

Using machine learning approach towards successful crowdfunding prediction

Sarifah Putri Raflesia¹, Dinda Lestarini^{1,2}, Rizka Dhini Kurnia³, Dinna Yunika Hardiyanti^{1,4}

¹Department of Computerized Accounting, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

²Database and Big Data Laboratory, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

³Department of Information Management, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

⁴Electronic Data Processing and Decision Support System Laboratory, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

Article Info

Article history:

Received Nov 15, 2022

Revised Dec 12, 2022

Accepted Jan 19, 2023

Keywords:

Crowdfunding
Linear regression
Prediction
Random forest
XGBoost

ABSTRACT

Crowdfunding is a concept that emerged due to difficulties in raising funds for community business projects, social activities, micro-enterprises, and start-ups conventionally. Crowdfunding uses internet technology as a bridge between the donor and the recipient of funds so that it can reach a wider range of donors. This study aims to compare the performance of machine learning approaches in predicting crowdfunding campaign success. Three machine learning algorithms were employed to predict crowdfunding campaign success, namely logistic regression, random forest, and extreme gradient boosting (XGBoost). The dataset used in this study contains data about all projects posted on Kickstarter from January 2020 to September 2022. To improve the prediction model's performance, experiments using principal component analysis (PCA) feature reduction and log transformation were conducted. The results show that the implementation of log transformation on the dataset can increase the prediction model's performance. Meanwhile, XGBoost algorithm performs better than linear regression and random forest.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Dinda Lestarini

Department of Computerized Accounting, Faculty of Computer Science, Universitas Sriwijaya

30862 Indralaya, Palembang, Sumatera Selatan, Indonesia

Email: dinda@unsri.ac.id

1. INTRODUCTION

The use of information technology to assist human activities is currently very massive. One of technology that is used massively and connecting human is internet. The emergence of the internet as one of the impacts of technological developments makes people can easily and quickly find whatever they are looking for. Starting from social media, shopping, studying, and discussing, to the internet as a medium to reap benefits as digital influencers. The use of internet technology has provided an understanding for some groups that the internet can be maximally utilized to get people's attention quickly and a lot. Not only done locally but also worldwide [1]–[3]. In the development of the industrial revolution 4.0, more innovations have been carried out in various fields including financial technology such as mobile payments, crowdfunding, peer-to-peer (P2P) lending, insurance, and wealth management [4]. In this research, the technology product that will be discussed is crowdfunding.

Starting from the difficulty of raising funds for community business projects, social activities, micro-enterprises, and start-ups, in recent years a platform called crowdfunding has emerged [5], [6]. Crowdfunding is an alternative form of traditional funding. Crowdfunding can be used to fund various

projects and is open to any individual or group. The principle of crowdfunding is the same as conventional funding, but what makes the difference is the use of the internet as a bridge between the donor and the recipient of funds [7], [8]. Using the internet, fundraisers can easily go viral, allowing large amounts of money to be raised. In addition, the internet allows information about projects to be funded to spread quickly. Crowdfunding has four different models [9]–[11] that can make it easier for funders to make the best fund placement decisions, namely i) donation-based crowdfunding that does not give anything in return for any contributions from the donor, ii) award-based crowdfunding that rewards in the form of rewards or various things that are not in the form of money or share ownership, such as clothes or merchandise, iii) equity-based crowdfunding which rewards funders in the form of share ownership, and iv) lending-based crowdfunding which provides reward in the form of interest in the loan provided by the funder.

There are several crowdfunding platforms like Kickstarter, Indiegogo, and GoFundMe. The data that is used in this research gathered from Kickstarter. Kickstarter is an American public benefit corporation based in Brooklyn, New York, that maintains a global crowdfunding platform focused on creativity. The company's stated mission is to "help bring creative projects to life". As of September 2022, Kickstarter has received \$6.8 billion in pledges from 21 million backers to successfully fund 227,158 projects, such as films, music, stage shows, comics, journalism, video games, technology, publishing, and food-related projects. This research aims to predict the success of projects funded through Kickstarter using machine learning approach between 2020–2022. Furthermore, the article describes the method used in section 2, the results are explained in section 3, and the conclusion is in section 4.

2. METHOD

Crowdfunding is also known as crowd financing, equity crowdfunding, or crowdsource fundraising. Crowdfunding can be defined as funding by a group of people [12] or a form of initiative from individuals/teams/organizations/entities to raise funds to realize a project. It has a characteristic that is collecting funds from very small to moderate amounts for an interest that is usually able to attract the attention of many people. The existence of crowdfunding occurs after crowdsourcing. Both of them use social media and the internet as intermediaries to the wider community, but there is a clear difference between crowdfunding and crowdsourcing. The difference between crowdsourcing and crowdfunding is related to the role of donors in a project. Crowdsourcing expects donors to be more involved in the projects they are helping [13], by providing feedback in the form of ideas and suggestions for the sustainability of the project. Meanwhile, crowdfunding only uses donors to raise funds for the implementation of a project. Crowdfunding is defined as the process of taking a project or business, in need of investment, and asking a large group of people to supply this investment [14] using internet [15], [16]. Based on what investors receive after contributing, crowdfunding platforms are categorized into four types: donation-based, lending, reward-based, and equity [17].

Figure 1 illustrates the complete process for comparative analysis and prediction of crowdfunding campaign success. The first phase is data collection. The dataset used in this study was obtained from a scrapping website called Web Robots. The dataset contains data about all projects posted on Kickstarter from April 2009 to September 2022. In this study, the dataset used was from January 2020–September 2022. Data pre-processing is carried out on the data that has been collected. The total number of projects after data aggregation is almost close to 2 million projects with 1.7 million of them having the same project id. After data preprocessing stages, 51,513 projects meet the criteria. This data will be split into training and testing data. The distribution of training and testing data is 70:30%. Several machine learning algorithms are used to classify whether the project is successful or not based on the available features.

Machine learning is an approach technique from artificial intelligent (AI) which is used to imitate to replace the role of humans in carrying out activities to solve problems [18], [19]. In short, machine learning is a machine that is made so that it can learn and do work without direction from its users. Machine learning is a branch of science that studies how to give computers the ability to learn without being explicitly programmed. Machine learning can do this if it is based on ideas obtained from previous data and identifies patterns and makes decisions using minimal human or user intervention. Machine learning algorithms are divided into three types, namely supervised learning, unsupervised learning, and reinforcement learning [20]. Supervised learning is an approach to AI creation. It is called "supervised" because, in this approach, machine learning is trained to recognize patterns between input data and output labels. The label means the tag of the data added to the machine learning model. For example, images of cats tagged with "cat" in each image of a cat and images of dogs tagged with "dog" in each image of a dog. Machine learning categories can be classification ("dog", "cat", and "bear") and regression (weight and height). Supervised learning is widely used in predicting patterns where there is already a complete sample of data, so the pattern formed is the result of learning the complete data. It is certain that if we enter new data after we do extract transform load (ETL) then we will get the feature information from the new sample. Then the features are compared with the

classification pattern of the model obtained from the labeled data. Each label will be compared to completion, and the one with the higher percentage will be taken as the final prediction. Also, machine learning is trained to identify the underlying relationships of input data connections with output labels.

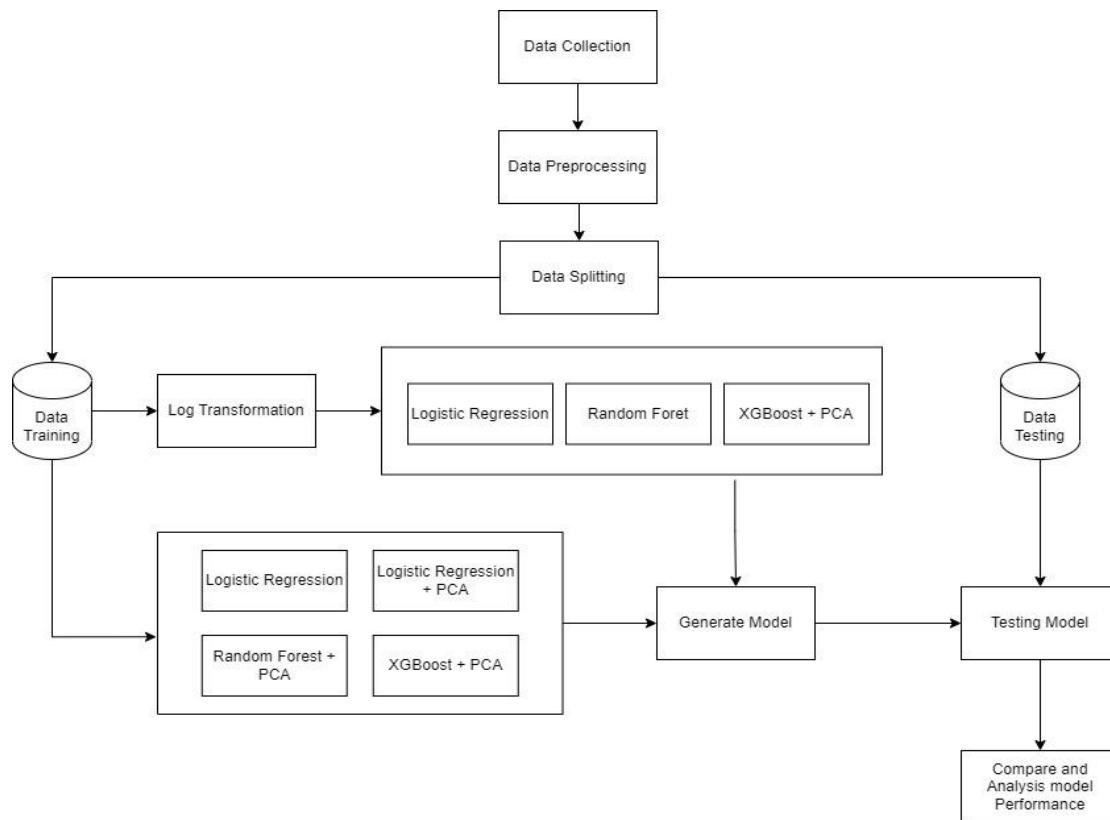


Figure 1. Research method

The unsupervised learning algorithm is an algorithm that does not require labeled data. It is used in pattern detection and descriptive modeling that does not require categories or labeled outputs on which the algorithm is based to find the right model [21], [22]. Also, this algorithm is used for clustering and association rules. The advantage of unsupervised learning is that because it does not require labels, the algorithm is more flexible to look for patterns that may not have been previously known. While the drawbacks of this algorithm are the difficulty of finding information in the data because there are no labels and it is more difficult to compare the output with the input.

The goal of reinforcement learning is to use observations gathered from interactions with the environment to take actions that will maximize output and minimize risk. This algorithm will continue to learn repeatedly. In this algorithm, there are agents who will learn from interactions with their environment. To generate a model, the reinforcement learning algorithm goes through several stages, including the agent observing the input data, after which the agent takes an action to decide. After the decision is made, the agent will receive a "reward" or reinforcement from the environment. Then, re-observe the input, and the decision-making process is carried out again but with additional reinforcement from the environment so that the results of the decisions taken are more accurate.

In this study, logistic regression, random forest, and extreme gradient boosting (XGBoost) algorithm were used. Logistic regression is a statistical method that is applied to model categorical response variables (nominal/ordinal scale) based on one or more predictor modifiers which can be either categorical or continuous variables (interval or ratio scale). If the response modifier consists of only two categories, the logistic regression method that can be used is binary logistic regression. Logistic regression is a part of regression analysis that can be used if the dependent variable (response) is dichotomous. Dichotomous variables usually consist of only two values, which represent the occurrence or absence of an event which is usually assigned a number 0 or 1. Unlike ordinary linear regression, logistic regression does not assume a linear relationship between the independent and dependent variables. Regression logistics is a non-linear

regression where the specified model will follow a linear curve pattern. Logistic regression will form predictor/response variables which are linear combinations of independent variables.

Random forest is the development of the classification and regression tree (CART) method by applying the bootstrap aggregating (Bagging) method and the random feature selection [23], [24]. Bagging is a method that can improve the results of the classification algorithm. Baggage is a method based on the ensemble method, which is a method that uses a combination of several models. Predictor bagging is a method used to generate multiple versions of predictors and use them to obtain a set of predictors. Multiple versions are formed by bootstrap replication of experimental data. The random forest has many trees that can reach hundreds, and each tree is planted the same way. Several learning functions generated by random forest are used as an ensemble bagging strategy to overcome the problem of overfitting when faced with small train data. Many algorithms can be used in the formation of decision trees such as ID3, CART, and C4.5. The random forest method itself has several advantages, among others, produces good classification results, produces lower errors, and can efficiently handle training data with very large amounts of data [25], [26].

XGBoost is an effective technique in machine learning for regression analysis and classification [27]–[29] based on the gradient boosting decision tree (GBDT) [30], [31]. Firstly, the main concept was introduced by [32], in his research Friedman connected boosting and optimization in building a gradient boosting machine (GBM). Subsequently, a new model was developed to predict the error from the previous model used in the boosting method. The addition of new models is carried out until no more errors are found. By using gradient descent to minimize errors when creating a new model, this algorithm is called gradient enhancement. XGBoost has many advantages, including being able to perform parallel processing which can speed up computations [33], having high flexibility in setting objectives [34], built-in cross validation, having regularization features, and overcoming splits during negative loss [35]. Therefore, XGBoost is very suitable for processing classification data. XGBoost will create a tree as a way to classify train data so that a specific target can be obtained. XGBoost has several parameters that we can set so that it is adjusted to the dataset obtained. Parameter tuning in XGBoost can be done using GridSearch CV and tuning manually.

3. RESULTS AND DISCUSSION

In this section, the comparison of model performance using several machine learning algorithms is described. The confusion matrix is presented as a basis for comparing model performance. The performance indicators used are accuracy, sensitivity, precision, and F1-score. Table 1 shows comparison of crowdfunding campaign success prediction using various algorithms.

Table 1. Crowdfunding campaign success prediction performance

Algorithm	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-score (%)
Logistic regression	84	52	74	61
Logistic regression + PCA	84	52	74	61
Random forest + PCA	82	41	73	53
XGBoost + PCA	83	49	70	58
logistic regression with log-transformed data	86	58	76	66
Random forest with log-transformed data	86	54	82	65
XGBoost with log-transformed data	86	59	77	67

Logistic regression can be used as a classifier for data with binary labels. In this study, the labels used are successful projects and unsuccessful projects. The testing process using the default parameters provides an accuracy of 84% with sensitivity, precision, and F1-score of 52%, 74%, and 61%, respectively. From Figure 2(a), it can be seen that the number of false negatives is quite large with 1776 data.

There are 106 features that can be used to predict the success of crowdfunding. Principal component analysis (PCA) method can be used to reduce features while still describing variations in the data. Table 2 describes the variance and accuracy using various numbers of feature. The desired data variation in this study is 99% with the number of PCA components of 90 features. The logistic regression model can be further improved by optimizing its parameters. The GridSearch method can be used to test several different regularization parameters (C values), penalties (l1 or l2), and models with or without intercepts. From the testing results, the optimal parameter for the logistic regression method is obtained with regularization of 10 and using the intercept and penalty l2. As can be seen from Table 1, this setting resulted in the same value of accuracy, sensitivity, precision, and F1-score with the prediction using logistic regression, but it succeeded in reducing false positive. But on the other hand, the false negative value also increased by 20 (as can be seen in Figure 2(b)).

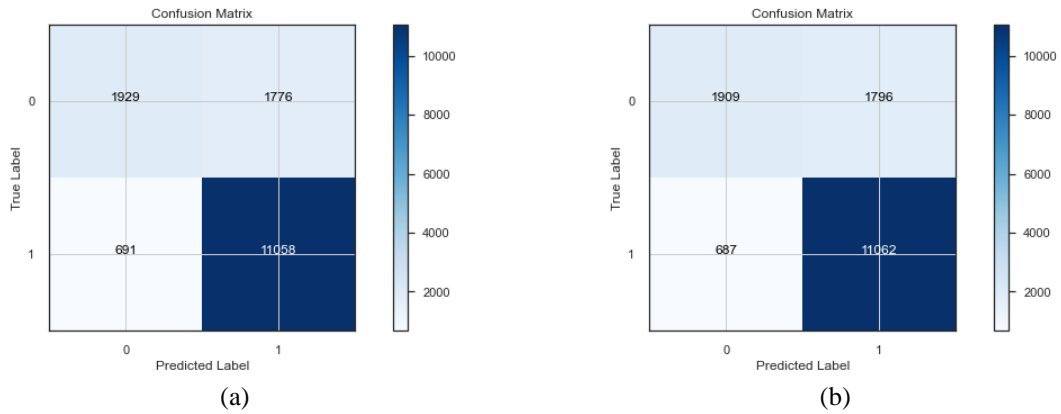


Figure 2. Confusion matrix of prediction model using (a) logistic regression and (b) logistic regression+PCA

Table 2. Data variance and accuracy using various numbers of feature

Numbers of feature	Variance (%)	Accuracy (%)
58	80	82.257
70	90	82.762
90	99	83.823

The next algorithm used is the combination of random forest and PCA. Random forest parameter used is a depth of 30, with several trees of 100. Based on the data in Table 1, the combination of random forest and PCA didn't generate better results than the previous algorithm. From the confusion matrix (Figure 3(a)), the test results show that the combination of random forest and PCA succeeded in reducing false positive, but the false negative value also increased significantly.

XGBoost is a form of gradient enhancement algorithm. Similar to the random forest algorithm, this algorithm is an ensemble method that generates several decision trees to improve the performance of the classification model, but the XGBoost method uses gradient descent to improve model performance on data that is very difficult to classify. The combination of XGBoost and PCA methods provides 83% accuracy, 49% sensitivity, 70% precision, and 58% F1-score. As can be seen in Figure 3(b), the number of false negatives and false positives increases significantly when compared to the results of the logistic regression.

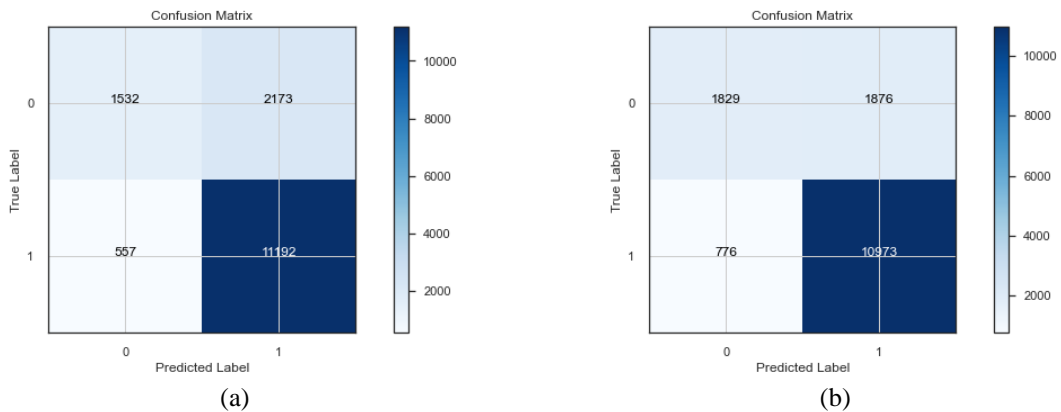


Figure 3. Confusion matrix of prediction model using (a) RF +PCA and (b) XGBoost+PCA

Furthermore, the logistic regression, random forest, and XGBoost algorithm were tested using log-transformed data, but the PCA feature reduction was not employed. This scenario shows better results on prediction model, as can be seen in Table 1. In addition, the overfitting phenomenon is overcome by implementing this scenario because there is a decrease of false positive and false negatives in the confusion matrix. As shown in Figure 4(a), the implementation of log-transformed data in logistic regression succeeded

in reducing false positive and false negative values by 20 and 226 when compared to vanilla logistic regression results. Meanwhile, the implementation of log-transformed data in random forest algorithm reduced false positive and false negative values significantly by 127 and 459 when compared to random forest algorithm and PCA results (as can be seen in Figure 4(b)). The XGBoost model with logarithmic data transformation becomes the best model compared to other models with an F1-score value of 67%. As shown in Figure 4(c), the implementation of log-transformed data succeeded in reducing false positive and false negative values by 135 and 362.

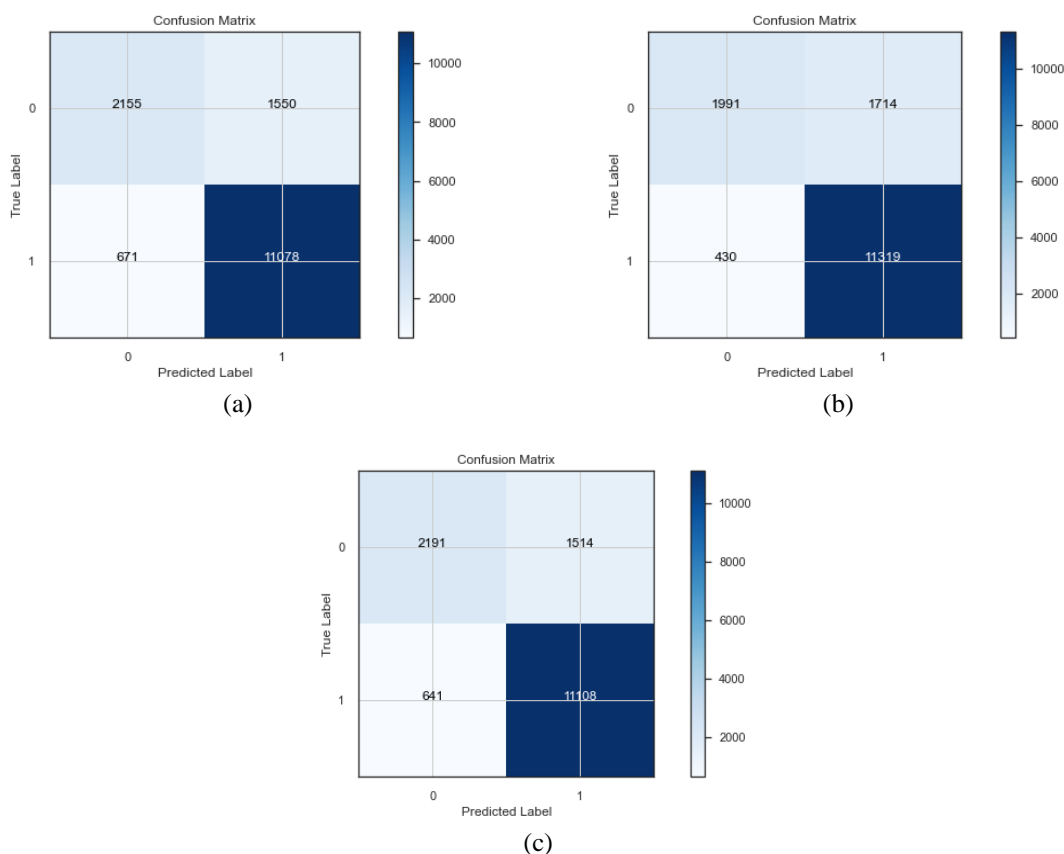


Figure 4. Confusion matrix of prediction model using (a) logistic regression, (b) random forest, and (c) XGBoost with log transformed data

4. CONCLUSION

This paper presents comparative analysis of machine learning performance in predicting crowdfunding campaign success. Three machine learning algorithms were employed to predict crowdfunding campaign success, namely logistic regression, random forest, and XGBoost. The experiments using PCA feature reduction and log transformation were also conducted to improve the performance of prediction model. Experimental results show that XGBoost algorithm has the best performance among the others. Moreover, the use of log-transformed data increased prediction model performance.

ACKNOWLEDGEMENTS

This work is funded by Universitas Sriwijaya through research grant program. The authors also acknowledge the support from Database and Big Data Laboratory, Faculty of Computer Science, Universitas Sriwijaya for providing the necessary resources to carry out this research.

REFERENCES

- [1] R. Sedgwick, S. Epstein, R. Dutta, and D. Ougrin, "Social media, internet use and suicide attempts in adolescents," *Current Opinion in Psychiatry*, vol. 32, no. 6, pp. 534–541, 2019, doi: 10.1097/YCO.0000000000000547.
- [2] F. Liu, E. T. K. Lim, H. Li, C. W. Tan, and D. Cyr, "Disentangling utilitarian and hedonic consumption behavior in online





- shopping: an expectation disconfirmation perspective,” *Information and Management*, vol. 57, no. 3, pp. 1–53, 2020, doi: 10.1016/j.im.2019.103199.
- [3] B. C. Y. Lo, R. N. M. Lai, T. K. Ng, and H. Wang, “Worry and permissive parenting in association with the development of internet addiction in children,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 21, pp. 1–12, 2020, doi: 10.3390/ijerph17217722.
- [4] Q. Zhao, P. H. Tsai, and J. L. Wang, “Improving financial service innovation strategies for enhancing China’s banking industry competitive advantage during the fintech revolution: a hybrid MCDM model,” *Sustainability*, vol. 11, no. 5, pp. 1–29, 2019, doi: 10.3390/su11051419.
- [5] J. Paschen, “Choose wisely: crowdfunding through the stages of the startup life cycle,” *Business Horizons*, vol. 60, no. 2, pp. 179–188, 2017, doi: 10.1016/j.bushor.2016.11.003.
- [6] G. K. C. Ahlers, D. Cumming, C. Günther, and D. Schweizer, “Signaling in equity crowdfunding,” *Entrepreneurship: Theory and Practice*, vol. 39, no. 4, pp. 955–980, 2015, doi: 10.1111/etap.12157.
- [7] M. J. Kim and C. M. Hall, “What drives visitor economy crowdfunding? The effect of digital storytelling on unified theory of acceptance and use of technology,” *Tourism Management Perspectives*, vol. 34, p. 100638, 2020, doi: 10.1016/j.tmp.2020.100638.
- [8] E. Battisti, F. Creta, and N. Miglietta, “Equity crowdfunding and regulation: implications for the real estate sector in Italy,” *Journal of Financial Regulation and Compliance*, vol. 28, no. 3, pp. 353–368, 2020, doi: 10.1108/JFRC-08-2018-0109.
- [9] G. Dushnitsky, M. Guerini, E. Piva, and C. R. -Lamastra, “Crowdfunding in Europe: determinants of platform creation across countries,” *California Management Review*, vol. 58, no. 2, pp. 44–71, 2016.
- [10] G. Yacoub, P. Mitra, T. Ratinho, and F. Fatalot, “Sustainable entrepreneurs: what drives them to engage in different crowdfunding types?,” *International Journal of Entrepreneurial Behaviour and Research*, vol. 28, no. 4, pp. 980–1000, 2022, doi: 10.1108/IJEBR-05-2021-0321.
- [11] H. Zhao *et al.*, “Voice of charity: prospecting the donation recurrence donor retention in crowdfunding,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1652–1665, 2020, doi: 10.1109/TKDE.2019.2906199.
- [12] Y. R. -Ricardo, M. Sicilia, and M. López, “What drives crowdfunding participation? The influence of personal and social traits,” *Spanish Journal of Marketing - ESIC*, vol. 22, no. 2, pp. 163–182, 2018, doi: 10.1108/SJME-03-2018-004.
- [13] T. Muliawati and F. Masya, “Fund raising and donation application system,” *International Research Journal of Computer Science*, vol. 6, no. 6, pp. 639–653, 2019.
- [14] H. Forbes and D. Schaefer, “Guidelines for successful crowdfunding,” *Procedia CIRP*, vol. 60, pp. 398–403, 2017, doi: 10.1016/j.procir.2017.02.021.
- [15] E. Mollick, “The dynamics of crowdfunding: an exploratory study,” *Journal of Business Venturing*, vol. 29, no. 1, pp. 1–16, 2014, doi: 10.1016/j.jbusvent.2013.06.005.
- [16] V. Chandna, “Social entrepreneurship and digital platforms: crowdfunding in the sharing-economy era,” *Business Horizons*, vol. 65, no. 1, pp. 21–31, 2022, doi: 10.1016/j.bushor.2021.09.005.
- [17] P. Belleflamme, T. Lambert, and A. Schwienbacher, “Crowdfunding: tapping the right crowd,” *Journal of Business Venturing*, vol. 29, no. 5, pp. 585–609, 2014, doi: 10.1016/j.jbusvent.2013.07.003.
- [18] H. Salehi and R. Burgueño, “Emerging artificial intelligence methods in structural engineering,” *Engineering Structures*, vol. 171, pp. 170–189, 2018, doi: 10.1016/j.engstruct.2018.05.084.
- [19] S. Dilek, H. Cakır, and M. Aydın, “Applications of artificial intelligence techniques to combating cyber crimes: a review,” *International Journal of Artificial Intelligence & Applications*, vol. 6, no. 1, pp. 21–39, 2015, doi: 10.5121/ijai.2015.6102.
- [20] I. Muhammad and Z. Yan, “Supervised machine learning approaches: a survey,” *ICTACT Journal on Soft Computing*, vol. 5, no. 3, pp. 946–952, 2015, doi: 10.21917/ijsc.2015.0133.
- [21] J. Yang and J. Kim, “Accident diagnosis algorithm with untrained accident identification during power-increasing operation,” *Reliability Engineering and System Safety*, vol. 202, p. 107032, 2020, doi: 10.1016/j.res.2020.107032.
- [22] J. Zurn, W. Burgard, and A. Valada, “Self-supervised visual terrain classification from unsupervised acoustic feature learning,” *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 466–481, 2021, doi: 10.1109/TRO.2020.3031214.
- [23] R. Yao, J. Li, M. Hui, L. Bai, and Q. Wu, “Feature selection based on random forest for partial discharges characteristic set,” *IEEE Access*, vol. 8, pp. 159151–159161, 2020, doi: 10.1109/ACCESS.2020.3019377.
- [24] M. Z. -Kermani, D. Stephan, M. Barjenbruch, and R. Hinkelmann, “Ensemble data mining modeling in corrosion of concrete sewer: a comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models,” *Advanced Engineering Informatics*, vol. 43, p. 101030, 2020, doi: 10.1016/j.aei.2019.101030.
- [25] V. Y. Kulkarni and P. K. Sinha, “Pruning of random forest classifiers: a survey and future directions,” in *2012 International Conference on Data Science & Engineering (ICDSE)*, 2012, pp. 64–68, doi: 10.1109/ICDSE.2012.6282329.
- [26] S. A. Naghibi, H. R. Pourghasemi, and B. Dixon, “GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran,” *Environmental Monitoring and Assessment*, vol. 188, no. 1, pp. 1–27, 2016, doi: 10.1007/s10661-015-5049-6.
- [27] Z. Ding, H. Nguyen, X. N. Bui, J. Zhou, and H. Moayedi, “Computational intelligence model for estimating intensity of blast-induced ground vibration in a mine based on imperialist competitive and extreme gradient boosting algorithms,” *Natural Resources Research*, vol. 29, no. 2, pp. 751–769, 2020, doi: 10.1007/s11053-019-09548-8.
- [28] H. Nguyen, C. Drebenstedt, X. N. Bui, and D. T. Bui, “Prediction of blast-induced ground vibration in an open-pit mine by a novel hybrid model based on clustering and artificial neural network,” *Natural Resources Research*, vol. 29, no. 2, pp. 691–709, 2020, doi: 10.1007/s11053-019-09470-z.
- [29] H. Xu, J. Zhou, P. G. Asteris, D. J. Armaghani, and M. M. Tahir, “Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate,” *Applied Sciences (Switzerland)*, vol. 9, no. 18, pp. 1–19, 2019, doi: 10.3390/app9183715.
- [30] Y. C. Chang, K. H. Chang, and G. J. Wu, “Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions,” *Applied Soft Computing Journal*, vol. 73, pp. 914–920, 2018, doi: 10.1016/j.asoc.2018.09.029.
- [31] R. Cai, S. Xie, B. Wang, R. Yang, D. Xu, and Y. He, “Wind speed forecasting based on extreme gradient boosting,” *IEEE Access*, vol. 8, pp. 175063–175069, 2020, doi: 10.1109/ACCESS.2020.3025967.
- [32] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2000.
- [33] J. Henriques, F. Caldeira, T. Cruz, and P. Simões, “Combining k-means and xgboost models for anomaly detection using log datasets,” *Electronics (Switzerland)*, vol. 9, no. 7, pp. 1–16, 2020, doi: 10.3390/electronics9071164.
- [34] S. S. Dhaliwal, A. A. Nahid, and R. Abbas, “Effective intrusion detection system using XGBoost,” *Information*, vol. 9, no. 7, pp.

1–24, 2018, doi: 10.3390/info9070149.





- [35] C. V. Priscilla and D. P. Prabha, "Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 1309–1315, doi: 10.1109/ICSSIT48917.2020.9214206.

BIOGRAPHIES OF AUTHORS







Sarifah Putri Raflesia     is well-experienced in system development, business process re-engineering, and service excellence field. She graduated from School of Electrical and Informatics (STEI), Institut Teknologi Bandung (ITB), She actively joint research project in ITB from 2014-2016. She also worked as research and information system laboratory assistant in ITB. Now, she is an active researcher and lecturer in Faculty of Computer Science, Universitas Sriwijaya since 2016. Sarifah also has already published many research works in reputable journals and also international conferences. Now she has Scopus H-index=5, Google Scholar H-index=7, and WoS H-index=2. She can be contacted at email: sarifah@unsri.ac.id.







Dinda Lestari     is an active lecturer in Faculty of Computer Science, Universitas Sriwijaya. She received master's degree in informatics from School of Electrical and Informatics, Institut Teknologi Bandung (ITB). Her research interest includes knowledge management, business process management, and information technology service management, and data science. She has already produced a number of research papers in reputable journal and conference. She has Scopus H-index of 4, WoS H-index of 2 and Google Scholar H-index of 6. Currently, she serves as the head of database and big data laboratory in Faculty of Computer Science, Universitas Sriwijaya. She can be contacted at email: dinda@unsri.ac.id.



Rizka Dhini Kurnia     is a lecturer in Faculty of Computer Science, Universitas Sriwijaya since end of 2009. Earned master's degree from Universiti Teknologi Malaysia (UTM) majoring in IT Management. She has already published many research papers in reputable journals dan international conferences. Her research interest includes IT project management, IT service management, and business process improvement. She can be contacted at email: rizkadhini@gmail.com.



Dinna Yunika Hardiyanti     received Bachelor degree from Faculty of Computer Science, Universitas Sriwijaya, and master degree from School of Electrical and Informatics (STEI) Institut Teknologi Bandung, Indonesia. She is an active lecturer in Computer Science Faculty, Universitas Sriwijaya. Currently, she is the head of Electronic Data Processing and Decision Support System Laboratory. Her research areas are data analytical, decision support system, and information system. She can be contacted at email: dinna.yunika@gmail.com.