Proceedings of the International Conference on Industrial Engineering and Operations Management
Pilsen, Czech Republic, July 23-26, 2019

Analysis of Accidental Deaths During Songkran Festival Using Data Mining

Pornpimol Chaiwuttisak
Statistics Department
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
pornpimol.ch@kmitl.ac.th

Abstract

The objective of this research is to analyse the deaths of people during Songkran holidays and to develop a model for the classification of deaths caused by road accidents. Data used in this research studies including Injuries and the loss of life in the accident between 2008 and 2014, a total of 2,875 people from the database of the Digital Government Development Office. The statistics used in the hypothesis testing are the Chi-square test statistic, Independent variables are behavior factors: drinking, not wearing a helmet, physical environmental factors such as the time when the road accident occurs, the type of road that caused the accident and the dependent variable was the death and injured person. The hypothesis testing at the significance level of 0.05 showed that all variables are associated with death during Songkran holidays. In addition, data mining techniques are applied to this research. Decision Tree, Bayesian Learning, Logistic Regression and Neural Network are applied to identify deaths described by a set of attributes and compare the accuracy of data classification with various data mining techniques. As the result, it was found that logistic regression can be correctly classified higher than other classification techniques with a precision of 72.20%.

Keywords
Road Accident, Songkran Festival, Chi square test, Data Mining Techniques

1. Introduction
The World Health Organization (2019) reported that road accidents are the major causes of the death of people around the world. In 2016, more than 1.35 million people are died with road accidents which are the biggest killer of young people and working people which adversely affects the economy and society. The road accident situation in Thailand has been ranked as the top first of death rate in Southeast Asia with a death rate of 32.7 people per 100,000 people and as the second rank in the world. Ocharoen (2017) said that during 2011 - 2013, the average cost of road accident was 545,435 million Baht per year, accounting for 6% of the Gross Domestic Product (GDP) each year. It was found that more than 10,000 people die in the road, and more than 1 million people are injured and tens of thousands of people with disabilities Especially during the New Year and Songkran Festivals due to it is the long holiday period, many people often use the road to travel back to their homeland and travel during the festival. It leads to the higher road accident risks than the normal situation.

Suangka (2016) studied the driver behaviour of adolescents affecting the risk of motorcycle accidents. The sample data are students in vocational schools in Nakhon Ratchasima, Surin and Chaiyaphum province about 933 people by using the Structural Equation Modelling. The results of the analysis revealed that the driver attitudes had a direct influence on risk-taking behaviour in accidents. The drivers who have the negative attitude toward traffic laws, speed limits, and driving responsibly lead to a high risk for accidents

Tanaboribun (2005) conducted a trend analysis of road accidents. The risk factors influencing road accidents in Thailand by using the descriptive analysis statistics are time of accident, drinking alcohol, wearing a helmet. The study indicated that motorcycles are also vehicles that are the main cause of accidents. It is often an accident without a party or Single Vehicle Crash and caused by drinking alcoholic beverages. The number of people suffering from accidents during the night will be close to the daytime, but the number of deaths in the nighttime is more than day time.

The research aims to investigate factors affecting deaths from road accidents during the Songkran Festival of Thailand in order to create and present a model to predict deaths from road accidents based on relevant factors by using Chi-square test and data mining techniques: Decision Tree, Bayesian Learning, Logistic Regression and Neural Network. Section 2 describes the data mining techniques. Section 3 is a method of conducting research which consists of data collection procedures, data preparation until the modeling process. Section 4 shows the experimental result and Section 5, which is the last part, is a summary.

2. Literature Reviews

2.1 Chi-square test

The chi-square test is used to test the relationship between two variables. The data are classified in the Nominal Scale. The data are in the form of frequencies, percentage, ratios. The formula can be written in Equation (1).

$$\chi^2 = \sum_{i=1}^{r}\sum_{j>1}^{c}\frac{(O-E)^2}{E} \tag{1}$$

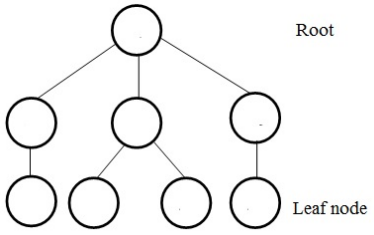
, where
O represents the observed frequency
E represents the expected frequency

We can find E from $E = \frac{r \times c}{N}$

, where
r represents the sum of the frequencies in that row
c represents the sum of the frequencies in that column.
N represents the total of frequencies

2.2 Decision tree

Decision Tree is a data mining technique based on the tree structure for data classification to support various decisions. It usually consists of rules in the form of "if <the condition> then <the result>" which is similar to the nature of the tree invert structure. The node in the first level of the tree is called “Root”. Each node describes the attribute and a branch shows the value of the attribute. The leaf node shows the class as shown in Figure 1.



Figures 1: A structure of Decision Tree

C4.5 algorithm can be applied to both continuous and discrete data. It can to customize the tree for making a decision, known as Pruning Trees. The information gain and the Entropy are calculated as equation (2) and (3).

$$Entropy(S) = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{2}$$

, where
S is the attribute that is used to measure Entropy
P_i is the ratio of the number of members of the group i to the total number of members of the sample group.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|Sv|}{|S|} Entropy(Sv) \tag{3}$$

, where
A is an attribute
$|S_v|$ is the number of members of attribute A with the value v
$|S|$ is the number of members of the sample group

Kaewta and Mahaweerawat (2010) have applied the decision tree for analysis to classify the offense into the appropriate section for the case. Srichaiwong et al. (2014) developed a decision support system by using the decision tree technique to diagnose longan leaf disease.

2.3 Bayesian learning

Bayesian learning applies the principle of probability by considering the probability distribution in data classification. In this paper, we refer to the Naïve Bayesian Learning algorithm.

Simple Bayes learning is based on Bayes' s rules, but it reduce complexity by adding the assumption that the properties of the data will not depend on each other. It can be said that the probability of data classified in the group C_i for data that have n attributes $(A_1, A_2, .., A_n)$ and can be represented by symbols as follows: $P(C_i \mid A_1, A_2, .., A_n)$

From Bayes's Theorem:

$$P(C_i \mid A_1, A_2, .., A_n) = \frac{P(A_1, A_2, .., A_n \mid C_i) \times P(C_i)}{P(A_1, A_2, .., A_n)}$$

Each attribute does not depend on each other. It can be written in the equation (4):

$$\frac{\prod_{j=1}^{n} P(A_j \mid C_i) \times P(C_i)}{P(A_1, A_2, .., A_n)} \tag{4}$$

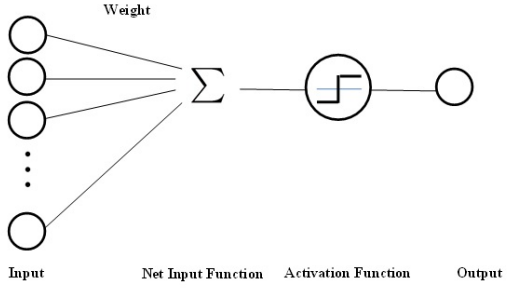
Domingos and Pazzani (1996) say that Bayes Learning is employed to classify data effectively, although the above assumptions are not true.

2.4 Logistic regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability or the likelihood of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X that are independent variables.

2.5 Neural network

Neural Network is a simulation based on the concept of the work in human brain cells. The smallest unit of the neural network called Perceptron by accepting input and calculating these values by giving the weight of each input. The output will be calculated by errors to adjust the input weight. It can be shown as Figure 2.



Figures 2: A structure of Neural Network

3. Research Methodology

In this study, CRISP-DM is the methodology applied to achieve objectives. Firstly, analysis of attributes and their subsets is conducted. Secondly, appropriate data mining tools are used in order to identify the underlying patterns.

3.1 Data preparation

The data used in the experimental study are secondary data from the database of the Digital Government Development Office. (Public Organization, 2018) during the year 2008 and 2014 about the accident occurring during the Songkran holiday season by selecting only the accident that knows the details of the accident, and then select the attributes to classify the injuries or deaths from the accidents There are 10 attributes and the total number of the possible data values is equal to 44. In addition, the symbols representing the features and the possible values of each attribute are shown in Table 1.

Table 1: The attributes or features for analysis

Symbol	**Value**
Accident time	1 = 24:01-06:00 2 = 06:01-12:00 3 = 12:01-18:00 4 = 18:01-24:00
Gender	1 = Male 2 = Female
Age	1 = < 20 2 = 20 – 30 3 = 31 – 40 4 = 41 – 50 5 = 51 – 60 6 = >60
Road	1 = City streets 2 = Rural road 3 = Highway
Status	1 = Passenger 2 = Driver 3 = Pedestrian
Vehicle	1 = No party/ Single Vehicle Crash 2 = Motorcycle 3 = Bicycle 4 = pickup 5 = Taxi 6 = wheeler buses 7 = Big bus 8 = Truck 9 =Tricycle 10 = Other
Litigant	1 = No party/ Single Vehicle Crash 2 = Motorcycle 3 = Bicycle 4 = pickup 5 = Taxi 6 = wheeler buses 7 = Big bus 8 = Truck 9 =Tricycle 10 = Other
Wearing a helmet	1 = Yes 2 = No
Drink	1 = Yes 2 = No
Death	1 = Yes 2 = No

3.2 Classification modeling

To create a model to predict the deaths of people in this road accident, Data mining techniques are employed in this research: Decision Tree using algorithm C4.5, Bayesian Learning using the Naïve Bayesian Algorithm., Logistic Regression and Neural Network using Multilayer Perceptron (MLP) algorithm with Learning Rate equal to 0.10.

Weka application version 3.7.5 is used for modeling and evaluation. The data set used in the research consists of 2,875 records. The data set is divided into 2 datasets by stratified random sampling. Firstly, training set for creating a model and testing dataset for testing a model in the proportion of 70:30 by using Microsoft Excel version 2010 application.

3.3 Performance evaluation

The evaluation of the efficiency of the model using data mining techniques is determined by the Confusion Matrix.

- Accuracy
- Precision
- Recall
- F-Measure

4. Experimental Result

4.1 Data analysis to test hypotheses

For Hypothesis 1, drinking alcohol affects deaths from road accidents that occur during Songkran Festival. The following statistical hypothesis is:

H_0: Drinking alcohol does not relate deaths from road accidents occurring during Songkran.
H_1: Drinking alcohol relates deaths from accidents that occur during Songkran.

From Table 2 studying factors of drinking alcohol influencing on death from road accidents, it is found that the people who drank alcohol have the highest impact on road death during Songkran Festival, with Pearson Chi-Square = 74.025 and P-Value = 0.000 ($< \alpha = 0.05$). That is rejecting the null hypothesis. It can interpret that the loss of life of people who drink and those who do not drink differently or alcohol drinking factors are associated with deaths from road accidents that occurred during Songkran Festival with statistical significance of 0.05**.**

Table 2: Two-way table between drinking alcohol and deaths

Y / X	Death	Not Death	Total
Drink	933 (64.88%)	704 (48.99%)	1,637
Not drink	505 (35.12%)	733 (51.01%)	1,238
Total	1,438 (100%)	1,437 (100%)	2,875

For Hypothesis 2: Wearing a helmet or safety belt is associated with death from road accident that occurred during the Songkran Holiday. A statistical hypothesis can be written as follows:

H_0: Wearing a helmet or safety belt does not affect the death from an accident that occurred during Songkran.
H_1: Wearing a helmet or safety belt affects the death from an accident that occurred during Songkran.

From Table 3, studying factors of wearing a helmet or safety belt influencing on death from road, it is found that the people that did not wear a helmet or safety belt have the highest impact deaths on road during the Songkran Festival. The result of the relationship test of wearing a helmet or safety belt affects the death from an accident that occurred during the Songkran holiday, with Pearson Chi-Square = 50.372 and P-Value = 0.000 ($< \alpha = 0.05$). That is rejecting the null hypothesis. It can interpret that the loss of life of the person who wears the helmet or safety belt and the person who does not wear a helmet or safety belt is different or the wearing a helmet or safety belt is related to the death from the road accident occurred during Songkran with statistical significance at the level of 0.05.

Table 3: Two-way table between wearing a helmet and deaths

Y / X	Death	Not Death	Total
Wearing a helmet	1,238 (86.09%)	1,351 (94.02%)	2,589
Without wearing a helmet	200 (13.91%)	86 (5.98%)	286
Total	1,438 (100%)	1,437 (100%)	2,875

Hypothesis 3, the type of vehicle is related to death from road accidents that occurred during Songkran holidays. A statistical hypothesis can be written as follows:

H_0: Vehicle type does not affect death from accidents occurring during Songkran Festival.
H_1: Vehicle type affects deaths from accidents occurring during Songkran Festival.

From Table 4, studying the factor of vehicle type influencing on death from road accidents It was found that the motorcycle has a significant impact on road accidents during Songkran Festival when comparing to other types of vehicles. The results of the type of vehicle have resulted in deaths from accidents occurring during the Songkran holidays with Pearson Chi-Square = 103.196 and P-Value = 0.000 ($< \alpha = 0.05$). That is rejecting the null hypothesis. It can interpret that the type of vehicles are associated to deaths from accidents that occurred during Songkran holiday with statistical significance at the level of 0.05.

Table 4: Two-way table between type of vehicles and deaths

Y / X	Death	Not Death	Total
Single Vehicle	1,157 (80.46%)	973 (67.50%)	2,127
Motorcycle	98 (6.82%)	199 (13.85%	297
Bicycle	55 (3.83%)	93 (6.47%)	148
Pickup	15 (1.04%)	61 (4.25%)	76
Taxi	61 (4.24%)	41 (2.85%)	102
Van	6 (0.42%)	15 (1.04%)	21
Bus	10 (0.70%)	7 (0.49%)	17
Truck	5 (0.35%)	14 (0.97%)	19
Tricycle	20 (1.39%)	18 (1.25%)	38
Other	11 (0.77%)	19 (1.32%)	30
Total	1,438 (100%)	1,437 (100%)	2,875

For Hypothesis 4, the time period is related to death from accidents that occurred during Songkran holidays. A statistical hypothesis can be written as follows:

H_0: The time period does not affect deaths from road accidents occurring during Songkran.
H_1: The time period affects deaths from road accidents occurring during Songkran.

From Table 5, studying factors of time period influencing on death from road accidents, it is found that during the period time of 18.01-24.00 has related to death from road accidents during the Songkran Festival. The results show the period of time has associated to deaths from accidents occurring during Songkran holidays with Pearson Chi-Square = 34.068 and P-Value = 0.000 (< α = 0.05). That is rejecting the null hypothesis. It can interpret that time is related to death from an road accident that occurred during Songkran, with statistical significance at the level of 0.05.

Table 5: Two-way table between the period of time and deaths

Y / X	Death	Not Death	Total
06:01-12:00	285 (19.82%)	314 (21.85%)	599
12:01-18:00	593 (41.24%)	462 (32.15%)	1055
18:01-24:00	442 (30.74%))	475 (33.05%))	917
24:01-06:00	118 (8.21%)	186 (12.94%))	304
Total	1,438 (100%)	1,437 (100%)	2,875 (100%)

Hypothesis 5, the type of road is related to death from accidents that occurred during Songkran holidays. A statistical hypothesis can be written as a statistical hypothesis as follows

H_0: Type of road has no effect on death from accidents that occurred during Songkran.
H_1: Type of road affecting deaths from accidents that occurred during Songkran

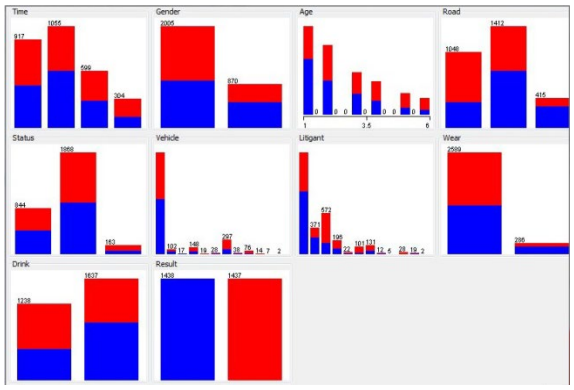
From Table 6 comparing road types and death from road accidents It was found that the sample who drove on the highway had an effect on road accident during Songkran Festival, which is the highest level compared to other road types. The road relationship test results have an effect on the deaths from accidents occurring during the Songkran holidays. With the Chi Square statistics From the results table, the results show the Chi-Square test statistic. It was found that Pearson Chi-Square equals 205.390 and the P-Value is equal to .000, which is less than the specified level (= 0.05). Of roads is associated with deaths from accidents that occurred during Songkran With statistical significance at the level of 0.05.

Table 6: Two-way table between the types of roads and deaths

Y / X	Death	Not Death	Total
City streets	295 (20.51%)	120 (8.35%)	415
Rural road	790 (54.94%)	622 (43.28%)	1,412
Highway	353 (24.55%)	695 (48.36%)	1,048
Total	1,438 (100.00%)	1,437 (100%)	2,875 (100%)

4.2 Data Classification By Data Mining Techniques

Figure 3 shows the data classified by the characteristics of various attributes: the time of road accident, gender of people who suffer from road accidents, age of people who suffer from road accidents, road types occurring accidents, driving status of people who suffer from road accidents, vehicle types of people who suffer from road accidents, vehicle types of the parties, wearing a helmet or a safety belt, drinking alcohol.



Figures 3: The value of attributes for analysis in the Weka application.

Classification rules are obtained from decision trees. The conditions of the victims from road accidents are the motorcyclists who drink alcohol and do not wear a helmet, with no parties during period of time 06.01 - 12.00 and 24:01 - 06:00 or the parties who drive motorcycles and drink alcohol.

Table 7: A comparisons of measures among classification techniques

Measures	Decision Tree	Naïve Bayes	Logistic Regression	NN
Accuracy	66.9%	70.4%	72.2%	64.3%
Precision	66.9%	70.6%	72.2%	64.3%
Recall	66.9%	70.4%	72.2%	64.3%
F	66.9%	70.4%	72.2%	64.2%

The comparisons of classification accuracy among data mining techniques are shown as in Table 7. It can be seen that logistic regression methods can be more accurately classified than other classification techniques.

5. CONCLUSIONS

From the analysis of the chi-square test, it was found that drinking alcohol, wearing a helmet, types of vehicles driving, time and road accidents that are associated with deaths during Songkran holidays. In addition, there are the following important conclusions.

- Injuries and deaths from road accidents that drink alcohol are most common on rural roads than other roads.
- Injuries and deaths from road accidents that do not wear a helmet or safety belt are most common on rural roads than other roads
- Any motorcyclists without wearing a helmet or safety belt have a risk of death more than those wearing a helmet or safety belt 1.59 times

When comparing data classification results among data mining techniques, it was found that logistic regression can be more accurately classified than other classification techniques with the accuracy of 72.20%.

References

Digital Government Development Agency (DGA), 2018, Injuries and death during 2008 - 2014 festival, Retrieved December, 2, 2018, from https://data.go.th/DatasetDetail.aspx?id=f33db12e-bfd5-4eec-aa8a-e5b3ff6e1cb1.

Domingos, P. and Pazzani, M., Beyond Independence Conditions for the Optimality of the Simple Bayesain Classifier. *Proceedings of the 13th International Conference of Machine Learning*, Morgan Kaufmann, pp. 105-112, 1996.

Kaewta, C. and Mahaweerawat, A. 2010. Diagnosis of cases with decision tree techniques, Retrieved December 2, 2018, from http://home.kku.ac.th/wichuda/DMining/CU/EX_Lawsuit.pdf. (in Thai).

Ocharoen, N. 2017. Safety on the road, Thailand Development Research Institute, Retrieved January 7, 2019, from https://tdri.or.th/2017/08/econ_traffic_accidents/ (in Thai).

Srichaiwong, C., Takoonsuk P. and Boonlue, S., Decision Support Systems for Longan Leaf Disease Diagnosis with Decision Tree, *Veridian E-Journal Science and Technology Silpakorn University*, vol. 1, no.6, pp. 1-14, 2014. (in Thai).

Suangka, K.. A study of young driver behaviour that affect the risk of accidents from the motorcycle, Research Report, Suranaree University of Technology, 2016. (in Thai).

Tanaboribun, Y. Trends in road accidents and relationships with risk factors related to road accidents, Research Report, Health Systems Research Institute, 2005. (in Thai).

World Health Organization (WHO), 10 January 2019, Strengthening Road Safety in Thailand, Retrieved from http://www.searo.who.int/thailand/areas/roadsafety/en

Biography

Pornpimol Chaiwuttisak is a lecturer in Department of statistics i at King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. She earned B.S. in Computer Science from Kasetsart University, Bangkok, Thailand and Masters in Information and Systems Management from National Institute of Development Administration, Bangkok, Thailand and Masters in Operational Research from University of Southampton, United Kingdom and PhD in Operational Research from University of Southampton, United Kingdom. She has published journal and conference papers. Dr Pornpimol worked research projects with Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) team, University of Lille, France. Her research interests include Heuristics, Optimization, Logistics, and Data Analytics.