Identifying Seriousness of Car Accidents using Data Mining in Jordanian Roads

Faisal Aburub, Wael Hadi

Abstract*: Deemed as the second most common cause of death in Jordan and the second leading cause of death amongst youths and infants, traffic collisions form a serious health problem in Jordan, claiming an average of 1.25 million lives per year and gravely injuring 50 million more globally. This study purposes to explore the utilisation of data mining strategies so as to identify the seriousness of car accidents in Jordan, and then studying them further so as to pinpoint a number of features that impact car accidents.

Keywords* : Data Mining, Car Accidents, Associative Classification.

I. INTRODUCTION

Deemed as the second most common cause of death in Jordan and the second leading cause of death amongst youths and infants, traffic collisions form a serious health problem in Jordan, claiming an average of 1.25 million lives per year and gravely injuring 50 million more globally. 1-2% of countries' gross national products go towards these accidents, not to mention the vast amount of grief and suffering it creates. It is for these reasons that it is in the best interests of all countries to attempt to minimise the number of car accidents occurring; thus, what may aid the deterrence of road accidents are to identify the geographical factors of road accidents, and to study the differing features of accident locations.

Referring to 'an analytic process aimed at exploring large amounts of data in search of structures, commonalities, and hidden patterns/rules amongst data', data mining techniques are regarded as one of the most essential strategies when it comes to studying the varying features—and discovering the links between—road accidents. Data mining techniques (e.g., Bayesian network analysis and the decision tree technique) offer a range of benefits, such as: there are no assumptions made on variable distributions; it provides the possibility of taking information from vast numbers of data; bicycle-crash-severity-knowledge is not necessarily needed; and a number of variables with a vast amount of categories/discrete variables are handled more effectively than that of regression models (Prati *et al*., [1]).

The job concerning data mining is to take knowledge (high-level information) from low-level data (commonly stores in vast databases) (Zucco, [2]); data mining can thus be de5fined as 'the non-trivial process of identifying patterns in data that are valid, original, potentially useful, and understandable' [2]. Additionally, it has proven to be a reliable strategy in studying road accidents, supplying accurate results, which is commonly undergone by data mining processes—such as classification, feature selection, and clustering—so as to identify the characteristics that tend to frequently cause accidents [3]. Further, data mining can additionally be explained as 'a novel technique to extract hidden and previously unknown information from the large amount of data'. Association rule mining, clustering, and classification are some examples of frequently used road accident analysis strategies globally [4]. The key objective of this paper is to successfully study road locations so as to identify specific traffic accident characteristics that appear to facilitate traffic accidents to commonly occur within these locations; further, this study purposes to explore the utilisation of data mining strategies so as to identify the seriousness of car accidents, and then studying them further so as to pinpoint a number of features that impact car accidents. The layout of this paper is split into three key sections: Section 1 puts across a short introduction; Section 2 explains the methodology and data that has been utilised; and Section 3 presents the conclusions drawn from this study.

II. LITERATURE REVIEW

When it comes to resource and economic costs, road accidents have resulted in a largely negatively impact; because of this, a number of studies have been constructed so as to examine this issue. One such example of this is that of Martin et al. [5], who studied the data concerning dangerous points on Andalusia, Spain's Complementary Road Network, whereby a link was, indeed, founded, and the Andalusian roads improved. In a similar vein, in Ethiopia, Tesema et al. [6] utilised an adaptive regression tree model in order to construct a decision support system for the road accidents on the Ethiopian roads; similarly, in Belgium, Geurts et al. ([7] utilised association rule mining technique to comprehend the many circumstances that come about within high-frequency accident locations. Further, in Spain, Abellan et al. created many decision trees to take varying decision rules for different trees in order t examine two-lane rural highway data, finding that dull light conditions and safety barriers negatively impact crash severity. Utilising data mining techniques in order to spot common accident locations (and examining them deeply so as to identify the varying features that impact road accidents within these locations), Kumar & Toshniwal [4] took some valuable, previously concealed information concerning the data, which could, indeed, be applied to attempt to prevent road accidents.

Revised Manuscript Received on January 15, 2020
***** Correspondence Author
Faisal Aburub*, MIS department, University of Petra, Amman, Jordan.
Wael Hadi, CIS department, University of Petra, Amman, Jordan.

Data mining algorithms were initially presented for the purpose of classification of road accident vehicle collision (Punithavilli, [3]), founded from the classification rules that can be used for the forecasting of vehicle collision patterns. To begin with, the training set was extracted, before the removal of incomplete, noisy, or inconsistent data; then, the pre-processed data would be transferred into a suitable form for mining. Following, the feature set attribute space was minimised for vehicle collision classification, which can be obtained by utilising the feature selection algorithms (e.g., Fast Correlation Based Filter (FCBF), Correlation based Feature Selection (CFS), Mutual Information Feature Selector (MIFS), and Multi valued Oblivious Decision Tree (MODTree)). The above chosen features are utilised in varying classification algorithms (e.g., Random Tree, Naïve Bayes, Rule Induction, C4.5, Decision List, Classification and Regression Trees (C&RT), and RndTree).

Derived from one of the busiest roads within Nigeria, Olutayo & Eludire [9] examined Nigerian traffic accidents by organising the obtained data into continuous—examined using the Artifical Neural Network—and categorical—examined using the decision tree technique—data. Implemented on three- and four-legged UK road junctions, Nitsche et al. [10] additionally suggested a new strategy on taking essential pre-crash situations from accident data. Considering it is capable of handling categorical data and is strong against outliers, the clustering method k-medoids was discovered to be the best for the dataset at hand [11].

So as to verify the predictors of the seriousness of bicycle collisions within Italy, Prati et al. [12] applied Bayesian network analysis and the CHAID decision tree technique; the conclusions derived from the latter suggested that the predictors of the highest importance were—in decreasing order of importance—: road type; crash type; cyclist age; road signage; cyclist gender; opponent vehicle type; month; and road segment type. Within Bayesian network analysis, the above eight variables were incorporated as target predictors (i.e., bicycle crash severity); additionally, opponent vehicle type, crash type, and road type were founded to be the most telling predictors of bicycle collision severity, when utilising Bayesian network. Directing a research so as to gain a model for forecasting accident types (damages/causalities) at an appropriate accuracy level, Alikhani et al. [13] additionally studied the Adaptive-Neuro Fuzzy Inference System (ANFIS) and the Artificial Neural Network (ANN) accuracy in Iran; for classification of road accident severity within Iran, they explored a heuristic concept of combined clustering-classification system.

The Classification and Regression Trees (CART) strategy is amongst one of the most popular prediction and classification strategies, and the conclusions they draw may be purposed within decision tree framework (Chang & Wang [14]); as a matter of fact, a number of researchers within traffic safety studios have modified it by leading analysis on 12,604 instances of data related to accidents in Taiwain (via the CART strategy). Alikhani et al. [13] suggested that the ANNs (Artificial Neural Networks) are one of the non-parametric strategies and data mining tools whereby the researchers have studied the seriousness of injuries and accidents amongst the individuals involved in such collisions. Using regression analysis, CART technique, and neural networks, Sohn, S.Y. & Shin [15] classified the collisions on the grounds of the seriousness in Korea; their conclusions suggested that the utilisation of safety helmets and seat belts are largely influential factors in collision severity.

The t¬-tests are to test considerable variations amongst injury-severity levels for specific factors—such as occupant ejections, blood alcohol concentration, seat belt use percentage (Ostrom & Eriksson [16]). So as to examine traffic characteristic and road geometry implications on accident rates for multilane and rural two-lane roads, Karlaftis & Golias [17] utilised hierarchical traits; in the same vein, Park & Saccomanno [18] utilised recursive partitioning method (RPART). This was for the purpose of stratifying highway-railway crossways in Canada into homogeneous classes of crossings; this was undergone by applying control variables of track number, highway class, and track type.

There have been various studies that have examined the links between vehicle collisions and road features, as discussed above; saying this, considering most of the prior studies have fixated on specific road accident types, new road features, and/or specific road users, there is still a great deal to be understood concerning this issue; especially on Jordanian roads, the underlying causes for serious collisions and particularly vulnerable groups on the road are not fully grasped. Thus, this paper's objective is to utilise data mining techniques so as to examine the data derived from car collisions occurring in Jordan, as well as to recognise a number of features that impact Jordanian car accidents.

III. METHOD

A. Car Accidents in Jordan

Amongst one of the most essential problems facing countries globally in the present day is traffic collisions and their implications, ranging from human injury to material loss; this problem is not limited to any one specific country, and results in large-scale suffering for humankind, both economically and socially. Almost 1.25 million people are lost per year due to these accidents, not to mention the tens of millions lost altogether (World Road Safety Report, 2015); the moral and psychological effects on families as a result of these accidents should be considered too.

The Jordan total population (2017) was 10,000,053; the total amount of registered cars was 1,583,458; and the total amount of car crashes was 150,226—10,446 resulting in human injury (1,495 serious, 14,751 minor) and a devastating 685 proving fatal. The total costs for Jordan, as a result, were JOD 3,000,008.

B. Car Accidents in Jordan

So as to identify the seriousness of car collision dataset, varying data mining classification algorithms are likened according to F1 measures, classification accuracy, recall, and precision measures; the data in question involves 3,450 car accidents derived from JPSD during the last two years.

This is to weigh up five popular data mining algorithms, which have been selected because of the differing training strategies they use when it comes to discovering rules.

1. Having been utilised in a number of fields of science, NB filter belongs is a statistical text classification system; this theory relies on the prior event so as to prove the conditions and supply the best solution to resolve any issue (Thabtah et al., 2009);

2. amongst one of the most effective and commonly used decision tree classifiers, J48 classifier is the enhanced version of C4.5 algorithms put forward by Quinlan (1993);

3. one of the rule induction algorithms that only deals with one character attribute and does not undergo any pruning using coverage search, PRISM applies a top-down approach, and has created an illustration for the classification conclusion, not to mention it is utilised directly for decision-making (Romero et al., 2010);

4. utilised for regression analysis and classification, SVM is a supervised learning model that additionally has linked learning algorithms—which identify patterns and examine data (Olivo et al., 2011);

5. MCAR can be described as a supervised learning model that assimilates between classification and association rule (Thabtah et al., 2005).

The exploration of these three algorithms were led utilising known Weka software (Hall et al., 2009); in addition, all classification conclusions that were drawn were all weighed up by 10-fold cross-validation.

C. Dataset

Supplied by the Jordanian Public Security Directorate (JPSD), the dataset utilised within this study comprises material concerning Jordanian road collisions between the years 2016-2017; notably, the JPSD database does not involve a distinction between the varying injury levels, resulting in distinctions only being made between road collisions ending in fatalities or injuries. The database was reorganised and 14 categorical variables were chosen (as demonstrated in Table 1): 1) month; 2) day of the week; 3) time; 4) age of driver; 5) gender of driver; 6) type of road; 7) location of accident; 8) condition of road surface; 9) segment of road; 10) amount of lanes; 11) lighting on road; 12) surrounding area; 13) condition of weather; 14) seriousness of accident.

The results listed below have been uncovered as a result of critical analysis of the data obtained:

- Out of the total amount of vehicle collisions, 4.84% resulted in minor injuries, 6.5% were fatal, and 10% resulted in serious injuries;
- the most common time period for car collision injuries—6.5%—was 16:00-16:59; on the other hand, the rarest time period for car collision injuries—0.7%—was 17:00-17:59;
- the most common day of the week for car collision injuries—7.15%—was Thursday; on the other hand, the rarest day of the week for car collision injuries—3.13%—was Friday;
- the most common month for car collision injuries—8.6%—was August; on the other hand, the rarest month for car collision injuries—4.9%—was February;
- the most common weather condition for car collision injuries—5.97%—was clear weather;
- the highest amount of accidents resulting in human injuries was 1.49%;
- the highest amount of accidents resulting in death in the age group 60+ years was 12.8%.

D. Accuracy

Accuracy can be defined as 'the rate of correct predictions that the model is capable of obtaining when compared with the actual classifications in the dataset'; the accuracy of the MCAR algorithm when it comes to forecasting the seriousness of car collision compared with four other algorithms is represented in Figure 1, which demonstrates clear-cut variations between the varying filter techniques and identifies the most capable of them to be utilised to prevent serious road collisions. We identified that the MCAR algorithm demonstrated the highest capability (after comparisons between the learning machines PRISM, MCAR, SVM, J48, and NB)—95.2% better than the others, as shown in Figure 1. Utilising the AC approach and being easy to grasp, the MCAR algorithm outperformed the other algorithms by the following amounts: PRISM by 10.1%; SVM by 2.7%, J48 by 3.3%, and NB by 6.8%.

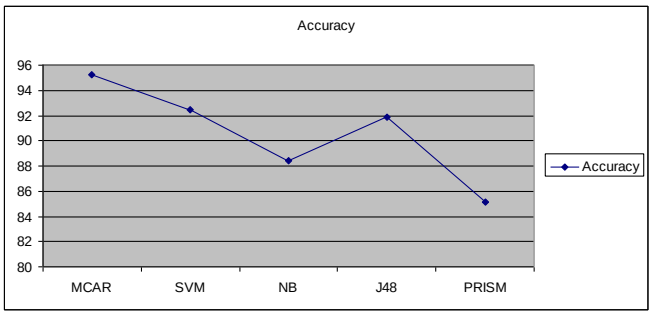

Fig. 1: The Accuracy Result

E. Precision and Recall Results

Utilised within the classification problem, recall and precision are two evaluation measures calculated on the grounds of a matrix dubbed 'confusion matrix' (Table 1); these evaluation measures are computed as listed below:

$$\textbf{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Whereby:

The number of correct hits in the positive instance are symbolised as True Positive (TP); the number of incorrect hits in the positive instance is symbolised as False Negative (FN); the number of incorrect hits in the negative instance is symbolised as False Positive (FP); and the number of correct hits in the negative instance is symbolised as True Negative (TN).

Table- I: Confusion Matrix

	Classified Positive	Classified Negative
Actual Positive	**TP**	**FN**
Actual Negative	**FP**	**TN**

Amongst the MCAR and other varying algorithms (dependant on the confusion matrix), Figure 2 demonstrates Average Precision and Recall.

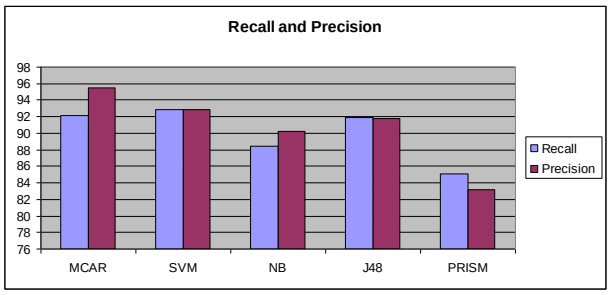

Fig. 2: Average Precision and Recall results

Meaning that the algorithm produced the highest amount of appropriate results (Recall result), the MCAR achieved the biggest average Precision and Recall results; conversely, high levels of Precision translates to the fact that the algorithm produced more appropriate results than irrelevant results (amongst other varying classification algorithms, which were applied within the experiments as the algorithm utilised many rules for the purpose of assigning the class for data-testing, as opposed to utilising a single rule prediction). It seems that machine learning algorithms and data mining undergo their purpose well (identifying seriousness of car collisions).

IV. CONCLUSION

Amongst some of the most essential issues being tackled by all countries in the present time are traffic accidents and their implications, from human injury to material loss; socially and economically, individuals, families, and whole economies have suffered greatly as a result of these collisions. As discussed within the introduction, almost 1.25 million people are lost annually as a result of car accidents, not to mention the tens of millions that have been lost altogether. Thus, the prime objective of this paper has been to utilise data mining techniques so as to examine data concerning Jordanian car accidents, as well as to spot the varying characteristics that impact these accidents.

The construction of an automated system that organises the seriousness of vehicle collisions has been one of the most pressing issues concerning a number of countries, especially Jordan; therefore, within this study, we demonstrated this issue by applying five popular data mining algorithms so as to control our classification problem against dataset, which was derived from JPSD. It was founded that the most capable algorithm was MCAR, which outperformed the other studied algorithms significantly; thus, it is reasonable to conclude that the AC approach is, indeed, capable of tackling this issue. Within the future, we propose to construct a novel, innovative algorithm, so as to better the automated seriousness of car accidents classification problem.

ACKNOWLEDGMENT

This work was supported by University of Petra – Dean of Scientific Research [2018/2/6].

REFERENCES

1. Prati, G., Pietrantoni, L. and Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accident Analysis & Prevention, 101, pp.44-54.
2. Zucco, C., 2018. Data Mining in Bioinformatics. Encyclopedia of Bioinformatics and Computational Biology, Volume 1, 2019, Pages 328-335
3. Punithavalli, M., 2018. REVIEW ON ROAD ACCIDENT DETECTION USING DATA MINING TECHNIQUES. International Journal of Advanced Research in Computer Science, 9(2).
4. Kumar, S. and Toshniwal, D., 2016. A data mining approach to characterize road accident locations. Journal of Modern Transportation, 24(1), pp.62-72.
5. Martín, L., Baena, L., Garach, L., López, G. and de Oña, J., 2014. Using data mining techniques to road safety improvement in Spanish roads. Procedia-Social and Behavioral Sciences, 160, pp.607-614.
6. Tesema, T.B., Abraham, A. and Grosan, C., 2005. Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation, 6(10-11), pp.80-94.
7. Geurts, K., Wets, G., Brijs, T. and Vanhoof, K., 2003. Profiling of high-frequency accident locations by use of association rules. Transportation Research Record: Journal of the Transportation Research Board, (1840), pp.123-130.
8. Abellán, J., López, G. and De OñA, J., 2013. Analysis of traffic accident severity using decision rules via decision trees. Expert Systems with Applications, 40(15), pp.6047-6054.
9. Olutayo, V.A. and Eludire, A.A., 2014. Traffic accident analysis using decision trees and neural networks. International Journal of Information Technology and Computer Science, 2, pp.22-28.
10. Nitsche, P., Thomas, P., Stuetz, R. and Welsh, R., 2017. Pre-crash scenarios at road junctions: a clustering method for car crash data. Accident Analysis & Prevention, 107, pp.137-151.
11. Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*(1), 81–106. http://doi.org/10.1023/A:1022643204877
12. Prati, G., Pietrantoni, L. and Fraboni, F., 2017. Using data mining techniques to predict the severity of bicycle crashes. Accident Analysis & Prevention, 101, pp.44-54.
13. Alikhani, M., Nedaie, A. and Ahmadvand, A., 2013. Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran. Safety science, 60, pp.142-150.
14. Chang, L.Y. and Wang, H.W., 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention, 38(5), pp.1019-1027.
15. Sohn, S.Y. and Shin, H., 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics, 44(1), pp.107-117.
16. Öström, M. and Eriksson, A., 2001. Pedestrian fatalities and alcohol. Accident Analysis & Prevention, 33(2), pp.173-180.
17. Karlaftis, M.G. and Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis & Prevention, 34(3), pp.357-365.
18. Park, Y.J. and Saccomanno, F., 2005. Collision frequency analysis using tree-based stratification. Transportation Research Record: Journal of the Transportation Research Board, (1908), pp.121-129.

AUTHORS PROFILE

Faisal Aburub is a professor at MIS department in University of Petra, Amman, Jordan. He holds PhD degree in Information Systems from the University of the West of England, 2006, Bristol, UK. Dr. Faisal has more than 8 years of experience in information systems including research and development. His research interests are mainly focused on business process modelling, business process improvement building on process modelling, bridging the gap between system models and business process models, and e-learning systems.

Weal Hadi is a head of CIS department at University of Petra. He is an associate professor. Dr. Wael hold a PhD from Arab Academy for Banking and Financial Sciences, Amman, Jordan. His main research is about developing new data mining techniques.