Advances and Applications in Statistics

http://www.pphmj.com
http://dx.doi.org/10.17654/AS062020185
Volume 62, Number 2, 2020, Pages 185-201 **ISSN: 0972-3617**

COMPARISON OF DIFFERENT SUPERVISED MACHINE LEARNING ALGORITHMS FOR THE PREDICTION OF TUBERCULOSIS MORTALITY

Apiradee Lim, Muhamad Rifki Taufik, Phattrawan Tongkumchum and Nurin Dureh

Department of Mathematics and Computer Science
Faculty of Science and Technology
Prince of Songkla University
Pattani Campus, Pattani 94000
Thailand

Abstract

Supervised machine learning (ML) algorithms are widely used in several areas for classification and prediction, especially with large datasets. However, it is important to investigate the application of ML algorithms to rare events. This paper compares the performance of four ML models, logistic regression (LR), recursive partitioning (RP), random forest (RF) and neural network (NN), in predicting tuberculosis (TB) mortality using LR as a benchmark. The models were applied to TB mortality data based on verbal autopsies, covering 9,495 deaths in Thailand of individuals aged five years and above, in 2005, using both the original data and also a double-sized dataset produced by the bootstrap sampling technique. The results revealed that LR performed best in predicting rare events (TB deaths) in the original data whereas RF performed best with a larger sample size. This paper concludes that the size of the dataset available for learning

Received: April 15, 2020; Accepted: May 25, 2020
2010 Mathematics Subject Classification: 62.
Keywords and phrases: classification, machine learning, predictive performance, training size.

greatly increases the performance of all the models except LR. Furthermore, the RP, RF and NN algorithms require large training datasets for the learning process. Moreover, the use of predictive accuracy alone is of limited value in distinguishing the performance of different models, particularly when the occurrence of the event is rare.

1. Introduction

The model generally used for predicting a binary outcome from a set of predictors is logistic regression (LR), which is widely used in the medical field [1]. However, since the 1990s, machine learning (ML) has been proposed as an alternative, and is gaining increasing popularity [2]. ML algorithms can be constructed using several techniques including statistical approaches and probability and optimization of learning from past experience, by extracting information from datasets [3]. Moreover, LR has also been adopted as a basic ML algorithm [4] and other well-known algorithms for ML classification or prediction include recursive partitioning (RP), random forest (RF) and neural network (NN). The application of ML algorithms in many fields, involves supervised algorithms, including, health science [4-9], environmental science [10], and social science [11]. Supervised ML is a process using a learning algorithm from a training dataset, which uses a set of inputs to predict a known output and then uses a new input data set or testing dataset, from which an unknown output can be predicted.

The performance of ML algorithms has been compared and reviewed in a number of previous studies [4-7, 9, 12] but different conclusions have been reached by these studies. Some studies have concluded that NN [13-14] and RF [15-16] have more predictive power than LR whereas other studies have claim that the prediction performance of LR is superior to that of RF [17-19] and NN [19]. Moreover, various studies have concluded that there is no difference in the predictions from LR and other forms of ML [5, 12, 20]. However, there have been only limited studies comparing the performance of ML algorithms for very low binary outcomes.

Mortality from tuberculosis (TB) is considered as a rare outcome, with the percentage of people who suffer from the disease in Thailand and die as a result ranging from 2.5 to 4%, which is a similar proportion to the mortality in developed nations such as those in Europe, and the USA [21]. Thus the study described in this paper aimed to compare the performance of different ML algorithms, namely LR, DT, RF, NN and LR, for the prediction of mortality from TB.

2. Methodology

2.1. Predictive models

Four ML algorithms were used in this study. Only one model, LR is a parametric model while all the rest are non-parametric models. The details of all the models used are described in the following section.

Logistic regression

LR is a traditional statistical model used for predicting binary outcomes from a number of predictors. A logistic function is known as a sigmoid function as it provides an S-shaped curve, and a logit function is commonly used to transform the sigmoid curve into a linear form. LR is a supervised ML tool used for classification purposes based on probability. It is used for classifying data into two classes at a certain cut point, such as 0.5.

The logistic regression model takes the form

$$p(Y = y \mid x_i) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{n} \beta_i x_i)}}, \tag{1}$$

where $p(Y = y \mid X)$ is the probability of binary outcome given a set of known predictors.

A linear form is called the logit or the log odds ratio is given below.

$$logit\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{n} \beta_i x_i. \tag{2}$$

Recursive partitioning

RP is a nonparametric statistical method of prediction and classification [22]. It creates a decision tree by splitting or not splitting each node on the tree into two daughter nodes [23]. Each node denotes a test on an attribute value, and each branch represents an outcome of the test. The tree leaves represent the final decisions as to class or class distribution [15]. A decision tree classifies the dataset appropriately by recursively picking the most important attribute that provides the best classification for each branch. To measure the quality of a split, a Gini criterion function or Gini impurity can be used, which measures how often a randomly chosen element (or determinant) from the set would be incorrectly classified.

Random forest

The RF approach was proposed by Breiman [24] and is conceptually similar to a decision tree. It is an ensemble classifier that produces multiple decision trees using randomly selected observations and a specific subset of variables (features). A large number of decision trees are trained separately and then combined into a single unified prediction, where each tree votes for the final decision as to class. The decision with the most votes by a simple majority is then the winner. A class label is represented by each leaf, based on decisions made at internal nodes of the tree. The RF approach also applies Gini impurity to determine the node splits.

Neural network

The NN approach was inspired by the system of neurons through which the brain functions. An NN consists of a large number of highly interconnected processing elements (neurons) that work in unison to solve specific problems, such as classification. A typical feed-forward neural network, known for its strong self-learning, has three layers with numerous neurons connecting to the neighboring layers. These three layers are the input layer, which is designed to obtain information from the outside world, which the network will attempt to learn about, recognize, or otherwise process, the hidden layer, which is designed to be the artificial brain, and the output layer, which presents the responses by way of classifications of the information provided [25].

2.2. Data Source

The data used in this study to compare the four different ML algorithms were verbal autopsy (VA) data collected in 2005. These data contained a total of 9,644 instances of deaths collected using a multistage stratified cluster sampling technique, in nine provinces in Thailand [26]. The variables in the data include province, gender, age, location of death, the International Classification of Diseases (ICD-10) code reported in the death registry (DR) certificate and the ICD-10 code assessed from the VA. These data were obtained from the Bureau of Health Policy and Strategy, Ministry of Public Health, Thailand.

Data exploration, cleaning, manipulation and preparation were performed before carrying out the analytical process. Mortality cases aged less than 5 years old were excluded since there were only a small number of deaths in this age group, leaving 9,495 deaths for data analysis. There were nine province represented in the dataset: Bangkok, Nakhon Nayok, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Suphan Buri, Chumpon and Songkhla. The sample was grouped by age at death into six age groups: 5 to 39, 40 to 50, 51 to 59, 60 to 69, 70 to 79, and 80 and above. An age-gender group variable was formed by combining the age groups with gender. Location of death was classified as either died in or outside hospital. ICD-10 codes of DR based on mortality tabulation [27] included 21 major causes with a chapter-block classification [28] which has been described elsewhere [29-32]. This variable was grouped into six groups: died from TB, other infectious diseases, respiratory diseases, digestive diseases, ill-defined and other diseases. The DR cause of death was combined with location of death to form a new variable with 12 groups called location-cause of death. The outcome of this study was TB death assessed from VA coded as 1 or death from other diseases coded as 0. Province, age-gender and location-cause of death were used as the predictors for each ML model.

2.3. Analytical process

The data were randomly split into training and testing sets at a 70:30 ratio. Because of the small proportion of TB deaths, the use of a ratio 70:30

helped to ensure that there would be sufficient subjects in both the training and testing sets. All four ML models were created with the variables of province, age-gender group and location-cause of death as the model predictors and death due to TB as an outcome. LR was the traditional statistical method used as a benchmark to compare with the predictive performance of RP, RF and NN. The complexity parameter (CP) value of the RP model was set at 0.0001. The number of trees in the RF model was assigned to be 2,000 and the number of nodes chosen for the NN model was six. All of these parameters were the same for both the training and testing datasets. Further, in order to confirm the prediction performance, the VA data were doubled in size using the bootstrap method and then split into training and testing datasets, also with a 70:30 ratio. All of the parameters were the same as those used in analyzing the original data.

2.4. Model evaluation

The parameters used to compare model performance were predictive accuracy (PA), true positive rate (TP or sensitivity), false positive rate (FP or 1 – specificity) and area under the curve (AUC) of the receiver operating characteristic (ROC) curve. All of these parameters can be calculated from a confusion matrix. The confusion matrix is a cross tabulation of predicted values versus observed values of the outcome. PA is the proportion of correct classifications. However, in the case of an imbalance in the sample, as was the case in the dataset used, where death due to TB was a rare event accounting for only 2% of the deaths, the PA might give a value close to 100%. Therefore, the TP and FP were also used to evaluate model performance. TP is the proportion of cases that are correctly predicted and FP is the proportion that are incorrectly predicted as positive but are actually negative. The ROC curve is the plot between TP and FP. The AUC of the ROC curve illustrates the overall performance of the model. The values of the AUC range from 0 to 1 and are generally multiplied by 100 to convert them into percentages. A model with a higher AUC value has a better predictive performance. Moreover, models with an AUC value > 0.5 perform better than models without any predictors.

In this study, confusion matrix plots were created to depict the results from all the models and ROC curves were constructed to illustrate the relationship between the TP and FP. The R program version 3.6.2 [33] was used to create the ML models and produce the confusion matrices, 95% confidence interval plot of TB mortality and the ROC curves.

3. Results

Comparison of PA

The PA of the four models based on the original data is depicted in the confusion matrices shown in Figure 1. In these matrices, the estimated risk predicted from the model is shown on the X axis and the actual status of mortality from TB is shown on the Y axis. Generally, a threshold of 0.5 is used as the cut-point of predicted probability and translated into a presence-absence classification map. However, this cut-point might not be appropriate for the data with a very low occurrence event [34]. The threshold criteria should, therefore, be dependent on prevalence and should maximize the sensitivity and specificity [35]. Thus, the cut-point of the predicted probabilities was chosen as 0.03. This cut-point was chosen based on the 2.1% (195 cases) of the cases in which death was due to TB in the dataset, and it provided the highest PA.

Each box on the plot covers 50% of the sample. The number preceded by a plus sign indicates the number of false positives and that preceded by a minus sign indicates the number of false negatives. In the training dataset from the original data, NN produced the highest PA (85.78%) followed by RF (85.69%), RP (84.85%) and LR (83.54%). The results were also very similar for the testing dataset with slightly lower PAs of 85.71%, 84.38%, 83.08% and 81.75% for NN, RF, RP and LR, respectively. Using the data produced by doubling the observations using the bootstrap technique, it was found that the PAs for all four models for both the training and testing datasets produced similar results within a range of 86 to 88% (Figure 2) and all the models showed slightly increased PAs from the values for the original datasets. For the training dataset, NN had the highest PA (87.96%) followed

by RF, LR and RP with PAs of 87.44%, 87.3% and 86.02%, respectively, while for the testing dataset the PAs were 87.48%, 86.89%, 86.2% and 85.52% for NN, RF, LR and RP, respectively.

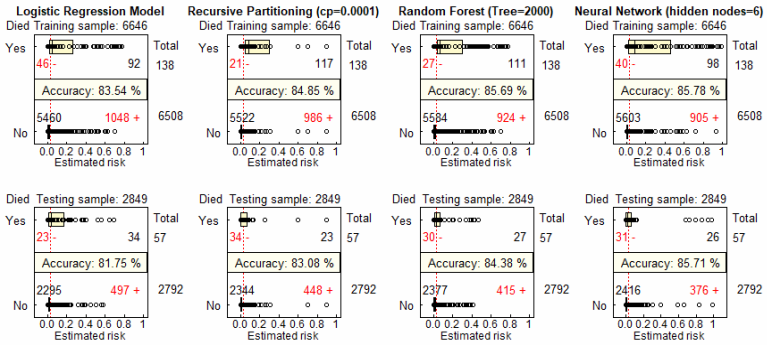

Figure 1. Confusion matrices from four ML models for training and testing datasets from the original data.

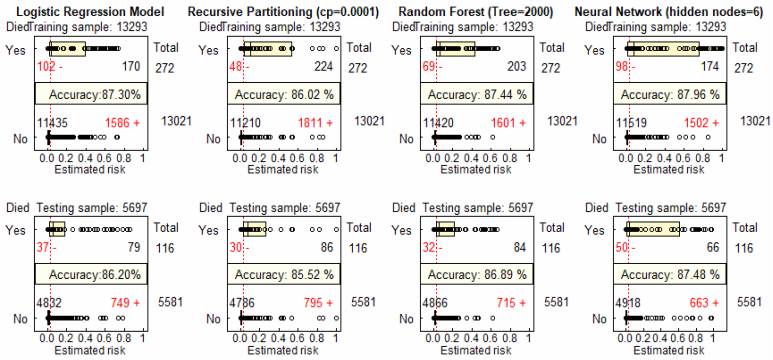

Figure 2. Confusion matrix from four ML models for training and testing datasets from the doubled-sized dataset based on the bootstrap method.

Comparison of sensitivity, specificity and AUC

The ROC curves were plotted as shown in Figure 3, with the curves for the original data on the left and those from the double-sized sample data on the right. The diagonal line represents the model without any predictors. Thus, the percentages of the AUC above the diagonal line are shown in the graph, rather than the AUC for the whole area, which clearly illustrates the difference in the predictive performance of the models.

Table 1. Performance of four ML models

Data	Values	%			
		LR	RP	RF	NN
Original	Training (n = 6,646)				
	Sensitivity	66.67	84.78	80.43	71.01
	Specificity	83.90	84.85	85.80	86.09
	Accuracy	83.54	84.85	85.69	85.78
	AUC	83.71	83.55	91.92	83.91
	Testing (n = 2,849)				
	Sensitivity	59.65	40.35	47.37	46.61
	Specificity	82.20	83.95	85.14	86.53
	Accuracy	81.75	83.08	84.38	85.71
	AUC	79.65	62.59	70.97	67.29
Double-sized sample	Training (n = 13,293)				
	Sensitivity	62.50	82.35	89.43	63.97
	Specificity	87.82	86.09	86.70	88.46
	Accuracy	87.30	86.02	87.44	87.96
	AUC	82.90	82.90	90.89	87.96
	Testing (n = 2,849)				
	Sensitivity	68.10	74.14	72.41	56.90
	Specificity	86.58	85.76	87.19	88.14
	Accuracy	86.20	85.52	86.89	87.48
	AUC	84.45	80.73	87.73	80.88

A summary of all the calculated metrics is shown in Table 1. Considering the AUC of the training dataset for the original data, RF performed best (91.92%) followed by NN (83.91%), LR (83.71%) and RP (83.55%) whereas for the testing dataset, LR performed best (79.65%) followed by RF (70.97%), NN (67.25%) and RP (62.59%). For the training dataset, RP had the highest sensitivity (84.78%) whereas NN had the highest specificity (86.09%). For the testing dataset, LR had the highest sensitivity (59.65%) whereas NN had the highest specificity (86.53%).

For the training dataset from the double-sized sample data, RF had the highest AUC (89.43%) followed by NN (87.96%) with 82.9% for both LR and RP. For the testing dataset, RF had the highest AUC (87.73%) followed by LR (84.45%), NN (80.88%) and RP (80.73%). The sensitivity of the training dataset was highest for RF (89.43%) while it was highest for RP in the testing dataset (74.14%). The specificity of all the datasets exhibited similar percentages within a range of 82-88%, with the highest for NN among all the datasets.

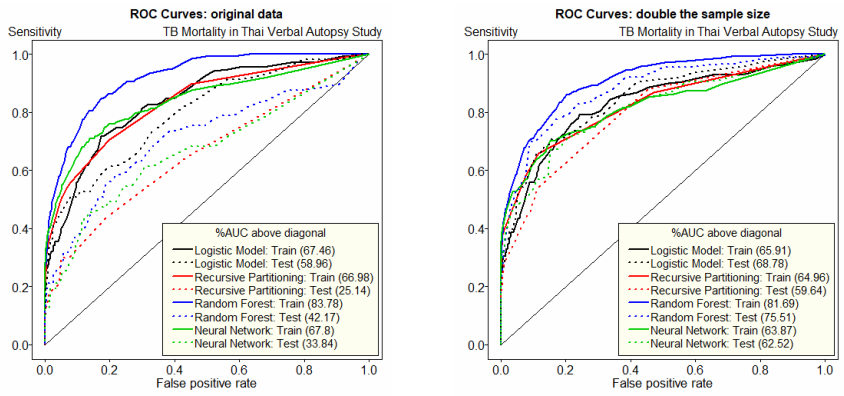

Figure 3. ROC curves of four models for original data (left) and double-sized sample based on the bootstrap method (right).

In order to explain the association between the predictors and the outcomes, the results from the LR model were used and the coefficients converted into percentages with standard errors appropriate for a 95% confidence interval (CI). The results are shown in Figure 4. The average percentage of TB mortality was 2.1 as shown by the horizontal line.

Province, age-gender and location-DR cause of death were significantly associated with TB deaths from VA. There were three provinces with higher than average, and two provinces with lower than average deaths from TB. Higher rates of TB deaths were found in males aged 50-59 years and 70 years and older whereas males aged less than 50 years, and females aged less than 40 and over 80 years had lower death rates from TB. DR deaths due to TB and from other infectious diseases were in accord with the TB deaths evaluated from VA.

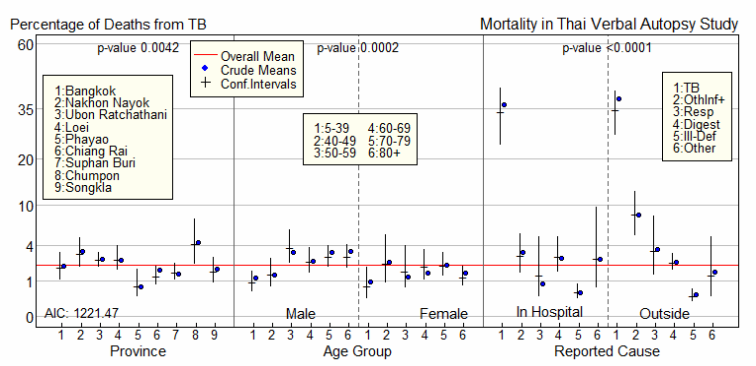

Figure 4. The 95% CI plot of TB mortality from the LR model.

4. Discussion

In this study, the predictive performance of four ML models was compared using TB mortality data from an original VA dataset and also a double-sized VA dataset produced using the bootstrapping method, and the performance of the models was considered based on testing datasets. The findings revealed that NN had the highest PA and specificity in both the original VA data and the VA data in the double-sized sample, followed by RF, LR and RP. For the original data, LR performed best as it had the highest AUC and sensitivity followed by RF, NN and RP. In the double-sized dataset, RF performed best as it had the highest AUC followed by LR, NN and RP.

The highest specificity found for the NN approach indicated that it had good performance in detecting non-cases (i.e., those who died from diseases

other than TB) as this group consisted of almost 98% of the data, which was a sufficiently large sample from which the NN could learn. This resulted in the highest PA for NN. However, PA is generally affected by threshold selection [34-36] and PA is, therefore, an ambiguous indicator, especially with extreme data as was used in this study, where the sought-after outcome represented a very small percentage from the overall sample. Therefore, the evaluation of predictive performance for a rare outcome should not rely overmuch on PA.

Another measure of the overall discrimination capacity of the models is AUC, which is obtained by integrating the area under the ROC curve of the plot between TP and FP [14]. By this measure, LR was superior to the other models based on the original data. This result supports the findings from the study by Faisal et al. [5], which compared several ML methods for low prevalence in-hospital mortality in emergency medical admissions, and found that LR performed best. This result is also in line with a study conducted by Dureh and Tongkumchum [37], who modified data by replacing zero counts by ones and doubled the values in cells with a non-zero count. They also concluded that LR performed reasonably well in those circumstances.

In the present study, the sample size is large but with low prevalence of the sought-after outcome. When the sample size was doubled, LR still provided a similar AUC to that produced with the original data. However, increasing the sample size had a greater effect on the other ML models investigated. Our findings support the suggestion of van der Ploeg et al. [38] that ML requires more data than LR. Some previous studies have pointed out that ML will not outperform LR if only a small number of predictors are considered since ML depends on there being a very large number of predictors [39-40]. With the sample size doubled, RF performed best followed by LR. This result is consistent with the study conducted by Deist et al. [41], who suggested that RF and LR should be the first choice for building classification models or for use as benchmarks.

However, one limitation of this study is that it did not compare different numbers of predictors in the models since only a small number of variables were available in the original dataset.

5. Conclusion

When dealing with a rare event, LR is a simple and suitable method which can be used for prediction although care should be taken when using PA to evaluate model performance without also assessing the AUC. The size of the sample was found not to have a great effect on the predictive performance of LR in this study, but had a large effect on that of RF, RP and NN. However a comparison of the optimal hyperparameters for ML needs to be conducted before comparing the performance of different models.

Acknowledgements

We would like to thank the Ministry of Public Health for granting permission to use the data for this study. We also highly appreciate the help of Professor Don McNeil, who provided valuable suggestions and advice. This work was supported by a Thailand's Education Hub for ASEAN Countries (TEH-AC) scholarship [grant number: THE-AC 112/2016] and a research grant provided by the Graduate School, Prince of Songkla University.

References

[1] S. Dreiseitl and L. Ohno-Machadob, Logistic regression and artificial neural network classification models: a methodology review, Journal of Biomedical Informatics 35 (2002), 352-359.

[2] N. Baghdadi and M. Zribi, Land Surface Remote Sensing in Agriculture and Forest, ISTE Press Ltd., London, 2016.

[3] T. M. Mitchell, Machine Learning, McGraw-Hill Science Inc., New York, 1997.

[4] S. S. Panesar, R. N. D'Souza, F. C. Yeh and J. C. Fernandez-Miranda, Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database, World Neurosurgery: X 2 (2019), 100012. https://doi.org/10.1016/j.wnsx.2019.100012

[5] M. Faisal, A. Scally, R. Howes, K. Beatson, D. Richardson and M. A. Mohammed, A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical

admissions via external validation, Health Informatics Journal (2018), 1-11. https://doi.org/10.1177/1460458218813600

[6] J. Z. Feng, Y. Wang, J. Peng, M. W. Sun, J. Zeng and H. Jiang, Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries, Journal of Critical Care 50 (2019), 110-116.

[7] S. Kuhle, B. Maguire, H. Zhang, D. Hamilton, A. C. Allen, K. S. Joseph and V. M. Allen, Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study, BMC Pregnancy and Childbirth 18 (2018), 333. https://doi.org/10.1186/s12884-018-1971-2

[9] S. Uddin, A. Khan, M. E. Hossain and M. A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Medical Informatics and Decision Making 19 (2019), 281. https://doi.org/10.1186/s12911-019-1004-8

[10] H. P. Sahragard, M. Ali and M. A. Z. Chahouki, Comparison of logistic regression and machine learning techniques in prediction of habitat distribution of plant species, Range Management and Agroforestry 37 (2016), 21-26.

[11] I. H. Sarker, A. S. M. Kayes and P. Watters, Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage, Journal of Big Data 6 (2019), 57.

[12] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakeland and B. V. Calstera, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, Journal of Clinical Epidemiology 110 (2019), 12-22.

[13] E. Shafiei, E. Fakharian, A. Omidi, H. Akbari, A. Delpisheh and N. Arash, Comparison of artificial neural network (ANN) and logistic regression (LR) models for prediction of psychological symptom six months after mild traumatic brain injury, Iranian Journal of Psychiatry and Behavioral Sciences 11 (2016), e5849. DOI: 10.17795/ijpbs-5849

[14] J. Huang and C. X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering 17 (2005), 299-310.

[15] C. Kuo, L. Yu, H. Chen and C. Chan, Comparison of models for the prediction of medical costs of spinal fusion in Taiwan diagnosis-related groups by machine learning algorithms, Healthcare Informatics Research 24 (2018), 29-37.

[16] H. Mansoor, I. Y. Elgendy, R. Segal, A. A. Bavry and J. Bian, Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach, Heart Lung 46 (2017), 405-411.

[17] H. Forssen, R. Patel, N. Fitzpatrick, A. Hingorani, A. Timmis, H. Hemingway and S. Denaxas, Evaluation of machine learning methods to predict coronary artery disease using metabolomic data, Studies in Health Technology and Informatics 235 (2017), 111-115.

[18] V. Taslimitehrani, G. Dong, N. L. Pereira, M. Panahiazar and J. Pathak, Developing EHR-driven heart failure risk prediction models using CPXR (log) with the probabilistic loss function, Journal of Biomedical Informatics 60 (2016), 260-269.

[19] M. M. Islam, C. C. Wu, T. N. Poly, H. C. Yang and Y. C. Li, Applications of machine learning in fatty live disease prediction, 40th Medical Informatics in Europe Conference, MIE, IOS Press, 2018.

[20] J. D. Frizzell, L. Liang, P. J. Schulte, C. W. Yancy, P. A. Heidenreich, A. F. Hernandez, D. L. Bhatt, G. C. Fonarow and W. K. Laskey, Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: Comparison of machine learning and other statistical approaches, JAMA Cardiology 2 (2017), 204-209.

[21] WHO, Global Tuberculosis Report 2017, Geneva: World Health Organization, 2017.

[22] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[23] A. J. Izenman, Recursive Partitioning and Tree-Based Methods, Modern Multivariate Statistical Techniques, Springer Texts in Statistics, Springer, New York, 2013.

[24] L. Breiman, Random forests, Machine Learning 45 (2001), 5-32.

[25] B. Krose and P. van der Smagt, An Introduction to Neural Networks, The University of Amsterdam, Amsterdam, 1996.

[26] C. Rao, Y. Porapakkham, J. Pattaraarchachai, W. Polprasert, N. Swampunyalert and A. D. Lopez, Verifying causes of death in Thailand: rationale and methods for empirical investigation, Population Health Metrics 8 (2010), 1-13.

[27] WHO, International Classification of Diseases and Related Health Problems (10th rev., ICD-10), World Health Organization, Geneva, 2016.

[28] N. Pipatjaturon, P. Tongkumchum and A. Ueranantasun, Estimating lung cancer

deaths in Thailand based on verbal autopsy study in 2005, Pertanika Journal of Science and Technology 25 (2017), 469-478.

[29] N. Klinjun, A. Lim and K. Bundhamcharoen, Estimating external causes of death in Thailand 1996-2009 based on the 2005 Verbal Autopsy study, Songklanakarin Journal of Science and Technology 36 (2014), 711-718.

[30] N. Klinjun, A. Lim and K. Bundhamcharoen, A logistic regression model for estimating transport accident deaths using verbal autopsy data, Asia-Pacific Journal of Public Health 27 (2015), 286-292.

[31] A. Chutinantakul, P. Tongkumchum and K. Bundhamcharoen, Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009, Population Health Metrics 12 (2014). https://doi.org/10.1186/s12963-014-0025-x

[32] S. Waeto, N. Pipatjaturon, P. Tongkumchum, C. Choonpradub, R. Saelim and N. Makaje, Estimating liver cancer deaths in Thailand based on verbal autopsy study, Journal of Research in Health Sciences 14 (2014), 18-22.

[33] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, 2019. https://www.R-project.org/

[34] E. A. Freeman and G. G. Moison, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, Ecological Modelling 217 (2008), 48-58.

[35] J. V. Alberto and J. Lobo, Threshold criteria for conversion of probability of species presence to either - or presence-absence, Acta Oecologica - International Journal 31 (2007), 361-369.

[36] M. Hassouna, A. Tarhini, T. Elyas and M. S. A. Trab, Customer churn in mobile markets: a comparison of techniques, International Business Research 8 (2015), 224-237.

[37] N. Dureh and P. Tongkumchum, A comparison of logistic regression and machine learning algorithms applied to zero counts data in contingency tables, Advances and Applications in Statistics 55 (2019), 67-76.

[38] T. van der Ploeg, P. C. Austin and E. W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Medical Research Methodology 14 (2014), 137.

[39] R. C. Deo and B. K. Nallamothu, Learning about machine learning: the promise and pitfalls of big data and the electronic health record, Circulation: Cardiovascular Quality and Outcomes 9 (2016), 618e20.

[40] B. A. Goldstein, A. M. Navar and R. E. Carter, Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges, European Heart Journal 38 (2017), 1805e14.

[41] T. M. Deist, F. J. W. M. Dankers, G. Valdes, R. Wijsman, I. C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, L. Wee, O. Morin, D. R. Raleigh, W. Bots, J. H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker and P. Lambin, Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers, Medical Physics 45 (2018), 3449-3459.