INTERNATIONAL JOURNAL OF
ACADEMIC RESEARCH IN
BUSINESS & SOCIAL SCIENCES

Published Online: 18 February 2023 Vol 13, Issue 2, (2023) E-ISSN: 2222-6990

# Reverse Migration Factor in Machine Learning Models

Azreen Anuar[1], Nur Huzeima Mohd Hussain[2], Suraya Masrom[3], Thuraiya Mohd[4], Suriati Ahmad[5], Nur Azfahani Ahmad[6]

[1]Centre of Graduate Studies, Universiti Teknologi MARA Perak Branch, Seri Iskandar Campus, Malaysia, [2,4,5,6]Department of Built Environment Studies and Technology, Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA Perak Branch, 32610 Seri Iskandar Campus, Malaysia, [3]Faculty of Computer and Mathematical Sciences Universiti Teknologi MARA Perak Branch, Tapah Campus Malaysia

Email: nurhu154@uitm.edu.my, azreenanuar0309@gmail.com

## Abstract

Models of human migration predictions are adequately important to Governments and society as comprehensive tools for making predictions on population changes in the near future. The prediction models would accommodate a database of the readiness and effect in labor markets, the risk of spreading infectious diseases, and the capacity in enduring economic fluctuations. However, as migrations are complex and unpredictable in the huge demographic processes, predicting migration can be notorious and have a high error rate. Using a machine learning approach that can intelligently predict the reverse migration based on the tested data set is a significant way to reduce errors. Thus, this paper aims to explore and compare machine learning models namely three gradient-boosted trees, a random forest, and a decision tree in reverse migration forecasting. Besides the performance of the comparisons, this paper presents the specific weight correlation in each of the machine learning models to describe the importance of the reverse migration factors to the model.

**Keywords:** Reverse Migration, Machine Learning, Factor, Model

## Introduction

Human migration has historically been an important process of urbanization in developing countries. Previous studies described various motives that make people move. The reason varies, including the potential or opportunity for education, jobs or employment, well-equipped infrastructure, marital status, inheritance responsibility, and being attracted to the modern way of living in cities (Hussain, 2015; Kang et al., 2015).

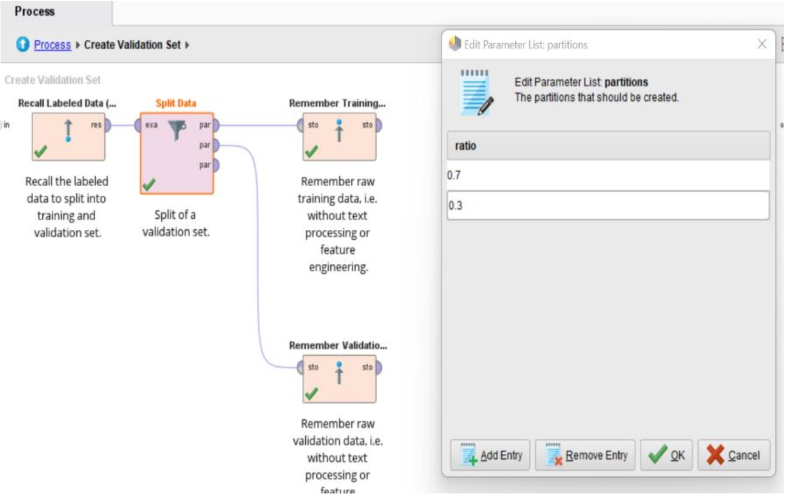
The motivations behind migration flow and the emergence of new types of migration that transcend short-term and long-term mobility have raised the importance of forecasting and conceptualizing migration in the contemporary 20th century. The raising concern in preparing comprehensive migration prediction data is demandable to accommodate the future. Furthermore, predicting migration trends are important in measuring land occupancy, spatial availability, population distribution, and food security that subsequently foreseeing economic trends, society growth and environmental needs. This information is sufficiently reliable for the local administrator, government, legal planners, investors and perhaps the community to strategies and make a survival living. Thus, understanding and predicting population mobility or human migration are crucial in every aspect.

The conventional method of measuring and predicting changes in migration trends requires more energy, cost, time and expertise. Several studies has established machine learning potential in showing better performance, especially in huge and complicated data (Adewale, 2005). Machine learning is useful in many fields, including business, computer science, industrial engineering, bioinformatics, medicine, pharmacology, physical science, and statistics to learn about forecast future events Mohd (2020), UN (2017)). Machine learning has received a lot of attention from the research and communities due to the availability of data, the variety of open-source machine learning tools, and powerful computing (Brohi et al., 2019). As migration involves huge and various empirical dataset, forecasting through machine learning are suitably acceptable. In machine learning deployment, selecting the machine learning algorithm is substantially important (Praveena and Jaiganesh, 2017). This paper explores a machine learning model namely three gradient-boosted trees (GBT), a random forest (RF), and a decision tree (DT) in reverse migration forecasting. The two common types of these machine learning; random forest (RF) and decision tree (DT) algorithms are reported as the best outperforming machine learning when tested on different cases of migration prediction. By viewing different multi-featured aspects, researchers are able to study the performances of the RF algorithm in migration prediction. In addition, this study applied the DT algorithm in forecasting the migration factor as the independent variable that triggered a migration. DL result has led to ranking the related factor based on statistical information among the advancements of human migration research through population census. Therefore, the machine learning approach is extremely important in adapting reverse migration prediction. The main purpose of this paper is to explore machine learning algorithms that were typically sufficient in the reverse migration study.

## Methodology

### Data Collection and Datasets

Machine learning algorithms were used to predict the probability of migration that were tested on data collected from the Department of Statistics Malaysia (DOSM). If the probability prediction value of a given data is above 0.5 and above, the machine learning algorithms will set the migration class to 1 otherwise the migration class is 0. Thus, the migration model of this research is a kind of machine-learning classification model. All the experiments were implemented with the Rapid Miner software tool in a computer with 8 GB RAM. The collected dataset consists of 105 records of migration data in Selangor, Malaysia. To implement the machine learning, the dataset was divided into training and validation with a ratio of 70:30 percentages respectively. Thus, from the 105 records, 75 of the datasets were used for machine learning training and 30 records were used for testing.



Recall the labeled data to split into training and validation set.

Split of a validation set.

Remember raw training data, i.e. without text processing or feature engineering.

Remember raw validation data, i.e. without text processing or feature

Edit Parameter List: **partitions**
The partitions that should be created.

| ratio |
|---|
| 0.7 |
| 0.3 |

Figure 1. The ratio of training and testing for all the machine learning models

Three machine learning algorithms used in this research were the Decision Tree, Random Forest and Gradient Boosted Tree. Preliminary experiments have been conducted to identify the optimal parameters of the machine learning algorithms. As for Decision Tree, the relevant parameter is the tree *maximal depth*, hence six values of the parameters have been observed, as shown in Table 1.

Table 1
*Decision Tree Optimal Perimeter*

| Maximal Depth | Error Rate |
|---|---|
| 25 | 48% |
| 15 | 48% |
| 10 | 48% |
| 7 | 48% |
| 4 | 48% |
| 2 | 48% |

It can be seen in the Table 1 that all the variant of *maximal depth* generated same error rate at 48%. Therefore, the minimum *maximal depth* (2) was used for the Decision Tree.

Table 2
*Random Forest*

| Number of Trees | Maximal Depth | Optimal Parameter |
|---|---|---|
| 20 | 2 | 10% |
| 60 | 2 | 6% |
| 100 | 2 | 6% |
| 140 | 2 | 8% |
| 20 | 4 | 2% |
| 60 | 4 | 2% |
| 100 | 4 | 5% |
| 140 | 4 | 2% |
| 20 | 7 | 3% |
| 60 | 7 | 3% |
| 100 | 7 | 3% |
| 140 | 7 | 3% |

Furthermore, Random Forest algorithm, which an extension algorithm of Decision Tree has *Number of Trees* and *maximal depth* that need to be observed. As seen in Table 2, the most ideal setting (2% error rate) has been presented by 20 *Number of Trees* and 4 *maximal depth* from the 12 configurations.

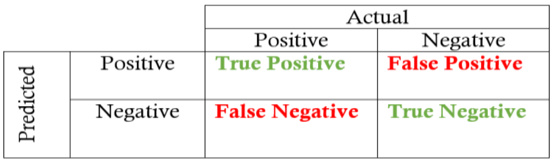
Table 3
*Gradient Boosted Tree*

| Number of trees | Maximal Depth | Learning Rate | Error rate |
|---|---|---|---|
| 30 | 2 | 0.001 | 6.3% |
| 90 | 2 | 0.001 | 6.3% |
| 150 | 2 | 0.001 | 6.3% |
| 30 | 4 | 0.001 | 6.3% |
| 90 | 4 | 0.001 | 6.3% |
| 150 | 4 | 0.001 | 6.3% |
| 30 | 7 | 0.001 | 6.3% |
| 90 | 7 | 0.001 | 6.3% |
| 150 | 7 | 0.001 | 6.3% |
| 30 | 2 | 0.01 | 6.3% |
| 90 | 2 | 0.01 | 6.3% |
| 150 | 2 | 0.01 | 6.3% |
| 30 | 4 | 0.01 | 6.3% |
| 90 | 4 | 0.01 | 6.3% |
| 150 | 4 | 0.01 | 6.3% |
| 30 | 7 | 0.01 | 6.3% |
| 90 | 7 | 0.01 | 6.3% |
| 150 | 7 | 0.01 | 6.3% |
| 30 | 2 | 0.1 | 4.8% |
| 90 | 2 | 0.1 | 4.8% |
| 150 | 2 | 0.1 | 4.8% |
| 30 | 4 | 0.1 | 4.8% |

| | | | |
|---|---|---|---|
| 90 | 4 | 0.1 | 4.8% |
| 150 | 4 | 0.1 | 4.8% |
| 30 | 7 | 0.1 | 4.8% |
| 90 | 7 | 0.1 | 4.8% |
| 150 | 7 | 0.1 | 4.8% |

Gradient Boosted Tree algorithms has three parameters. Using a low learning rate can significantly improve the performance of gradient boosting models. The effective learning rate is typically between 0.1 and 0.3. Among the 0.1 *learning rate* (Refer to Table 3), the *Number of Trees* 30 with *maximal depth* 2 has generated the lowest error rate at 4.8%.

How frequently the machine learning classification models can classify the migration into 1 or 0 can be measured by looking at the algorithm's accuracy. Out of all the data points, accuracy is the percentage migration occurrence can be detected by the machine learning. Refer to confusion matrix in Figure 2, there will be four possible cases.



| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | True Positive | False Positive |
| | Negative | False Negative | True Negative |

Figure 2: Confusion Matrix

True Positive (TP) is the actual migration occurrence that can be correctly classified as migration by the machine learning. On the other hand, False Positive (FP) is the actual migration occurrence that incorrectly classified as no migration by the machine learning. For the case of no migration in the actual case, False Negative (FN) refers to the wrongly classified as migration by the machine learning while True Negative (TN) denotes the case when machine learning can correctly classify the no migration cases. So, the formula for accuracy is calculated with Equation 1 (Brohi et. al., 2019).

$$Accuracy = \frac{TP+TN}{all\ cases} \qquad (1)$$

The classification error rate can be calculated with Equation 2.

$$Classification\ Error = \frac{FP+FN}{all\ cases} \qquad (2)$$

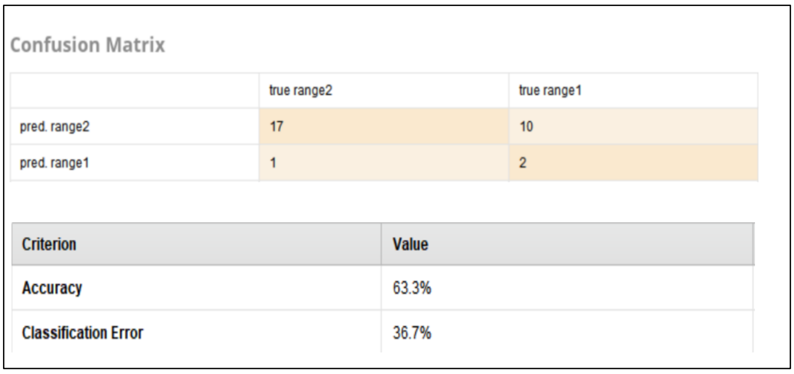
**Results and Discussion**

In this section, the results of the machine learning models are presented in two divisions. Firstly, the performance results of each machine learning algorithm are given regarding the prediction accuracy, classification of error, and processing time. Secondly, the weight of contributions of each independent variable to the reverse migration in Selangor, Malaysia.

**The Machine Learning Performances**

According to the tables below, which are tables 4, 5, and 6, show that predication range 1 represents no migration, while predication range 2 represents migration.

Table 4
*Accuracy and Classification Error in Gradient Boosted Tree*



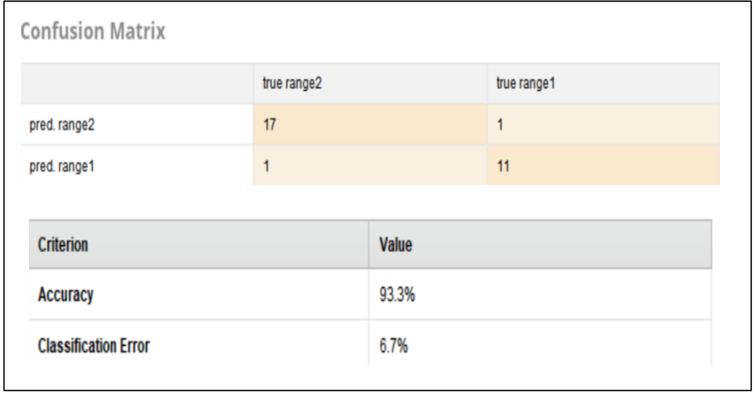
Confusion Matrix

| | true range2 | true range1 |
|---|---|---|
| pred. range2 | 17 | 10 |
| pred. range1 | 1 | 2 |

| Criterion | Value |
|---|---|
| Accuracy | 63.3% |
| Classification Error | 36.7% |

The classification error is 36.7 percent, while the accuracy of the gradient boosted tree technique is 63.3 percent, as shown in Table 4. On 30 data sets, machine learning was tested to predict reverse migration. The model accurately predicted the occurrence of reverse migration, which occurred with 17 true positives and 2 true negatives, respectively, in which there was no reverse migration.

Table 5
*Accuracy and Classification Error in Random Forest*



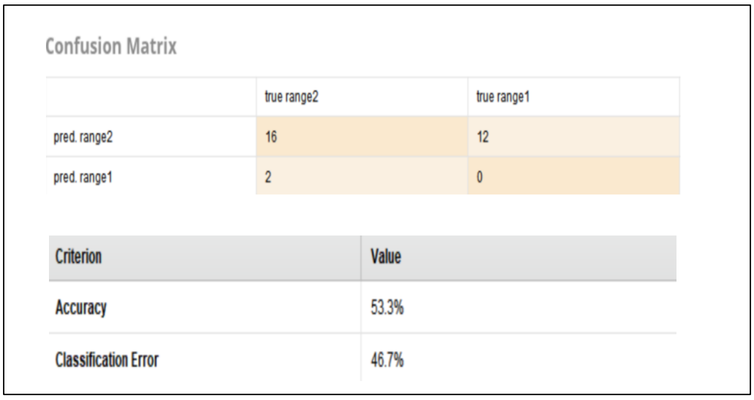
Confusion Matrix

| | true range2 | true range1 |
|---|---|---|
| pred. range2 | 17 | 1 |
| pred. range1 | 1 | 11 |

| Criterion | Value |
|---|---|
| Accuracy | 93.3% |
| Classification Error | 6.7% |

Table 5 shows, the accuracy of the Random Forest Algorithm is 93.3 %, while the classification error is 6.7 %. The accuracy results indicate that true positives were 17 and true negatives were 11. Meanwhile, false negatives and false positives both contributed to 1. The true range 1 in that table, which is a false positive, indicates that the machine learning model was mistaken to predict the occurrence of reverse migration.

Table 6
*Accuracy and Classification Error in Decision Tree*



Confusion Matrix

| | true range2 | true range1 |
|---|---|---|
| pred. range2 | 16 | 12 |
| pred. range1 | 2 | 0 |

| Criterion | Value |
|---|---|
| Accuracy | 53.3% |
| Classification Error | 46.7% |

According to Table 6, the decision tree algorithm's accuracy is 53.3 %, while the classification error is 46.7 %. The accuracy of the results indicates that true positives were 16 and true negatives were 0. False positives managed to reach 12 and false negatives reached 2. Reverse migration was correctly predicted by the decision tree model at the true range 2 with 12 true positives.

**The Correlations of Variables in the Machine Learning Models**

The significance of each independent variable in each of the machine learning models is discussed in this section. The correlation weights for each independent variable/attribute in the Gradient Boosted Tree, Random Forest, and Decision Tree models are shown in Figures 3, 4, and 5.

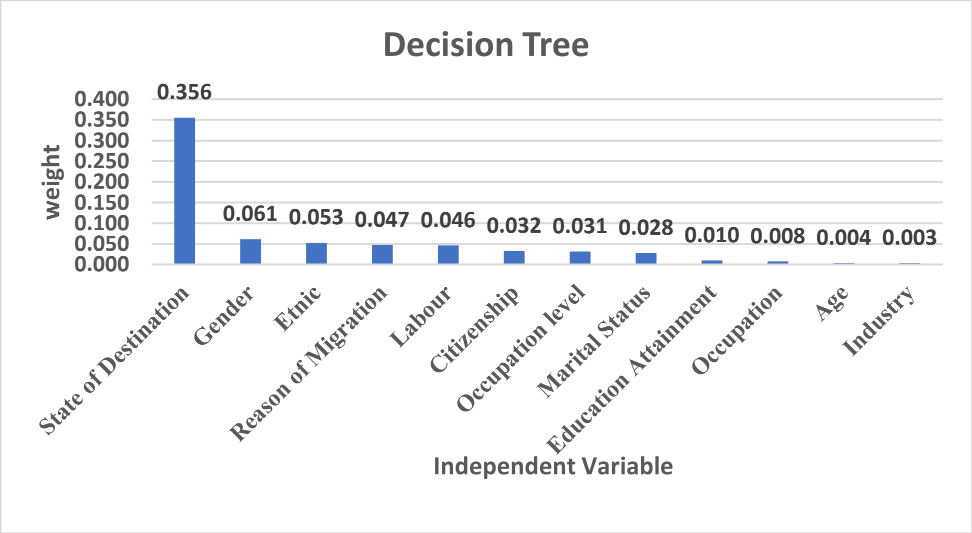


Figure 3: Weight of each independent variable in Decision Tree

As can be seen in Figure 3, States of Destination weights 0.356 which has a very strong correlation to the prediction of reverse migration in the decision tree model. In contrast, Industry has a weight of 0.003 and a very low correlation of weight prediction reverse migration in the decision tree model.

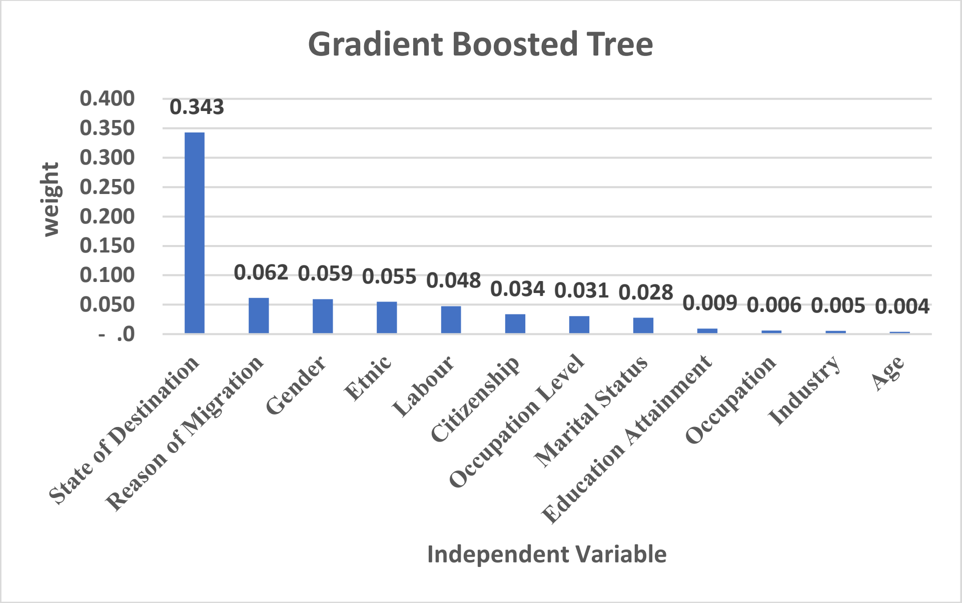


Figure 4: Weight of each independent variable in Gradient Boosted Tree

The Gradient Boosted Tree model also shows in Figure 4 that the States of Destination have a weight of 0.343 and a significant correlation. In contrast, the Age weight in the gradient-boosted tree model is 0.004 and its correlation with weight prediction reverse migration is extremely low.

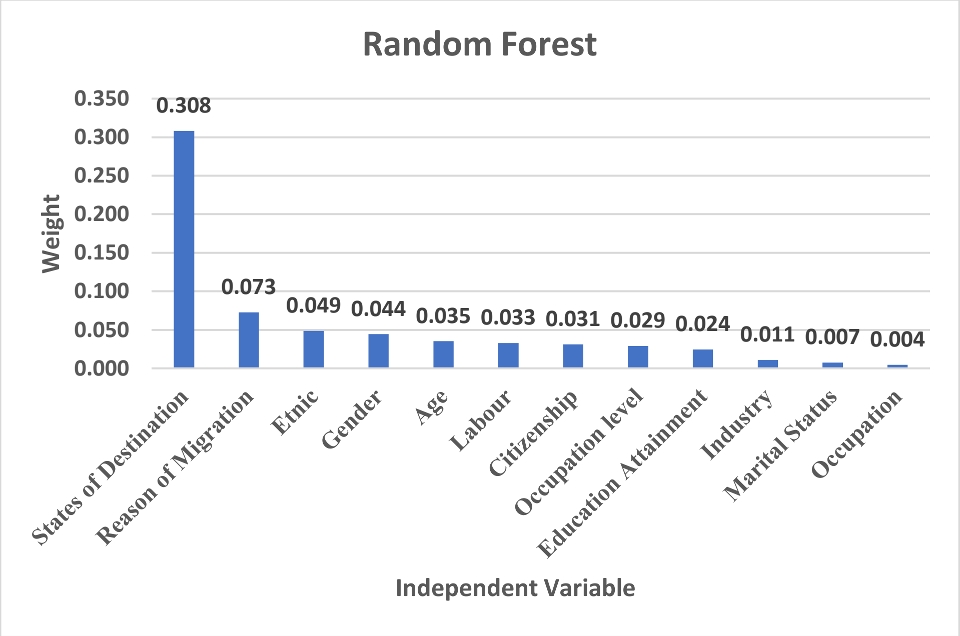


Figure 5: Weight of each independent variable in Random Forest

In Figure 5, Random Forest also demonstrates and shows the States of Destination have a very strong correlation of weight in prediction reverse migration, similar to Decision Tree and Gradient Boosted Tree models. Meanwhile, Occupation has a very low weight which is 0.004.

## Conclusion

This paper addresses the reverse migration factors of population mobility in predicting future migration patterns and capacity. This research focuses on DT, GBT and RF algorithms, interpretation of the finding was described with some limitations on the methodology as well as on the tested dataset. The main challenge of this study is the limited data from DOSM due to MCO restrictions. Therefore, more extensive work is needed in line with the prediction model and the robust machine-learning approaches. Thus, with the accuracy of this model, the results indicated that random forest is one of the suggested algorithms to pursue the development of machine learning models for reverse migration in Malaysia. Compared to the conventional method, this machine learning model is more effective at forecasting reverse migration. Therefore, parallel with the industrial revolution 4.0, the machine learning model can effectively predict and pursuit many significant issues in the industry.

## Acknowledgements

The authors would like to express their gratitude to the Ministry of Higher education Malaysia and Universiti Teknologi MARA for the full support of this study that was funded under FRGS grant number (FRGS/1/2019/SS06/UITM/02/16).

## 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.

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.

Hussain, N. H. M. (2015). From Kampong to City and Back Again: A Case Study of De-urbanisation in Malaysia, Unpublished Thesis, University of Auckland, NZ

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.

Mohd, T., Jamil, S., & Masrom, S. (2020). Machine learning building price prediction with green building determinant. *IAES* International *Journal of Artificial Intelligence*, 9(3), 379–386.Retrieved December 2, 2020 from https://doi.org/10.11591/ijai.v9.i3.pp379-386

United Nations, Department of Economic and Social Affairs, *Population Division. (2017).* Trends in International Migrant Stock: The 2017 Revision. (United Nations database, POP/DB/MIG/Stock/Rev.2017)