IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 4, December 2024, pp. 4054~4061
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i4.pp4054-4061

Enhancing interpretability in random forest: Leveraging inTrees for association rule extraction insights

Fatma Hilali Moh'd, Khairil Anwar Notodiputro, Yenni Angraini
Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Dec 29, 2023
Revised Feb 29, 2024
Accepted Mar 21, 2024

Keywords:

Association rule
Extraction rule
InTrees
Random forest
Unemployment

ABSTRACT

The random forest model is a powerful supervised learner, recognized for its ability to learn the pattern within data with superior predictive accuracy. However, it is a black box model because it lacks interpretability. This study addressed the interpretable challenge by employing the inTree framework. The rules were extracted from each decision tree in a random forest model, and the association rules were determined through measured matrix support and confidence to reveal the frequent variable interactions for predicting unemployment. This approach provided insight into the relationships between specific variables and unemployment outcomes. The developed method used data set from the integrated labor force survey (ILFS) 2020/2021 in Zanzibar. Zanzibar's unemployment rate consistently increased across surveys conducted in 2006, 2014, and 2020/2021. Results have shown that the rules that most predict unemployment for individuals are female and lack of health insurance and secondary education level, female and youth age group and lack of health insurance and secondary education level with a high confidence level. This study provides practical insights for Zanzibar's government to develop effective interventions, programs, and policies. Improving the interpretability of the random forest model enhances decision-making to address unemployment challenges.

Corresponding Author:

Khairil Anwar Notodiputro
Department of Statistics, Faculty of Mathematics and Natural Science Sciences, IPB University
Bogor, Indonesia
Email: khairil@apps.ipb.ac.id

1. INTRODUCTION

Ensemble trees [1]–[4] are very powerful supervised learners, recognized for their highly exceptional predictive precision due to capturing information in data. The random forest model is the most used ensemble tree due to their adjustable parameters, ease of use, and capacity to manage limited sample size and high-dimensional feature spaces [5], [6]. Due to their high predictive performance, they are considered a must-try method for solving real problems. They have achieved outstanding scores in various data mining competitions, including predicting the employment status of job applicants [7], [8].

However, random forest is the most competitive model; it has been criticized for being a 'black box' model due to its complexity [9]–[11]. Because random forests combine multiple decision trees with different rules, it becomes difficult to interpret or understand the relationship between predictors and predictions [12]. The model's accuracy grows with rule complexity, although this complexity hinders interpretability. Interpretability is essential for learning from the model to understand the reasons behind decisions [13], [14]. An interpretable model allows a qualitative understanding of the relationship between predictor and outcome variables [15].

Like decision trees, tree-based models offer interpretability by translating their trees into rules, making them easier to understand and apply [16]. Decision rules take the form of if(condition)-then(prediction) statements [17], [18]. The condition outlines specific input variable values, while the prediction indicates the expected value of the outcome variable when an observation meets the given requirement in that condition. These rules operate on a conjunctive basis, requiring fulfillment of every condition; if an observation fails to meet the condition, the rule is not used for that observation. Random forests, an ensemble of decision trees, demonstrate various rule patterns.

Association rules, introduced by Agrawal *et al.* [19], extract correlations, frequent patterns, and associations among sets of items in databases. Association rules are "if-then" statements identifying combinations of items frequently occurring together in large datasets. In the context of this research, the item set refers to the decision rules within the random forest model [20]. Proposed an inTrees framework to interpret random forests by extracting rules, measuring rules, extracting frequent variable interactions, and handling the rules produced by the decision trees in the forest. The inTrees framework has been utilized in various studies for interpreting tree ensembles, including applications in breast cancer diagnosis [21] and the poverty status of households [12].

Unemployment is a severe problem that has an impact on various aspects, such as finances, mental health, and increased suicide rates [22]. We observed a notable increase in Zanzibar's unemployment rate in ILFS surveys conducted in 2006, 2014, and 2020/2021 at 5.5%, 14.3%, and 19.6%, respectively [23]. Given this situation, prediction is needed to identify significant variables that contribute to unemployment to help create better job opportunities. This study is the first in Zanzibar to use a random forest model to predict unemployment status. It is also the initial application of the inTrees framework in the field of unemployment prediction, enhancing interpretability and providing a unique contribution to the literature.

This study has utilized the integrated labor force survey (ILFS) data from Zanzibar in 2020/2021 to build the random forest model for predicting unemployment status and used the inTrees (interpretable trees) framework to extract and assess rules from random forest models, aiming to uncover frequent variable interactions that contribute to unemployment. This document is organized as follows: section 2 describes the research method for the utilized data, the variables, and the random forest model and inTree framework approaches. In section 3, results and discussion are presented. Finally, the conclusion and suggestion of this paper are briefly outlined in sections 4 and 5, respectively.

2. METHOD

2.1. Data and variables

The utilized data for this investigation is secondary data from the ILFS 2020/2021 conducted by the office of the Chief Government Statistician Zanzibar (OCGS). The dataset has 9608 observations, including individuals from the surveyed households within the working age range of 15 to 64 years. The outcome variable is binary: unemployment status (employed=0, unemployed=1). The ten predictor variables used to predict unemployment status are displayed in Table 1.

Table 1. Data variables

Variables	Description of variables	categories
sex	Sex of respondent.	Male=1 Female=2
age	Age of respondent	Youth=1 Adult=2
insu	Health insurance	Yes=1 No=2
disab	Disability	Yes=1 No=2

marit	Marital status	Single=1
		Married=2
		Widowed=3

citiz	Citizenship	Yes=1 No=2

read	Language proficiency	Kiswahili =1
		English & Kiswahili=2
		Cannot=3

edulev	Education level	Never attended=1
		Primary=2
		Secondary=3
		Vocational training=4
		Tertiary/non-university
		University=6

traini	Train attended	None=1 Yes=2
urbrur	Residence	Rural=1 Urban=2

2.2. Model

A random forest model was fitted by employing optimal parameters obtained from a grid search approach for predicting unemployment status. This model was trained with 100 decision trees and 4 predictors for splitting at each node, with 10 minimum sizes of terminal nodes and 200 maximum numbers of terminal nodes. For each decision tree, randomly select a subset of the data by bootstrap method and select 4 sets of predictors for each tree in the ensemble. After the model was trained, the model used out-of-bag (OOB) data, which means data not used for training, to estimate model performance. The overall prediction for each observation is obtained by taking a majority vote from individual tree predictions [24]. The model was evaluated on testing data using accuracy, sensitivity, specificity, and area under the curve (AUC).

However, random forest is difficult to interpret because it's like a black box. We suggested using inTrees for association rule extraction to make it more interpretable. This approach provides insights into the model's decisions and improves its interpretability [20].

2.3. Rule extraction from decision trees

Decision rules, as described by [17], are invaluable for simplifying decision trees and improving their interpretability. A decision tree can be transformed from a tree form into a decision rule by taking the form "if [condition] then [prediction]." Figure 1 is an example of extracting decision rules from a decision tree. The number of rules formed in random forests varies based on the number of final nodes in each decision tree. According to Figure 1, the result of the rules formed is as follows:

rule1= {([salary≥$100]) ⇒ ([Accept job])}
rule2= {([salary<$100])&([Gender=Male]) ⇒ ([Reject job])}
rule3= {([salary<$100])&([Gender=Female]) ⇒ ([Accept job])}

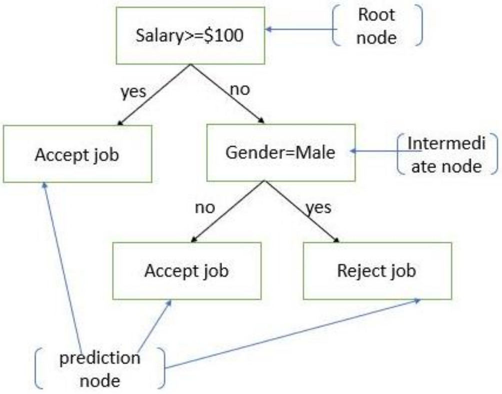

Figure 1. Decision tree structure

2.4. The intrees framework

The inTrees framework was proposed by [20] to interpret random forests by extracting rules, measuring rules, extracting frequent variable interactions, and handling the rules formed by the decision trees in the forest. The rule quality is measured by the length of the number of items in each condition, frequency is the popularity of the rule based on matching data rows, and error is the proportion of rows with mismatched response variable values. Mathematically, the frequency and error values are expressed as follows [12]:

$$frequency = \frac{the\ number\ of\ rows\ that\ match\ the\ condition\ that\ occurs}{the\ number\ of\ data\ rows} \tag{1}$$

$$error = \frac{the\ number\ of\ rows\ that\ match\ the\ condition\ but\ have\ a\ different\ response}{the\ number\ of\ data\ rows} \tag{2}$$

2.5. Association rule

From [25]–[27], association rules involve uncovering frequent associations or relationships between items or variables in a dataset. Association rules take the form "If antecedent, then consequent", along with a measure of the support and confidence associated with the rule. Support represents the percentage of database conditions that satisfy the given rule. Confidence measures the level of certainty in the identified association. Mathematically, the support and confidence formula are expressed as follows [27]:

$$support(X \Rightarrow Y) = \frac{number\ of\ rules\ containing\ both\ X\ and\ Y}{total\ number\ of\ rules\ formed} \tag{3}$$

$$confidence(X \Rightarrow Y) = \frac{number\ of\ rules\ containing\ both\ X\ and\ Y}{number\ of\ rules\ containing\ X} \tag{4}$$

where $X\ and\ Y$ represent condition and prediction in the rule, respectively.

2.6. Data analysis procedure

The data analysis procedure used in this study is demonstrated in Figure 2. As shown in Figure 2, data was imported, and then data exploration was done. After data was split into 80% for training and 20% for testing data, the training data was balanced using the synthetic minority over-sampling technique (SMOTE); then a random forest model was fitted. After the model was fitted, rules were extracted from random forests, and the most frequent rules were used to identify frequent variables. The construction random forest model was facilitated using the regularized random forest (RRF) package, and the inTrees package [28] was employed to extract the rule in the trees and compute the measure of the rule interest.

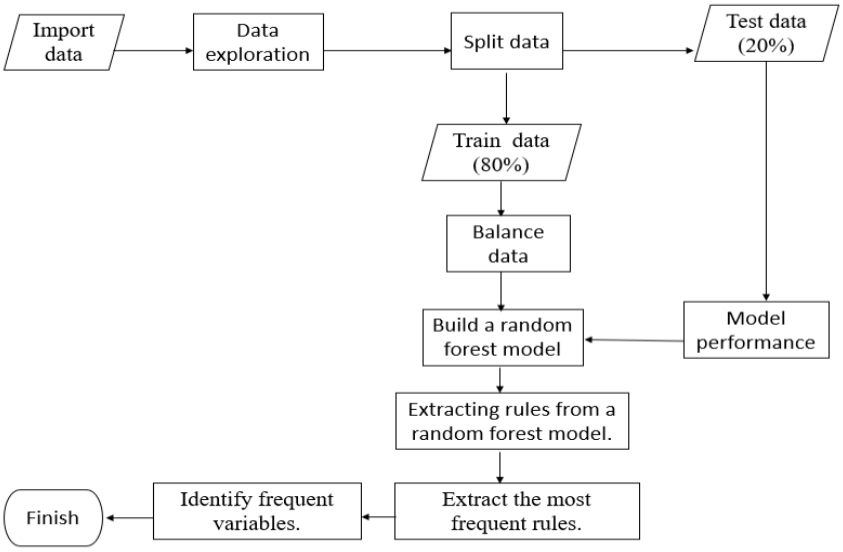
Import data
Data exploration
Split data
Test data (20%)
Train data (80%)
Balance data
Build a random forest model
Model performance
Extracting rules from a random forest model.
Finish
Identify frequent variables.
Extract the most frequent rules.

Figure 2. Flowchart of the procedure of data analysis

3. RESULTS AND DISCUSSION

3.1. Data exploration

There are no instances of missing data in any of the predictor variables or the outcome variable. The analysis of the outcome variable ‘unemployment status’ showed a distinct class distribution, where ‘employed’ accounts for 81.5 % and ‘unemployed’ for 18.5% of the dataset, as shown in Table 2. These observations indicate a significant data imbalance.

Table 2. Class distribution of an unemployment status

Unemployment status	Proportion
Employed	0.815
Unemployed	0.185

Unbalanced data can harm prediction accuracy because models are typically biased toward the dominant class in the dataset. To mitigate this issue, we employed a technique called SMOTE in training data to make data balance. SMOTE is designed to address imbalanced data by creating artificial samples for the minority class, thus enhancing the model's capacity to predict both classes within our 'unemployment status' variable more accurately [27]–[30].

3.2. Random forest model

The random forest model was built using the optimal parameter from the grid search approach. The model was trained with 100 decision trees and 4 predictors, with 10 minimum sizes of terminal nodes and 200 maximum numbers of terminal nodes. Subsequently, a model was evaluated using test data, presenting the performance outcomes in Table 3.

Table 3. Model performance on test data

Performance metrics	Value (%)
Accuracy	66.51
Sensitivity	74.72
Specificity	64.65
AUC	69.68

The model designed for predicting unemployment status achieved an overall accuracy of 66.51%. With a sensitivity of 74.72%, it demonstrates the ability to predict the unemployed class, with a specificity of 64.65%, highlighting its accuracy in predicting the employed class. The AUC is 69.68%, indicating a moderate level of accuracy in distinguishing between unemployed and employed individuals.

3.3. Rule extraction and measurement

This research obtained 544 unique rules from 5,531 total rules extracted from the first 100 trees. Among 544 unique rules, 325 rules predict employed while 219 rules predict unemployed. Table 4 demonstrates unique rules based on length extracted from a random forest.

This study focused on variable interactions with a minimum length of 2 (length=2), following the approach used in poverty status prediction literature [12], highlighting the significance of considering interactions between two variables to understand better the dynamics influencing the predicted outcomes. Referring to Table 4, there are 203 unique rules derived from the rules with 2 ≤ length ≤ 4 that contribute to predicting unemployment among individuals. Table 5 shows the statistical description of four rule measures generated by the inTrees framework: frequency, error, support, and confidence of 203 unique rules formed in predicting unemployment.

Table 5 displays the statistical summary of the rule's measurement involving frequency and error obtained from the dataset. Additionally, support and confidence are used for the association rule to determine frequent variables for predicting an outcome variable (unemployment status). The frequency values range between 0.2% and 89.6%, averaging 25.3%. However, error values vary from 21.3% to 50.0%, averaging 38.4% across the same dataset rows. For rules predicting unemployment status, the support values span from 0.010 to 0.101, averaging at 0.024 and with a median of 0.017. Confidence values range between 0.500 and 1.00, with an average of 0.707 and a median of 0.695.

Table 4. Unique rule counts by length for random forest model predictions

Length	Employed	Unemployed	Total unique rules
1	24	16	40
2	130	82	212
3	129	96	225
4	39	25	64
5	3	0	3
Total	325	219	544

Table 5. Descriptive statistics of rule measures for predicting unemployment (Rules with 2≤Length≤4)

Measures	Frequency	Error	Support	Confidence
Min	0.002	0.213	0.010	0.500
1st Qu	0.132	0.321	0.012	0.610
Median	0.209	0.389	0.017	0.695
Mean	0.253	0.384	0.024	0.707
3rd Qu	0.344	0.458	0.030	0.792
Max	0.896	0.500	0.101	1.000

3.4. The most variable frequent interactions

The analysis focused on finding frequent variable interactions that significantly contribute to predicting unemployment. Table 6 displays the top ten frequent variable interactions, outlining their respective metrics within the inTrees framework. Length (Len) shows the number of items or variables in each condition or rule. Frequency (Freq) and error (Err) are measurements that describe the rule based on the dataset. Support and confidence are measurements that describe the rule based on prediction.

Table 6. Top ten interactions of variables predicting unemployment sorted by highest support, with confidence ≥ 88.0%

Len	Condition	Freq	Err	Label	support	confide	pred
4	Sex=2 & age=1 & insu=2 & urbrur=2	0.195	0.223	1	0.029	0.891	1
2	Edulev=(1,2,3)& urbrur=2	0.475	0.408	1	0.018	0.882	1
2	Sex=2 & edulev=2	0.124	0.415	1	0.018	0.882	1
4	Sex=2 & age=1& edulev=3& urbrur=1	0.130	0.264	1	0.017	0.933	1
3	Sex=2 & edulev=3 & traini=1	0.257	0.278	1	0.016	0.964	1
4	Sex=2 & age=1 & edulev = (2,3) & urbrur=1	0.164	0.283	1	0.016	0.900	1
4	Sex=2 & age=1& edulev=3& traini=1	0.194	0.213	1	0.015	0.962	1
3	Sex=2 & insu=2 & edulev=3	0.366	0.282	1	0.015	1.000	1
3	Sex=2 & age=1& edulev=2	0.060	0.321	1	0.015	0.929	1
4	Sex=2 & age=1& insu=2 & edulev=3	0.278	0.227	1	0.014	1.000	1

Based on Table 6, the frequency and error of the first condition (Sex=2 & age=1 & insu=2 & urbrur=2) are 0.195 and 0.223, respectively. This indicates that the condition's popularity in the dataset is 19.5%, but the error is 22.3%, implying that the condition did not align with the dataset's outcome class variable (unemployed=1). Conversely, the support and confidence are 0.029 and 0.891, respectively, predicting unemployment. This means that a 2.9% proportion of rules is formed, satisfying the rule for predicting unemployment with a confidence of 89.1%. In other words, when this condition occurs, there is an 89.1% probability of predicting unemployment. Table 7 shows the characteristics of unemployed individuals in Zanzibar based on a confidence value of 88.0% above and sorted by most excellent support and lowest error.

In Table 7, the confidence value is the probability of an individual being classified as unemployed based on the interaction variable in a rule formed. Among 10 rules formed, the four dominant variables are being female, age group youth, having secondary education, and having primary education. These rules reveal key characteristics that strongly indicate unemployment among individuals.

- Females with secondary education and lacking health insurance showcase a 100.0% probability of unemployment.
- Youthful females lacking health insurance and secondary education display a 100% probability of unemployment.
- Females with secondary education and no training attendance have a 96.4% likelihood of unemployment.
- Youthful females with secondary education and no training attendance display a 96.2% probability of unemployment.
- Youthful females with secondary education and residing in rural areas display a 93.3% probability of unemployment.
- Youthful females with primary education have a 92.9% probability of unemployment.
- Youthful females with primary or secondary education and residing in rural areas have a 90.0% probability of unemployment.

Table 7. Characteristics of unemployed individuals based on confidence values ≥ 88.0%

Rule	sex	age	Health insurance	Education level	Attended training	Residence	Confidence (%)
1	female	-	no	secondary	-	-	100.0
2	female	youth	no	secondary	-	-	100.0
3	female	-	-	secondary	no	-	96.4
4	female	youth	-	secondary	no	-	96.2
5	female	youth	-	secondary	-	rural	93.3
6	female	youth	-	primary	-	-	92.9
7	female	youth	-	Primary/secondary	-	rural	90.0
8	female	youth	no	-	-	urban	89.1
9	-	-	-	Never attended/primary/secondary	-	urban	88.2
10	female	-	-	primary	-	-	88.2

4. CONCLUSION

This study successfully applied the interpretable random forest model to predict unemployment at the individual levels in Zanzibar. The interpretable model generated using the inTrees framework revealed insightful patterns and frequent variable interactions that contribute to unemployment. The key findings that contribute to unemployment are female sex and having no health insurance and her education level is secondary, female and her age group is youth and having no health insurance and her education level is secondary, both with confidence of 100%. These findings present the opportunity to tailor interventions and formulate targeted policies that address these variables, potentially leading to more effective employment strategies. These results provide a better understanding of the unemployment dynamic in Zanzibar.

5. SUGGESTION

In the future, this research can be extended to develop the interpretability of the double random forest. The double random forest can successfully model the underfitting data and has been preliminary investigated by previous studies, but needs improvement.

ACKNOWLEDGEMENTS

The author would like to extend heartfelt gratitude to those who have supported this research. Special thanks to the Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, and Ministry of Education of Indonesia for their invaluable sponsorship of my studies in Indonesia through the Partnership for Developing Country (KNB) scholarship program, offered by the Indonesian government, as well as Office of Chief Government Statistician (OCGS) of Zanzibar for provision of dataset were instrumental in this research.

REFERENCES

[1] B. Siswoyo, Z. A. Abas, A. N. C. Pee, R. Komalasari, and N. Suryana, “Ensemble machine learning algorithm optimization of bankruptcy prediction of bank,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 679–686, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp679-686.

[2] A. B. Gumelar, A. Yogatama, D. P. Adi, F. Frismanda, and I. Sugiarto, “Forward feature selection for toxic speech classification using support vector machine and random forest,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 717–726, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp717-726.

[3] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[4] A. Thebelt, J. Kronqvist, M. Mistry, R. M. Lee, N. S. -Merx, and R. Misener, “ENTMOOT: A framework for optimization over ensemble tree models,” *Computers and Chemical Engineering*, vol. 151, 2021, doi: 10.1016/j.compchemeng.2021.107343.

[5] G. Biau and E. Scornet, “A random forest guided tour,” *Test*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[6] F. Farooq *et al.*, “A comparative study of random forest and genetic engineering programming for the prediction of compressive strength of high strength concrete (HSC),” *Applied Sciences*, vol. 10, no. 20, Oct. 2020, doi: 10.3390/app10207330.

[7] O. Awujoola, P. O. Odion, M. E. Irhebhude, and H. Aminu, “Performance evaluation of machine learning predictive analytical model for determining the job applicants employment status’,” *Malaysian Journal of Applied Sciences*, vol. 6, no. 1, pp. 67–79, Apr. 2021, doi: 10.37231/myjas.2021.6.1.276.

[8] M. G. Celbiş, “Social networks, female unemployment, and the urban-rural divide in Turkey: Evidence from tree-based machine learning algorithms,” *Sosyoekonomi*, vol. 29, no. 50, pp. 73–93, Oct. 2021, doi: 10.17233/sosyoekonomi.2021.04.04.

[9] C. Bénard, G. Biau, S. D. Veiga, and E. Scornet, “Interpretable random forests via rule extraction,” *Arxiv-Statistics*, vol. 130, pp. 937–945, Apr. 2020, doi: 10.48550/arXiv.2004.14841.

[10] A. J. Sage, Y. Liu, and J. Sato, “From black box to shining spotlight: Using random forest prediction intervals to illuminate the impact of assumptions in linear regression,” *The American Statistician*, vol. 76, no. 4, pp. 414–429, Oct. 2022, doi: 10.1080/00031305.2022.2107568.

[11] S. Wenck, M. Creydt, J. Hansen, F. Gärber, M. Fischer, and S. Seifert, “Opening the random forest black box of the metabolome by the application of surrogate minimal depth,” *Metabolites*, vol. 12, no. 1, Dec. 2021, doi: 10.3390/metabo12010005.

[12] H. Ilma, K. A. Notodiputro, and B. Sartono, “Association rules in random forest for the most’,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 1, pp. 0185–0196, Apr. 2023, doi: 10.30598/barekengvol17iss1pp0185-0196.

[13] Tim Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[14] A. Vellido, “The importance of interpretability and visualization in machine learning for applications in medicine and health care,” *Neural Computing and Applications*, vol. 32, no. 24, pp. 18069–18083, 2020, doi: 10.1007/s00521-019-04051-w.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An introduction to statistical learning,” in *Springer Texts in Statistics*, vol. 103, no. 10, New York: Springer, 2013. doi: 10.1007/978-1-4614-7138-7.

[16] T. Elomaa, “In Defense of C4.5: notes on learning one-level decision trees,” in *Machine Learning Proceedings 1994*, Elsevier, 1994, pp. 62–69. doi: 10.1016/B978-1-55860-335-6.50016-7.

[17] M. Fokkema, N. Smits, H. Kelderman, and B. W. J. H. Penninx, “Connecting clinical and actuarial prediction with rule-based methods,” *Psychological Assessment*, vol. 27, no. 2, pp. 636–644, 2015, doi: 10.1037/pas0000072.

[18] O. Sagi and L. Rokach, “Approximating XGBoost with an interpretable decision tree,” *Information Sciences*, vol. 572, pp. 522–542, Sep. 2021, doi: 10.1016/j.ins.2021.05.055.

[19] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, Jun. 1993, doi: 10.1145/170036.170072.

[20] H. Deng, “Interpreting tree ensembles with inTrees,” *International Journal of Data Science and Analytics*, vol. 7, no. 4, pp. 277–287, Jun. 2019, doi: 10.1007/s41060-018-0144-8.

[21] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, “An improved random forest-based rule extraction method for breast cancer diagnosis,” *Applied Soft Computing*, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105941.

[22] F. Bianchi, G. Bianchi, and D. Song, “The long-term impact of the COVID-19 unemployment shock on life expectancy and mortality rates,” *Journal of Economic Dynamics and Control*, vol. 146, Jan. 2023, doi: 10.1016/j.jedc.2022.104581.

[23] NBS, “Integrated labour force survey 2020-2021,” *Tanzania National Bureau of Statistics*, Dodoma, Tanzania: NBS, 2021.

[24] H. V. T. Mai, T. A. Nguyen, H. B. Ly, and V. Q. Tran, “Prediction compressive strength of concrete containing GGBFS using random forest model,” *Advances in Civil Engineering*, vol. 2021, 2021, doi: 10.1155/2021/6671448.

[25] N. F. Idris, M. A. Ismail, M. S. Mohamad, S. Kasim, Z. Zakaria, and T. Sutikno, “Breast cancer disease classification using fuzzy-ID3 algorithm based on association function,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 448–461, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp448-461.

[26] T. Daniel, C. Larose, and D. Larose, *Data mining and predictive analytics*, New York: John Wiley & Sons, 2015.

[27] M. J. Huang, H. S. Sung, T. J. Hsieh, M. C. Wu, and S. H. Chung, “Applying data-mining techniques for discovering association rules,” *Soft Computing*, vol. 24, no. 11, pp. 8069–8075, 2020, doi: 10.1007/s00500-019-04163-4.

[28] H. Deng, X. Guan, and V. Khotilovich, “inTrees: interpret tree ensembles,” *CRAN R Project*, 2014, doi: 10.32614/CRAN.package.inTrees.

[29] D. Elreedy and A. F. Atiya, “A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance,” *Information Sciences*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[30] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, “SMOTE for handling imbalanced data problem : a review,” in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, IEEE, Nov. 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.

BIOGRAPHIES OF AUTHORS

Fatma Hilali Moh’d is a statistician at the Office of Chief Government of Statisticians in Zanzibar-Tanzania. She completed her Bachelor of Science in Statistics in 2016 from the University of Dodoma, Tanzania. She is pursuing a Master of Science in Statistics and Data Science at IPB University, Bogor, Indonesia. Since 2018, she has been actively involved in various roles at OCGS Statistics Zanzibar, contributing as a staff member across multiple national surveys and censuses. Her contributions encompass a range of areas, including employment and earnings surveys, household budget surveys, integrated labor force surveys, population censuses, and industrial production censuses. She can be contacted at email: fatmahmohd@apps.ipb.ac.id.

Khairil Anwar Notodiputro earned B.S. and Master’s degrees from IPB University and Ph.D. in Statistics from Macquarie University, Australia. His thesis title is “Modified fisher scoring algorithms for image reconstruction from projection”. He has been invited as a “Visiting professor at Prince of Songkla University and Kasetsart University, Thailand”. He is a Professor of Statistics and Data Science study Program at IPB University. He is a member of the International Statistical Institute and a former Chairman of the Indonesia Statistical Association. He is actively doing competitive research and supervising undergraduate and postgraduate students. His research interests include mixed models, small area estimation, time series analysis, and statistical machine learning. During the last ten years, he has taught the following subjects: time series analysis and forecasting, advanced data analysis, statistical analysis, statistical analysis, statistical method, and data science. He can be contacted at email: khairil@apps.ipb.ac.id.

Yenni Angraini is academic journey led her to excel in statistics, earning her Bachelor’s, Master’s, and Doctorate degrees in Statistics from IPB University. Currently serving as a dedicated faculty member. She is a respected lecturer at the Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University. Her commitment to education extends beyond the classroom, as evidenced by her active involvement in professional organizations. She is a proud member of the Ikatan Statistics Indonesia (Indonesian Statistician Association) and the Forum Pendidikan Tinggi Statistika Indonesia (Indonesian Higher Education Statistics Forum). For inquiries and collaboration opportunities. She can be contacted at email: y_angraini@apps.ipb.ac.id.