INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Customer Loyalty Prediction for Hotel Industry Using Machine Learning Approach

Iskandar Zul Putera Hamdan [a], Muhaini Othman [a,*], Yana Mazwin Mohmad Hassim [a], Suziyanti Marjudi [a], Munirah Mohd Yusof [a]

[a] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Malaysia*
Corresponding author: [*]muhaini@uthm.edu.my

***Abstract*— Today, machine learning is utilized in several industries, including tourism, hospitality, and the hotel industry. This project uses machine learning approaches such as classification to predict hotel customers' loyalty and develop viable strategies for managing and structuring customer relationships. The research is conducted using the CRISP-DM technique, and the three chosen classification algorithms are random forest, logistic regression, and decision tree. This study investigated key characteristics of merchants' customers' behavior, interest, and preference using a real-world case study with a hotel booking dataset from the C3 Rewards and C3 Merchant systems. Following a comprehensive investigation of prospective preferences in the pre-processing phase, the best machine learning algorithms are identified and assessed for forecasting customer loyalty in the hotel business. The study's outcome was recorded and examined further before hotel operators utilized it as a reference. The chosen algorithms are developed utilizing Python programming language, and the analysis result is evaluated using the Confusion Matrix, specifically in terms of precision, recall, and F1-score. At the end of the experiment, the accuracy values generated by the logistic regression, decision tree, and random forest algorithms were 57.83%, 71.44%, and 69.91%, respectively. To overcome the limits of this study method, additional datasets or upgraded algorithms might be utilized better to understand each algorithm's benefits and limitations and achieve further advancement.**

***Keywords*—Machine learning; classification; CRISP-DM.**

Manuscript received 6 Nov. 2022; revised 28 May 2023; accepted 6 Jun. 2023. Date of publication 10 Sep. 2023.

I. Introduction

Current applications of machine learning include agricultural information management, consumer loyalty programs, customer profile management, and others. The human brain's ingenuity resulted in the development of several devices. These technologies facilitated human existence by allowing individuals to satisfy various demands, such as travel, industry, and computing, one of which is machine learning [1]. Machine learning is the branch of computer science that studies artificial intelligence structures and adapts computational learning theories [2]. Machine learning is the adjustment of systems capable of performing artificial intelligence (AI)-related tasks, which include recognition, evaluation, organization, robot navigation, and forecasting [3]. It has garnered great interest due to its ability to predict many difficult occurrences [4] precisely. The demand for machine learning is expanding in several industries, including the tourist and hospitality business or, more precisely, the hotel industry. In addition to classification, clustering, and regression, the hotel and tourist sectors have applied machine learning for financial management, customer experience development, and organizational data analysis [5]. As revealed by Parvez [6], the goal of machine learning in the hospitality sector is to establish preparations for collecting data and extracting knowledge from it while also striving to continue boosting self-capability via observation without human intervention or basic reconfiguring. It can be implemented in a staged process where specialists collect, select, organize, pre-process, and incorporate datasets into the machine before establishing a statistical model.

Findbulous Technology Sdn. Bhd. has launched the C3 Rewards and C3 Merchant applications for merchants' consumers and merchants, respectively. Merchants are enterprises such as New York Hotel and Gloria Hotels and Resorts Johor Bahru that employ the C3 Merchant application solution provided by Findbulous. The application provides merchants with customer relationship management, a client coupon platform, a loyalty and achievements campaign, an

online booking system, and a gateway administrator. When clients of merchant desire to reserve a room for the night, they can use the application to create a profile and proceed with the reservation procedure without engaging in person with hotel employees. Upon registration, the program requires the client's personal information, which will be saved through Amazon Web Service Data Lake (AWS). The applications were developed using Python, and the developers of Findbulous continually updated them. To enhance the C3 Merchant application, functionalities for predicting consumer loyalty in the hotel business must be introduced. The existing program does not offer merchants data mining techniques to estimate the loyalty status of their customers.

This research aims to investigate merchants' clients' behavior, interests, and preferences and assess them using stated machine learning techniques, such as random forest, logistic regression, and decision trees. This research was completed by discovering and analyzing the optimal algorithm for machine learning in forecasting client loyalty in the hotel business. Using the Python programming language, the chosen algorithms were implemented. Before hotel companies utilized the findings of the study analysis as a reference, they were reviewed thoroughly.

II. Materials and Method

A. Machine Learning in Hospitality Industry

The advancement of artificial intelligence and robots and increased digital connectivity influence all business sectors, including services [7]. Artificial intelligence enables workers to work smarter, which leads to greater business outcomes, but it also necessitates the development of new competencies and capabilities, ranging from technological knowledge to social and emotional abilities, as well as creative ability [8]. Machine learning also can be classified as a field of artificial intelligence that analyzes massive volumes of data to continually refine models and generate plausible predictions using algorithms [9]. Utilizing big data in hotel companies will assist them in making the best tactical and strategic decisions, increasing corporate value. There are three major strategies in machine learning where those strategies are semi-supervised machine learning, unsupervised machine learning, and supervised machine learning. This study employed supervised machine learning and classification approaches to develop prediction models from the dataset. Labeled datasets to train algorithms that consistently categorize data or anticipate outcomes are characterized as supervised machine learning. Various algorithms will build a function to turn the inputs into the required outputs [10].

We can use machine learning to address our challenge in several scenarios or conditions. According to Brynjolfsson and Mitchell [11], eight factors may be used to determine if a job is acceptable for using a machine learning approach. The task begins with employing a function that translates well-defined inputs to well-defined outputs. Second, the job may be designed with huge datasets or input-output pairs. Third, the job may give unambiguous feedback with well-defined goals and metrics. Fourth, the challenge does not call for lengthy sequences of logic or reasoning that rely on a broad variety of background knowledge or human decency. Fifth, extensive explanations of how the judgment was made are not required. Sixth, the job allows for mistakes and does not need responses that are probably correct or ideal. Seventh, the component or characteristic under consideration should not vary significantly over time. Eighth, no specific dexterity, physical aptitude, or mobility is required. As big data increases and grows, so will the market demand for data analysts and scientists who contribute to identifying the most critical business challenges and, ultimately, solutions to those challenges.

A. Decision Tree

The decision tree is a complicated and widely utilized machine learning technique for predicting and classifying huge amounts of data and can be utilized in various fields such as machine learning, image processing, and identification of patterns [12]. It is one of the various analytic methodologies. A decision tree is a tree-based approach in which data splitting decides every path from the root toward the leaf node until a Boolean result is attained at the leaf node [13]. It is a hierarchical interpretation of knowledge relationships with connections and nodes. When relations are used to classify, nodes identify the intent. Machine learning classification methods can handle large volumes of data. It may be used to make predictions regarding the category of the class names, classify data due to class labels and training sets, and classify newly accessible data. The decision tree technique has the advantage of categorizing categorical and numerical outcomes; however, the feature created must be categorical. Next, the decision tree approach is basic and easy to understand since the process workflow is similar to how the human brain operates and analyzes. Based on Simon [2], decision trees, unlike algorithms such as nearest neighbor (NN), support vector machine (SVM), and others which can be described as black box algorithms, help us understand the logic underlying data analysis.

B. Random Forest

According to Breiman [14], they presented the random forest methodology, a form of ensemble strategy intended to forecast the mean of multiple alternative regular models in regression and classification approaches for the random forest architecture. The dataset is randomly divided into two pieces in the Random Forest algorithm: the training dataset (the in-Bag) for learning and the testing dataset (the out-of-Bag) for assessing the learning level [15]. The ensemble approach is a strategy that utilizes various learning algorithms to enhance expected classification and regression results. During the training phase, the ensemble classifiers approach generates many decision trees and outputs class labels that receive the most votes [16]. Bagging is a form of ensemble method mostly used in the random forest (Bootstrap Aggregation), and also, the techniques could be employed to reduce the variance of a decision tree. The algorithm then trains trees on every one of the 1,..., B sub-samples and combines the outputs of each tree into one overall prediction. Let B be the total number of trees grown, {φb, b = 1, ..., B} denote individual trees, and bags denote the collection of all these trees. One advantage of the random forest approach is that it can accommodate incomplete data elements while maintaining the data's reliability. Moreover, Random Forest is capable of effectively processing data with a large number of features

and classes. Additionally, attribute values are unaffected by scaling (or, more broadly, any monotonous change) [17].

C. Logistic Regression

Logistic regression is a classification approach used to assess the connection between many predictor factors, either categorical or continuous, and a binary (dichotomous) result [18], [19]. Due to the dichotomous nature of the objective or dependent variable, there are typically just two classes; that is, the dependent variable is frequently binary, in which data is represented as 1 or 0 for yes or no. It is an extension of conventional regression in that it could only model a dualistic variable that essentially offers the chance of an event occurring or not occurring, with the outcome ranging from 0 to 1 [20]. Theoretically, logistic regression can anticipate p (y = 1) as a factor of x. It is among the greatest fundamental machine learning approaches and can be applied to many classification problems, such as clinical diagnosis and spam email detection. Using logistic regression to categorize the dataset has various advantages. In addition to revealing the suitability of a coefficient size predictor, it may also provide the direction of the association, whether positive or negative. Next, by utilizing multinomial regression, this classifier is easily expandable to numerous classes and gives a natural probabilistic view of class predictions. Lastly, logistic regression is one of the simplest fundamental machine learning techniques, and although it is simple to use, it may sometimes offer exceptional training efficiency. Due to these factors, developing a model using this method does not need a substantial amount of computer available resources.

D. Comparison Existing Research

Several previous studies are chosen and discussed to get more information that may be used to carry out the planned study. Table 1 summarizes the associated works.

TABLE I
THE ANALYSIS OF RELATED WORKS

No	Article	Algorithm	Total of instances	Accuracy (%)
1	"The prediction of Hotel Customer Loyalty using Machine Learning Technique" [3]	Decision Tree	119,386 samples	98.90
2	"Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification" [21]	Support Vector Machine	386 samples	76.42
		Naïve Bayes	386 samples	72.54
3	"Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study" [22]	Multilayer Perceptron	127 million samples	83.00
		Decision Tree	127 million samples	87.00
		Random Forest	127 million samples	87.00
		Gradient Boosted Tree	127 million samples	87.00
4	"Implementation of Data Mining Using C4.5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Persero) Pati Area Office" [23]	Decision Tree (Experiment 1)	166 samples	89.95
		Decision Tree (Experiment 2)	166 samples	94.07
5	"Using Decision Tree to Predict Response Rates of Consumer Satisfaction, Attitude and Loyalty Surveys" [24]	Decision Tree	244 samples	78.26
		Logistic Regression	244 samples	73.10

B. Methodology

The Cross-Industry Standard Process for Data Mining, also known as CRISP-DM [25], is one of the methods for managing data mining operations and development. The CRISP-DM is a general data mining procedure framework that overviews data mining project life cycles [26]. Following research undertaken by Martinez-Plumed et al. [27], the CRISP-DM approach remains the norm for building data mining and information discovery systems relying on several users' questionnaires and surveys. The methodology's iterative execution also involves communication among business specialists and data mining experts [28]. Using the CRISP-DM methodology's six processes or phases, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, this study elucidated the implementation of several classification models using the multiple methods or algorithms enclosed in project scopes. The life cycle of the CRISP-DM reference model is depicted in Fig. 1.

Before experimenting, the research life cycle begins with the business understanding stage. This phase's purpose is to identify objectives from a business perspective and convert them into machine learning objectives, gather and validate data quality, and determine the project's feasibility. Checking the project's viability before initiating it is regarded as the best strategy for the overall success of the machine learning technique [29] and can lessen the danger of premature failures caused by unreasonable expectations. After identifying the project's objectives and scope, the data is acquired via a data collecting method and undergoes data quality validation, which involves three tasks: data description, data needs, and data verification, during the data understanding stage. Subsequently, the dataset is chosen during the data preparation stage, and missing attributes or elements are replaced.

Additionally, the dataset underwent a cleansing phase in which the tasks were to repair, impute, or eliminate inaccurate values. If necessary, irrelevant attributes can be eliminated to reduce space and processing time. After that, the dataset was pre-processed using normalization.

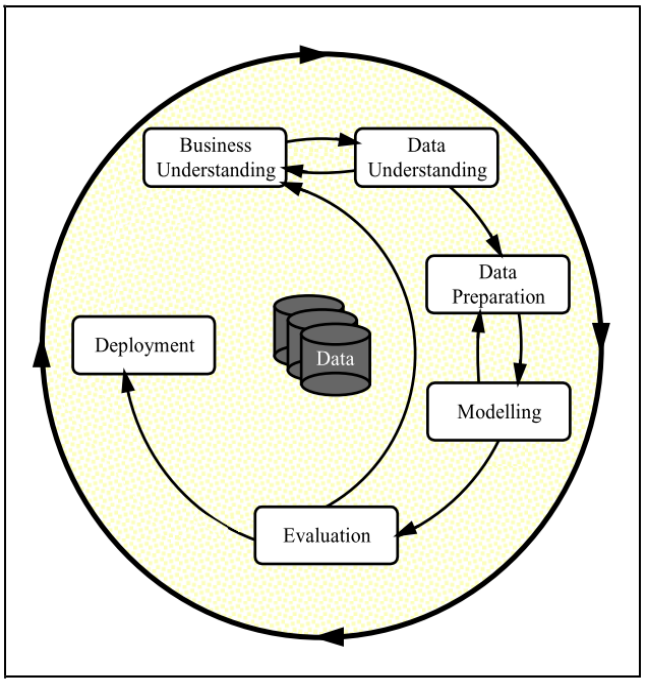

Fig. 1 Process of CRISP-DM [25]

Data normalization is a method for transforming the dataset where the hotel's room reservation dataset is converted to sequential values between 0 and 1. It is crucial to remember that identical normalization values must be implemented in both training and testing sets [30]. The Min-Max normalization formula is shown in Eq. 1 below.

$$MnMx = \frac{(v-Mn\,x)}{Mx\,x-Mn\,x}(nwMx - nwMn) + nwMn \quad (1)$$

Based on Eq. 1, Mn represents the minimum value of the features from the dataset, Mx represents the maximum value of the features from the dataset, V is the selected value of the row on every feature of the dataset, nwMx is utilized to change the maximum value to 1, and nwMn is utilized to change the minimum value to 0. Throughout the modeling phase, the dataset must be prepared for training using the project's specified methods, such as random forest, decision tree, and logistic regression. The objective of the modeling phase is to develop one or more models that best meet the stated criteria. Whether the dataset must be divided into training, test, and validation sets during test design construction depends on the modeling technique. The dataset is divided into testing and training datasets for this study. The model's algorithms are written in Python and implemented in any integrated development environment (IDE) that supports Python, such as Jupyter and RapidMiner. To keep track of the machine learning algorithm and research procedure, documentation is performed. In the assessment step, the outcome is examined and contrasted after the dataset has been trained with chosen algorithms. The performance of classification algorithms may be quantified using precision, accuracy, recall, and f-measure to create a confusion matrix.

- Accuracy. The fraction of total forecasts that were accurate.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

- Precision. Positive samples are determined by dividing the number of samples accurately identified as positive by the total amount of samples.

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

- Recall. The number of positive samples divided by the total amount of positive samples in the testing set.

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

- F-measure. The weighted mean of Precision and Recall.

$$F - Measure = \frac{2*(Recall*Precision)}{(Recall+Precision)} \quad (5)$$

In the last stage, also known as the deployment stage, the selected algorithm from the results of the comparison is trained and applied with the dataset. Any modifications applied to the algorithm's parameter setting or the dataset are recorded.

III. Results and Discussion

This research employs random forest, decision tree, and logistic regression as its chosen classification methods. Each approach includes its own tool engine and preset configuration values, allowing it to produce precise results to any dataset. This study's primary objective is to identify the algorithm that provides the highest accuracy score based on the suggested methods and given dataset. As a solution for implementation, classification algorithms are used to the dataset with the selected parameter settings in this study. Designing the phases of the experiment is a crucial aspect of ensuring the experiment's success. In the case that the experiment yields an unsatisfactory outcome, following the experiment's workflow might reveal which portion of the workflow may need to be modified or redone from the beginning of the operation. The phases of the research project that should be carried and finished for the study to be effective are depicted in Fig. 2.

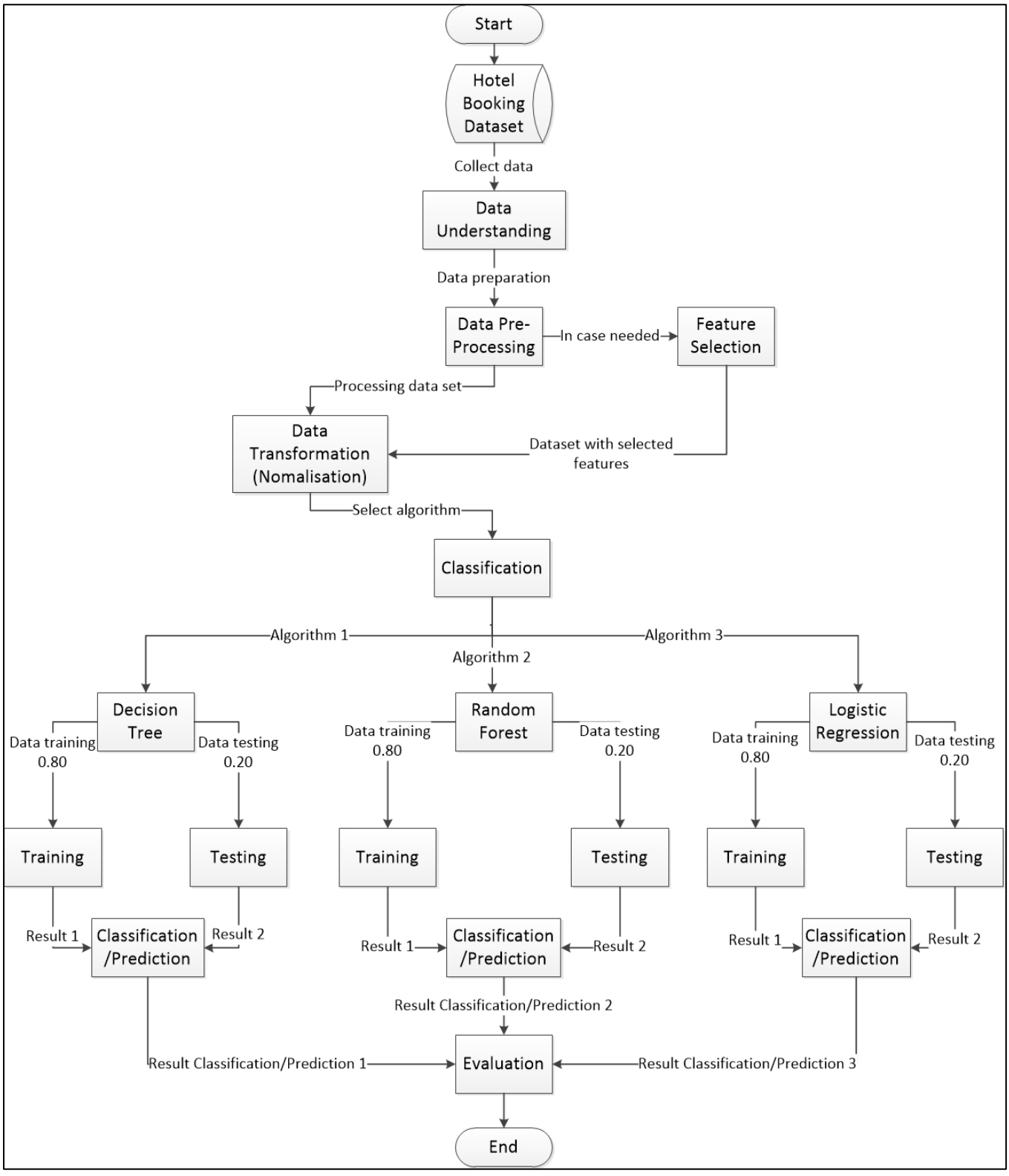

Fig. 2 Flowchart of the experiment phases

Fig. 2 depicts the flowchart of the research phases that served as a reference for this study. Upon identifying the datasets that to be examined, the testing phase commences. As for this study, hotel booking data and related attributes are chosen. After safeguarding the data, it is analyzed to determine its attributes and relationship to the business process. Next, the dataset undergoes data preparation to make it more readable by converting it from JavaScript Object Notation (JSON) format to Comma-separated values (CSV) format. The dataset is next subjected to data pre-processing to verify that it includes no null values or redundant data. During the pre-processing stage, the feature selection procedure eliminates irrelevant attributes in the dataset. This step can be omitted if the dataset has a small number of relevant attributes.

The next phase is normalization, in which each value in the dataset must be tuned so that each classification operation may be clearly comprehended. For further analysis, the normalized data are fed into the appropriate algorithms. As stated earlier, the algorithms decision tree, random forest, and logistic regression are chosen to examine the data. During this stage, the dataset was divided into training datasets of 80% and testing datasets of 20%. Each classification technique will utilize This training-to-testing ratio in the same environment. After the operation was finished, each result was recorded and assessed to see which algorithm gave the most accurate score to be chosen as the best.

In this study, the execution of this experiment needs the usage of a test bed or platform that will act as the experiment's setting. There are several available platforms, including Microsoft Azure, Matlab, RapidMiner, Jupyter, and Rstudio. In this study, Jupyter was utilized as a platform, and the Python programming language was used to construct each classification method suggested. According to various research articles on data science that conducted experiments, Python and R are well-known programming languages for statistical analysis and data exploration. Python is an excellent programming language for machine learning and computer science since it provides several data-oriented function modules that expedite and improve data manipulation and processing, thus saving time. Therefore, the Python programming languages and Jupyter IDE was utilized in this experiment.

As noted earlier in the test bed, Jupyter and the Python programming language was employed as the bare minimum need prior to commencing the study and experiment. However, the classification technique is not currently accessible in the Jupyter and Python programming languages. To overcome this issue, it is necessary to import appropriate library packages, such as the Pandas library, which provides a number of data modification operations, including combining, choosing, resizing, data filtering, and data wrangling. Next, NumPy module can be utilized for array manipulation, linear algebra, the Fourier transform, and matrix manipulation. In addition, the Scikit-learn or sklearn library module was utilized, since it includes effective analytical and machine learning modeling abilities, such as classification, random forest, decision tree, and logistic regression. After importing all required library modules, the research may be performed without difficulty.

A. Parameter and Testing Methods

As stated previously, one of the primary goals of this research study is to evaluate the accuracy of specified classification algorithms based on a dataset of hotel reservations. The dataset offers 4881 hotel reservation records with 22 features for this research. A variable or parameter is a component that is used to generate a forecast in classification models like logistic regression, random forest, and decision trees. In addition, testing techniques relate to the evidence that each algorithm's true or false result is supported. The strategy employs a confusion matrix, which demonstrates experiment error, the correctness of the algorithms and others.

TABLE II
LIST OF ATTRIBUTES IN HOTEL BOOKING DATASET

No	Attributes	Explanation
1	booking_id	The unique number for hotel booking
2	customer_id	The unique number for customer
3	customer_voucher_id	The unique number for used voucher ID
4	voucher_id	Category of voucher used
5	date_from	Date start for the customer book the hotel
6	date_to	Date end for the customer book the hotel
7	currency	Type of currency that customer used for transaction
8	room_price	Price for the room that user want to book
9	extra_price	Any extra charge for customer (If any)
10	total	Total price for the booking transaction
11	deposit	Deposit charge that customer need to pay (If any)
12	tax	Total tax charged to the customer transaction (If any)
13	discount	Total discount customer get for transaction (If any)
14	status	The status of booking, whether it is confirmed or cancel
15	room_names	Types of rooms that customer want to book
16	customer_gender	Customer gender, whether it is male or female
17	customer_race	Customer race for customer background purpose
18	c_arrival	Time recorded when customer arrived at the hotel
19	c_title	Preferred customer title
20	book_created	The date of the customer book the room
21	booking_reason	The reason of the booking, whether it is a vacation trip or business trip
22	loyalty_status	Customer loyal level whether they are not loyal or loyal customer (**Class label**)

The properties or attributes supplied in the dataset are displayed in Table 2. The dataset comprises categorical and numerical data. In order to apply the classification to the dataset, certain categorical variables are translated from String to Integer format. As a class label, the discrete property named "loyalty status" forecasts client loyalty based on other factors. Before the dataset can be utilized for classification or forecasting, it must undergo data understanding, data preparation, data pre-processing, data cleaning, data transformation, and data normalization. In the data understanding stage, the dataset is subjected to an Exploratory Data Analysis (EDA) to comprehend the dataset's structure, elements, and others.

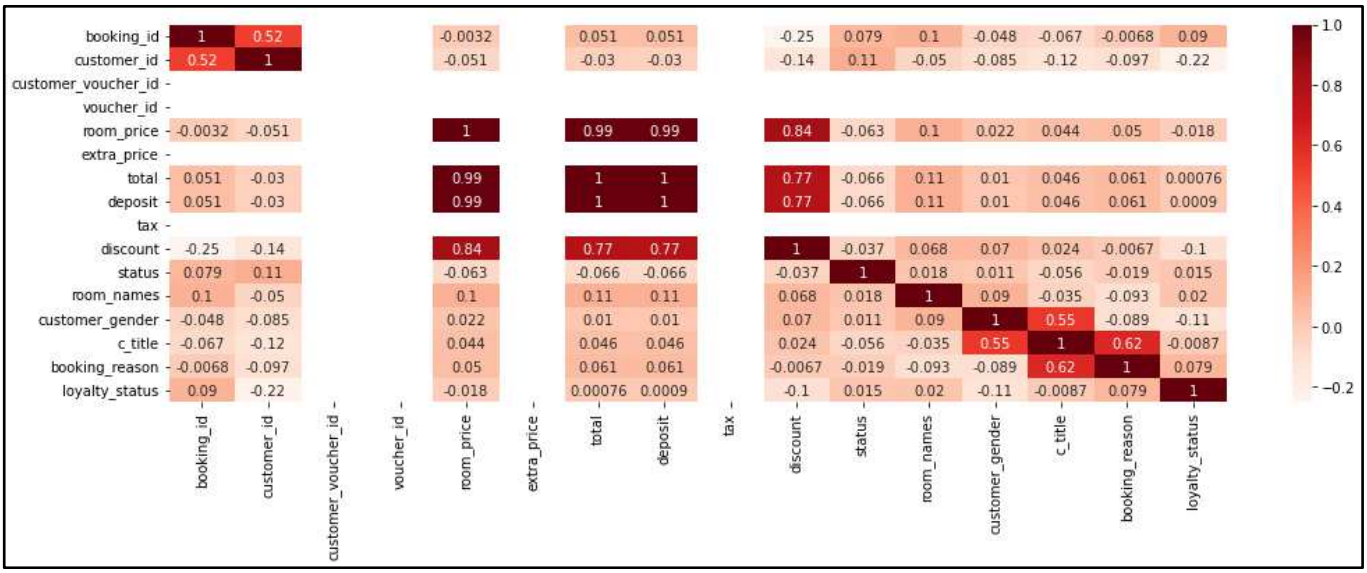

Fig. 3 Heat map for Pearson Correlation Coefficient

In the data preparation stage, the dataset will initially be transformed into a readable dataset format, such as a Comma Separated Values (CSV), to facilitate comprehension and accessibility. For this study, the dataset will be transformed from JSON to CSV to be utilized in Jupyter easily. Every categorical data in the dataset will be converted from String to Numeric or Integer format during the second phase of data pre-processing. In the dataset, the following features were changed from String to Integer: "status", "room names", "customer gender", "c title", and "booking reason".

The heat map depicting the Pearson Correlation Coefficient value is seen in Fig. 3. The Pearson Correlation Coefficient is applied to determine the connection between two attributes within the same dataset. Pearson Correlation Coefficient value will be generated from -1.0 to 1.0. Correlation ranges from -1 to +1. Values closer to zero mean no linear trend between the two variables. The closer to 1, the more the correlation. For example, "room_price", "total" and "deposit" attributes display a high positive correlate value with each other, while "customer_gender" and "loyalty_status" attributes display a low negative correlate value with each other. A score of 0 for an attribute implies that the attribute has no relationship with other attributes. During the data pre-processing step, any attribute or data containing too many null values or biased data that is unrelated to other attributes will be deleted from the dataset.

The "customer id" attribute has been eliminated to prevent data bias, and "customer race" must also be dropped because it includes 70% more null data. Based on the Pearson Correlation Coefficient result, the "customer voucher id", "voucher id", "date from", "date to", "currency", "extra price", "tax", "c arrival", and "book created" features were eliminated since they have no correlation with other data. After the data cleansing procedure, the new total number of features, which includes the class label, is 11. The mutual Information function may be applied to a dataset to determine whether features substantially influence the class label by calculating the statistical dependency between two variables.

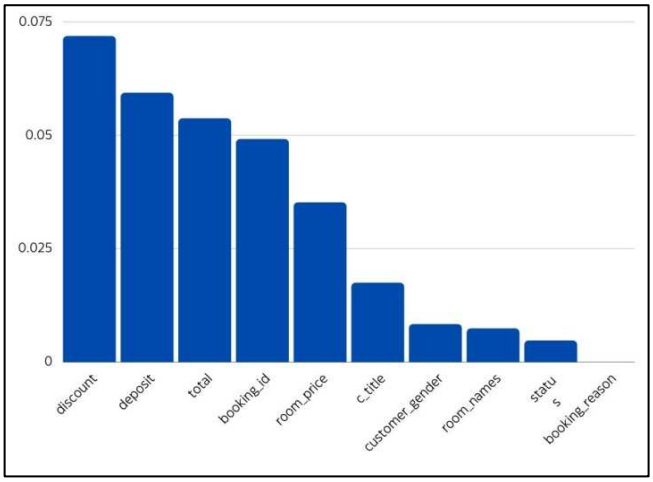

Fig. 4 Mutual Information value for hotel booking dataset.

The Mutual Information value for every feature is depicted in Fig. 4. Based on the analysis, "discount" attribute generates 0.071805 value while "deposit", "total", "booking_id", "room_price", "c_title", "customer_gender", "room_names", "status" and "booking_reason" generates 0.059280, 0.053662, 0.049098, 0.035127, 0.017426, 0.008329, 0.007357, 0.004710 and 0.0000 value respectively. It indicates that the "discount" feature was the most predictive of client loyalty, as a client who gained a discount during their reservation process is likely to become a loyal customer.

B. Result Analysis and Evaluation

Processes engaged in the chosen algorithm may have a comparable approach for maintaining the data collection but different ways of executing experiments. Before doing classification, the dataset will be standardized using min-max normalization. Min-max normalization is one of the most used data normalization techniques. For each feature, the minimum value is changed to zero, the maximum value to one, and the remaining values to a decimal between zero and one. Once the dataset has been normalized, methods such as decision trees, logistic regression, and random forest may be used to classify it. All ten features and one class label are

utilized in the classification phase. Following this, the normalized dataset is separated into two halves for training and testing, with 80% of the dataset designated for training and the remaining 20% designated for testing.

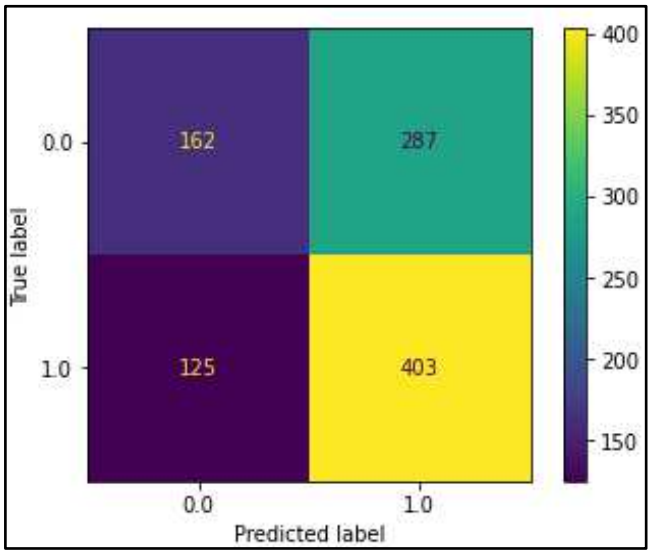

Fig. 5 Confusion matrix for logistic regression

In Confusion Matrix, there are True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP) values that have been generated. The outcome of the Confusion Matrix for the logistic regression technique is depicted in Fig. 5. TN value produces 162, FP value produces 287, FN value produces 125, and TP value produces 403.

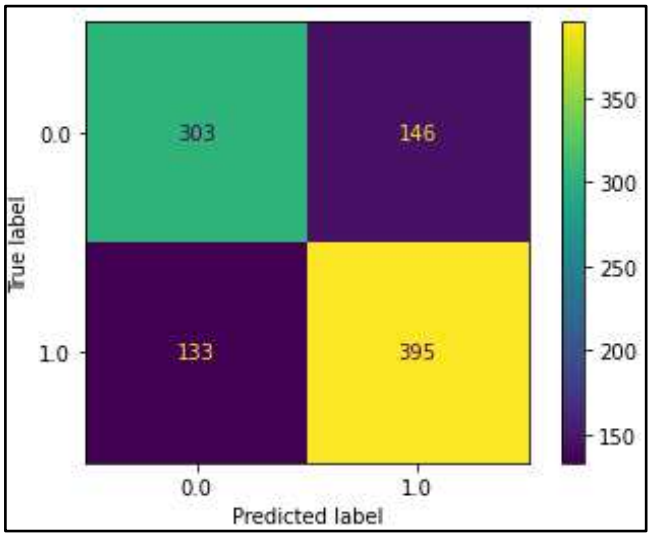

Fig. 6 Confusion matrix for decision tree

Fig. 6 depicts the Confusion Matrix for the decision tree method. TN value produces 303, FP value produces 146, FN value produces 133, and TP value produces 395. Fig. 7 depicts the output of the random forest algorithm's Confusion Matrix. According to the analysis, TN value produces 296 results, FP value produces 153 results, FN value produces 141 results, and TP value produces 387 results.

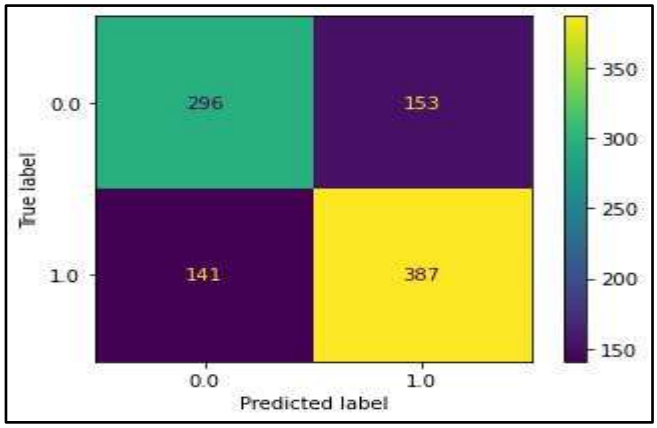

Fig. 7 Confusion matrix for random forest

TABLE III
RESULT COMPARISON BETWEEN ALGORITHMS

No	Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	Logistic Regression	57.83	58.41	76.33	66.00
2	Decision Tree	**71.44**	73.01	74.81	74.00
3	Random Forest	69.91	71.67	73.30	72.00

According to Table 3, the decision tree method yields the maximum accuracy, 71.44 %. Random forest was the second-best approach, with an accuracy score of 69.91%, while logistic regression yielded an accuracy score of 57.83%. All methods were implemented using the same dataset. Normalized and divided into two halves, 80% of the dataset is training data, and 20% is testing data.

IV. CONCLUSION

This study achieved its objective of forecasting customer loyalty in the hotel business by employing three classification techniques: logistic regression, random forest, and decision tree. The analysis results were meticulously recorded. By comparing the outcomes of the three algorithms, it is possible to establish that the decision tree method is the optimal approach for evaluating the hotel booking dataset, as it provides the greatest accuracy score (71.44%) among the three techniques considered. This research project can be enhanced with more training on many datasets and the incorporation of new or alternative approaches. In addition, this study project can be enhanced by employing more classification techniques to comprehend each approach's benefits and limits better.

ACKNOWLEDGMENT

This research project is under SEPADAN RE-SIP Grant (vot M074) from Universiti Tun Hussein Onn Malaysia (UTHM) and Findbulous Technology Sdn. Bhd.

REFERENCES

[1] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/ART20203995.

[2] J. P. Simon, "Artificial intelligence: scope, players, markets and geography," *Digit. Policy, Regul. Gov.*, vol. 21, no. 3, pp. 208–237, 2019.

[3] Y. Choi and J. W. Choi, "The prediction of hotel customer loyalty using machine learning technique," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 5, pp. 7908–7915, 2020, doi: 10.30534/ijatcse/2020/143952020.

[4] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-asl, and B. Yu, "methods , and applications," pp. 1–11, 2018.

[5] R. S. Ganga, P. C. P. Reddy, and B. C. Mohan, "System for Intelligent Tourist Information using Machine Learning Techniques Proof Logic Ontology vocabulary digital signatures," *Int. J. Appl. Eng. Res.*, vol. 13, no. 7, pp. 5321–5327, 2018.

[6] M. O. Parvez, "Use of machine learning technology for tourist and organizational services : high-tech innovation in the hospitality industry," *J. Tour. Futur.*, vol. 7, no. 2, pp. 240–244, 2021, doi: 10.1108/JTF-09-2019-0083.

[7] E. Mingotto, F. Montaguti, and M. Tamma, "Challenges in re-designing operations and jobs to embody AI and robotics in services. Findings from a case in the hospitality industry," *Electron. Mark.*, vol. 31, no. 3, pp. 493–510, 2021, doi: 10.1007/s12525-020-00439-y.

[8] H. Ruel and E. Njoku, "AI redefining the hospitality industry," *J. Tour. Futur.*, vol. 7, no. 1, pp. 53–66, 2020, doi: 10.1108/JTF-03-2020-0032.

[9] J. Wei *et al.*, "Machine learning in materials science," *InfoMat*, vol. 1, no. 3, pp. 338–358, 2019, doi: 10.1002/inf2.12028.

[10] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons.B*, vol. 4, pp. 51–62, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[11] E. Brynjolfsson and T. Mitchell, "What can machine learning do? Workforce implications," *Science (80-.).*, vol. 358, no. 6370, pp. 1530–1534, 2017.

[12] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 02, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[13] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," *Citizen-Based Mar. Debris Collect. Train. Study case Pangandaran*, vol. 6, no. 10, pp. 74–78, 2018.

[14] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[15] C. M. Yeşilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110210.

[16] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine , Random Forest , and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[17] E. Nazarenko, V. Varkentin, and T. Polyakova, "Features of Application of Machine Learning Methods for Classification of Network Traffic (Features , Advantages , Disadvantages)," *Int. Multi-Conference Ind. Eng. Mod. Technol.*, pp. 1–5, 2019, doi: 10.1109/FarEastCon.2019.8934236.

[18] P. Ranganathan, C. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Measures of agreement," *Perspect. Clin. Res.*, vol. 8, no. 3, pp. 148–151, 2017, doi: 10.4103/picr.PICR_123_17.

[19] N. A. M. R. Senaviratna and T. M. J. A. Cooray, "Diagnosing Multicollinearity of Logistic Regression Model," *Asian J. Probab. Stat.*, vol. 5, no. 2, pp. 1–9, 2019, doi: 10.9734/ajpas/2019/v5i230132.

[20] S. Uddin, A. Khan, E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 8, pp. 1–16, 2019.

[21] H. Sulistiani, K. Muludi, and A. Syarif, "Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification," *J. Phys. Conf. Ser.*, vol. 1338, pp. 1–8, 2019, doi: 10.1088/1742-6596/1338/1/012050.

[22] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, "Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study," *J. Big Data*, vol. 7, no. 1, pp. 1–24, 2020, doi: 10.1186/s40537-020-00290-0.

[23] R. Muttaqien, M. G. P, and A. Pramuntadi, "Implementation of Data Mining Using C4 . 5 Algorithm for Predicting Customer Loyalty of PT . Pegadaian (Persero) Pati Area Office," *Int. J. Comput. Inf. Syst.*, vol. 02, no. 03, pp. 64–68, 2021.

[24] J. Han, M. Fang, S. Ye, C. Chen, Q. Wan, and X. Qian, "Using Decision Tree to Predict Response Rates of Consumer Satisfaction , Attitude , and Loyalty Surveys," *Sustainability*, vol. 11, no. 2306, pp. 1–13, 2019.

[25] R. Wirth and J. Hipp, "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.

[26] W. Y. Ayele, "Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas using a Textual Dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020, doi: 10.14569/IJACSA.2020.0110603.

[27] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[28] R. Ribeiro, A. Pilastri, C. Moura, F. Rodrigues, R. Rocha, and P. Cortez, "Predicting the Tear Strength of Woven Fabrics Via Automated Machine Learning: An Application of the CRISP-DM Methodology," *ICEIS 2020 - Proc. 22nd Int. Conf. Enterp. Inf. Syst.*, vol. 1, pp. 548–555, 2020, doi: 10.5220/0009411205480555.

[29] Y. Watanabe *et al.*, "Preliminary Systematic Literature Review of Machine Learning System Development Process," *IEEE 45th Annu. Comput. Software, Appl. Conf.*, pp. 1407–1408, 2019.

[30] S. Studer *et al.*, "Towards CRISP-ML (Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, 2021.