Heliyon 8 (2022) e10302

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A knowledge elicitation approach to traffic accident analysis in open data: comparing periods before and after the Covid-19 outbreak

ChienHsing Wu [a,*], Shu-Chen Kao [b], Chia-Chen Chang [a]

[a] National University of Kaohsiung, Kaohsiung, Taiwan
[b] Kun Shan University, Tainan, Taiwan

HIGHLIGHTS

- Determinants associated with traffic injuries are revealed using knowledge elicitation approach.
- Open data before and after Covid-19 pandemics are considered.
- Determinants in Human and Vehicle categories revealed higher classification ranks.
- Motorcycle accidents and injury on leg or foot were 5.13% and 3.46% higher after Covid-19 pandemic.
- Support of rules and simplicity of decision tree were higher after Covid-19 pandemic.

ARTICLE INFO

Keywords:
Knowledge elicitation
Traffic accidents
Open government data
Covid-19 pandemic

ABSTRACT

Extracting knowledge from open data of traffic accidents has been attracting increasing attention to policymakers responsible for road safety. This article presents a knowledge elicitation approach to exploring the determinants of traffic accidents from open government data of an urban area in Taiwan. The collected open dataset contains 34 decisional attributes and one predictive attribute (i.e., type of injury, including head, breast, leg), and 47,974 cases. Prediction models using a classification-oriented mechanism and generated rules that considered datasets from before (*B-dataset*; 30,116 cases) and after (*A-dataset*; 17,868 cases) beginning to combat the Covid-19 pandemic in an urban area of Taiwan were compared. The findings showed that prediction accuracy was acceptable but not high, at 70.73% for *B-dataset* and 74.77% for *A-dataset*. Determinants in the human and vehicle categories revealed higher classification ranks than those in the temporal and environment categories. Traffic accidents involving motorcycles were 5.13% higher in *A-dataset*, whereas those involving cars were 4.11% lower. Injury on leg or foot was 3.46% higher in *A-dataset*, whereas other types of injury were up to 1.00% lower. The average support for rules in the *A-dataset* rule base and the simplicity of the *A-dataset* decision tree were higher than those of *B-dataset*. The research demonstrates the value of open government data in prediction model development and knowledge elicitation to support policymaking in the traffic safety domain.

1. Introduction

As road traffic accidents are among the leading contributors of injuries and fatalities, the development of traffic accident analysis and prediction models is an important field of research (Chand et al., 2021; Sangkharat et al., 2021; Tavakoli and Heydarian, 2022). All traffic accidents have a cost, and this cost can even be immeasurable when human injuries and fatalities are involved (French et al., 2009; Kaygisiz et al., 2017). The total cost of traffic accidents is often expressed as a certain percentage of a country's gross domestic product (Elvik, 2000; Connelly and Supangan, 2006; Law et al., 2009), showing the obviously negative impact of accidents on national economies (Vipin and Rahul, 2021). Exploring the key factors behind traffic accidents that lead to injuries and fatalities is one of the main objectives of research seeking to develop insights to support traffic safety policymaking. There are a variety of reasons for the serious problem of a high volume of traffic accidents, and some possible solutions with respect to policies and strategies, such as helmet and drink-driving laws, have been proposed and implemented (Vorel et al., 2014; George et al., 2017; Alcaniz et al., 2021). Prediction models that can be used to disclose causes of traffic accidents in historical

* Corresponding author.
E-mail addresses: chwu@nuk.edu.tw, ch9509@gmail.com (C. Wu).

https://doi.org/10.1016/j.heliyon.2022.e10302
Received 30 March 2022; Received in revised form 2 June 2022; Accepted 11 August 2022

data are proposed by researchers to aid in developing such strategies to reduce traffic accidents (Li and Zhao, 2022; Roy et al., 2021; Valent, 2022; Sangkharat et al., 2021; Olowosegun et al., 2022; Antona-Makoshi et al., 2018).

The models that have been proposed to predict accidents have adopted various data sources and analysis techniques (Chand et al., 2021; Li and Zhao, 2022; Olowosegun et al., 2022; Roy et al., 2021). For example, classification-oriented models (decision trees) were proposed to discover decision rules to detect traffic accidents (de Oña et al., 2014; da Cruz da Cruz Figueira et al., 2017). Statistical models were used to estimate the frequencies and risk of brain injuries (Antona-Makoshi et al., 2018), and the examination of the effect of road markings on drivers' compliance with speed limits was conducted (Charlton et al., 2018). Moreover, the analysis of road traffic accidents was presented using exploratory data analysis and time series regression (Vipin and Rahul, 2021), while regression models were employed to predict traffic accidents (Kaygisiz et al., 2017) and identify determinants of accident severity (Ratanavaraha and Suangka, 2014). More recently, accident detection and road condition analysis using a social network-based real-time monitoring framework was proposed (Ali et al., 2021), and the examination of the relationships between temperature and traffic accidents using a generalized additive model and meta-analysis was conducted (Park et al., 2021). The impact of COVID-19 travel-restriction policies on road traffic accident patterns was also examined using statistical models (Li and Zhao, 2022).

The different variables used in these models can be grouped into various categories. For example, legal variables mainly include drink-driving laws and helmet laws, whereas socioeconomic variables have number of vehicles and price of fuel (Chen and Liu, 2012; Vorel et al., 2014). Variables in the social influence category contain comments and complaints (Ali et al., 2021; Parady et al., 2020) whereas seatbelts, bumpers, and vehicle types are included in vehicle category (da Cruz da Cruz Figueira et al., 2017; Antona-Makoshi et al., 2018; Li et al., 2021). Such variables as age, alcohol intake, and speed are used in the human category (Ratanavaraha and Suangka, 2014; de Oña et al., 2014; Alcaniz et al., 2021). Variables of temporal category are month, day, and time zooms (Vipin and Rahul, 2021; Park et al., 2021; Li et al., 2021; Li and Zhao, 2022), whereas environment category includes traffic volume, weather, road condition, and lighting (Ratanavaraha and Suangka, 2014; Charlton et al., 2018; Kaygisiz et al., 2017; Park et al., 2021; Sangkharat et al., 2021).

Moreover, since 2019, the Covid-19 pandemic has had a considerable impact on behavior in using transportation systems and has thereby influenced road traffic accidents (Chen and Pan, 2020; Valent, 2022; Li and Zhao, 2022). For example, the high risk perception of Covid-19 infection produced more self-restriction behaviors (e.g., deciding not to travel, avoiding crowds), especially in relation to eating out and leisure activities (Parady et al., 2020). To reduce the likelihood of infection, workers or travelers began to prefer personal vehicles over public transportation systems. It is estimated that 5.3% of commuters shifted from public to private transport modes due to the Covid-19 pandemic (Pawar et al., 2020). These changes are likely to have an influence on traffic accidents, and this needs to be investigated to adapt relevant safety policies.

Although studies have presented important general aspects and identified the most remarkable antecedents of traffic accidents, understanding the ranking of influence for different types of variables, such as temporal, human, vehicle, and environmental factors, is particularly important to road safety policymaking. Moreover, the use of knowledge elicitation techniques on real cases, such as with open government data, is a valuable approach to developing insights for devising road safety policies and strategies.

To address these issues, a classification-oriented prediction model applied to open government data to reveal determinants of traffic accidents is proposed and implemented in this paper. The association strength with respect to the classification power (CP) of influencing variables is analyzed. The elicited knowledge is presented in the form of decision rules. The study has three objectives. The first objective is to develop a classification-oriented traffic accident prediction model. The second objective is to conduct knowledge elicitation from the open government data of Taoyuan city, Taiwan, a municipality with a population of 2,272,452 and population density of 1,861.21 persons per square kilometer. The third objective is to compare the results before and after the date (January 24, 2020) that the Taiwan government started combating the Covid-19 pandemic. The open dataset contains 48,055 valid accident cases with confirmed injuries within the period of January 1, 2017 to June 30, 2021.

2. Literature review

2.1. Prediction of traffic accidents

The literature identifies the growing importance of traffic accidents as the main cause of injuries and fatalities (Olowosegun et al., 2022; Roy et al., 2021; Valent, 2022; Ratanavaraha and Suangka, 2014; Xin et al., 2020). Various determinants have been considered and examined in association with traffic accidents and models have been developed to explain these determinants. The ultimate aim of this body of research is to recognize and reduce the likelihood of traffic accidents and to provide suggestions to support road safety policymaking. The relevant literature is summarized in this section.

The factors are generally divided into up to four main operating categories: human, vehicle, environment, and legal and socioeconomic. Whereas the early studies often considered two or three categories, more recent studies have tended to focus on a single category. For example, some studies considered human, vehicle, and environment factors together for their contribution to traffic accidents or road safety risks (de Oña et al., 2014; Kwon et al., 2015; Altwaijri et al., 2012). Other studies have analyzed only the environmental factors in road accidents (Ali et al., 2021; Kaygisiz et al., 2017; van Wee et al., 2019; Charlton et al., 2018). Furthermore, one study considered legal and socioeconomic factors from a governmental perspective (Vorel et al., 2014), one proposed a social network-based model to detect traffic accidents and road flow (Ali et al., 2021), two analyzed traffic accident mortality based on temporal factors (Vipiv & Rahul, 2021; Park et al., 2021), and one looked into motorcyclist injuries using a spatiotemporal analysis (Li et al., 2021).

The influence of road characteristics on traffic accidents has also been examined using a linear model (Fernandes & Neves, 2013), Lighting was found to be one of the determinants of traffic accidents by de Oña et al. (2014), and another study confirmed the benefits of road marking to drivers' compliance with speed limits (Charlton et al., 2018). Online community opinions have been used to detect traffic accidents and road flow using a social network-based real-time monitoring framework (Ali et al., 2021), through which environmental factors were found to be considered the most important determinant of traffic accidents. Despite the limitations of this method (e.g., data sources, data quality, labeling quality, and reliance on overly subjective opinions), it is clear that these opinions do have value for examining traffic accidents.

To provide input into road safety policies based on time zone, temporal variations have been examined by forecasting specific road traffic accidents in particular time zones based on exploratory data analysis and time series regression (Vipiv & Rahul, 2021). The effects of heat at different times of the day have also been examined to provide information conducive to traffic accident reduction policies (Park et al., 2021). In the study of Li et al. (2021), the relationships of spatial and temporal variables to motorcyclist injury severity were examined using non-stationary tests and it was found that the helmet, engine size, vehicle age, pillion passenger, at-fault striking, and speeding were significant factors.

In terms of method, all studies have used quantitative approaches, including binary logic, regression models, naïve Bayes, decision trees,

linear models, text mining, machine learning, and general statistics. The variables used in each category vary greatly with availability, and the findings are not consistent. For example, age and other drivers' characteristics have been identified as important factors in traffic accidents (de Oña et al., 2014) and road crashes (Altwaijri et al., 2012), but also as non-significant determinants of safety risk (Kwon et al., 2015) and accident severity (Ratanavaraha and Suangka, 2014). As a consequence, the implications and suggestions lack adequate focus. Furthermore, the data sources are diverse and can be either linguistic or numeric in form, or both; various studies have drawn on online comments and opinions, open government data, or survey data. This review shows that studies have explored a range of possibilities to derive findings from various proposed models drawing on various types of data collected from multiple channels to support traffic safety policymaking.

2.2. Effects of the Covid-19 pandemic on traffic conditions

The Covid-19 pandemic has fundamentally influenced the lifestyles and travel behavior of individuals across communities in several ways, such as remote working, social distancing, self-isolation, eating at home, and changes to leisure activities (Pawar et al., 2020; Hotle et al., 2020). These changes may have unpredictable consequences for traffic accidents. Studies that have explored the general effects of the Covid-19 pandemic are contributing to the adaptation to the impact of the disease, the progress of which remains unpredictable. For example, the factors that lead to behavioral changes have been examined in an online panel survey focusing on risk perception and social influence (Parady et al., 2020). The main finding suggested that targeting the avoidance of non-essential travel might be effective in addressing the severity of Covid-19, particularly among groups that have difficulty maintaining social distancing. Based on protection motivation theory, the effect of risk perception on travel to various locations has also been examined (Hotle et al., 2020). It was found that individuals would reduce the number of trips they took if they perceived medium or high levels of risk, but this was not the case for travel to workplaces, even when the perceived risks were high.

A study of the impact of Covid-19-related lockdowns on traffic accidents from March 16 to April 26, 2020 estimated that the number of accidents per day decreased by 74.3% in comparison with the previous week and by 76% with the previous year (Saladié et al., 2020) due to the decrease of mobility. This shows that a reduction in traffic associated with measures to control the Covid-19 pandemic drastically decreases the number of traffic accidents. To deepen the exploration of the effects of Covid-19 on traffic accidents, however, the changes in the determinants of accidents before and after lockdown need to be explored in more detail.

2.3. Classification-based technique

Classification-oriented knowledge elicitation techniques have been successfully applied in various domains (Quinlan, 1986; Lausch et al., 2015; Wu and Kao, 2021), and have the advantageous features of an entirely data-driven approach, learnability, high classification accuracy, and multi-context datasets, particularly when the data is characterized by multi-dimensionality, multi-collinearity, and non-homogeneity. Although it is not a new model, variants of the classification tree algorithm, such as ID3, C4.5, CHAID, CART (Ture et al., 2009), random forest (RF; Breiman, 2001), and their extensions (Rao et al., 2019), have demonstrated high applicability.

A decision tree is a promising mechanism when considering a classification-based prediction model (Rao et al., 2019; Wu and Kao, 2021). By using the ID3 algorithm to calculate the CP of an attribute, the C4.5 considers the pruning and non-pruning ability by computing the gain ratio of an attribute to exclude nodes and leaves that are unable to expand for a defined goal (Quinlan, 1993). It is a data-driven and top-down classification technique to return a decision tree determined by the entropy of an attribute, with higher entropy implying a higher power of classification. The C4.5 was used in the present study as the mechanism to reveal the CP of decisional attributes from which decision rules can be derived.

2.4. An urban area in Taiwan as a case study site

Taiwan's rapid economic growth since the 1980s has increased its traffic volume and thereby made it prone to high numbers of traffic accidents. There were 362,393 vehicle accident cases reported in Taiwan in 2020, which was an increase over the 341,972 reported in 2019 (Statistics of Vehicle Accidents, 2019). An urban area in Taiwan was adopted as the case study site for this study to demonstrate the achievement of the three research objectives and the presentation of findings and implications. The original dataset was collected from the open government repository of traffic accident data for the municipality of Taoyuan for the January 1, 2017 to June 30, 2021 period (https://data.gov.tw/). The dataset held 300,376 cases involving injuries and no injuries. After removing the non-injury cases, there were 47,974 records of reported accidents with confirmed injuries to the head, leg, breast, etc. The collected dataset contains 34 decisional attributes in four categories (temporal, environmental, human, vehicle) and one predictive attribute (i.e., injury). The pre-processed original dataset was split into two datasets based on the date January 24, 2021, which is when the government of Taiwan started implementing measures to contain the Covid-19 pandemic.

3. Method

3.1. Research design

This study was conducted in four phases: preparation, implementation, validation and comparison, and knowledge elicitation in the form of decision rules. The implementation procedure of conducting the study is presented in Table 1, which summarizes the characteristics and required tasks of the experimental design in detail regarding research objectives, open data collection, data pre-processing, attribute dimension reduction, elicitation mechanism, training and testing, and knowledge discovery.

The preparation stage involved data collection and pre-processing, which included the elimination of missing data in either cases or attributes, the granulation of continuous data types, and the splitting of the original dataset into before and after datasets according to the critical date of the containment of the Covid-19 pandemic in Taiwan. In the implementation stage, the information entropy algorithm was used to rank the CP of attributes (or variables), and the validation stage used the criteria of 70% for training and 30% for testing to evaluate the mined rules, with accuracy representing how accurate the mined rules predicted the 30% test cases. The comparison stage analyzed the differences between the two datasets with respect to the determinants of traffic accidents. The elicited knowledge presents the main findings in the form of decision rules with high levels of support and simplicity, followed by implications and suggestions.

3.2. Variables

To reveal the importance ranks of the attribute categories to the predictive attribute for the dataset, the decisional attributes were grouped into four categories, namely *temporal*, *environmental*, *human*, and *vehicle*. The dependent variable (class) was *injury*, with eight possible values: *head*, *neck*, *breast*, *abdomen*, *waist*, *back*, *hand* (including wrist), and *leg* (including foot). There were five variables in the temporal category (*year*, *month*, *day*, *week*, and *hour*), 15 variables in the environmental category (e.g., *weather*, *light*, *road type*, *road condition*, *speed limit*, *signal*), 12 variables in the human category (*nationality*, *gender*, *age*, *license type*, *alcohol*, *protection*), and two in the vehicle category (*vehicle type* and *collision point*). The range of values was specific to each variable: for

Table 1. Research design.

Feature	Description
Objectives	(1) Disclose determinants of traffic accidents using a data mining approach (2) Discover knowledge for traffic accident prediction (3) Compare traffic accidents before and after the outbreak of the Covid-19 pandemic regarding the main determinants of traffic accidents, the vehicle types (car, bus, motorcycle) and injury types (e.g., head, breast, leg, etc.), and the decision rules extracted.
Open dataset collection	(1) Collect open government data of vehicle accidents in Taoyuan city from January 1, 2017 to June 30, 2021 (2) Original dataset comprises 34 decisional attributes (e.g., weather, city road, speed limit, collision point, age) and one predictive attribute (injury) (3) Original dataset comprises 47,974 original cases with injury labels (e.g., head, breast, leg, etc.)
Data pre-processing	(1) Granulate continuous attributes using the equal with interval technique (2) Group decisional attributes into four categories: temporal (5 attributes), environmental (15 attributes), human (12 attributes), and vehicle (2 attributes) (3) Divide original dataset into two subsets (B-dataset and A-dataset) according to the critical date of the impact of Covid-19 in Taiwan (January 24, 2020)
Attribute dimension reduction	(1) Reduce attribute dimensions based on the CP using the C4.5 algorithm (2) Remove inconsistent granulated data based on the 60% consistency acceptance level (Wu et al., 2013)
Mining mechanism	(1) Conduct data mining with a tree-based algorithm (C4.5) with size of leaves equal to 2 (2) Consider entire consistent (cleaned) granulated dataset
Training and testing	(1) Use criteria of 70% for training and 30% for testing (2) Obtain prediction accuracy
Knowledge discovered	(1) Obtain decision rules with depth, support, and reliability (2) Present findings and discuss implications of generated rules with high levels of support (e.g., 50)

example, week (temporal) took one of seven values corresponding to the days of the week, road type (environmental) took one of eight values (e.g., "national highway," "provincial highway," "country road"), license type (human) took one of five values (e.g., "professional," "regular," "motorcycle"), vehicle type (vehicle) took one of three values ("car," "bus," or "motorcycle").

3.3. Data pre-processing

Three main pre-processing tasks were undertaken: the granulation of continuous data types, data separation, the reduction of decisional attributes, and the removal of inconsistent data. First, the original dataset contained three continuous attributes (day and hour in the temporal category, and age in the human category), which were granulated using equal width intervals of 3, 6, and 10, respectively, to meet the requirement of discrete data types for the operation of the classification-oriented mining mechanism.

Second, the new granulated dataset was divided into two subsets. *B-dataset* covered the period before the Covid-19 containment efforts began, from January 1, 2017 to January 23, 2020, and *A-dataset* covered the later period, from January 24, 2020 to June 30, 2021. The date of January 24, 2020 was selected as the critical data for Taiwan's efforts to contain the pandemic as this was when the Ministry of Health and Welfare announced that the export of medical-grade and N95 face masks was banned and when customs checks began to increase (Crucial Policies for Combating Covid-19, 2020). This announcement conveyed an important message that greatly influenced many subsequent government policies, the entire field of social and economic activities, and public and private social behavior in the direction of protecting public and private safety, and marked the key moment when individuals began to adapt to the need to avoid infection and maintain their health. *B-dataset* contained 30,116 cases and *A-dataset* contained 17,868 cases, with both datasets analyzed to achieve the research objectives.

Third, the prediction accuracy (PA) of the two datasets was estimated and attribute selection used to reduce the dimensions. The CP of the 34 decisional attributes was obtained using the C4.5 algorithm (Quinlan, 1986, 1993; Ture et al., 2009) for both datasets. The reduction of decisional attributes was then conducted according to the CP ranks given by the algorithm. The reduced datasets were labelled *B-dataset-topN* and *A-dataset-topN*. The C4.5 algorithm is presented in Eq. (1) to compute information entropy (InEn), in Eq. (2) to compute the expected information entropy for an attribute [EIE (Attribute)], in Eq. (3) to determine the CP of the attribute [CP(Attribute)], and in Eq. (4) to compute the gain ratio for the attribute [GaRa(Attribute)].

$$InEn(nc_1, nc_2, \ldots nc_n) = \left(-\frac{nc_1}{T}\log 2\frac{nc_1}{T}\right) + \ldots + \left(-\frac{nc_n}{T}\log 2\frac{nc_n}{T}\right) \quad (1)$$

n_{C_i}: The number of records that return to class C_i, $i = 1, 2, \ldots, n$

T: The total number of tuples.

$$EIE(Attribute) = \sum_{i=1}^{t}\left[\left(\frac{nvi}{T}\right)InEn(avic_1, avic_2, \ldots, avic_m)\right] \quad (2)$$

t: The number of different values that *Attribute* can take on.

n_{vi}: The total number of records that *Attribute* takes on value V_i, $i = 1,2,\ldots,t$.

avicj: The total number of records that *Attribute* takes on value V_i and returns to class C_j, $i = 1, 2, \ldots, t$, $j = 1,2,\ldots,m$.

T: The total number of records.

$$CP(Attribute) = InEn(nc_1, nc_2, \ldots nc_n) - EIE(Attribute) \quad (3)$$

$$GaRa(\text{Attribute}) = \frac{CP(Attribute)}{EIE(Attribute)} \quad (4)$$

Finally, inconsistent data means that the same conditions can produce different conclusions. Given the pre-defined attribute spaces, inconsistent data will very possibly lead to prediction failure and meaningless findings (Wu et al., 2013). For example, the decision rule derived from the inconsistent cases that "*if* sit in the back seat and fasten seat belt *then* you may or may not get your head injured in a traffic accident" appears useless because of its contradictory inferences and low reliability. In reality, such cases may not be uncommon because decisional attributes are ill-defined or not pre-defined, but they are not considered in our research.

To avoid discovering meaningless or unreliable decision rules, detection and removal for the purpose of data inconsistency was conducted for both datasets (*B-dataset* and *A-dataset*) using the model proposed by Wu et al. (2013) with a defined consistency acceptance level (CAL). To navigate the dilemma that too much inconsistent data is removed if CAL is set too high and too few reliable rules are generated if it is set too low, the CAL was set at 60%. For example, assume an inconsistent subset (IS) containing 10 cases with the same decisional attribute values, but three different predictive attribute values (PAV1, PAV2, and PAV3) for which PAV1 has seven cases, PAV2 has one, and

PAV3 has two. The inconsistency rate for PAV1, PAV2, and PAV3 would be 0.70, 0.10, and 0.20, respectively. With a CAL of 60%, the seven cases belonging to PAV1 would remain while one belonging to PAV2 and two belonging to PAV3 would be removed from the IS. The datasets after cleaning to remove inconsistent data were labelled *B-dataset-topN-clean* and *A-dataset-topN-clean.*

3.4. Experimental settings

3.4.1. Training and testing

B-dataset-topN-clean and *A-dataset-topN-clean* were trained and tested to examine the PA based on a 70/30% split. Each dataset was randomly divided into a training subset with 70% of cases to be used for training the mining algorithm and a testing subset with the remaining 30% cases to be used for testing. On the one hand, the mining algorithm was J48, carried out with the WEKA machine learning tool, which is an open source software that is easy to use. The J48 is based on the C4.5 algorithm (Quinlan, 1993) to generate a pruned or unpruned decision tree. The WEKA contains tools for data pre-processing, classification, regression, clustering, association rules mining, and visualization. Despite drawbacks, such as inefficiency for large volume of datasets, WEKA is considered suitable for the research because of the granulation of continuous data, data volume of less than 50, 000, and decisional attributes that are less than 40. The prediction accuracies from C4.5 for before and after cleaning inconsistent data are revealed to ensure the suitability of the knowledge elicitation process. On the other hand, the research develops a rule generation algorithm to elicit knowledge in the form of a decision rule, which is not covered by WEKA.

3.4.2. Knowledge elicitation

To comprehensively consider the entire dataset when generating decision rules, the dataset was divided into several subsets according to the unique combinations of decisional attributes. To make this division, the decisional attributes of the dataset were re-ordered according to the CP determined by C4.5. It was then sorted using the entire body of decisional attributes. A rule was generated when a unique combination of attribute values returned a single predictive attribute class. The number of attributes involved (*depth*), the number of cases (support), *reliability*, and *simplicity* [support $\times$ (1/*depth*)] were computed for every generated rule. The rule generation algorithm (RGA), using a depth-first approach to generate rules from the dataset (Quinlan, 1986, 1993), is shown below.

The RGA contains three main steps. First, it retrieves the features of the dataset being processed (e.g., value space of each decisional attributes). Second, it generates rules from cases where the first attribute appears to have the same attribute value and class, after which we removed these cases associated with the rule from the original dataset. It processes the remining dataset by using a depth-first approach that generates a subset based on the first-rank attribute and re-rank the remaining decisional attributes using the C4.5 algorithm. Finally, it uses the same procedure until same attribute value reaches the same class. This procedure does not stop until the remaining dataset is empty.

A generated rule contains a set of attribute values (e.g., *speed limit* = "50 km/h" and *collision point* = "left side,"), the type of injury (e.g., head, leg), depth (e.g., 9), support (e.g., 40), reliability (=1.0), and simplicity (e.g., 20.3). The support represents the volume of cases that are described by the rule. The reliability of every generated rule is 1.00 because

```
Begin
  Read dataset named DT;
  Compute gain for attributes of DT using classification algorithm;
  Reorganize DT attributes according to attribute gain, denoted by DT-gain;
  Create an empty rule base with the same attribute of DT-gain, named RuleBase;
                                        //Store generate rules
  Sorting DT-gain with all attributes;
  Compute size of value space for each attribute of DT-gain;
                                        //denoted N1, N2, ... Nm, m=the i^th attribute
  Do while not-empty for DT-gain;
    i=1;
    Do while i < m
      DFRG=If attribute value leads to same class; // Depth-first rule generation
            Generate rules and store rules in RuleBase in corresponding attribute and class;
                                                    //Rule generation
          Else
            Remove and create the subset (S1) using the attribute value from the remaining
              dataset; //First iteration is DT-gain and attribute size decreases
            Compute gain for attributes of the subset; //Re-rank attributes
          Endif
        Repeat DFRG until rule generation completion for the subset (S1) with the ith attribute
          value;
        i=i+1;
    EndDo
  EndDo
End
```

Table 2. Ranks of CP of decisional attributes for B-dataset and A-dataset.

Category and attribute code	Attribute name	Ranks from C4.5	
Temporal		B-dataset	A-dataset
X01	Year	33	34
X02	Month	30	24
X03	Day	34	33
X04	Week	32	32
X05	Hour	26	29
Environmental			
X06	Weather (e.g., rain, storm, cloudy)	31	30
X07	Light (e.g., dawn, dusk)	25	27
X08	Road category (e.g., city road, country road)	14	21
X09	Road type (e.g. railway, bridge, multiple intersection, overpass)	18	14
X10	Speed limit (e.g., 50 km/h)	17	12
X11	Road condition (e.g., railroad crossing, single lane)	22	19
X12	Accident location (e.g., intersection, road section)	19	18
X13	Accident site (e.g., left turn waiting zone, U-turn lane)	15	15
X14	Road pavement (e.g., concrete, gravel)	9	5
X15	Road pavement condition (e.g., snow or ice, wet, slippery)	29	26
X16	Road pavement defect (e.g., bumpy, soft surface)	13	7
X17	Road obstacles (e.g., under construction, parked vehicle)	21	22
X18	Line of sight (e.g., curve, slope)	24	23
X19	Signal type (e.g., traffic light, flashing signal)	28	31
X20	Signal condition (e.g., normal, abnormal)	27	28
Human			
X21	Nationality (e.g., Taiwan, non-Taiwan)	23	25
X22	Gender (e.g., male, female)	16	16
X23	Age	12	13
X24	Occupation (e.g., business man, affair worker)	11	17
X25	Travel purpose (e.g., for work, for school)	20	20
X26	Behavioral condition (e.g., parking, left-turning)	10	11
X27	License status (e.g., legal, suspended)	8	10
X28	License type (e.g., professional, regular)	2	2
X29	License vehicle type (e.g., trailer, car, heavy, motorcycle)	3	3
X30	Alcohol (e.g., zero or less than 0.15 mg/L, between 0.16 and 0.25)	5	6
X31	Device use (e.g., cellphone)	7	9
X32	Safety device use (e.g., helmet, no seatbelt)	6	8
Vehicle			
X33	Vehicle type (e.g., bus, car, motorcycle)	1	1
X34	Collision point (e.g., front end, left side)	4	4

inconsistent data was removed at the pre-processing stage. Simplicity is based on the concept that more support and fewer conditions of a rule implies better results for knowledge elicitation. It represents how precisely a decision tree is generated; in other words, the higher the simplicity, the better the generated decision tree. From the generalization viewpoint, the generated rules with high support values are presented and their implications are addressed.

In general, given a defined number of decisional attributes, a decision tree with deeper leaves (thus more complex) produces rules with more conditions. Similarly, given a fixed number of cases, a rule with more support returns a simpler decision tree. The simplicity of a decision tree is determined by the sum of simplicity of individual rules, considering depth and support. The greater the simplicity, the better the decision tree. Therefore, given the certain size of dataset size, the averaged simplicity (AS) of a generated decision tree (DeTr) is determined by Eq. (5).

$$AS(DeTr) = \left(\sum_{i=1}^{r} \frac{1}{Depth_i} \times Support_{Ri} \right) \Big/ N_R \tag{5}$$

where $Depth_i$: The number of attributes of the ith decision rule, $i = 1, 2, 3, \ldots, r$, where r is the total number of rules. $Support_{Ri}$: The support of the ith rule. N_R: The number of rules generated.

4. Results

4.1. Dataset processing

B-dataset held 30,116 cases over a period of 3 years and 23 days and *A-dataset* held 17,868 cases over a period of 1 year, 5 months, and 7 days; thus there were an average of 819.04 cases per month in *B-dataset* and 1037.03 cases per month in *A-dataset*. This implies an 26.62% monthly increase after the critical date of the Covid-19 pandemic in Taiwan. The PA of *B-dataset* and *A-dataset* from C4.5 was estimated at 56.11% and 55.01%, respectively. As these results were unacceptable, the decisional attribute dimensions were reduced based on the CP of decisional attributes using C4.5. The rank of CP is presented in Table 2 and Figure 1. Three main findings from the attribute selection process are as follows.

(1) Considering the top 15 ranked attributes, the human and vehicle categories were more important than the environment and temporal categories as decisional attributes to classify traffic injuries. The vehicle category shows the highest CP (2 of 2 for both datasets), followed by human (9 of 12 for *B-dataset* and 8 of 12 for *A-dataset*), environment (4 of 15 for *B-dataset* and 5 of 15 for *A-dataset*), and finally temporal (0 of 5 for both datasets).
(2) Some attributes were ranked inside the pool for one dataset but outside of the pool for the other. For example, the road (X08) and occupation (X29) attributes were ranked only in *B-dataset*, and road type (X09) and speed limit (X12) were ranked only in *A-dataset*. This implies that speed limit was not a key predictor of traffic injuries before the critical date of Taiwan's reaction to the Covid-19 pandemic but was an important predictor after this date.
(3) Simply considering the environment as the influential factor in traffic accidents is insufficient when developing a road accident prediction model. This finding is consistent with the report in Ratanavaraha and Suangka (2014) that speed was an important determinant of accidents and the findings of de Oña et al. (2014) that drivers' characteristics were the main cause of traffic accidents when multiple determinant categories were studied (vehicle, environment, and human).

The top 15 attributes were selected as the final decisional attributes for both datasets. *B-dataset-top15* and *A-dataset-top15* were thus determined and cleaned by removing inconsistent data based on the CAL of 60%. *B-dataset-top15-clean* and *A-dataset-top15-clean* were obtained as shown in Table 3, which shows that the volume of inconsistent data removed was 8,626 records from *B-dataset-top15* and 5,803 records from *A-dataset-top15*. Table 3 also presents the PA for the datasets before and after cleaning. The PA of *B-dataset-top15* and *A-dataset-top15* were 70.73% and 74.77%, respectively, after cleaning, indicating that the

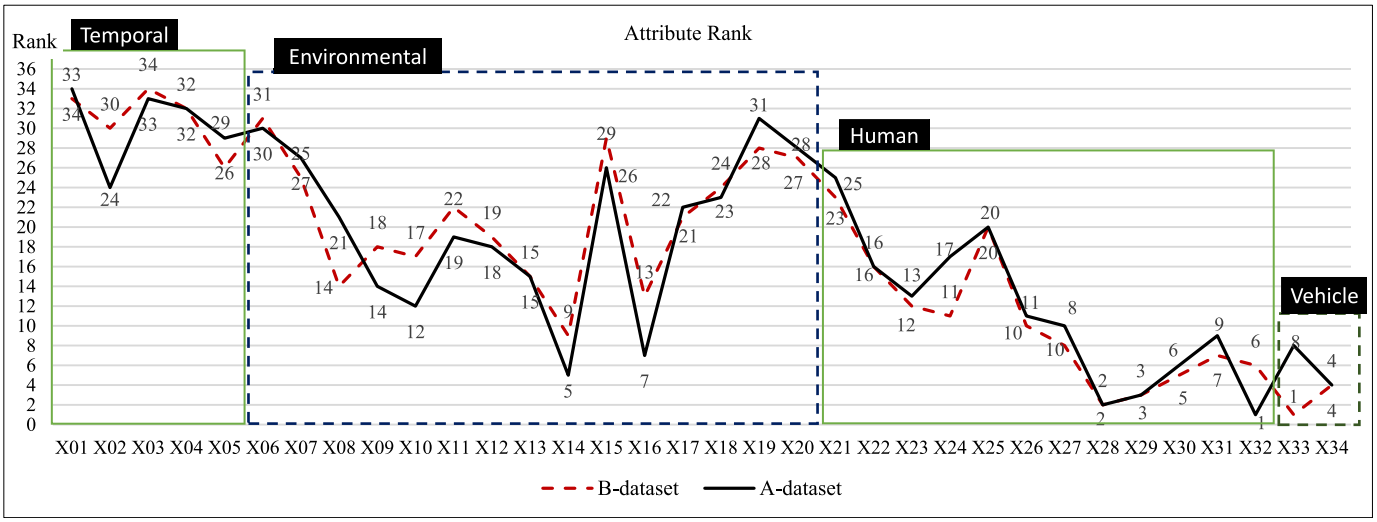

Figure 1. Ranks of attributes for B-dataset (before pandemic response) and A-dataset (after pandemic response).

Table 3. Removal of inconsistent data and prediction accuracy.

Datasets	Dataset size after inconsistency removal		Prediction accuracy
Before cleaning	B-dataset-top15	30,116	55.52%
	A-dataset-top15	17,868	55.30%
After cleaning	B-dataset-top15-clean	21,490	70.73%
	A-dataset-top15-clean	12,065	74.77%

removal of inconsistent data increased the PA by 15.21% on *B-dataset* and 19.47% on *A-dataset*. Despite the PA still not being high for either dataset, the accuracy was acceptable and therefore the datasets were used for knowledge elicitation.

Collision points and vehicle types were highly ranked. The collision points varied among the 16 options but front end, right side, back end, and left side were most prominent; in regard to vehicle type, motorcycles were the main type. Table 4 presents the results of comparing the relative frequency of vehicle types and injury types in the period before the Covid-19 pandemic and the period after its impact was felt. The comparison reveals some remarkable findings. Traffic accidents involving motorcycles increased from 16.31% to 21.44% whereas those involving buses and cars decreased by 1.02% and 4.11%, respectively. Head injuries decreased from 8.04% to 6.89% but leg injuries increased from 72.10% to 75.57%, with the proportion of other types also changing but by no more than 0.93%.

Table 4. Relative frequency of vehicle types and injury types for B-dataset (before pandemic response) and A-dataset (after pandemic response).

Vehicle type (VT)	B	Relative frequency	A	Relative frequency	Difference
VT1 (bus)	1699	7.91%	831	6.89%	−1.02%
VT2 (car)	16286	75.78%	8647	71.67%	−4.11%
VT3 (motorcycle)	3505	16.31%	2587	21.44%	5.13%
Total	21490		12065		
Injury type (IT)					
IT1 (head)	1727	8.04%	691	5.73%	−2.31%
IT2 (neck)	209	0.97%	121	1.00%	0.03%
IT3 (breast)	380	1.77%	197	1.63%	−0.14%
IT4 (abdomen)	95	0.44%	55	0.46%	0.01%
IT5 (waist)	302	1.41%	155	1.28%	−0.12%
IT6 (back)	162	0.75%	90	0.75%	−0.01%
IT7 (hand/wrist)	3120	14.52%	1639	13.58%	−0.93%
IT8 (leg/foot)	15495	72.10%	9117	75.57%	3.46%
Total	21490		12065		

Table 5. Main features of B-Rulebase and A-Rulebase

Feature	B-Rulebase	A-Rulebase	Note
Dataset size	21,490	12,065	
Rules generated	9,622	4697	
Mean support	2.2334	2.5687	A > B
Mean simplicity	0.3799	0.4613	A > B
No. of highly supported rules	31	21	≧50 supports
Top 5 support values	740, 453, 219, 218, 178	599, 377, 173, 153, 132	

4.2. Rule base generation

By implementing the RGA, rule bases containing decision rules were generated from *B-dataset-top15-clean* and *A-dataset-top15-clean* and labelled *B-Rulebase* and *A-Rulebase*, respectively. Each rule base includes the rule characteristics of decisional attributes, predictive attribute, depth, support, reliability, class distribution, and simplicity. The features of the generated rules are presented concisely in Table 5. The ranks of decisional attributes in both rule bases changed as the leaves of the decision tree were generated. There are 9,622 rules in *B-Rulebase* and 4,697 in *A-Rulebase*. The reliability of each rule is 1.00 because of the removal of inconsistent data. *B-Rulebase* contains 31 rules with a support value greater than 50 and *A-Rulebase* has 21, but the depth of these rules ranges from 4 to 12. The top 5 support (depth) values in *B-Rulebase* and *A-Rulebase* are 740(12), 453(10), 219(9), 218(8), 178(7), and 599(9), 377(8), 173(9), 153(5), 132(9), respectively. The average support and simplicity of *B-Rulebase* are 2.2334 and 0.3799, and those of *A-Rulebase* are 2.5687 and 0.4613.

Figure 2 illustrates the support and simplicity against depth for *B-Rulebase* and *A-Rulebase*. Despite some oscillation, the support and simplicity values are almost normally distributed. The values of support and simplicity in *B-Rulebase* increase at depth 3, reach a peak at depth 6, and decrease to depth 13. Traffic injury is normally predicted using four to nine decisional attributes for both periods. However, the rules in *A-Rulebase* end at depth 11, indicating that most rules in *B-Rulebase* need more decisional attributes to predict traffic injury than those in *A-Rulebase*, as is also indicated by the average simplicity.

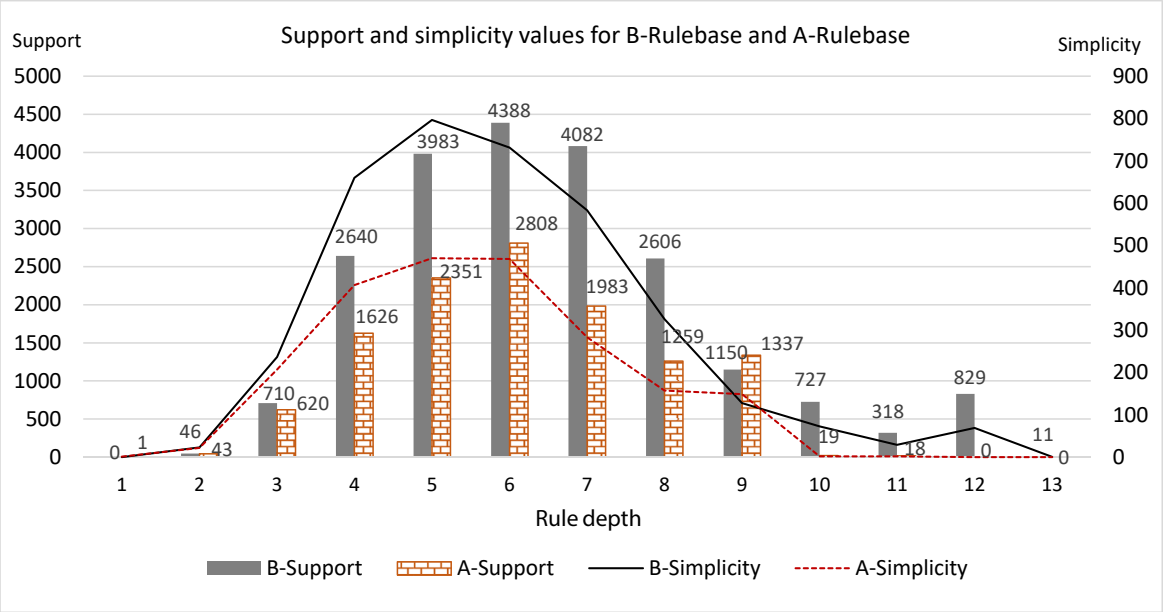

Figure 2. Support and simplicity for B-Rulebase and A-Rulebase.

The generated rules with a support value greater than 50 are revealed to explore the key determinants associated with traffic injury in *B-Rulebase* (part in Table 6) and *A-Rulebase* (part in Table 7). In particular, in Table 6, there are 15 decisional attributes (X33–X13) and one predictive attribute (class). Each record represents a decision rule with an identifier. For example, rule #2772 is generated: *When {Collision point = front end} and {Occupation = general worker} and {Age = between 21 and 30} and {Road pavement defect = negative} and {Road category = city road} and {Accident cite = intersection} then {Main injury = leg}*. It also reveals that a rule has six decisional attributes (column Dep.), has 91 kinds of support (column Sup.), is a perfect classification (column Class distribution), and has simplicity of 15.1667 (column Simplicity). Part of A-Rulebase is presented in Table 7, the structure of which is the same as that of Table 6. Moreover, extracts from each rule base in the form If conditions-Then conclusion (injury) are precisely presented in Table 8. This reveals that some decisional attributes are included in both B-Rulebase and A-Rulebase, others in either one rule base. For example, Collision spot, Driver's license types, Alcohol test, and License status are included in both rule bases. Occupation, Road category, Device use, Road pavement, and Road pavement defect only appear in the B-Rulebase, whereas Speed limit and Road pavement condition only appear in the A-Rulebase.

4. Implications and discussion

There are five major points of discussion related to the method and findings of this study. First, the original datasets had large numbers of decision attributes and the PA was found unacceptable for both *B-dataset* (before the critical pandemic date) and *A-dataset* (after the critical pandemic date) (56.11% and 55.01%, respectively). The attribute dimensions were reduced according to rank and the top 15 attributes in each dataset were selected, but the PA remained low (55.52% and 55.30%, respectively). Inconsistent data was then removed to avoid high failure classifications or unreliable generated rules, which increased the PA to 70.73% and 74.77%, respectively. This data cleaning sacrificed 8,626 cases in B-dataset and 5,803 in A-dataset that revealed high inconsistency rates (28.64% and 32.48%, respectively), but by producing an acceptable PA, the study was able to move on to knowledge elicitation with reliable outcomes.

Second, real-world data on traffic accidents is quite unstructured and contains decision attributes that are the same but produce different conclusions. This causes a certain degree of difficulty in modeling traffic accidents using classification-oriented prediction models. There are inherent cognitive, subjective, or transitionary errors in policy accident reports (Li et al., 2021), and traffic accident datasets collected from public open data repositories will therefore inevitably contain mistakes. It is suggested that knowledge elicitation from open data requires thorough data cleaning and pre-processing to ensure the outcome quality of each stage of analysis right down to the penultimate stage; in particular, inconsistent granulated data needs to be well managed (Wu et al., 2013). The involvement of domain specialists and police officers is required to enhance data quality. In our case study, data cleaning was performed to ensure there was no missing data and no inconsistent cases.

Third, the expected entropy of attributes of both datasets and their ranks (Table 3) reveal that the main factors to predict accidents are *cars* and *collision point* (e.g., "front end," "front left end") in the vehicle category and *motorcycle riders* in the human category, which mainly lead to leg (or foot) injuries. This supports findings in the literature that drivers' characteristics and behaviors are the main determinants of traffic accidents (de Oña et al., 2014; Xin et al., 2020) and that collision types influence the severity level of traffic accidents (Kwon et al., 2015). It also echoes the suggestion of Chen and Liu (2012) that a car bonnet leading edge is probably helpful in decreasing femur/pelvis injury risk. According to our findings in Taiwan, cars and motorcycles, moving forward or turning left, hitting at the front end or right and left side are the major factors associated with vehicle accidents, with the main outcome of leg (or foot) injuries. The finding that turning left was a major cause of traffic accidents stands out as one that is not covered in the literature.

Fourth, when considering the environmental category in isolation, *road category* (e.g., "city road") was ranked 14 in determinants of accidents in *B-dataset* but was ranked outside the top 15 (specifically, 21) in *A-dataset*. Meanwhile, *road type* (e.g., "intersection and overpass") did not appear in *B-dataset* but was ranked 14 in *A-dataset* (ranked 14). This implies that before the Covid-19 period, traffic accidents occurred mostly on non-specific types or sections of city roads. However, after the city began to adapt to the pandemic, the road type had stronger CP, especially at or near intersections. A possible explanation is the increased volume of motorcycles as travelers attempted to maintain social distancing. This interpretation is in line with the finding of Pawar et al. (2020) that commuters shifted from public to private modes of transport to avoid exposure to the coronavirus.

Table 6. Extract from B-Rulebase (support >50).

RID	X33	X28	X29	X34	X30	X32	X31	X27	X14	X26	X24	X23	X16	X08	X13	Class	Dep.	Sup.	Rel.	Class distribution	Simplicity
1752			DLS10	VHS11	DAS2	PEC1	FDC1	SLS1	RS1	SBC9	OC21		RSD4	RC5	AS9	MW8	12	740	1	0,0,0,0,0,0,0,740	61.6667
1585			DLS10	VHS11	DAS2			SLS1		SBC9	OC21	\'B3of10\"	RSD4	RC5	AS1	MW8	10	453	1	0,0,0,0,0,0,0,453	45.3000
1588			DLS10	VHS11	DAS2		FDC1			SBC9	OC21	\'B4of10\"		RC5	AS1	MW8	9	219	1	0,0,0,0,0,0,0,219	24.3333
1598			DLS10	VHS11	DAS2					SBC9	OC21	\'B2of10\"		RC5	AS1	MW8	8	218	1	0,0,0,0,0,0,0,218	27.2500
5737			DLS10	VHS14	DAS2					SBC9	OC21	\'B3of10\"			AS1	MW8	7	178	1	0,0,0,0,0,0,0,178	25.4286
1683			DLS10	VHS11	DAS2				RS1	SBC9	OC21			RC6	AS2	MW8	8	150	1	0,0,0,0,0,0,0,150	18.7500
3529				VHS12	DAS2	PEC1		SLS1		SBC9	OC21	\'B3of10\"			AS1	MW8	8	127	1	0,0,0,0,0,0,0,127	15.8750
1591			DLS10	VHS11	DAS2	PEC1		SLS1		SBC9	OC21	\'B5of10\"		RC5	AS1	MW8	10	114	1	0,0,0,0,0,0,0,114	11.4000
2683				VHS11							OC22			RC7	AS1	MW8	4	114	1	0,0,0,0,0,0,0,114	28.5000
1672			DLS10	VHS11	DAS2			SLS1	RS1	SBC9	OC21	\'B3of10\"	RSD4	RC5	AS2	MW8	11	112	1	0,0,0,0,0,0,0,112	10.1818
3902			DLS10	VHS12					RS1	SBC9	OC21	\'B3of10\"	RSD4		AS9	MW8	8	99	1	0,0,0,0,0,0,0,99	12.3750
5738			DLS10	VHS14	DAS2					SBC9	OC21	\'B4of10\"			AS1	MW8	7	93	1	0,0,0,0,0,0,0,93	13.2857
2430			DLS10	VHS11		PEC1				SBC9	OC22	\'B3of10\"		RC5	AS1	MW8	8	92	1	0,0,0,0,0,0,0,92	11.5000
2772				VHS11							OC4	\'B3of10\"	RSD4	RC5	AS1	MW8	6	91	1	0,0,0,0,0,0,0,91	15.1667
3538				VHS12	DAS2					SBC9	OC21	\'B4of10\"		RC5	AS1	MW8	7	81	1	0,0,0,0,0,0,0,81	11.5714
5799			DLS10	VHS14	DAS2	PEC1		SLS1		SBC9	OC21	\'B3of10\"	RSD4	RC5	AS9	MW8	11	79	1	0,0,0,0,0,0,0,79	7.1818
4774			DLS10	VHS13						SBC9	OC21	\'B3of10\"		RC5		MW8	6	74	1	0,0,0,0,0,0,0,74	12.3333
1720			DLS10	VHS11	DAS2	PEC1		SLS1		SBC9	OC21	\'B3of10\"			AS8	MW8	9	73	1	0,0,0,0,0,0,0,73	8.1111

RID: Rule identifier, Dep.: Depth, Sup.: Support, Rel: Reliability.

Table 7. Extract from A-Rulebase (support >50).

RID	X33	X28	X29	X34	X14	X30	X16	X32	X31	X27	X26	X10	X23	X9	X13	Class	Dep	Sup.	Rel.	Class distribution	Simplicity
915			DLS10	VHS11		DAS2	RSD4		FDC1	SLS1	SBC9	SL50			AS9	MW8	9	599	1	0,0,0,0,0,0,0,599	66.5556
834			DLS10	VHS11		DAS2	RSD4				SBC9	SL50	\'B3of10\"		AS1	MW8	8	377	1	0,0,0,0,0,0,0,377	47.1250
835			DLS10		RS1	DAS2				SLS1	SBC9	SL50	\'B4of10\"		AS1	MW8	9	173	1	0,0,0,0,0,0,0,173	19.2222
2540			DLS10	VHS14							SBC9		\'B3of10\"		AS1	MW8	5	153	1	0,0,0,0,0,0,0,153	30.6000
864			DLS10	VHS11		DAS2		PEC1	FDC1		SBC9	SL50	\'B3of10\"		AS2	MW8	9	132	1	0,0,0,0,0,0,0,132	14.6667
527			DLS10	VHS11	RS1	DAS2		PEC1	FDC1		SBC9	SL40			AS1	MW8	9	127	1	0,0,0,0,0,0,0,127	14.1111
842			DLS10	VHS11		DAS2					SBC9	SL50	\'B2of10\"		AS1	MW8	7	122	1	0,0,0,0,0,0,0,122	17.4286
838			DLS10	VHS11		DAS2				SLS1	SBC9	SL50	\'B5of10\"		AS1	MW8	8	115	1	0,0,0,0,0,0,0,115	14.3750
1421			DLS10	VHS12					FDC1		SBC5	SL50			AS1	MW8	6	96	1	0,0,0,0,0,0,0,96	16.0000
750			DLS10	VHS11					FDC1		SBC5	SL50			AS1	MW8	6	91	1	0,0,0,0,0,0,0,91	15.1667
1609				VHS12		DAS2					SBC9	SL50	\'B3of10\"		AS1	MW8	6	83	1	0,0,0,0,0,0,0,83	13.8333
2509			DLS10	VHS14					FDC1		SBC9	SL50	\'B4of10\"		AS1	MW8	7	69	1	0,0,0,0,0,0,0,69	9.8571
869			DLS10	VHS11	RS1	DAS2					SBC9	SL50	\'B4of10\"		AS2	MW8	8	69	1	0,0,0,0,0,0,0,69	8.6250
2648			DLS10	VHS14		DAS2					SBC9	SL50	\'B3of10\"	RT3	AS9	MW8	8	69	1	0,0,0,0,0,0,0,69	8.6250
2366				VHS14		DAS2				SLS1	SBC5	SL50	\'B3of10\"			MW8	6	68	1	0,0,0,0,0,0,0,68	11.3333
402			DLS10	VHS11		DAS2			FDC1		SBC9	SL30			AS9	MW8	7	68	1	0,0,0,0,0,0,0,68	9.7143
1753			DLS10	VHS12		DAS2	RSD4				SBC9	SL50	\'B3of10\"	RT3	AS9	MW8	9	68	1	0,0,0,0,0,0,0,68	7.5556
561			DLS10	VHS11		DAS2			FDC1		SBC9	SL40	\'B3of10\"		AS9	MW8	8	63	1	0,0,0,0,0,0,0,63	7.8750

RID: Rule identifier, Dep.: Depth, Sup.: Support, Rel: Reliability.

Table 8. Extracts from B-Rulebase and A-Rulebase with a support value more than 50.

	B-RB	Values in B	Value in A	A-RB
Predictive attribute	V	{leg or foot}		V
Decisional attributes				
Collision point	V	{front end, right side}		V
Driver license	V	{motorcycle license}		V
Occupation	V	{unknown}		
Alcohol test	V	{pass}		V
Age	V	{from 11 to 30}	{from 21 to 30}	V
Behavior condition	V	{moving forward}		V
Safety device use	V	{helmet or fasten seatbelt}		V
Accident cite	V	{regular lane, on intersection}	{regular lane, near intersection, on intersection}	V
License status	V	{legal}		V
Road category	V	{city road}		
Device use (mobile phone)	V	{no}		
Road pavement	V	{asphalt}		
Road pavement defect	V	{None}		
Speed limit			{50}	V
Road pavement condition			{wet}	V

B-RB: B-Rulebase, A-RB: A-Rulebase

As shown in Table 4, accidents due to collisions between cars and motorcycles were 5.13% higher in *A-dataset* than *B-dataset*, and injuries to the legs increased by 3.46%. To maintain social distancing and cope with parking restrictions, travelers are likely to have shifted from using public buses (public transportation systems) or cars to individual motorcycles, thus increasing the likelihood of traffic accidents in the already heavy traffic conditions in the city. Further support for this interpretation is the finding that speed limits were not a major determinant of accidents before Covid-19 but became prominent in *A-dataset*. Speed has been identified in the literature as one of the main determinants of the severity of crashes (Altwaijri et al., 2012; Ratanavaraha and Suangka, 2014). It is therefore suggested to the government and companies that remote work should be encouraged, but where this is not possible segmenting working hours to reduce traffic load is a worthwhile alternative.

Fifth, *collision point* and *vehicle type* ("motorcycle") held the highest classification ranks in both datasets after cleaning, and the most frequent outcome was injury to the leg or foot. This echoes the evidence that motorcycle use increased after the pandemic outbreak. Moreover, despite no remarkable differences between *B-Rulebase* and *A-Rulebase*, the latter showed higher average levels of support and simplicity (2.5687 vs. 12.2334 and 0.4613 vs. 0.3799, respectively). This implies that on average a rule generated from A-dataset holds 5.240 decisional attributes with a support value of 2.5768 and a rule generated from B-dataset holds 5.636 attributes and a support value of 2.2334; the average rule from A-dataset is therefore superior on both measures, with a lower number of decisional attributes and a higher support value.

In general, these numbers seem too low to confidently verify the generated decision tree, although there is no evidence in the literature to draw on. The rules support values of 50 or above of various depths generated from *B-Rulebase* and *A-Rulebase* have been used to present the key determining attributes and their associated injuries as the main outcomes of the proposed prediction model. Nonetheless, these numbers do reveal that the structure of the decision tree represented by *A-Rulebase* is more concise than that of *B-Rulebase*. As shown in Figure 2, 87.71% (17,699 of 21,490) of cases in *B-Rulebase* have a depth of four to nine, whereas 95.19% of cases in *A-Rulebase* meet this criterion. This indicates that given the same decisional attributes, prediction of traffic accidents is more likely for the period after the pandemic hit than the period beforehand. As mentioned above, this is probably due to the increase of accidents between cars and motorcycles, which generate a high occurrence of leg injuries and therefore make for more consistent rules in the prediction model.

In summary, the Covid-19 pandemic and associated containment measures has changed the factors and outcomes of traffic accidents. The findings of the present study indicate that to maintain social distancing and avoid infection, travelers in Taiwan became more likely to ride motorcycles, with a 5.13% increase in motorcycle involvement in accidents. This echoes the findings of Pawar et al. (2020) that 5.3% of commuters shifted from public to private transport modes. This shift increases traffic load, and shows that in addition to the impacts of human and vehicle factors, environmental variables are indisputably important and significantly influence the occurrence and nature of traffic accidents. One promising future research direction is to examine how technologies can be used to support remote work and how working hours for non-remote workers can be rearranged to spread traffic load. To support related policies, traffic policymakers should implement innovative technologies (e.g., surveillance systems and data analysis techniques) to collect data and frequently update management on traffic accident trends and factors.

5. Limitation and future works

Some limitations of the study are worth mentioning with a view to supporting further research. Similar with most previous studies, only a single data source (government open data) was used in the prediction model for traffic accidents developed in the research. Further insight could be provided by analyzing additional data sources, such as technological surveillance systems. Data selection is the key to success in this endeavor: issues such as privacy, cost, and consent from vehicle users should be considered, especially in the post-pandemic era. From the perspective of policymaking support, a prediction model must focus on the sources that are available, which determine the contexts relevant to the prediction or description of traffic accidents at an acceptable validation level, given that traffic accidents are becoming increasingly difficult to predict.

This study demonstrates that the accuracy of the prediction model of both original datasets with 34 decisional attributes was not sufficiently high. This shows the need for application-oriented research to include in the preprocessing of real-world open data measures to deal not only with missing data and granulation but also with the selection of attributes, pre-examination of the prediction model, and inconsistency management to ensure the output quality of later stages of analysis. Moreover, the original dataset was split into two by considering the date that the government announced it would begin to combat the incoming threat of Covid-19. Future studies may examine major government policy announcements and changes in Covid-19 pandemic trends to generate additional findings conducive to traffic policymaking.

The top 15 attributes by classification ranking were considered in this study, and inconsistent granulated data was removed at the 60% level. Although these steps were necessary for the feasibility of the research, they limit the findings and implications due to some fundamental theoretical and practical considerations. The results may vary by the selection of an acceptable inconsistency level. Given its links to PA, future studies should carefully manage the selection policy with respect to inconsistent granulated data removal. The unsupervised granulation technique (i.e., equal width interval) was used here to convert attributes from continuous to discrete measures. Other unsupervised techniques, such as equal frequency intervals (Liu and Setiono, 1997; Wu et al., 2013), and supervised techniques, such as minimum distance length (Grünwald, 2007), could be used to examine whether alternative granulation methods enhance the accuracy of the prediction model.

Furthermore, this study adopted a classification-oriented model using C4.5 to develop the prediction model and generate decision rules for traffic accidents. This prediction model can be extended to developing prediction models for other traffic issues (e.g., transportation systems) in other regions. Other prediction models, such as random forest (Breiman, 2001) and support vector machine (Keerthi et al., 2001), are also available (Chand et al., 2021) and may have merit for knowledge elicitation. Comparisons between C4.5 used in the present study and other techniques to deepen the understanding of knowledge extraction technique applications may be one of the future works in the decision making of road safety context.

6. Conclusion

This study used the classification-oriented technique on open government data of Taoyuan city in Taiwan to model traffic accidents. In this paper, the literature on the use of open data is reviewed and the procedure is reported for pre-processing the collected dataset, splitting the dataset into two subsets for comparison, applying the classification-oriented technique to predict traffic accidents, and generating decision rules to achieve the research objectives. Four categories of variables were considered as potential determinants of traffic accidents: temporal, environmental, human, and vehicle. With various combinations of categories having been used in the literature, the findings of the present study disclose that human and vehicle factors are more important than the other two. Although it is worth placing emphasis on how to develop a safe traffic environment, vehicle drivers or users are key to the reduction of traffic risks.

Data pre-processing, including cleaning and granulation, was a time-consuming task. The use of the classification technique C4.5 to develop the prediction model on the original datasets produced an unacceptable level of PA. However, accuracy was improved with the removal of inconsistent granulated data. This reveals that there are uncertainties involved in knowledge elicitation from a real-world dataset, such that confirmation of the processing quality at each stage is necessary to ensure that the next stage is meaningful. The significance and uniqueness of this research are evident. The proposed prediction model and research findings presented above reveal that there is a transition gap, not well covered by the existing literature, in the proposed prediction model for traffic accidents in Taiwan. Although this research mainly contributes to the domain of traffic safety, particularly in respect to the human and vehicle factors in traffic safety policymaking, the method can be applied and refined to advance technological innovation and to support existing approaches beyond this domain and in different contexts.

Declarations

Author contribution statement:

Wu, C.H.: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper. Kao, S.C.: Conceived and designed the experiments; Analyzed and interpreted the data. Chang, C.C.: Performed the experiments; Analyzed and interpreted the data.

Funding

Dr. ChienHsing Wu was supported by Ministry of Science and Technology, Taiwan [110-2410-H-390-001-MY2].

Data availability

Data associated with this study has been deposited at Taiwan government open data https://data.gov.tw/dataset/107386; https://data.gov.tw/dataset/129320; https://data.gov.tw/dataset/143048

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Alcaniz, M., Guillen, M., Santolino, M., 2021. Differences in the risk profiles of drunk and drug drivers: evidence from a mandatory roadside survey. Accid. Anal. Prev. 151.
Ali, F., Ail, A., Imran, M., Naqvi, R.A., Siddiqi, M.H., Kwak, K.S., 2021. Traffic accident detection and condition analysis based on social networking data. Accid. Anal. Prev. 151.
Altwaijri, S., Quddus, M., Bristow, A., 2012. Analyzing the severity and frequency of traffic crashes in Riyadh City using statistical models. Int. J. Transport. Sci. Technol. 1 (4), 351–364.
Antona-Makoshi, J., Mikami, K., Lindkvist, M., Davidsson, J., Schick, S., 2018. Accident analysis to support the development of strategies for the prevention of brain injuries in car crashes. Accid. Anal. Prev. 117, 98–105.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Chand, A., Jayesh, S., Bhasi, A.B., 2021. Road traffic accidents: an overview of data sources, analysis techniques and contributing factors. Mater. Today Proc. 47 (15), 5135–5141.
Charlton, S.G., Starkey, N.J., Malhotra, N., 2018. Using road markings as a continuous cue for speed choice. Accid. Anal. Prev. 117, 288–297.
Chen, C.S., Liu, T.C., 2012. Medical cost and motorcycle helmet law in Taiwan. Econ. Res. Int. 2012, 1–8.
Chen, Q., Pan, S., 2020. Transport-related experiences in China in response to the Coronavirus (COVID-19). Transp. Res. Interdiscip. Perspect. 8.
Connelly, L.B., Supangan, R., 2006. The economic costs of road traffic crashes: Australia, states and territories. Accid. Anal. Prev. 38, 1087–1093.
da Cruz Figueira, A.C., Pitombo, C.S., de Oliveira, P.T.M.S., Larocca, A.P.C., 2017. Identification of rules induced through decision tree algorithm for detection of traffic accidents with victims: a study case from Brazil. Case Studies on Transp. Policy 5, 200–207.
de Oña, J., de Oña, R., Eboli, L., Forciniti, C., Machado, J.L., Mazzulla, G., 2014. Analyzing the relationship among accident severity, drivers' behavior and their socio-economic characteristics in different territorial contexts. Proc. Soc. Behav. Sci. 160, 74–83.
Elvik, R., 2000. How much do road accidents cost the national economy? Accid. Anal. Prev. 32 (2000), 849–851.
Fernandes, A., Neves, J., 2013. An approach to accidents modeling based on compounds road environments. Accid. Anal. Prev. 53, 39–45.
French, M.T., Gumus, G., Homer, J.F., 2009. Public policies and motorcycle safety. J. Health Econ. 28, 831–838.
George, Y., Athanasios, T., George, P., 2017. Investigation of road accident severity per vehicle type. Transport. Res. Procedia 25, 2076–2083.
Grünwald, P., 2007. The Minimum Description Length Principle. MIT Press, Cambridge, MA.
Hotle, S., Murray-Tuite, P., Singh, K., 2020. Influenza risk perception and travel-related health protection behavior in the US: insights for the aftermath of the COVID-19 outbreak. Transp. Res. Interdiscip. Perspect. 5.
Kaygisiz, O., Senbil, M., Yildiz, A., 2017. Influence of urban built environment on traffic accidents: the case of Eskisehir (Turkey). Case Studies on Transp. Policy 5, 306–313.
Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput. 13 (3), 637–649.
Kwon, O.H., Rhee, W., Yoon, Y., 2015. Application of classification algorithms for analysis of road safety risk factor dependencies. Accid. Anal. Prev. 75, 1–15.
Lausch, A., Schmidt, A., Tischendorf, L., 2015. Data mining and linked open data - new perspectives for data analysis in environmental research. Ecol. Model. 295, 5–17.
Law, T.H., Noland, R.B., Evans, A.W., 2009. Factors associated with the relationship between motorcycle deaths and economic growth. Accid. Anal. Prev. 41, 234–240.
Li, J., Zhao, Z., 2022. Impact of COVID-19 travel-restriction policies on road traffic accident patterns with emphasis on cyclists: a case study of New York City. Accid. Anal. Prev. 167.

Li, X., Liu, J., Zhang, Z., Parrish, A., Jones, S., 2021. A spatiotemporal analysis of motorcyclist injury severity: findings from 20 years of crash data from Pennsylvania. Accid. Anal. Prev. 151.
Liu, H., Setiono, R., 1997. Feature selection via discretization. IEEE Trans. Knowl. Data Eng. 642–646.
Olowosegun, A., Babajide, N., Akintola, A., Fountas, G., Fonzone, A., 2022. Analysis of Pedestrian Accident Injury-Severities at Road Junctions and Crossings Using an Advanced Random Parameter Modelling Framework: the Case of Scotland, 169. Accident Analysis & Prevention.
Parady, G., Taniguchi, A., Takami, K., 2020. Travel behavior changes during the COVID-19 pandemic in Japan: analyzing the effects of risk perception and social influence on going-out self-restriction. Transp. Res. Interdiscip. Perspect. 7.
Park, J., Choi, Y., Chae, Y., 2021. Heatwave impacts on traffic accidents by time-of-day and age of casualties in five urban areas in South Korea. Urban Clim. 39.
Pawar, D.S., Yadav, A.K., Akolekar, N., Velaga, N.R., 2020. Impact of physical distancing due to novel coronavirus (SARS-CoV-2) on daily travel for work during transition to lockdown. Transp. Res. Interdiscip. Perspect. 7.
Quinlan, J.R., 1986. Induction of decision tree. Mach. Learn. 1, 81–106.
Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo.
Ratanavaraha, V., Suangka, S., 2014. Impacts of accident severity factors and loss values of crashes on expressways in Thailand. IATSS Res. 37, 130–136.
Rao, T.R., Mitra, P., Bhatt, R., Goswami, A., 2019. The big data system, components, tools, and technologies: a survey. Knowl. Inf. Syst. 60, 1165–1245.
Roy, S., Delwer Hossain Hawlader, M., Hayatun Nabi, M., Ananyo Chakraborty, P., Zaman, S., Morshad Alam, M., 2021. Patterns of injuries and injury severity among hospitalized road traffic injury (RTI) patients in Bangladesh. Heliyon 7 (3).
Saladié, Ò., Bustamante, E., Gutiérrez, A., 2020. COVID-19 lockdown and reduction of traffic accidents in Tarragona province, Spain. Transp. Res. Interdiscip. Perspect. 8.
Sangkharat, K., Thornes, J.E., Wachiradilok, P., Pope, F.D., 2021. Determination of the impact of rainfall on road accidents in Thailand. Heliyon 7 (2).
Statistics of vehicle accidents. National Police Agent, Taiwan. Accessed on January, 2019, https://www.npa.gov.tw/NPAGip/wSite/ct?xItem=78478&ctNode=12878&mp=1.
Tavakoli, A., Heydarian, A., 2022. Multimodal Driver State Modeling through Unsupervised Learning, 170. Accident Analysis & Prevention.
Ture, M., Tokatli, F., Kurt, I., 2009. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert Syst. Appl. 36 (2), 2017–2026.
Valent,F.,2022.RoadtrafficaccidentsinItalyduringCOVID-19.TrafficInj.Prev.23(4),193–197.
van Wee, B., De Vos, J., Maat, K., 2019. Impacts of the built environment and travel behaviour on attitudes: theories underpinning the reverse causality hypothesis. J. Transport Geogr. 80.
Vipin, N., Rahul, T., 2021. Road traffic accident mortality analysis based on time of occurrence: evidence from Kerala, India. Clinical Epidemiology and Global Health, 11.
Vorel, G., Kao, S.C., Wu, C.H., Wu, C.C., 2014. Determinants of traffic fatalities in Taiwan. Int. J. Inf. Manag. Sci. SI-August (2014), 233–249.
Wu, C.H., Kao, S.C., 2021. Knowledge discovery in open data for epidemic diseases prediction. Health Policy and Technology 10 (1), 126–134.
Wu, C.H., Kao, S.C., Okuhara, K., 2013. Examination and comparison of conflicting data in granulated datasets: equal width interval vs. equal frequency interval. Inf. Sci. 239 (1), 154–164.
Xin, Z., Vu, L.H., Huang, H., 2020. Fifty years of Accident Analysis & Prevention: a bibliometric and scientometric overview. Accid. Anal. Prev. 144, 105568.