Data Mining technique: Application of Apriori algorithm for road accident analysis

Ryan Clifford Larraquel Perez[1*]

[1]Marinduque State College, Tanza, Boac, Marinduque, Philippines
[*]Corresponding author: ryancliffordperez@gmail.com

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.13.2.2831.2023

Received: July 05th, 2023
Revised: August 10th, 2023
Accepted: August 14th, 2023

Keywords:

apriori; association rule analysis; CRISP-DM; data mining; road accidents

ABSTRACT

Road accidents can happen due to various factors. These factors that contribute to road accidents have cost damage to properties, injuries or deaths and most road accidents are attributable to the lack of knowledge on road safety. To provide safe driving and road safety plans, critical analysis of road accident data is needed, to identify the causes of road accidents. Annually, 1,250,000 people die and 50,000,000 are injured in road accidents worldwide, and fatal road accidents are caused by human error. Improving road conditions is not sufficient, but significantly understanding human errors that cause road accidents, and negligence of corrective and safety driving protocols provided by the concerned government agencies or private organizations. The study aimed to help get insights about the causes of road accidents, and to provide knowledge of road accidents for road safety using Association Rule Mining with the application of the Apriori Algorithm. Association rule mining using the Apriori Algorithm produces significant patterns and insights that help identify the causes of road accidents.

1. Introduction

Numerous vehicles travel the streets and highways every day, it is one of the functional methods of transportation in any place, and these numbers are not getting any lower (El Tayeb, Pareek, & Araar, 2015), so the odds of having a road accident can significantly increase (Luhach, 2017; Solanke & Gotmare, 2018). Road accidents are one of the most fatal causes of injuries, disabilities, and even death in modern society, not only can cause physical damage but in other aspect as well as psychological, economical, and property damages (Atnafu & Kaur, 2017a, 2017b; Kumar & Toshniwal, 2016). The cost of these road accidents can inflict a substantial problem for the victims, the government, and insurance companies for it can cost millions (Luhach, 2017). These road accidents can take place at any time (Tiwari & Kalitin, 2017; Xi, Zhao, Li, & Wang, 2016). It has become a big concern of the government, private groups, and even individuals, considering the prevention and for classification of various reasons for road accidents (Solanke & Gotmare, 2018). However, accidents in nature are uncertain or unsure but understanding why they occur can be considered to inhibit or lessen their occurrence (Bhardwaj, Ridhi, & Kumar, 2017).

Road accidents can take place due to various factors, such as a collision of vehicles, an unseen walking pedestrian or crossing animals, natural obstacles on roads and highways, and the increasing number of vehicles (Comi, Polimeni, & Balsamo, 2022; El Tayeb et al., 2015; Solanke & Gotmare, 2018) increasing number of populations, careless driving, attitude, and behavior are

also accumulating the odds. Driver's attitude and behavior also contribute a connection in road accidents and age increases a factor for the driver's change in attitude and behavior (Luhach, 2017). It is also acknowledged that what people eat, and health issues can offer factors in driving performance (Mulay & Mulatu, 2016). But the causes of road accidents can also be drawn from the negligence of road safety and disregarding traffic rules while driving such as over-speeding, driving while in the influence of alcohol or drugs, and even misjudging driving skill (Gupta, Solanki, & Singh, 2017b; Luhach, 2017; Solanke & Gotmare, 2018). Annually, 1,250,000 people die and 50,000,000 are injured in road accidents worldwide (Punay, 2017b), and fatal road accidents are caused by human error (Punay, 2017b). Improving road conditions is not sufficient, but significantly understanding human errors that cause road accidents, and negligence of corrective and safety driving protocols provided by the concerned government agencies or private organizations.

The published statistical report of July 30, 2018, from the National Quick Statistics Philippine Statistics Authority (Philippine Statistics Authority, 2018) Year 2016, a total of 9,251,565 motor vehicles were registered in the Philippines. 2,055,098 were newly registered vehicles and 7,196,467 were renewed registration vehicles. This number of registered vehicles has increased from previous years. With these numerous vehicles traveling the busy roads and streets every day, road accidents statistics show that an average of 34 Filipinos are killed daily in road accidents, it is also identified as one of the main causes of death for the Filipino youth in the Philippines. In 2016 road accident statistics from the Philippine National Police - Highway Patrol Group (PNP-HPG), 10,000 road accidents are already recorded in the first four (4) months resulting in 549 recorded deaths (Ager, 2016). Ninety percent (90%) of documented road accidents in Metro Manila alone based on Police Blotters are attributable to human errors. The increasing number of vehicles, miscalculated movements while driving, and vehicle malfunction caused by negligence are also recognized as human errors.

The study primarily focused on identifying the causes of road accidents, using data mining techniques with the application of the Apriori Algorithm to produce a significant pattern that helps identify the causes of road accidents (Gupta, Solanki, & Singh, 2017a). Data mining is the technique of extracting unseen knowledge from a huge amount of known and observed datasets by digging out meaningful patterns and association between different sets of variables using various algorithms (Larose & Larose, 2014; Montella, Aria, D'Ambrosio, & Mauriello, 2011). This technique has been applied in many research domains such as fraud detection (Solanke & Gotmare, 2018), financial analysis, bank transactions, crime analysis, and health for digging out useful and unseen information. To provide road safety plans, critical analysis of road accident data is needed, to identify the associations of causes of road accidents (Li, Shrestha, & Hu, 2017). This study primarily focused on road accidents by extracting unknown information and relationships between road accidents and their causes using the Data Mining Technique with the application of the Apriori Algorithm.

2. Related literature

Traffic accident datasets from the Dubai Traffic Department are applied with Association Rule Discovery using the Apriori algorithm and Predictive Apriori to explore the link between recorded accident factors to accident severity in Dubai. The result showed that the association rules generated by the Apriori algorithm are more efficient than the Predictive Apriori algorithm (El Tayeb et al., 2015).

By applying K-Mode Clustering and Association Rule Mining using the Apriori algorithm on road accident datasets, relevant variables were identified that contribute to the severity of road accidents. Using K-mode Clustering, road accident data were clustered into groups and applied improved Apriori algorithm on the clustered data for extracting interesting rules. The results of the proposed technique with K-mode clustering and Apriori are measured by: Clustering Time, Accuracy Rate (AR), and Association Rule mining time and proposed a framework that uses K-Mode's clustering and Improved Apriori to boost the result for analyzing patterns (Kaur, Luhach, & Pooja, 2017).

Gathered road accident data from Emergency Management Research Institute (EMRI) with 9,640 road accident instances from 2009 to 2014 are used for road accident analysis. These records were from 108 ambulances services running across the state with 17 different variables but 13 where only identified as suitable for analysis; and data mining cluster analysis K-modes clustering and association rule mining using Apriori were used for analysis and extraction of unseen knowledge. Data were clustered to find similarities in road accident occurrences and association rules were applied in each cluster to classify causes. The results were identified and utilized by the traffic officers for road safety and accident prevention processes. It was also recognized that 108 ambulance services are a rather lifesaving system (Kumar & Toshniwal, 2015).

By utilizing Statistics and Data Mining algorithms in Fatal Accident datasets, relationships are discovered between fatal rates and various attributes such as collision, weather, surface and light conditions, and drunk driving. The application of the Apriori algorithm, and cluster analysis was used to build a model for prediction. The result shows that environmental factors do not entirely result in fatal rates, and human factors like drunk driving and collision type have much stronger effects on fatal rates of vehicle accidents (Li et al., 2017). Road traffic incident data from UKDA datasets collected from the Department of Transportation of England in 2009 were used to identify different factors in road accidents. Market Basket Analysis with Apriori using a software called WEKA was used for the extraction of interesting rule patterns. The result shows that the characteristics and behavior of an individual are very important in the event of road accidents. The results can also be utilized by highway and transportation engineers for a safer road.

Vehicle accident data from data.gov.uk is gathered for road accident analysis. Cluster Analysis using K-modes Algorithm is applied to create clusters to identify similarities within the accident data. Association Rule Mining Using the Apriori Algorithm were applied to identify event in each cluster that caused vehicle accidents. The results can be used to identify different categories of vehicle accidents for prevention. Also, the approach was compared with the existing method in time and accuracy. Results proved that the used method has better performance (Comi et al., 2022).

573 road accident data from the year 2012 were used to classify the effect of diet routine on driving in decreasing the probabilities of road accidents using Data mining techniques. Association rule discovery with the Apriori algorithm, patterns were extracted from road accidents data and nutrition data from the National Health and Nutrition Examination Survey in the United States, and the results reveal that a driver's diet routine can provide factors and can be one of the causalities of road accidents (Mulay & Mulatu, 2016). Various studies that use data mining techniques with different approaches to road accidents in identifying their severity are gathered and briefly deliberated. Every method that was discussed gives promising and productive results in different ways in decreasing the number of casualties in the event of a vehicle accident. The

research only aimed to find better mining techniques in the application of accident circumstances (Gupta et al., 2017b).

Association Rule Mining using Apriori Algorithm was used in discovering hidden patterns from rainy weather crash data recorded from eight (8) years in Louisiana (2004 - 2011). These crash types were associated with some roadway features i.e., on grade-curve, aligned roadways, curved roadways, and roadways with no streetlights during the evening. Results showed that during rainy weather, Property Damage Only (PDO) and sideswipe (same direction) road crashes are significant in numbers while moderate injuries leading to single-vehicle crashes. Poor illumination on the street is associated with straight level aligned roadways crashes during rainy weather. Drivers aged 15 - 24 are open to run-off road crashes on roads with poor illuminations and curved-aligned roads. These results help highway authorities in identifying solutions for road safety improvement (Das & Sun, 2014).

Association rule mining with the Apriori algorithm was one of the widely used approaches that identify the correlations of various attributes in road accident analysis. It was applied in road accident occurrences in a certain location. By analyzing this location, it can help identify the frequency of road accidents. Before the application of association rule discovery using the Apriori algorithm, data are segmented using K-means clustering to identify high, moderate, and low road accident frequency, then association rule discovery using apriori is applied. Each cluster produces different rules and factors that were associated with road accidents in certain locations, though some clusters reveal the same rules discovered, they have different interest scores. Data with more attributes can uncover more hidden rules with this approach.

3. Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) was applied in the study; this involves obtaining data from various sources and using mathematical quantification tools and techniques to derive results. Research works show the application of CRISP-DM in providing a structured and step-by-step approach for analytical tasks in Data Mining in the road accident domain. CRISP-DM has been applied as a guide in road accident analysis to develop predictive models using road accidents data from Road Traffic Office at Addis Ababa, Ethiopia with 14, 254 accident cases from 2004 - 2005 to analyze driver's behavior and road accidents (Beshah, Ejigu, Abraham, Snasel, & Kromer, 2013).

CRISP-DM is a categorized process model consisting of tasks described with four (4) levels of abstraction; phases, generic task, specific task, and process instances, with six (6) phases as the top level shown in Figure 1. Each phase consists of a second-level generic task; adequately generic to shelter all conceivable data mining conditions. In the third-level, the particular level job defines how the generic task must be accepted in a precise and detailed situation. The fourth level, is the process instance where activity records, results, and decisions of the data mining task are engaged. CRISP-DM provides a map from a generic process to a specific process model (Larose & Larose, 2014).

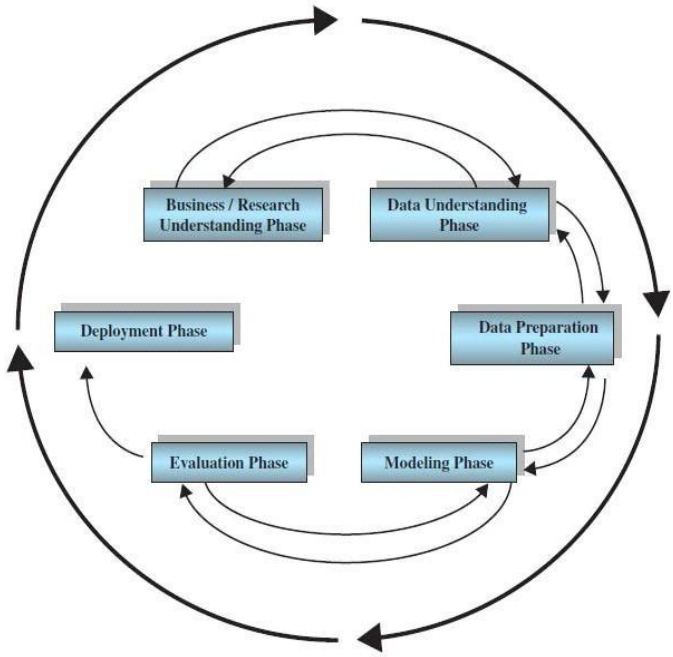

Figure 1. CRISP-DM Stages

Source: Larose and Larose (2014)

Business Understanding Phase: Project objectives and requirements are identified, gathered, and understood to provide an initial plan to get insights about the causes associated with road accidents. For other government agencies or private organizations, a basis for providing necessary, rightful actions and knowledge on road safety.

Data Understanding Phase: Data were gathered from e-Blotter System reported road accidents from 2016 - 2019, to be familiarized with the data in order to determine the quantity and problems within the data. To come up with the initial perception to get insights about interesting subsets and hidden information.

Data Preparation Phase: Data attributes were identified in this phase suitable for the data mining task and problem to create the final dataset that will undergo a series of conversions, binning, and cleaning from the raw data.

Modeling Phase: The technique and algorithm used for modeling that is suitable for the data mining task are identified in this phase. Association Rule Discovery is the technique applied in data mining tasks. Association Rule Discovery is the technique used in data mining for discovering interesting relations between different sets of variables; it is intended to identify interesting rules and patterns with measurements.

Association Rule discovery generates various sets of rules that define a correlation between different sets of attributes in the data set. It assumes that all subsets of a frequent item set must be frequent. Support and Confidence are the measurements that will identify how strong the rules are generated from the road accident data. The support will indicate the frequency of occurrence of a certain rule and confidence will define how reliable the rules are from the road accident data. The association rule that will be generated from road accident data with high confidence and support value is the main interest (Kumar & Toshniwal, 2015; Zhang, 2012). The Support and Confidence is illustrated as:

$$support\ (A \rightarrow B) = P\ (A \cup B) \quad \text{Eq. (1)}$$

$$confidence\ (A \rightarrow B) = P(B|A) \quad \text{Eq. (2)}$$

Support is a measure of the frequency or prevalence of a particular item set in a dataset. In the context of the Apriori algorithm, an item set refers to a set of items that appear together in a given event or transaction (Li et al., 2017). On the other hand, confidence is a measure of the reliability or strength of an association rule. An association rule consists of an antecedent (a set of items) and a consequent (a single item). Confidence quantifies the likelihood that the consequent will be present in an event given that the antecedent is present (Kaur et al., 2017). These metrics help Apriori discover significant patterns and reliable rules from data, supporting market basket analysis and decision-making (Comi et al., 2022).

Evaluation Phase: The model's result was evaluated if it achieved the desired result from the research objective.

Deployment Phase: Final results were reported which includes detailed findings, model explorations, and other desired outputs to be presented to determine and discuss if the goal of the data mining task has been come across.

4. Result and discussion

Association Rule discovery generates an arbitrary set of rules that describes a correlation between sets of features in the given datasets using the Apriori Algorithm. The strength of each generated rule was measured using support and confidence.

a) minimum Support = Support (0.01)

b) minimum confidence = Confidence (0.7)

The minimum support and confidence were set to the Apriori algorithm to identify the association to the 'Probable Cause' of road accidents using python 3.0. The generated rules with the highest association with the 'Probable Cause' are chosen.

Table 1 shows the generated rule from road accidents that involved drivers who were under the influence of intoxicated substances while driving. The rule shows that these road accidents involve riding a motorcycle or any vehicle while drunk. Driving while under the influence of intoxicated substances or liquor is too dangerous for both drivers, incoming traffic, and pedestrians, due to the fact that the driver's senses are not as sharp resulting in worse case in a fatal road accident.

Table 1

'Intoxicated' as identified probable cause

Antecedent	Consequent	Conf.	Min. Support
{Motorcycle, Male Suspect} =>	{Intoxicated}	0.90	0.10

Table 2 shows generated rules from road accidents that involve vehicles driving at high speed. This road accident involved drivers that were driving above the speed limit. The rule shows motorcycle-related accidents causing physical injury for both the suspect and the victim. On some occasions, it could result in a fatal accident.

Table 2

'Over Speeding' as identified probable cause

Antecedent	Consequent	Conf.	Min. Support
{Male Suspect, Physical Injury, Motorcycle, speeding} =>	{Over-speeding}	0.80	0.10

Table 3 shows generated rules from road accidents that have the consequence of 'Premature Overtaking' as the 'Probable Cause'. Premature overtaking are case in which the vehicle is in the process of taking the lead of the in-front vehicle but unfortunately miscalculating the process resulting in sideswiping or bumping with another vehicle. This road accident mostly involved motorcycles. Overtaking is one of the most dangerous actions while driving and should be performed with extreme precaution.

Table 3

'Premature overtaking' as identified probable cause

Antecedent	Consequent	Conf.	Min. Support
{Male Suspect, motorcycle, overtake bump} =>	{Premature Overtaking}	0.88	0.10

Table 4 shows generated rules from road accidents that involved Improper turning of vehicles. It has been identified that the proper method for a vehicle to make a turn is by turning on the signal lights for at least ten (10) seconds prior to warning any incoming traffic that the vehicle will change its direction. The generated rule shows that Improper turning are case in which a vehicle is taking a turn or changing direction without turning on the signal lights or changing direction without giving signs resulting in the incoming traffic accidentally hitting or being hit by the vehicle while making a turn. These road accidents mostly involved motorcycles. Though this road accident is typically non-fatal but could cause physical injuries.

Table 4

'Improper turn' as identified probable cause

Antecedent	Consequent	Conf.	Min. Support
{Unharmed Victim, hit turn} =>	{Improper Turn}	0.80	0.10
{Motorcycle, hit turn} =>	{Improper Turn}	0.70	0.10

Table 5 shows generated rules from road accidents that involve vehicular malfunctions. Vehicular Malfunctions involve vehicles typically losing brakes or control. These road accidents are due to negligence of proper and regular checking of the vehicle condition before driving. And should frequently be done in advance.

Table 5

'Mechanical Malfunction' as identified probable cause

Antecedent	Consequent	Conf.	Min. Support
{Male Suspect, Unharmed Victim, hit, lost brake} =>	{Mechanical Malfunction}	0.70	0.10

5. Conclusions & recommendations

The association rule discovery using the Apriori algorithm shows insights into identifying the causes of road accidents. Motorcycles are generally the most ideal type of vehicle today since it is significantly easier to acquire. But it is also the vehicle most involved in a road accident. Motorcycle riders have the tendency to be more aggressive in riding which results in road accident collisions with incoming traffic and unseen pedestrians. Motorcycles are also significantly easier to maneuver in traffic compared with other vehicle type, and due to that, motorcycle riders have a greater leaning to take the lead or overtake and pass through even in bad traffic conditions, and miscalculations or wrong judgment from overtaking results in road accidents which causes road accidents. The study mainly focused on identifying the causes of road accidents, with the application of association rule mining using the Apriori Algorithm to produce significant patterns that help identify the causes of road accidents.

1. Explore the potential of classification tasks in data mining with the Hidden Markov Theorem to predict and classify road accidents from fatal or non-fatal using the significant attributes in road accidents.

2. Use road accident data with road conditions and weather to evaluate the occurrence of fatal and non-fatal road accidents from different weather and road conditions.

Explore the application of classification tasks in data mining with decision trees in giving insights for the public in making decisions if it is going to be safe to drive and provide recommendations.

References

Ager, M. (2016). *10,000 vehicular accidents in first 4 months of 2016 - police*. Retrieved October 10, 2022, from Inquirer.Net. website: https://newsinfo.inquirer.net/821966/10000-vehicular-accidents-in-first-4-months-of-2016-police

Atnafu, B., & Kaur, G. (2017a). Analysis and predict the nature of road traffic accident using data mining techniques in Maharashtra India. *International Journal of Engineering Technology Science and Research (IJETSR), 4*(10), 1153-1162.

Atnafu, B., & Kaur, G. (2017b). Survey on analysis and prediction of road traffic accident severity levels using data mining techniques in Maharashtra, India. *International Journal of Current Engineering and Technology*, *7*(6), 2277-4106.

Beshah, T., Ejigu, D., Abraham, A., Snasel, V., & Kromer, P. (2013). Mining pattern from road accident data: Role of road user's behaviour and implications for improving road safety. *International Journal of Tomography and Simulation*, *22*(1), 73-86.

Bhardwaj, R., Ridhi, R., & Kumar, R. (2017). Modified approach of cluster algorithm to analysis road accident. *International Journal of Computer Applications*, *166*(2), 24-28.

Comi, A., Polimeni, A., & Balsamo, C. (2022). Road accident analysis with data mining approach: Evidence from Rome. *Transportation Research Procedia*, *62*(2022), 798-805.

Das, S., & Sun, X. (2014). Investigating the pattern of traffic crashes under rainy weather by association rules in data mining. *Transportation Research Board 93rd Annual Meeting*, *12*(14), 14-1540.

El Tayeb, A. A., Pareek, V., & Araar, A. (2015). Applying association rules mining algorithms for traffic accidents in Dubai. *International Journal of Soft Computing and Engineering*, *5*(4), 1-12.

Gupta, M., Solanki, V. K., & Singh, V. K. (2017a). A novel framework to use association rule mining for classification of traffic accident severity. *Ingeniería Solidaria, 13*(21), 37-44.

Gupta, M., Solanki, V. K., & Singh, V. K. (2017b). Analysis of datamining technique for traffic accident severity problem: A review. *RICE, 10*(19), 197-199.

Kaur, I., Luhach, A. K., & Pooja. (2017). Proposing k-mode based methodology for road accidents with improved apriori. *International Journal of Engineering Applied Sciences and Technology, 2*(5), 66-74.

Kumar, S., & Toshniwal, D. (2015). Analysing road accident data using association rule mining. *2015 International Conference on Computing, Communication and Security (ICCCS)*, 1-6.

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data*, *3*(13), 1-11.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining* (Vol. 4). Hoboken, NJ: John Wiley & Sons.

Li, L., Shrestha, S., & Hu, G. (2017). Analysis of road traffic fatal accidents using data mining techniques. *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, 363-370.

Luhach, A. K. (2017). Mining of road accident data using K-Mode clustering and improved apriori. *International Journal of Computer Science and Information Security, 15*(4), 235-249.

Montella, A., Aria, M., D'Ambrosio, A., & Mauriello, F. (2011). Data-mining techniques for exploratory analysis of pedestrian crashes. *Transportation Research Record, 2237*(1), 107-116.

Mulay, P., & Mulatu, S. (2016). What you eat matters road safety: A data mining approach. *Indian Journal of Science and Technology, 9*(15), 1-8.

Philippine Statistics Authority. (2018). *Quickstat*. Retrieved October 10, 2022, from http://psa.gov.ph/statistics/quickstat

Punay, E. (2017a). *Human error: Leading cause of road mishaps in Metro Manila*. Retrieved October 10, 2022, from https://www.philstar.com/headlines/2017/05/21/1702367/34-pinoys-die-daily-road-mishaps

Punay, E. (2017b). *34 Pinoys die daily in road mishaps*. Retrieved October 10, 2022, from https://www.philstar.com/headlines/2017/05/21/1702367/34-pinoys-die-daily-road-mishaps

Solanke, N. A., & Gotmare, A. D. (2018). Analysis of roadway traffic using data mining techniques for providing safety measures to avoid fatal accidents. *International Journal on Future Revolution in Computer Science and Communication Engineering, 4*(6), 45-50.

Tiwari, P., & Kalitin, D. (2017). A conjoint analysis of road accident data using k-modes clustering and sayesian networks (Road accident analysis using clustering and classification). *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, 10*(15), 53-56.

Xi, J., Zhao, Z., Li, W., & Wang, Q. (2016). A traffic accident causation analysis method based on AHP-Apriori. *Procedia Engineering, 137*(216), 680-687.

Zhang, S.-L. (2012). A new mining algorithm of association rules and applications. *Bio-Inspired Computing and Applications: 7th International Conference on Intelligent Computing. 7*(6840), 123-128.