International Journal of Information Science and Management
Vol. 20, No. 2, 2022, 55-65
https://dorl.net/dor/20.1001.1.20088302.2022.20.2.4.9

Original Research

Features Analysis of the Research and Development Industry in Indonesia

Endang Febrian Khusnul Hidayati
Analyst at Indonesian National Research and Innovation Agency, Indonesia.
efkhusnulh@gmail.com
ORCID iD: https://orcid.org/0000-0002-7289-5419

Bagus Sartono
Lecturer in Department Statistics and Data Science, IPB University, Indonesia.
bagusco@apps.ipb.ac.id
ORCID iD: https://orcid.org/0000-0003-1115-4737

Agus Mohamad Soleh
Lecturer in Department Statistics and Data Science, IPB University, Indonesia.
agusms@apps.ipb.ac.id
ORCID iD: https://orcid.org/0000-0002

Received: 04 April 2021
Accepted: 21 June 2021

Abstract

R&D is one of the key drivers of technological progress and contributes to increased productivity and profit growth. Indonesian percentage of Gross Domestic Expenditure on R & R&D (GERD) to GDP in 2018 is one of the Global Competitiveness Index indicators, only reaches 0.28% and is dominated by the government sector, while the industrial sector is only 7.34%. One of the reasons for this small value is that the data collection of R&D on the business sector in Indonesia has not been carried out optimally. A classification model is needed to determine the data collection target so that the results are more optimal. The main objective of this study is to classify R&D industries actors in Indonesia using XGBoost and then analyze the features for R&D industries actors using SHAP. XGBoost is one of the black-box models that is difficult to interpret, and SHAP is one of the interpretation methods. The classification results using XGBoost obtained the accuracy, AUC, and F1-Score values of 79.61%, 0.7646, and 84.44%, respectively. Based on the Shapley value of the SHAP method, it was found that the average growth in R&D expenditure had the highest contribution. The feature's contribution to the estimation will be even higher if the mean of R&D expenditure growth is higher (more than 0). The other one is the ratio of researchers to R&D human resources. If the ratio is more than 75%, it will negatively contribute. Finally, exports and State-Owned Enterprise (BUMN) feature with the smallest contribution.

Keywords: Research and Development (R&D), Industry, XGBoost, SHAP, Feature Analysis.

Introduction

Research and development (R&D) is one of the key drivers of technological progress and contributes to increased productivity and profit growth (Barreto & Kypreos, 2004). Meanwhile, R&D expenditure positively impacts economic growth in developing countries, especially the upper-middle-income (Inekwe, 2015). R&D expenditure is also one of the

indicators of the Global Competitiveness Index, which is the percentage of R&D spending on GDP or Gross Expenditure on Research and Development (GERD). Spending or investment in R&D is the most critical determinant in boosting scientific and technological progress (Asmara, Achelia, Simamora & Sartono, 2019). GERD refers to R&D expenditures to GDP originating from a country's industry public and private research institutions (Cardozo, Luzuriaga, Miranda, Lopez, Pajuelo & Japura, 2021). GERD indicates how much a country pays attention to R&D. One component of the GERD is the percentage of R&D spending on GDP in the industrial sector or Business Expenditure on Research and Development (BERD).

The percentage of R&D spending on GDP in the business sector (BERD) determines how much attention the business sector has in conducting R&D in a country. The BERD value in Indonesia in 2018 was 7.34%. This value is considered far from developed countries such as South Korea, which is 78.2% (OECD, 2020). One of the reasons for this happening is that the collection of R&D data in the business sector in Indonesia has not been performed optimally. Since 2016, the government has attempted to collect R&D data through the National Research and Innovation Agency (BRIN). The number of large industries (manufacturing) in Indonesia can reach 10,289 industries (BPS, 2020). Meanwhile, every year the target for collecting R&D data in the industrial sector is only 500-600 industries. Based on existing data, there are two types of industries that conduct R&D, namely industries that carry out R&D regularly and not regularly every year. Therefore, it is necessary to create a classification model for the R&D industry to obtain data on industries that regularly conduct R&D as targets for data collection. Industry classifications that perform R&D regularly and not regularly every year can be used as material for policymaking by the government. The government will consider industries that do not regularly provide R&D incentives. The most important thing for making policy that must be considered are the factors that influence the industry to carry out R&D regularly and not regularly every year. The main objective of this study is to classify R&D industries actors in Indonesia using XGBoost and then analyze the features for R&D industries actors using SHAP.

The classification method that can be used to create a classification model for the two types of industry is the ensemble tree. Several studies conclude that, in general, classification and prediction with ensembles can produce higher accuracy and stability than the use of a decision tree individually. The prediction of compressive strength of fly ash-based concrete using individual and ensemble algorithm produce that an ensemble tree gives a robust performance compared to a decision tree individually (Ahmad et al., 2021). One of the ensemble tree-based methods is XGBoost (Chen & Guestrin, 2016). XGBoost is a popular machine learning used because of its performance, which is better than other classification tree-based methods due to efficient computing and more accurate classification results (Zhang, Li, Wu, Li, Liu & Liu, 2020). A study on the vulnerability of debris flow in China concluded that XGBoost was better than a single classification tree (Zhang, Ge, Tian & Liou, 2019). Compared to Random Forest and Deep learning methods, XGBoost achieved the best performance with a very low cost of time (Joharestani, Cao, Ni, Bashir & Talebiesfandarani, 2019). However, the ensemble tree is a black-box model, where the resulting model cannot be known how the mechanism is in it (Molnar, 2019).

Several studies used Shapey Additive Explanations (SHAP) to explain the black-box model that is often difficult to interpret (Parsa, Movahedi, Taghipour, Derrible & Mohammadian, 2020). The SHAP algorithm is powerful, and this new framework also

provides desirable interpretations of the model performance and highlights the most important features for identifying m7G sites (Bi, Xiang, Ge, Li, Jia & Song, 2020). SHAP produces the conclusions derived from the proposed framework to provide important scientific information for government policymakers in disease control strategies (Kristjanpoller et al., 2021). Shapley initially proposed SHAP in 1953, based on game theory (Shapley, 2016). SHAP is a method based on Shapley's value to determine each player's contribution fairly. SHAP can show the contribution of each explanatory variable or feature to the estimated value of a model (Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014).

Materials and Methods

Data

The data used in this research are industrial data that conducted R&D activities from 2015 to 2018. The data are secondary data obtained from the Data Center and Information Technology of the Ministry of Research and Technology/ the National Research and Innovation Agency. Moreover, several features were added to complement the secondary data obtained from several sources from other Ministries/Agencies. The dependent feature used in this study is routine (1: Yes, 0: No) obtained from the longitudinal data. The ratio binary class of dependent features is 60% Yes and 40% No. Based on this ratio, this data still tends to be balanced. The data was categorized as not extreme, with a ratio of less than 90% (Hakim, Sartono & Saefuddin, 2017). The R&D industry that routinely carries out R&D annually is the industry at the time of data collection consistently stating that it conducted R&D. Meanwhile, industries that do not routinely carry out R&D are those stating that they did not conduct R&D during data collection. The list of explanatory features and descriptions of each feature used in this study are displayed in Table 1. Imputation in missing data for several features was calculated using the miss forest method. Miss forest is a non-parametric method missing value imputation for mixed type data (Stekhoven & Bühlmann, 2012).

Table 1
Description of explanatory features

Features	Description	Statistics
Numeric		***Mean***
growth. investment	Average of Growth total investment on R&D	4.92
manpower	Average of total manpower/labour	1448.00
R&D personnel	Average of R&D Personnel (Researchers, Technicians, and equivalent staff, and other supporting staff)	21.00
researchers	Average of researchers	10.00
ratio.person.to.mp	The ratio of R&D Personnel to Manpower	5.00
ratio.research.to.person	The ratio of Researchers to R&D Personnel	46.00
Categoric		***Mode***
coop. univ	Ever been cooperation with University (1:Yes, 0:No)	0
extramural	Ever been given funding for another R&D unit outside the company (1:Yes, 0:No)	0
bumn	State own enterprises (1:State own enterprises, 2:Regional own enterprises, 0:Other)	0
export	Ever been export product (1:Yes, 0:No)	1
branch	Having branches for the same country (1:Yes, 0:No)	0
multinational	Multinational company (1:Yes, 0:No)	0

Extreme Gradient Boosting (XGBoost)

Chen and Guestrin (2016) developed extreme Gradient Boosting or XGBoost. XGBoost is a boosting-based ensemble tree algorithm that creates a new model to predict errors from the previous model. The final goal of this process is to obtain the function closest to $\hat{F}(x)$ to the constructor functions *f (x)* through the minimization of the loss function value $L(f, f(x))$, which is defined by the equation:

$$\hat{F} = argmin_f E_{x,y}[L(y, f(x))]$$

In the training process, each iteration is to minimize the loss function value based on the initial function$F_0(x)$. In general, the equation of the gradient boosting algorithm is shown as follows:

$$\{\gamma_m, h_m\} = argmin \sum_{m-1}^{M} L\left(y_i, f^{(m-1)}(x_i) + \gamma_m, h_m(x_i)\right)$$

Optimization performed by the XGBoost algorithm is 10 times faster than other Gradient Boosting implementations (Chen & Guestrin, 2016).

Model Parameters

In the XGBoost method, several parameters are used to optimize the model performance. Parameter optimization is important for XGBoost to prevent overfitting and high complexity (Wang & Liu, 2020). Table 2 shows the parameters that will be optimized in order to obtain a better accuracy value.

Table 2
List of XGBoost Parameters

Parameter	Description	Range
nrounds	Number of iterations	200:5000
eta	Learning rate at training process	0.01, 0.025, 0.05, 0.1, 0.3
gamma	Penalty parameter at regularization	0
max_depth	The depth level of a tree, the deeper a tree will be more complex	2:4
min_child_weight	The minimum value of weight needed by child node	1
subsample	The number of samples used for the training process. For example, 0.5 means using half of the data randomly in making new trees	1
colsample_bytree	The number of column samples for making new trees	1

Following the optimization of the XGBoost parameters through cross validation and three time repetitions, the optimal parameters obtained are nrounds: 3750, eta: 0.01, gamma: 0, max_depth: 3, min_child_weight: 1, subsample: 1, and colsample_bytree: 1.

Model Evaluation

The classification model generated from various methods is expected to classify all data correctly, yet it is almost impossible for a classification system to have 100% correctness and accuracy. The measure performance of the classification model can be measured through a

confusion matrix, which is a cross-tabulation between response feature data included in the prediction and observation class (Kuhn & Johnson, 2013). Cases with 2 cross-tabulation classes will form at Table 3.

Table 3
Confusion Matrix

		Prediction	
		Positive	Negative
Actual	Positive	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

The performance of a classification model can be measured by employing three values; accuracy, sensitivity, and specificity. Accuracy is the percentage of the correct model in making predictions with the following equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

F1-Score is a weighted comparison of average precision and recall, with the equation:

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \times 100\%$$

The Receiver Operating Characteristic (ROC) curve is another technique for evaluating the accuracy of the classification model prediction. The ROC curve illustrates the probability plot between the sensitivity and (1-specificity) values. The ROC curve can be converted into a scalar value, one of which is AUC. The AUC value ranges from 0 to 1. The better the performance of the classification model, the closer the AUC value is to 1.

Shapley Additive exPlanations (SHAP)

Shapley Additive exPlanations (SHAP) is an interpretation method for individual predictions based on Shapley Values of optimal game theory (Lundberg and Lee, 2017). Shapley Value is used to fairly measure each player's contribution in a game (Shapley, 1953). The purpose of the SHAP is to explain the predictions of an individual x by calculating the contribution of each variable or feature. Based on several equations to determine the contribution of each feature, Shapley's value on SHAP is presented as follows:

$$\phi_i = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$

The linear function in the variable g can be obtained based on the additive feature attribution method:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

Where $z' \in \{0,1\}^M$, equals 1 if the variable is observed and 0 if the variable is not observed. M is the value of a variable of a type of SHAP algorithm, which is TreeSHAP used for classification tree-based machine learning (Lundberg, Erion & Lee, 2018).

Results

XGBoost Model

In this study, the making of the XGBoost model used 75% of the training data, namely random data from all data for the training process, and 25% of the test data was used for process evaluation (Ahuja, Bansal, Prakash, Venkataraman & Banga, 2018). Then, the 10 fold cross-validation process was carried out in making the model obtain the stability of the model performance (Jung, Bae, Um, Kim, Jeon & Park, 2020). The process of 10-fold cross-validation will divide the training data into 10 random parts. From the 10 sections, 10 models will be trained using 9 parts of the data and tested using 1 part (subsample). Data analysis was performed using the caret, xgboost, SHAPforxgboost, and treeshap packages on software R version 4.0.3.

Based on this process, optimal parameters were obtained, and three performance measures were acquired: accuracy: 79.61, f1-score: 84.44%, and AUC: 76.46%. This modeling stage also compared a bagging-based ensemble tree, a random forest. Given the random forest, the results of the performance measure show accuracy: 74.76%, f1-score: 81.16%, and AUC: 72.8%. The comparison results imply that XGBoost produced better accuracy, f1-score, and AUC values than random forest.

Feature Analysis

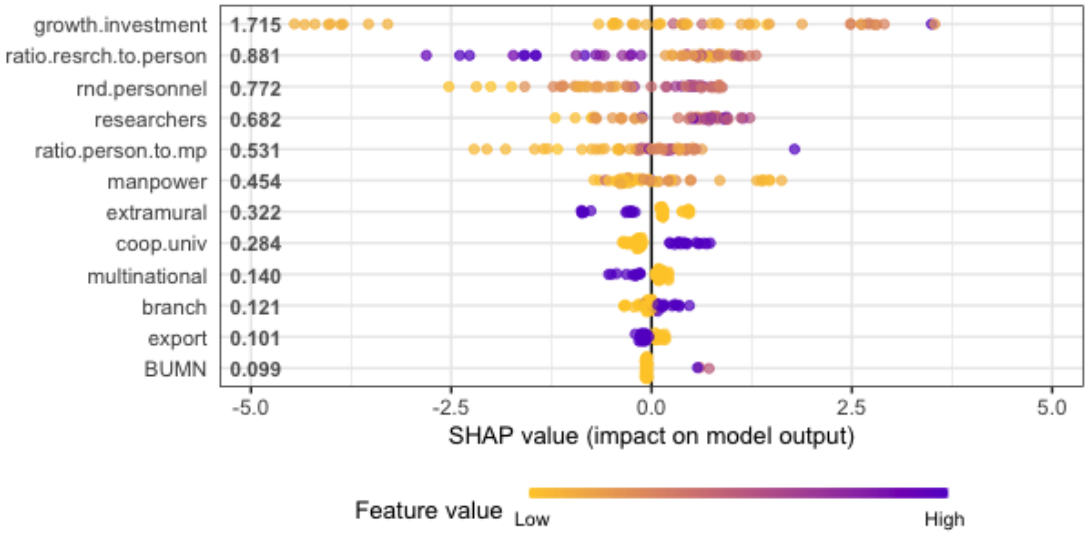


Figure 1: SHAP summary plot

Figure 1 is a SHAP summary plot in which the order of the features is based on the level of importance or contribution in predicting the industries that conduct R&D regularly and irregularly every year. The most important feature in the first order is the average growth in investment on R&D. In this feature, the higher the investment growth, the higher Shapley's value is. It denotes that the higher the investment growth, the higher the possibility for the industry to perform R&D regularly every year. Conversely, if the investment growth is lower, it will reduce the possibility of the industry conducting R&D regularly. It is very reasonable because once an industry experiences an increase in R&D expenditure from year to year; the industry already regards R&D to be an important thing for its business.

The other feature that contains the highest contribution is the ratio of researchers to R&D

human resources. In this feature, a higher value of the ratio of researchers to R&D human resources will reduce the opportunity for an industry to conduct R&D regularly. Conversely, if the ratio is smaller, it will increase the opportunity for an industry to conduct R&D regularly. However, given the plot, specifically for the ratio of researchers to R&D personnel, there is a high ratio value in the right of the 0 limits. It assumes that there is a certain limit in this feature that can determine when the ratio can increase or reduce opportunities. This limit can be observed in more detail in the description of the dependency analysis feature. One of the reasons for this negative linear relationship is that it is extremely rare for an industry to have such a high proportion of researchers compared to other R&D personnel. R&D personnel such as technicians and administrative personnel are also required in R&D activities. Therefore, if an industry has an adequate number of researchers compared to R&D personnel, it is necessary to check and reconfirm it to the industry.

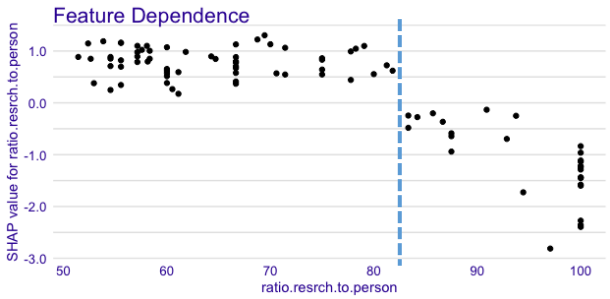
The next feature is R&D personnel, researchers, and ratio R&D personnel who have the same pattern as the first important feature: the average growth of investment on R&D. The value of the three features has a positive linear relationship with the contribution results or Shapley value. Meanwhile, the manpower feature has a negative linear relationship with Shapley's value, the same as the ratio of researchers to R&D personnel feature. Therefore, it is necessary to look deeper into the limit of how many manpower features can increase or decrease the opportunity to carry out R&D routinely.

Based on the explanation above, the feature with a high contribution comes from a feature of numeric type. Meanwhile, categorical features have a smaller contribution than the numeric feature. The SHAP summary plot shows that extramural, multinational and export interpret the same. If an industry carries out R&D financing to other units or industries, or if an industry is a multinational company and/or the industry experts, it will reduce the opportunity for an industry to conduct R&D regularly every year. Industries that conduct R&D financing to other units tend not to carry out R&D activities themselves.

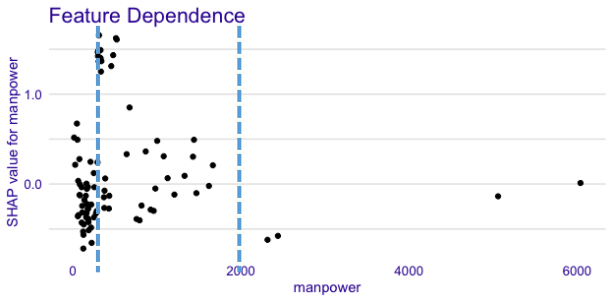
Further, multinational industries tend to carry out their R&D activities at the headquarters or Home Office, most in developed countries. Meanwhile, cooperation with universities, branches, and State-Owned enterprises has the same interpretation. If an industry collaborates with a university or has branches and/or is a non-governmental company, it will increase the opportunity for the industry to carry out R&D regularly. It shows that the opportunities for private and non-government industries to carry out R&D regularly are more favorable than state-owned industries.

Feature Dependency Analysis

Figure 2 shows the SHAP dependence plot deeper than the previous SHAP summary plot. Based on the explanation on the summary plot, the feature ratio of researchers to R&D personnel and manpower negatively relates to Shapley's value. In (a) shows that the limit for the ratio of researchers to R&D personnel is approximately 82%. If the ratio is less than 82%, this feature will increase the opportunity for an industry to conduct R&D regularly every year. On the other hand, if the ratio is more than 82%, it will reduce the opportunity for the industry to conduct R&D regularly. It reinforces the statement that it is necessary to double-check industries with more than an 82% ratio.



(a) **Impact ratio researchers to personnel R&D**

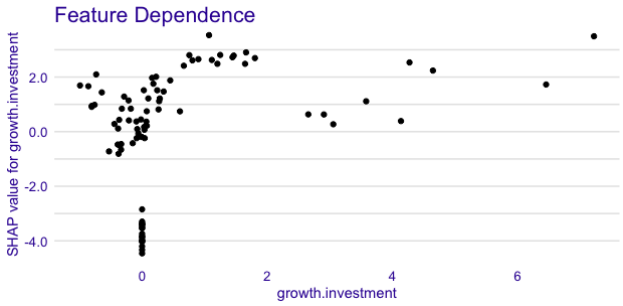


(b) **Impact manpower on Model Output**

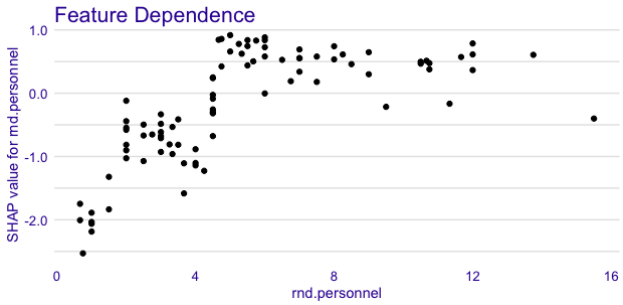
Figure 1: SHAP Dependence Plot

The same interpretation is shown in the manpower feature (b) that the average limit of the number of workers is around 150 people and 2000 people. If the average number of workers is less than 150 people or more than 2000 people, it is more likely to reduce the opportunity for the industry to carry out R&D regularly. Meanwhile, if an industry has an average number of workers between 150 and 2000 people, it increases the opportunity for the industry to carry out R&D regularly.

Figure 3 shows the feature's positive linear relationship with their contribution or Shapley's value. Based on the explanation on the SHAP summary plot, the higher the average growth value of the investment in R&D and R&D personnel, it will increase the opportunity for an industry to conduct R&D routinely. It is reinforced by the SHAP dependence plot in (c) and (d) in which the plot shows a positive relationship but is non-linear.



(c) **Impact growth investment on Model Output**



(d) **Impact personel R&D on Model Output**

Figure 2. SHAP Dependence Plot (1)

Discussion

This research is limited to using a black-box machine learning model with XGBoost that ensemble tree-based classification. The ensemble tree used the TreeSHAP algorithm to the use of SHAP. Hence, it is urged to compare it with the KernelSHAP algorithm using other machine learning such as Deep Neural Networks. The study to compound potency and multi-target activity prediction used machine learning models ensemble tree and Deep Neural Network then interpreting using SHAP with TreeSHAP algorithm and KernelSHAP algorithm. Kernel and Tree SHAP analyses were found to yield very similar results in the assessment of activity and potency predictions, with a high correlation between prioritized features (Rodríguez-Pérez & Bajorath, 2020). The comparison of the two algorithms, indeed,

is not only in measuring the performance of each model but also in the computational speed of each algorithm.

Furthermore, the interpretation using SHAP has not been detailed because it has not explained the relationship between the features. It needs to be carried out to determine how the features interact with each other so that the information obtained from the analysis of these features is deeper. Explaining how a prominent feature interacts with other features at study to analyze what makes an online review more helpful, (Meng, Yang, Qian & Zhang, 2021) used SHAP interactions that mapped the value of the prominent feature against its SHAP value in the samples of the whole dataset and colored the value of several other features with strong interactions on the prominent feature.

Based on the explanation of the analysis of these features, to determine the target for collecting industrial data from R&D players, the government needs to consider the size of the average growth investment, the ratio of researchers to R&D personnel, R&D personnel, researchers, the ratio of R&D personnel to manpower with an average Shapley value of 0.5. Besides, the government can also make policies to intervene in industries with a small opportunity to conduct R&D regularly.

Conclusion

XGBoost produces an accuracy of 79.61%, an f1-score of 84.44%, and an AUC of 76.46% to classify the R&D industries actors in Indonesia using. SHAP is one method for analyzing the contribution of each feature from black-box models. Based on Shapley's value, the feature with the highest contribution is the average growth investment in R&D, so the State-Owned Enterprise (BUMN) feature becomes the feature with the lowest contribution. In addition, the feature of ratio researchers to R&D personnel and manpower has a negative relationship with Shapley's values. If the ratio is less than 82%, this feature will increase the opportunity for an industry to conduct R&D regularly every year. On the other hand, if the ratio is more than 82%, it will reduce the opportunity for the industry to conduct R&D regularly. Likewise, the manpower feature with an average limit of the employee numbers is around 150 people and 2000 people.

References

Ahmad, A., Farooq, F., Niewiadomski, P., Ostrowski, K., Akbar, A., Aslam, F. & Alyousef, R. (2021). Prediction of compressive strength of fly ash based concrete using individual and ensemble algorithm. *Materials*, 14(4), 794. https://doi.org/10.3390/ma14040794

Ahuja, R., Bansal, S., Prakash, S., Venkataraman, K. & Banga, A. (2018). Comparative study of different sarcasm detection algorithms based on behavioral approach. *Procedia Computer Science*, 143, 411-418. https://doi.org/10.1016/j.procs.2018.10.412

Asmara, I. J., Achelia, E., G. Simamora, N. & Sartono, B. (2019). Measuring R&D performance using data envelopment analysis (DEA): Case Indonesia. *International Journal of Social Science and Humanity*, 9(4), 91–96. https://doi.org/10.18178/ijssh.2019.V9.997

Barreto, L. & Kypreos, S. (2004). Endogenizing R&D and market experience in the “bottom-up” energy-systems ERIS model. *Technovation*, 24(8), 615-629. https://doi.org/10.1016/S0166-4972(02)00124-4

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C. & Song, J. (2020). An interpretable prediction model for identifying n7-methylguanosine sites based on XGBoost and SHAP. *Molecular Therapy - Nucleic Acids*, 22,362-372. https://doi.org/10.1016/j.omtn.2020.08.022

[BPS] Badan Pusat Statistik. 2020. *Statistik Industri Manufaktur Indonesia 2018*. Jakarta (ID): Badan Pusat Statistik Republik Indonesia.

Chen, T. & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cardozo, M. L., Luzuriaga, A. G., Miranda, D. G., Lopez, H. R. P., Pajuelo, M. L. T. & Japura, G. A. (2021). Characterization of gross domestic expenditure on r&d in latin american countries during 2008-2017. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(5), 684-688. https://doi.org/10.17762/turcomat.v12i5.1469

Hakim, L., Sartono, B. & Saefuddin, A. (2017). Bagging Based Ensemble Classification Method on Imbalance Datasets. *IJCSN -International Journal of Computer Science and Network*, 6(6), 670-676.

Inekwe, J. N. (2015). The Contribution of R&D expenditure to economic growth in developing economies. *Social Indicators Research*, 124(3), 727-745. https://doi.org/10.1007/s11205-014-0807-3

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B. & Talebiesfandarani, S. (2019). PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere*, 10(7), 373. https://doi.org/10.3390/atmos10070373

Jung, K., Bae, D. H., Um, M. J., Kim, S., Jeon, S. & Park, D. (2020). Evaluation of nitrate load estimations using neural networks and canonical correlation analysis with K-fold cross-validation. *Sustainability (Switzerland)*, 12(1), 400. https://doi.org/10.3390/SU12010400

Kristjanpoller, W., Michell, K. & Minutolo, M. C. (2021). A causal framework to determine the effectiveness of dynamic quarantine policy to mitigate COVID-19. *Applied Soft Computing*, 104, 107241. https://doi.org/10.1016/j.asoc.2021.107241

Kuhn M. & Johnson K. (2013). Measuring performance in regression models. In *Applied Predictive Modeling*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3_5

Lundberg, S. M., Erion, G. G. & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M. & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, (pp. 4768–4777). https://dl.acm.org/doi/pdf/10.5555/3295222.3295230

Meng, Y., Yang, N., Qian, Z. & Zhang, G. (2021). What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3), 466-490. https://doi.org/10.3390/jtaer16030029

Molnar, C. (2019). Interpretable *machine learning. A guide for making black box models explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/

OECD (2020). *OECD main science and technology indicators*. R&D Highlights in the February 2020 Publication, Directorate for Science, Technology and Innovation. Retrieved from www.oecd.org/sti/msti2020.pdf.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S. & Mohammadian, A. (K.). (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136, 105405. https://doi.org/10.1016/j.aap.2019.105405

Rodríguez-Pérez, R. & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34(10), 1013-1026. https://doi.org/10.1007/s10822-020-00314-0

Shapley, L. S. (2016). 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II. Princeton: Princeton University Press*, Kuhn, Harold William and Tucker, Albert William (Eds.). https://doi.org/10.1515/9781400881970-018

Stekhoven, D. J. & Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3),647-665. https://doi.org/10.1007/s10115-013-0679-x

Wang, X.-W. & Liu, Y.-Y. (2020). Comparative study of classifiers for human microbiome data. *Medicine in Microecology*, 4, 100013. https://doi.org/10.1016/j.medmic.2020.100013

Zhang, Y., Ge, T., Tian, W. & Liou, Y. A. (2019). Debris flow susceptibility mapping using machine-learning techniques in Shigatse area, China. *Remote Sensing*, 11(23), 2801. https://doi.org/10.3390/rs11232801

Zhang, W. G., Li, H. R., Wu, C. Z., Li, Y. Q., Liu, Z. Q. & Liu, H. L. (2020). Soft computing approach for prediction of surface settlement induced by earth pressure balance shield tunneling. *Underground Space*, 6(4), 353-363. https://doi.org/10.1016/j.undsp.2019.12.003