provides the conclusions and suggestions for future work.

II. METHODOLOGY

In this case, we use the CRoss-Industry Standard Process for Data Mining (CRISP-DM) framework. The CRISP-DM stage chart is shown in Fig. 6.

There are several steps in CRISP-DM including [11]:

A. Business Understanding

This initial phase focuses on understanding the objectives and requirements in the business side which are then converted into knowledge in determining the definition of the main problem that can be solved by data mining.

The cancellation rate of hotel room bookings has an effect on the company's profitability. It can be seen from Table I, the level of cancellation ratio has increased every year, this certainly causes loss to the company. In 2017, based on our datasets [8], total lost revenue reached 7 Million Euros. Therefore, a good prediction by machine learning model can help the hotel management to take preventive steps in dealing with the problem of canceling hotel reservations. The hotel management can reduce losses caused by canceling bookings by not confirming bookings from customers who are predicted to cancel the booking.

B. Data Understanding

This phase is in the form of a stage where the researcher makes the process of getting acquainted with the data, knowing the quality of the data, getting the initial insight from the data, and getting several hypotheses to find hidden information in the data. At this stage, data visualization is done to understand the data and clean the data from the lost data or delete the problematic features to produce a better and more general machine learning model.

The data we used comes from an article written by Antonio et al, 2019 entitled "Hotel Booking Demand Datasets"[8], so it is appropriate to do classification analysis. This dataset is a combination of two different hotel data. The two hotels are in Portugal where the H1 hotel is a resort hotel located in the Algarve and the H2 hotel is a city hotel located in Lisbon. Both hotels have the same structure, with 31 variables describing 40,060 H1 observations and 79,330 H2 observations. Each observation represents a hotel reservation. Both datasets record orders that arrived between July 1, 2015 and August 31, 2017, including orders that arrived and orders that were canceled. Before the modelling steps, the dataset needs to be clean and also modified.

1) Data Visualization

There are several data visualizations done to help the authors in understanding the case, including:

TABLE I. HOTEL REVENUE YEARLY

Year	Earned Revenue (Euro)	Missing Revenue (Euros)
2015	6,818,116.56	2,306,557
2016	18,870,601.20	7,197,099
2017	17,034,779.77	7,223,579

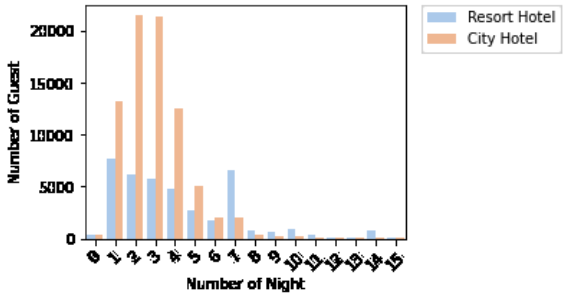

Fig. 1. Visualization of Number of Night and Number of Guest Relations

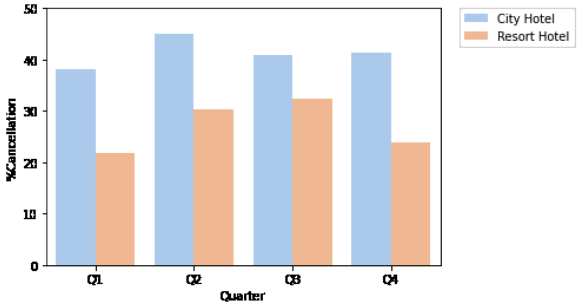

Fig. 2. Visualization of Arrival Time and Cancellations Relations

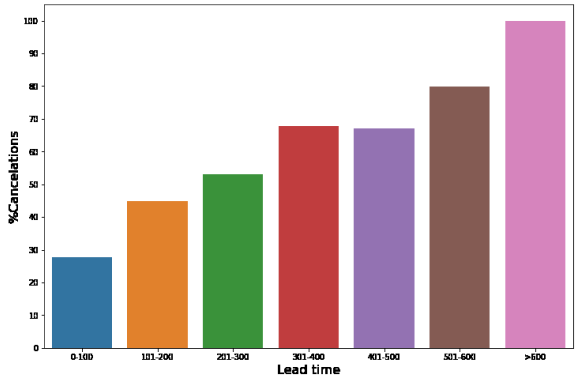

Fig. 3. Visualization of Lead Time and Cancellation Relations

a) Number of Night and Number of Guest Relations

From Fig. 1, it is found that for city hotels, many guests stay between one to four night, whereas for resort hotels, it is quite common to stay more than 7 night. This is quite reasonable, because resort hotels are usually used for vacation purposes that require more time.

b) Relation between Arrival Time and Cancellation

As shown in Fig. 2, for city hotels, the average cancellation rate is around 40% every quarter and there is no seasonality in the ratio of the cancellation rate, while for resort hotels there is seasonality, where the cancellation rate occurs at its peak in the third quarter (July until September) and also low in the first and forth quarter, this means that the customer arrival time throughout the year will be an important feature in machine learning model because there are seasonal trends in certain arrival time.

c) Lead Time and Cancellation Relations

Lead time is the number of days between when a booking is made by the customer and the date when the customer arrives at the hotel. It can be seen in Fig. 3 that orders that have a short lead time are very rare to cancel, while bookings made more than one year before are often get canceled.

2) Missing Data

In the hotel booking dataset there are some features that have missing or undefined data. It can be seen in Table II.

Based on Table II, several steps are taken to overcome the missing data. First, we changed the undefined value in the meal feature to "SC" which have a same value with undefined according to the dataset description [8]. Second, we changed the value of agent and company features that lost or undefined to 999 to represent the missing features. Third, deletion on rows that have missing country, distribution_channel, market_segment and children feature values. The third step is done because the ratio of undefined data on the feature is very small, which is below 0.5% so that it does not delete a lot of data.

3) Data Leakage

Based on business understanding, the following is the process that occurs when a customer makes an order. Based on Fig. 4, it can be seen that there are several features that cause data leakage in the dataset including the assigned_room_type, booking_changes, days_in_waiting_ list, reservation_status, reservation_status_ date and country. The assigned_room_type and days_in_waiting_list features are obtained after the management confirms the booking. These features are leakage because the prediction of the model is done before the booking is confirmed by the hotel management. The booking_changes, reservation_status and reservation_status_date features can still change after the booking is confirmed, so this feature has the potential to experience leakage. According to Antonio [9], the country feature has a leakage caused by charging the country by default to "PRT" or portugal as the country of origin, and that information is only confirmed when the customer checks in.

To prevent data that has not been obtained when predictions are made, the features that leakage are not used in the next step.

TABLE II. MISSING OR UNDEFINED DATA

Feature	Number of Missing Features	Missing Feature Ratio (Percent)
Company	112593	94,307
Agent	16340	13,686
Meal	1169	0,979
Country	488	0,409
Distribution_channel	5	0,004
Children	4	0,003
Market_segment	2	0,002

Fig. 4. Booking Process

C. Data Preparation

In this phase, all activities are carried out to create the final dataset, this dataset will be entered into the model created. Data preparation relates to all activities to construct the dataset so that it can be used in the model. Data preparation in this case will be focus on the selection of features used in the model and we also doing several feature engineering. The final feature used in the model consists of 24 features. There are 6 features not used due to data leakage. Apart from these 6 features, there are 3 time-series features related to arrival time which are also not used and changed into one feature, namely day which is the distance of the day from January 1 to the time of booking desired by the customer at the same year. For categorical features namely meal, market_segment, distribution_channel, deposit, agent, customer_type, agent and company features will be encoded using the one hot encoding technique. Except for reserved_room_type feature, we changed this feature into integer by converting the feature by following alphabetical order. For example, we convert value of 'A' to 1, 'B' to 2, etc. Then, we also combine the stays_in_weekend_nights and stays_in_week_nights. We convert these two features into one feature namely stay_in_night by adding these two features together. Finally, we changed the values of previous_cancellations and previous_bookings_not_canceled to be percentage of each of the two, but if both have 0 value, then the value is still going to be 0. We calculate previous cancellations percentage feature by using this formula:

$$\%Previous\_cancellation = \frac{previous\_cancellation}{previous\_cancelation + previous\_booking\_not\_cancelled} \quad (1)$$

$$\%Previous\_booking\_not\_cancel = \frac{previous\_booking\_not\_canceled}{previous\_cancelation + previous\_booking\_not\_cancelled} \quad (2)$$

D. Modelling

In this phase, the selection and application of variations in machine learning models are carried out, so that the best model is obtained. The model we used is a tree-based algorithm including XGBoost (XGB), Catboost, Light Gradient Boosting Machine (LGBM), and Random Forest (RF). The authors decides to use tree-based algorithms in solving prediction problems in hotel booking cancellations because they prove to work well in previous studies [10], some of the models that will be used include:

1) XGBoost

XGBoost is a sparsity-aware algorithm. XGBoost implements an end-to-end tree boosting system that is very

popular among data scientists to achieve satisfying results. XGBoost is a Gradient Boosting algorithm that has been optimized with the use of parallel processes, the existence of tree pruning, can accept missing values, and can avoid overfitting by using regularization. This tree lined algorithm uses the techniques of cache access patterns, data compression and sharding so as to make this algorithm scalable [12].

2) Catboost

CatBoost is a gradient boosting based algorithm. CatBoost outperforms other algorithms because it is robust in a variety of dataset. There are two things that make catboost superior, because of using permutation algorithms and using special algorithms to process categorical data. Both techniques were created to be robust in avoiding prediction shifts due to leakage that is commonly experienced by other boosting algorithms [13].

3) Light Gradient Boosting Machine

LightGBM is a model that use Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques that can save processing time because it avoids calculating information gain estimates at all possible split points. Therefore, LightGBM is also said to be light because it is fast. The LightGBM algorithm is capable of receiving large data using small memory [14].

4) Random Forest

Random Forest is a machine learning model that combines multiple decision trees into just one model. Random Forest depends on the value of a random vector that has the same distribution in each tree in each decision tree and has a maximum depth. [15]. Random Forest is an ensemble learner, a method that produces several classifiers and combines the results into one model [16].

Random Forest uses Gini impurity to select splits, selecting the node with the least impurity. The value of Gini impurity is in the range 0 to 1, where 0 if a node belongs to the same class. If we want to look Gini impurity for variable X = $\{x_1, x_2, ..., x_j\}$ in node t, where N is the amount of samples, j is the total of children in node t, n_{ci} is the total of samples with value x_i corresponding to class c, and a_i is the amount of samples in value x_i in node t. Accordingly, Gini impurity can be calculated by formulation as follows [16]:

$$I(t_{x_i}) = 1 - \sum_{c=0}^{C} \left(\frac{n_{ci}}{a_i}\right)^2 (3)$$

Gini index is the average of the Gini measure over the different value of X, which can be formulated by formulation as follows.

$$Gini\ (t, X) = \sum_{i=1}^{j} \frac{a_i}{N} I(t_{x_i}) (4)$$

The splitting criterion is based on the smallest Gini impurity value computed among m variables. In Random Forest, every tree utilizes a distinct set of *m* variables to produce splitting rules. One of the outcomes of the Random Forest is variable importances. Variable importances indicate the level of correlation between a variable and its classification results.

To compute variable importances for the variable *j*, out-of-bag (OOB) samples are computed in the tree and the accuracy of prediction is obtained. Then the values for variable *j* are permuted again in OOB sample, the accuracy is also measured again. This calculation is performed on all trees when the tree is built. The average reduction in accuracy when permutation happen is then averaged over all trees and is used to calculate the importance of the variable *j*. If the accuracy value drops significantly, then this indicates that the variable has a strong correlation with the classification results [17].

After the importance value of all the variables calculated, random forrest then return a ranked list of the variable importances. Formally, let β_t is the OOB samples at tree t, $t \in \{1, ,, ntree\}$, γ_i^t is the predicted class before the permutation for instance i and $\gamma_{i,\alpha}^t$ is the predicted after the permutation class for instance i. So that, the variable importances *VI* for variable j in tree t is given by:

$$VI_j^t = \frac{\sum_{i=1}^{N} \beta_t I(\gamma_i = \gamma_t')}{|\beta_t|} - \frac{\sum_{i=1}^{N} \beta_t I(\gamma_i = \gamma_{i,\alpha}')}{|\beta_t|} (5)$$

The raw variable importances for variable j then averaged accross all trees in the Random Forest, as shown as follows.

$$VI_j = \frac{\sum_{t=1}^{ntree} VI_m^t}{ntree} (6)$$

E. Evaluation

At this stage, the model is evaluated and we also analyzed several steps that previously taken to ensure that the model chosen has the best quality that can achieve the objectives of the existing business problem. The evaluation is based on accuracy, recall, precission, and F1 values generated by each model.

The authors make a model to predict the target variable namely IsCanceled which is a binary value (0 means no cancellation, and 1 if it is canceled), several algorithms used by the authors including XGBoost, LightGBM, CatBoost and Random Forest. The authors use K-fold Cross Validation, a model evaluation technique that is quite popular and widely used. The authors choose 10 for the K value. This means, for each iteration, the K-fold algorithm evaluates the model using 10% test data and 90% training data. The K-fold mechanism can be seen in Fig. 5. After we get the best model, then we fine-tuned the model hyperparameter to get the best score.

F. Deployment

At the deployment stage, the results of data science or data mining that have been carried out are displayed in the form that the user needs. The deployment phase will not be discussed in detail in this study. However, we suggest that the machine learning model that has been created can be integrated into the hotel room reservation system with aims to reduce the loss of income at the company.

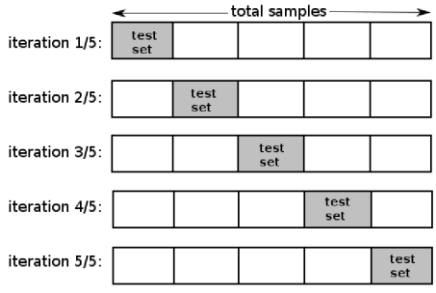

Fig. 5. Kfold Cross Validation Process

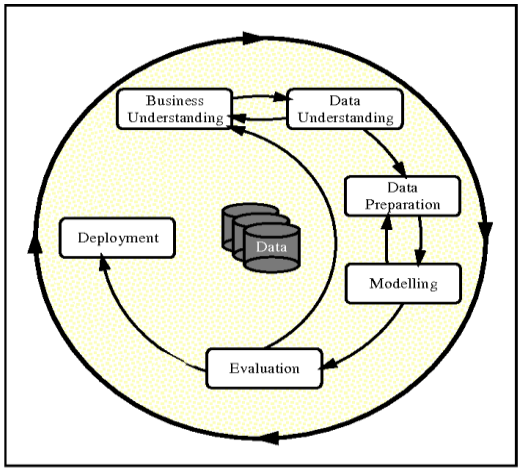

Fig. 6. Chart of Stages in CRISP-DM

III. Result And Discussion

In this case, modelling is done on tree-based algorithms, namely XGBoost (XGB), Catboost, Light Gradient Boosting Machine (LGBM), Random Forest (RF). We use default hyperparameter for all models. Modelling results can be seen in Table III.

With accuracy score of 0.8717, Random Forest outperformed other algorithm on predicting hotel booking cancellation. Then, we fine-tuned the hyperparameter for Random Forest and got the best value for the hyperparameter at 250 for n_estimators, 50 for max_depth, and also with 0.5 for the max_feature value which determines 50% percent of the features the Random Forest that will be split. The after tuned score with accuracy value of 0.8725 is our final score that is slightly increased from before we conducted tuning. The final after tuned score can be seen at Table IV.

The results shown in Table IV is for all datasets, to compare our model with the state-of-the-art model, we also conducted an experiment by training and testing data separately between H1 and H2 as was done in the previous paper [10]. Table V shows comparison between our experiment and the state-of-the-art model. It shows that our proposed models have slightly better accuracy score than previous study.

With Random Forest, the most influential features in predicting hotel booking cancellations can be extracted by using variable importances. This is shown in Fig. 7.

TABLE III. Modelling Result

No.	Modelling Result				
	Model	**Accuracy**	**Recall**	**Precision**	**F1**
1.	**XGB**	0.8227	0.6141	0.8688	0.7195
2.	**Catboost**	0.8484	0.7105	0.8557	0.7763
3.	**LGBM**	0.8380	0.6739	0.8580	0.755
4.	**RF**	0.8717	0.7750	0.8646	0.8173

TABLE IV. Tuning Modelling Result

Modelling Result				
Model	**Accuracy**	**Recall**	**Precision**	**F1**
RF	0.8717	0.7750	0.8646	0.8173
After Tuning RF	0.8725	0.7826	0.8606	0.8197

TABLE V. Model Comparison With Previous Study

Model	Hotel	Accuracy	Precision	F1
Previous Study Model	H1	0.8646	0.8484	0.7410
	H2	0.8701	0.8849	0.8460
Propose RF	H1	0.8724	0.8117	0.7539
	H2	0.8712	0.8735	0.8396

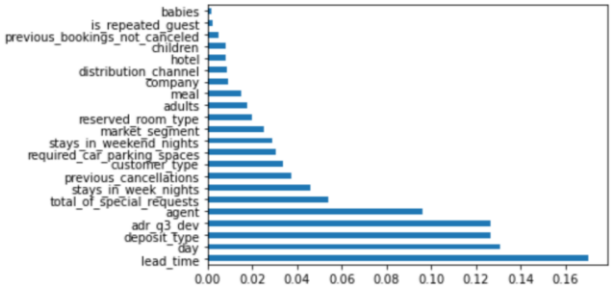

Fig. 7. The Most Influential Feature

Fig. 7 shows that the most influential features on the model are the lead_time feature, which is the number of days between when a booking is made by the customer and the date when the customer arrives at the hotel and followed by a day feature that represents what day of the year the customer will arrive at the hotel.

IV. Conclusion

Predicting cancellation of hotel room reservations is important in order to minimize the loss of income to the company. The use of predictive analysis in cases of cancellations of hotel rooms can be done using the CRISP-DM framework. By using the stages in CRISP-DM, the most appropriate features for the machine learning model are built, the best machine learning model is Random Forest, with an accuracy value of 0.8725 and the time difference between bookings made and time arrival is the most influential feature to predict the cancellation rate of a hotel booking. Further work of this paper are to do dataset preprocessing in other ways or techniques, make deployment strategies, and also perform other modelling or hyperparameter tuning technique to get better accuracy.

Acknowledgment

The authors would like to thank the Department of Informatics and Diponegoro University for supporting this research.

REFERENCES

[1] Statistica, "Digital travel sales worldwide from 2014 to 2020," 2020. https://www.statista.com/statistics/499694/forecast-of-online-travel-sales-worldwide/.

[2] C. C. Wong and P. L. Hiew, "Correlations between factors affecting the diffusion of mobile entertainment in Malaysia," *ACM Int. Conf. Proceeding Ser.*, vol. 113, pp. 615–621, 2005, doi: 10.1145/1089551.1089661.

[3] ILO (International Labour Organization), *Developments and challenges in the hospitality and tourism sector*, no. November. Geneva, 2010.

[4] D. Romero Morales and J. Wang, "Forecasting cancellation rates for services booking revenue management using data mining," *Eur. J. Oper. Res.*, vol. 202, no. 2, pp. 554–562, 2010, doi: 10.1016/j.ejor.2009.06.006.

[5] P. H. Liu, "2004," *Revenue Manag. pricing Case Stud. Appl.*, no. ISBN 1-84480-062-8, pp. 91–108, 2004.

[6] S. J. Smith, H. G. Parsa, M. Bujisic, and J. P. van der Rest, "Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry," *J. Travel Tour. Mark.*, vol. 32, no. 7, pp. 886–906, 2015, doi: 10.1080/10548408.2015.1063864.

[7] N. Antonio, A. de Almeida, and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," *Tour. Manag. Stud.*, vol. 13, no. 2, pp. 25–39, 2017, doi: 10.18089/tms.2017.13203.

[8] N. Antonio, A. de Almeida, and L. Nunes, "Hotel booking demand datasets," *Data Br.*, vol. 22, pp. 41–49, 2019, doi: 10.1016/j.dib.2018.11.126.

[9] N. P. P.-I. U. L. S. of T. and A. Antonio, "Hotel Revenue Management: Using Data Science to Predict Booking Cancellations," 2019.

[10] N. Antonio, A. De Almeida, and L. Nunes, "An automated machine learning based decision support system to predict hotel booking cancellations," *Data Sci. J.*, vol. 18, no. 1, 2019, doi: 10.5334/dsj-2019-032.

[11] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000, doi: 10.1.1.198.5133.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," vol. 42, no. 8, p. 665, 2016.

[13] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Section 4, pp. 6638–6648, 2018.

[14] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[15] Y. Aditya, "Random Forest," 2018. http://machinelearning.mipa.ugm.ac.id/2018/07/28/random-forest.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.