OPEN ACCESS
ISSN 2356-5462
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 9, No. 2, Dec 2023. pp. 120-136
doi: doi.org/10.21108/ijoict.v9i2.864

XGBoost for Predicting Airline Customer Satisfaction Based on Computational Efficient Questionnaire

Nur Ghaniaviyanto Ramadhan [1*], Aji Gautama Putrada [2]

[1]*Software Engineering, Institut Teknologi Telkom Purwokerto, Purwokerto, Indonesia*
[2]*Telkom University, Bandung, Indonesia*
[*] ghani@ittelkom-pwt.ac.id

Abstract

Customer satisfaction can be created through a well-crafted service quality strategy, which forms the cornerstone of a successful business-customer relationship. Establishing and nurturing these relationships with customers is vital for long-term success. Within the airline industry, a persistent challenge lies in enhancing the passenger experience during flights, necessitating a comprehensive understanding of customer demands. Addressing this challenge is crucial for airlines aspiring to thrive in a competitive landscape, thus underlining the significance of providing top-notch services. This study addresses this issue by leveraging predictive airline customer satisfaction data analysis. We forecast customer satisfaction levels using a powerful Extreme Gradient Boosting (XGBoost) ensemble-based model. An integral aspect of our methodology involves handling missing values in the dataset, for which we utilize mean-value imputation. Furthermore, we introduce a novel logistic Pearson Gini (Log-PG) score to identify the factors that significantly influence airline customer satisfaction. In our predictive model, we achieved notable results, showing an accuracy and precision of 0.96. To ascertain the efficiency of our model, we conducted a comparative analysis with other boosting-type ensemble prediction models, such as gradient boosting and adaptive boosting (AdaBoost). The comparative assessment established the superiority of the XGBoost model in predicting airline customer satisfaction. The findings of this research have substantial implications and offer invaluable insights into enhancing airline customer satisfaction. Airlines can tailor their services to align with customer expectations by understanding customer contentment factors, ultimately fostering stronger relationships and enduring success in the industry.

Keywords: XGBoost, Prediction, Airline Customer, Missing Value

I. INTRODUCTION

IMPROVING the quality of in-flight services drives an airline's success [1]. Examples of existing services are in-flight food service, punctuality, friendliness of the flight attendants, cleanliness, and comfort of the airplane seats. One way to improve service quality is to assess the satisfaction of travel passengers. Customer satisfaction can bring a high level of loyalty for airlines so that it will be profitable for airlines [2],[3]. Several previous studies in the aviation industry have tried to classify the effect of service quality [4]. Several other studies address airline customer satisfaction using artificial intelligence models, such as deep and ensemble learning [5]–[7]. K. Hulliyah *et al.* [1] used several classification models such as k-nearest neighbor (KNN), Logistic Regression, Gaussian Naïve Bayes (NB), Decision Trees, and Random Forests. Eboli *et al.* [8], in their paper, used the logistic regression model which showed that several service aspects did not significantly affect

Received on 07 Nov 2023. Revised on 30 Nov 2023. Accepted and Published on 29 Dec 2023.

passenger satisfaction. W. Baswardono *et al.* [9], in their research, conducted a comparative analysis of the C4.5 algorithm and random forest for the classification of aircraft customer satisfaction. Ouf *et al.* [10] researched to improve classification results on airline customer satisfaction data using a deep neural network (DNN) with Adam optimization parameters. Several other classification models are used as comparisons, such as Random Forest, Support Vector Machine (SVM), and artificial neural network (ANN). Tan *et al.* [11] discussed predicting airline customer satisfaction levels using the bidirectional-long short-term memory (LSTM) model. Ezmaeilzadeh *et al.* [12] predicted factors using the nonlinear SVM model. Sankaranarayanan *et al.* analyzed airline customer satisfaction data for future predictions using the logistic model tree model [13]. Jiang, Xuchu, et al. discussed identifying the most important factors in airline customer satisfaction data using a recursive feature elimination (RFE) model based on random forest [14].

On the other hand, boosting-type ensemble methods are also an effective method on prediction cases. Abdurohman *et al.* [15] proved that extreme gradient boosting (XGBoost) performs better than autoregression-integrated moving average (ARIMA) on electricity load forecasting when handling sequential dataset with dynamic trend, while performing better than LSTM with limited dataset. Fauzan *et al.* [16] proved that adaptive boosting (AdaBoost), together with principal component analysis (PCA) can provide a low-bias and low-dimensional solution on sentiment analysis. Pane, *et al.* [17] proved that gradient boosting can predict department store sales based on department and holiday information better than other regression methods. There is a research opportunity to apply boosting-type ensemble methods on predicting airline customer satisfaction. In addition, this study will take advantage of the gap with previous research, namely that there is still no one who carried out the selection feature using the Log-PG technique. The research aim of our research is to apply XGBoost to predict airline customer satisfaction. The XGBoost model is used because it has been proven in several previous studies to be robust for prediction cases. This research starts from the availability of airline customer satisfaction datasets. The dataset is then processed to detect whether it has a missing value. If it does, it is replaced using statistical techniques. After the data has no missing value, we identify the features in the dataset that have an effect and do not have an effect, then look at the correlation between features. Predictions are made using the XGBoost model. Finally, we measure the performance of the prediction model.

Based on the presentation of previous research problems related to measuring airlines customer satisfaction, the scientific contributions to this research are:

1. A computationally efficient airline customer satisfaction dataset that is compressed using a validated random sampling method.
2. Log-PG, a novel feature selection model that can identify what factors are most influential in increasing customer satisfaction using a combination of PCC and Gini score.
3. A solution that can predict airline customer satisfaction using the XGBoost model.

II. LITERATURE REVIEW

K. Hulliyah et al. [1] conducted study to examine the aviation industry's rivalry and the elements that contribute to its success. This study used several classification models such as k-nearest neighbor (KNN), Logistic Regression, Gaussian Naïve Bayes (NB), Decision Trees, and Random Forests. The results of this study indicate that the Random Forest model produces the highest accuracy. The paper found that Wi-Fi Service is important for customer satisfaction. Eboli *et al.* [8], in their paper, aimed to provide a tool to measure aircraft passenger satisfaction and identify service aspects available at the terminal to offer services that are characterized by the best quality. The logistic regression model showed that several service aspects such as personnel assistance, airport appearance, airport directions, toilets in the terminal, availability, and frequency of arrivals, and connecting buses did not significantly affect passenger satisfaction.

Moreover, W. Baswardono *et al.* [9], in their research, conducted a comparative analysis of the C4.5 algorithm and random forest for the classification of aircraft customer satisfaction. The best results obtained by the C4.5 algorithm were 92.83%. However, the difference with the random forest algorithm was only 0.01%. Ouf *et al.* [10] researched to improve classification results on airline customer satisfaction data using a DNN with Adam optimization parameters. In that study, the dataset quality used should have been addressed. Several other classification models are used as comparisons, such as random forest, SVM, and ANN. The results obtained by the proposed method are superior to the comparison method, with an accuracy of 99.3%. Tan *et al.* [11] discussed predicting airline customer satisfaction levels using the bidirectional-LSTM (bi-LSTM) model. This research also identified what factors can increase customer satisfaction. The results obtained with the proposed model were 91.27%.

Furthermore, Ezmaeilzadeh *et al.* [12] examined one of the factors in measuring airline customer satisfaction, namely flight departure delay. This factor was predicted using the nonlinear SVM model. They conducted analytical data research to determine loyal customers to airlines using a modified multi-layer perceptron (MLP) model, which is a hybrid of several machine learning models [18]. Sankaranarayanan *et al.* [13] analyzed airline customer satisfaction data for future predictions using the logistic model tree model. Jiang, Xuchu, *et al.* [14] discussed identifying the most important factors in airline customer satisfaction data using a RFE model based on random forest. Even when the COVID-19 pandemic occurred, airline customer satisfaction was also observed [19]. Using AdaBoost as an airline passenger satisfaction prediction model is a research opportunity.

Several studies have proven that boosting-type ensemble learning performs better than other machine learning methods in certain cases. Abdurohman *et al.* [15] proved that XGBoost performs better than ARIMA on electricity load forecasting when handling sequential datasets with dynamic trends while performing better than LSTM with limited datasets. Fauzan *et al.* [16] showed that AdaBoost, together with PCA, can provide a low-bias and low-dimensional solution for sentiment analysis. Pane *et al.* [17] researched that gradient boosting could predict department store sales better based on department and holiday information than other regression methods. There is a research opportunity to apply boosting-type ensemble methods to predict airline customer satisfaction.

Some studies use random sampling in satisfaction surveys for reasons such as computation efficiency. Gopinath *et al.* [20] took 100 survey data from nurses with proportionate random sampling in a hospital case study of employee satisfaction. Prasetya *et al.* [21] applied random sampling in an online questionnaire to a case study of student satisfaction with e-learning during COVID-19. In their next study, Gopinath *et al.* [22] used stratified purposive random sampling to collect 250 respondents in a survey to find a link between job involvement and job satisfaction. Recent research by Gopinath *et al.* [23] used random sampling to collect 10% of the survey respondents looking for a relationship between organizational commitment and job satisfaction. Using random sampling in airline customer satisfaction prediction for computational efficiency is a research opportunity.

TABLE 1.
RELATED WORKS ON AIRLINE CUSTOMER SATISFACTION SURVEY PREDICTION

Ref.	**Airline Customer Satisfaction**	**Random Sampling**	**Feature Selection**	**Boosting Ensemble Prediction**
[1]	Yes	No	Yes	No
[8]	Yes	No	No	No
[9]	Yes	No	No	No
[10]	Yes	No	Yes	No
[12]	Yes	No	No	No
[13]	Yes	No	No	No
[14]	Yes	No	Yes	No
[15]	No	No	No	Yes
[16]	No	No	No	Yes
[17]	No	No	No	Yes
[20]	No	Yes	No	No
[21]	No	Yes	No	No
[22]	No	Yes	No	No
[23]	No	Yes	No	No
Proposed Method	Yes	Yes	Yes	Yes

The problem of missing values is a common problem in a data set. Several other studies also have this problem in the data used, such as medical datasets [24], [25], Meteorology, Climatology, and Geophysical Agency

(BMKG) [26], network intrusion [27], and ideal customers [28]. The techniques applied to this problem vary, such as statistics (mean and median) and machine learning (naïve Bayes and KNN) [29], [30]. Table 1 shows a comparison of state-of-the-art papers while highlighting the contributions of our research. In Table 1 comparison with the same features on dataset.

III. RESEARCH METHOD

Seen in **Error! Reference source not found.** is the research method used. This research starts from the availability of airline customer satisfaction datasets. The dataset is then processed to detect whether it has a missing value. If it does, it will be replaced using statistical techniques. After the data has no missing value, we identify the features in the dataset that have an effect and do not have an effect, then look at the correlation between features. Predictions are made using the AdaBoost model. Finally, we measure the performance of the prediction model.

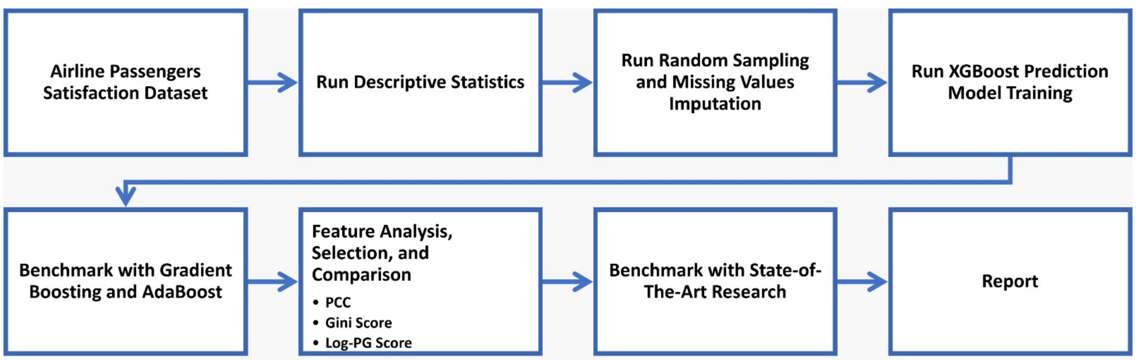

Fig. 1 Proposed Methodology Process

A. *Airline Customer Satisfaction Dataset with Missing Values and Feature Identification*

The airlines' passenger satisfaction dataset used in this study comes from the Kaggle repository. The dataset consists of 22 features and 25976 rows. In the dataset, there are categorical variables such as gender (female or male), type of customer (loyal or disloyal), type of travel (personal or business), class from airlines (business, economy, economy plus), and satisfaction (yes or no). The dataset we use is not imbalanced. Detailed characteristics related to datasets with numeric features can be seen in Table 2. The data used shows that the data collection technique used is a survey. It can be seen from the min and max values as a range.

Our dataset consists of demographics and survey items. Demographic data is categorical, while survey items are ordinal. Demographic data can be valuable for predicting satisfaction. Therefore we convert demographic data into ordinal data using a label encoder [31].

Random sampling is used to take a subset of a larger dataset in an unbiased and random way [32]. Fig. 2 shows the algorithm. X is our dataset, while Q is the desired number of samples. Proportion can also be used by using the proportion formula = Q/N. N is our dataset size. V_{min} is the smallest number in random number generation. V_{max} is the largest number in random number generation. Finally, X' is the dataset after random sampling.

This study uses the mean model to replace the missing row value [33]. The choice of the mean technique in handling missing values is due to adjusting to the characteristics of the data used and the number of rows of missing values. The airline satisfaction dataset used has a missing value in the arrival delay feature of 83 lines.

TABLE 2.
DATASET CHARACTERISTICS

Code	Features Name	Mean	Stdev	Min	Max
0	"Age"	39.6	15.13	7	85
1	"Inflight Wi-Fi service"	2.72	1.33	0	5
2	"Departure/Arrival service time convenient"	3.04	1.53	0	5
3	"Ease of online booking"	2.75	1.41	0	5
4	"Gate location"	2.97	1.28	1	5
5	"Food and drink"	3.21	1.33	0	5
6	"Online boarding"	3.25	1.35	0	5
7	"Seat comfort"	3.44	1.32	1	5
8	"Inflight entertainment"	3.35	1.33	0	5
9	"On-board service"	3.38	1.28	0	5
10	"Leg room service"	3.35	1.31	0	5
11	"Baggage handling"	3.63	1.17	1	5
12	"Check-in service"	3.31	1.26	1	5
13	"Inflight service"	3.64	1.18	0	5
14	"Cleanliness"	3.28	1.31	0	5
15	"Departure delay in minutes"	14.3	37.42	0	5
16	"Arrival delay in minutes"	14.74	37.5	0	5
17	"Satisfaction"	-	-	-	-

Algorithm 1: The random sampling algorithm.

Data: $X, Q, N, V_{min}, V_{max}$
Result: X'
Assign a unique ID to all data items in X;
for $n = 1$ *to* N **do**
  $r \leftarrow$ Generate a random number between V_{min} and V_{max};
  Assign r to X_n;
end
$S \leftarrow$ Sort X based on r;
$X' \leftarrow \{S_1, S_2, \ldots, S_Q\}$;

Fig. 2. The random sampling algorithm.

Identifying features of a set of datasets in the prediction process is an important step. Identifying correlations between features in the dataset can also lead to classification prediction accuracy results [34]. One way to identify correlations between features is to use heatmap visualization [35], [36]. This study uses a novel logistic Pearson Gini (Log-PG) score for feature selection. First we use the heatmap of Pearson correlation coefficient (PCC) to see the correlation between features in the dataset [37]. In addition, it uses the Gini score calculation to identify influential and non-influential factors. Lastly, we calculate the LPG formula to determine the feature selection [38]. The Equation 1 is the calculation of the novel Log-PG ($l(p,g)$) index:

$$l(p,g) = \frac{1}{1 + e^{-(0.1 + p + 0.2g)}} + 0.33 \tag{1}$$

where p is the absolute PCC score, while g is the Gini score. Scores lower than 1 are filtered.

B. Boosting-Type Ensemble Prediction Model

Boosting is a type of ensemble learning that makes weak learners a sequential series, where the weak learner after is an improvement from the previous weak learner [39]. The final decision of a booster is the aggregate of all these weak learners. We compared three types of boosting, namely XGBoost, AdaBoost, and gradient boosting.

1) XGBoost: XGBoost is a gradient-type boost that fixes the previous weak learner by minimizing its loss using gradients and hessian [40]. *Hessian* is the second-order gradient used in the Newton-Raphson method

Algorithm 2: The XGBoost algorithm.

Data: x, y, N, M

Result: $f(x)$

Initialize the ensemble using the following equation:

$$f_0(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta)$$

for $m = 1$ *to* M **do**

Calculate the gradient using the following equation:

$$g_m(x) = \left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]$$

Calculate the hessian using the following equation:

$$h_m(x) = \left[\frac{\partial^2 L(y, f(x))}{\partial f(x)^2}\right]$$

Approximate next prediction with the first-order Taylor expansion:

$$\phi_m = \sum_{i=1}^{N} \left[g_m(x_i) \times f_m(x_i) + \frac{1}{2} \times h_m(x_i) \times f_m(x_i)^2\right] + \Omega f_{m-1}(x_i)$$

Iteratively train weak learners with the following equation:

$$f_m(x) = f_{m-1}(x) + \alpha \phi_m$$

end

Return the final model, which has the following equation:

$$f(x) = \sum_{m=0}^{M} f_m(x)$$

[41]. [41].

Fig. 3 show the algorithm, where *x* is the feature, *y* is the label, *N* is the number of data items, *M* is the number of weak learners, θ is a constant, *L* is the loss function, *g* is the gradient, *h* is the hessian, ϕ is the optimization function, Ω is regularization, $\alpha$is the learning rate, and *f(x)* is prediction function.

Algorithm 2: The XGBoost algorithm.

Data: x, y, N, M

Result: $f(x)$

Initialize the ensemble using the following equation:

$$f_0(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta)$$

for $m = 1$ *to* M **do**

Calculate the gradient using the following equation:

$$g_m(x) = \left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]$$

Calculate the hessian using the following equation:

$$h_m(x) = \left[\frac{\partial^2 L(y, f(x))}{\partial f(x)^2}\right]$$

Approximate next prediction with the first-order Taylor expansion:

$$\phi_m = \sum_{i=1}^{N} \left[g_m(x_i) \times f_m(x_i) + \frac{1}{2} \times h_m(x_i) \times f_m(x_i)^2\right] + \Omega f_{m-1}(x_i)$$

Iteratively train weak learners with the following equation:

$$f_m(x) = f_{m-1}(x) + \alpha\phi_m$$

end

Return the final model, which has the following equation:

$$f(x) = \sum_{m=0}^{M} f_m(x)$$

Fig. 3. The XGBoost algorithm

The XGBoost algorithm performs several processes such as: model initialization using a constant model of each feature, Second, creating a weak model which is able to predict errors from previous models. Third, it calculates the residual error between the predicted value of the previous model and the original value of the target. Fourth, rearrange training data and create new models. Fifth, combine models and iterate repeatedly until they produce significant results.

2) Gradient Boosting: In contrast to AdaBoost, gradient boosting does not increase the *weight* of incorrectly predicted data. Instead, it applies gradient descent to the loss function so that the following weak learner loss function decreases according to the gradient descent [42]. Like XGBoost, gradient boosting aggregates all its weak learners, unlike AdaBoost, which does majority voting as the final step. Nevertheless, unlike XGBoost, gradient boosting does not take advantage of the hessian value of its loss function.

How Gradient Boosting works in general is as follows [42]: The first step, the first weak model, is built on the data. The initial predictions produced may not be accurate. Then, the following model is built to correct the prediction errors made by the previous model. This new model is focused on data that still has prediction errors. This process is repeated by adding subsequent models until it reaches a predetermined number or when predictions are not significantly improved.

The advantages of Gradient Boosting include its ability to produce accurate predictions, even from weak models, and its ability to handle complex and diverse data. One widespread implementation of Gradient Boosting is the XGBoost (eXtreme Gradient Boosting) algorithm, which has been successful in many fields, from data science to machine learning competitions.

3) AdaBoost: The AdaBoost model is used in this work to provide predictions. AdaBoost is a popular ensemble technique that creates a composite classifier before successively training the classifiers while emphasizing a specific pattern [43]. AdaBoost's central concept is to give each sample in the training set a weight [44].

Every sample in training set S is given the same weight $1/N$ at the beginning, which ensures that every sample has an equal probability of being chosen in the first step. It takes T rounds of basic training for learners using T separate training sample groups S_t, (t = 1, 2, ..., T), to generate the AdaBoost model. The function for calculating the weight of n observations in round t is designated by the symbol $D_t(n)$. The value of $D_t(n)$ is modified in accordance with how the observations were classified by the C_t classifier in each round after the learner construct C_t, which gives the F_t function to map x to {1, 1}, and the group training sample is then generated in the form of D_t on S by sample replacement.

4) Performance Metrics: One sample T-test can be used to prove the representativeness of the new dataset from random sampling [45]. The result of the T-test is the value of the t-statistic (t) (Equation 2).

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{Q}}} \tag{2}$$

where $\bar{x}$ is the sample mean, μ is the population mean, and s is the sample standard deviation. The degree of freedom (*df*) value is $Q - 1$ in the one sample t-test. The p-value is obtained from the t-table based on the t-value and the df value. We use a value of 0.05 for the significance level (α). If the p-value < α, then H_1 is accepted, alias the sample cannot represent the population. Conversely, when H_0 is accepted, the sample can represent the population.

We apply random sampling for more efficient computation. Knowing the time complexity of XGBoost is important to measure the performance of random sampling [46]. XGBoost has time complexity (T(t)) according to Equation 3.

$$T(t) = \mathrm{O}(\mathrm{M.Q.F.log(Q)}) \tag{3}$$

where M is the number of trees, Q is the number of samples, and F is the number of features.

We test the performance of our prediction model using a confusion matrix [47]. TABLE 3 shows an illustration of a confusion matrix. The target of the confusion matrix is the accumulated true negative (TN) and true positive (TP) values. Large values accumulating on false negatives (FN) and false positives (FP) indicate that the prediction performance is not good.

TABLE 3.
CONFUSION MATRIX

		Actual	
		Yes	No
Predicted	Yes	TP	FP
	No	FN	TN

Accuracy and precision are two metrics that can objectively measure the confusion matrix [48],[49]. The Equation 4 – 7 are used to calculate accuracy and precision:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} x\ 100\% \tag{4}$$

$$Precision = \frac{TP}{TP + FP} x\ 100\% \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

$$F1 - Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

In the case of predicting airline customer satisfaction, it is very important to ensure that the prediction model is balanced. If the model experiences overfitting, the consequences that can occur include poor generalization of new data, particularly when carrying out the next questionnaire. We calculate the four metrics above to prevent overfitting by applying k-fold cross-validation. We use k=3, dividing the data into 66.7% training and 33.3% testing data.

IV. RESULTS AND DISCUSSION

A. Result

The first result of the experiment is to handle the missing value. After that, we identify features that have an effect and do not have an effect using a heatmap and Gini score. We then conduct random sampling for computational efficiency. We use a sample fraction value of 20%, which means that the dataset is reduced from 129,880 data items to 25,976 data items. We conduct the one-sample T-Test to prove that the null hypothesis or H_0. If H_0 is accepted, the conclusion is that our dataset sample statistically represents the dataset population. Table 4 Shows the one-sample T-Test results. All variables are representative based on the test results while the number of data items is acceptable for airline customer satisfaction prediction.

In Table 4, we use the value $\alpha = 0.05$. Seeing that the overall P-Value result exceeds the α value, then H0 is declared accepted. While the new dataset is acceptable, the original time complexity as mentioned in Eq. (8) is O(M. Q.F.log(Q)), whereas the new time complexity becomes O(M.(0.2Q).F.log(0.2Q)). We test and compare the two datasets training time. The original dataset has a training time of 19.95 seconds, whereas the new dataset has a training time of 2.68 seconds. Both of which have the same accuracy score: 0.96.

Next, predictions are made using the proposed XGBoost model. Fig. 4 shows the confusion matrix value obtained by the XGBoost model. This study uses a dataset division of 70% for training and 30% for testing. The value of FP in the confusion matrix is 188, while FN is 131. Then the TP and TN values are 3200 and 4274, respectively.

TABLE 4.
RANDOM SAMPLING ONE-SAMPLE T-TEST FOR COMPUTATIONAL EFFICIENCY

Name	T-Stat	P-Value	H_0
“id”	0.968215	0.332946	Accepted
“Gender”	0.990272	0.32205	Accepted
“Customer Type”	-0.45023	0.652546	Accepted
“Age”	-0.06563	0.947675	Accepted
“Type of Travel”	0.463861	0.642751	Accepted
“Class”	-0.42021	0.674336	Accepted
“Flight Distance”	1.086325	0.277345	Accepted
“Inflight WI-FI service”	0.391214	0.695642	Accepted
“Departure/Arrival time convenient”	1.413376	0.157557	Accepted
“Ease of Online booking”	0.255183	0.798583	Accepted
“Gate location”	-0.68909	0.490773	Accepted
“Food and drink”	-0.18767	0.851133	Accepted
“Online boarding”	0.869094	0.384804	Accepted
“Seat comfort”	0.396061	0.692063	Accepted

"Inflight entertainment"	0.674591	0.499942	Accepted
"On-board service"	1.113969	0.265303	Accepted
"Leg room service"	-0.71574	0.474159	Accepted
"Baggage handling"	0.273272	0.784646	Accepted
"Checking service"	1.032882	0.301669	Accepted
"Inflight service"	1.03459	0.30087	Accepted
"Cleanliness"	-0.65666	0.511405	Accepted
"Departure Delay in Minutes"	-0.58837	0.556289	Accepted
"Arrival Delay in Minutes"	-0.6893	0.490641	Accepted
"satisfaction"	-0.44577	0.655768	Accepted

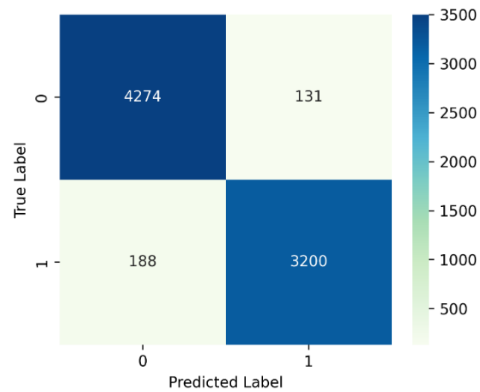

Fig. 4. Confusion matrix of XGBoost prediction.

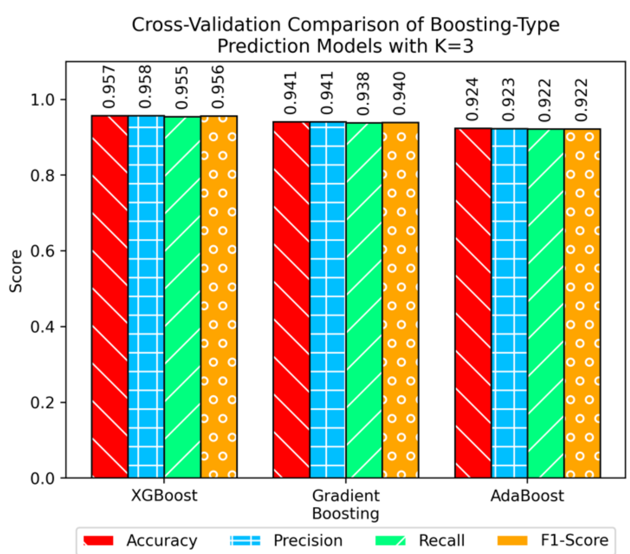

Fig. 5 Boosting-type ensemble methods performance comparison

TABLE 5.
FEATURE SELECTION BASED ON THE NOVEL LOG-PG SCORE METHOD

Name	PCC		Gini		Log-PG	
	Score	**Selected**	**Score**	**Selected**	**Score**	**Selected**
"Gender"	0.01	FALSE	0.01	FALSE	0.646192	FALSE
"Customer Type"	-0.19	FALSE	0.04	FALSE	0.689765	TRUE
"Age"	0.15	FALSE	0.03	FALSE	0.679253	TRUE
"Type of Travel"	-0.45	FALSE	0.1	TRUE	0.748011	TRUE
"Class"	-0.45	FALSE	0.09	TRUE	0.748621	TRUE
"Flight Distance"	0.3	TRUE	0.04	FALSE	0.713677	TRUE
"Inflight wifi service"	0.29	TRUE	0.14	TRUE	0.714821	TRUE
"Departure/Arrival time convenient"	-0.06	FALSE	0.02	FALSE	0.657979	FALSE
"Ease of Online booking"	0.17	TRUE	0.04	FALSE	0.685189	TRUE
"Gate location"	0	FALSE	0.02	FALSE	0.644934	FALSE
"Food and drink"	0.21	TRUE	0.01	FALSE	0.692579	TRUE
"Online boarding"	0.51	TRUE	0.17	TRUE	0.761763	TRUE
"Seat comfort"	0.35	TRUE	0.04	TRUE	0.724949	TRUE
"Inflight entertainment"	0.39	TRUE	0.06	TRUE	0.734893	TRUE
"On-board service"	0.32	TRUE	0.03	FALSE	0.718275	TRUE
"Leg room service"	0.31	TRUE	0.04	FALSE	0.717135	TRUE
"Baggage handling"	0.24	TRUE	0.02	FALSE	0.699879	TRUE
"Checkin service"	0.23	TRUE	0.02	FALSE	0.698309	TRUE
"Inflight service"	0.24	TRUE	0.02	FALSE	0.700646	TRUE
"Cleanliness"	0.3	TRUE	0.02	FALSE	0.714278	TRUE
"Departure Delay in Minutes"	-0.06	FALSE	0.01	FALSE	0.659279	FALSE
"Arrival Delay in Minutes"	-0.07	FALSE	0.01	FALSE	0.660646	FALSE

Next, we tested the performance of XGBoost using accuracy, precision, recall, and f1-score. We used k-fold cross-validation with k=3 to check for the model's overfitting symptoms. To highlight the performance of XGBoost, we compared it with two other boosting models, gradient boosting and AdaBoost. Fig. 5 shows the comparison, wherein the figure XGBoost outperforms the other two models in accuracy, precision, recall, and f1-score, with values of 0.957, 0.958, 0.955, and 0.956, respectively. The several reasons of why XGBoost performs better is that XGBoost implements advanced regularization techniques like shrinkage (learning rate) and column subsampling, reducing overfitting by controlling model complexity more effectively than traditional gradient boosting or AdaBoost. XGBoost's ability to capture complex nonlinear relationships between features and target variables through its sophisticated tree-based ensemble approach allows it to extract more nuanced patterns from the data, potentially resulting in better predictive performance compared to simpler boosting methods like AdaBoost.

We carry out calculations for PCC, Log-PG, and Gini on training data. The results obtained by visualizing the correlation between data features using a heatmap can be seen in Fig. 6. The correlation value search uses the PCC method. Comparisons are made only between two features which aims to find out whether the two features have a good or bad correlation. The five highest correlations between features are as follows: (1) "Ease of Online booking" and "Inflight WI-FI service" with a correlation value of 0.71. (2) The correlation between the "Inflight entertainment" and "Cleanliness" features is 0.69. (3) The correlation between "Seat comfort" and "Cleanliness" features is 0.68. (4) The "Food and drink" and "Cleanliness" feature has a correlation value of 0.66. (5) Lastly, the "Inflight entertainment" and "Food and drink" feature correlation is 0.62.

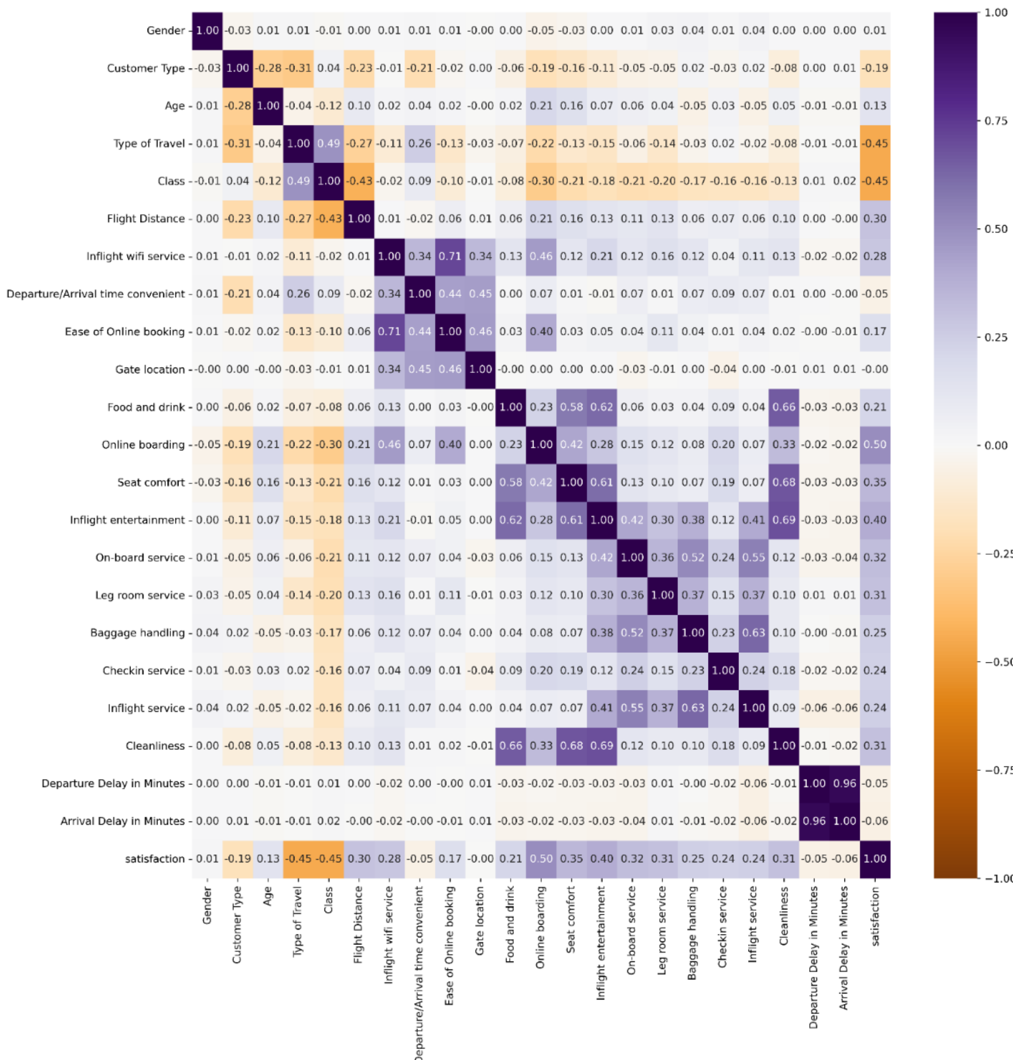

Fig. 6. PCC matrix of airline customer satisfaction features.

Next, the calculation of the Log-PG index from Equation (1) to determine the order of feature strength. Table 4 is a visualization of feature sequences. The top five features that significantly impact customer satisfaction are "Online boarding," "Inflight Wi-Fi service," "Type of travel," "Class of travel," and "Inflight entertainment." Meanwhile, the five features that do not affect customer satisfaction are "Gender," "Departure delay in minutes," "Food and drink," "Arrival delay in minutes," and "Gate location."

We again compared the performance of XGBoost by using it on some new feature sets selected by three methods: Log-PG, PCC, and Gini score. Fig. 7 shows the results. XGBoost with Log-PG does not degrade XGBoost performance before feature selection. On the other hand, applying feature selection with PCC and Gini scores reduces XGBoost's performance. XGBoost + PCC feature selection has the worst performance, with an accuracy of 0.790 and a precision of 0.793. Based on the results in Fig. 7, the proposed feature selection technique is far superior to other traditional techniques. This causes Log-PG to produce high accuracy and precision because the approach used is a combination of statistical between logistic regression and Gini coefficient. Thus, Log-PG when identifying influential features is clearer and more accurate.

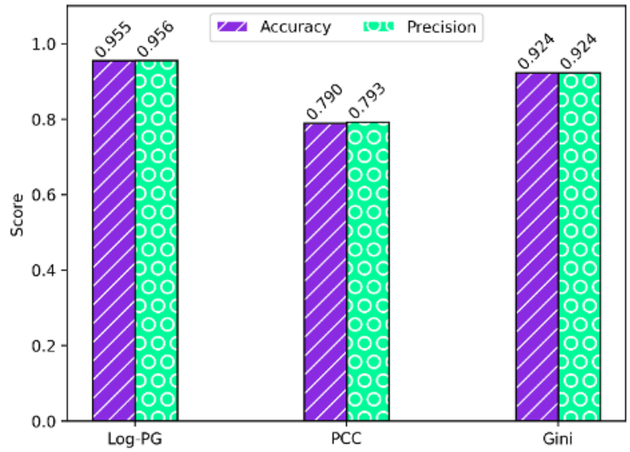

Fig. 7. XGBoost performance with three different feature selection methods.

B. Discussion

Several studies have used random sampling in surveys, such as papers [20]–[23]. In this study, we apply random sampling with a fraction rate 0.2. The new dataset reduces the training time from 19.95 seconds to 2.68 without reducing the prediction accuracy. The dataset has also been shown to be representative through the one-sample t-test. Our research contribution is a new dataset for airline customer satisfaction prediction, which is computationally efficient.

TABLE 6.
PERFORMANCE COMPARISON OF METHODS FROM STATE-OF-THE-ART AND OUR PROPOSED METHOD

Method	Proposed in	Accuracy	Precision
KNN	[1], [14]	0.91	0.92
Logistic regression	[1], [8], [14]	0.81	0.81
Decision tree	[1]	0.93	0.93
Random forest	[1], [10], [14]	0.95	0.95
DNN	[10]	0.92	0.92
ANN	[10]	0.89	0.90
SVM	[10], [12]	0.94	0.94
Gaussian NB	[14]	0.84	0.84
XGBoost	Proposed Method	**0.96**	**0.96**

Research on airline customer satisfaction prediction already exists, such as [1], [8], [10], [12], [14]. However, these studies have not used boosting-type ensemble methods such as XGBoost, gradient boosting, and AdaBoost. In addition, these studies have not used the Kaggle questionnaire data for airline customer satisfaction. We reused the methods proposed in each research: KNN, logistic regression, decision tree, random forest, DNN, ANN, SVM, and Gaussian NB, on the Kaggle dataset. TABLE 6 compares the prediction methods from state-of-the-art research with our proposed method. Comparison done with previous studies by looking at the dataset used, the same dataset used derived from the online questionnaire. It shows that our proposed method has superior performance. An interesting fact is that DNN, which leverages feature learning, shows better performance than ANN. The results could be a research opportunity to use stacked learning ensembles in the future to enhance the performance of our studies.

Several studies have used airline customer satisfaction prediction using the Kaggle dataset, such as papers [9], [11], which use random forest and Bi-LSTM, respectively. Here we use XGBoost and get better performance. XGBoost can outperform other models due to its ability to handle complex data, flexibility in tunable parameters, and handling the possibility of overfitting. The comparison accuracy is 95.85% versus 92.83% and 91.27%, respectively. TABLE 7 shows a direct comparison of the performance of our proposed method with state-of-the-art methods. Between boosting-type ensemble methods such as gradient boosting and AdaBoost, XGBoost is the method proven to increase the prediction performance of airline customer satisfaction based on a questionnaire. Our research contribution is an XGBoost model to apply predictions to airline customer satisfaction based on a questionnaire.

TABLE 7

COMPARISON WITH OTHER RESEARCH WITH KAGGLE AIRLINE CUSTOMER SATISFACTION DATASET

Reference	**Model**	**Accuracy (%)**
[9]	Random forest	0.928
[11]	Bi-LSTM	0.913
Proposed Method	XGBoost	**0.959**

Several studies have used feature selection as a part of the methodology in airline customer satisfaction prediction. Papers [1], [10] used the PCC score method, while paper [14] used the Gini score. In our research, we use Log-PG, a novel method combining PCC and Gini scores. Unlike the PCC and Gini scores, the results show that XGBoost's performance on the Log-PG result feature can be maintained. Our research contribution is a novel Log-PG method that combines PCC and Gini scores for feature selection in airline customer satisfaction prediction. The advantage of the novel Log-PG method is that it does not depend on data distribution and carefully considers interactions between features.

V. CONCLUSION

This study aims to predict airline customer satisfaction using the XGBoost model. In this study, the missing values are handled using mean imputation and determining what factors had and did not affect customer satisfaction results using the novel Log-PG score. The results show that the proposed model produced an accuracy of 0.96 and a precision of 0.96. The results are also compared with other models where the proposed model produces superior accuracy and precision. This study also identified the most influential factor, "Online boarding," while the factor with the least Log-PG score is "Gender." Based on the results obtained, this study aims to successfully predict airline customer satisfaction and determine the factors that influence customer satisfaction. XGBoost ensemble-based models are also superior to other models AdaBoost, DNN, NB, and Random Forest. This study also has better results than state-of-the-art research. In future research, it is possible to predict airline customer satisfaction using other meta-learning methods or stacked ensemble-based and deep learning-based model approaches.

REFERENCES

[1] K. Hulliyah, “Predicting Airline Passenger Satisfaction with Classification Algorithms,” *IJIIS Int. J. Inform. Inf. Syst.*, vol. 4, no. 1, pp. 82–94, Mar. 2021, doi: 10.47738/ijiis.v4i1.80.

[2] C.-F. Chen, “Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: Evidence from Taiwan,” *Transp. Res. Part Policy Pract.*, vol. 42, no. 4, pp. 709–717, May 2008, doi: 10.1016/j.tra.2008.01.007.

[3] S. Tiernan, D. L. Rhoades, and B. Waguespack, “Airline service quality: Exploratory analysis of consumer perceptions and operational performance in the USA and EU,” *Manag. Serv. Qual. Int. J.*, vol. 18, no. 3, pp. 212–224, May 2008, doi: 10.1108/09604520810871847.

[4] B. K. Behn and R. A. Riley, “Using Nonfinancial Information to Predict Financial Performance: The Case of the U.S. Airline Industry,” *J. Account. Audit. Finance*, vol. 14, no. 1, pp. 29–56, Jan. 1999, doi: 10.1177/0148558X9901400102.

[5] P. Kunekar, M. Deshpande, A. Gharpure, V. Gokhale, A. Gore, and H. Yadav, “Evaluating the Predictive Ability of the LightGBM Classifier for Assessing Customer Satisfaction in the Airline Industry,” in *2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India: IEEE, Jan. 2023, pp. 1–6. doi: 10.1109/ICONAT57137.2023.10080120.

[6] S.-H. Park, M.-Y. Kim, Y.-J. Kim, and Y.-H. Park, “A Deep Learning Approach to Analyze Airline Customer Propensities: The Case of South Korea,” *Appl. Sci.*, vol. 12, no. 4, p. 1916, Feb. 2022, doi: 10.3390/app12041916.

[7] R. Pranav and H. S. Gururaja, “Explainable Stacking Machine Learning Ensemble for Predicting Airline Customer Satisfaction,” in *Third Congress on Intelligent Systems*, S. Kumar, H. Sharma, K. Balachandran, J. H. Kim, and J. C. Bansal, Eds., in Lecture Notes in Networks and Systems, vol. 608. Singapore: Springer Nature Singapore, 2023, pp. 41–56. doi: 10.1007/978-981-19-9225-4_4.

[8] L. Eboli and G. Mazzulla, “An ordinal logistic regression model for analysing airport passenger satisfaction,” *EuroMed J. Bus.*, vol. 4, no. 1, pp. 40–57, May 2009, doi: 10.1108/14502190910956684.

[9] W. Baswardono, D. Kurniadi, A. Mulyani, and D. M. Arifin, “Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification,” *J. Phys. Conf. Ser.*, vol. 1402, no. 6, p. 066055, Dec. 2019, doi: 10.1088/1742-6596/1402/6/066055.

[10] S. Ouf, “An Optimized Deep Learning Approach for Improving Airline Services,” *Comput. Mater. Contin.*, vol. 75, no. 1, pp. 1213–1233, 2023, doi: 10.32604/cmc.2023.034399.

[11] C. Tan, “Bidirectional LSTM Model in Predicting Satisfaction Level of Passengers on Airline Service,” in *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, Hangzhou, China: IEEE, Nov. 2021, pp. 525–531. doi: 10.1109/ICAICE54393.2021.00107.

[12] E. Esmaeilzadeh and S. Mokhtarimousavi, “Machine Learning Approach for Flight Departure Delay Prediction and Analysis,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2674, no. 8, pp. 145–159, Aug. 2020, doi: 10.1177/0361198120930014.

[13] H. B. Sankaranarayanan, B. V. Vishwanath, and V. Rathod, “An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees,” in *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Kolkata, India: IEEE, Sep. 2016, pp. 285–290. doi: 10.1109/ICRCICN.2016.7813672.

[14] X. Jiang, Y. Zhang, Y. Li, and B. Zhang, “Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model,” *Sci. Rep.*, vol. 12, no. 1, p. 11174, Jul. 2022, doi: 10.1038/s41598-022-14566-3.

[15] M. Abdurohman and A. G. Putrada, “Forecasting Model for Lighting Electricity Load with a Limited Dataset using XGBoost,” *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, pp. 197–206, Jul. 2019, doi: 10.22219/kinetik.v4i3.841.

[16] M. N. Fauzan, A. G. Putrada, N. Alamsyah, and S. F. Pane, "PCA-AdaBoost Method for a Low Bias and Low Dimension Toxic Comment Classification.," in *2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS)*, IEEE, 2022, pp. 1–6, doi: 10.1109/ICACNIS57039.2022.10055017.

[17] S. F. Pane, A. G. Putrada, N. Alamsyah, and M. N. Fauzan, "A PSO-GBR Solution for Association Rule Optimization on Supermarket Sales," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, IEEE, 2022, pp. 1–6, doi: 10.1109/ICIC56845.2022.10007001.

[18] S. J. Parvez, "Data Analytics for Finding Loyalty of International Airline Passengers Using Deep Network MLP Combining with Machine Learning Algorithms on Python," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. SP7, pp. 2886–2891, Jul. 2020, doi: 10.5373/JARDCS/V12SP7/20202431.

[19] S. M. Zahraee *et al.*, "A study on airlines' responses and customer satisfaction during the COVID-19 pandemic," *Int. J. Transp. Sci. Technol.*, p. S2046043022001009, Dec. 2022, doi: 10.1016/j.ijtst.2022.11.004.

[20] O. Atika, A. T. Junaedi, A. A. Purwati, and Z. Mustafa, "Work Discipline, Leadership, and Job Satisfaction on Organizational Commitment and Teacher Performance of State Junior High School in Bangko District, Rokan Hilir Regency," *J. Appl. Bus. Technol.*, vol. 3, no. 3, pp. 251–262, Sep. 2022, doi: 10.35145/jabt.v3i3.109.

[21] T. A. Prasetya, C. T. Harjanto, and A. Setiyawan, "Analysis of student satisfaction of e-learning using the end-user computing satisfaction method during the Covid-19 pandemic," *J. Phys. Conf. Ser.*, vol. 1700, no. 1, p. 012012, Dec. 2020, doi: 10.1088/1742-6596/1700/1/012012.

[22] D. R. Gopinath and D. R. Kalpana, "RELATIONSHIP OF JOB INVOLVEMENT WITH JOB SATISFACTION," *Adalya J.*, vol. 9, no. 7, Jul. 2020, doi: 10.37896/aj9.7/029.

[23] Md. A. I. Gazi, Md. A. Islam, J. Shaturaev, and B. K. Dhar, "Effects of Job Satisfaction on Job Performance of Sugar Industrial Workers: Empirical Evidence from Bangladesh," *Sustainability*, vol. 14, no. 21, p. 14156, Oct. 2022, doi: 10.3390/su142114156.

[24] N. G. Ramadhan, Adiwijaya, and A. Romadhony, "Preprocessing Handling to Enhance Detection of Type 2 Diabetes Mellitus based on Random Forest," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, 2021, doi: 10.14569/IJACSA.2021.0120726.

[25] M. Liu *et al.*, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artif. Intell. Med.*, vol. 142, p. 102587, Aug. 2023, doi: 10.1016/j.artmed.2023.102587.

[26] I. D. Oktaviani and A. G. Putrada, "KNN imputation to missing values of regression-based rain duration prediction on BMKG data," *J. INFOTEL*, vol. 14, no. 4, Nov. 2022, doi: 10.20895/infotel.v14i4.840.

[27] M. Szczepański, M. Pawlicki, R. Kozik, and M. Choraś, "The Application of Deep Learning Imputation and Other Advanced Methods for Handling Missing Values in Network Intrusion Detection," *Vietnam J. Comput. Sci.*, vol. 10, no. 01, pp. 1–23, Feb. 2023, doi: 10.1142/S2196888822500257.

[28] N. G. Ramadhan, "Data Mining Techniques in Handling Personality Analysis for Ideal Customers," *J. OfInformation Syst. Eng. Bus. Intell.*, vol. 8, no. 2, pp. 175–181, Oct. 2022, doi: 10.20473/jisebi.8.2.175-182.

[29] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/s10462-019-09709-4.

[30] T. Aittokallio, "Dealing with missing values in large-scale studies: microarray data imputation and beyond," *Brief. Bioinform.*, vol. 11, no. 2, pp. 253–264, Mar. 2010, doi: 10.1093/bib/bbp059.

[31] K. Shyamala and C. S. Padmasini, "Mtanh-Attention-BiLSTM model for prediction of Automobile Export Customer Satisfaction," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, 2021, pp. 652–660, doi: 10.1109/ICAIS50930.2021.9395837.

[32] C. Kadilar and H. Cingi, "Ratio estimators in simple random sampling," *Appl. Math. Comput.*, vol. 151, no. 3, pp. 893–902, Apr. 2004, doi: 10.1016/S0096-3003(03)00803-8.

[33] S. Elmståhl, J. Sanmartin Berglund, C. Fagerström, and H. Ekström, "The life satisfaction index-A (LSI-A): normative data for a general Swedish population aged 60 to 93 years," *Clin. Interv. Aging*, pp. 2031–2039, 2020, doi: 10.2147/CIA.S275387.

[34] K. Patidar, R. K. Gour, A. Dixit, M. Verma, and A. K. Pal, "An Improved Method for the Data Cluster Based Feature Selection and Classification," in *2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India: IEEE, Jan. 2023, pp. 1–6. doi: 10.1109/ICONAT57137.2023.10080669.

[35] S. Buyrukoğlu and A. Akbaş, "Machine Learning based Early Prediction of Type 2 Diabetes: A New Hybrid Feature Selection Approach using Correlation Matrix with Heatmap and SFS," *Balk. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 110–117, Apr. 2022, doi: 10.17694/bajece.973129.

[36] F. Viton, M. Elbattah, J.-L. Guerin, and G. Dequen, "Heatmaps for Visual Explainability of CNN-Based Predictions for Multivariate Time Series with Application to Healthcare," in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, Oldenburg, Germany: IEEE, Nov. 2020, pp. 1–8. doi: 10.1109/ICHI48887.2020.9374393.

[37] M. B. Satrio, A. G. Putrada, and M. Abdurohman, "Evaluation of Face Detection and Recognition Methods in Smart Mirror Implementation," in *Proceedings of Sixth International Congress on Information and Communication Technology*, Springer, 2022, pp. 449–457, doi: 10.1007/978-981-16-2380-6_39.

[38] Hong Han, Xiaoling Guo, and Hua Yu, "Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest," in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China: IEEE, Aug. 2016, pp. 219–224. doi: 10.1109/ICSESS.2016.7883053.

[39] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, "Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey," *IEEE Access*, vol. 10, pp. 45137–45178, 2022, doi: http://doi.org/10.1109/ACCESS.2022.3169765.

[40] N. K. Le *et al.*, "Fedxgboost: Privacy-preserving xgboost for federated learning," *ArXiv Prepr. ArXiv210610662*, 2021, doi: https://doi.org/10.48550/arXiv.2106.10662.

[41] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, "XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data," in *2022 International Symposium on Electronics and Smart Devices (ISESD)*, Nov. 2022, pp. 1–7. doi: http://doi.org/10.1109/ISESD56103.2022.9980630.

[42] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, and O. Lyashevska, "Predictive analytics with gradient boosting in clinical medicine," *Ann. Transl. Med.*, vol. 7, no. 7, 2019, doi: 10.21037/atm.2019.03.29.

[43] A. G. Putrada and D. Perdana, "Improving Thermal Camera Performance in Fever Detection during COVID-19 Protocol with Random Forest Classification," in *2021 International Conference*

Advancement in Data Science, E-learning and Information Systems (ICADEIS), Bali, Indonesia: IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ICADEIS52521.2021.9702045.

[44] L. Zhou and K. K. Lai, “AdaBoost Models for Corporate Bankruptcy Prediction with Missing Data,” *Comput. Econ.*, vol. 50, no. 1, pp. 69–94, Jun. 2017, doi: 10.1007/s10614-016-9581-4.

[45] N. A. C. Cressie and H. J. Whitford, “How to Use the Two Samplet-Test,” *Biom. J.*, vol. 28, no. 2, pp. 131–148, 1986, doi: 10.1002/bimj.4710280202.

[46] X. Deng, M. Li, S. Deng, and L. Wang, “Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification,” *Med. Biol. Eng. Comput.*, vol. 60, no. 3, pp. 663–681, 2022, doi: https://doi.org/10.1007/s11517-021-02476-x.

[47] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: https://doi.org/10.1016/j.patcog.2019.02.023.

[48] N. G. Ramadhan and F. D. Adhinata, “Sentiment analysis on vaccine COVID-19 using word count and Gaussian Naïve Bayes,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, p. 1765, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1765-1772.

[49] N. G. Ramadhan, M. Wibowo, N. F. L. Mohd Rosely, and C. Quix, “Opinion mining indonesian presidential election on twitter data based on decision tree method,” *J. INFOTEL*, vol. 14, no. 4, Nov. 2022, doi: 10.20895/infotel.v14i4.832.