Penerbit UTHM

© Universiti Tun Hussein Onn Malaysia Publisher's Office

JSCDM

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/jscdm

Journal of Soft Computing and Data Mining

Predicting Customer Loyalty Using Machine Learning for Hotel Industry

Iskandar Zul Putera Hamdan[1], Muhaini Othman[1*]

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jscdm.2022.03.02.004
Received 12 July 2022; Accepted 20 October 2022; Available online 01 November 2022

Abstract: The popularity of machine learning is growing and the demand for it is increasing in various fields including tourism and hospitality industry specifically hotels industry. The purpose of this research is to apply machine learning classification techniques to predict customers' loyalty in hotel company so that hotel company can use the result to create possible solutions for customer relationship management. The experiment will be performed by implementing CRISP-DM methodology and three proposed algorithms such as decision tree, random forest and logistic regression and the result will be compared with each other to obtain the best algorithm among them by using confusion matrix. The dataset that will be used is obtained from Findbulous technology company. From the analysis result, logistic regression, decision tree and random forest algorithms generate 57.83%, 71.44% and 69.91% accuracy score respectively. For further improvement, this research approach can be used with other dataset or implement a new algorithm to identify each algorithm strengths and limitations.

Keywords: Machine learning, classification, CRISP-DM, confusion matrix

1. Introduction

Nowadays, machine learning has been used in many sectors such as farm information management, customer's loyalty program, customer profiling management and others. Machine learning is one of the fields in computer science where it identifies the patterns of artificial intelligence and learns the computational learning theories [1]. Machine learning is a change in systems that can perform artificial intelligence (AI) related tasks and some of the examples of these tasks are recognition, analysis, planning, robot control and prediction [2]. It has attracted a lot of attention due to its ability to accurately forecast a wide range of complicated phenomena [3]. The popularity of machine learning is growing and the demand for it is increasing in any fields and that also include tourism and hospitality industry or specifically hotels industry. In basic terms, machine learning can be used for clustering, classification, and regression but it has also been used in hospitality and tourism for revenue management, customer satisfaction improvement and operational analytics [4]. According to [5], the goal of machine learning in hospitality is to put together arrangements for gathering data and learning from it, as well as to improve self-capability through experience without the interference of humans or basic reprogramming and it can be conducted step by step where experts first gather, select, organize, pre-process, and convert data sets to the machine before building analytical models.

In Findbulous Technology Sdn Bhd, the company has developed applications named C3 Rewards for merchants' customers and C3 Merchant for merchants. Merchants are other companies such as Gloria Hotels & Resorts, New York Hotel Johor Bahru and others who use Findbulous application service named C3 Merchant. The application or system provide merchants with features such as customer relationship management, customer voucher system, loyalty and rewards program and advanced booking engine and channel manager. When merchants' customers want to book a room

*Corresponding author: muhaini@uthm.edu.my

in the hotel, they can use the application by registering their account and continue their booking process without interacting via face to face with the hotel's employees. The application requires personal information from the customer during registration and the data will be kept in the database using Amazon Web Service (AWS) Data Lake. The applications have been developed using Python programming language and it have been maintained by Findbulous's developers regularly and they will fix any bugs that have been reported by merchants and merchants' customers. To improve C3 Merchant application, forecasting customer loyalty for hotel industry function need to be added into the application. Existing application does not provide merchant with data mining tools to predict their customers' loyalty.

Furthermore, no data mining research has been conducted to analyze the data from the application's database. To solve the problem, a project or research titled "Predicting Customer Loyalty Using Machine Learning for Hotel Industry" will be conducted to implement machine learning on the data that have been collected through the application. Objectives of this study are to investigate relevant variables of merchants' customers' behavior, interest, and preference and to analyze them using the proposed machine learning methods which are decision tree, random forest and logistic regression. Finally, the research will select and evaluate the best machine learning algorithm mentioned above for predicting customer loyalty in hotel industry. The selected algorithms will be implemented using Python programming language. The result from the research analysis will be evaluated further before being used as reference by hotel companies.

2. Related Work

This section will discuss about the overview of case study, machine learning, decision tree, random forest, and logistic regression. This section helps to educate those who are unfamiliar with machine learning. Moreover, comparison with related study also will be discussed in this section.

2.1 Overview of Case Study

In this research, the hotel booking dataset will be used as the chosen data sample that will be trained and tested using the proposed algorithms such as decision tree, random forest, and logistic regression to predict customers' loyalty and to get the accuracy and performance of each algorithm. Then, selected algorithms will be compared with each other to determine which one has the best expected result.

2.1.1 Machine Learning

In machine learning, there are three primary methods that have been identified. Those methods are supervised machine learning, unsupervised machine learning and semi supervised machine learning. In this research, supervised machine learning will be used along with classification technique to create prediction model from the dataset. Supervised machine learning is described through use of labelled datasets to train algorithms that reliably classify data or forecast the outcomes. The inputs will be converted into desired outputs through a function generated by various algorithms [6].

2.1.2 Decision Tree

Among numerous analysis techniques, decision tree is a sophisticated and widely used machine learning algorithm for forecasting and classifying large amounts of data. A decision tree is a tree-based approach in which each path starting from the root is characterized by a data splitting process until a Boolean outcome is attained at the leaf node. It is a hierarchical representation of knowledge relationships with nodes and links. Nodes indicate purposes when relations are used to categorize. Classification algorithms in machine learning can manage massive amounts of data. It can be utilized to form assumptions about category class names, to categorize knowledge based on class labels and training sets and to categorize recently available data. Some of the benefits from using decision tree approach is it can be used to classify both categorical and numerical outcomes however the attribute produced need to be a categorical one. Next, decision tree procedure is simple and easy to comprehend because the workflow of the process follows the similar method the human brain works and thinks. According to [1], decision tree, as opposed to being a black box algorithm such as support vector machine, nearest neighbor, and others, assist us to comprehend the reasons behind the data interpretation.

2.1.3 Random Forest

Random forest technique introduced by [7] is a form of ensemble approach that was developed for the goal of predicting the average of numerous independent base models in classification and regression methods for random forest framework. The ensemble technique is a strategy that employs numerous learning algorithms to improve predicted performance in classification and regression. Bagging is one of the ensemble methods used in random forest (Bootstrap Aggregation) and the approaches also can be used in a decision tree to reduce the variation of the decision tree. It continues to train trees on each of these 1,…,B subsamples and then aggregates all of the individual tree results into a single final estimation. Let B represent the total amount of trees grown, {φb, b = 1,...,B}depict all individual trees and bag depict the collection of all these trees. One of the advantages from implementing random forest technique is it can handle missing data values and retain the accuracy of the missing data. Next, random forest is capable to process data

with a high number of attributes and classes in an efficient manner. Furthermore, attribute values are insensitive to scaling (and, in general, to any monotonous transformations) [8].

2.1.4 Logistic Regression

Logistic regression is a classification procedure that estimates the likelihood of a target variable using supervised machine learning. Because the goal or dependent variable's structure is dichotomous, there are usually only two viable classes, in other words, the dependent variable is often binary, with data expressed as 1 (for yes) or 0 (for no). A logistic regression method can forecast p (y = 1) as a function of x mathematically. It is one of the most basic machine learning techniques that may be used to solve a variety of classification issues such as email spam detection, cancer diagnosis and others. There are several advantages from utilizing logistic regression to classify the dataset. One of them is that it can reveal not only the suitability of a predictor (coefficient size), but also the direction of the association (positive or negative). Next, this classifier is easily extensible to several classes by using multinomial regression and provides a natural probabilistic perspective of class predictions. Finally, logistic regression is one of the most basic machine learning algorithms and while it is simple to construct, it can yield excellent training efficiency in some circumstances. Because of these factors, training a model with this technique does not necessitate many computational resources.

2.2 Comparison with Existing Research

In this section, several existing research are selected and discussed to obtain more information that can be used to conduct the proposed research. Table 1 shows a summary of the related works.

Table 1 - The analysis of related works

No	Article	Algorithm	Total of instances	Accuracy (%)
1	"The prediction of Hotel Customer Loyalty using Machine Learning Technique" [2]	Decision Tree	119,386 samples	98.90
2	"Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification" [9]	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" [10]	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" [11]	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" [12]	Decision Tree	244 samples	78.26
		Logistic Regression	244 samples	73.10

3. Methodology

This section will discuss about the methodology that are used in conducting this study. CRISP-DM [13], which refers for Cross-Industry Standard Process for Data Mining is an industry-proven method for directing any data mining activities or research. According to research conduct by [14], CRISP-DM method is still the default standard for constructing data mining and knowledge discovery applications based on various user surveys and polls. By using six processes or phases includes Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment in CRISP-DM methodology, this research explained the implementation of various classifications models while employing the various techniques or algorithms covered in project scopes. Fig. 1 below is the life cycle of CRISP-DM reference model.

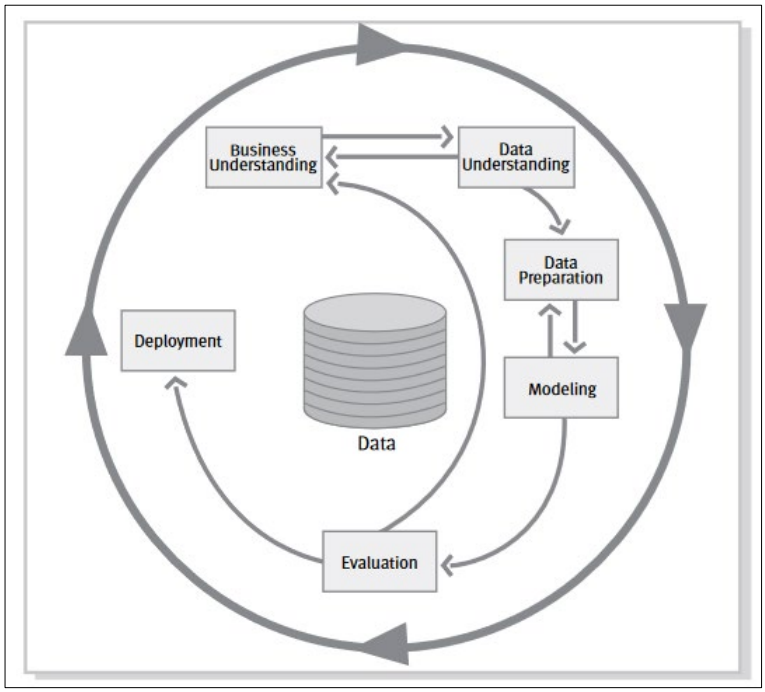


Fig. 1 - Processes of CRISP-DM [13]

3.1 Business Understanding Phase

Business understanding phase is a starting point of the research life cycle before continue to conduct the experiment. The objective of this phase is to concern with tasks like defining business objectives and translating them into machine learning objectives, collecting, and verifying data quality and finally assessing project feasibility. Checking the feasibility of the project before starting it is considered best practise for the overall success of the machine learning approach [15] and can reduce the risk of premature failures due to unrealistic expectations.

3.2 Data Understanding Phase

After objectives and scope of project have been identified, the data is collected through data collection process and goes through data quality verification where it includes three tasks such as data description, data requirements and data verification during data understanding phase.

3.3 Data Preparation Phase

During data preparation phase, the dataset is selected and the data that contain missing attributes or values are replaced. The dataset went through a cleansing phase as well where the tasks are to correct, impute or remove invalid values. Irrelevant features can be removed if needed to save space and time for computation. After that, normalization was used to pre-process the dataset. Data normalization is a data transformation method in which the hotel booking dataset need to be transformed into continuous values with values ranging from 0 to 1. It is important to note that the normalization that applied to the training set must also be applied to the test set using the same normalization parameters [16]. Eq. 1 below shows the formula for Min-Max normalization.

$$MinMax = \frac{(v - Min\ x)}{Max\ x - Min\ x}\ (newMax - newMin) + newMin \quad (1)$$

From the Eq. 1, Min is representing the minimum value of the attributes from the dataset, Max is representing the maximum value of the attributes from the dataset, V is the pick value of the row on every attribute of the dataset, newMax is used to set the maximum value into 1 and newMin is setting the minimum value as 0.

3.4 Modelling Phase

In modelling phase, the dataset needs to be prepared to train by using selected algorithms that have been stated in the project scope such as decision tree, random forest and logistic regression. The modelling phase's goal is to create one or more models that best suit the given requirements. In generating test design, it depends on the modelling approach whether the dataset need to be split into training, test and validation sets. For this study, the dataset is split into training and testing dataset. The algorithms for the model is implemented using Python programming language in any integrate development environment (IDE) that support Python language such as RapidMiner and Jupyter. The documentation is conducted to keep track of the experiment process and machine learning model.

3.5 Evaluation Phase

After the dataset have been trained with selected algorithms, the result is evaluated and compared in evaluation phase. The algorithms of classification's performance can be measured in terms of accuracy, precision, recall and f-measure to form classification's confusion matrix.

- **Accuracy**. The proportion of the total number of predictions was correct.

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

- **Precision**. The positive samples are defined as the number of samples correctly classified as positively divide by the total number of samples.

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

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

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

* Where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

- **F-measure**. Calculated as the weighted average of Precision and Recall

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

3.6 Deployment Phase

The last phase also known as deployment phase is the step where the selected algorithm from comparison results is deployed and trained with the dataset. Any changes that have been made whether in the dataset or the parameter selection of the algorithm is documented.

4. Results and Discussion

This section explained about the research design and implementation according to the proposed methodology. The methodology model will act as a guideline in conducting this research. In this section, there are a few important points that are discussed such as the proposed solution, research workflow or experiment design and parameter and testing methods to conduct classification techniques on the dataset.

4.1 Proposed Solution

Decision tree, random forest and logistic regression are the selected classification techniques that are implemented in this research. Each technique has its own engine and default parameter settings that allow it to generate accurate results with any dataset. The main goal of this study is to find the best algorithm that generates more score in term of accuracy based on the proposed algorithms and provided dataset. Thus, the classification techniques are conducted on the dataset with chosen parameter setting in this study as a solution for implementation.

4.2 Experiment Design, Test Bed and Simulation Setup

Designing the experiment phases is one of the critical parts to ensure the experiment's success. In the event that the experiment produces a bad result, tracking the workflow of the experiment can discover which section of the workflow may need to be changed or restarted from the beginning of the procedure. Fig. 2 below depicts a flowchart of the experiment stages that must be followed and completed for the experiment to be successful.

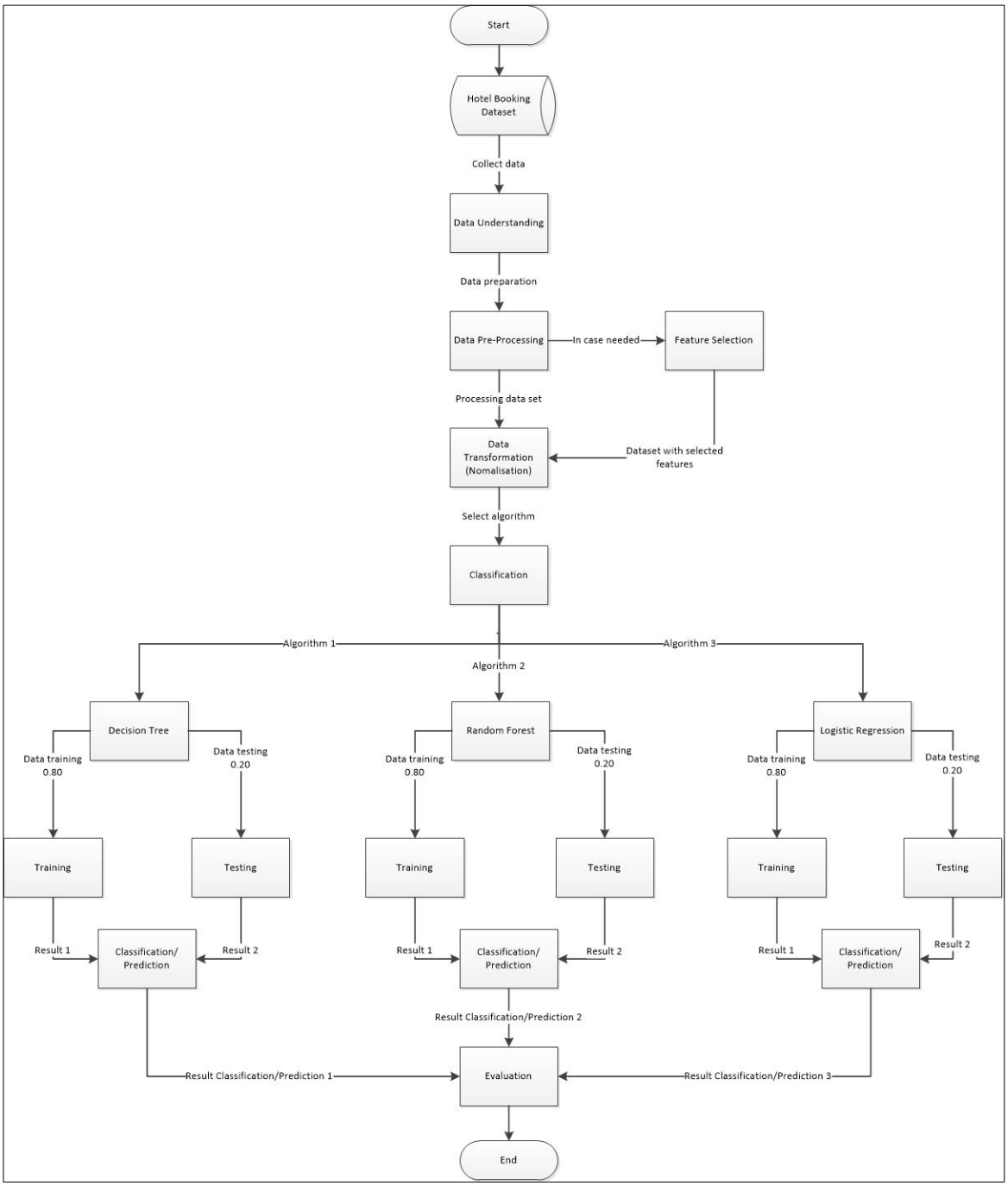


Fig. 2 - Flowchart of the experiment stages

Fig. 2 shows the flowchart of the experiment stages that has been used as a guideline for this research. The experimentation phase start after identifying the datasets that are going to be analyzed. As for this research, hotel booking data and related features are selected. Then, after securing the data, data are studied to understand the data features and its relation with the business process. Next, the dataset goes thru data preparation in order to make it easier to read by transforming the dataset from JavaScript Object Notation (JSON) format into Comma-separated values (CSV) format. Then the dataset undergo the data pre-processing stage to ensure that neither the dataset contains any missing values or

duplicate data. During pre-processing phase, the feature selection process is done to eliminate features that are not relevant to the model. If the dataset contains only a few relevant features, this process can be ignored.

The next step is the normalization phase where each value in the dataset must be optimised so that each procedure in the classification strategy may be easily interpreted. After data is processed and normalized, the data are feed into the selected classification algorithm. As previously stated, decision tree, random forest and logistic regression algorithms are selected to analyzed the data. During this phase, the dataset will be split into 80% for training dataset and 20% for testing dataset. This training-to-testing ratio will be applied to the identical setting for every classification approach. After the procedure is completed, each result will be recorded and must be evaluated to determine which one of the algorithms generate the higher score in term of accuracy to be selected as the best one.

In this study, executing this experiment requires the use of a test bed or also known as a platform which will serve as an environment for the experiment. There are numerous systems available including Microsoft Azure, Matlab, RapidMiner, Jupyter and Rstudio. In this research, Jupyter will be used as a platform and Python programming language will be used to programme every proposed classification algorithm. According to several research papers that conducted an experiment on data science, Python and R programming languages are well known for statistical analysis or exploratory data analysis. Python is a great programming language in computer science because it includes many data-oriented feature packages that help accelerate and optimise data processing and thus saving time. That is why Jupyter and Python programming language will be employed in this study.

As mentioned previously in test bed, the platform that will be used is Jupyter as well as Python programming language, which was the minimal prerequisite before initiating the research and experiment. However, the classification approach in Jupyter and Python programming language is not yet available. To solve this error, specific library packages must be imported including the pandas library, which supports a variety of data manipulation procedures such as merging, selecting, reshaping, data cleaning and data wrangling. Next, NumPy is a Python package that is used to work with arrays, linear algebra, the Fourier transform and matrices. Moreover, Scikit-learn or sklearn library package will be used as it provides efficient machine learning and statistical modelling capabilities such as classification including decision tree, random forest, and logistic regression. After all the required library packages have been imported, the experiment can run without any issues right away.

4.3 Parameter and Testing Methods

This section will discuss about the parameters used in the algorithms. As previous stated, one of the main goals for this research is to determine the accuracy of selected classification approaches based on the dataset of hotel booking. In this study, the dataset provides 4881 records for hotel booking with 22 attributes. A variable or parameter is a factor that is used to produce a prediction, which refers to a classification model such as logistic regression, decision tree and random forest. Furthermore, testing methods refer to what the assessment uses to support the outcome of each algorithm, whether true or false. The approach utilizes a confusion matrix, which illustrates the accuracy of the algorithms, experiment error and others.

Appendix A shows the list of features or attributes provided in the dataset. The dataset contains categorical and numerical data. To implement the classification on the dataset, some of the categorical data is converted from String data type into Integer data type. The “loyalty status” attribute has been selected as a class label which was a discrete attribute where the value is used to predict customer loyalty based on the values of other attributes. Before the dataset can be used for classification or prediction, it will need to go through data understanding, data preparation, data pre-processing, data cleaning, data transformation and data normalization. In data understanding phase, the dataset undergoes a Data Exploratory Analysis (EDA) to understand the format of the dataset, contents of the dataset and others.

In data preparation phase, first the dataset will be converted into any dataset format such as Comma Separated Values (CSV) file for easy understanding and access. For this experiment, the dataset will be converted from JSON to CSV file, so it can be easily used in Jupyter. Next, during data pre-processing, any categorical data in the dataset will be converted from String data type into a Numerical or Integer data type. The attributes that have been converted from String into Integer in the dataset were “status”, “room_names”, “customer_gender”, “c_title” and “booking_reason”.

Fig. 3 shows the heat map that represents Pearson Correlation Coefficient value. Pearson Correlation Coefficient is used to identify a relationship value between one attribute with other attributes in the same dataset. The attribute that has value between 0.0 and 1.0 indicate a high relationship while a value between 0.0 and -1.0 indicate a low relationship. For attribute with 0 score, it indicates that the attribute does not has any relation with other attributes. During data pre-processing phase, any data or attribute that contain more null values or bias data and do not have any connection with other attributes will be removed from the dataset in data cleaning process.

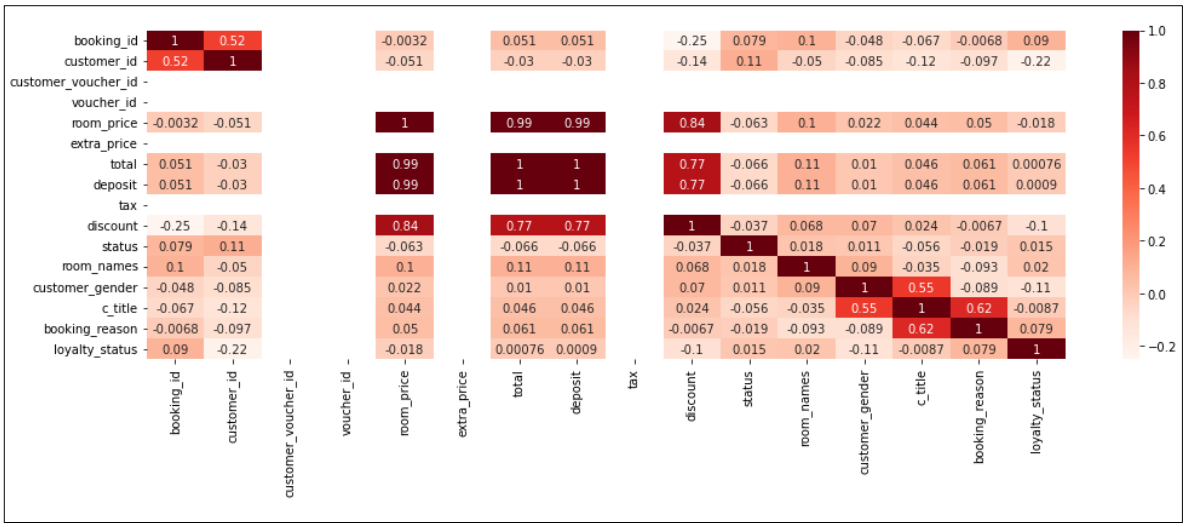


Fig. 3 - Heat map for Pearson Correlation Coefficient

The "customer_id" attribute has been removed to avoid any data bias in the dataset and "customer_race" need to be removed because it contains 70% more null values. The "customer_voucher_id", "voucher_id", "date_from", "date_to", "currency", "extra_price", "tax", "c_arrival" and "book_created" attributes will be removed because they do not have any relationship with another attributes based on Pearson Correlation Coefficient value. After the data cleaning process, new total number of attributes is obtained which are 11 attributes including the class label. In order to identify which attributes, have significant impact to the class label, Mutual Information function can be used on the dataset where it can calculate the statistical dependence between two variables.

Fig. 4 illustrate the Mutual Information value for each attribute. From the analysis, "discount" attribute generate 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 shows that "discount" attribute was the most important attribute in predicting customer loyalty because customer who got a discount in their booking transaction will be likely to become a loyal customer.

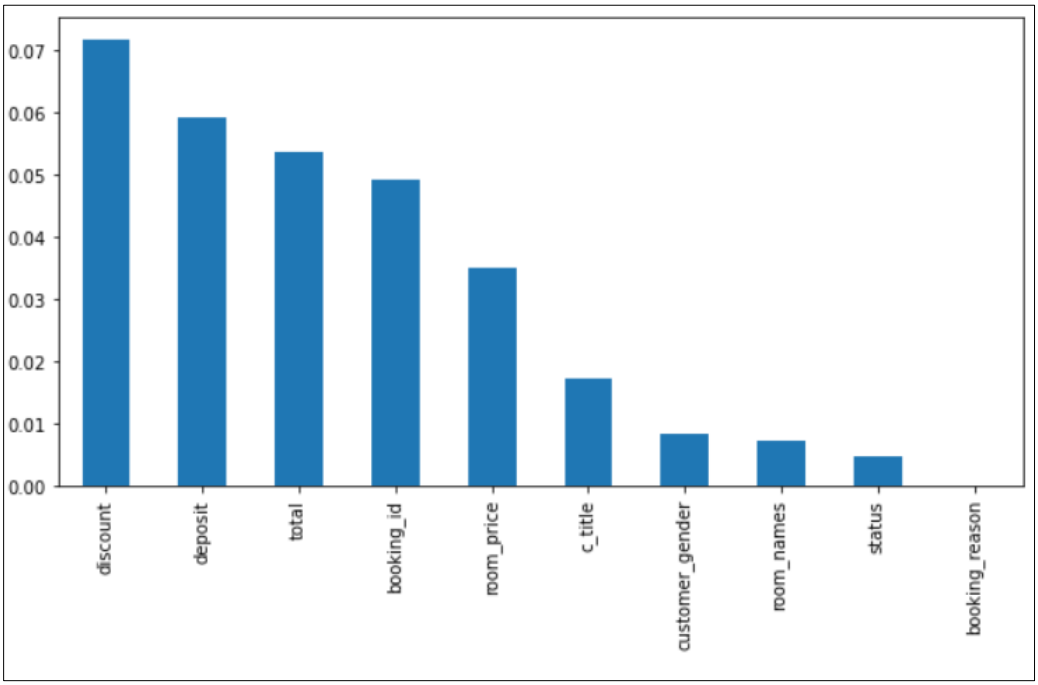


Fig. 4 - Mutual Information value for hotel booking dataset

4.4 Result Analysis and Evaluation

In this section, result analyses for all algorithms are identified and the outcomes of the analysis are presented in the tables below. Each process of each algorithm may have the similar procedure on managing the dataset but has a various method of conducting the experiments. Before conducting the classification process, the dataset will be normalized by using min-max normalization. One of the most prevalent methods for normalizing data is min-max normalization. For each attribute, the minimum value is converted to 0, the highest value is converted to 1 and all other values are converted to a decimal between 0 and 1. After the dataset has been normalized, it now can be used for classification by using proposed algorithms such as logistic regression, decision tree, and random forest. All 10 attributes and one class label are used for the classification procedure. After that, the normalized dataset is divided into two halves for training and testing purposes where 80% of the dataset is for training and another 20% for testing. Fig. 5 shows the result of confusion matrix for logistic regression algorithm. Based on the analysis, True Negative (TN) value generates 162, False Positive (FP) value generates 287, False Negative (FN) value generates 125 and True Positive (TP) value generates 403.

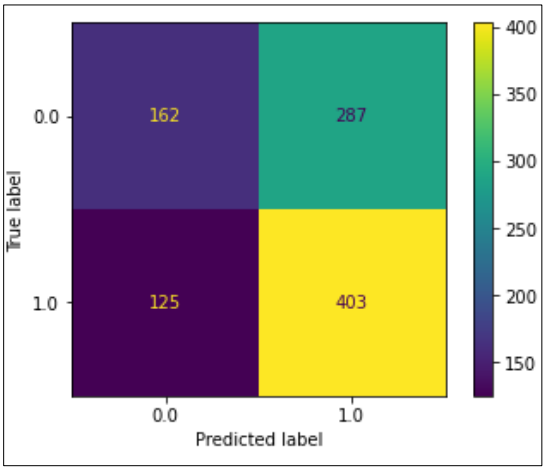


Fig. 5 - Confusion matrix for logistic regression

Fig. 6 shows the result of confusion matrix for logistic regression algorithm. Based on the analysis, True Negative (TN) value generates 303, False Positive (FP) value generates 146, False Negative (FN) value generates 133 and True Positive (TP) value generates 395.

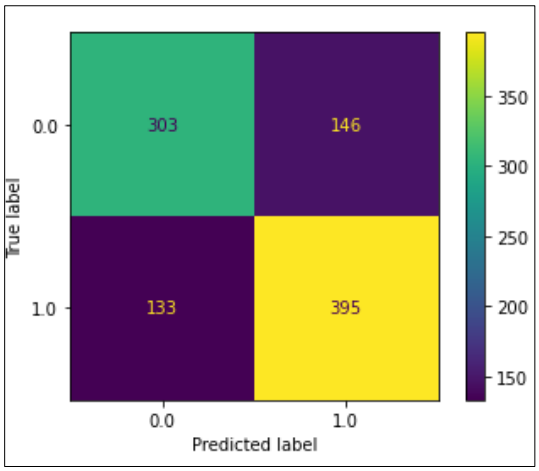


Fig. 6 - Confusion matrix for decision tree

Fig. 7 shows the result of confusion matrix for logistic regression algorithm. Based on the analysis, True Negative (TN) value generates 296, False Positive (FP) value generates 153, False Negative (FN) value generates 141 and True Positive (TP) value generates 387.

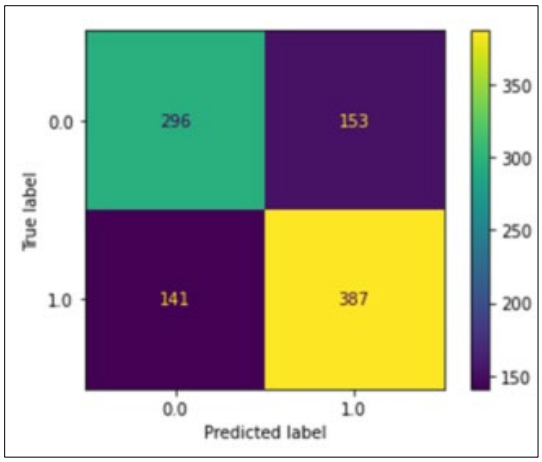


Fig. 7 - Confusion matrix for random forest

Based on Table 2, decision tree algorithm generates the highest accuracy which was 71.44%. The second-best algorithm was random forest with 69.91% accuracy score while logistic regression generates 57.83% accuracy score. All algorithms have been conducted by using the same dataset. The dataset has been normalized and split into two halves, 80% data for training data and 20% for testing data.

Table 2 - 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

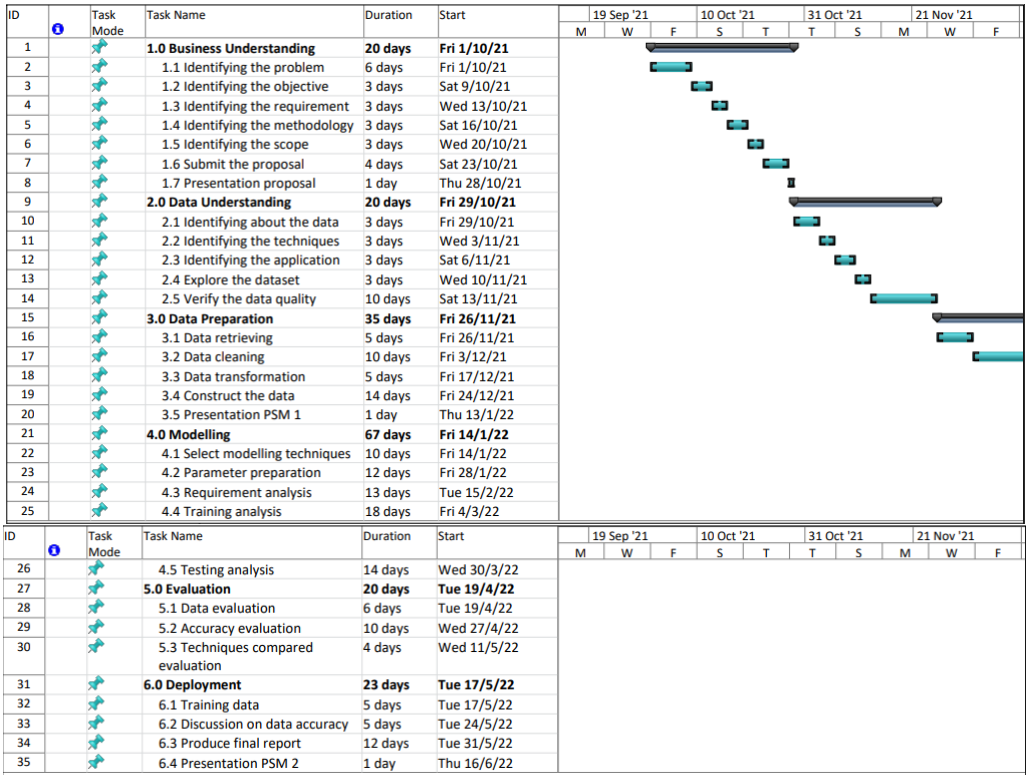
5. Conclusion

This research accomplished its purpose and goals by analyzing the data of hotel booking for predicting customer loyalty in hotel industry using three selected classification algorithms which are logistic regression, decision tree, and random forest and the findings were thoroughly documented and analyzed. By comparing the results from the three algorithms, it can be concluded that decision tree algorithm is proven to be the best algorithm to be utilized in analyzing the hotel booking dataset because it generates 71.44% score in term of accuracy which was the highest score among the selected algorithms. This project might be expanded in the future with further training on various datasets with the addition of new or different methodologies. Furthermore, this research project can be improved in the future by utilizing other classification algorithms to better understand the advantages and limitations of each algorithm.

Acknowledgement

This research is supported by the Universiti Tun Hussein Onn Malaysia under the Matching Grant Scheme Vot M074.

Appendix A: Gantt Chart of Predicting Customer Loyalty Using Machine Learning for Hotel Industry



ID	Task Mode	Task Name	Duration	Start
1		**1.0 Business Understanding**	**20 days**	**Fri 1/10/21**
2		1.1 Identifying the problem	6 days	Fri 1/10/21
3		1.2 Identifying the objective	3 days	Sat 9/10/21
4		1.3 Identifying the requirement	3 days	Wed 13/10/21
5		1.4 Identifying the methodology	3 days	Sat 16/10/21
6		1.5 Identifying the scope	3 days	Wed 20/10/21
7		1.6 Submit the proposal	4 days	Sat 23/10/21
8		1.7 Presentation proposal	1 day	Thu 28/10/21
9		**2.0 Data Understanding**	**20 days**	**Fri 29/10/21**
10		2.1 Identifying about the data	3 days	Fri 29/10/21
11		2.2 Identifying the techniques	3 days	Wed 3/11/21
12		2.3 Identifying the application	3 days	Sat 6/11/21
13		2.4 Explore the dataset	3 days	Wed 10/11/21
14		2.5 Verify the data quality	10 days	Sat 13/11/21
15		**3.0 Data Preparation**	**35 days**	**Fri 26/11/21**
16		3.1 Data retrieving	5 days	Fri 26/11/21
17		3.2 Data cleaning	10 days	Fri 3/12/21
18		3.3 Data transformation	5 days	Fri 17/12/21
19		3.4 Construct the data	14 days	Fri 24/12/21
20		3.5 Presentation PSM 1	1 day	Thu 13/1/22
21		**4.0 Modelling**	**67 days**	**Fri 14/1/22**
22		4.1 Select modelling techniques	10 days	Fri 14/1/22
23		4.2 Parameter preparation	12 days	Fri 28/1/22
24		4.3 Requirement analysis	13 days	Tue 15/2/22
25		4.4 Training analysis	18 days	Fri 4/3/22
26		4.5 Testing analysis	14 days	Wed 30/3/22
27		**5.0 Evaluation**	**20 days**	**Tue 19/4/22**
28		5.1 Data evaluation	6 days	Tue 19/4/22
29		5.2 Accuracy evaluation	10 days	Wed 27/4/22
30		5.3 Techniques compared evaluation	4 days	Wed 11/5/22
31		**6.0 Deployment**	**23 days**	**Tue 17/5/22**
32		6.1 Training data	5 days	Tue 17/5/22
33		6.2 Discussion on data accuracy	5 days	Tue 24/5/22
34		6.3 Produce final report	12 days	Tue 31/5/22
35		6.4 Presentation PSM 2	1 day	Thu 16/6/22

Appendix B: 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 confirm 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 for vacation or business trip
22	loyalty_status	Customer loyal level whether they are not loyal or loyal customer (**Class label**)

References

[1] J. Simon. Artificial intelligence: scope, players, markets and geography. Digital Policy, Regulation and Governance, 21(3), 208-237, 2019.

[2] Choi, Y., & Choi, J. W. The Prediction of Hotel Customer Loyalty using Machine Learning Technique. International Journal of Advanced Trends in Computer Science and Engineering, 9(5), 7908–7915, 2020. https://doi.org/10.30534/ijatcse/2020/143952020.

[3] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592, 2019.

[4] Ganga, R.S., Reddy, P.C.P. and Mohan, B.C, "System for intelligent tourist information using machine learning techniques", International Journal of Applied Engineering Research, Vol. 13 No. 7, pp. 5321-5327, 2018.

[5] Parvez, M. O. Use of machine learning technology for tourist and organizational services: High-tech innovation in the hospitality industry. Journal of Tourism Futures, 7(2), 240–244, 2020. https://doi.org/10.1108/jtf-09-2019-0083.

[6] Nasteski, V. An overview of the supervised machine learning methods. HORIZONS.B, 4, 51–62, 2017. https://doi.org/10.20544/horizons.b.04.1.17.p05

[7] Breiman, L. Random Forests. Machine Learning, 45, 5–32, 2001. https://doi.org/https://doi.org/10.1023/A:1010933404324

[8] Nazarenko, E., Varkentin, V., & Polyakova, T. Features of Application of Machine Learning Methods for Classification of Network Traffic (Features , Advantages , Disadvantages). International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 1–5, 2019. https://doi.org/10.1109/FarEastCon.2019.8934236

[9] Sulistiani, H., Muludi, K., & Syarif, A. Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification. Journal of Physics: Conference Series, 1338, 1–8, 2019. https://doi.org/10.1088/1742-6596/1338/1/012050.

[10] Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00290-0

[11] Muttaqien, R., P, M. G., & Pramuntadi, A. (2021). Implementation of Data Mining Using C4 . 5 Algorithm for Predicting Customer Loyalty of PT. Pegadaian (Parser) Pati Area Office. International Journal of Computer and Information System (IJCIS), 02(03), 64–68.

[12] Han, J., Fang, M., Ye, S., Chen, C., Wan, Q., & Qian, X. (2019). Using Decision Tree to Predict Response Rates of Consumer Satisfaction, Attitude, and Loyalty Surveys. Sustainability, 11(2306), 1–13.

[13] Wirth, R., & Hipp, J. Crisp-dm: towards a standard process modell for data mining, 2000.

[14] Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061, 2021. https://doi.org/10.1109/TKDE.2019.2962680.

[15] Watanabe, Y., Washizaki, H., Sakamoto, K., Saito, D., Honda, K., Tsuda, N. & Yoshioka, N. Preliminary systematic literature review of machine learning system development process. arXiv preprint arXiv:1910.05528, 2019.

[16] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. Towards crisp-ml(q): A machine learning process model with quality assurance methodology. Machine Learning and Knowledge Extraction, 3(2), 392–413, 2021.