A comparison of statistical techniques in predicting violent outcomes in Thailand's deep South

Bunjira Makond [1,2]

Abstract

The classification of violent outcomes helps monitor their causes to prevent further loss among the public. The ability of statistical techniques to accurately predict the outcomes needs to be investigated. This study applied logistic regression (LR) and chi-squared automatic interaction detection decision tree (CHAID) techniques to predict physical and non-physical injuries which are considered violent outcomes. A set of 21,424 data about violent events from 2004 to 2016 were obtained from the Deep South Coordination Centre database and were divided into, and used as, training and testing datasets. Nine significant predictors, including arson, gun, bomb, time, province, day, quarter, zone, and district, were identified by LR as predicting violent outcomes. Likewise, only five factors, gun, zone, bomb, time, and arson, were represented in the CHAID results. However, the performances of LR and CHAID were not significantly different in terms of overall classification accuracy and area under the receiver operating characteristic curve.

Keywords: logistic regression, decision tree, Thailand's deep South, violent outcomes

[1]Leturer, Faculty of Commerce and Management, Prince of Songkla University,Trang campus, Trang, 92000, Thailand.
[2]Centre of Excellence in Mathematics, Commission on Higher Education (CHE), Ministry of Education, Ratchathewi, Bangkok, 10400 Thailand.
E-mail: bunjira.m@psu.ac.th

Introduction

Recently, violence has been a global problem and public concern because of its significant impact on the health and well-being of all humans. The World Health Organization has defined violence as the intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in, or has a high likelihood of resulting in, injury, death, psychological harm, maldevelopment, or deprivation (Krug et al., 2002). According to the definition violence includes both physical and non-physical violence.

The consequences of the violence are related to the type of violence. Physical violence causes physical injury which affects the health of the individuals involved. However, physical injury not only affects the individuals directly involved, but also influences the entire healthcare system, the delivery of healthcare, and surrounding people (The world medical association, inc, 2012). Although non-physical violence does not have a direct physical effect, it can lead to loss of faith in one's own perception of reality, constant fear of attack, fear of self-assertion, depression, and feeling frightened, for example (Morris, 2007).

In the far southern provinces of Thailand, specifically, Pattani, Yala, and Narathiwat provinces and parts of Songkhla province (Nathawi, Sabayoi, Chana, and Thepa districts), continuously occurring violence causes serious, large-scale impacts on public health. Between January 2004 and March 2013, approximately 13,000 violent events were noted, resulting in 15,574 victims, which include 5,614 people who died and 9,960 people who injured (Burke, Tweedie, & Poocharoen, 2013). Although there are people who are not directly affected by the violence, the continuing violence not only limits the prospects for economic strengthening and job creation and increases poverty. Likewise, the children, the youth, and the female members of the families need psychological counselling because they are unable to cope with the situation that they face (Sabur, 2017).

There are several different kinds of loss impact related to this situation due to the various types of violence, ranging from less serious offences, like breach of the peace, to murder. Understanding the factors associated with violence is an essential step in the public health approach to preventing violence and protecting people from violence. The understanding of the relationships and characteristics of violence becomes very important for decreasing the burden of injuries, the decision making process, and predicting the effects of those decisions.

Several researches have circumstances related to the violence data in the southernmost provinces of Thailand (Marohabout, Choonpradub, & Kuning, 2009; Inyaem, Haruechaiyasak, Meesad & Tran, 2010; Kengpol & Neungrit, 2012; Khongmark & Kuning, 2013; Chirtkiatsakul, Kuning, McNeil, & Eso, 2014; Kengpol & Neungrit, 2014). Interestingly, this study is the pioneering research to study the factors associated with violence to rely on the event incidents to predict the violent outcomes (i.e. physical and non-physical injury). The study aims to employ two well-known statistical techniques, logistic regression (LR) and Chi-squared automatic interaction detection (CHAID), analyse the violent event data to predict the outcomes. Finally, the performances of models' ability to predict outcomes will be compared.

Literature Review

Logistic regression and decision tree analyses have often been used for analogous purposes. However, several researchers have demonstrated that the performance of the two models varied substantially for different datasets. Some researchers have shown that decision tree analysis performs better than logistic regression (Delen, Walker, & Kadam, 2005; Yu, Chao, Cheng, & Kuo, 2009) while others have concluded that logistic regression outperforms decision tree analysis [such as (Long, Griffith, Selker, & D'Agostino, 1993; Zurada & Lonial, 2005)], and some have found that they are indifferent (Rudolfer & Peers, 1999; Liu, Yang, Ramsay, Li, & Coid, 2011). Accordingly, the choice of technique appears to be strongly reliant on the application. In this study LR and CHAID applied to the violent event data.

1. Logistic Regression

Logistic regression (LR) is a well-known statistical method that is also considered the gold standard method for prediction tasks (Badriyah, Briggs, & Prytherch, 2012). This method is used to describe the relation between predictor variables, denoted by $X' = (x_1, x_2, ..., x_p)$, and a dependent variable, which is a dichotomous variable represented by Y (Hosmer & Lemeshow, 2000). Due to the dichotomous dependent variable, it is also called "Binary Logistic Regression Analysis". It is critical that the categories of dependent variable are assigned as 0 and 1 in the analysis. In this study, the dichotomous dependent variable is set as $Y = 1$ if physical injury occurred, and $Y = 0$ otherwise.

The conditional probability that $Y = 1$, given the value of $X' = (x_1, x_2, ..., x_p)$ is $P(Y|X) = \pi(X)$ and the probability that $Y = 0$ is $1 - \pi(X)$ where $\pi(X)$ can be expressed as the following formula:

$$\pi(X)=\frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p)}} \quad , \quad \text{when}$$

$$0 \le \pi(X) \le 1 \qquad (1)$$

The ratio $\pi(X)/1-\pi(X)$ is called the odds. A useful transformation of LR that is taking the natural logarithm of the odds ratio is called the logit transformation, defined as:

$$g(X)=\ln\left(\frac{\pi(X)}{1-\pi(X)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p \qquad (2)$$

An odds ratio (OR) associated with the effect of a one unit change in x_j in the predicted odds ratio with the other variables in the model held constant is represented as e^{β_j}.

2. Chi-squared automatic interaction detection

Decision tree (DT) methodology is a universally used data mining method. The methodology was developed for classification systems based on multiple variables, or for developing and predicting outcomes for a target variable (i.e. dependent variable). In implementation, the DT method classifies a data into hierarchical groups of data comprised of a root node, internal nodes, and leaf nodes. The algorithm is a type of non-parametric method and is capable of coping with large, complicated datasets without any restrictions on parameters (Song & Lu, 2015).

The Chi-squared automatic interaction detection (CHAID) decision tree algorithm is one of the multivariate methods which were first proposed by Kass in 1975. The algorithm is used to detect the relationship between a categorical dependent variable and multiple independent variables which possibly are categorical, numerical or both. However, in the case of numerical variables, the coding and transformation into categorical variables must be completed beforehand (Milanovic & Stamenkovic, 2016)

Conceptually, the decision tree is a data mining technique based on a criterion that recursively divides the heterogeneous input data set into homogenous groups with respect to the dependent variable categories. For CHAID, the Pearson's Chi-square statistic test is utilized as a criterion for division. The procedure of testing and expressing the conclusions is like the traditional procedure for statistical hypothesis testing.

CHAID analysis is carried out as follows (Milanovic & Stamenkovic, 2016; "CHAID and Exhaustive CHAID Algorithms," n.d.):

Step1: Merging

In this step, the non-significant categories for each independent variable were merged so that each final category of an independent variable would result in one child node if the independent variable is used to split the node. To perform the merging, the two variable categories with the largest p-values in

relation to the dependent variable were determined to be most similar based and merged. The search for a new merging pairs continues until no new pairs are identified, or if the p-value for all potential pairs is smaller than the defined level of significance, α. In addition, the adjusted p-value is calculated so it can be used in the splitting step. When non-binary independent variables are involved, the p-value above is computed for the merged categories by using Bonferroni adjustment.

Step2: Splitting

In splitting, the independent variable is selected to be best split node by comparing the adjusted p-value related with each variable. If the adjusted p-value is less than or equal to the predefined level of significance, α, then the node is split into sub-nodes based on the merged categories. Otherwise, the node is considered to be the terminal node. The tree building process stops when p-values of all the input independent variables are higher than the specified split threshold.

3. Receiver Operating Characteristics Curve

A receiver operating characteristics curve (ROC) is a methodology that has been used to envisage, form, and select classifiers by relying on their performance. It was developed from signal detection theory in order to depict the trade-off between the hit rates and false alarm rates of classifiers during World War II. In recent years, ROC has commonly been used in various research areas, for example, medical and radiology, psychiatry, manufacturing inspection systems, finance and database marketing, machine learning, and data mining (Fawcett, 2006).

The establishment of the ROC curve is related to the confusion matrix's construction (see Table 1) and the calculation of sensitivity and specificity measures (Gajowniczek, Zabkowski, & Szupiluk, 2014). An ROC curve is a set of points (x, y) where x = 1-specificity, y = sensitivity. Sensitivity is the ability of a classifier to correctly classify positives. Likewise, specificity is the quantity of negatives that are correctly classified. Sensitivity and specificity can be calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (4)$$

Table 1 Confusion matrix

		Predicted class	
		yes	no
Actual class	yes	TP	FN
	no	FP	TN

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. In this study, the class of interest (i.e. physical injury) is "yes"; which is therefore denoted as "positive" with others as “negative”.

4. Area under the curve

The area under the curve (AUC) is a measure to summarize the total area of the whole ROC curve. The AUC can be interpreted as having the following meanings: (1) the probability that a randomly chosen instance of positive is ranked as more likely to be positive than a randomly chosen negative instance, and (2) the average value of sensitivity for total possible values of specificity. The maximum value of the AUC is 1, meaning that the classification, or prediction, model is perfect in the differentiating between positive and negative instances. An AUC which equals 0.5 means the chance discrimination that curve becomes a diagonal line through the ROC space. When all instances are incorrectly classified, then the AUC equals 0; however, this happening is an extremely rare to occurrence (Hajian-Tilaki, 2013). The AUC can be calculated via trapezoidal approximation (Fawcett, 2006).

Methodology

1. Data, data pre-processing and variables

This was a retrospective study in which the data consisted of information on violent events which were recorded between 2004 and the beginning of January 2016 in the Deep South Coordination Centre (DSCC) database, Prince of Songkla University, Pattani, Thailand. Though there are various types of violent events, they are considered as equally violent in this study.

Nine related variables were included in this study, namely, “arson”, “gun”, “bomb”, “time”, “province”, “day”, “quarter”, “zone”, and “district” . The variables were selected according to the studies’ results of (Marohabout, Choonpradub, & Kuning, 2009;

Chirtkiatsakul, Kuning, McNeil, & Eso, 2014;) and epidemiological bases (Kuning, Eso, Sornsrivichai, & Chongsuvivatwong, 2014).

Normally, the data contained some noise, such as missing values, uncertainty, and replication. Noise effects cause destructiveness in any data analysis (Xiong, Pandey, Steinbach, & Kumar, 2006). Consequently, data preparation processes included data cleaning, data integration, data transformation, and data reduction were implemented so that the quality of the data and the accuracy of the classification and prediction processes were enhanced.

Therefore, data cleaning was employed for missing variable values, namely, “day of the week”, “time of the day”, “place of the incident”, and “province” were removed from the data set. Data transformation was implemented to convert data into the required form; for example, the variable “day of the week” was determined from the date of the event. Moreover, values for four variables, “time of day”, “quarter of year”, “place of the incident”, and “district” were grouped into categories as specified in Tables 2 and 3 to facilitate analysis. Finally, the dataset was pre-classified into two classes according to the violence’s impacts on health. Violence that caused a physical injury (i.e. the victims were injured or died) was labelled as “1”; otherwise, it was labelled as “0”. All the steps of data pre-processing were performed using SQL and Microsoft Excel. Ultimately, the total number of incidents included in the study was 21,424. The variables with descriptions and their values are presented in Tables 2 and 3.

Table 2 Variables with descriptions and the values

Variables	Descriptions	Values	Number of events
time	time of the day	"1" represents time period from 00:01 a.m. to 03.00 a.m.	1,437
		"2" represents time period from 03:01 a.m. to 06:00 a.m.	1,912
		"3" represents time period from 06:01 a.m. to 09:00 a.m.	3,824
		"4" represents time period from 09:01 a.m. to 12:00 a.m.	2,538
		"5" represents time period from 12.01 p.m. to 15.00 p.m.	2,116
		"6" represents time period from 15:01 p.m. to 18.00 p.m.	2,564
		"7" represents time period from 18:01 p.m. to 21.00 p.m.	4,841
		"8" represents time period from 21:01 p.m. to 24.00 a.m.	2,192
day	day of the week	"1" represents Sunday	2,773
		"2" represents Monday	3,316
		"3" represents Tuesday	3,130
		"4" represents Wednesday	3,358
		"5" represents Thursday	3,323
		"6" represents Friday	3,030
		"7" represents Saturday	2,494
quarter	quarter of the year	"1" represents January to March	5,081
		"2" represents April to June	5,824
		"3" represents July to September	5,599
		"4" represents October to December	4,920

Table 2 Variables with descriptions and the values (continued)

Variables	Descriptions	Values	Number of events
zone	place of the incident	"1" represents road / highway	11,792
		"2" represents residential area/personal area or shop	4,603
		"3" represents other / unspecified	5,029
district	district in this study	"1" represents district that is adjacent to neighbour province/country	14,250
		"0" represents district that is not adjacent to neighbour province/country	7,174
province	province in this study	"1" represents Narathiwat	7,614
		"2" represents Pattani	7,040
		"3" represents Songkhla	1,092
		"4" represents Yala	5,678
arson	means used in the incident was arson	"1" represents yes	2,538
		"0" represents no	18,886
gun	weapon used in the incident was one or more guns	"1" represents yes	10,342
		"0" represents no	11,082
bomb	weapon used in the incident was one or more bombs	"1" represents yes	3,987
		"0" represents no	17,437

Table 3 Class distribution of dependent variable

Variables	Descriptions	Values	Number of events
outcome	The outcome of violence	"1" represents physical injury	10,740
		"0" represents otherwise	10,684

2. Model evaluation criteria

The amassed data was split into two data sets: the training set contained 80% of the data was used to build the models (determine their parameters) and the test set (20% of the data) was used to measure their performance. The performance of the models was measured using overall classification accuracy and area under the receiver operating characteristic curve (AUC). Overall accuracy is usually expressed as a percentage that expresses what proportion of the complete set of reference points were classified correctly. Meanwhile, a Receiver Operating Characteristic (ROC) curve is made up of a plot of the true positive rate (sensitivity) and the false positive rate (1-specificity). The area under the ROC curve (AUC) quantifies how correctly a parameter is able to classify between two classes (physical injury / non-physical injury).

Experiment Results

1. Results from the logistic regression analysis

LR was used to build the model. In order to develop an equation that maximized the descriptive capacity using the lowest number of statistically significant independent variables (i.e. risk factors), a backward stepwise variable selection method was used. Backward stepwise regression analysis starts with a full model and variables are removed from the model in an iterative process using the Wald statistic test. The results presented in Table 4 show that all of the risk factors were significantly associated with physical injury. The risk of the occurrence of physical injury was significantly higher in violent incidents where guns or bombs were used. Moreover, events which occurred during time periods between 06:01 a.m. and 09:00 a.m., 09:01 a.m. and 12:00 p.m., 12.01 p.m. and 15.00 p.m., and 15:01 p.m. and 18.00 p.m. had a higher risk of physical injury.

Events which took place on a road / highway or in residential/personal areas or a shop had a higher risk of physical injury than those which occurred in other / unspecified areas. Events causing physical injury were more likely to take place in Pattani and Yala compared to Narathiwat. The performances of the LR in terms of accuracy value and AUC (also see Figure 2) demonstrated in Table 5 are 82.3% and 0.891, respectively.

Table 4 Logistic regression analysis of risk factors of physical injury

Risk factor	B	Wald	Sig.	Exp(B)	95% C.I. for EXP(B)	
					Lower	Upper
arson	-1.439	139.546	.000*	.237	.187	.301
gun	3.713	4027.976	.000*	40.981	36.541	45.960
bomb	1.660	689.508	.000*	5.257	4.645	5.951
time (ref = 21:01 p.m. to 24.00 a.m.)		99.678	.000*			
00:01 a.m. to 03.00 a.m.	.043	.136	.712	1.043	.832	1.309
03:01 a.m. to 06:00 a.m.	-.310	8.066	.005*	.733	.592	.908
06:01 a.m. to 09:00 a.m.	.314	13.600	.000*	1.368	1.158	1.616
09:01 a.m. to 12:00 a.m.	.335	13.576	.000*	1.398	1.170	1.670
12.01 p.m. to 15.00 p.m.	.264	7.820	.005*	1.302	1.082	1.566
15:01 p.m. to 18.00 p.m.	.578	39.003	.000*	1.782	1.487	2.137
18:01 p.m. to 21.00 p.m.	.072	.788	.375	1.075	.916	1.261
province (ref = Narathiwat)		12.385	.006*			
Pattani	.162	9.162	.002*	1.176	1.059	1.306
Songkhla	.054	.814	.367	1.055	.939	1.185
Yala	.247	5.176	.023*	1.280	1.035	1.583
day (ref = Saturday)		19.125	.004*			
Sunday	-.248	8.084	.004*	.780	.657	.926
Monday	-.299	12.866	.000*	.741	.629	.873
Tuesday	-.113	1.795	.180	.893	.757	1.054
Wednesday	-.286	11.781	.001*	.751	.638	.885
Thursday	-.223	7.043	.008*	.800	.679	.943
Friday	-.197	5.368	.021*	.821	.695	.970
quarter (ref = October to December)		20.713	.000*			
January to March	.117	3.435	.064	1.124	.993	1.272
April to June	-.158	6.803	.009*	.854	.759	.962
July to September	-.028	.209	.648	.972	.863	1.096
zone (ref = other / unspecified area)		189.977	.000*			
road / highway	.711	177.952	.000*	2.035	1.833	2.259
residential area/personal area or shop	.671	107.149	.000*	1.956	1.723	2.221
district	-.364	52.506	.000*	.695	.630	.767
constant	-2.172	197.890	.000*	.114		

*p-value < 0.05

2. Results from the chi-squared automatic interaction detection analysis

In this study, after setting the defaults, CHAID was implemented to develop a classification tree model, which is illustrated in Figure 1. As shown, the tree model included 5 risk factors: gun, zone, time, bomb, and arson. The model contained a total of 20 nodes, including 12 terminal nodes. Using the tree model, as the results in Table 5 demonstrates, it is able to predict physical injury with the accuracy rate is 82.8%. As well, the AUC (also see in Figure 2) is 0.894.

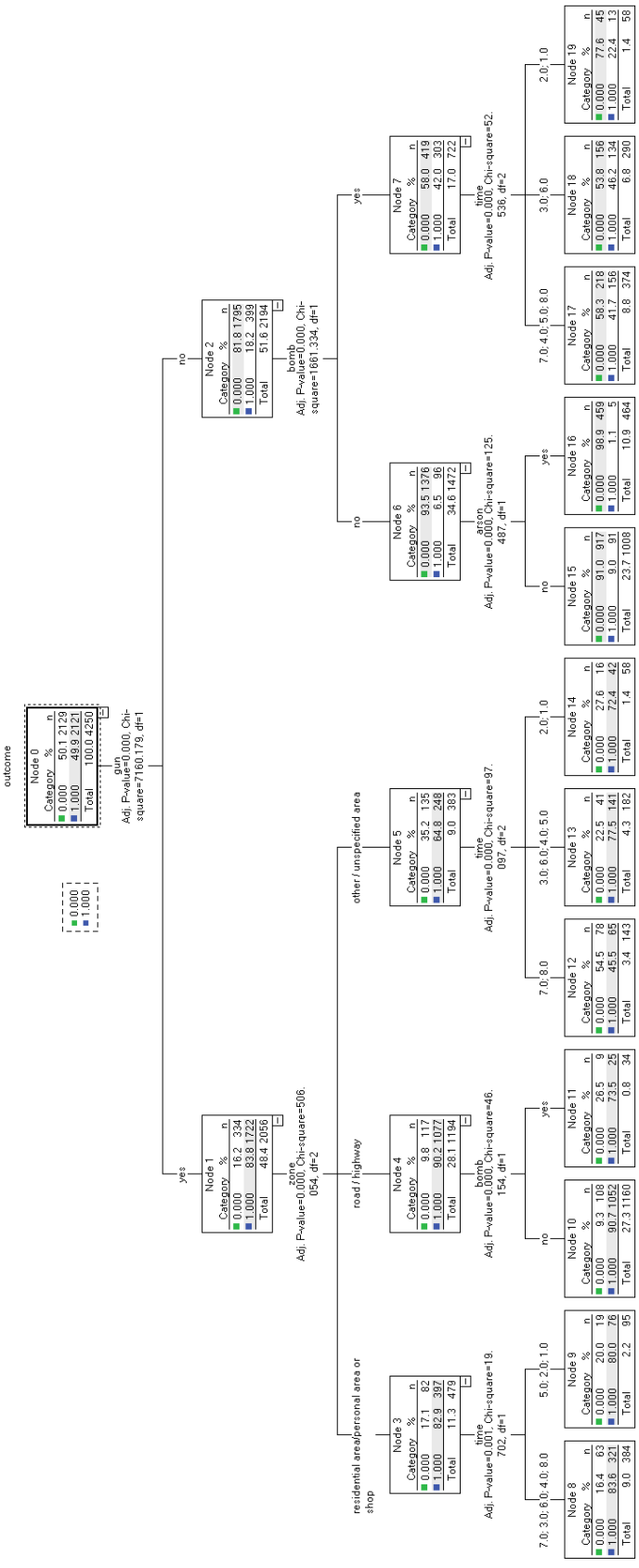

Figure 1: The tree model of CHAID

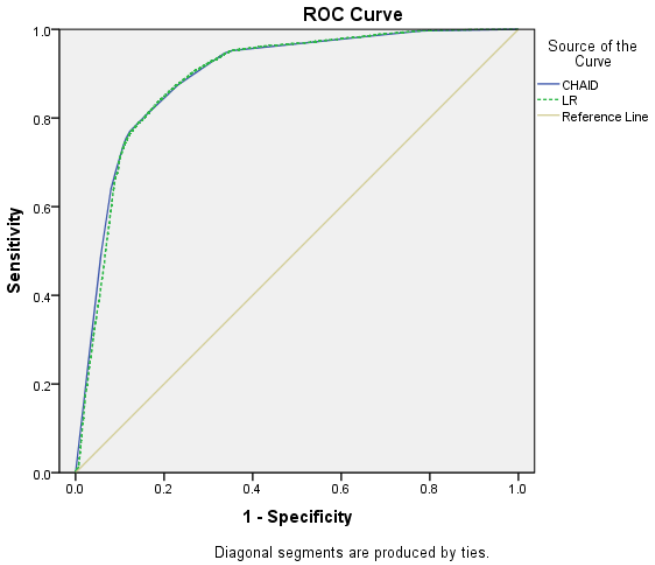

Figure 2: The area under the ROC curve

Table 5 Performances of models

Models	Accuracy	AUC
Logistic regression	82.3%	0.891
CHAID	82.8%	0.894

Conclusions and Discussions

This study applied two statistical models, LR and CHAID to predict violent outcomes (i.e. physical injury and non-physical injury). The data obtained from the DSCC database included nine independent variables. The results showed that the nine variables are significant in the LR model; while only five variables were included in the tree model. Nevertheless, the performance of LR and CHAID were not appreciably different in terms of overall classification accuracy (82.3% and 82.8% for LR and CHAID, respectively) and area under the receiver operating characteristic curve (0.891 and 0.894 for LR and CHAID, respectively).

This is pioneering research studying the factors associated with violence to rely on the event incidents to predict violent outcomes (i.e. physical and non-physical injury). The findings from both LR and CHAID showed consistent results (Chirtkiatsakul, Kuning, McNeil, & Eso, 2014) which concluded that there was a

higher probability for death to occur as a result of attacks using guns rather than bombs. However, some researches have shown that most deaths were the result of bombings (Pusponegoro, 2003; Peleg, Daniel, & Stein, 2004). It is essential to recognize that this study did not explore the question of why most physical injuries were caused by guns. Therefore, further studies are needed which include more variables related to victims and injuries for clarification of this issue.

In terms of application, CHAID has the advantage over LR; CHAID is a non-parametric method without assumptions about multicollinearity like those which are inherent in LR. It can incorporate both categorical and continuous variables to build the model and has the ability of modeling complex relationships between variables, which is a phenomenon which generally happens in real data. Moreover, the results of CHAID presented in graphs are easy to interpret.

Acknowledgement

The author is very grateful for the help of Assistant Prof. Metta Kuning, the former Director of the DSCC, Prince of Songkla University, Pattani campus. My thanks also goes to the Centre of Excellence in Mathematics, Commission on Higher Education, Thailand for their financial support.

References

Badriyah, T., Briggs, J. S., & Prytherch, D. R. (2012).Decision trees for predicting risk of mortality using routinely collected data. **International Journal of Social and Human Sciences**, 6, 303-306.

Burke, A., Tweedie, P., & Poocharoen, O. (2013). **Understanding the Subnational Conflict Area. The Contested Corners of Asia: Subnational Conflict and International Development Assistance The Case of Southern Thailand**, (p11–24). The Asia Foundation,San Fransisco, CA, U.S.A.

CHAID and Exhaustive CHAID Algorithms. (n.d.) Available from URL:ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf [Accessed 2018 Feb.]

Chirtkiatsakul, B., Kuning, M., McNeil, N., & Eso, M. (2014). Risk Factors for Mortality among Victims of Provincial Unrest in Southern Thailand. **Kasetsart J.** *(Soc. Sci)*, 35, 84 - 91 (2014)

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. **Artificial Intelligence in Medicine**, 34, 113-127.

Fawcett, T. (2006). An introduction to ROC analysis. **Pattern Recognition Letters**, 27, 861-874.

Gajowniczek, K., Zabkowski, T., & Szupiluk, R. (2014). Estimating the ROC curve and its significance for classification models' assessment. **Quantitative Methods in Economics,** 15(2), 382 -391.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. **Caspian Journal of Internal Medicine**, 4(2), 627-635.

Hosmer, D.W., & Lemeshow, S. (2000). **Applied logistic regression** (2nd ed.). New York, New York, USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.

Inyaem, U., Haruechaiyasak, C., Meesad, P., & Tran, D. (2010). Terrorism Event Classification Using Fuzzy Inference Systems. **International Journal of Computer Science and Information Security**, 7(3), 247-256.

Kengpol, A., & Neungrit, P. (2012). A Prediction of Terrorist Distribution Range Radius and Elapsing Time: A Case Study in Southern Parts of Thailand. Proceedings of Intelligence and Security Informatics Conference (EISIC), 2012 European, Odense, 180-188. doi: 10.1109/EISIC.2012.19

Kengpol, A., & Neungrit, P. (2014). A decision support methodology with risk assessment on prediction of terrorism insurgency distribution range radius and elapsing time: An empirical case study in Thailand. **Computers & Industrial Engineering**, 75, 55-67.

Khongmark, S., & Kuning, M. (2013). Modeling Incidence Rates of Terrorism Injuries in Southern Thailand. **Chiang Mai Journal of Science**, 40(4), 743-749.

Krug, E. G., Dahlberg, L. L., Mercy, J. A., Zwi, A.B., & Lozano, R. (2002). **World report on violence and health.** Geneva, *World Health Organization*.

Kuning, M., Eso, M., Sornsrivichai, V., & Chongsuvivatwong, V. (2014)**. Epidemiology of the Violence in the Deep South. In Chongsuvivatwong,V. , Boegli, L. C. & Hasuwannakit, S (Eds). Healing under fire the case of southern Thailand (pp. 41-49).** The Deep South Relief and Reconciliation Foundation and the Rugiagli Initiative, Thailand, Bangkok.

Liu, Y.Y., Yang, M., Ramsay, M., Li, X.S., & Coid, J.W. (2011). A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending. **J Quant Criminol**. doi: 10.1007/s10940-011-9137-7

Long, W.J., Griffith, J.L., Selker, H. P., & D'Agostino, R.B. (1993). A comparison of logistic regression to decision tree induction in a medical domain. **Computers in Biomedical Research**, 26, 74-97.

Milanovic, M., & Stamenkovic, M. (2016**).** Chaid decision tree: methodological frame and application. **Economic Themes,** 54(4), 563-586. doi: 10.1515/ethemes-2016-0029

Marohabout, P. , Choonpradub, C., & Kuning M. (2009). Terrorism Risk Modeling in Southern Border Provinces of Thailand during 2004 to 2005. **Songklanakarin J. of Social Sciences & Humanities**, 15(6), 883-895.

Morris, S. C. (2007). The Causes of Violence and the Effects of Violence On Community and Individual Health. **Global Health Education Consortium.** September 2007.

Peleg, K., Daniel, L., & Stein, M. (2004). Gunshot and explosion injuries characteristics, outcomes, and implications for care of terror-related injuries in Israel. **Annals of Surgery,** 239(3), 311-318.

Pusponegoro, A. D. (2003). Terrorism in Indonesia. **Prehospital and Disaster Medicine**, 18(2), 100-105.

Rudolfer, S.M., & Peers, I.S. (1999). A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. **Computer and Biomedical Research**, 32, 391-414.

Sabur, M. A. (2017). **Minorities in Thailand and current Issues**. Available from URL:http://www.iiipeace.org/Southern%20Thailand%20Present%20Situation%20and%20Its%20Impact%20by%20Abdus%20Sabur%2005_07_12.htm [Accessed 2017 Jan.]

Song, Y.-Y. & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. **Shanghai Arch Psychiatry**, 27(2), 130-135. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.215044

The world medical association, inc. (2012)**. WMA statement on violence in the health sector by patients and those close to them. Adopted by the 63rd WMA General Assembly, Bangkok, Thailand, October 2012**. Available from URL: https://www.med.or.jp/jma/jma_infoactivity/jma_activity/2012wma/2012_06e.pdf [Accessed 2018 Mar.]

Yu, Y.-W., Chao, C.-M., Cheng, B.-W., & Kuo, Y.-L. (2009). Predicting breast cancer survivability: a comparison of three data mining methods. Proceedings of Asia Pacific Industrial Engineering & Management Systems Conference. 14-16.

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. **IEEE Transactions on Knowledge and Data Engineering**, 18(3), 304-319.

Zurada, J., & Lonial, S. (2005). Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. **The Journal of Applied Business Research**, 21, 37-53.