Available online at
https://myse.my/

e-ISSN: 0128-2417

Malaysian Journal of Sustainable Environment 13(2) 2026, 234–244

Malaysian Journal of Sustainable Environment

Recalibrating reverse migration factors through a machine learning model in Malaysia

Nur Huzeima Mohd Hussain[1*], Suraya Masrom[2], Thuraiya Mohd[3], Azreen Anuar[4]

[1,3]*Faculty of Built Environment, Universiti Teknologi MARA Perak Branch, Seri Iskandar Campus, 32600 Seri Iskandar, Perak, MALAYSIA*
[2]*Faculty of Computer and Mathematical Sciences , Universiti Teknologi MARA Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, MALAYSIA*
[4]*Perbadanan Bekalan Air Pulau Pinang, Level 4, Komtar, Georgetown 10300, Penang MALAYSIA*

ARTICLE INFO

Article history:
Received 29 November 2025
Revised 31 December 2025
Accepted 25 January 2026
Online first 08 July 2026
Published 31 August 2026

Keywords:
Reverse
Migration
Machine Learning
Recalibration
Malaysia

DOI:
10.24191/myse.v13i2.12548

ABSTRACT

Reverse migration is the catalyst that reverberates across social, economic, and demographic landscapes, influencing both the destination and origin regions. With an increase of understanding that continues to grow, it is not possible in a resource-constrained world, to assume that rural-to-cities migration will continue to grow indefinitely. It needs to be re-examined. This paper presents a comprehensive analysis of the factors that motivate migrants to move from rural-to-cities and back again. As migration predictions are notorious and complex, this study adopted a machine learning technique in predicting reverse migration based on the dataset by the Department of Statistical Malaysia (DOSM). These features and the significance of the reverse migration factors were compared between the three types of tree-based machine learning (Gradient Boosted Trees, Random Forest, and Decision Tree). The results depicted that the 'destination state' stood out as the most important feature in all the machine learning reverse migration prediction models. Notably, although this factor exhibited the lowest correlation coefficient in the Random Forest algorithm, the combined contributions of other factors, each with lower coefficient values, collectively rendered Random Forest the most exceptional algorithm, achieving a remarkable accuracy rate of 93.3%. The findings of this study are valuable for authorities in formulating effective strategies for sustainable development, allocating resources, and human settlements in both urban and rural regions.

[2*]Corresponding author. *E-mail address*: nurhu154@uitm.edu.my,
Co-authors. *E-mail address* : suray078@uitm.edu.my, thura231@uitm.edu.my

1. INTRODUCTION

In the evolving landscape of migration studies, the dynamics of reverse migration have emerged as a compelling area of exploration. Reverse migration shows the return of individuals or groups to their region of origin after residing elsewhere and reflects a nuanced interplay of factors that shape and redefine the socio-economic fabric of both destination and origin regions.

This research delves into the intricate task of recalibrating reverse migration factors, leveraging the transformative capabilities of machine learning models. As a catalytic force, reverse migration embodies a multifaceted spectrum of motivations, encompassing economic, cultural, and personal dimensions. Understanding and accurately recalibrating these factors are imperative to grasp the underlying patterns and making informed predictions, thereby facilitating more responsive and nuanced policy frameworks toward future developments.

In this pursuit, machine learning has allowed the incorporation of machine learning models to predict the origin and destination; of any number of exogenous feature flows of human migration (Robinson, 2018). Unlike traditional methods such as the gravity and radiation model, the algorithm of machine learning is considered an emerging technology that can study data without the clear specification of relationships, and machine learning models based on the data set can discover more multifaceted relationships or patterns between variables during the learning phase (Robinson, 2018; Anuar, 2023). Machine learning is a robust and valuable method widely used by researchers but has been limited in migration prediction trends. By harnessing the computational prowess of these models, the objectives will uncover hidden correlations, extract factors, and ultimately contribute to a more inclusive understanding of the factors leading to reverse migration.

Therefore, this study aims to compare the importance of factors in the ML reverse migration models in accordance with the ML prediction accuracies. The findings will determine significant factors through comparative charts and the Pareto chart towards predicting reverse migration factors that would be useful to government agencies and the national agenda.

2. LITERATURE REVIEW

2.1 Reverse Migration and Related Factors

Migration has been a pervasive and transformative phenomenon throughout human history, shaping societies, economies, and cultures across the world. Jobs, education, family, and health have always been influential factors for movements for people (Hussain, 2015; Anuar, 2023). However, an important aspect that has gained increasing attention is the phenomenon of reverse migration.

Reverse migration refers to the movement back to the region of origin after a period of residing in another location. This process is dynamic and can be influenced by multiple factors, including economic opportunities, cultural ties, changing political landscapes, health, and other personal considerations (Dandekar, 2020; Takahashi, 2021; Anuar, 2013). While reverse migration has been observed in various contexts, its significance lies in its potential to impact the social, economic, and demographic fabric of both the destination and origin regions (Qingjiu,2018; Danial, 2022).

According to Statistical Malaysia, the urban-to-rural migration rate is rising from 15.2% in 2016 to 19.5% in 2018 (DOSM, 2018). These figures showed the culmination of a trend of increasing urban-to-rural migration and is expected to gradually increase especially after the strike of the COVID-19 pandemic. Anuar et al. (2021) in a study added that migrants returned to their hometowns to seek peace, sympathy, and survival. In this context, the study also explores the recalibration of reverse migration factors through the lens of machine learning. Leveraging advanced computational models, this research seeks to uncover factors, correlations, and predictive insights through automated ML screening on the statistical data year 2018 and conceptual framework.

These ML screening factors are aligned with a study done by Anuar et al. (2021) that focuses on reverse migration factors using a Systematic Literature review (SLR). The research established family, occupation, environment, economic, social aspects, and quality of life to be the factors that 'pull' migrants back to their hometown. Refer to Fig. 1.

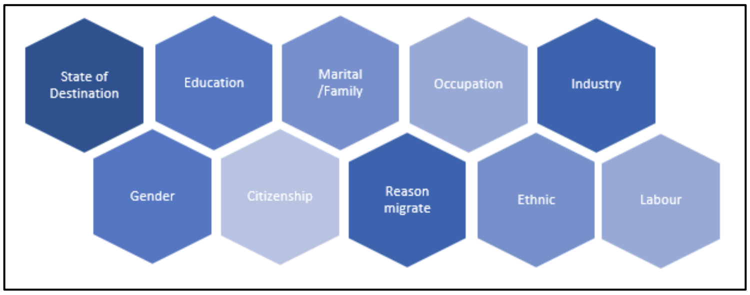

Fig. 1. Factors of migration
Source: Authors

Thus, understanding and recalibrating the factors that drive reverse migration is crucial for policymakers, researchers, and communities. The dynamics of reverse migration can inform strategies for sustainable development, resource allocation, and community integration. Furthermore, in an era marked by increasing interconnectedness and mobility, comprehending the patterns and drivers of reverse migration is essential for fostering inclusive and resilient societies.

By doing so, this study provides a foundation for evidence-based statistical decisions and migrant engagement in response to the evolving dynamics of migration and reverse migration.

2.2 Machine Learning

Machine learning has raised widespread attention in current years due to its ability to analyse huge amounts of data and recognise patterns and insights that could reduce errors in human practices. Machine learning was applied crosswise of domains, highlighting its potential to revolutionize various industries (Ja'afar, 2021). Machine learning algorithms can be broadly grouped into two categories: linear-based and tree-based. While linear algorithms have been commonly used in the past, tree-based algorithms have increased significant popularity in recent years due to the capability to handle non-linear relationships and interactions between variables (Hussain, 2021). Despite this, there are limited studies that compare the performance of linear and tree-based models, particularly in the context of reverse migration prediction. The emerging evidence of rising reverse migration has been alarming in recent years, particularly during economic downturns, political instability, natural disasters, and the COVID-19 pandemic. These events have led to the displacement of migrant workers in urban areas, causing them to return to their rural homes (Micevska, 2021). The consequences of reverse migration can be significant, affecting not only the social and economic well-being of the returnees but also the industries and employers that rely on migrant workers based on a study done by Caulifield (2019). Predicting reverse migration can help policymakers and stakeholders design and implement appropriate interventions to mitigate its negative effects. Despite the potential benefits of applying machine learning algorithms for predicting reverse migration, only a limited number of studies have explored these shifting trends in population mobility (Robinson, 2018). Furthermore, Robinson (2018) added that machine learning models have shown better performance than conventional models in calculating intercountry migration flows, related factors, and prediction, especially when the data is complicated. Therefore, adopting machine learning models in recalibrating factors is significantly important for current and future references.

3. RESEARCH METHODOLOGY

This study employed a set of 104 secondary data prescribed by the Department of Statistical Malaysia (DOSM) for the migration in Selangor in the year 2018. Rapid Miner software implemented the dataset and experiments for understanding the data sampling, splitting the dataset into training and testing sets, and evaluating the ML algorithms. Rapid Miner provided an Auto Model to allow this research to observe the features and importance of each migration factor and to compare the performances of machine learning algorithms. The migration factors are represented as the independent variables (IVs) or features for the machine learning reverse migration models. Meanwhile, the dependent variable (DV) for the prediction model was the target label classified as 1 or 0, representing the 'migration' or 'no migration'.

3.1 ML Training and Testing Datasets

Based on the 104 records from the DOSM dataset, a 70:30 ratio was employed to partition the data into training and testing datasets. Consequently, 73 records were allocated for testing, while the remaining 31 records were set aside as holdout samples, as illustrated in Fig. 2.

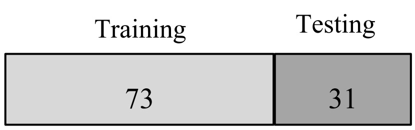

Fig. 2. ML training and testing sets
Source: Authors

3.2 Selecting a Machine Learning Algorithm

Selecting a reliable machine learning algorithm for a specific task is considered a critical step in the model development process. The choice of algorithm depends on various factors, including the nature of the problem, the characteristics of the data, the available computational resources, and the desired outcome. The use of Auto Model Rapid Miner offers the advantage of automating this critical process of algorithm selection. Out of the nine (9) suggested algorithms by Auto Model, Decision Tree, Random Forest, and Gradient Boosted Trees were the three most accurate algorithms recommended by the Auto Model. These three (3) suggested algorithms are in line with the commonly identified reviewed from established literature. Based on previous studies, these selected algorithms have presented promising results in predicting the movement of people and reverse migration (Hussain, 2024). For example, Hussain, (2021) implemented a Random Forest machine-learning technique for movement in people, while Loh (2011) used a Decision Tree for prediction in reverse migration

4. RESULT AND DISCUSSION

The significance of implementing the process with Auto Model is to identify the best machine learning algorithms and optimal parameters to develop the reverse migration models based on machine learning. Results generated by the Auto Model suggested three (3) types of machine learning algorithms that most pursue the nature of population statistical datasets. Besides this Auto Model generation, the study has also determined an appropriate machine learning algorithm for reverse migration through a systematic literature review (SLR) (Hussain, 2021). Subsequently, the SLR determination models are aligned with the Auto Model algorithm suggestions. This remark that the selection is significant and well-established in the body of knowledge. The suggested algorithms employed in this study are the Decision Tree, Random Forest, and Gradient-Boosted Tree.

4.1 The Factors in Machine Learning Reverse Migration Models

The significance of each independent variable (IVs) in each of the Machine Learning reverse migration models is presented in Fig. 3. The correlation weights for each independent variable in the Decision Tree, Gradient Boosted Tree, and Random Forest models present the importance of the factors in the Machine Learning reverse migration models.

https://doi.org/10.24191/myse.v13i2.12548

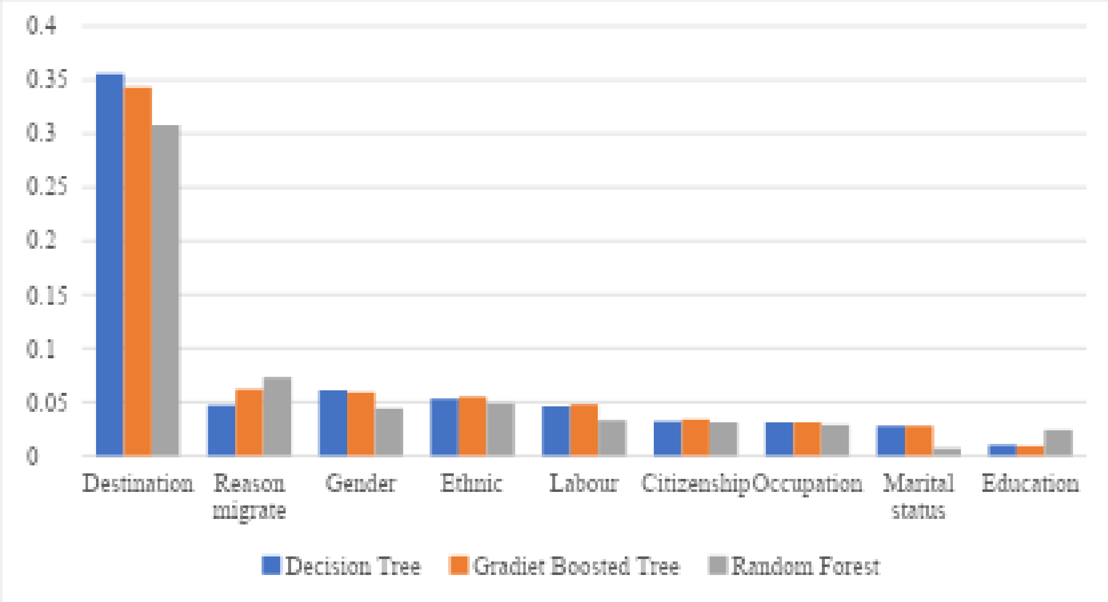

Fig. 3. The IVs correlation coefficients in three machine-learning reverse migration models
Source: Authors

According to the correlation coefficients provided in Fig. 3, it presented that the IVs are correlated to migration and show nonlinearity in their association with the DV. The low correlation coefficients below 0.1 suggested that there was no strong linear correlation between the migration IVs and the DV. Notably, the ‘Destination’ variable shows a strong correlation coefficient for all ML reverse migration models with an average weightage of 0.3. These findings suggest the pivotal role of the destination state as the most important feature in all ML algorithms, particularly in the Decision Tree algorithm. Regardless of the destination, Random Forest prioritizes 'Reason for migration' and 'Education' as the most influential factors, while in Gradient Boosted Trees, 'Ethnic' emerges as the second most significant contributor. This diversity in feature importance across different ML algorithms underscores the varied importance of factors in predicting reverse migration.

Despite the presence of nonlinearity and the indication of low correlation coefficients, there is merit in incorporating these variables into the machine learning reverse migration model. These variables might encapsulate intricate relationships that linear-based machine learning struggles to capture. They could comprehend valuable information that enhances the overall accuracy of the model. The results in Fig. 3 show that machine learning algorithms possess the capability to discern complex and non-linear relationships between variables, which may elude detection through simple correlation coefficients. Table 1 illustrates that utilizing all factors results in accurate predictions by machine learning algorithms.

https://doi.org/10.24191/myse.v13i2.12548

Table 1: Results of Performance

ML Algorithm	Accuracy
Decision Tree	53.3%
Random Forest	93.3%
Gradient Boosted Trees	63.3%

Source: Authors

Table 1 depicted that Random Forest as the most outstanding ML algorithm in the reverse migration prediction model. Despite having the lowest correlation coefficient of the ‘Destination’ factor in the Random Forest algorithm, the aggregated impacts of other factors, each with lower coefficient values (less than 0.01), collectively distinguished Random Forest as the most outstanding algorithm, attaining an impressive accuracy rate of 93.3%. On the contrary, despite 'Destination' attaining the highest correlation coefficient in the Decision Tree, the accuracy of this algorithm is the lowest at 53.3%. Notably, Decision Tree employs lower correlation coefficients from 'Reason migrate' and 'Education' compared to Random Forest. As the second-best performer, Gradient Boosted Trees utilizes a slightly higher correlation weight for 'Reason migrate' and 'Education' than Decision Tree. These findings highlight the meaningful contributions of all variables incorporated in the reverse migration models.

Besides adopting ML, this study also seeks to analyse the significant distribution of the factors by using a Pareto chart. According to Alkiayat (2021), a Pareto chart enables quality improvement in making decisions and prioritizes suitable interventions to identify the highly weighted factors and achieve the desired outcome, with less effort and unnecessary costs. The Pareto chart shown in Figure 4 presented ‘Destination’ to be the most significant factor in correlation coefficient weightage, which also concluded that ‘Destination’ is the major factor that pulls migrants to reverse migration.

These findings revealed the factors influencing reverse migration based on the Rapid Miner analysis. By reviewing the previous empirical analysis through the Systematic Literature Review (SLR) together with the extraction of the factors discussed in the Rapid Miner analysis, this paper highlighted 8 variables as the most influential factors in reverse migration.

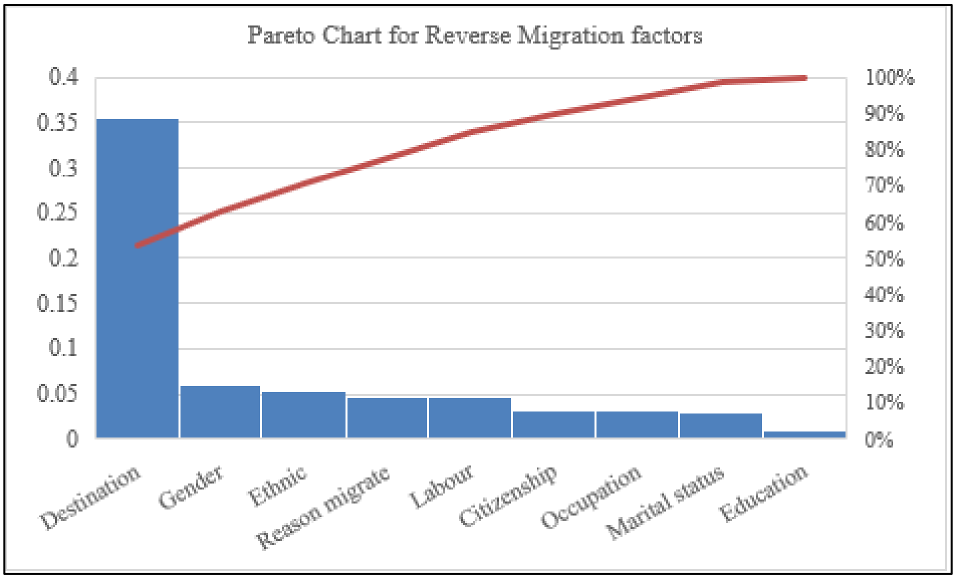

Fig. 4. Pareto chart for reverse migration factors
Source: Authors

5. CONCLUSION

In conclusion, with shreds of evidence from both SLR and machine learning, this study recalibrates the reverse migration factors of population movement in predicting future migration trends and dimensions. The study presents three (3) machine learning models including Decision Tree, Gradient Boost Tree, and Random Forest algorithms to interpret the tested dataset. The findings established the machine learning prediction on population changes and patterns that impacted future rural and urban settlements in Malaysia. The most influential factors that recalibrated the urban-to-rural migration are 'destination', 'reason to migrate', and 'social and cultural attachment'. This is opposed to the former reasons of rural-to-urban migrants who were attracted to the urbanization of the economy, education, and quality of life. As emerging evidence portrays challenges in cities, the reasons that 'pull' migrants back to resilience are compromisable. With the accurateness of this model, this study also indicated Random Forest as one of the recommended algorithms to pursue the development of machine learning models for reverse migration in Malaysia. Despite the conventional method, this machine learning model is sufficiently effective at forecasting reverse migration. Therefore, with regards to the Industrial Revolution 4.0, the Machine Learning Model can effectively predict and pursue many substantial issues for society, and the industry. Through this exploration, this study attempts to not only advance academic discourse but also offer practical implications for policymakers, planners, community leaders, and researchers navigating the evolving dynamics of migration. The recalibration of reverse migration factors through machine learning models stands as a transformative endeavor, promising to enrich our understanding of this phenomenon and, in turn, inform strategies that foster resilience, inclusivity, and sustainable development in our interconnected world.

https://doi.org/10.24191/myse.v13i2.12548

6. ACKNOWLEDGEMENTS/FUNDING

The authors would like to express gratitude to the research team, PJIM UiTM Perak Branch, the Universiti Teknologi MARA, and the Ministry of Higher Education Malaysia for the full support of this study.

7. CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

8. AUTHORS' CONTRIBUTIONS

All authors jointly contributed to the research design, questionnaire development, data analysis, and manuscript preparation. All authors have read and approved the final version of the manuscript.

9. DECLARATION OF GENERATIVE AI IN THE WRITING PROCESS

AI tools (e.g, ChatGPT, Grammarly) were used for drafting, assistance, paraphrasing and grammar checking. All content was critically reviewed and edited by researcher.

10. DATA AVAILABILITY/SUPPLEMENTARY MATERIALS

The datasets used and/or analysed during this study are available from the corresponding author upon reasonable request.

11. ETHICS STATEMENT

The authors declare that this research did not involve human or animal subjects.

12. REFERENCES

Adewale, J. G. (2005). Socio-economic factors associated with urban-rural migration in Nigeria: A case study of Oyo State, Nigeria. *Journal of Human Ecology, 17*(1), 13–16. https://www.researchgate.net/publication/267716974

Alkiayat, M. (2021). A practical guide to creating a Pareto chart as a quality improvement tool. *Global Journal on Quality and Safety in Healthcare, 4*(2), 83–84. https://doi.org/10.36401/JQSH-21-X1

Anuar, A., Hussain, N. H. M., & Byrd, H. (2023). Tree-based machine learning in classifying reverse migration. *Mathematical Sciences and Informatics Journal (MIJ), 4*(1), 49–56. https://ir.uitm.edu.my/id/eprint/78335/

Anuar, A., Hussain, N. H. M., Mohd, T., Masrom, S., & Byrd, H. (2021). Reverse migration determinant factor: A systematic literature review. *Environment-Behaviour Proceedings Journal, 6*(16), 309–315. https://doi.org/10.21834/ebpj.v6i16.2731

Anuar, A., Mohd Hussain, N. H., Mohd, T., Masrom, S., & Ahmad, S. (2022). Migration intentions in Selangor, Malaysia: A descriptive analysis. Virtual Go Green: *Conference and Publication (v-GOGREEN 2021)*

"Rethinking Built Environment: Towards a Sustainable Future". https://ir.uitm.edu.my/id/eprint/65616/

Brohi, S. N., Pillai, T. R., Kaur, S., Kaur, H., Sukumaran, S., & Asirvatham, D. (2019). Accuracy comparison of machine learning algorithms for predictive analytics in higher education. *In International Conference for Emerging Technologies in Computing* (pp. 254–261). Springer, Cham. https://doi.org/10.1007/978-3-030-23943-5_19

Caulfield, M., Bouniol, J., Fonte, S. J., & Kessler, A. (2019). How rural out-migrations drive changes to farm and land management: A case study from the rural Andes. *Land Use Policy*, 81, 594–603. https://doi.org/10.1016/j.landusepol.2018.11.030

Dandekar, A., & Ghai, R. (2020). Migration and reverse migration in the age of COVID-19. *Economic and Political Weekly, 55*(19), 28–31. https://www.epw.in/journal/2020/19/commentary/migration-and-reverse-migration-age-covid-19.html

Danial, M. H., & Williamson, P. (2022). Differential urbanisation in Malaysia, 1980-2010. *Planning Malaysia Journal, 20*(23). https://doi.org/10.21837/pm.v20i23.1158

Hussain, N. H. M. (2015). *From Kampong to city and back again: A case study of de-urbanisation in Malaysia.* University of Auckland. https://doi.org/10.13140/RG.2.2.20331.34082

Hussain, N. H. M., Anuar, A., Masrom, S., Mohd, T., & Ahmad, N. A. (2021). Machine learning of the reverse migration models for population prediction: A review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(5), 1830–1838. https://turcomat.org/index.php/turkbilmat/article/view/2197

Hussain, N. H. M., Anuar, A., Masrom, S., Mohd, T., Ahmad, N. A., & Byrd, H. (2024). Machine learning in reverse migration classification. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 38*(2), 45–55. https://doi.org/10.37934/araset.38.2.4555

Hussain, N. H. M., Byrd, H., & Ahmad, N. A. (2017). The urban migrant future desire: A study of reverse migration in Malaysia. *Malaysian Journal of Sustainable Environment, 3*(2), 177–192. https://ir.uitm.edu.my/id/eprint/29530/

Ja'afar, N. S., & Mohamad, J. (2021). Application of machine learning in analysing historical and non-historical characteristics of heritage pre-war shophouses. *Planning Malaysia Journal, 19*(16). https://doi.org/10.21837/pm.v19i16.953

Kang, C., Liu, Y., Guo, D., & Qin, K. (2015). A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint. *PLOS ONE, 10*(11), e0143500. https://doi.org/10.1371/journal.pone.0143500

Loh, W. (2011). Classification and regression trees. *WIREs Data Mining and Knowledge Discovery, 1*(1), 1–14. https://doi.org/10.1002/widm.8

Micevska, M. (2021). Revisiting forced migration: A machine learning perspective. *European Journal of Political Economy*, 70, Article 102044. https://doi.org/10.1016/j.ejpoleco.2021.102044

Qingjiu, S., & Maliki, N. Z. (2018). Rural-urban migrants sense of place in the city of Baoding China. *Planning Malaysia Journal, 16*(8). https://doi.org/10.21837/pm.v16i8.539

Robinson, C., & Dilkina, B. (2018). A machine learning approach to modeling human migration. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies,* 1–8. https://doi.org/10.1145/3209811.3209868

Takahashi, Y., Kubota, H., Shigeto, S., Yoshida, T., & Yamagata, Y. (2021). Diverse values of urban-to-rural migration: A case study of Hokuto City, Japan. *Journal of Rural Studies,* 87, 292–299. https://doi.org/10.1016/j.jrurstud.2021.09.013