EasyChair Preprint

№ 6624

Feature Selection and Adaptive Synthetic Sampling Approach for Optimizing Online Shopper Purchase Intent Prediction

Rizal Dwi Prayogo and Siti Amatullah Karimah

September 16, 2021

Feature Selection and Adaptive Synthetic Sampling Approach for Optimizing Online Shopper Purchase Intent Prediction

1st Rizal Dwi Prayogo
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
rizaldp@itb.ac.id

2nd Siti Amatullah Karimah
School of Computing
Telkom University
Bandung, Indonesia
karimahsiti@telkomuniversity.ac.id

***Abstract*—This paper proposes a novel approach for optimizing online shopper purchase intent prediction using feature selection combined with Adaptive Synthetic Sampling (ADASYN). A supervised learning technique is applied to predict whether the customer visits ending with shopping or not based on the features. However, not all features are important to predict the classes. In addition, a suboptimal performance may occur due to the class imbalance problem. Therefore, we propose Information Gain and Correlation feature selection to select the most important features. ADASYN is additionally used to deal with the class imbalance problem by adaptively generating new synthetic samples of the minority class with considering density distribution. The proposed approach is run using Random Forest classifier. The results indicate that ADASYN effectively improves the classification performances in terms of accuracy, precision, recall, and F1-score. The use of feature selection combined with ADASYN has been compared to previous works, the results indicate that our proposed approach outperforms all. We additionally use a statistical test to show that our results are statistically significant. By these results, our proposed approach is promising in optimizing classification performances.**

***Index Terms*—ADASYN, class imbalance problem, feature selection, online shoppers' purchasing intention, Random Forest.**

I. Introduction

Recently, most business activities provide online shopping services due to the pandemic. Many e-commerce and companies set strategies to maintain engagement with customers including invest in advanced technology. One of the strategies is the application of machine learning using supervised learning methods. It can be implemented for online shopper purchase intent (OSPI) prediction whether the customer visits ending with shopping or not. Many studies on OSPI prediction have been done with different approaches. The use of some classifiers was applied to predict the online shopper behavior.

As in [1], different classification algorithms, as well as ensemble methods, were used to identify a suitable model. Decision Tree, Naïve Bayes, and SVM were run with accuracy, respectively, 85.9%, 84.17%, and 83%. In addition, ensemble methods, i.e. Random Forest, Stacking, Voting, Bagging, and Gradient Boosting were used with accuracy, respectively, 89.55%, 89.65%, 88.58%, 90.25%, and 90.34%. Naïve Bayes, MLP, SVM, Random Forest, and Decision Tree were additionally used in [2] with accuracy 80.88%, 88.56%, 88.07%, 90.13%, and 89.29%, respectively.

Several classifiers were additionally used in [3] for the OSPI dataset. KNN, Logistic Regression, LDA, and Stacking were implemented with accuracy 86%, 87%, 87%, 94%, respectively. However, previous works mentioned above worked with an imbalanced dataset and without feature selection. A high-accuracy result would be biased when the dataset is imbalanced. The use of feature selection is also important to increase the accuracy with more time-efficient [4].

A cat boost classification algorithm was applied in [5] to the imbalanced dataset to classify the actual purchase customers. The proposed algorithm was able to measure and sort the features during the model training process and select the most important features for the model. However, this work made class prediction from the perspective of an imbalanced dataset, thus the trained model of the minority class needs to be improved. The resulting accuracy was 88.51%.

The use of the oversampling technique was introduced in [6] to overcome the class imbalance problem on the OSPI dataset before continuing to the classifiers. This work adopted Synthetic Minority Oversampling Technique (SMOTE) as in [7] to generate new synthetic samples for the minority class. Several classifiers were used involving Naïve Bayes, C4.5, and Random Forest with accuracy, respectively, 86.66%, 86.59%, and 86.78%. However, the weakness of SMOTE is the overgeneralization of the minority class without considering the majority class, this causes the overlapping of the classes [8].

A real-time online shopper behavior analysis system was proposed by [9] that can simultaneously predict visitor's purchase intent and the possibility to leave the site. Oversampling technique was used to handle the class imbalance problem by selecting multiple samples of the minority class more than once. The filter-based feature selection involving Correlation, Mutual Information (MI), and mRMR was used before passing to the learning stage using C4.5, SVM, and MLP that produced accuracy 82.34%, 84.88%, and 87.24%, respectively. However,

the results were lower than the previous work mentioned above and need to be improved.

The public dataset used on OSPI prediction has been provided by [10]. However, the dataset has the class imbalance problem as the number of samples for a certain class is less than the other class. This problem leads to a suboptimal classification model as the minority class is frequently misclassified [11]. This study proposes ADASYN to deal with the class imbalance problem by adaptively generating new synthetic samples of the minority class with considering density distribution. ADASYN was first proposed by [12], which is an improvement of SMOTE. ADASYN was utilized in [8], [13], [14] in handling an imbalanced dataset for supervised learning tasks.

In this study, we apply ADASYN combined with filter-based feature selection using Information Gain and Correlation to optimize the classification performances. The proposed approach is evaluated using Random Forest classifier to measure the accuracy, precision, recall, and F1-score. We determine the optimal number of features on the OSPI dataset to avoid high dimensionality and provide the statistical test to validate our results.

II. Material and Method

A. Dataset Description

The online shoppers' purchasing intention dataset used in this study containing 12330 samples comprises 10 numerical and 7 categorical features with a label in binary class, i.e. negative (false) and positive (true). A negative class represents samples that did not end with shopping and a positive class represents samples ending with shopping. The detailed dataset descriptions are explained in [9] and available online at the UCI Machine Learning repository [10]. We run categorical encoding and min-max normalization in the data preprocessing stage.

As the dataset contains 10422 samples of negative class (85%) and 1908 samples of positive class (15%), this causes the class imbalance problem with the imbalanced ratio (IR) as follows

$$\text{IR} = \frac{\text{Number of Minority Class Sample}}{\text{Number of Majority Class Sample}} = \frac{1908}{10422} = 0.18 \tag{1}$$

where the range of IR from 0 to 1 indicating the class distribution ratio. If an IR value close to 0 indicates an imbalanced class, while an IR value close to 1 indicates a balanced class [15].

B. Feature Selection

In this study, we use Information Gain and Correlation feature selection methods to increase the classification performances by selecting the most important features.

1) Information Gain: Entropy is first calculated to measure the heterogeneity in a dataset as follows

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \tag{2}$$

where c is the number of classes, p_i is the proportion number of samples in class i over all samples S. Then, we obtain Information Gain as follows

$$Gain(S) = Entropy(S) - \sum_{n \in f} \frac{|S_n|}{|S|} \cdot Entropy(S_n) \tag{3}$$

where n is the possible value in feature f, $|S_n|$ is the number of samples for n, and $|S|$ is the total samples [16].

2) Correlation: This uses Pearson's correlation method to evaluate the features. Consider a and b are two features, the Pearson's correlation coefficient is derived by [4] as follows

$$r(a,b) = \frac{n\sum a_i b_i - \sum a_i \sum b_i}{\sqrt{n\sum a_i^2 - (\sum a_i)^2}\sqrt{n\sum b_i^2 - (\sum b_i)^2}} \tag{4}$$

where n is the size, a_i and b_i denote the i-th feature values. The Correlation coefficient range $r(a,b)$ from -1 to 1.

C. ADASYN

Since the imbalanced ratio of the dataset is close to 0 as in (1), this leads to the class imbalance problem and can be resolved using ADASYN. ADASYN is an oversampling approach by synthetically generating new samples based on interpolation from existing minority class samples. This aims to reduce the bias caused by imbalanced classes and adaptively learning. The detailed ADASYN algorithm for binary class classification is presented in [12]. The new samples are generated corresponding to k-nearest neighbors as follows

$$S_{\text{new}} = S_i + (S_k - S_i) \cdot \lambda \tag{5}$$

where S_{new} is the new synthetic sample, S_i is the original sample, S_k is the nearest neighbor sample, and λ is a random number in the range $[0, 1]$.

ADASYN algorithm has the main approach that works with density distribution as a measure to adaptively generate the number of new synthetic samples for the minority class, by changing the weights of different minority samples to balance the skewed class distribution [13]. The dataset generated by ADASYN does not only provides a balanced class distribution but will also enforce the classification algorithm to focus on learning the difficult samples. This is the advantage of ADASYN compared to SMOTE [7].

D. Random Forest

Random Forest is one of the ensemble supervised learning methods that comprise a set of individually base classifiers. The advantage of ensemble learning often provides more accurate results than any of the single classifiers [17]. Random Forest comprises a set of tree-based classifiers. Each tree depends on the values of random vectors that are independently sampled with the same distribution and use majority voting for class prediction.

The strength of Random Forest applies random feature selection to select each node, thus it provides a relatively low error rate. Random Forest only uses a subset of the features to train the model (usually 20% of the feature number), thus it will be more time-efficient for a large dataset with a more varied set of independent models [16].

E. Model Evaluation

The classification performances are measured using the confusion matrix as shown in Table. I representing True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) as follows:

- TP is a positive class that is predicted as positive.
- FP is a negative class that is predicted as positive.
- FN is a positive class that is predicted as negative.
- TN is a negative class that is predicted as negative.

We evaluate the classification performances using the confusion matrix with the following measures:

1) Accuracy: The percentage number of correct prediction.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{6}$$

2) Precision: The ratio of the true positive over all positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

3) Recall: The ratio of the true positive over all class predictions.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

4) F1-score: The harmonic value of precision and recall.

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

III. Result and Discussion

A. Evaluation of k Values for ADASYN

ADASYN generates new synthetic data samples based on the k-nearest neighbors as in (5), thus we evaluate synthetic data generation using different k-values to select the best one. In this study, k-values are odd numbers to avoid a draw as the class is binary, where $k = 1, 3, 5, 7, 9$ [18]. The evaluation of k values is given in Table. II, the results show that $k = 7$ provides the closest IR to 1. The closer the IR value to 1, the more balanced the class distribution.

TABLE I
The Confusion Matrix

Confusion Matrix		**Actual Class**	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

TABLE II
Evaluation of k-Nearest Neighbor Values for ADASYN

k	**Class Distribution**		**Imbalanced Ratio (IR)**
	Negative	**Positive**	
1	10422	9986	0.958
3	10422	10152	0.974
5	10422	10389	0.996
7	**10422**	**10449**	**0.997**
9	10422	10485	0.993

B. Evaluation of Feature Selection and ADASYN

In this study, we apply Information Gain and Correlation feature selection before processing to the classifier. The features are evaluated corresponding to the relevancy to the class prediction. The features are then ranked to select the most important features as shown in Table. III. We have the top-5 rank feature that is highly correlated with the class label in different values, i.e. *Page value, Exit rate, Bounce rate, Product related, Product related duration*.

We evaluate the effect of ADASYN in optimizing the classification performances by conducting simulation with several schemes of K-fold Cross-validation (CV) value, where $K = 3, 5, 7, 10$. The objective is to compare the behavior of the proposed approach when implemented using Random Forest with various options. The comparison of classification performances is given in Table. IV, in which the best results are marked in bold. The results show that Random Forest with ADASYN consistently outperforms Random Forest without ADASYN in terms of accuracy, precision, recall, and F1-score.

We additionally evaluate the use of Information Gain and Correlation feature selection combined with ADASYN using Random Forest associated with accuracy in the different number of features. The experiment is run using the 10-fold CV. Fig. 1 and Fig. 2 demonstrate that the classification accuracy with Information Gain and Correlation combined with ADASYN is higher than that of without ADASYN. The proposed approach is then validated with other methods in

TABLE III
Features Rank by Information Gain (IG) and Correlation (CORR)

Feature [9]	**IG Value**	**Rank**	**CORR Value**	**Rank**
Administrative	**0.023**	**6**	**0.139**	**6**
Administrative duration	**0.02**	**9**	0.094	9
Informational	0.008	10	0.095	8
Informational duration	0.008	11	0.07	12
Product related	**0.034**	**5**	**0.159**	**3**
Product related duration	**0.041**	**3**	**0.152**	**4**
Bounce rate	**0.034**	**4**	**0.151**	**5**
Exit rate	**0.057**	**2**	**0.207**	**2**
Page value	**0.237**	**1**	**0.493**	**1**
Special day	0.007	13	0.082	10
Operating Systems	0.005	14	0.015	15
Browser	0	16	0.024	14
Region	0	17	0.012	16
Traffic Type	**0.022**	**8**	0.005	17
Visitor Type	0.007	12	**0.103**	**7**
Weekend	0.001	15	0.029	13
Month	**0.022**	**7**	0.077	11

TABLE IV
COMPARISON OF CLASSIFICATION PERFORMANCES USING RANDOM FOREST AND ADASYN

K-fold Value	Random Forest				Random Forest with ADASYN			
	Accuracy (%)	Weighted Average			Accuracy (%)	Weighted Average		
		Precision	Recall	F1-score		Precision	Recall	F1-score
3	90.26	0.897	0.903	0.898	**92.792**	**0.928**	**0.928**	**0.928**
5	90.081	0.895	0.722	0.584	**93.185**	**0.932**	**0.932**	**0.932**
7	90.17	0.896	0.902	0.897	**93.271**	**0.933**	**0.933**	**0.933**
10	90.365	0.898	0.904	0.9	**93.271**	**0.933**	**0.933**	**0.933**

classifying OSPI dataset [10] as shown in Table. V. The results show that our proposed approach outperforms all previous works.

Furthermore, the features can be reduced based on the threshold value of feature selection to avoid high dimensionality. The threshold value for Information Gain is 0.01 [19] and Correlation is 0.1 [20]. Therefore, we mark the selected features in bold as shown in Table. III while remaining features can be discarded. In other words, we only select the top-9 features for Information Gain and top-7 features for Correlation feature selection. With less number of features, our proposed approach provides maximum accuracy of 93.34% that still outperforms all.

C. Statistical Test

To validate our results and avoid any biases, we provide statistical comparisons to evaluate whether our proposed approach tends to have values higher than compared technique. In this study, we use a nonparametric statistical test for the reasons [14]: can deal with normally or non-normally distributed data and is more reliable than the parametric test. As we have two independent groups, we apply The Mann-Whitney U to evaluate our proposed approach by comparing the accuracy in Table. IV, Fig. 1 and Fig. 2.

As in [21], the Mann-Whitney U statistics are mathematically expressed as follows

$$U_1 = n_1 n_2 + \left(\frac{n_1(n_1+1)}{2}\right) - r_1 \tag{10}$$

$$U_2 = n_1 n_2 + \left(\frac{n_2(n_2+1)}{2}\right) - r_2 \tag{11}$$

where n_1 is the number of samples in the first group, n_2 is the number of samples in the second group, r_1 is the ranking sum for the first group, and r_2 is the ranking sum for the second group. Furthermore, the p corresponding to U statistics and statistical threshold ($\alpha = 0.05$) are used to conclude whether the null hypothesis (H_0) is rejected or accepted by the following terms

if p of $min(U_1, U_2) < \alpha$ **then** Reject H_0 **else** Accept H_0

With the null hypothesis, the two groups should have similar value, it means that our proposed approach is equal to the other ones. On the contrary, the null hypothesis is rejected when the two groups are statistically different. The statistical comparison is given in Table. VI, the results show that our proposed approach is statistically significant.

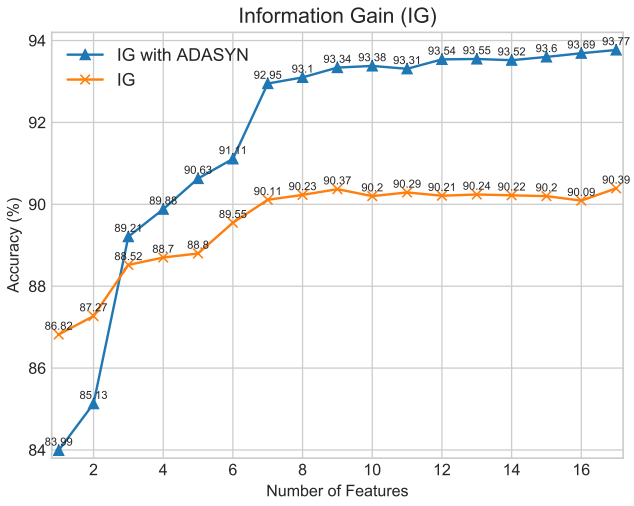

Fig. 1. The effect of Information Gain and ADASYN in different number of features using Random Forest.

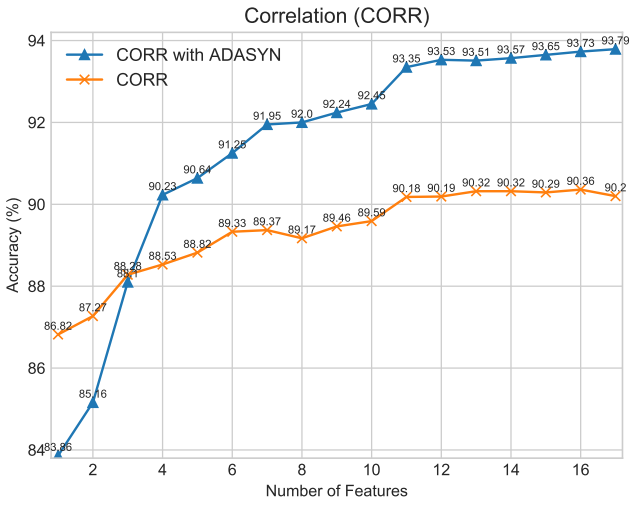

Fig. 2. The effect of Correlation and ADASYN in different number of features using Random Forest.

TABLE V
COMPARISON OF DIFFERENT TECHNIQUES ON OSPI DATASET TO VALIDATE OUR PROPOSED APPROACH

Reference	Number of Features	Number of Samples	Feature Selection	Oversampling	Max. Accuracy (%)
[1]	17	12330	-	-	90.34
[2]	17	12330	-	-	90.13
[5]	17	12330	-	-	88.50
[6]	17	12330	-	SMOTE	86.78
[9]	17	12330	Correlation, MI, mRMR	Random Oversampling	87.24
Proposed approach	17	12330	Information Gain, Correlation	ADASYN	**93.78**

IV. CONCLUSION

A new approach to optimize OSPI prediction has been applied using filter-based feature selection combined with ADASYN. This study has been working with an imbalanced dataset that leads to suboptimal results. ADASYN has been used to deal with the class imbalance problem by adaptively generating new synthetic samples of the minority class with considering density distribution. Information Gain and Correlation have been involved to assess the most important features. We have been using Random Forest classifier to evaluate the proposed approach. The results indicate that ADASYN has effectively increased the classification performances in terms of accuracy, precision, recall, and F1-score. The use of ADASYN together with feature selection has been compared to previous works with the same dataset, the results indicate that our proposed approach outperforms all. Validation has been additionally presented using a statistical test to show that our results are statistically significant. By these results, our proposed approach is promising in optimizing the performances of OSPI prediction.

TABLE VI
STATISTICAL COMPARISON TO VALIDATE OUR PROPOSED APPROACH

Group Sample	*p*-Value	Hypothesis
Table. IV	0.0147	H_0 rejected
Fig. 1	0.0013	H_0 rejected
Fig. 2	0.0008	H_0 rejected

REFERENCES

[1] M. R. Kabir, F. Bin Ashraf, and R. Ajwad, "Analysis of different predicting model for online shoppers purchase intention from empirical data," 2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019, pp. 18-20, 2019, doi: 10.1109/ICCIT48885.2019.9038521.

[2] Y. Christian, "Comparison of machine learning algorithms using weka and sci-kit learn in classifying online shopper intention," Journal of Informatics and Telecommunication Engineering, 3(1), pp. 58-66, 2019, doi: 10.31289/jite.v3i1.2599.

[3] S. Mootha, S. Sridhar, and M. S. K. Devi, "A stacking ensemble of multi-layer perceptrons to predict online shoppers purchasing intention," 2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020, pp. 721-726, 2020, doi: 10.1109/ISRITI51436.2020.9315447.

[4] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70-79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[5] X. Dou, "Online purchase behavior prediction and analysis using ensemble learning," 2020 IEEE 5th Int. Conf. Cloud Comput. Big Data Anal. ICCCBDA 2020, pp. 532-536, 2020, doi: 10.1109/ICCCBDA49378.2020.9095554.

[6] K. Baati and M. Mohsil, "Real-time prediction of online shoppers purchasing intention using random forest," vol. 583 IFIP. Springer International Publishing, pp. 43-51, 2020, https://doi.org/10.1007/978-3-030-49161-1_4.

[7] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Int. Res. 16, pp. 321-357, 2002.

[8] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," 2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017, vol. 2017-January, pp. 79-85, 2017, doi: 10.1109/ICACCI.2017.8125820.

[9] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Comput. Appl., vol. 31, no. 10, pp. 6893-6908, 2019, doi: 10.1007/s00521-018-3523-0.

[10] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, Online Shoppers Purchasing Intention Dataset, 2018, Retrieved from https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset, [accessed Feb 2, 2021].

[11] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with oversampling and undersampling techniques: Overview study and experimental results," 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, pp. 243-248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[12] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congr. Comput. Intell. (pp. 1322 1328), no. 3, pp. 1322-1328, 2008.

[13] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci. (Ny)., vol. 250, pp. 113-141, 2013, doi: 10.1016/j.ins.2013.07.007.

[14] A. Amin et al., "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," IEEE Access, vol. 4, no. MI, pp. 7940-7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[15] R. D. Prayogo and S. A. Karimah, "Optimization of Phishing Website Classification Based on Synthetic Minority Oversampling Technique and Feature Selection," 2020 International Workshop on Big Data and Information Security (IWBIS), 2020, pp. 121-126, doi: 10.1109/IWBIS50925.2020.9255562.

[16] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers, Elsevier, 2012.

[17] V. Y. Kulkarni and P. K. Sinha, "Random forest classifier: A survey and future research directions," Int. J. Adv. Comput., vol. 36, no. 1, pp. 1144-1156, 2013.

[18] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," SN Appl. Sci., vol. 1, no. 12, pp. 1-15, 2019, doi: 10.1007/s42452-019-1356-9.

[19] H. Sulistiani and A. Tjahyanto, "Comparative analysis of feature selection method to predict customer loyalty," IPTEK J. Eng., vol. 3, no. 1, pp. 1-5, 2017, doi: 10.12962/joe.v3i1.2257.

[20] E. M. Karabulut, S. A. Özel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technol., vol. 1, pp. 323-327, 2012, doi: 10.1016/j.protcy.2012.02.068.

[21] N. Nachar, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," Tutor. Quant. Methods Psychol., vol. 4, no. 1, pp. 1320, 2008, doi: 10.20982/tqmp.04.1.p013.