



Genetic Programming based Machine Learning in Classifying Public-Private Partnerships Investor Intention

Ahmad Amin

Faculty of Economics and Business, Universitas Gadjah Mada, Yogyakarta, Indonesia
amien@ugm.ac.id

Rahmawaty

Faculty of Economics and Business, Universitas Syiah Kuala, Aceh, Indonesia
rahmawaty@unsyiah.ac.id

Maya Febrianty Lautania

Faculty of Economics and Business, Universitas Syiah Kuala, Aceh, Indonesia
mayahaidar@unsyiah.ac.id

Rahayu Abdul Rahman

Faculty of Accounting, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak, Malaysia
rahay916@uitm.edu.my

Article Info

Article history:

Received Feb 18, 2023

Revised Mac 25, 2023

Accepted Apr 22, 2023

Keywords (minimum 5):

Genetic Programming
Machine Learning
Public-private Partnership
Investor Intention
Classification

ABSTRACT

To accelerate the growth of public infrastructure development, the government employs public private partnerships (PPP). However, this scheme exposes the private sector to various risks, including political risks, which can negatively impact the financial performance and reporting of participating firms. A significant challenge for the government is the insufficient private sector engagement in PPP arrangements. Hence, the purpose of this study is to evaluate the effectiveness of machine learning prediction models in categorizing private investor interest in PPP programs based on Indonesia evidences. The PPP data was analyzed in this study using two machine learning approaches, Genetic Programming and conventional machine learning, with testing results showing that all machine learning algorithms from both approaches achieved high accuracy rates of over 80%, with the Genetic Programming machine learning outperformed the conventional approach. This study highlights the potential of machine learning algorithms in predicting private investor interest in PPP programs, providing a tool for managing political risks and encouraging greater private sector participation.

Corresponding Author:

Ahmad Amin

Faculty of Economics and Business, Universitas Gadjah Mada, Yogyakarta, Indonesia

email: amien@ugm.ac.id

1. Introduction

Public-private partnerships (PPP) have emerged as a popular mechanism for accelerating public infrastructure development globally. Despite its potential benefits, PPP schemes pose various risks to private sector participants, including political risks, which can adversely affect their financial performance and reporting[1],[2]. One of the challenges facing governments is the lack of private sector engagement in PPP arrangements, particularly in Indonesia, where there is a higher risk due to economic volatility and the possibility of natural disasters[3],[4]. The possibility of natural disasters in Indonesia creates a higher risk environment for PPP arrangements, as these events can result in project delays, cost overruns, and disruptions in revenue streams, making it more challenging to attract private sector participation in PPP projects. As a result, identifying potential private sector investors' interests in PPPs becomes critical to ensure the successful implementation of infrastructure development projects[5].



Given the risks associated with PPPs, it is essential to identify potential private sector investors' interests in participating in these arrangements. Previous studies have attempted to predict private sector interest in PPPs using machine learning algorithms, but limited research has been conducted on Indonesia. Furthermore, these studies suggest that more advanced approaches are necessary to improve the accuracy of predictions.

Therefore, this research aims to fill this gap by introducing a Genetic Programming (GP) approach to predict private investor interest in PPP programs in Indonesia, and compare its efficacy with the commonly used machine learning approach of AutoModel RapidMiner. This study is significant as it presents an alternative method for identifying potential private sector investors in PPP arrangements, thereby offering new insights for the development of effective strategies to encourage private sector participation in infrastructure development. The proposed method that used GP machine learning was found to be user-friendly and not overly complex, thus making it accessible to data scientists with varying levels of expertise across multiple domains.

2. Literature Review

Genetic Programming (GP) is a type of artificial intelligence optimization algorithm that involves creating computer programs by using evolutionary algorithms inspired by biological evolution[6],[7]. It starts with a set of random programs that are evaluated based on a fitness function, and then the best programs are selected and combined through genetic operations such as mutation, crossover, and reproduction. This process is repeated over multiple generations, gradually improving the fitness of the programs and evolving them to solve the problem at hand. GP is often used for problems that are difficult to solve using traditional programming methods or require creative and innovative solutions. GP has been increasingly utilized in the field of machine learning.

One issue in machine learning that has attracted researchers to include GP is the challenge of selecting the appropriate features or variables for a given problem, which can be time-consuming and require domain-specific knowledge. GP offers a method for automatically selecting and optimizing the features or variables used in machine learning models, thereby improving their performance and reducing the need for human expertise. For an example, researchers [8] examines the effectiveness of constructing multiple features using GP in high-dimensional data classification o, with a focus on comparing single-feature construction and multi-tree GP representation. The study found that multiple-feature construction through multi-tree GP representation resulted in significantly better performance and improved interpretability of the constructed features, highlighting the potential of GP-based FC for real-world applications. The research in [9] demonstrated that GP can be an effective alternative for developing intelligent systems, particularly in the field of pattern recognition. The advantage of GP over other techniques such as regression and artificial neural networks is that it does not require an a priori definition of its structure. The study also found that feature engineering techniques can significantly improve the accuracy of the model. Overall, the research highlights the potential of GP for solving binary classification problems and suggests that it should be considered as an option for the development of intelligent systems. In [10], the researcher proposes a solution to the problem of class imbalance in datasets using a genetic algorithm (GA) with a fitness function based on entropy and information gain. The proposed solution aims to improve the impurity of the dataset and achieve a more balanced result without modifying the original dataset. The experiments were conducted on different datasets to evaluate the effectiveness of the proposed solution, and the results were compared with several other state-of-the-art algorithms. The study highlights the potential of genetic algorithms as a promising approach to address the problem of class imbalance in datasets.

Despite the increasing use of machine learning in various fields such as finance[11],[12], properties[13], banking[14], human behavior and social[15],[16], there has been limited research on its application to predicting and classifying PPP investments mainly with advance machine learning such as the inclusion of GP. Only a few of previous studies have employed machine learning to predict successful PPP projects [17]–[20], which have highlighted the potential benefits of using this intelligent approach to address various challenges related to PPP. Therefore, further research is needed to the use of machine learning in PPP, and accelerate the development of computational systems that can facilitate PPP projects and investments.

3. Methodology

3.1 The dataset

The effectiveness of the GP in machine learning was evaluated on a dataset comprising information gathered from 165 high-level executives of Indonesian publicly traded companies. To determine the correlation coefficient weights of each PPP attribute as independent variables (IVs), Pearson correlation tests were conducted using the actual collected data. Figure 1 depicts the heat map between all the IVs and dependent variable (DV).

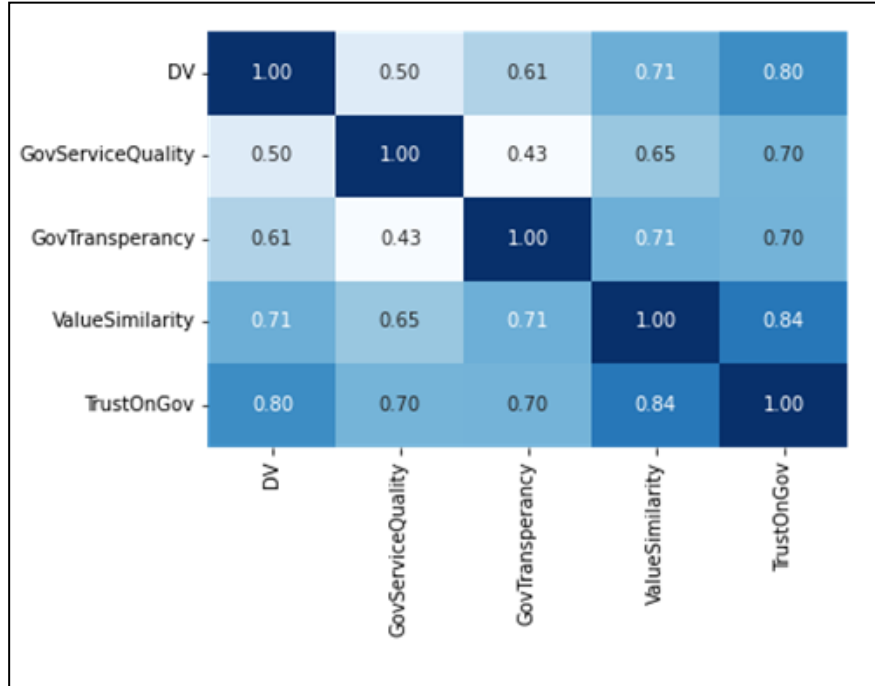


Figure 1. Heat Map of the PPP IVs and DV Correlation Coefficient

The PPP attributes used in this study represent factors that influence investment intention, specifically, *TrustOnGov* which is based on government trust, *GovServiceQuality* which reflects the perceived quality of government services, *GovTransperancy* which pertains to the perceived level of government transparency, and *ValueSimilarity* which measures the similarity of values between the public and private sectors. The DV in this study is the investor intention, which is classified as either 1 to indicate intention or 0 to indicate no intention. Equation 1 illustrates the formulation used to classify the dependent variable into either 1 or 0.

$$\text{Total IVs} = \begin{cases} 3 \text{ and above, intention}=1 \\ \text{Below 3, intention}=0 \end{cases} \quad (1)$$

3.2 GP Machine Learning Optimization

GP algorithm is the optimization method used for identifying the best machine learning pipelines, including the model features selection, the best outperforms machine learning algorithm and the suitable hyper-parameters of the selected machine learning. This study utilized Tree-Based Pipeline Optimization Tool (TPOT) library in Python[21] that enable GP machine learning. Figure 2 shows the inclusion of TPOT library in the execution that used computer notebook with 16GB RAM.

```
import pandas as pd
import numpy as np
import random as rnd

import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
!pip install tpot
from tpot import TPOTClassifier
```

Figure 2. Declaration of libraries in Python

TPOT is capable of supporting both regression and classification problems. As demonstrated in Figure 2, *TPOTClassifier* was employed in this study to address the investor intention classification problem. TPOT can save significant time and effort in the machine learning process by automating the pipeline optimization stage. When implementing machine learning, TPOT splits the dataset into training and testing subsets based on the designated *test_size*. According to information presented on line 5 of Figure 3, the testing dataset size used in this study was set at 40% of the total dataset, meaning that 50 out of the 165 data points were reserved for testing purposes. The remaining 115 data points were used for training the model.

```
X_train = train_df.drop(["DV"], axis=1).values
Y_train = train_df["DV"]
Features = X_train
Class = Y_train
Feature_Train, Feature_Test, Class_Train, Class_Test = train_test_split(X_train, Y_train, test_size=0.4)
tpot = TPOTClassifier(generations=3, population_size=40, mutation_rate=0.9, crossover_rate=0.1, verbosity=2, cv=5)
tpot.fit(Feature_Train, Class_Train)
print(tpot.score(Feature_Test, Class_Test))
```

Generation 1 - Current best internal CV score: 0.8484210526315789

Generation 2 - Current best internal CV score: 0.8484210526315789

Generation 3 - Current best internal CV score: 0.8578947368421053

Best pipeline: RandomForestClassifier(BernoulliNB(RandomForestClassifier(input_matrix, bootstrap=True, criterion=entropy, 0.8787878787878788
/usr/local/lib/python3.9/dist-packages/sklearn/metrics/_scorer.py:794: FutureWarning: sklearn.metrics.SCORERS is deprecating warnings.warn(

Figure 3. Simple Python codes to execute TPOT

Additionally, the training subset is further divided into K-Fold cross-validation sets, which are used for both training and validation purposes. In this study, 5 K-Fold cross-validation was used as seen in line 6 of Figure 3. The parameters of GP in TPOT are *generations*, *population_size*, *mutation_rate*, *crossover_rate* and *verbosity*. Although the default values of these parameters have produced favorable results in preliminary experiments (Refer Figure 3), this study specifically aims to compare the impact of varying the *population_size* parameter. Moreover, Figure 3 displays the output of each validation for every generation of GP, specifically for the three generations analyzed. The best pipeline, which is the most efficient machine learning model identified by TPOT, is also indicated in Figure 3. The testing accuracy of the best pipeline is reported as 0.89 in the figure.

3.3 Testing performance metrics

In this study, the accuracy of the model was measured from the validation and testing processes, and the performance was evaluated with different population sizes. In addition to accuracy,

the Area Under Curve (AUC) from the Receiver Operating Characteristic (ROC) plot was also compared to assess the model's performance.

The ROC curve is a graphical representation that illustrates the trade-off between sensitivity and specificity for a binary classifier as the decision threshold varies. AUC is a single-number summary of the ROC curve and represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Compared to accuracy, AUC is a more robust metric for evaluating the performance of a model in imbalanced datasets. This is because accuracy can be misleading in cases where the data is imbalanced and the negative class is dominant. In such scenarios, a model that always predicts the negative class can still achieve high accuracy, despite being practically useless. AUC, on the other hand, considers the true positive rate (sensitivity) and the false positive rate (1-specificity) across all possible decision thresholds and is not affected by imbalanced data. Therefore, AUC is a better metric than accuracy for evaluating the performance of models in imbalanced datasets.

4. Results and Discussion

The obtained results have been presented in two categories. The first category pertains to the validation and testing accuracies of the machine learning algorithm optimized by the GP, whereas the second category focuses on the Area Under Curve (AUC) results. The testing accuracies of GP machine learning at different population sizes are given in Figure 4.

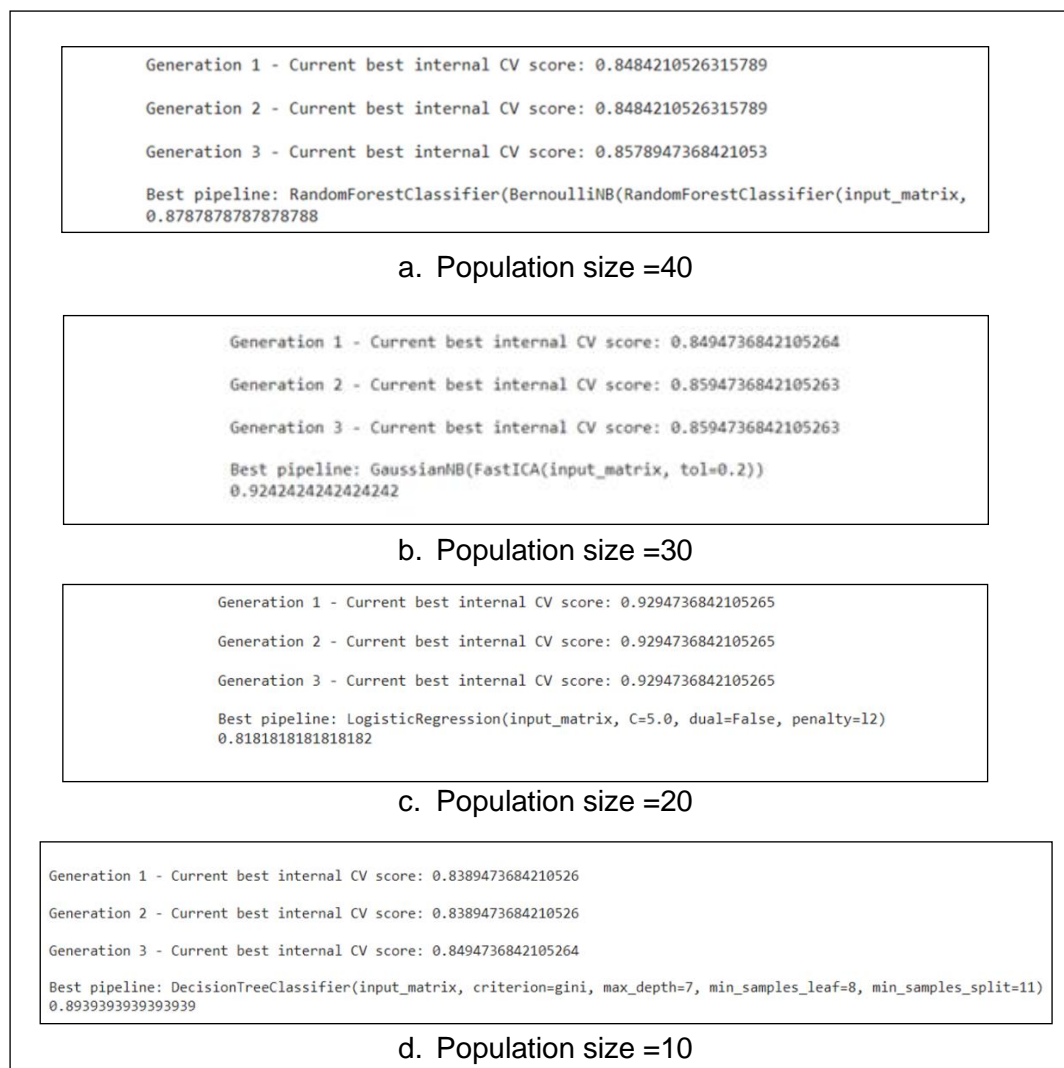


Figure 4. Output from TPOT

As depicted in Figure 4, the results obtained from the various validation stages across all GP generations, as well as the testing phase, indicate that the machine learning algorithm developed using GP is highly accurate, with all validation and testing accuracies exceeding 80%. Notably, the most optimal testing accuracy was achieved when the GP population size was set to 30. Figure 5 presents the testing results from conventional machine learning on the same dataset. The results demonstrate that the GP-based machine learning approach outperformed the non-GP algorithms. Figure 6 presents the AUC values for comparing the performance of the GP machine learning and the conventional machine learning.

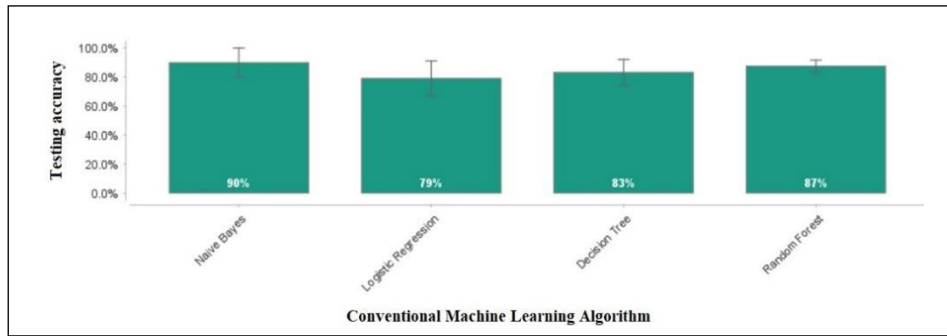


Figure 5. Accuracy of testing from conventional machine learning

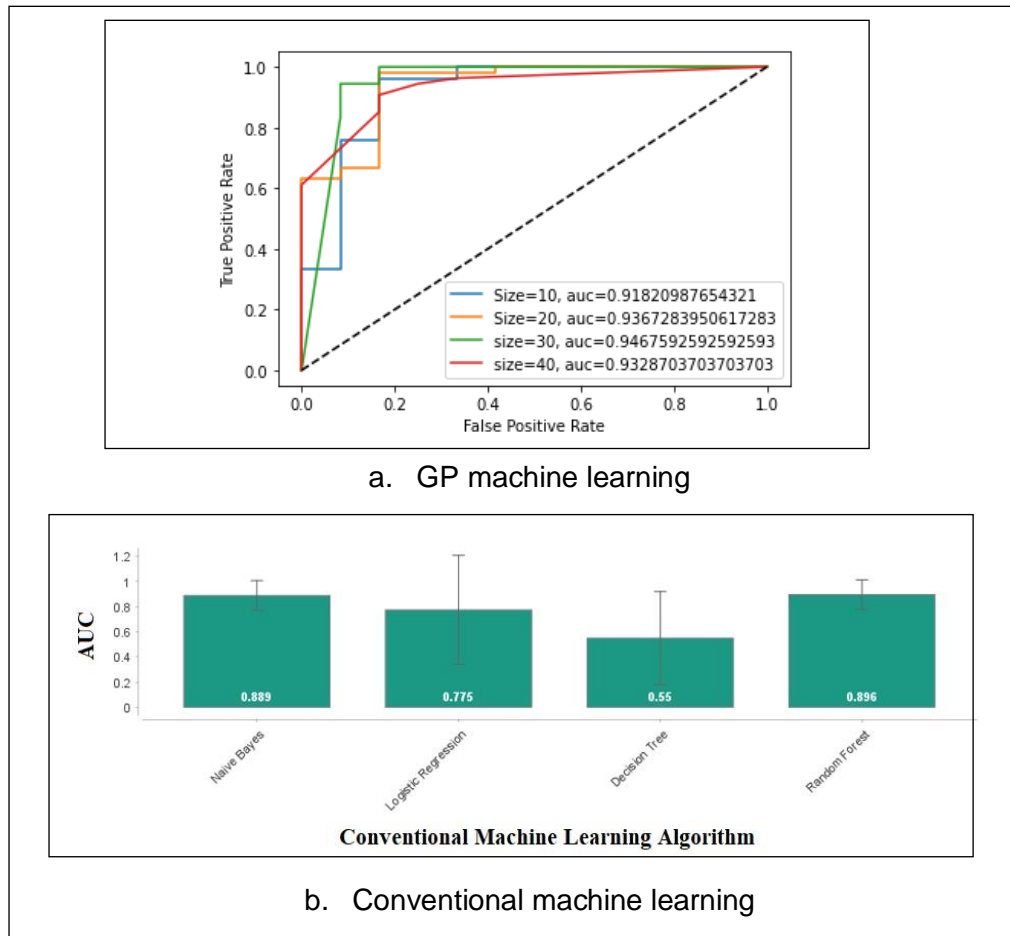


Figure 6. AUC results

The results in Figure 6a reveal that the efficacy of the GP machine learning techniques is consistent with the AUC scores. The optimal AUC score was obtained by GP machine learning using the *population_size* of 30, which aligns with the highest testing accuracy achieved by the GP-based models in Figure 4. Both Figure 6a and Figure 6b inform that The GP machine learning outperform traditional machine learning algorithms in terms of AUC values.

5. Conclusion

This study has introduced an advanced machine learning framework based on genetic programming (GP) that is easy to implement, even for inexpert data scientists from various domains, including policy makers and stakeholders in the domain of Public-Private Partnership (PPP). The results obtained from the framework using the tested dataset of PPP Indonesia cases have shown improved performance over conventional machine learning methods. The proposed GP-based approach has achieved higher accuracy and AUC values compared to non-GP methods.

These findings have important implications for the PPP domain as the proposed GP-based framework can assist in decision-making processes and improve project success rates. Future work can focus on extending the framework to other PPP datasets and evaluating its effectiveness in real-world settings. Moreover, the proposed framework can be extended by including additional features or using more advanced GP techniques to enhance its performance further.

Acknowledgements

We would like to express our gratitude towards Universitas Gadjah Mada, Yogyakarta, Indonesia and Universiti Teknologi Mara, Malaysia for their valuable support in conducting this collaborative research.

Conflict of Interest


The authors declare no conflict of interest in the subject matter or materials discussed in this manuscript.




References

- [1] C. B. Casady, K. Eriksson, R. E. Levitt, and W. R. Scott, "(Re) defining public-private partnerships (PPPs) in the new public governance (NPG) paradigm: an institutional maturity perspective," *Public Manag. Rev.*, vol. 22, no. 2, pp. 161–183, 2020.
- [2] G. Y. Debela, "Critical success factors (CSFs) of public-private partnership (PPP) road projects in Ethiopia," *Int. J. Constr. Manag.*, vol. 22, no. 3, pp. 489–500, 2022.
- [3] E. Endri, Z. Abidin, T. P. Simanjuntak, I. Nurhayati, and others, "Indonesian stock market volatility: GARCH model," *Montenegrin J. Econ.*, vol. 16, no. 2, pp. 7–17, 2020.
- [4] R. Djalante, "A systematic literature review of research trends and authorships on natural hazards, disasters, risk reduction and climate change in Indonesia," *Nat. Hazards Earth Syst. Sci.*, vol. 18, no. 6, pp. 1785–1810, 2018.
- [5] H. Yurdakul, R. Kamacsak, and T. Y. Öztürk, "Macroeconomic drivers of Public Private Partnership (PPP) projects in low income and developing countries: A panel data analysis," *Borsa Istanbul Rev.*, vol. 22, no. 1, pp. 37–46, 2022.
- [6] J. C. Bansal, P. K. Singh, N. R. Pal, and others, *Evolutionary and swarm intelligence algorithms*, vol. 779. Springer, 2019.
- [7] T. Bäck, D. B. Fogel, and Z. Michalewicz, *Evolutionary computation 1: Basic algorithms and operators*. CRC press, 2018.
- [8] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognit.*, vol. 93, pp. 404–417, 2019.
- [9] L. W. Santoso, B. Singh, S. S. Rajest, R. Regin, and K. H. Kadhim, "A genetic programming approach to binary classification problem," *EAI Endorsed Trans. Energy Web*, vol. 8, no. 31, pp. e11–e11, 2021.
- [10] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor, "A classification model for class imbalance dataset using genetic programming," *IEEE Access*, vol. 7, pp. 71013–71037, 2019.
- [11] A. Priyadarshini, S. Mishra, D. P. Mishra, S. R. Salkuti, and R. Mohanty, "Fraudulent credit card transaction detection using soft computing techniques," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 3, pp. 1634–1642, Sep. 2021, doi: 10.11591/ijeecs.v23.i3.pp1634-1642.

- [12] R. A. Rahman, S. Masrom, and N. Omar, "Tax Avoidance Detection Based on Machine Learning of Malaysian Government-Linked Companies," no. 2, pp. 535–541, 2019, doi: 10.35940/ijrte.B1083.0982S1119.
- [13] S. Jamil, T. Mohd, S. Masrom, and N. A. Rahim, "Machine Learning Price Prediction on Green Building Prices," in *2020 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2020*, 2020. doi: 10.1109/ISIEA49364.2020.9188114.
- [14] A. Muneer, R. F. Ali, A. Alghamdi, S. M. Taib, A. Almaghthawi, and E. A. Abdullah Ghaleb, "Predicting customers churning in banking industry: A machine learning approach," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 1, pp. 539–549, 2022, doi: 10.11591/ijeecs.v26.i1.pp539-549.
- [15] N. L. Saad and R. Ibrahim, "Context-Aware Recommender System based on Machine Learning in Tourist Mobile Application," vol. 3, no. 1, pp. 19–28, 2022.
- [16] S. Masrom, R. A. Rahman, N. Baharun, S. Redzwan, and S. Rohani, "Machine learning with task-technology fit theory factors for predicting students' adoption in video-based learning," vol. 12, no. 3, pp. 1666–1673, 2023, doi: 10.11591/eei.v12i3.5037.
- [17] S. Ghorbany, S. Yousefi, and E. Noorzai, "Evaluating and optimizing performance of public-private partnership projects using copula Bayesian network," *Eng. Constr. Archit. Manag.*, no. ahead-of-print, 2022.
- [18] Y. Wang and R. L. K. Tiong, "Public-Private Partnership Contract Failure Prediction Using Example-Dependent Cost-Sensitive Models," *J. Manag. Eng.*, vol. 38, no. 1, p. 4021079, 2022.
- [19] M. Eskandari, M. Taghavifard, I. R. Vanani, and S. S. G. Noori, "An Intelligent Hybrid Model for Determining Public-Private Partnership in Iranian Water and Wastewater Industry Based on Collective Tree Algorithms," *J. Water Wastewater*, vol. 32, no. 1, pp. 69–90, 2021.
- [20] Y. Wang, Z. Shao, and R. L. K. Tiong, "Data-driven prediction of contract failure of public-private partnership projects," *J. Constr. Eng. Manag.*, vol. 147, no. 8, p. 4021089, 2021.
- [21] R. S. Olson and J. H. Moore, "TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning," *Automated Machine Learning: Methods, Systems, Challenges*. Springer International Publishing, pp. 151–160, 2019. doi: 10.1007/978-3-030-05318-5_8.

Biography of all authors

Picture	Biography	Authorship contribution
	Ahmad Amin is a lecturer at the Accounting Department, Faculty of Economics and Business Universitas Gadjah Mada, Indonesia. In 2006, he received his master's degree from Universitet i Agder in Kristiansand, Norway. He is currently pursuing his doctorate at the Accounting Research Institute (ARI) Universiti Teknologi Mara, Shah Alam, Malaysia. Public sector accounting, public finance, PPP, governance, and ethics are among the research topics addressed. Apart from research, Ahmad Amin is involved in community service by assisting Indonesian local governments in implementing good government governance, such as the development of financial reports and public accountability reports. In terms of professional organizations, he participates in several forums organized by the Indonesian Institute of	Design the research work, evaluation

	Accountants. Ahmad Amin can be reached through amien@ugm.ac.id .	
	<p>Rahmawaty, SE., M.Si., Ak., CA is a senior lecturer in the Accounting Department at Universitas Syiah Kuala, USK. She completed her master's degree in Postgraduate Program of Universitas Syiah Kuala in 2008. Currently, she is taking a doctoral program at USK. Her research projects cover topics about Public Sector Accounting, Taxation, Sharia Accounting. Currently, she is interested in research about Performance of Local Government issues. Rahmawaty has published a number of papers in preferred Journals and chapters in books and participated in a range of forums on Accounting.</p> <p>She can be contacted through email at rahmawaty@unsyiah.ac.id</p>	Drafting article, camera ready
	<p>Maya Febrianty Lautania is a lecture at the faculty of economy and business, university of syiah kuala in 2006. I received my bachelor degree in 1999 from FEB, Unsyiah. She holds two master degree, i.e. master of management which is completed in 2003 and master of accountancy completed in 2011. Both of my master degrees were received from unsyiah. My research interests are in micro-entrepreneur, accountancy, stock, and exchange. She can be contacted through email at mayahaidar@unsyiah.ac.id</p>	Literature review, data collection
	<p>Dr. Rahayu Abdul Rahman is an Associate Professor at the Faculty of Accountancy, UiTM. She received her PhD in Accounting from Massey University, Auckland, New Zealand in 2012. Her research interest surrounds areas, like financial reporting quality such as earnings management and accounting conservatism as well as financial leakages including financial reporting frauds and tax aggressiveness. She has published many research papers on machine learning and its application to corporate tax avoidance. She is currently one of the research members of Machine Learning and Interactive Visualization Research Group at UiTM Perak Branch. She can be contacted through email at rahay916@uitm.edu.my.</p>	Genetic Programming TPOT experimental works

Surat kami : 700-KPK (PRP.UP.1/20/1)

Tarikh : 20 Januari 2023

Prof. Madya Dr. Nur Hisham Ibrahim
Rektor
Universiti Teknologi MARA
Cawangan Perak



Tuan,

PERMOHONAN KELULUSAN MEMUAT NAIK PENERBITAN UiTM CAWANGAN PERAK MELALUI REPOSITORI INSTITUSI UiTM (IR)

Perkara di atas adalah dirujuk.

2. Adalah dimaklumkan bahawa pihak kami ingin memohon kelulusan tuan untuk mengimbas (*digitize*) dan memuat naik semua jenis penerbitan di bawah UiTM Cawangan Perak melalui Repositori Institusi UiTM, PTAR.

3. Tujuan permohonan ini adalah bagi membolehkan akses yang lebih meluas oleh pengguna perpustakaan terhadap semua maklumat yang terkandung di dalam penerbitan melalui laman Web PTAR UiTM Cawangan Perak.

Kelulusan daripada pihak tuan dalam perkara ini amat dihargai.

Sekian, terima kasih.

“BERKHIDMAT UNTUK NEGARA”

Saya yang menjalankan amanah,

Setuju.

27.1.2023

SITI BASRIYAH SHAIK BAHARUDIN
Timbalan Ketua Pustakawan

PROF. MADYA DR. NUR HISHAM IBRAHIM
REKTOR
UNIVERSITI TEKNOLOGI MARA
CAWANGAN PERAK
KAMPUS SERI ISKANDAR

nar