Sains Malaysiana 49(2)(2020): 447-459
http://dx.doi.org/10.17576/jsm-2020-4902-24

Ensemble Learning for Multidimensional Poverty Classification

(Pembelajaran Ensembel untuk Pengelasan Kemiskinan Pelbagai Dimensi)

Azuraliza Abu bakar*, Rusnita Hamdan & Nor Samsiah Sani

ABSTRACT

The poverty rate in Malaysia is determined through financial or income indices and measurements. As such, periodic measurements are conducted through Household Expenditure and Income Survey (HEIS) twice every five years, and subsequently used to generate a Poverty Line Income (PLI) to determine poverty levels through statistical methods. Such uni-dimensional measurement however is unable to portray the overall deprivation conditions, especially based on the experience of the urban population. In addition, the United Nation Development Programme (UNDP) has introduced a set of multi-dimensional poverty measurements but is yet to be applied in the case of Malaysia. In view of this, a potential use of Machine Learning (ML) approaches that can produce new poverty measurement methods is therefore of interest, which must be triggered by the existence of a rich database collection on poverty, such as the eKasih database maintained by the Malaysian Government. The goal of this study was to determine whether ensemble learning method (random forest) can classify poverty and hence produce multidimensional poverty indicator compared to based learner method using eKasih dataset. CRoss Industry Standard Process for Data Mining (CRISP-DM) methods was used to ensure data mining and ML processes were conducted properly. Beside Random Forest, we also examined decision tree and general linear methods to benchmark their performance and determine the method with the highest accuracy. Fifteen variables were then rank using varImp method to search for important variables. Analysis of this study showed that Per Capita Income, State, Ethnic, Strata, Religion, Occupation and Education were found to be the most important variables in the classification of poverty at a rate of 99% accuracy confidence using Random Forest algorithm.

Keywords: Machine learning; multidimensional poverty; random forest

ABSTRAK

Kadar kemiskinan di Malaysia ditentukan melalui pengukuran perspektif kewangan atau pendapatan. Pengukuran berkala dilakukan melalui Bancian Perbelanjaan Rumah dan Penyiasatan Pendapatan (HEIS) dua tahun sekali digunakan untuk menghasilkan Paras Garis Kemiskinan (PGK) dalam menentukan tahap kemiskinan menggunakan kaedah statistik. Pengukuran uni-dimensi itu bagaimanapun tidak dapat menggambarkan keadaan kekurangan keseluruhan yang terutamanya dialami penduduk bandar. Program Pembangunan Bangsa-Bangsa Bersatu (PBB) telah memperkenalkan satu kaedah pengukuran kemiskinan pelbagai dimensi yang belum digunakan di Malaysia. Oleh itu, potensi penggunaan pendekatan Pembelajaran Mesin (ML) untuk menghasilkan kaedah pengukuran kemiskinan yang baru adalah tinggi disebabkan oleh adanya pengumpulan pangkalan data kemiskinan yang utama seperti pangkalan data eKasih yang dikendalikan oleh Kerajaan Malaysia. Tujuan kajian ini untuk membuktikan kaedah pembelajaran mesin bergabung (hutan rawak) boleh mengkelaskan kemiskinan dengan ketepatan yang tinggi dan dapat menyenaraikan indikator pelbagai dimensi kemiskinan berbanding dengan kaedah pembelajaran asas menggunakan dataset eKasih. Metod CRoss Industry Standard Process for Data Mining (CRISP-DM) digunakan untuk memastikan perlombongan data dan proses ML dijalankan dengan baik. Di samping Hutan Rawak, kami juga mengkaji pokok keputusan dan kaedah linear am untuk menanda aras prestasi mereka dan menentukan kaedah terbaik dengan ketepatan tertinggi. Lima belas pemboleh ubah disusun menggunakan kaedah varImp untuk mencari pemboleh ubah penting. Analisis kajian ini menunjukkan bahawa Pendapatan Perkapita, Negeri, Etnik, Strata, Agama, Pekerjaan dan Pendidikan didapati sebagai faktor yang paling penting dalam mengkelaskan kemiskinan pada kadar kepercayaan ketepatan 99% dengan menggunakan algoritma hutan secara rawak.

Kata kunci: Hutan rawak; kemiskinan pelbagai dimensi; pembelajaran mesin

Introduction

Poverty reduction is one of the main agenda of the United Nations Development Programme (UNDP). It begun since 1957, to ensure that the development of a country is broad to the bottom level of citizens (poor people). In 2016, Malaysia managed to reduce poverty to 0.4% as compared to 49.3% in 1970 (Prime Minister Office Malaysia 2015) (The Economic Planning Unit 2017). This showed that Malaysia has successfully eradicated poverty level to its bare minimum. However, the economic growth indices have shown not much reduction in the level of poverty of the poor, hence, there is a need for inclusiveness (World-

Bank 2013). In order to achieve inclusiveness, multidimensional poverty index was introduced by Oxford Poverty and Human Development Initiative (OPHI) and United Nation Development Programme (UNDP), which is being published annually by the Human Development Report Office from 2010 to measure poverty from various perspectives (Lucci et al. 2018).

The measurement method of poverty is crucial to the government for developing and empowering policy. As such, there is a need for a good and trusted method to establish a strong accuracy in classifying poverty. Due to this, the Malaysia poverty measurement, a poverty line income (PLI) was created to employ the use of statistic method Gini Coefficient through based on basic costs of the items (Jamil & Mat 2014). The government of Malaysia made initiative to improve the cost of living, quality of life, and wellbeing of the nation by applying multidimensional poverty index that is comparable to the relative poverty measurement approach practiced by developed countries in the Eleventh Malaysian Plan (11MP) 2016-2020 (Economic Planning Unit 2015). The initiative was aimed to precisely identify the group of lower income that is below the Bottom 40% (B40) income group. Table 1 shows the multidimensional indicator listed by government to classify poverty using the statistic method.

In 2007, the eKasih - Poverty Bank of Malaysia was developed to keep all information about poor, hard core poor and B40 income group. The B40 community in the 11MP is defined as a household with a mean monthly income of MYR2,537 (Unit Perancang Ekonomi 2015) and according to latest Household Expenditure and Income Survey (HEIS) 2016, the B40 mean income is MYR4,360 (DOSM 2017. These bulks of data may have potential knowledge to classify new poverty indicator using machine learning (ML) method.

Classification problems have been widely discussed by researches in many contexts and domain. It reflects the benefits and discovery of new technique in data analysis. Accuracy and precision in data classification is vital and has been applied in many disciplines, such as medical (Husam et al. 2017; Pavithra & Sudha 2018; Nor Samsiah et al. 2018a), meteorology (Chen et al. 2018; Doycheva et al. 2017; Natita et al. 2017; Wrzesień et al. 2019; Zhong et al. 2019), image recognition (Albashish et al. 2016; Wu et al. 2019; Yang et al. 2019; Zheng et al. 2019), customer segmentation (Adomavicius & Tuzhilin 2001; Alsac et al. 2017; Vafeiadis et al. 2015) and increasingly popular in socio-economic (poverty, household, living standard) fields. Some methods of ML that have been experimented in socio-economic domain are random forest (Sohnesen & Stender 2016; Thoplan 2014), logistic regression (Kshirsagar et al. 2017), linear regression (Sohnesen & Stender 2017), convolutional neural network (Jean et al. 2016; Perez and Azzari 2017, K-means (Deng et al. 2016; Sano & Nindito 2011) and K-nearest neighbour (Santoso & Mohammad Isa 2016). However, the success of ML in the studies discussed above led to the usage of the ML methods in developing a poverty classification model in this research.

This study attempts to investigate random forest (RF) method, which was claimed to have good performance in Sohnesen and Stender (2016)'s as well as Thoplan (2014) studies using Malaysian data. Apart from that, this study was also conducted to extract important variables from the model that can contribute to multi-dimensional poverty indicator.

TABLE 1. Multidimensional Poverty Indicator 11MP

Dimension	Indicator	Deprivation cut-off	Weight
Education	Years of schooling	All household members aged 17-60 have less than eleven years of education	1/8
Health	School attendance	Any school-aged children (aged 6-16) not schooling	1/8
	Access to health facility	Distance to health facility is more than 5 kilometres away and no mobile health facility is provided	1/8
	Access to clean water supply	Other than treated pipe water inside house and public water pipe/ stand pipe	1/8
Living Standards	Conditions of living quarters	Dilapidated or deteriorating	1/24
	Number of bedrooms	More than 2 members/room	1/24
	Toilet facility	Other than flush toilet	1/24
	Garbage collection facility	No facility	1/24
	Transportation	All members in the household do not use private or public transport to commute	1/24
	Access to basic communication tools	Does not have consistent fixed line phone or mobile phone	1/24
Income	Mean monthly household income	Mean monthly household income less than PLI	1/4

(Source from Economic Planning Unit, Malaysia 2015)

Based on this, this current study is organized as follows: Next section will discuss the related work of poverty classification, machine learning modelling and related algorithms, subsequent sections present the methods of the proposed work as well as experimental result and analysis, respectively. The final section would conclude the overall findings and suggestion for future work.

Related Work In Poverty

POVERTY MEASUREMENT IN MALAYSIA

Absolute poverty is defined as the number of people who are unable to order adequate assets to fulfill their essential needs (Mohamed Saladin et al. 2011). However, economists have concurred that poverty does not have one direct idea. Therefore, the poverty measurement approach also varies by countries.

There are many poverty measurement approach, such as monetary approach, capability approach, social exclusion and poverty participatory assessment (PPA) (Harun & Abdullah 2007). Poverty in Malaysia is often conceptualised and operationalised from the monetary approach, according to basic costs of items (Jamil & Mat 2014). The amount of money needed to fulfill basic needs is known as Poverty Line Income (PLI). Poverty occurs when the income of the household's head is lower than PLI. These measurements are revised once every two years through survey findings, Household Expenditure and Income Survey (HEIS). The PLI or commonly known as the poverty threshold in Malaysia is determined by the Economic Planning Unit (EPU) of the Prime Minister's Department. Currently, PLI in Peninsular Malaysia is MYR960, Sabah MYR1,180 and Sarawak MYR1,020 (Jabatan Perangkaan Malaysia 2017). Malaysia uses Gini Coefficient (Jabatan Perangkaan Malaysia 2017) as the main method for measuring poverty level. Table 2 shows the PLI of 2016.

POVERTY CLASSIFICATION USING MACHINE LEARNING

Several studies have shown that random forest (RF) could contribute better prediction for poverty. Sohnesen and Stender (2017) experimented using RF in six countries, which are Ethiopia, Malawi, Uganda, Albania, Tanzania, and Rwanda where the study found that RF is more accurate than multi imputation (MI) method using Stata. RF used 25 variables with highest importance score rather than MI and selected 81 to 132 variables. This small RF model leads to improved accuracy in four out of the six countries.

McBride and Nichols (2016) analysed RF's performance and compared the result with the existing regression-based models for developing proxy-means-test targeting models. The assessment was created for United States Agency for International Development (USAID) to investigate out-of-sample accuracy in three countries, which are Bolivia, Timor-Leste, and Malawi. In which they noticed that quantile RF is not considerably higher at predicting the economic condition standing of households McBride. Thus, it concluded that RF considerably improves out-of-sample performance by 2-18 percent. Even if quantile RF is higher at properly estimating a poor house as poor, it still has higher wrong classification of a non-poor houses to be poor.

Bambang Widjanarko and Dian Seftiana (2015) noticed that an RF technique is working correctly in distinguishing qualified poor households for social insurance packages in Indonesia, whereas Thoplan (2014) found that associate application in Mauritius uses RF to identify economic condition predictors and found out that RF predicts economic condition accurately. However, none of these studies discussed about the feature's importance.

Unlike other literatures, in the study of Nor Samsiah et al. (2018b), eight features were determined and ranked by feature selection to improve bottom 40 percent (B40) household in Malaysia. According to 11MP, B40 is a household that earns income less than RM3,855 per month (Economic Planning Unit 2015), which covers poor and hardcore poor household. The eight features selected were total income, average monthly income, income per capita, state, date of record, area, ethnic and household number (Nor Samsiah et al. 2018b). It was observed that Decision Tree (J48) performed better accuracy rather than Naïve Bayes and K-Nearest Neighbour. Feature selection defining were importance in having model higher accuracy according to Nor Samsiah et al. (2018b).

TABLE 2. Poverty Line Income (PLI) for Malaysia, 2016

Region	Strata	Household (MYR)
West Malaysia	Urban	970
	Rural	880
Sabah/W.P. Labuan	Urban	1170
	Rural	1220
Sarawak	Urban	1070
	Rural	940

METHODS

This study employs CRoss Industry Standard Process for Data Mining (CRISP-DM) methods, to give comprehensive instructions and procedures for applying data mining algorithms in order to solve real-world problems. Figure 1: Phases of CRISM-DM Methodology shows the six steps of data mining methods; Business Understanding, Data Understanding, Data Preparation, Model Development, Model Evaluation, and Deployment (Wirth 2000).

BUSINESS UNDERSTANDING

Conversely, understanding the objective and requirement of business/domain may lead to identifying problem of certain data-mining task. For many years, Malaysia used census data to determine the level of poverty income, current poverty status and aid distributions (Siwar & Yusof 1997). This census involved huge government expenditure, manpower and time consuming. Information such as household demographic, income, occupation, health and members of the household have been kept in databases without any further analysis. By knowing the capabilities of data mining in prediction and classification, these databases can be explored to discover new knowledge of poverty classification.

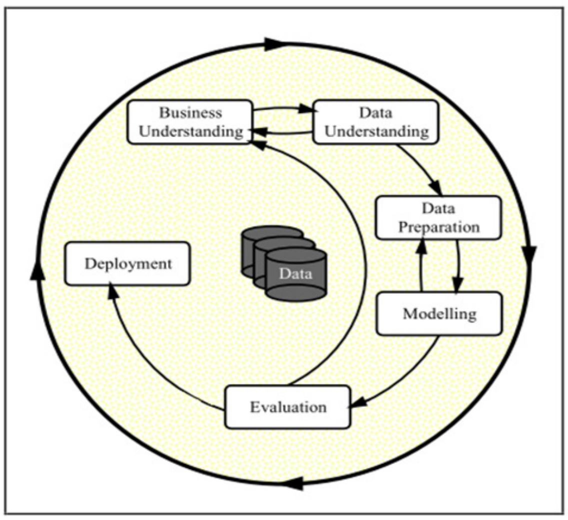


FIGURE 1. Phases of CRISM-DM Methodology
(Source from Wirth 2000)

DATA UNDERSTANDING

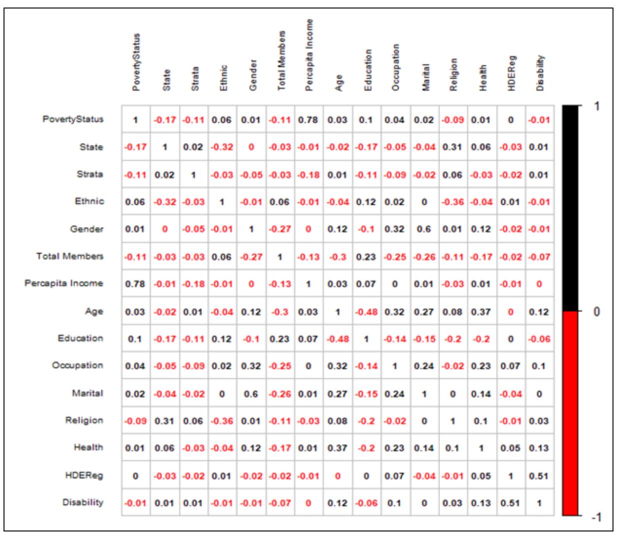
In this study, data are obtained from the Information Coordination Unit, Prime Minister Department (ICU JPM) known as eKasih for the year 2017. A total of 196,650 observations and 24 variables were used; where 2 variables represent household information, 2 variables represent income, 3 variables represents health information, 9 variables represent the location of household and others represent household demographic. Out of these 24 variables, 15 variables were selected based on literature review. Detail information about eKasih dataset can be seen in Table 3.

DATA PREPARATION

In this dataset, there are 1,105 missing values. All these missing values occur in 3 variables which are; per capita income, health and HDEReg. According to literature review, per capita income is one of the main variables acting as predictor in poverty classification. Thus, the subject matter expert suggests replacing missing value for per capita income with zero. For another two variables health and HDEReg, NA imputation was imposed. All data type in dataset was converted to numeric for modelling purposes. Description of before and after pre-processing data are shown in Table 4.

In order to have a better understanding about the dataset after pre-processing, exploratory analysis was conducted to see the correlations among the variables using Pearson's Correlation technique. The variables were plotted to check if there are a strong collinearity. According to Pearson's, 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation between two variables (Laradji et al. 2014).

As shown in Figure 2, it was observed that not all variables were correlated. Variables that have positive correlation with poverty status as well as strong relationship are; per capita income, education, ethnicity, occupation, age, marital status, gender and health. These strong positive correlation means for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. While negative correlation variables are; disability, religion, strata, total members and state. It means for every positive increase in one variable, there is a negative decrease on a fixed proportion in the other. However, HDEReg has zero correlation with poverty status, which means for every increase, there is not a positive or negative increase.



	PovertyStatus	State	Strata	Ethnic	Gender	Total Members	Percapita Income	Age	Education	Occupation	Marital	Religion	Health	HDEReg	Disability
PovertyStatus	1	-0.17	-0.11	0.06	0.01	-0.11	0.78	0.03	0.1	0.04	0.02	-0.09	0.01	0	-0.01
State	-0.17	1	0.02	-0.32	0	-0.03	-0.01	-0.02	-0.17	-0.05	-0.04	0.31	0.06	-0.03	0.01
Strata	-0.11	0.02	1	-0.03	-0.05	-0.03	-0.18	0.01	-0.11	-0.09	-0.02	0.06	-0.03	-0.02	0.01
Ethnic	0.06	-0.32	-0.03	1	-0.01	0.06	-0.01	-0.04	0.12	0.02	0	-0.36	-0.04	0.01	-0.01
Gender	0.01	0	-0.05	-0.01	1	-0.27	0	0.12	-0.1	0.32	0.6	0.01	0.12	-0.02	-0.01
Total Members	-0.11	-0.03	-0.03	0.06	-0.27	1	-0.13	-0.3	0.23	-0.25	-0.26	-0.11	-0.17	-0.02	-0.07
Percapita Income	0.78	-0.01	-0.18	-0.01	0	-0.13	1	0.03	0.07	0	0.01	-0.03	0.01	-0.01	0
Age	0.03	-0.02	0.01	-0.04	0.12	-0.3	0.03	1	-0.48	0.32	0.27	0.08	0.37	0	0.12
Education	0.1	-0.17	-0.11	0.12	-0.1	0.23	0.07	-0.48	1	-0.14	-0.15	-0.2	-0.2	0	-0.06
Occupation	0.04	-0.05	-0.09	0.02	0.32	-0.25	0	0.32	-0.14	1	0.24	-0.02	0.23	0.07	0.1
Marital	0.02	-0.04	-0.02	0	0.6	-0.26	0.01	0.27	-0.15	0.24	1	0	0.14	-0.04	0
Religion	-0.09	0.31	0.06	-0.36	0.01	-0.11	-0.03	0.08	-0.2	-0.02	0	1	0.1	-0.01	0.03
Health	0.01	0.06	-0.03	-0.04	0.12	-0.17	0.01	0.37	-0.2	0.23	0.14	0.1	1	0.05	0.13
HDEReg	0	-0.03	-0.02	0.01	-0.02	-0.02	-0.01	0	0	0.07	-0.04	-0.01	0.05	1	0.51
Disability	-0.01	0.01	0.01	-0.01	-0.01	-0.07	0	0.12	-0.06	0.1	0	0.03	0.13	0.51	1

FIGURE 2. Variables correlations

TABLE 3. Dataset description

eKasih 2017 dataset 196,650 obs. of 15 variables			
Variables	Description	Data type	Examples of variables values
Poverty status	Poverty categories of household head	chr	‘P: Poor’, ‘HP: Hardcore Poor’ ...
State	Household state of living	chr	‘Sarawak’ ‘Kelantan’ ‘Kelantan’ ‘Sabah’ ...
Strata	Type of human settlement	chr	‘2:Rural’ ‘1:Urban’ …
Ethnic	Category of people	chr	‘Iban’ ‘Melayu’ ‘Melayu’ ‘Dusun’ ...
Gender	Sex of household head	chr	‘1: Male’ ‘2: Female’ …
Total Members	Families	num	4 5 11 4 7 7 8 10 7 5 ...
Per capita income	Total of monthly income divided by total of members	num	159 180 182 234 150 ...
Age	Age of household head (in years)	num	65 59 57 56 55 54 52 51 51 51 ...
Education	Level of study	chr	‘Primary’ ‘Secondary’ ‘Post-Secondary’ ‘Higher Education’ ‘None’
Occupation	Employment	chr	‘Self-Employed’ ‘Wage earner’ ‘Unemployed’
Marital	Marital	chr	‘Married’ ‘Single’ ‘Divorced’ ‘Widow’ ...
Religion	Believing	chr	‘Kristian’ ‘Islam’ ‘Buddhis’ ‘Ateisme’ ...
Health	Health conditions	chr	‘Good’ ‘Poor’
HDEReg	Registration as Human Deficient Effort	chr	‘Yes’ ‘No’
Disability	Type of disable	chr	‘Yes’ ‘No’

TABLE 4. Result of data preparation

eKasih 2017 dataset 196,650 observations of 15 variables				
Variables	Original data type	Originalmissing value	New data type	New missing value
Poverty Status (Class)	chr	0	num	0
State	chr	0	num	0
Strata	chr	0	num	0
Ethnic	chr	0	num	0
Gender	chr	0	num	0
TotalIR	num	0	num	0
Per capita Income	num	17	num	0
Age	num	0	num	0
Education	chr	0	num	0
Health	chr	0	num	0
Marital	chr	0	num	0
Religion	chr	0	num	0
Health	chr	1009	num	0
HDEReg	chr	79	num	0
Disability	chr	0	num	0

MODELLING

RANDOM FOREST

RF is an ensemble machine learning classifier. It consists of a collection of tree-structured classifiers *{h(x, ?k), k = 1,...}* where the *{?k}* are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x* (Breiman 2001).

The RF algorithm is bagging ensemble classifier. It runs fast and is considered to have relatively high accuracy compared to other classification algorithm (Thoplan 2014). Leo Breiman was the first to formally introduce the RFs after the bagging method which is a combination of models in view of increasing classification accuracy. RF can overcome the overfitting problem because of a large number of trees, the generalization error converges to a limiting value under the strong law of large number (Breiman 2001).

Steps of RFs algorithm are outlined as follows:

A random sample of observations is taken and subsequent bootstrap samples for other trees are taken; A subset of m variables that is much less than the total number of variables in the dataset is randomly selected using the Gini score, and thus the best split is determined; and The out-of-bag (OOB) prediction is obtained through a majority vote across trees whose observation is not included in the bootstrap sample.

Also, RF is capable of providing a ranking of variable importance. In order to evaluate the importance of a variable, Louppe et al. (2013) proposed to evaluate, for all trees in the forest, the average of an impurity decrease measure for all nodes where the variable is concerned. The variable with the largest decrease in impurity will be considered as the most important variable. This can be achieved through the Mean Decrease Gini (MDG) or the Mean Decrease Accuracy (MDA). In this paper, we focus mainly on the MDG to identify important variables.

Using the notations from Louppe et al. (2013), any mean decrease impurity measure can be mathematically represented as follows:

$$\mathrm{Imp}(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T : \upsilon(s_t) = X_m} p(t) \Delta i(s_t, t) \tag{1}$$

From (1), represents the X_m variable, N_T is the number of trees in the forest, is the variable at split s_t, p(t) is the proportion of records at node t out of the total number of records in the data and

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{2}$$

p_L represents the number of records in the left child node of *t* out of the total number of records at node *t*. For this study, we shall consider the impurity measure *i(t)* as the Gini index. The Gini index, *i(t)* is defined as follows for a node *t*:

$$i(t) = 1 - \sum_j p(j \mid t)^2 \tag{3}$$

where *j* = 1, 2 for this study representing poverty class.

DECISION TREE

The decision tree is a well-known classifier that presents the output in a tree structure. The tree represents a test on a variable, where each branch denotes an outcome of a test and each leaf at the end of the branch is the output of a class label.

The topmost node in a tree is the root node (Wu et al. 2015). Given a tuple, *X*, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is then traced from the root to a leaf node, which holds the class prediction for that tuple. Decision tree classifiers have good accuracy (Yang & Fong 2011).

Steps of this algorithm are given as follows.

Input: Data partition, which is a set of training tuples and their associated class labels; Variables list, the set of candidate variables; and Variables selection method, a procedure to determine the splitting criterion that 'best' partitions the data tuples into individual classes.

Output: A decision tree.

MODEL EVALUATION

Accuracy: It is a ratio of ((no. of correctly classified instances) / (total no. of instances)) × 100) (Nor Samsiah et al. 2018a) and it can be defined as:

$$Accuracy = \frac{(TruePositive + TrueNegative)}{(TruePositive+FalsePositive)+(FalseNegative+TrueNegative)} * 100$$

Confusion Matrix: Show the number of correct and incorrect classification of test dataset break into each class (Ahmad & Abu Bakar 2018). The confusion matrix table of principal in Table 5 can be explained as follows:

True positives (TP): There are data predicted as Class 1 and actual data are also Class 1. True negatives (TN):There are data predicted as Class 2 and actual data are also Class 2. False positives (FP): There are data predicted as Class 1 but actual data are in Class 2 (Also known as a 'Type I error'). False negatives (FN): There are data predicted as Class 2 but actual data are in Class 1 (Also known as a 'Type II error.').

TABLE 5. Confusion matrix

	Prediction	
Actual	Class 1	Class 2
Class 1	True Positive (TP)	False Negative (FN)
Class 2	False Positive (FP)	True Negative (TN)

Receiver Operating Characteristic (ROC): Is a measurement of prediction sensitivity. It is generated from test dataset by plotting the TP Rate and FP Rate. The formula for ROC is as follows (Othman et al. 2018):

$$\text{ROC for TP rate} = \frac{TP}{TP+FN}\times 100$$

$$\text{ROC for FP rate} = \frac{FP}{FP+TN}\times 100$$

Within the ROC, different threshold can be determined by the users, where it will show either the classification increases to FP or TP. Also, ROC graph is used to visualise the result. The quality of ROC is often summarized as a single number using the area under the curve (AUC), but higher AUC scores are better. Figure 3 shows the example of ROC graph.

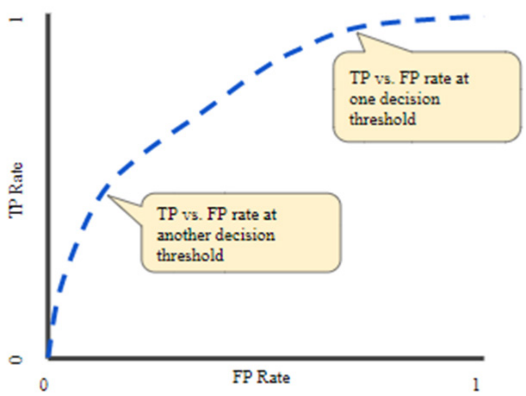


FIGURE 3. ROC Graph

Results and Knowledge Analysis

MACHINE LEARNING PERFORMANCE: ACCURACY, CONFUSION MATRIX AND ROC

The classification of poverty starts by dividing a poverty dataset into two sets (training and test set). The training set consists of a 75% sample of variables and targeted class from the dataset. While others are used as test sets. Experiment output from this dataset is discussed in this section.

Modelling poverty classification starts with RF method by setting numbers of tree (n) to grow set to 100. Poverty Status variables were selected as a class label to train the training dataset.

With n=100, confusion matrix for RF shows that 46,985 data were predicted as TP and 100,135 data as FN, which means correctly predicted. However, only 53 data were predicted as TN and 314 data as FP mean incorrectly predicted.

This gave accuracy of 99% to the model with out of bag (OOB) estimate error calculated as 0.25% within 21.88 second processing time. This small error of OOB shows fewer mistakes in the prediction of overall training sample.

According to Breiman (2001), 500 number of tree is a default value of having a good RF modelling. However, it may consume time and require high computational power. Figure 4 shows that the errors will decrease when more trees are iterated for this experiment. Green line shows the error rate decrease when the number of variables randomly samples as candidate at each split (*mtry*) is equal to 1. However, the error rate is lesser when the *mtry* is equal to 0. Hyper parameter, such as *mtry* and n can be tuning for having a better performance of model.

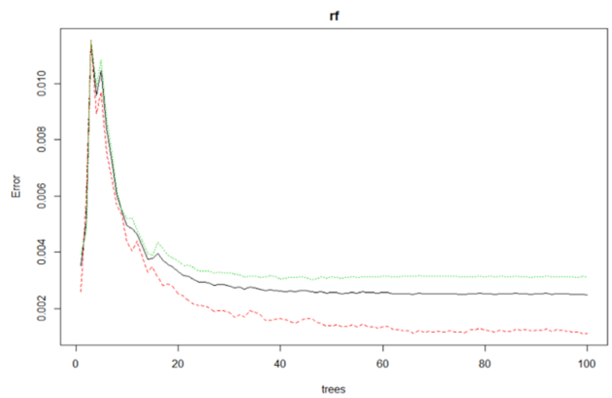


FIGURE 4. Error rate reduced when the number of trees is larger

The poverty classification model is the decision tree (J48), the based learner. The output form decision tree method is a tree, as the tree is very easy to interpret and understand especially for domain expert. Figure 5 shows the decision tree for this experiment and Table 6 summarises all eight rules extracted from the tree.

This decision tree model also can be pruned according to strata either urban or rural. The urban tree in Figure 6 and rural tree in Figure 7 simplifies the diagram to classify poverty status.

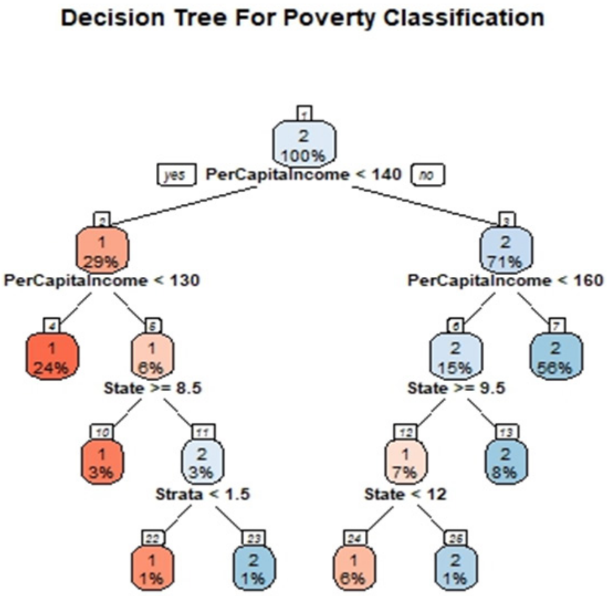


FIGURE 5. Decision Tree Diagram for poverty classification

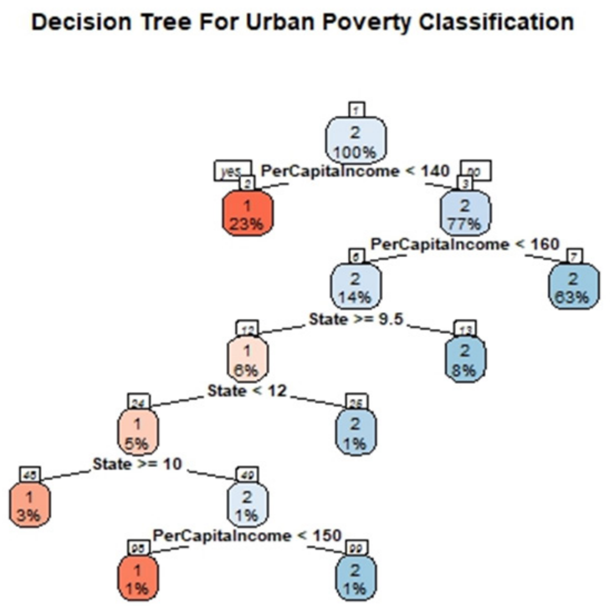


FIGURE 6. Decision Tree Diagram for urban poverty classification

From these figures, it is observed that the eight rules from the tree are divided into two main poverty class (hardcore poor and poor). Four rules were used to classify hardcore poor and four rules to classify poor.

These rules are able to translate the Malaysia Poverty Line Income (PLI) 2014 as presented in Table 7. However, the per capita income threshold is slightly higher, especially for poor status in Table 7 due to sampling error in the data set.

The confusion matrix for decision tree shown in Table 8 clearly indicates that this model is able to predict correctly 15,683 as TP and 32,642 as TN, and incorrectly predicted only 832 as FN and 6 as FP. With this very high TP and TN correctly predicted, accuracy for this model is 98% with only 2.17 s processing time. Table 8 shows the comparative analysis of accuracy, confusion matrix, processing time of the method.

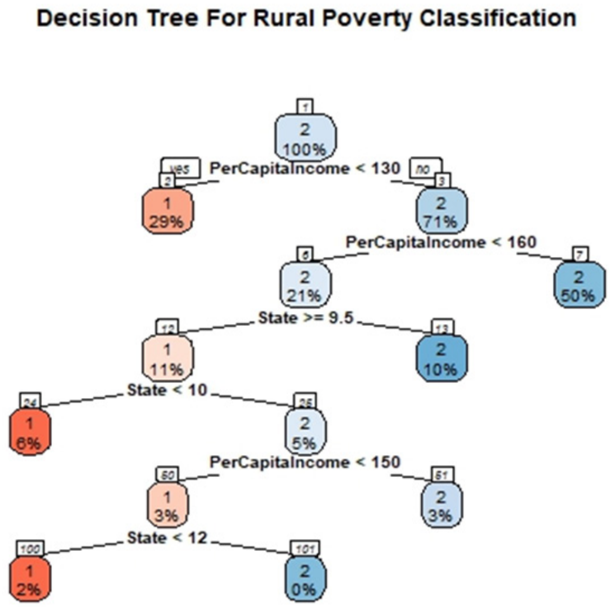


FIGURE 7. Decision Tree Diagram for rural poverty classification

Performance for both of these classifications can also be measured using ROC. ROC can show the sensitivity of the model towards correct and incorrect prediction by the model. Figure 8 shows ROC for decision tree and RF model. Both ROC model is closer to left hand and top border, representing higher accuracy and sensitivity. AUC value for RF is 0.9999, while AUC for decision tree is 0.9975. Even when the value different is quit slim, RF model performs slightly better than decision tree.

IMPORTANT VARIABLES

Determining important variable in this experiment is crucial to identify if there are variables that influence poverty classification other than income. This is important in leading us to build a multidimensional poverty indicator to classify poverty. Conversely, RF algorithm has capability to list important variables by using MDG impurity calculation. Table 9 shows the rank of important variables from the experiment. It can be observed that from out of 14 variables, per capita income, states, ethnic, strata and religion are the top five important variables in classifying poverty.

While in decision tree model, important variables are calculated using Information Gain for chosen important variable as stated in Table 10. Although Figure 5 decision tree diagram only needs three variables to identify poverty class, which are per capita income, strata and state, other variables can as well be used when the variable is not available in certain situation.

TABLE 6. Decision tree rules

No.	Rules	Poverty status class	Information
1	per capita income < MYR130 AND < MYR140	Hardcore Poor	Household with per capita income between less than MYR130, household is classify as hardcore poor status
2	per capita income < MYR140 AND >=MYR130 AND State ID >=8	Hardcore Poor	Household with per capita income is equal and more than MYR130 but less than MYR140 AND State ID is equal or more than 8, which are 8:Perlis, 9:Pulau Pinang,10:Sabah,11:Sarawak,12:Selangor,13:Terengganu,14:WP Kuala Lumpur,15: WP Labuan, 16: WP Putrajaya, household is classify as hardcore poor status
3	per capita income < MYR140 AND >=MYR130 AND State ID <8 AND Strata ID=1	Hardcore Poor	Household with per capita income is more than MYR130 but less than MYR140 AND State ID that is less than 8 are 1:Johor, 2:Kedah,3:Kelantan,4:Melaka,5:Negeri Sembilan, 6:Pahang,7:Perak AND Strata ID is 1:Urban, household is classify as hardcore poor status
4	per capita income < MYR140 AND >=MYR130 AND State ID < 8 AND Strata ID=2	Poor	Household with per capita income is more than MYR130 but less than MYR140 AND State ID is less than 8 which are 1:Johor, 2:Kedah,3:Kelantan,4:Melaka,5:Negeri Sembilan, 6:Pahang,7:Perak AND Strata ID is 2:Rural, household is classify as poor status
5	per capita income >=MYR140 AND < MYR180 AND State ID>=8 AND State ID<12	Hardcore Poor	Household with per capita income is equal or more than MYR140 but less than MYR180 AND State ID more than 8 but less than 12 which are 8:Perlis, 9:Pulau Pinang,10:Sabah,11:Sarawak, household is classify as hardcore poor status
6	per capita income > =MYR140 AND < MYR180 AND State ID>=8 AND State ID>=12	Poor	Household with per capita income is equal or more than MYR140 and less MYR180 AND State ID equal or more than 8 which are 8:Perlis, 9:Pulau Pinang,10:Sabah,11:Sarawak,12:Selangor,13:Terengganu,14:WP Kuala Lumpur,15: WP Labuan, 16: WP Putrajaya, household is classify as poor status
7	per capita income > =MYR140 AND <MYR180 AND State ID<8	Poor	Household with per capita income is equal or more than MYR140 but less than MYR180 AND State ID less than 8 which are 1:Johor, 2:Kedah,3:Kelantan,4:Melaka,5:Negeri Sembilan, 6:Pahang,7:Perak, household is classify as poor status
8	per capita income > =MYR140 AND >=MYR180	Poor	Household with per capita income equal or more MYR140, household is classify as poor status

TABLE 7. Poverty Line Income 2014

Region	Strata	Household Income (MYR)	Per Capita Income (MYR)	Household Income (MYR)	Per Capita Income (MYR)
		Poor		Hardcore Poor	
West Malaysia	Urban	940	240	580	140
	Rural	870	200	580	130
Sabah/ W.P.Labuan	Urban	1,160	260	690	150
	Rural	1,180	260	760	160
Sarawak	Urban	1,040	250	700	160
	Rural	920	240	610	150

(Source from Economic Planning Unit, Malaysia 2014)

TABLE 8. Comparative analysis of random forest and decision tree

Method	Confusion matrix			Accuracy (%)	Processing time (s)
		Predicted			
	Actual	Hardcore Poor	Poor		
Decision Tree	Hardcore Poor	15,683	832	98%	3.34s
	Poor	6	32,642		
Random Forest	Hardcore Poor	46,985	53	99%	31.64s
	Poor	314	100,135		

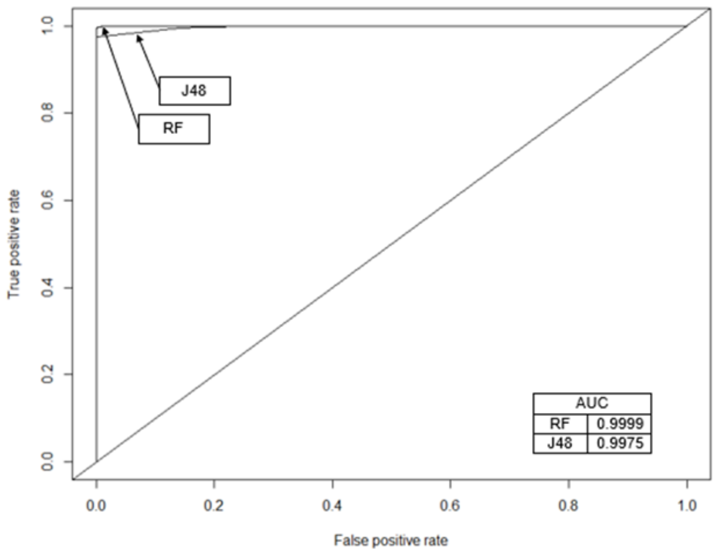


FIGURE 8. ROC Curve

TABLE 9. Ranking of important variables using RF Model

Variables	Rank
Per capita income	5.293
State	0.364
Ethnic	0.162
Strata	0.137
Religion	0.079
Total Members	0.069
Age	0.059
Occupation	0.058
Education	0.026
Marital Status	0.013
Health	0.009
Gender	0.008
Disability	0.006
HDReg	0.002

TABLE 10. Ranking of important variable using linear model

Variables	Rank
Per capita income	5.213
State	1.187
Ethnic	0.563
Religion	0.480
Strata	0.294
Occupation	0.101
Education	0.042
Total Members	0.003
Age	0.001
Gender	0.000
Marital Status	0.000
Health	0.000
HDReg	0.000
Disability	0.000

Furthermore, we also evaluated important variables using *varImp* function available in R language using linear model as comparison. It shows that per capita income, state, strata, occupation, education and ethnic have high value among others. Table 11 shows the ranking of important variables using linear model.

Since Pearson's Correlation coefficient also shows the correlation between variables displayed in Figure 2, therefore, it can indicate the important variable by listing the ascending value of each variable. The rank of important variables according to correlation coefficients is; per capita income, education, ethnic, occupation, age, marital status, gender, health, disability, religion, strata, total members, state and HDEReg. Table 12 shows the ranking of important variables using Pearson's Correlation.

In concluding the important variable result in this experiment, mean for each variable rank is calculated. The result shown in Table 13 shows that the rank for important variables in classifying poverty as follows; Per Capita Income, State, Ethnic, Strata, Religion, Occupation, Education, Age, Marital Status, Disability, Gender, HDEReg, Health and Total Members. Despite that, and median mean value for rank is also calculated in order to choose the best variable influence for the poverty classification. The median for mean rank is also calculated as 0.065. Hence, variable with mean rank equal or more than 0.065 were chosen as the most important variables for classifying poverty, given 7 variables in total. These are Per Capita Income, State, Ethnic, Strata, Religion, Occupation and Education.

TABLE 11. Ranking of important variables using linear model

Variables	Rank
Per capita income	5.679
State	1.013
Strata	0.296
Education	0.199
Occupation	0.199
Ethnic	0.141
Disability	0.068
Age	0.065
HDEReg	0.059
Marital Status	0.057
Total Members	0.053
Gender	0.050
Religion	0.048
Health	0.045

TABLE 12. Ranking of important variable using Pearson’s Correlation

Variables	Rank
Per capita income	0.785
Education	0.098
Ethnic	0.062
Occupation	0.040
Age	0.033
Marital Status	0.018
Gender	0.009
Health	0.006
HDEReg	0.001
Disability	-0.005
Religion	-0.089
Total Members	-0.105
Strata	-0.114
State	-0.170

TABLE 13. Ranking of important variables by multi method

Method	Linear Model	Random Forest Model	Decision Tree Model	Pearson's Correlation	Important Variable Rank	
Variables	Rank	Rank	Rank	Rank	Mean	Rank
Per capita income	5.679	5.293	5.213	0.785	4.243	1
State	1.013	0.364	1.187	-0.170	0.598	2
Ethnic	0.141	0.162	0.563	0.062	0.232	3
Strata	0.296	0.137	0.294	-0.114	0.153	4
Religion	0.048	0.079	0.480	-0.089	0.129	5
Occupation	0.199	0.058	0.101	0.040	0.099	6
Education	0.199	0.026	0.042	0.098	0.091	7
Age	0.065	0.059	0.001	0.033	0.039	8
Marital Status	0.057	0.013	0.000	0.018	0.022	9
Disability	0.068	0.006	0.000	-0.005	0.017	10
Gender	0.050	0.008	0.000	0.009	0.017	11
HDEReg	0.059	0.002	0.000	0.001	0.015	12
Health	0.045	0.009	0.000	0.006	0.015	13
Total Members	0.053	0.069	0.003	-0.105	0.005	14

TABLE 14. Comparison of model performance

Method	Before feature selection (14 variables)		After feature selection (7 variables)	
	Accuracy	Time	Accuracy	Time
Decision Tree	98%	3.34s	98%	1.39s
Random Forest	99%	31.64s	99%	14.97s

MODEL PERFORMANCE WITH 7 IMPORTANT VARIABLES

In the final part of this study, we also conducted experiment using seven important variables selected in previous section. The performance comparison presented in Table 14 shows that, accuracy percentage for the model remain the same. However, processing time to predict the poverty class is faster. Therefore, it is better to use these seven important variables to classify poverty rather than selecting all.

CONCLUSION

Poor and hardcore poor classifications using ML is a viable method to determine and identify the poverty class. Specifically, the RF algorithm was shown to achieve higher accuracy than a decision tree in poverty classification. Experiments also showed that seven features were identified to be important variables, according to the mean rank multi-method. These are Per Capita Income, State, Ethnic, Strata, Religion, Occupation and Education. Further experiments using these seven variables show similar accuracy results with the advantage of less ML runtime. Therefore, we conclude that dimension reduction of the variables for ML is beneficial. Furthermore, multi-dimensional poverty variables were able to classify poverty with higher accuracy compared to uni-dimensional poverty classification. The seven variables chosen are also in line with indicators outlined by the Malaysian Government in the 11th Malaysian Economic Plan. Leveraging from the impact of the recent data explosion, sectors involved with poverty management stand to gain the benefit of improved accuracy in poverty classification using ML technology. This allows poverty alleviation programs to be implemented by government agencies, in order to identify the poor and hardcore poor more effectively. Finally, aids can be given to those in need with better clarity which will reduce the issues of deprivation. It is therefore suggested that there should be further research that would be channelled towards the improvements of the RF method using greater number of trees of more variables and multi-sourced data to obtain more important variables for poverty classification. Furthermore, variables that are not selected as important in this experiment can be fused with other dataset encompassing education, expenses and health domains to gain a more useful knowledge.

ACKNOWLEDGEMENTS

Special thanks to UKM for providing the funding for this project under the grand challenge LAB40 research grant of DCP-2017-015/1 and Implementation Coordination Unit, Department of Prime Minister, Malaysia for the positive cooperation given in this research.

REFERENCES

Adomavicius, G. & Tuzhilin, A. 2001. Using data mining methods to build customer profiles. *Computer* 34(2): 74-81.

Ahmad, W.D. & Abu Bakar, A. 2018. Classification models for higher learning scholarship. *Asia-Pacific Journal of Information Technology and Multimedia* 7(2): 131-145.

Albashish, D., Sahran, S., Abdullah, A., Shukor, N.A. & Pauzi, S. 2016. Ensemble learning of tissue components for prostate histopathology image grading. *International Journal on Advanced Science, Engineering and Information Technology* 6(6): 1134-1140.

Alsac, A., Colak, M. & Keskin, G.A. 2017. An integrated customer relationship management and data mining framework for customer classification and risk analysis in health sector. *6th International Conference on Industrial Technology and Management, ICITM 2017.* pp. 41-46.

Bambang Widjanarko Otok. & Dian Seftiana. 2015. The classification of poor households in jombang with random forest classification and regression trees (RF-CART) approach as the solution in achieving the 2015 Indonesian MDGs' targets. *International Journal of Science and Research* 3(8): 1497-1503.

Chen, G.B., Li., S.S., Knibbs, L.D., Hamm, N.A.S., Cao, W., Li, T.T., Guo, J.P., Ren, H.Y., Abramson, M.J. & Guo, Y.M. 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. *Science of The Total Environment* 636: 52-60.

Deng, H.L., Zhang, L.J. & Su, W.K. 2016. Clustering the families successfully applying for minimum living standard security system based on K-means algorithm. *12th International Conference on Computational Intelligence and Security.* pp. 494-498.

DOSM. 2017. Department of Statistics Malaysia Press Release Report of Household Income and Basic Amenities Survey 2016. *Report of Household Income and Basic Amenities Survey 2016*. doi:10.1021/ja064532c.

Doycheva, K., Horn, G., Koch, C., Schumann, A. & König, M. 2017. Assessment and weighting of meteorological ensemble forecast members based on supervised machine learning with application to runoff simulations and flood warning. *Advanced Engineering Informatics* 33: 427-439.

Husam, I.S., Abuhamad, Azuraliza Abu Bakar, Suhaila Zainudin, Mazrura Sahani. & Zainudin Mohd Ali. 2017. Feature selection algorithms for Malaysian dengue outbreak detection model. *Sains Malaysiana* 46(2): 255-265.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. & Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353(6301): 790-794.

Kshirsagar, V., Wieczorek, J., Ramanathan, S. & Wells, R. 2017. Household poverty classification in data-scarce environments: A machine learning approach. *NIPS 2017 Workshop on Machine Learning for the Developing World*. http://arxiv.org/abs/1711.06813.

McBride, L. & Nichols, A. 2016. Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review* 32(3): 531-550.

Natita Wangsoh, Wiboonsak Watthayu. & Dusadee Sukawat. 2017. A hybrid climate model for rainfall forecasting based on combination of self- organizing map and analog method. *Sains Malaysiana* 46(12): 2541-2547.

Nor Samsiah Sani, Mariah Abdul Rahman, Azuraliza Abu Bakar, Shahnurbanon Sahran. & Hafiz Mohd Sarim. 2018. Machine learning approach for bottom 40 percent households (B40) poverty classification. *International*

Journal on Advanced Science, Engineering and Information Technology 8(4-2): 1698.

Nor Samsiah Sani, Illa Iza Suhana Shamsuddin, Shahnorbanun Sahran, Abdul Hadi Abd Rahman. & Ereena Nadjimin Muzaffar. 2018. Redefining selection of features and classification algorithms for room occupancy detection. *International Journal on Advanced Science, Engineering and Information Technology* 8(4-2): 1486-1493.

Othman, Zalinda, Soo Wui Shan, Ishak Yusoff. & Chang Peng Kee. 2018. Classification techniques for predicting graduate employability. *International Journal on Advanced Science, Engineering and Information Technology* 8(4-2): 1712-1720.

Pavithra, R. & Sudha, P. 2018. A survey on classification in R programming using data mining. *International Journal of Research in Engineering, Science and Management* 1(9): 401-403.

Perez, A. & Azzari, G. 2017. Poverty prediction with public landsat 7 satellite imagery and machine learning. *NIPS 2017 Workshop on Machine Learning for the Developing World*. https://arxiv.org/abs/1711.03654.

Sano, A.V.D. & Nindito, H. 2011. Application of K-Means algorithm for cluster analysis on poverty of provinces in Indonesia. *ComTech: Computer, Mathematics and Engineering Applications* 7(6): 141-150.

Santoso & Mohammad Isa Irawan. 2016. Classification of poverty levels using k-nearest neighbor and learning vector quantization. *International Journal of Computing Science and Applied Mathematics* 2(1): 8-13.

Sohnesen, T.P. & Stender, N. 2017. Is random forest a superior methodology for predicting poverty? An empirical assessment. *Poverty and Public Policy* 9(1): 118-133.

Thoplan, R. 2014. Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research* 4531(8): 252-259.

Unit Perancang Ekonomi. 2015. *Rancangan Malaysia Ke-11 (2016-2020)*. *Unit Perancang Ekonomi, Jabatan Perdana Menteri*. Kuala Lumpur: Percetakan Nasional Malaysia Berhad. http://www.epu.gov.my.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. & Chatzisavvas, K.C. 2015. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory* 55: 1-9.

Wirth, R. 2000. CRISP-DM: Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* 24959: 29-39.

Wrzesień, M., Waldemar, T., Klamkowski, K. & Rudnicki, W.R. 2019. Prediction of the apple scab using machine learning and simple weather stations. *Computers and Electronics in Agriculture* 161: 252-259.

Wu, R., Yan, S., Shan, Y., Dang, Q. & Sun, G. 2019. Deep image: Scaling up image recognition. *Arxiv.Org*. Accessed by May 15. https://arxiv.org/vc/arxiv/papers/1501/1501.02876v1.pdf.

Yang, X., Liu, W., Tao, D. & Cheng, J. 2019. Canonical correlation analysis networks for two-view image recognition. *Information Sciences* 385-386: 338-352.

Zheng, H., Fu, J., Mei, T. & Luo, J. 2019. Learning multi-attention convolutional neural network for fine-grained image recognition. *The IEEE International Conference on Computer Vision (ICCV)* 2017: 5209-5217.

Zhong, J., Zhang, X. & Wang, Y. 2019. Relatively weak meteorological feedback effect on $PM_{2.5}$ mass change in winter 2017/18 in the Beijing area: Observational evidence and machine-learning estimations. *Science of The Total Environment* 664: 140-147.

Center for Artificial Intelligence Technology
Faculty of Information Science & Technology
46300 UKM Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: azu1328@yahoo.com

Received: 13 March 2019
Accepted: 10 November 2019