International Journal of Software Engineering and Computer Systems (IJSECS)
ISSN: 2289-8522, Volume 5 Issue 1, pp. 77-92, February 2019

https://doi.org/10.15282/ijsecs.5.1.2019.6.0056

PREDICTION OF TRAFFIC ACCIDENT SEVERITY USING DATA MINING TECHNIQUES IN IBB PROVINCE, YEMEN

Muneer A.S. Hazaa[a], Redhwan M.A. Saad[b], and Mohammed A. Alnaklani[*a]

[a] *Faculty of Computer Sciences and Information Systems, Thamar University, Thamar, Yemen.*

Email:muneer_hazaa@yahoo.com

[b] *Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen.*

Email:redhwan@nav6.usm.my

[*a] *Correspondence Author: Mohammed A. Alnakhlani,*

Email:m.alnaklani@gmail.com.

ABSTRACT

Traffic accidents are the leading causes beyond death; it is the concern of most countries that strive for finding radical solutions to this problem. There are several methods used in the process of forecasting traffic accidents such as classification, assembly, association, etc. This paper surveyed the latest studies in the field of traffic accident prediction; the most important tools and algorithms were used in the prediction process such as Back-propagation Neural Networks and the decision tree. In addition, this paper proposed a model for predicting traffic accidents based on dataset obtained from the Directorate General of Traffic Statistics, Ibb, Yemen.

Keywords: Traffic Accidents, Neural Network, Decision Tree, Back-Propagation Algorithm.

INTRODUCTION

Traffic accidents can be considered as a direct threat to human life and property. There are many variables and factors that contribute directly or indirectly to traffic accidents. Although many studies have been conducted to investigate this problem, it is still difficult to find a radical solution to it .According to statistics produced by the World Health Organization (Organization, 2015), the number of traffic accidents increases dramatically and worries most of countries. Based on Road Safety Report (2015), the death cost was 1.25 million, and road accidents scored is the9th major cause death. It is expected to become the 7th leading cause for death by 2030. Most for of the victims were young people between the age of 15 and 29.

Therefore, in this paper, a comprehensive survey of the most recent studies that deal with this problem was studied in order to conduct a deep investigation into it. However, this study aimed to identify the best and most accurate techniques used in data extraction,

which contained a number of algorithms used in the process of prediction and the relationship between both dependent and independent variables.

RELATED WORK

There are several studies in the field of traffic accident prediction conducted by various researchers, focusing on the most important factors that cause traffic accidents and variables. This variable refers to the risks that describe the relationships between them to generate the necessary results of methodology, techniques, and standards for developing the proposed system. The classification of traffic accident prediction techniques is shown in Figure 1.below:

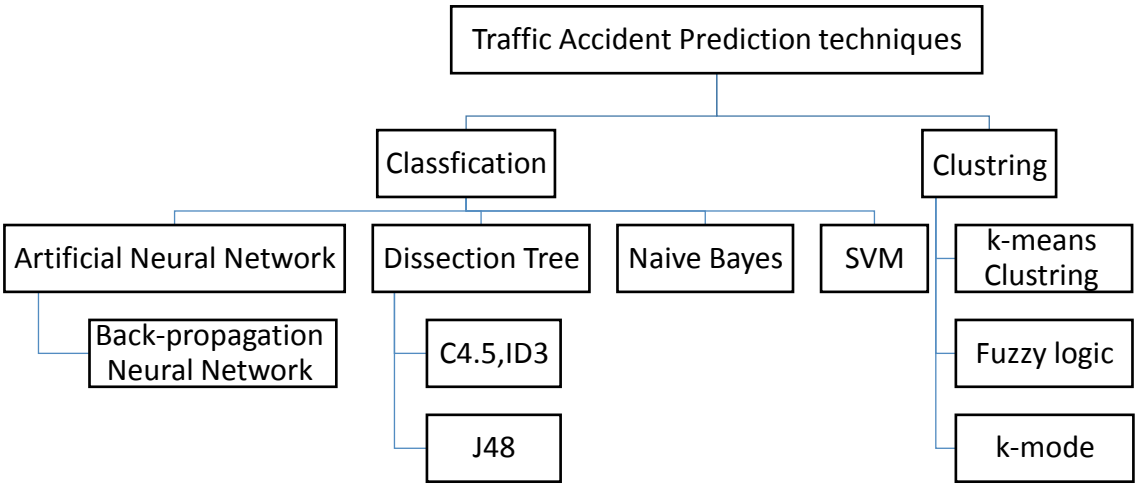

Figure 1: Traffic Accident Prediction Techniques Classification

Classification Technique

Classification is very important in the process of predicting traffic accidents and recently there are many researchers trying to perform a forecast for irrigated incidents, which can be reviewed in the following studies:

Artificial Neural Network:

Based on an excellent review of traffic accident prediction presented by (Alkheder, Taamneh, & Taamneh, 2017), which aimed to use artificial neural network and PROPIT model to forecast traffic accidents, the results showed that k-means and neural networks algorithm could predict accidents accurately as compared with PROPIT model. The accuracy of the network prediction of ANN was 74.6% while the PROPIT model scored a lower accuracy of 59.5%. The model was developed by. (Jadaan, Al-Fayyad, & Gammoh, 2014)for predicting traffic accidents using neural networks and determining their suitability to predict traffic accidents. The results showed that the model of accident prediction using neural networks was developed with an error coefficient, R = 0.992, by

analyzing the relationship between accidents and features that affected accidents. This model was validated and obtained good results which could be relied on to predict expected traffic accidents in Jordan.

Another study was done by(Ghani, Raqib, Sanik, Mokhtar, & Aida, 2011), which compared between two models, Multiple Linear Regression (MLR) model and Artificial Neural Network (ANN) in Malaysia. The results indicated that the MLR model was better in R 2 (99.92%) and at the same time, the model ANN (82.40%) indicated lower R2 than that of the MLR model. Therefore the MLR model was better than the ANN model.

A comparison between (ANN) and multivariate analysis (MVA)was done by (De Luca, 2015), which occurred in southern Italy between 2001 and 2005, it used cluster analysis with binary partition algorithm hard-c-mean. Two models were obtained, namely ANN and MVA. The conclusion of comparing the two models showed that the model ANN is better than the MVA model while the MVA model was the best in describing the darker and dangerous spots.

Another study conducted by (Ali & Bakheit, 2011) in which they made a comparative analysis of traffic accident prediction in Sudan using neural networks and statistical methods. The study concluded that the analysis and prediction that used neural networks were better than R regression technique. A recent study was done by (Contreras, Torres-Treviño, & Torres, 2018) which aimed to predict car accidents using the maximum sensitivity of the neural network was advanced, trained and verified using the Scilab development program. The result was concluded with the neuronal network of the maximum sensitivity in that it was possible to predict the occurrence of events weighting them by the times in which they were presented in the historical data.

A study by (Y. Li, Ma, Zhu, Zeng, & Wang, 2018), which aimed to identify the most important factors affecting the occurrence of accidents using the genetic algorithm (NSGA-11) multi-objective optimization and neural networks. It was found that the most important factors in terms of temporal and spatial perspective were the hour and day. This method also provides a new vision in the pattern of road injuries that can be used to raise awareness and improve understanding of prevention from future accidents.

A study by (Odhiambo, Wanjoya, & Waititu, 2015) that aimed to find out the causes of accidents and how to reduce them. They started appropriate safety measures in Nairobi province by using neural networks and compared them with Negative binomial regression, It was concluded that neural networks gave the best accuracy while high-performance of negative binomial regression. This study is not recommended for spatial modeling in future research because of not taken into account explanatory variables.

Back-Propagation Neural Network:

The back-propagation algorithm is one of the most widely used methods in the process of predicting traffic accidents because of its efficiency and high predictability. The most important of these studies are as follows. A Study was done by (Mussone, Bassani, & Masci, 2017), which aimed at determining the most important factors that affect the occurrence of accidents prediction using environmental variables and movement variables using the back- propagation network and generalized linear mixed model. The study concluded that BPNN scored the best performance of GLMMs.

A study was conducted by (Wenqi, Dongyu, & Menghua, 2017) that aimed to predict traffic accidents and find the factors of prediction by using neural networks CNN and back propagation neural network. It concluded that using neural networks CNN gave a better

performance than back propagation neural network. Although this study predicted the incidents well, there is still a lack of study because it does not include more road characteristics such as road alignment, road grade, and lanes to get more accurate of prediction. However, the data for the training process were few.

Dissection Tree

Decision tree plays an important role in the prediction process because of its ability to address the problems related to the classification and prediction of independent values. It was used by many researchers in the process of predicting traffic accidents, such as:

1. C4.5

In a major study in the field of classification conducted by (Olutayo & Eludire, 2014), which aimed to analyze traffic accidents using neural networks and decision tree resolutions for approximately 41,770 traffic accidents that occurred in the USA during 1995-2000. The results showed that the decision tree model gave better results than neural networks, and the most important factors leading to death were lack of wearing seatbelts, light on the road and drinking alcohol while driving.

Another study conducted by (Zhang & Fan, 2013) using low-resolution algorithms ID3, C4.5 as the main contributor to the occurrence of traffic accidents. The results indicated that the data extraction model using the decision tree cloud effectively classify the main factors contributing to the occurrence of traffic accidents. The most prominent ones were drinking, and non-compliance with traffic rules (e.g. distraction negligence and lack of experience in the driving process). Although the program self-developed by the researchers gave more accurate and reliable results, it still needs to develop and introduce criteria vehicles and drivers to give good results.

Another study by (Hashmienejad & Hasheminejad, 2017), which aimed to predict a traffic accident severity according to users preferences instead of conventional DTs using a multi-objective genetic algorithm known as NSGA-11. The study concluded that the proposed method was superior in terms of accuracy (88, 2%) and in terms of rules of support and confidence (0.79) and (0.74) as compared to the rest of methods that provided less accuracy and fewer rules of support and confidence.

2. J48:

A study done by (Al-Turaiki, Aloumi, Aloumi, & Alghamdi, 2016), which aimed at applying the classification in order to understand the most important factors in traffic accidents in Riyadh using the algorithms of CHAID, J48, and Naive Bayes. The study concluded that distraction during the use of driving was the most important factor leading to death and injury. Although the study clearly and explicitly determined that distraction during driving was an important factor in accidents, it is necessary to incorporate more road data for better results.

Naïve Bayes

A study conducted by (Kashyap & Singh, 2016) that aimed to identify the causes of accidents and how to reduce them with a focus on contribution of different inputs such as environment and the animals that are abruptly cut using the naive Bayes algorithm. The results showed that the naive Bayes model, when used with the Weka, was accurate by (45%). This study differs from previous studies in that it added new characteristics such as animal collisions, weather conditions and the condition for vehicle and good results. Here, we understand that when more properties are provided we get more accuracy and good results.

Another study investigated the most important factors affecting traffic accidents conducted by (Atnafu & Kaur, 2017a), which aimed to analyze and predict the nature of road traffic accidents using data mining techniques. The results showed that there were five main important factors emerging (straight road - four ordivier -unmanned rail crossing - fine (variable weather)).

(Zong, Xu, & Zhang, 2013) has compared the Bayzen network and the linear regression model to forecast traffic accidents. The results indicated that Biyzen Network was more suitable for predicting the risk of accidents than the linear regression model. The disadvantage of this model was the lack of some factors which affected the occurrence of accidents such as the characteristics of the driver, the characteristics of the vehicle, and the condition of traffic itself.

Support Vector Machine

A study done by (Tiwari, Kumar, & Kalitin, 2017), which aimed to analyze road accidents and find the most important factors that contributed to the accident using SVM and naive Bayes. It concluded that the decision tree using the k-modes algorithm gave the best performance compared to the rest of the methods used. In terms of comparison, a study was conducted by (Yu, Wang, Yao, & Wang, 2016), that aimed to predict traffic accident by comparing the performance of ANN and SVM models concluded that both models had the ability to predict traffic accidents at the time of the accident within acceptable limits. ANN gave better performance than SVM in long-term accidents while SVM gave better performance in the overall performance of forecasting the time of traffic accidents

Clustering

K-means Clustering

A study done by (Janani & Devi, 2018), that aimed to predict traffic accidents by using data mining and find the most important factors that caused most of accidents at the time of accidents, a predictive model was constructed using Naive Bayes and k-means clustering and association rule. The result showed that the Naive Bayes model gave the best accuracy of 92.45% as compared to other models.

Another study done by (Gaber, Wahaballa, Othman, & Diab, 2017), developed a model for predicting traffic accidents of the Western Desert Road in Aswan using fuzzy logic where the main objective was to detect the factors affecting traffic accidents. The study concluded that there was a correlation coefficient of 88% when compared prediction of the use of fuzzy logic with the actual data of accidents. The researchers recommended

increasing the width of the road and enhancing efficiency of traffic signals and removing and repairing road defects with full control of side entrances.

A study conducted by (Žunić, Djedović, & Đonko, 2017) aimed to use a clustering model to categorize the causes of traffic accidents and analyze the impact of road, vehicle, environment, and drivers on traffic accident using time series by k-means clustering. The results obtained were positive and satisfactory, using the prescribed method.

Fuzzy Logic

(Perone, 2015), predicted the risk of traffic accidents in the city of Porto Alleger, Brazil. The experimental results showed that the prediction could be established to assess the risk of injury models with better accuracy, even with limited data sets. The disadvantage of this model is that it does not use geospatial data.

K-Modes

A study done by (Kumar, Toshniwal, & Parida, 2017), aimed to compare the analysis of heterogeneity in road accident data, using the techniques of data extraction (k-modes clustering and latent class clustering and FP growth algorithms association rules). The study concluded that both methods were suitable for the absence of homogeneity of road accidents and the rules established. There was no homogeneity in the entire dataset

Table 1 showed that many algorithms used in the process of data to predict traffic accidents. The selection of these algorithms depends on the characteristics of these data and the main objective of the extraction of data was that the most commonly used algorithms were Neural Networks and Naive Bayes which generated positive results in the prediction process.

Table 1**:** Summary of the Most Important Algorithms Used to Predict Traffic Accidents.

Authors	Techniques Used	Algorithm Performance	Objective	Result
(Janani & Devi, 2018)	• k-means • naïve Bayes • fuzzy logic	• Naïve • Bayes = 92.45%	•To predict traffic accidents by using techniques for data mining and find the most important factors that cause most accidents at the time of accidents.	• Naïve Bayes gave the best accuracy of 92.45% over the other models and recommended that the authorities used this study to enhance road safety.
(Alkheder et al., 2017)	• MLP • Probit model • k-means clustering	• ANN = 74.6% • Propit = 59.5%	• To predict traffic accidents using neural networks.	• Neural networks can predict better accuracy than gamma propit
(Hashmienejad & Hasheminejad, 2017)	• NSGA-II • C4.5 • CART • ID3 • NAIVE	• 88.20% • 55.78% • 61.43% • 44.65% • 45.21%	• To predict traffic accident severity according to users preferences instead of conventional DTs and using a multi-objective genetic algorithm	• The suggested method NSGA-II It gave superior performances such as precision 88.20%, as well as support, rules 0.79, and

	BAYES • KNN • SVM • ANN	• 34.33% • 81.24% • 85.37%		trust 0.74.
(L. Li, Shrestha, & Hu, 2017)	• Aprior • Naive Bayes • K-Means	• Naïve Bayes = 67.95%	• To Know the most factors that affect accidents using data mining techniques	• The southern region of the United States had more than350% of people involved in the accident compared to the east of the country. The human factor also affected accidents more in the occurrence of accidents
(Delen, Tomak, Topuz, & Eryarsoy, 2017)	• ANN • SVM • C5 • LR	• ANN = 85.77% • SVM = 90.41% • C5 = 87.61% • LR = 76.96%	• To check the most important factors that affect the severity of the incident that affect the level of severity of the injury	• Non - use of seat belt and the method of collision and drugs are the most important factors that affect the severity of the injury.
(Kumar & Toshniwal, 2017)	• CART • Naive Bayes • SVM	• CART = 87.10% • Naive Bayes = 74.14% • SVM = 79.79%	• To analyze newly available PTWs road accident data from UTTARAKHAND state in India	• That tree CART = 87.10 % show better accuracy than other techniques and therefore was selected to extract the factors that affected the severity of accidents
(Atnafu & Kaur, 2017a)	• Random tree • J48 • Naive Bayes	• Random tree = 98.3% • J48 = 97.5%	• To analyze and predict the nature of road traffic accident using data mining techniques and find the most influential factors on the accident.	• Using an algorithm prior The five most important factors emerged (straight road-four ordivier - unmanned rail crossing-fine (variable weather))
(Žunić et al., 2017)	• k-means	• 80-95%	• To analyze the impact of road, environment, vehicles, and drivers on traffic accidents using time series	• The cases performed were satisfactory and it was possible to establish a list of causes of the problems in hierarchy using the method described.
(Wenqi et al., 2017)	• BP • CNN	• BP = 70.8% • CNN = 78.5%	• To predict traffic accidents based on Convolutional neural network and to find the most influential factors during the accident.	• Neural Networks CNN ware better than neural networks BP
(Mussone et al., 2017)	• BPNN • GLMM	• BPN Gave the best performance of GLMM	• To analyze the factors affecting the severity of crashes in urban road intersections.	• BPNN performed better than GLMM as it had the ability to predict and search for the relationship between variables

(Yu et al., 2016)	• MLP • SVM • MAE • RMSE	• ANN(MLP) Better for long-term accidents • SVM Best For the overall	• Aimed to predict traffic accident by comparing the performance of ANN and SVM models	• It concluded that both models had the ability to predict traffic accidents at the time of the accident within acceptable limits
(Olutayo & Eludire, 2014)	• RBF Networks • Id3 • FT • MLP Networks	• RBF =0.547 • Id3=0.777 • FT=0.703 • MLP =399	• Aimed to analyze traffic accidents using neural networks and decision tree resolutions for approximately 41,770 traffic accidents occurred in the United States of America for the period 1995-2000.	• Decision tree model gave better results than neural networks .The most important factors leading to death ware not wearing seatbelts, light on the road and drinking alcohol while driving.

TECHNIQUES USED FOR PREDICTING

Due to the importance of the subject, the use of techniques that assist in the process of predicting traffic accidents and determining the most important factors has various impact on accidents. Listed below are the different techniques used in this study to predict traffic accidents:

Classification Techniques

Classification is one of the most important techniques used to analyze data. It extracts models that classify categories and classifications of important data (Nikam, 2015). The classification techniques are used to analyze traffic accidents that occurred in the province of Ibb, Yemen which can be considered as supervised learning algorithms. The model is based on a set of previously known records called training data set, the model is then evaluated using unknown records called test dataset (Al-Turaiki et al., 2016). The techniques used for predicting is shown in Figure 2 below:

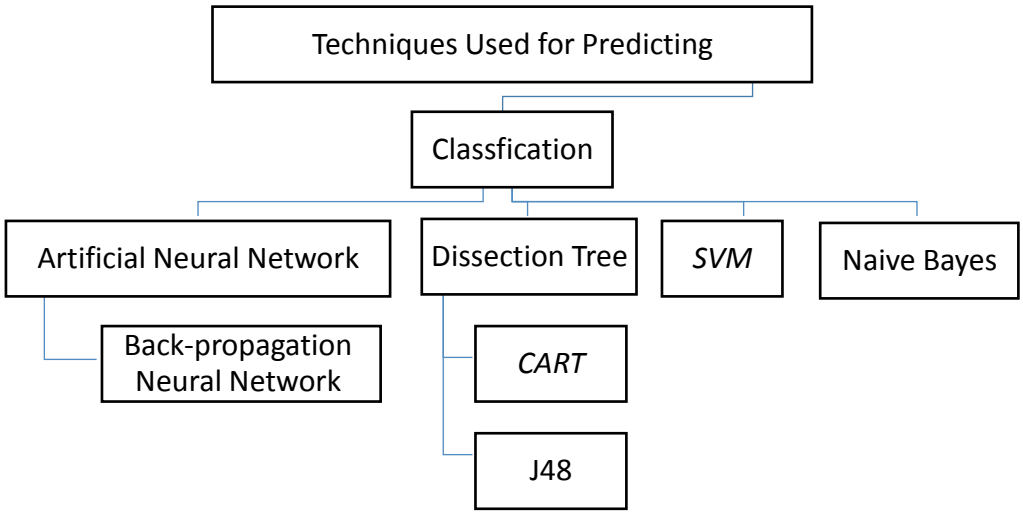

Figure 2: Techniques Used For Predicting

1. Back Propagation Neural Network

Back Propagation Algorithm is a member of the family of gradient descent algorithms and is an implementation of Delta rule. By iteratively adding the negative of the slope of the function to the value of the function at a given point, the algorithm tends to reach the maxima or the minima. This is how the gradient descent algorithm minimizes the error (Al-Maqaleh, A. Al-Mansoub, & N. Al-Badani, 2016). Back Propagation is a supervised learning method, since, the output for a given set of data inputs should be known before proceeding with the algorithm. The output is the key with which the algorithm ascertains the loss function gradient. It is considered a fully connected artificial neural network with one hidden layer (Sikka, 2014).

2. Decision Tree

The decision tree has an important role in the process of classifying particular data through a number of different levels of decisions that help in reaching a final decision. Where they are represented in a tree structure. To make a series of decisions, the tree structure is used to classify unknown data records. The tree resolution is characterized as easy to interpret and convert to understandable IF-THEN rules as they have higher performance and greater accuracy compared to other class models(Al-Turaiki et al., 2016).

a) Classification and Regressing Tree (CART)

CART is one of the decision tree algorithms developed by (Gokgoz & Subasi, 2015). The technique of repeated division according to certain criteria is used to create the contract. The tree is created using established and split nodes. Before applying the criteria of the repeated division according to certain criteria, the contract must have a better split point by processing variance function where the generated function is applied to each split point to calculate the best point of segmentation(Gokgoz & Subasi, 2015).

b) J48

J48 is one of the supervised learning algorithm used to predict the value of the target variable using the decision rules. Each internal axis in this tree has a characteristic property, and each leaf hub is compared to a class name. The record's attribute values are continuously compared with other internal nodes of the tree until a leaf node is reached with predicted class value. It uses a pruning method for construction of the tree. This method reduces the size (Prabakaran & Mitra, 2018).

3. Support Vectors Machine (SVM)

SVM is a method of classification which has the ability to deal with both linear and non-linear data. The support vectors is found to provide a compact description of the learned model. SVMs can be used for numeric prediction as well as classification. They have been applied to a number of areas, including handwritten digit recognition, object recognition, and speaker identification, as well as benchmark time-series prediction tests (Han, Pei, & Kamber, 2011).

4. Naive Bayes

The Naive Bayesian classifier is one of the most effective and widely used supervised learning algorithms which can be used to classify data. It is a statistical model that predicts class membership probabilities based on Bayes' theorem. The Naive Bayes classification algorithm is the probability-based methods used for classification and prediction based on the Bayes' hypothesis with the assumption of independence between each pair of variables (Atnafu & Kaur, 2017b).

DATA MINING TECHNIQUES AND TOOLS

When conducting data mining, especially with large data, we need special tools to analyze them and know patterns and relationships between them. The tools are as follows:

Weka

Weka is one of the most popular open sources of data mining software. It consists of many data mining components, included in many other tools such as Rapid Miner, Rattle and KNIME. It was developed at Waikato University in New Zealand in 1992. It is a Java-based tool that can be used to implement many automated learning algorithms and explore data written in Java. Weka offers three ways to use the tool: Java API, GUI, and CLI. WEKA. It contains classification, compilation, assembling rules for mining algorithms and data processing tools (Atnafu & Kaur, 2017b).

Orange

The Orange tools are characterized by their simplicity and creative graphical interface that requires limited knowledge of data mining, as compared to other data mining tools, its strength lies in the interactive display function that allows beneficiaries to display flags and then select data points or nodes directly from the charts. It supports programming languages like C, C++, and Python. This data mining tool supports Mac OS, Windows, and Linux(Kukasvadiya & Divecha, 2017).

Rapid Miner

RAPID MINER is a software platform developed by the same company previously known as YALE on Weka. It includes additional powerful functions for data analysis such as processing data visualization presentation and additional machine learning algorithms. This tool is more innovative than Weka, Providing an integrated environment for automated learning, data mining, text mining, predictive analysis, and business analysis. Rapid Miner uses the client/server model with the server provided as a software or cloud infrastructure (Slater, Joksimović, Kovanovic, Baker, & Gasevic, 2017).

Ratlee

Ratlee is a programming language for statistical computing and graphics. It is widely used among statisticians and miners to develop statistical software and analyze data. One of the strengths of R is that it is easy to use. Rattle also provides a custom graphical user interface

to explore data although understanding the tool does not require starting to use it in jobs of core data mining. But it is commensurate with the usual beneficiaries on R, The tool is also integrated with two tools specialized in the analysis of interactive graphics data. Latticist (Bhinge, 2015).

Table 2: comparison between Tools Used Data Mining

Total	RAPID MINER	WEKA	ORANG	KNIME	RATEEL	MATLAB
Usability	Easy to use	Easy to use	Easy to use	Easy to use	Complicated as coding required	Easy to use
OS platform	Windows ،Mac OS X Linux	Windows ،Mac OS X Linux	Windows ،Mac OS X Linux	Windows ،Mac OS X Linux	Windows ،Mac OS X Linux	Windows, Mac, Linux
Speed	Requires more memory to operate	Works faster on any machine.	Works faster	-	Works fast on any machine	Specifically optimized for best possible performance
Language	Java	Java	C, C++ and Python	Java	C, Fortran and R	Matlab C, Fortran
visualization	More options but less than Tableau	Fewer options	More options	Better visualization	Fewer options as compared Rapid Miner	Better visualization
Algorithms supported	Classification and Clustering	Classification and Clustering	Classification and registration	Classification and Clustering	Very few Classification and Clustering algorithms	Classification and Clustering
Data Set Size	Supports large and small dataset	Supports only small datasets	Supports average data	Supports average data	Supports large and small dataset	Supports large and small dataset
Memory Usage	Requires more memory	Less Memory hence works faster	More Memory	-	More Memory	More Memory
Primary Usage	Data Mining, Predictive Analysis	Machine Learning	Machine Learning	Data Mining, Predictive Analysis	Statistical Computing	Machine Learning
Interface Type Supported	GUI	GUI / CLI	GUI	GUI	CLI	GUI

Knime

It is an open source data analysis, used for reporting and integrating platform. It is based on the Eclipse platform and, through its modular API, is easily extensible. Custom nodes and types can be implemented in KNIME within hours to extend KNIME to comprehend and provide first-tier support for highly domain-specific data format. KNIME is also one of the best internal tools that support the tools which help new beneficiaries to build data mining. It supports R script and Python (Atnafu & Kaur, 2017b).

Matlab

MATLAB is a software development environment that offers a high-performance numerical computation, data analysis, visualization of capabilities, and application development tools developed by Cleve Muller in 1970. The initial programming language was written in FORTRAN and a new set of libraries was rewritten to process the matrix in

2000. The dynamically downloadable file files created by assembling the assembler functions "MEX-files" (for MATLAB executable). Since 2014, bidirectional interaction with Python has been added. Libraries written in Perl, Java, ActiveX, or .NET can be called directly from MATLAB.In MATLAB. It can be used for classification and regression, decision trees, Bayesian, logical clusters, association rules, and another algorithm (Amardeep, 2017). Table 2 shows comparison of tools used in data mining.

PROPOSED SYSTEM FOR PREDICTING TRAFFIC ACCIDENTS

In the proposed system, the use of IGR Technique and PCA Technique occur in Weka to analyze road accident data to extract the most important factors that affect traffic accidents. The propagation neural network and Decision tree model were used to predict traffic accidents in Ibb, Yemen. The two models were compared with each other to determine which accuracy is better in the prediction process. This in turn, contributes to the reduction of traffic accidents as shown in Figure 3.

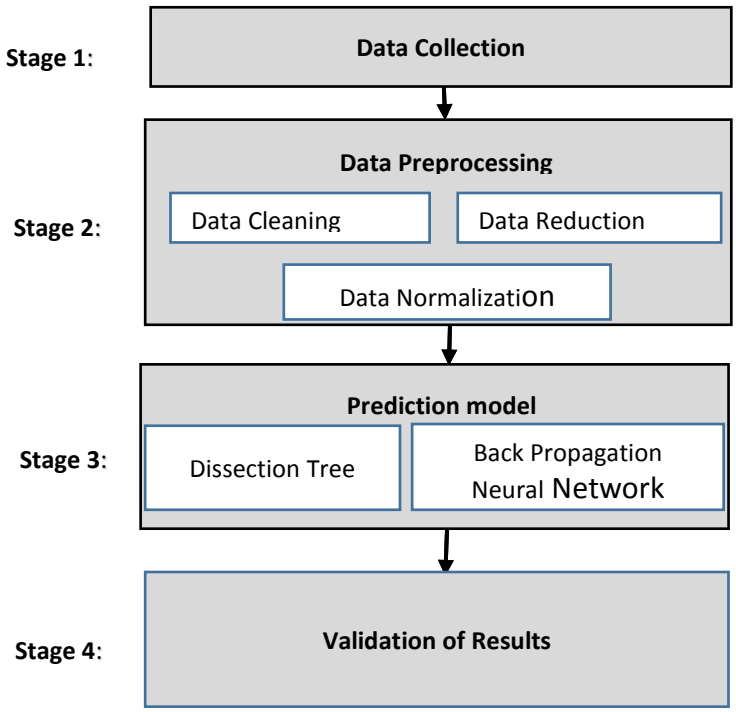

Figure 3. Architectural Design for Proposed System.

METHODOLOGY

According to studies, traffic accidents are affected by many factors. Many features and characteristics were collected during the affect directly or indirectly the occurrence of accidents prediction whether related to the driver, vehicle, or light. The important point is to do a traffic accident prediction in an attempt to obtain deeper characteristics that have a greater impact on accidents than among a large number of data obtained during accidents

The program used was PCA and IGR technology to extract the most important factors that affected the time of the accident using Neural networks and decision trees. Then they were used to create a model for traffic accident prediction comprising of the following:

Extraction Data

The process of extracting data from a dataset which contains a large amount of data is issued to discover hidden relationships and patterns between those data.

The Weka program was used to analyze the data collection of road accidents that occurred in the province of Ibb, Yemen for during 2011-2016. Then the process of designing and building the model us one of the programming languages C # or MATLAB.

Data Collection

Traffic accidents data were obtained from the General Directorate of Traffic accident in Ibb province for during 2015-2016 1530 traffic accidents were surveyed in the duration noted.

Data Preprocessing

After data get obtained in the form of an Excel spreadsheet and prior to the data mining process, the data were firstly checked for the exclusion of disturbing data which negatively might affect the quality of the results. Extraction data was the process of analyzing a large amount of data to extract and discover the hidden patterns in those data which were used in the prediction process. In this study, WEKA program used the IGR algorithm and PCA to discover the most important factors that affected traffic accidents. Then, it built ANN and decision tree to predict traffic accidents. The steps are as follows:

Data Cleaning

After obtaining the required data, the data processing started by deleting the excess columns, metadata, and missing and confusing data then identifying the most important features required and eliminating the duplication of data and values lost, extreme and distorted. The proposed system is planned to be carried out in the following manner. There were several ways to clean the data either by ignoring rows containing missing values or filling the data with duality. This gives more complexity, the more missing data or the use of a unified constant instead of the missing values or the use of one of the central tendency measures instead of the missing values, It measures the central tendency of the data category to which the missing values, belong. This is the best method which provides more accurate results by classifying data into different categories (Al-Turaiki et al., 2016).

Data Reduction

It is a process of reducing the size or representation of the dataset. So that, it results in the same analytical result but in a smaller size by removing the irrelevant attributes or creating a derivative attribute of more features and replacing the data using the models as regression models, linear or non-linear model as graphs and sampling and data collection, In this

paper, the excess features that do not need to reduce the data size to increase the efficiency of the model will be eliminated.

Data Normalization

It is the process of converting selected data into an appropriate format for the algorithms and applications to be used in the prediction. Some algorithms may require data to be present in a particular format before it is applied (ARFF, CSV).

Prediction Model

Once the database is ready for data exploration, the IGR algorithm, PCA is used to analyze data and extract the most important factors that affect the time of the accident using the IGR algorithm and PCA. Then, it started the process of designing and building prediction Back-propagation using neural networks and decision tree and detecting the best techniques in the process of prediction after comparing processes in the results.

Validation of Results

It is a process of the reasonably representative data representation of the model where the results were evaluated and the accuracy of the prediction was done, either for the training or testing of data to verify the validity of the results.

The classification algorithm was applied to the data set that was divided into a training group and a test group to obtain satisfactory results to find predictive results that would assist and contribute to the reduction of traffic accidents after appropriate evaluation and discussions.

CONCLUSION

In this study, a survey of the latest work in the field of traffic accident studies in regard to the analysis and the seriousness of predicting the traffic accidents used data extraction technique and applied them to the data collected at the time of the accident in the province of Ibb, Yemen. The phenomenon of irrigated accidents was constantly increasing due to several factors with regard to the circumstances in which the incidents occurred As a result of a failure to follow the general rules of passage, including to the political situation of the capital, the war affected the infrastructure of the main lines and sub lines since four years. Many researchers have tried to find out any serious radical solutions to this phenomenon, but there was still a lack of finding the right solutions due to the lack of knowledge of all factors affecting Traffic Accidents. In order to bridge this gap, this survey aimed to determine which algorithms and tools were better and more suitable for the process of prediction. It reviewed most recent studies and related models that might help to reduce the incidence of accidents in the future.

ACKNOWLEDGMENTS

We would like to thank the University Malaysia Pahang, which provided a platform for our research.

REFERENCES

Al-Maqaleh, B., A. Al-Mansoub, A., & N. Al-Badani, F. (2016). Forecasting using Artificial Neural Network and Statistics Models. *6*, 20-32. doi:10.5815/ijeme.2016.03.03

Al-Turaiki, I., Aloumi, M., Aloumi, N., & Alghamdi, K. (2016). *Modeling traffic accidents in Saudi Arabia using classification techniques.* Paper presented at the Information Technology (Big Data Analysis)(KACSTIT), Saudi International Conference on.

Ali, G. A., & Bakheit, C. S. (2011). Comparative analysis and prediction of traffic accidents in Sudan using artificial neural networks and statistical methods. *SATC 2011*.

Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity Prediction of Traffic Accident Using an Artificial Neural Network. *Journal of Forecasting, 36*(1), 100-108. doi:10.1002/for.2425

Amardeep, R. (2017). The MATLAB Data Mining Software. *International Journal of Recent Innovation in Engineering and Research*.

Atnafu, B., & Kaur, G. (2017a). Analysis and Predict the Nature of Road Traffic Accident Using Data Mining Techniques in Maharashtra, India. *Analysis*.

Atnafu, B., & Kaur, G. (2017b). Survey on Analysis and Prediction of Road Traffic Accident Severity Levels using Data Mining Techniques in Maharashtra, India.

Bhinge, A. V. (2015). *A Comparative Study on Data Mining Tools.* California State University, Sacramento.

Contreras, E., Torres-Treviño, L., & Torres, F. (2018). Prediction of Car Accidents Using a Maximum Sensitivity Neural Network *Smart Technology* (pp. 86-95): Springer.

De Luca, M. (2015). A comparison between prediction power of artificial neural networks and multivariate analysis in road safety management. *Transport, 32*(4), 379-385. doi:10.3846/16484142.2014.995702

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health, 4*, 118-131.

Gaber, M., Wahaballa, A. M., Othman, A. M., & Diab, A. (2017). TRAFFIC ACCIDENTS PREDICTION MODEL USING FUZZY LOGIC: ASWAN DESERT ROAD CASE STUDY.

Ghani, A., Raqib, A., Sanik, M. E., Mokhtar, M., & Aida, R. (2011). Comparison of accident prediction model between ANN and MLR models.

Gokgoz, E., & Subasi, A. (2015). Comparison of decision tree algorithms for EMG signal classification using DWT. *Biomedical Signal Processing and Control, 18*, 138-144.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hashmienejad, S. H.-A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International journal of crashworthiness, 22*(4), 425-440.

Jadaan, K. S., Al-Fayyad, M., & Gammoh, H. F. (2014). Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN). *Journal of Traffic and Logistics Engineering, 2*(2), 92-94. doi:10.12720/jtle.2.2.92-94

Janani, G., & Devi, N. R. (2018). Road Traffic Accidents Analysis Using Data Mining Techniques. *JITA-JOURNAL OF INFORMATION TECHNOLOGY AND APLICATIONS, 14*(2).

Kashyap, J., & Singh, C. P. (2016). Mining road traffic accident data to improve safety on road-related factors for classification and prediction of accident severity. *International research journal of engineering and technology, 3*(10), 221-226.

Kukasvadiya, M. S., & Divecha, N. H. (2017). Analysis of Data Using Data Mining tool Orange.

Kumar, S., & Toshniwal, D. (2017). Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India. *European transport research review, 9*(2), 24.

Kumar, S., Toshniwal, D., & Parida, M. (2017). A comparative analysis of heterogeneity in road accident data using data mining techniques. *Evolving Systems, 8*(2), 147-155.

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

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis & Prevention, 111*, 354-363.

Mussone, L., Bassani, M., & Masci, P. (2017). Analysis of factors affecting the severity of crashes in urban road intersections. *Accident Analysis & Prevention, 103*, 112-122.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology, 8*(1), 13-19.

Odhiambo, J. N., Wanjoya, A. K., & Waititu, A. G. (2015). Modeling Road Traffic Accident Injuries in Nairobi County: Model Comparison Approach. *American Journal of Theoretical and Applied Statistics, 4*(3), 178-184.

Olutayo, V., & Eludire, A. (2014). Traffic accident analysis using decision trees and neural networks. *International Journal of Information Technology and Computer Science, 2*, 22-28.

Organization, W. H. (2015). *Global status report on road safety 2015*: World Health Organization.

Perone, C. S. (2015). Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil. *arXiv preprint arXiv:1502.00245*.

Prabakaran, S., & Mitra, S. (2018). *Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning.* Paper presented at the Journal of Physics: Conference Series.

Sikka, S. (2014). Prediction of Road Accidents in Delhi using Back Propagation Neural Network Model. *International Journal of Computer Science & Engineering Technology (IJCSET), 5*(08).

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85-106.

Tiwari, P., Kumar, S., & Kalitin, D. (2017). *Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques.* Paper presented at the International Conference on Computational Intelligence, Communications, and Business Analytics.

Wenqi, L., Dongyu, L., & Menghua, Y. (2017). *A model of traffic accident prediction based on convolutional neural network.* Paper presented at the Intelligent Transportation Engineering (ICITE), 2017 2nd IEEE International Conference on.

Yu, B., Wang, Y., Yao, J., & Wang, J. (2016). A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. *Neural Network World, 26*(3), 271.

Zhang, X.-F., & Fan, L. (2013). *A decision tree approach for traffic accident analysis of Saskatchewan highways.* Paper presented at the Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on.

Zong, F., Xu, H., & Zhang, H. (2013). Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models. *Mathematical Problems in Engineering, 2013*, 1-9. doi:10.1155/2013/475194

Žunić, E., Djedović, A., & Đonko, D. (2017). *Cluster-based analysis and time-series prediction model for reducing the number of traffic accidents.* Paper presented at the ELMAR, 2017 International Symposium.