Songklanakarin J. Sci. Technol.
44 (2), 481-487, Mar. - Apr. 2022

http://www.sjst.psu.ac.th

*Original Article*

# A comparison of multiple linear regression and random forest for community concern of youth and young adults survey

Nurin Dureh*, Attachai Ueranantasan, and Mayuening Eso

*Department of Mathematics and Computer Sciences, Faculty of Sciences and Technology,
Prince of Songkla University, Pattani Campus, Mueang, Pattani, 94000 Thailand*

Received: 14 July 2021; Revised: 1 October 2021; Accepted: 7 October 2021

## Abstract

The youth and young adults are an essential part of a community's development. Therefore, an assessment of their concerns and related factors could help reflect the overall situation in the community. In this study, the community problems of concern to youth and young adults in three districts of Pattani province are addressed. The data were collected using a questionnaire consisting of 31 items for the problems of concern, and targeting 460 youth and young adults in the focus area. This study aimed to compare the performances of two methods to explore the related factors in the survey data. Those two methods are multiple linear regression (MLR), representing a conventional statistical method, and random forest (RF), representing a machine learning approach. In the results, the random forest regression models seemed superior to the multiple linear regression models in predictive performance and errors. The findings indicate that using RF for data analysis of survey results can be an alternative to a conventional approach in social sciences research.

**Keywords**: modeling, applied social science, multiple linear regression, random forest, survey

## 1. Introduction

Youth and young adults are taking amazing steps to improve their conditions and their communities. Their attitudes and actions are a critical window to development through opportunities to participate in meaningful activities. Since the young people feel that one matters, and form warm and supportive relationships with adults, positive development of their communities can emerge. However, in most developing countries, a problem related to youth's rights is a lack of chances to express their opinions or a lack of opportunity to be an active society member. This could lead to both psychological and physical problems for the youth in the society. The study by Zulkefly and Baharudin (2010) reveals that 47 percent of Malaysia's students face psychological problems. Sharma and Verma (2013) found that most of the street children in India, Philippines, Indonesia, and South Africa are stressed and suffer from also other psychological problems. The needs and social problems concerning youth in Thailand, especially in the Deep South, a region that suffers enormously from an ongoing violent insurgency, are not much scrutinized. Therefore, finding the problems and related factors influencing the problems of concern to youth and young adults is needed.

Many statistical models have been used in survey research. For example, multiple linear regression is a traditional parametric statistical method. It provides a model representing a valid approximation of the true function f(x) (e.g., the relationship between predictors and continuous outcome variables). Under the context of parametric regression, this implies careful model specification. However, prior knowledge about the correct functional form might not always be available (Kern *et al.*, 2019).

The machine learning (ML) methods have proposed several approaches to learning explainable models from data. The ML approach creates a model that often does not require prior knowledge about the functional form of the relationship and can be applied to complex non-linear and non-additive interrelations between outcome and covariates. Popular among these are regression methods and decision trees and their ensemble variants, such as random forests. This approach learns a model as an average of individual decision trees

*Corresponding author
Email address: nurin.d@psu.ac.th

trained on subsets of the data, and averaging in this way reduces overfitting and optimizes performance on held-out test sets (Fennell *et al.*, 2019). Moreover, the random forest algorithm was outstandingly good at predicting when compared to another classical multivariate algorithm, regression analysis. Another advantage of this technique compared to traditional methods is that it can deal with unbalanced and missing data.

Although the random forest was applied to several studies in social science, as Mollina and Garib (2019) mention, the results provided by random forest do not always have a practical sense for social sciences researchers when evaluating the interactions of the predictor variables with the variable explained. Therefore, most social sciences studies have used traditional techniques that provide models where the interaction of the predictor variables is precise, and that evaluate these predictor variables' roles (Victor & Javier, 2019).

For aforementioned reasons, this study aimed to compare the performances of these two techniques applied to survey data related to the problem and needs concerned by youth in Pattani province. Therefore, the research findings will benefit policymakers and guide the application of random forest approach in social sciences survey studies**.**

## 2. Methodology

### 2.1 Dataset

The original data used in this study are survey data. They were collected using structured questionnaires comprising two parts. The first part contains the demographic factors, and the second part focused on 31 items covering self-assessed needs and problems. The variables used in the study are listed in Figure 1. The 460 samples were selected from three districts in Pattani province (Muang, Mayo, and Yaring districts) and distributed in 15 sub-districts using simple random sampling. The inclusion criteria for the sample were age between 15-25 years old, and being resident in a selected village. The subjects were interviewed face-to-face by an experienced and well-trained team, fluent in the village vernacular. As the random forest is a machine learning method, the sample size for this method should be large enough. Thus, to increase size of the original data, the bootstrap approach has been applied. This approach uses repeated sampling from the data to generate an empirical sampling distribution for a statistic (Zadkarai, 2008). Then the data were divided into two sets called the “training sample” used to fit the model and the “testing sample” used to validate the model.

### 2.2 Ethics approval

Ethical approval for the study was obtained from Ethical Review Committee for Human Research, Prince of Songkla University, Pattani Campus, No. PSU.PN.1-006/61, 3 August 2018.

### 2.3 Methods and analysis

The first stage is to check whether missing data and outliers are present, and that no abnormal data is found in the survey data. The data analysis in this paper started by reducing the 31 response variables or problem items into smaller groups. This was followed by modeling the association between demographic variables and those clustered response variables using multiple linear regression and random forest regression. The models were trained on a random selection of 70% of the dataset and then validated using the remaining 30%. Finally, the performances of models from these two methods were compared based on predictive accuracy and root mean square error. The overall process is shown in Figure 2.

Factor analysis is a statistical technique used to identify a relatively small number of underlying dimensions, or factors, which can be used to represent relationships among interrelated variables (Bartholomew, 1999). The response variables in this study were the list of needs and problems with 31 items. Therefore, exploratory factor analysis fitted by

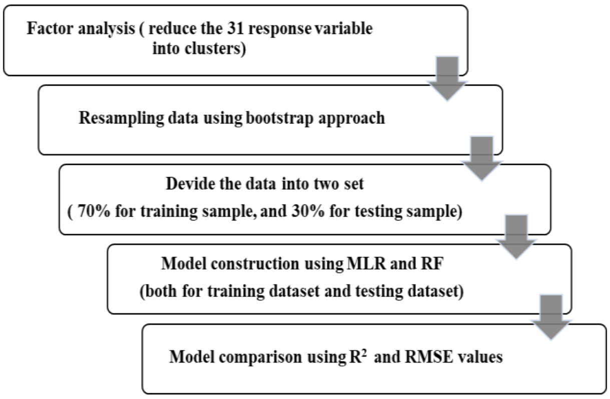
Factor analysis ( reduce the 31 response variable into clusters)
Resampling data using bootstrap approach
Devide the data into two set
( 70% for training sample, and 30% for testing sample)
Model construction using MLR and RF
(both for training dataset and testing dataset)
Model comparison using R² and RMSE values

Figure 2. Conceptual framework for the study

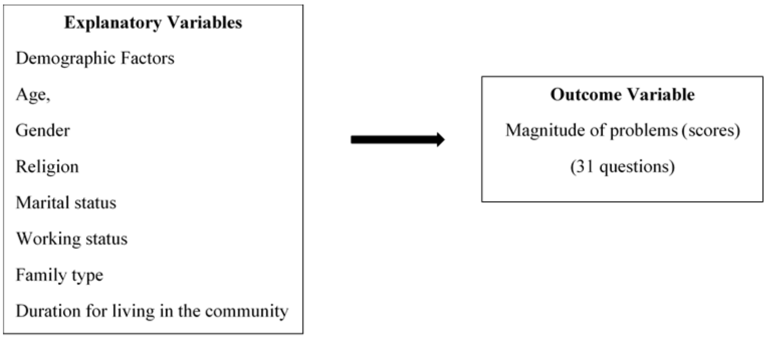

Figure 1. Path diagram

maximum likelihood with promax rotation (Venables & Ripley, 2002) was applied for clustering the 31 response variables (each coded as 0, 1, 2, or 3 to denote the severity of the problem) into smaller groups, using loadings greater than 0.35 to allocate variables to these groups.

Multiple linear regression and random forest were applied to access the relationships between selected demographic factors and the magnitude of problems or factor score. A multiple linear regression model with three exploratory variables might be specified as y~x1+ x2+ x3. It would correspond to a model with a familiar algebraic specification

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \qquad i= 1,2,...,n$$

where $y_i$ is the response variable, $x_i$ is an exploratory variable, $\varepsilon_i$ is and error and $\beta_0 \beta_i$ corresponding to intercept term and coefficients, respectively. (Venables & Ripley, 2002)

Random forest is an effective tool for classification, prediction, regression, and some other tasks. It randomly samples the training data to create many decision trees and chooses those that predict best by a bagging procedure. An extension of this algorithm was described by Breiman (2001). For this analysis, the number of trees was set to 500 and the package randomForestSRC in R program was applied for data analysis.

## 3. Results

### 3.1 Factor analysis

The results of factor analysis show that for this dataset there were 5 clusters, named as factor 1: “Lack of Chance and Opportunity”, factor 2: “Lack of Safety”, factor 3: “Lack of Social Space”, factor 4: “Lack of Facilities”, and factor 5: “Lack of Guidance”, as shown in Table1.

The factor score was calculated and treated as the outcome for model creation and model comparison based on these five clusters. The size of the sample needs to be large for applying a machine learning method, so the sample size was increased using the bootstrap approach. The data were divided into two sets called the “training sample” used to fit the model and the “testing sample” used to validate the model.

### 3.2 The association between factors and magnitude of problems using multiple regression analysis

The regression fitting was conducted to access the association between demographic factors and the magnitude of the problems for each aspect. Figures 3A to 3D show the results from regression models for the whole data set. The crude (unadjusted) means are plotted as green points, so the differences between them and confidence interval centers

Table 1. Loadings for factor analysis with promax rotation. The values in the same box are for one of the factors.

| Problem | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Uniqueness |
|---|---|---|---|---|---|---|
| Lack of education opportunity | 0.771 | | | | | 0.419 |
| Lack of chance to learn and practice new skills | 0.751 | | | | | 0.483 |
| Loss of high expectations | 0.624 | 0.106 | | | | 0.569 |
| Lack of career/education opportunity | 0.589 | -0.156 | -0.133 | 0.31 | | 0.503 |
| Unhealthy diet | 0.556 | | | | 0.134 | 0.667 |
| Children do not live with their father and mother | 0.498 | 0.2L01 | -0.205 | -0.162 | 0.392 | 0.505 |
| Illegal drug use | 0.456 | -0.151 | 0.151 | | | 0.683 |
| Child abuse, assault, bullying | | 0.814 | | | | 0.312 |
| Crime in the community | | 0.688 | | | | 0.408 |
| Lack of emotional safety | | 0.56 | | | 0.231 | 0.415 |
| Lack of physical safety | 0.317 | 0.505 | 0.13 | -0.179 | | 0.516 |
| No young people’s rights | | 0.479 | 0.202 | 0.219 | | 0.489 |
| Lack of child care | | 0.427 | -0.312 | 0.689 | | 0.556 |
| Bad nurturance and friendship | 0.213 | 0.353 | | | 0.169 | 0.538 |
| Do not have a group membership | | 0.182 | 0.863 | -0.238 | | 0.325 |
| Do not have chance to express and be creative | | 0.135 | 0.611 | -0.145 | 0.179 | 0.451 |
| Lack of decision making | | | 0.608 | | 0.174 | 0.504 |
| Lack the skills to listen and learn consciously | 0.301 | 0.159 | 0.491 | | -0.259 | 0.521 |
| Lack of quality media | 0.274 | | 0.449 | 0.176 | -0.194 | 0.504 |
| Little access to resources | 0.16 | 0.232 | 0.365 | 0.133 | -0.198 | 0.632 |
| Weak law enforcement | 0.268 | -0.258 | 0.341 | 0.225 | | 0.470 |
| Lack of playgrounds and picnic areas | | | -0.17 | 0.885 | | 0.511 |
| Do not have part-time employment | 0.189 | -0.165 | 0.169 | 0.443 | | 0.582 |
| Lack of facilities for keeping fit | 0.207 | -0.238 | | 0.38 | 0.161 | 0.599 |
| No recycling facility | 0.185 | 0.164 | | 0.372 | | 0.647 |
| Lack of supervision and training | 0.148 | 0.128 | -0.117 | | 0.798 | 0.269 |
| Lack of guidance | | 0.188 | | | 0.605 | 0.458 |
| No standards and boundaries | -0.197 | 0.162 | 0.136 | 0.217 | 0.496 | 0.483 |
| Gap between adults and children | | 0.154 | 0.169 | -0.132 | 0.48 | 0.589 |
| Children lack the opportunity to create a group activity | | | 0.23 | 0.204 | 0.391 | 0.542 |
| The community was hopeless to children | 0.166 | 0.138 | | 0.111 | 0.369 | 0.532 |

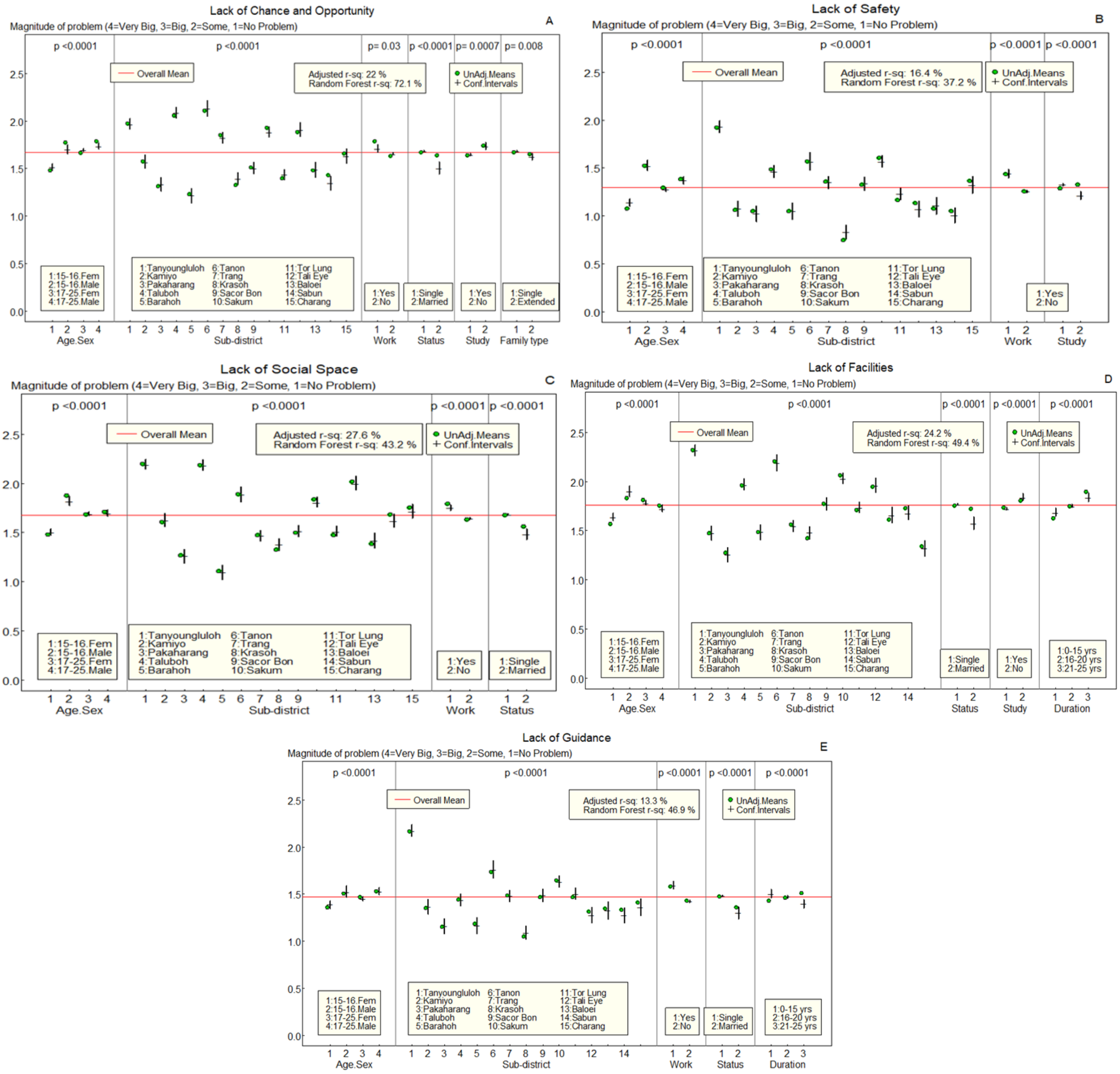

Figure 3. The confidence intervals for the relationship between demographic determinants and each factor of problems: (A) concern for factor 1, (B) concern for factor 2, (C) concern for factor 3, (D) concern for factor 4 and (E) concern for factor 5

indicate confounding bias due to associations between factors (McNeil, 2015). We found that there are age-sex variables and sub-district (Tambon) that are significantly related to factor 1: lack of chance and opportunity (Figure 3A), factor 2: lack of safety (Figure 3B), and factor 4: lack of facilities (Figure 3D). There is also evidence of an age-gender effect: girls aged 15-16 years old are less concerned about these problems, whereas men aged 17-25 are more concerned. However, for factor 3: lack of social space (Figure 3C), the results show that girls aged 15-16 years old and those who have married are less concerned about these problems. In addition, no significant differences were found in magnitudes of problems when concerning factor 5: lack of guidance, except by sub-district, as shown in Figure3D. Of these models, the random forest seems to provide higher $R^2$ compared to those from traditional multiple linear regression, as shown in Figures 3A to 3D.

### 3.3 Model comparison using R-squared ($R^2$) and root mean square error (RMSE)

The performances of the models were assessed from the coefficient of determination $R^2$, which indicates goodness of fit to data by the model. The root mean square error (RMSE) expresses the size of errors in model predictions. The larger the RMSE, the poorer the model fit.

Table 2 shows the $R^2$ and RMSE for the multiple linear regression model and the random forest model. We found that the random forest regression model provides a

Table 2. $R^2$ and root mean square error

| Factor | Linear regression | | Random forest | |
|---|---|---|---|---|
| | $R^2$ (%) | RMSE | $R^2$ | RMSE |
| Factor 1 | 21.9 | 0.6 | 72.0 | 0.3 |
| Factor 2 | 16.3 | 0.7 | 37.2 | 0.6 |
| Factor 3 | 27.5 | 0.5 | 43.2 | 0.5 |
| Factor 4 | 24.2 | 0.6 | 49.3 | 0.5 |
| Factor 5 | 13.2 | 0.6 | 46.9 | 0.5 |

better prediction and smaller errors than the traditional multiple linear regression model.

Figure 4 (A to E) compares the predictive accuracies of linear regression and random forest models, in plots of observed values versus predicted values. Based on these graphs, it is clear that the random forest model performs better in prediction than the linear regression model.

## 4. Discussion

This study compared the performances of two methods for predicting the factors related to community problems of concern to youth and young adults. The methods tested were multiple linear regression and random forest, one

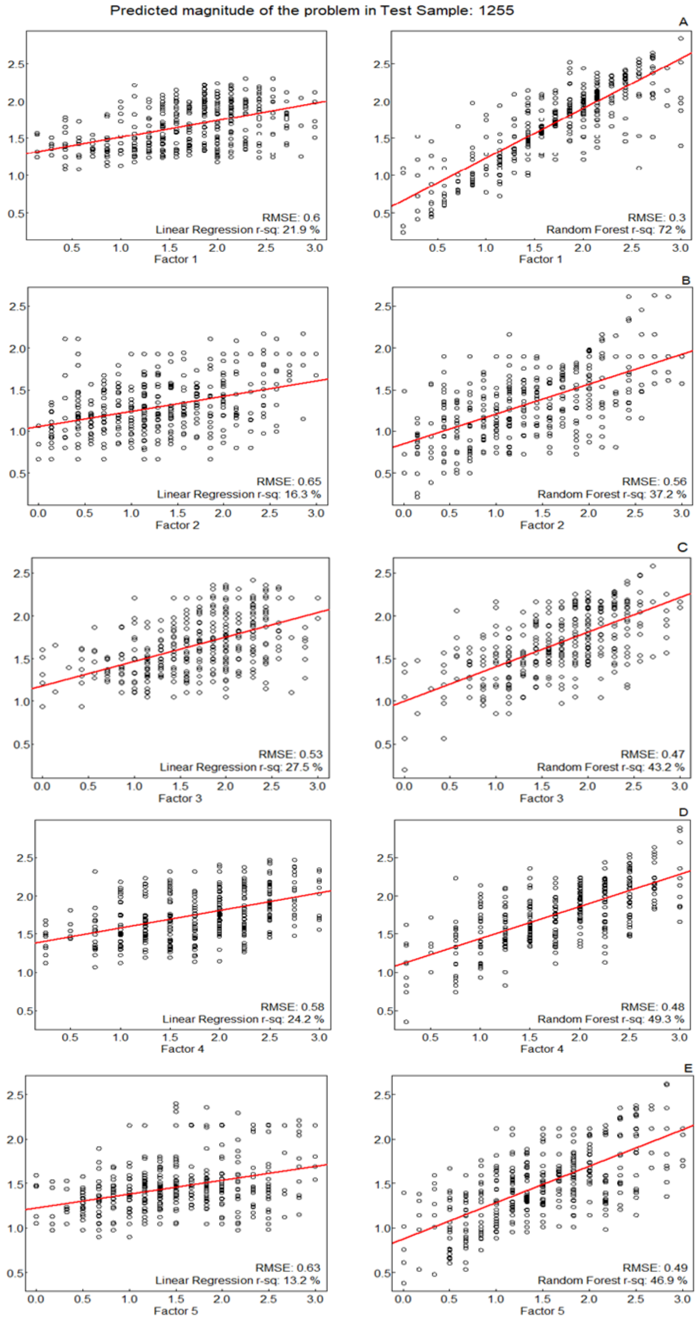

Figure 4. Predictive accuracy and error for linear regression model and random forest

of the machine learning approaches. Considering the $R^2$ values, a simple tool for assessing the model quality (Coskuntuncel, 2013), the random forest provided higher (better) values than those for linear regression. The difference between the $R^2$ values offered by these two methods was about 30%. Also, the RMSE values revealed that the random forest provided a bit lesser errors than those for linear regression. This might be because the machine learning method succeeds on large data sets and makes fewer assumptions about the data. Therefore, it can make use of non-normally distributed variables or data (Gahegan, 2003). Moreover, in some settings the classical linear regression may provide poor results; this might be because of the assumption of normality and of the absence of outliers, which are difficult to establish. Therefore, other procedures for estimation and inference than linear regression may provide a suitable alternative (Coskuntuncel, 2013; Renaud & Feser, 2010).

Many studies in social sciences have applied machine learning techniques and provided outstanding results compared to the traditional statistical methods. For example, Arpino, Moglie and Mencarini (2018) showed that random forests were able to classify the determinants of divorce according to their importance, highlighting the most powerful ones. The results found by Best, Gilligan and Baroud (2021) revealed that random forest models and other machine learning methods could help improve the predictive accuracy of migration models and identify patterns in complex social datasets. This performance is an essential strength of random forests over traditional regression because it determines what predictors of the outcome under study are most strongly related to it in a non-parametric way. Moreover, another key advantage of RF is its high predictive accuracy, and its ability to determine variable importance (Ouedraogo, 2019). However, there is no guarantee that the use of random forest will deliver the best results. Some machine learning methods do not improve prediction or fit beyond the simpler models (Seligman, Tuljapurkar, & Rehkopf, 2018). Although several social sciences studies have used random forest or other machine learning techniques, there are still a limited number of applications of machine learning in social sciences. There are two main reasons that might cause this. The first one is practical and related to the complexity of these techniques. These are not intuitive algorithms, and there is a lack of accessible resources for social scientists to learn about these techniques.

Moreover, most machine learning algorithms are computationally demanding and challenging to implement. The second reason is more fundamental. Some researchers may be doubtful about machine learning because the results are often seen as "black boxes", and findings are considered difficult to interpret in practical terms (Arpino *et al.*, 2018).

## 5. Conclusions

This study compared the performances of the models fitted to the survey data by using classical multiple linear regression and by using random forest. The clusters of problems of concern to youth and young adults in Pattani province were: lack of chance and opportunity, lack of safety, lack of social space, lack of facilities, and lack of guidance. The main finding was that age, gender, and living area were the main factors affecting the magnitude of problems. However, when comparing the model performances, the random forest performed better based on the coefficient of determination $R^2$ and the RMSE. It can be concluded that in social sciences research the random forest can be an alternative for use in data analysis.

## Acknowledgements

This research received financial support from Faculty of Science and Technology, Prince of Songkla University. We are grateful to Emeritus Professor Don McNeil for his guidance and to referees for their helpful comments.

## References

Arpino, B. Moglie, M. L., & Mencarini, L. (2018). Machine-learning techniques for family demography: An application of random forests to the analysis of divorce determinants in Germany. *Research and Expertise Centre for Survey Methodology*, Working paper, 56.

Bartholomew, DJ., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.), London, England: Arnold Publishing.

Best, KB., Gilligan, JM., Baroud, H., Carrico, A. R., Donato, K. M., Ackerly, B. A., & Mallick, B. (2021). Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. *Journal of Computational Social Science, 4*, 77–100. doi:10.1007/s42001-020-00066-9.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Coskuntuncel, O. (2013). The use of alternative regression methods in social sciences and the comparison of least squares and M estimation methods in terms of the determination of coefficient, education sciences. *Theory and Practice*, *13*(4). 2151-2158. doi:10. 12738/estp.2013.4.1867.

Fennell, P. G. Zuo, Z., & Lerman, K. (2019). Predicting and explaining behavioral data with structured feature space decomposition. *EPJ Data Science*, *8*(23). doi:10.1140/epjds/s13688-019-0201-0.

Gahegan, M. (2003), Is inductive machine learning just another wild goose (or might it lay the golden egg)? *International Journal of Geographical Information Sciences*, *17*, 69–92.

Kern, C. Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93.

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology, 45*, 27–45.

Ouedraogo, I. Defourny, P., & Vanclooster, M. (2019). Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeol Journal*, *27*, 1081–1098.

Renaud, O., & Feser, M. P. V. (2010). A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference, 140*. 1852-1862.

Seligman, B. Tuljapurkar, S., & Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*, *4*, 95–99. doi:10.1016/j.ssmph. 2017.11.008

Sharma, D., & Verma, S. (2013) Street girls and their fight for survival across four developing countries. *Psycho logical Studies*, *58*(4), 365–373.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with* S (4th ed.). Berlin, Germany: Springer.

Victor, S. L., & Javier, A. G. (2019). A random forest approach to study social determinants of depression: turning the black box into a white box in social sciences. Cadiz, Spain: University of Cadiz.

Zadkarami, M. R. (2008). Bootstrapping: A Nonparametric approach to identify the effect of sparsity of data in the binary regression models. *Journal of Applied Sciences*, *8*, 2991-2997.

Zulkefly, S. N., & Baharudin, R. (2010) Using the 12-item general health questionnaire(GHQ-12) to assess the psychological health of Malaysian College Students. *Global Journal of Health Science*, *2*(1), 73-79.