Mathematical Sciences and Informatics Journal
Vol. 4, No. 1, May 2023, pp. 49-56
http://www.mijuitmjournal.com DOI : 10.24191/mij.v4i1.22138

Tree-based Machine Learning in Classifying Reverse Migration

Azreen Anuar
Centre of Graduate Studies, Universiti Teknologi MARA, Perak Branch, Seri Iskandar Campus, Malaysia
2020864206@student.uitm.edu.my

Nur Huzeima Mohd Hussain
Department of Built Environment and Technology, Universiti Teknologi MARA, Perak Branch, Seri Iskandar Campus, Malaysia
nurhu154@uitm.edu.my

Hugh Byrd
Lincoln School of Architecture, University of Lincoln, Lincoln, United Kingdom
hbyrd@lincoln.ac.uk

Article Info

Article history:

Received Feb 16, 2023
Revised Apr 08, 2023
Accepted Apr 26, 2023

Keywords:

Tree-based machine learning
Linear-based machine learning
Reverse migration
Classification
Accuracy
Area Under the Curve

ABSTRACT

Reverse migration is an increasingly urgent issue as it is influenced by various factors such as economic crises, political turmoil, natural disasters, and the COVID-19 pandemic. Predicting reverse migration can provide valuable insights for policymakers and stakeholders to design appropriate interventions. However, there is a scarcity of studies that have applied machine learning algorithms to this problem. This paper aims to fill the gap in the literature by discussing the application of machine learning algorithms for predicting reverse migration. The study compares the performance of three types of tree-based machine learning (Decision Tree, Random Forest, Gradient Boosted Trees) with linear-based algorithms (Logistic Regression, Fast Last Margin, Generalized Linear Model). In addition to accuracy, this study also measured the area under the curve (AUC) metric, which has been seldom explored in previous research of reverse migration prediction. The findings revealed that tree-based machine learning algorithms performed slightly better than linear-based algorithms in terms of accuracy of prediction, with an improvement of approximately 1%. Based on the accuracy and AUC results, Gradient Boosted Trees is selected as the best algorithm. The findings of this study suggest that machine learning can provide valuable insights into predicting reverse migration. With the use of appropriate machine learning algorithms, policymakers and stakeholders can make more informed decisions to address the challenges posed by reverse migration.

Corresponding Author:
Nur Huzeima Mohd Hussain
Department of Built Environment and Technology, Universiti Teknologi MARA, Perak Branch, Seri Iskandar Campus, Malaysia.
nurhu154@uitm.edu.my

1. Introduction

Machine learning, a subset of artificial intelligence, has gained widespread popularity in recent years due to its ability to analyze vast amounts of data and identify patterns and insights that might be difficult for humans to detect. In healthcare, for example, machine learning algorithms are being used to analyze patient data[1]. In finance, machine learning is being used to detect tax avoidance [2] and financial decision-making[3]. These are just a few examples of how machine learning is being applied across a range of domains, highlighting its potential to revolutionize various industries.

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 gained significant popularity in recent years due to their ability to handle non-linear relationships and interactions between variables. Despite this, there are limited studies that compare the

PENERBIT PRESS UNIVERSITI TEKNOLOGI MARA

performance of linear and tree-based models, particularly in the context of reverse migration prediction.

The phenomenon of reverse migration has gained more attention in recent years, particularly in the wake of 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[4]. 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[5]. Accurately 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 this problem[6]. Thus, the present study aims to address this research gap by comparing the performance of linear-based and tree-based machine learning algorithms in predicting reverse migration.

The contributions of this study are two-folds. First, the findings provide valuable insights for policymakers and stakeholders in addressing the problem of reverse migration. Second, the comparison between tree-based and linear-based machine learning algorithms on both accuracy and area under the curve (AUC) performance metrics. AUC is a better evaluation metric than accuracy because it takes into account the overall performance of a model across different thresholds of classification. AUC measures the model's ability to correctly classify both positive and negative cases, regardless of the threshold used for classification. Accuracy, on the other hand, only measures the percentage of correct predictions made by the model. It is sensitive to the class imbalance problem, where the majority class can dominate the prediction results and lead to misleading accuracy scores. Therefore, AUC is a better metric to use when the class distribution is imbalanced or when the cost of false positives and false negatives is not equal. It provides a more comprehensive view of the model's performance and is less affected by class distribution than accuracy.

2. Literature Review

Tree-based machine learning[7],[8] is a type of algorithm that constructs decision trees to make predictions or classify data. Random Forest, Gradient Boosted Trees, and Decision Trees are examples of tree-based machine learning algorithms. Random Forest is an ensemble of decision trees, where each tree is constructed using a subset of the training data and a random subset of features[9]. Gradient Boosted Trees is another ensemble method, where each tree is constructed to correct the mistakes of the previous tree[8]. Decision Trees are single trees that recursively split the data based on the most informative feature until the tree reaches the maximum depth or a stopping criterion [8]. Decision trees are a flowchart-like structure, where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents a class label or a prediction.

TPOT (Tree-based Pipeline Optimization Tool) is a machine learning algorithm that is also based on the tree paradigm[10],[11]. However, TPOT is an automated machine learning tool that uses Genetic Programming to optimize the pipeline of the machine learning model. The goal of TPOT is to automate the process of selecting the best pipeline and hyperparameters for a given dataset. While TPOT has shown promising results in other applications, it requires extensive work to be used effectively and was therefore not included in this study.

Linear machine learning is a class of algorithms that are widely used for modeling relationships between a dependent variable and one or more independent variables. These algorithms are based on a linear equation that relates the input variables to the output variable. Some of the commonly used linear machine learning algorithms include linear regression, logistic regression, and generalized linear models. These algorithms are simple to understand, easy to implement, and fast to train on large datasets. However, they have certain limitations in terms of their predictive performance, especially when the relationship between the input and output variables is nonlinear or complex. To overcome these limitations, tree-based machine learning algorithms have been developed, which can handle nonlinearity and complex relationships between input and output variables.

Several research have investigated the use of machine learning for human migration domain but limited that deal with reverse migration. For example, researcher in [12] proposed a deep learning approach for modeling return migration patterns in the Netherlands. The study used several

variables, including demographic and socio-economic factors, to predict the likelihood of return migration for different regions in the country. In previous studies, Random Forest has been applied to mine large-scale human mobility data to predict long-term crime patterns [13]. In [14], researchers examined the challenges associated with human mobility prediction and discussed the potential of deep learning techniques to address these challenges. To predict the staying time of international migrants, a hybrid model of Particle Swarm Optimization and Support Vector Machine was proposed in [15]. These studies demonstrate the potential of machine learning for predicting reverse migration and identifying the key factors influencing migration decisions.

In the context of binary classification models such as reverse migration, accuracy, classification error, sensitivity, and specificity are the commonly used performance metrics. However, the AUC is considered a more robust measure of model performance[8]. AUC is calculated as the area under the receiver operating characteristic (ROC) curve. ROC is a chart that shows how well a binary classification model performs at different threshold levels for determining which class an input belongs to. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The ROC curve shows the trade-off between sensitivity and specificity, and a classifier with high accuracy will have an ROC curve that hugs the upper left corner of the plot. AUC provides an evaluation of the model's ability to distinguish between positive and negative classes across a range of possible classification thresholds. AUC ranges from 0 to 1, with 0.5 indicating random guessing and 1 representing perfect classification or discrimination between the two classes. A high AUC value indicates that the model has good predictive performance across a range of classification thresholds, while a low AUC value indicates poor performance.

3. Methodology

3.1 The dataset

The dataset is a set of 104 secondary data provided by the Department of Statistics Malaysia (DOSM) on the migration of Malaysian peoples in the year of 2018. Out of the entire dataset, 63 instances were allocated for training, while the remaining 41 instances were reserved for testing. To ensure that the distribution of independent variables in the testing data is representative of the overall dataset, the stratified sampling was used for splitting the training and testing dataset. Stratified sampling is a sampling technique where the population is divided into subgroups or strata based on one or more variables, and then random samples are taken from each stratum[16]. In this way, the sample is more representative of the overall population. RapidMiner software is the platform to process the dataset.

The dependent variable (DV) for the prediction model is class of reverse migration either 1 or 0 for presenting migration and no migration respectively. The independent variables (IVs) are given in Figure 1 that also presents the correlation coefficients of each IV to the DV based on Pearson Correlation test run in the RapidMiner.

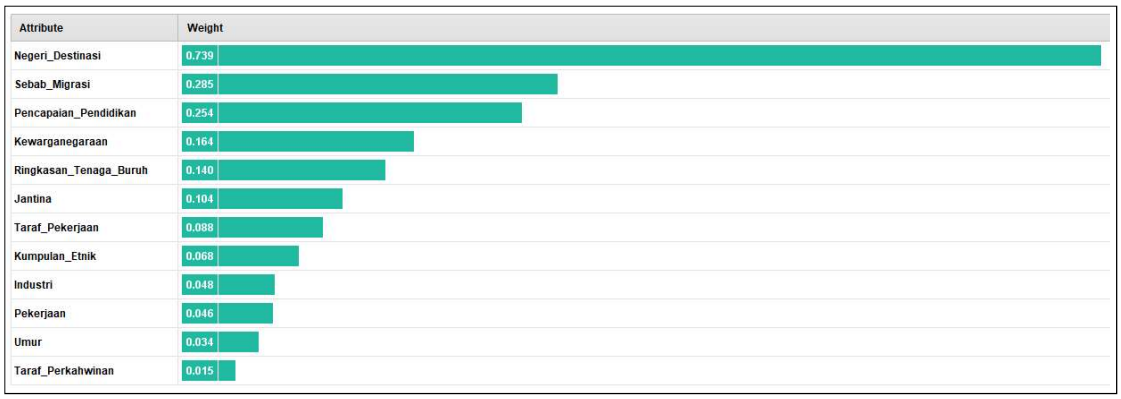


Figure 1. Correlation between each IV to DV

Based on the correlation coefficients provided in Figure 1, it is evident that the IVs related to migration exhibit nonlinearity in their relationship with the DV. The low correlation coefficients below 0.3 suggest that there is no strong linear correlation between the migration IVs and the DV. This nonlinearity could indicate the presence of more complex relationships between the variables that may not be easily captured by linear-based machine learning. Only the *Negeri_destinasi* variable shows a strong correlation coefficient of 0.74 with the dependent variable.

However, despite the low correlation coefficients between the independent variables and the dependent variable, it is still worthwhile to include them in the machine learning model. While they may not individually have a strong impact on predicting reverse migration, they may contain valuable information that can contribute to the overall accuracy of the model. Furthermore, machine learning algorithms are capable of identifying complex and non-linear relationships between variables that may not be captured by simple correlation coefficients.

3.2 Machine learning algorithm

The study used three types of tree-based machine learning (Decision Tree, Random Forest, Gradient Boosted Trees) and three types of linear-based algorithms (Logistic Regression, Fast Last Margin, Generalized Linear Model). The implementation platform for all the algorithms is RapidMiner. AutoModel was used as the preliminary for determining the algorithm specifications. AutoModel in RapidMiner is an automated machine learning tool that simplify the machine learning process by automating tasks such as algorithm selection, and hyperparameter tuning based on a combination of heuristics. Table 1 displays the configuration that achieved the highest accuracy results, as determined by the AutoModel.

Table 1. Configuration of parameters

Algorithm	Optimal Parameters
Fast Large Margin	C=1000
Decision Tree	Maximal depth=4
Random Forest	Number of trees=100 Maximal depth=4
Gradient Boosted Trees	Number of trees=30 Maximal depth=2 Learning rate=0.001

For Fast Large Margin, the value of C was set to 1000, which yielded the best accuracy. As for Decision Tree, a maximal depth of 4 was found to be the best setting. Random Forest, on the other hand, produced the highest accuracy when the number of trees was set to 100 and the maximal depth was set to 4. Finally, Gradient Boosted Trees were optimized with a number of trees of 30, maximal depth of 2, and learning rate of 0.001. These settings were found to be the most suitable for predicting reverse migration based on the dataset used in this study. Based on the configuration in Table 1, the models for developing the regression classification were constructed in RapidMiner manually, as depicted in Figure 2. The same processes were also used for the linear based machine learning that replaced the three algorithms with Logistic Regression, Fast Large Margin and Generalized Linear Model.

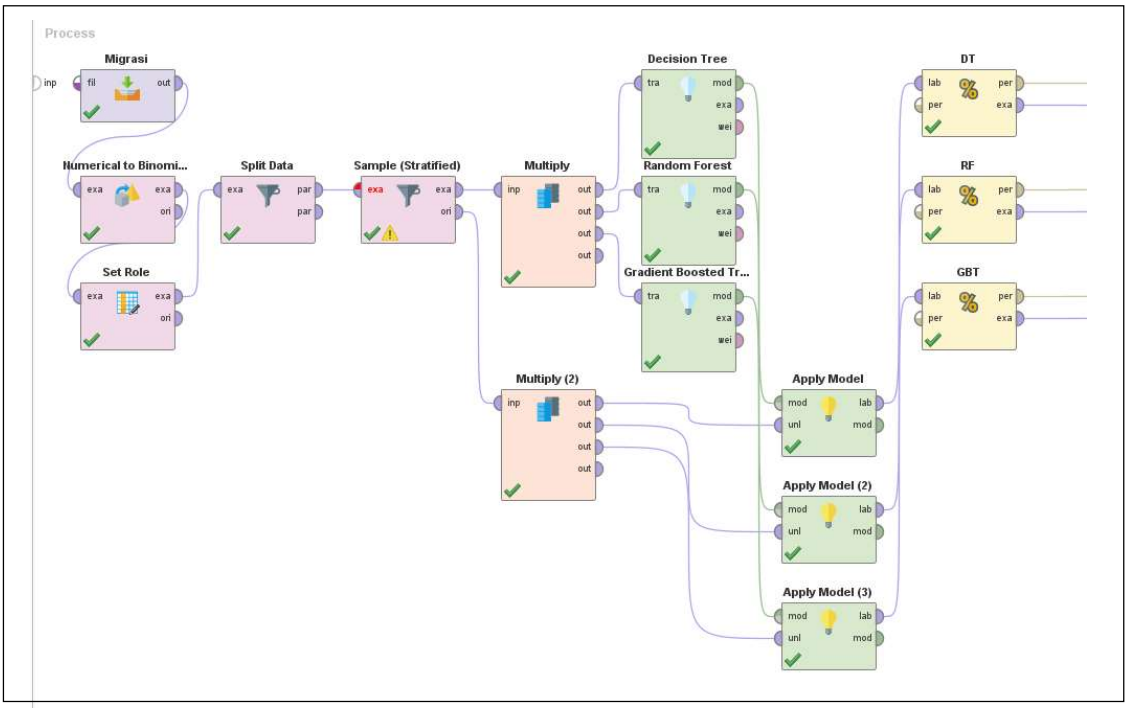


Figure 2. Correlation between each IV to DV

4. Results and Discussion

Table 2 list the accuracy and AUC results from the testing dataset. Based on the preliminary results, both Logistic Regression and Random Forest models produced 100% accuracy on the testing dataset. However, it is highly unlikely to have a perfect model in real-world scenarios, and this raises suspicion about the model's validity. Therefore, we excluded these models from further analysis to avoid overfitting and to ensure that the selected model is generalizable to new data. Instead, we focused on the models that had lower accuracy scores, such as Fast Large Margin, Generalized Linear Model, Decision Tree, and Gradient Boosted Trees, as they may have more room for improvement and could potentially provide a more realistic and robust model for the problem at hand.

Table 2. Performance results

Algorithm	Testing accuracy (%)	AUC of migration class
Fast Large Margin	90.5	0.93
Generalized Linear Model	90.5	0.92
Decision Tree	92.1	0.50
Gradient Boosted Trees	92.1	0.96

Refer to Table 2, Fast Large Margin and Generalized Linear Model have achieved good accuracy of 90.5%. Additionally, Fast Large Margin has an AUC score of 0.93, while Generalized Linear Model has an AUC score of 0.92. As both of these algorithms have AUC scores above 0.5, it indicates that they perform better than random guessing.

Decision Tree has an accuracy of 92.1%, which is slightly higher than Fast Large Margin and Generalized Linear Model. However, Decision Tree AUC score is only 0.5, which is the worst possible value. This suggests that Decision Tree may not be the best algorithm for this particular problem.

Gradient Boosted Tree also has an accuracy of 92.1%, which is the same as Decision Tree. However, its AUC score is much higher at 0.96, indicating that it is better at classifying between the two classes of reverse migration. Furthermore, Figure 3 and Figure 4 present the tree models of the Decision Tree and Gradient Boosted Trees.

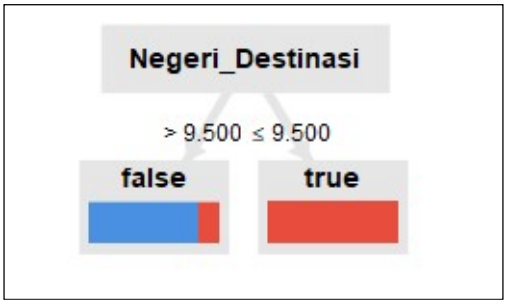


Figure 3. The tree model generated from Decision Tree

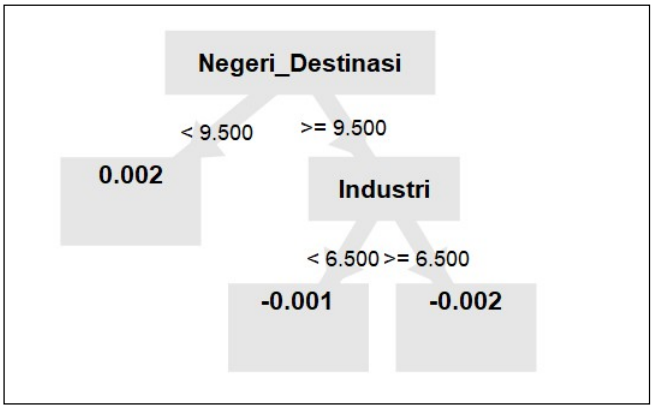


Figure 4. The tree model generated from the Random Forest

Based on the tree models, it was found that the *Negeri_Destinasi* had the highest feature importance for both algorithms. In addition, the type of *Industri* factor was the second most important contributor in the Gradient Boosted Trees. Further analysis and investigation of these factors could provide valuable insights for the reverse migration problem.

5. Conclusion

The research found no significant difference in performance between the tested tree-based and linear-based machine learning algorithms. This finding suggests that both types of algorithms can be useful for similar classification tasks. However, it is important to note that the AUC results revealed that the Decision Tree algorithm needs further investigation due to its poor performance in this metric. Further research is needed to confirm these findings with larger datasets and more diverse classification tasks. Within the scope of this works, these findings suggest that both types of algorithms can be useful for similar classification tasks, but careful consideration should be given to the selection of the appropriate algorithm based on the specific requirements of the task.

The benefits of this study include providing insights into the performance of different machine learning algorithms for classification tasks. The findings can be used to guide the selection of appropriate algorithms for similar classification tasks, potentially saving time and resources. Additionally, the study identified the features importance in the machine learning algorithms thus Further analysis and investigation of these factors could provide valuable insights for the reverse

migration field and help policy-makers and stakeholders to make data-driven decisions to improve the effectiveness of their strategies and programs.

Acknowledgements
The authors appreciate the financial support provided for this project through the FRGS grant (FRGS/1/2019/SS06/UITM/02/16) from the Malaysian Ministry of Higher Education and Universiti Teknologi MARA (UITM).

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

References

[1] A. H. Attia and A. M. Said, "Brain seizures detection using machine learning classifiers based on electroencephalography signals: a comparative study," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, pp. 803–810, 2022, doi: 10.11591/ijeecs.v27.i2.pp803-810.

[2] R. A. Rahman, S. Masrom, and N. Omar, "Tax avoidance detection based on machine learning of malaysian government-linked companies," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, 2019, doi: 10.35940/ijrte.B1083.0982S1119.

[3] P. H. Damia Abd Samad, S. Mutalib, and S. Abdul-Rahman, "Analytics of stock market prices based on machine learning algorithms," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 2, pp. 1050–1058, 2019, doi: 10.11591/ijeecs.v16.i2.pp1050-1058.

[4] M. Micevska, "Revisiting forced migration: A machine learning perspective," *Eur. J. Polit. Econ.*, vol. 70, p. 102044, 2021, doi: 10.1016/j.ejpoleco.2021.102044.

[5] M. Caulfield, J. Bouniol, S. J. Fonte, and A. Kessler, "How rural out-migrations drive changes to farm and land management: A case study from the rural Andes," *Land use policy*, vol. 81, pp. 594–603, 2019, doi: https://doi.org/10.1016/j.landusepol.2018.11.030.

[6] C. Robinson and B. Dilkina, "A machine learning approach to modeling human migration," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018, pp. 1–8. doi: 10.1145/3209811.3209868.

[7] C. Kern, T. Klausch, and F. Kreuter, "Tree-based machine learning methods for survey research," in *Survey research methods*, 2019, vol. 13, no. 1, p. 73.

[8] C. Halimu, A. Kasem, and S. H. S. Newaz, "Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification," in *Proceedings of the 3rd international conference on machine learning and soft computing*, 2019, pp. 1–6.

[9] A. V Joshi, "Decision Trees," *Machine Learning and Artificial Intelligence*. Springer, pp. 53–63, 2020.

[10] 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.

[11] A. S. A. Rahman, S. Masrom, R. A. Rahman, and R. Ibrahim, "Rapid Software Framework for the Implementation of Machine Learning Classification Models," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 11, pp. 8–18, 2021, doi: 10.46338/ijetae0821_02.

[12] J. Li and Y. Chen, "A deep learning method for solving third-order nonlinear evolution equations," *Commun. Theor. Phys.*, vol. 72, no. 11, p. 115003, 2020, doi: 10.1371/journal.pone.0232414.

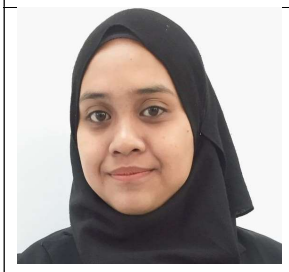
[13] C. Kadar and I. Pletikosa, "Mining large-scale human mobility data for long-term crime prediction," *EPJ Data Sci.*, vol. 7, no. 1, pp. 1–27, 2018, doi: 10.1140/epjds/s13688-018-0150-z.

[14] M. Luca, G. Barlacchi, B. Lepri, and L. Pappalardo, "A survey on deep learning for human mobility," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1–44, 2021, doi: 10.1145/3485125.

[15] P. K. Shukla *et al.*, "A novel machine learning model to predict the staying time of international migrants," *Int. J. Artif. Intell. Tools*, vol. 30, no. 02, p. 2150002, 2021, doi: 10.1142/S0218213021500020.

[16] R. Iliyasu and I. Etikan, "Comparison of quota sampling and stratified random sampling," *Biom. Biostat. Int. J. Rev*, vol. 10, pp. 24–27, 2021.

Biography of all authors

Picture	Biography	Authorship contribution
	Azreen Anuar received the Bachelor degree in Quantity Surveying (Hons.) from UiTM Perak in 2019. She has received the First-class degree and showed great achievement throughout her study. Azreen is working at the QS firm while pursuing Master degree as part-time student in UiTM Perak. Her research interest is technologically advancement namely the digital appliances, computational initiatives and machine learning. She can be contacted at email: 2020864206@student.uitm.edu.my.	Design the research work, data collection, data analysis and drafting article.
	Nur Huzeima Mohd Hussain Graduated with PhD in Architecture in 2015 from The University of Auckland, New Zealand after obtained MSc in Landscape Architecture in 2004 and Bachelor's Degree of Science in Housing, Building & Planning (Architecture) in 2002 from University Science of Malaysia. Her field(s) of interest focuses on Landscape Sociology, Green Initiatives and Innovative, Subsistence Living and Cultural Studies. She can be contacted at email: nurhu154@uitm.edu.my.	Design and supervising the research work and camera ready.
	Hugh Byrd is a Professor of Architecture in Faculty of Art Architecture and Design. He obtained a PhD in 1981 and became a registered architect in the UK in 1983. His research interests are in the future form of buildings and cities around the world as we enter an era characterised by resource depletion and climate change. He can be contacted at hbyrd@lincoln.ac.uk.	Supervising the article writing and final checking.