Send Orders for Reprints to reprints@benthamscience.net

Open Access

Analysis and Monitoring of the Traffic Suburban Road Accidents Using Data Mining Techniques; A Case Study of Isfahan Province in Iran

Mehdi Mansouri[1] and Mohammad Javad Kargar[*,2]

[1]*Department of Computer Engineering, Najafabad Branch , Islamic Azad University ,Isfahan, Iran*

[2]*Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran*

Abstract: Driving accidents have been always counted as one of the most ostensible causes of deaths in the societies today. Statistics and reports indicate that the road accidents in Iran rank several times more than the ones in the developed countries. In the current paper, the rules and factors influencing the traffic road accidents of Iran have been extracted along with extracting a local data model after collecting the data from a variety of sources followed by data aggregation and combination, data cleaning, and separating the inappropriate data. This was done by employing appropriate data mining methods, such as clustering and decision tree. The utilized data was based on 10000 accidents during 2011 to 2013 in Isfahan Province, Iran.

The experimental results have revealed that of the Decision Tree approaches, C5.0 algorithm outperforms the other algorithms with a lower error rate and a higher accuracy rate. Our research analysis also shows that in determining the accident type, three most important attributes include the type of the faulty vehicle, type of the vehicle hit, and the accident reason. The results and findings obtained in this study are significant and interesting which can provide the authorities with invaluable information on reducing the road accidents.

Keywords: Clustering, data mining, decision tree, traffic road accidents, transport.

1. INTRODUCTION

Road accidents are well known as the chief causes of deaths in the world, particularly among the individuals aging below 50. This causes an annual deaths rate of over 1 million people and injuring more than 50 million people around the globe [1]. In line with this, the road deaths rate in Iran is twentyfold the developed countries and fivefold the countries similar to Iran. Because of the traffic accidents, very weighty physical and economical losses are imposed on the society yearly so that every hour more than two people are killed while 87 are being injured. Isfahan Province is one major and central province of Iran, neighboring nine other provinces and has a noticeable share of the nationwide roads (20% freeway, 14% highway, 11% main way, 3% By-way) making it to have a particular status in terms of traffic.

There are several influential factors involved in accidents, among which the human factors, road factors, vehicle factors, and environmental factors. Due to the fact that this issue is of remarkable significance, there is an urge in doing necessary investigations, analyses, and interpretations in order to reduce such accidents. By employing the data mining knowledge along with the available data of the past traffic accidents, we will be able to obtain inimitable findings which can be used for diagnosing, predicting, and probability of their incidences in future. To fulfill such an aim, the first and the most fundamental step would be to obtain the local data set of Iran as well as the practical models compliant with the current road and cultural conditions. The main objectives in this study were to identify, analyze and interpret the important factors affecting the intensity of road accidents in Iran in addition to determining the most appropriate data mining techniques which yield the best results for modeling the accidents dataset.

The data sites have been come into existence over the last decades about various fields while the volume of such information sources is constantly increasing so that analyzing such data through a traditional method without using the modern technology and automated equipment will be very arduous and impossible. In fact, the mentioned information sites are in thirst for discovering the knowledge.

While all decision makings lend their basis to the past information, a more appropriate decision with fewer flaws will be made if this information is more precise and knowledge-based. Data mining commonly referred to as the "Knowledge Discovery in Database" is a powerful tool for data analysis and signposts extracting the unidentified and expedient information from the raw data set available in the data sites. Data mining is formed through a combination of various disciplines such as artificial intelligence, data site management, machine learning, mathematical algorithms, and statistics. Data mining was firstly identified in 1996 as the knowledge discovery from the data site, which was later developed rapidly [2].

*Address correspondence to this author at the Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran;
Tel: 00983527770952; Fax: 00983527770954; E-mail: Kargar@usc.ac.ir

2. LITERATURE REVIEW

This section elaborates on some of the investigations related to the data mining in road accidents in different countries. Nonetheless, because of having a problem in accessing the accidents data in Iran, the conducted surveys have dealt with the data of other countries.

Lord and Mannering [3] provided a detailed review of the key issues associated with crash-frequency data as well as the strengths and weaknesses of the several methodological approaches that researchers have used to address these problems.

Prato *et al.* employed [4] Descriptive (i.e., K-means and Kohonen clustering) and predictive (i.e., decision trees, neural networks and association rules) data mining techniques for the analysis of traffic accidents occurred in Israel between 2001 and 2004.

In 2010, Xi Jianfeng and his associates [5] conducted a research on the cause of accidents on the basis of traffic accidents information system in China. A noticeable percentage of the world accidents belongs to China because it has a huge population. The accident analysis based on 20000 road traffic accidents data caused in one region has indicated the relative significance degree of the road among four factors of individuals, vehicles, road, and environment is equal to 1. Moreover, in the road factor, three components have the highest degree of relative significance including the road guards beside the roads, physical separation of the roads, and the road surface, respectively.

However, Regorio Gecchele and *et al.* [6] established a comparison on various methods of data mining clustering for road classification in one analysis in 2011 with the purpose of estimating the average daily traffic per year. These analyses were fulfilled using the accessible data from 54 sites which automatically recorded the traffic in Italy as well as the factors related to the passenger and vehicle in the classification process. They concluded that in comparison with the other experimented methods, the model-based clustering methods yield superior results, specifying an important classification. While analyzing the type of day, Saturdays and Sundays revealed the highest amount of errors in the weekdays. Moreover, considering the year period, summer indicated more error compared to winter. In terms of vehicle, errors were more for heavy vehicles than the transport vehicles.

In order to investigate the phenomenon of road traffic accidents number, Rui and *et al.* [7] introduced a method for analyzing the road traffic accident based on data mining. In their research, they primarily analyzed the specifications and factors related to the rad traffic accidents and then introduced two data mining theories, namely the Rough Sets Theory and Theory of Association Rules. Finally, they presented a method for analyzing the reasons of the road traffic accidents based on data mining. The results of this research revealed that the study of reasons for road traffic accidents can help to quickly and efficiently identify the key factors and provide instructional methods for evading and reducing the road traffic accidents which can prominently lessen the individual death and financial losses caused by the mentioned accidents.

The classification and regression tree (CART), one of the most widely applied data mining techniques, has been commonly employed in business administration, industry, and engineering [8]. Tibebe Beshah in company with some other researchers [9] analyzed the road accidents data which had been collected from the Traffic Department of Addis Ababa (the capital city of Ethiopia) from various dimensions by using CART and RandomForest method with the aim of identifying the relevant pattern and presenting the techniques performance for the road safety field. Among the research objective was to explore and predict the drivers' roles in possible injury risk. Their experimental results demonstrated that this model could classify the accidents with an acceptable precision. Also, some of the studies combine the classification approaches such as Ruimin *et al.* [10] which In their study, Classification and Regression Tree (CART), CHAID and Exhaustive CHAID were employed to model the incident duration.

In addition, a research was conducted by Benoit Depaire and two colleagues [11] to present the effects of a clustering technique (i.e. latent clustering) for identifying different types of traffic accidents. They firstly converted a heterogeneous traffic accidents data set into seven clusters and then into seven accident types. Secondly, injury analysis is performed for each cluster. In the recent study, De one *et al.* [12], Latent Class Cluster (LCC) was used as a preliminary tool for segmentation of 3229 accidents on rural highways in Granada (Spain) between 2005 and 2008.

Finally, in another survey, So Yung Soh and Song Ho Lee attempted to combine the classification and clustering methods to enhance the accuracy of the classification of the road accidents intensity in Korea [13]. Nevertheless, Evrim Bayam and two other researchers investigated the effect of drivers' age on the accidents in America by using the decision tree method [14]. In 2014, two researchers called Eludire and Olutayo used the decision and artificial neural network with the intention of discovering new knowledge out of the data related to road accidents in one of the most crowded roads of Nigeria. Their focal aim was to determine the reason behind fatal accidents in the studied highway. Their data were organized into two categories, namely continuous and classified. They made use of decision tree for analyzing and interpreting the classified data whereas for the continuous data they used the artificial neural network. The results observed in their study proved that the decision tree method outperformed the artificial neural network with lower error rate and a higher precision. Their study indicated that the three noticeable factors for the accidents included the tyre burst, loss of control, and overspeeding [15].

In 2013, Xue-Fei Zhang conducted a research in which the traffic collision data was used that had been collected over the last 20 years on the rural highways and urban streets from Saskatchewan, Canada. In order to determine the major factors contributing to traffic collisions and their severity, they presented a data mining model using ID3 and C4.5 decision tree algorithms to analyze the traffic collision data. The experiment results from their study demonstrated that the developed data mining model using decision tree could effectively classify the major contributing factors to traffic collisions and their collision severity for different groups of people with good accuracy. The data mining model was evaluated and compared with a commercial software package Weka. Further recommendations drawn from their study results for traffic safety improvements were also detailed in their paper [16].

3. METHODOLOGY

Because the objective in the current study is to discover the hidden patterns and present a behavioral model of traffic data for the roads of Isfahan Province, a constructive data mining methodology is required to be used. In this section, the steps from the beginning of data mining operations to the end have been succinctly elucidated.

3.1. The Data Set

One of the obstacles which the researchers face when doing data mining projects and surveys using the local data in Iran is to access the data and information sites. In most cases, organizations and associations that keep hold of such data avoid providing them for diverse reasons. One of the reasons is that the data is confidential and private. In case the referred organization is Police and Army Force, this problem becomes more serious. When they face such problems, the researchers will try to use the data available from the foreign countries which are easily accessible while they are also refined data. In this way, the researcher's need for one of the most time taking stages of data mining i.e. data preparation will be eliminated.

The data needed for this research is related to the accident field while the main center for collecting such data is the information of the Police Department of Isfahan. For obtaining the needed data, a long way was taken and a huge time was spent. The studied case of this research is Isfahan Province and the data collected is hence for this province. The dataset used in the current research is annual and related to years 2011 to 2013.

Prior to analyzing the data, the data is needed to be preprocessed. Preprocessing the data implies preparing the data to run the main process which is knowledge discovery. In this stage, we investigate the data in terms of errors such as missing values, noise, repetitive data, and wrong data. It needs to be asserted that the dataset of this study contained many of such cases so that a long time was spent on preprocessing it and finally an appropriate dataset was obtained. Other tasks to be done during the preprocess stage includes data unity, data selection, and data conversion. Table **1** presents a description of the dataset selected from the other data.

3.2. Preprocessing and Data Preparation

Before knowledge discovery, the data should be prepared for data mining. To do so, data cleaning, data integration, data selection, and data conversion were performed. What follows is a brief description of the operations done as explained above:

- Data cleaning is one the steps in the preprocessing stage which can be a time taking process due to the data volume, data type, and the rate of the data error. Data cleaning encompasses the diagnosis, edit, and omission of the existing errors in the data. The data errors include the missing values, noise and repetitive data, wrong (incomplete) data, and contradictory data (or the data with inappropriate structure) [17]. Due to the fact that the data used in this research is extracted from the information in the Police officers and experts' reports which have been inserted into the Excel manually in Persian Language, we observed many human errors while entering the information. For this reason, a long time was spent on data cleaning, accordingly. These errors briefly include frequent misspells, space between two words or lack of necessary spaces, having hyphens, and synonymous words or sentences with different writings which caused separating such synonymous words or sentences.
- The next step in the preprocessing process is the data integration. The data mining projects mostly require combining the data from the multiple data sources. It is possible for data analysis to undergo data unity since the data typically originate from various sources. Some issues might be faced while integrating the data. One if such issues is adaptation of the homogenous objects. For example, how can we identify the *Customer-id* variable in one data site is the *Customer-number* variable in another data site? Consequently, adaptation of the homogenous objects is one of the actions fulfilled during the data integration [5]. In this stage, the data investigated in the current research was combined and integrated. To fulfill this, the data which is in 5 categories was combined in *Clementine Software* by performing the *Append* process, turning into an integrated dataset.
- Data selection was the next step in which the data related to the analysis of the data source will be restored. The data set entails various sorts of data but not all of them are required for data mining. In fact, the required data for data mining is required to be selected [18]. The data used in this research encompasses some records which are not influential in its result and were not used and selected in this research; for example, some of these variables contain personal information of the accidents wounded and dead individuals, as well as the vehicles plate number while using such information will violate the privacy. Moreover, in some records there were many missing variables such as the vehicles plate number, the faulty vehicle, the driver's education or his/her type of driving license.
- The last step of the preprocessing stage is converting the data in which the data is converted into appropriate forms for extracting the information. What follows is a list of activities done on the data set used in the current study:
  - ○ **Summarization** was done in this stage. For instance, we collected the data related to those who died on the spot, the ones died while being transferred, and the ones who died in the hospital and summarized them into the "Death Number Variable".

Table 1. A description of the dataset.

No.	Variable	Description
1	Name of the Police station	Name of the related Police station
2	Year	2011-2013
3	Months	The entire Months of a year
4	Day in Month	1-31 of Months
5	Weekday	Monday through Sunday
6	Season	Spring, Summer, Fall, Winter
7	Time	12 am-3 am, 3 am-6 am, 6 am-9 am, 9 am to 12 pm, …
8	Name of the Road	Name of the road where the accident occurred
9	Number of the wounded individuals	Number of the accidents' wounded individuals
10	Type of the accident	Casualty/fatal/car crash
11	Number of the dead individuals	Number of the accident dead individuals
12	Type of the car hit	Type of the hit of the vehicles involved in the accident
13	The road status	Type and status of the road where the accident occurred
14	The weather condition	Weather conditions of the accident day
15	The Accident reason	The reason diagnosed by the accident police expert
16	Type of the faulty vehicle	Type of the faulty vehicle
17	System of the faulty vehicle	System of the faulty vehicle
18	Airbag	Having the airbag
19	ABS Brake	Having ABS brake
20	Gender	Male/Female
21	Age	The driver's age

- **Generalization of the data** was also performed in which some data was replaced with other data; for example, the accident time was converted into the following intervals: 0-3, 3-6, 6-9, 9-12, 12-15, 15-18, 18-21, 21-24.
- **Generating the variables** was fulfilled wherein new variables were generated by exploiting some of the variables of the dataset in order to facilitate the extraction process. For example, we generated the variables "season", "month", "weekday" by making use of the variable of accident date.

4. MODELING AND FINDINGS

Following the summarization and data cleaning, data description was performed as a part of preprocessing of data mining. In this stage, useful information was obtained while only the most interesting information is visualized and provided in form of Figs. (**1-7**).

As it can be discerned in Fig. (**1**), accidents have mostly occurred between 3 and 6 pm, indicating that the drivers undergo the accidents during the first hours of the afternoon, particularly after lunch.

Based on Fig. (**2**), accidents have mostly occurred during September. This figure illustrates that contrary to the public, the lowest accident rates belong to April and May.

Fig. (**3**) clearly exhibits that in comparison with other seasons, summer ranks the first in terms of accident occurrence whereas spring shows the lowest rate, notwithstanding a high rate of New Year (Nowrooz) journeys in spring.

As it is clearly observed in Fig. (**4**), the Police Station of Isfahan-Shiraz Road has reported the highest rate of accidents occurred on the roads of Isfahan Province.

Fig. (**5**) clearly indicates that the most important reason for accidents is lack of attention to the front, followed by lack of yielding, and lack of control on the vehicle because of exceeding the Normal speed.

Male drivers caused more than 93% of the accidents as demonstrated by Fig. (**6**) but this high rate might be due the fact that the majority of the suburban motorists are male in Iran.

A more detailed sampled chart is provided by Fig. (**7**) showing the accident rate in Isfahan Province roads during various times. This figure clearly shows the accident difference in different roads. By comparing Figs. (**1-7**), it can

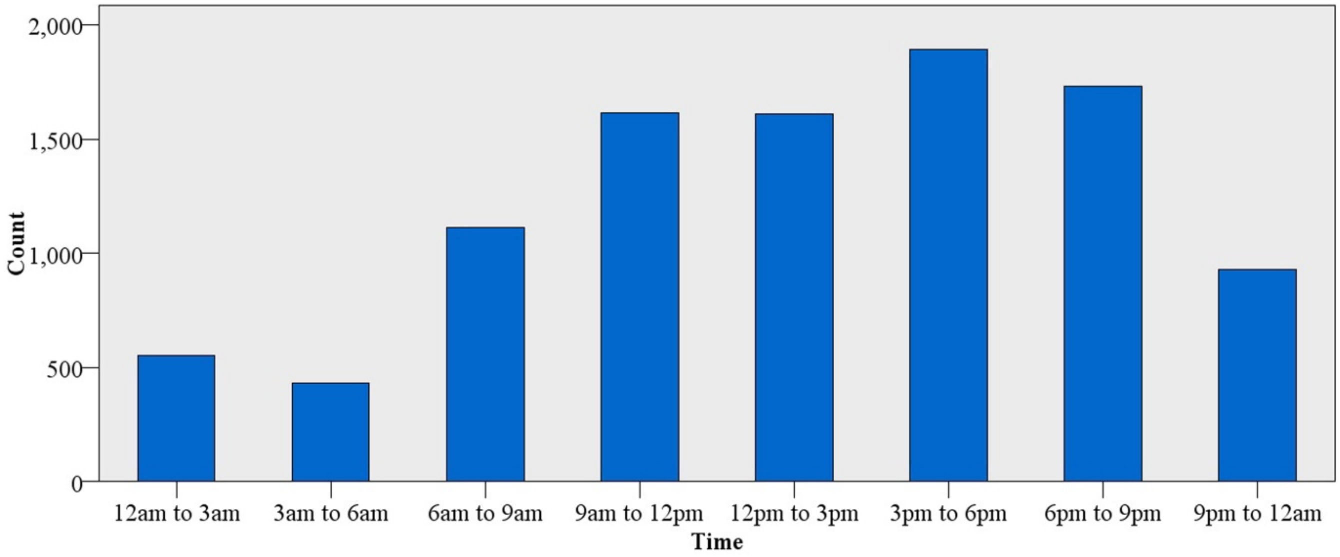


Fig. (1). Accident distribution Time different time interval.

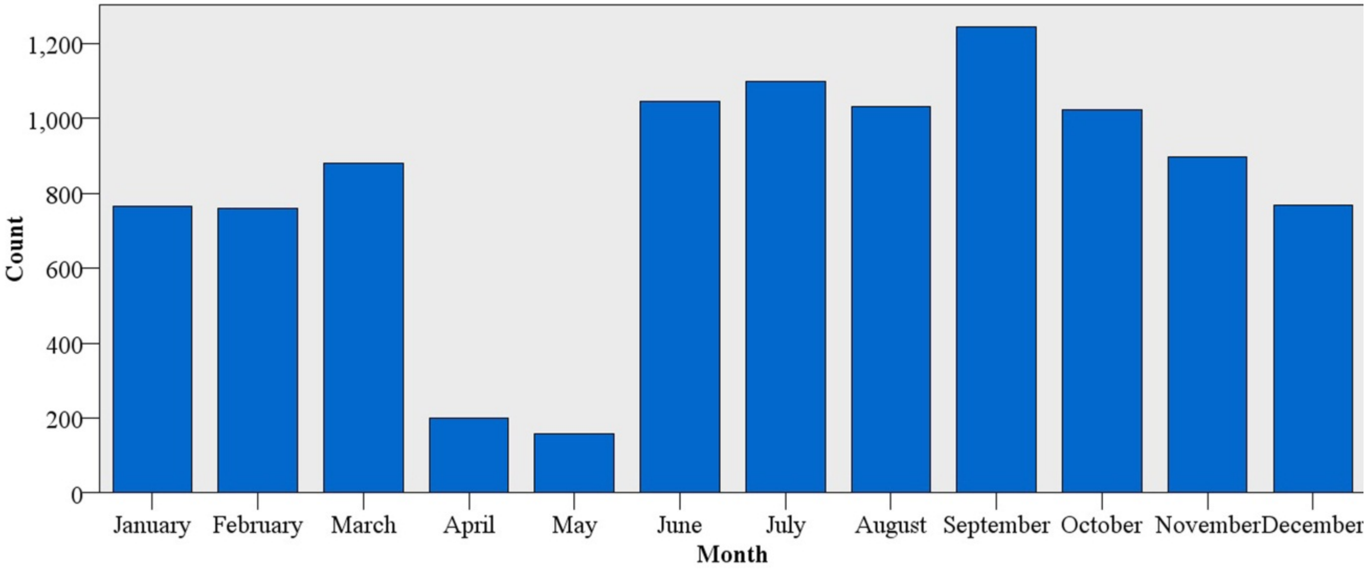


Fig. (2). Accident distribution in different months.

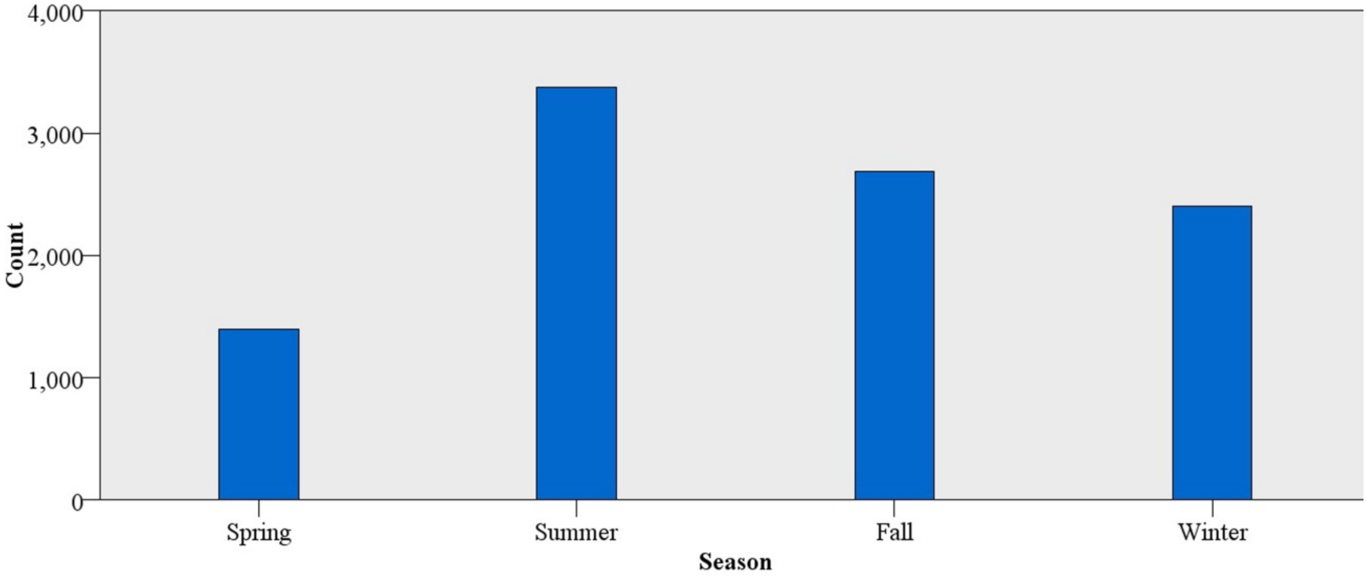


Fig. (3). Accident distribution in different seasons.

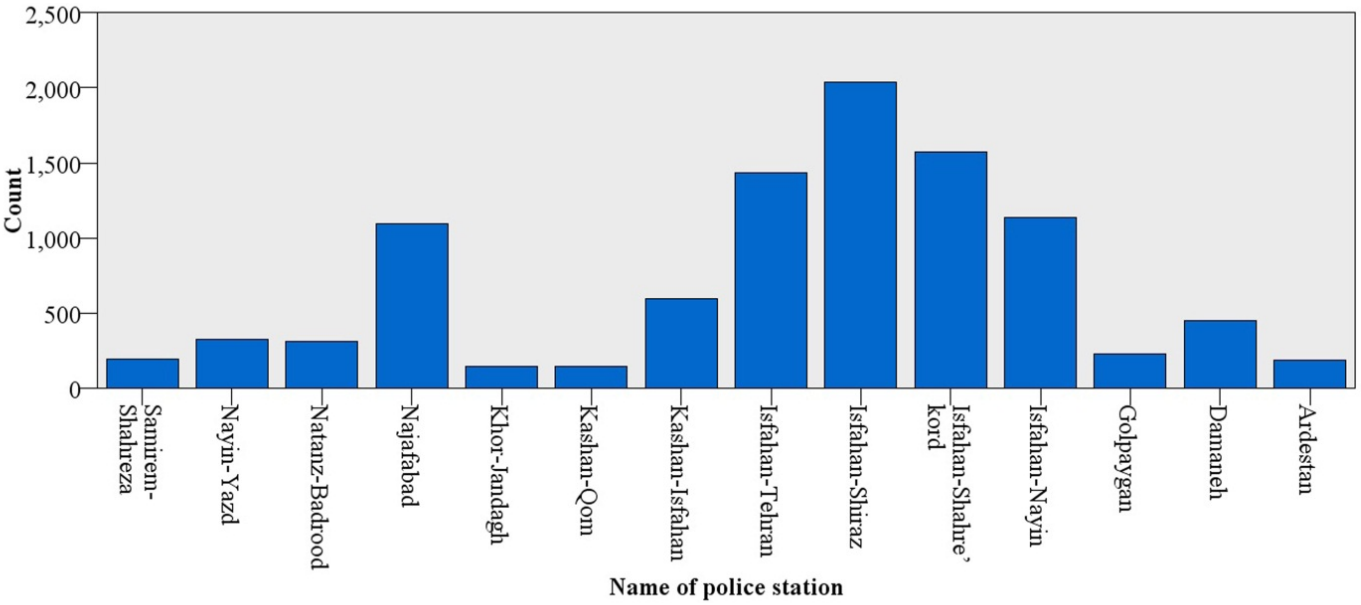


Fig. (4). The accident rate in the vicinity of police stations.

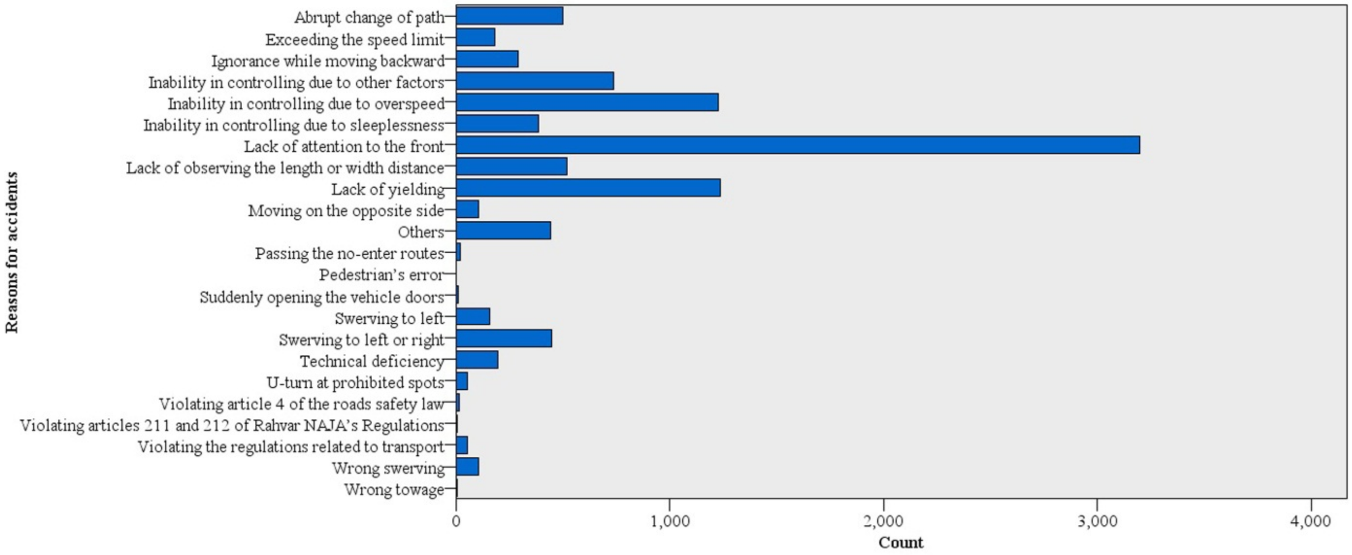


Fig. (5). The reasons for accidents.

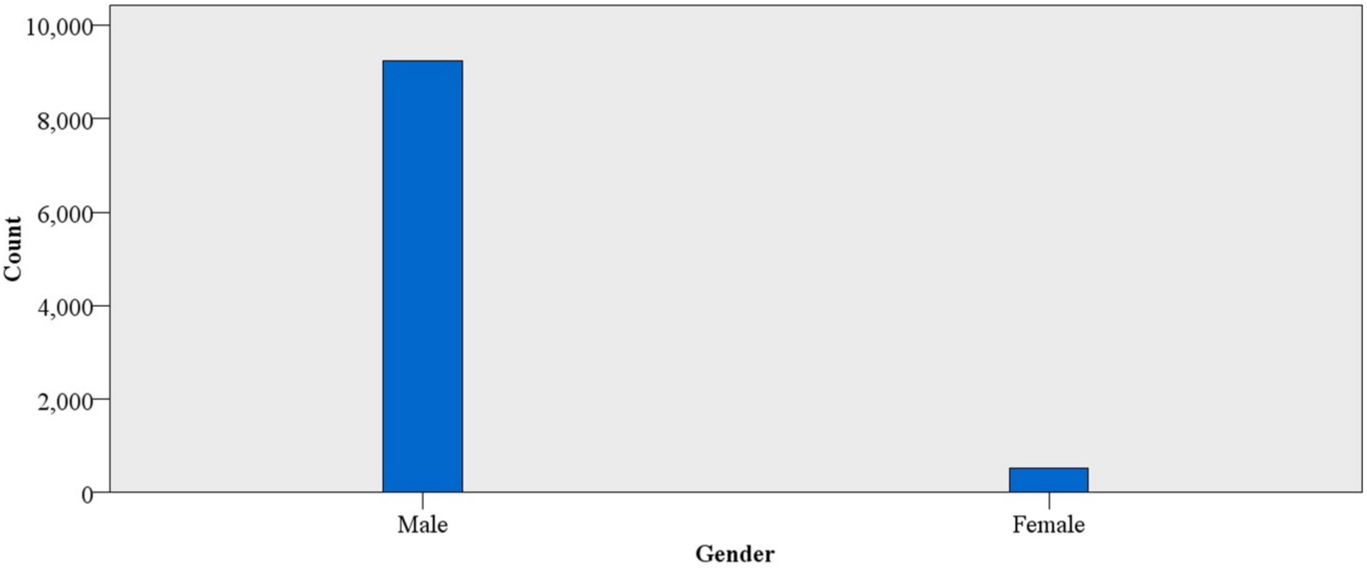


Fig. (6). The drivers' gender.

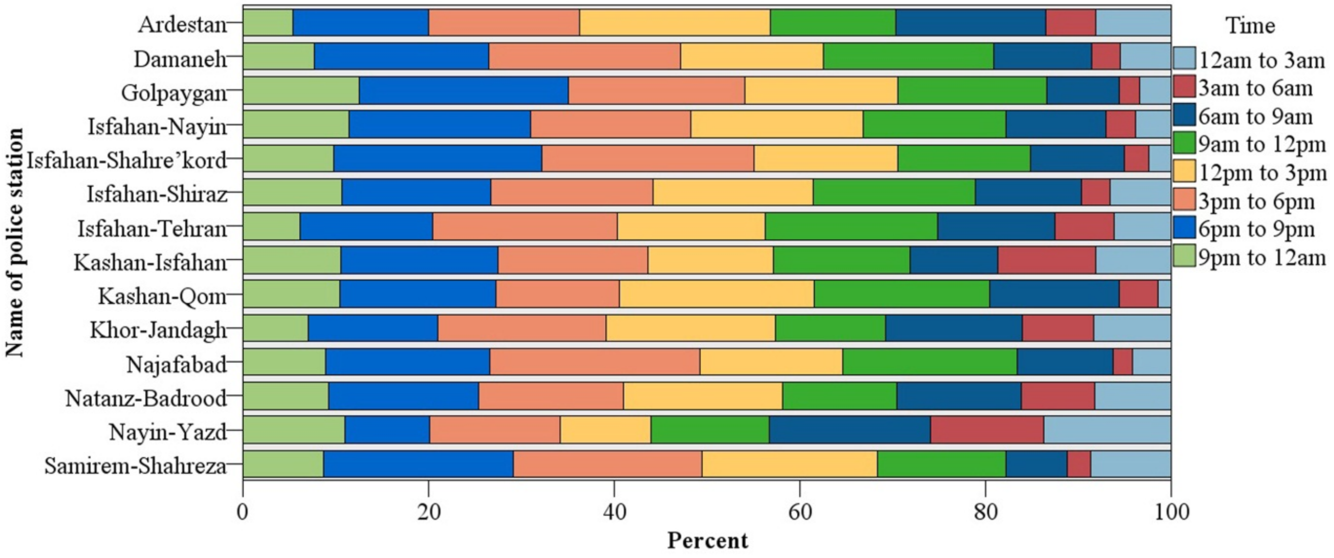


Fig. (7). Color chart of the Police Stations distinguished by the accident time.

be concluded that the accidents have mostly occurred between 3 and 6 pm as reported by Fig. (**1**); however, this fact is not confirmed in some Police stations. For example, the accident rate in Isfahan-Nayin Police Stations is reported to be significantly different compared to the other Stations during 12 midnight and 3 AM. Each piece of information obtained here has its own relevant explanation. For instance, having such a high accident rate between 12 midnight and 3 AM is probably because this road is a part of the transit road of Tehran-Bandar Abbas.

In the Table **2** and the Fig. (**8**), the number of fatal accidents, the number of the dead individuals, as well as the average of fatality in each accident has been presented for each police station separately. It is clearly discerned that in the regions wherein there are desert roads such as Kashan, Qom, and Khor-Jandagh the average of fatal accidents is higher in each accident.

5. MODELING

In this research, the organized data as well as rules for each accident were extracted by selecting the "type of accident" variable as the target variable and by using the decision trees. The Decision Trees models are appropriate for this aim because of the type of the target variable. Other merits of these models include high understandability and extracting rules from them. For this reason, all the data were divided into the training and Testing with a portion of 70% to 30%, following the preparation. By utilizing four different methods on the data of the Training part, modeling of the decision tree was accomplished and then this data was tested using the Testing data which had not been used in the modeling, yielding the subsequent results as tabulated in Table **3**.

In this comparison, the number of the Training data is 6881 (70% of the whole data) while the number of Testing data is 2988 (30% of the whole data). The following sections elaborate on the results obtained through the four decision tree methods:

The First Method: Classification & Regression Trees

This algorithm is a ranking and predicting model which is based on decision tree. This model resembles the C5.0 tree in that it uses recursive segmentation for analyzing the educational records into parts with equal output characteristics. This algorithm divides the educational records into two subcategories in each stage.

Results: 45.53% of the training data is confirmed in this model while this model shows 43.98% accuracy on the testing data.

Table 2. Details of fatal accidents.

	Ardestan	Isfahan-Tehran	Isfahan-Shahrekord	Isfahan-Shiraz	Isfahan-Naein	Khor-Jandagh	Samirem-Shahreza	Kashan-Isfahan	Kashan-Qom	Golpaygan	Nayin-Yazd	Najafabad	Damaneh	Natanz-badrood
Fatal Accidents	43	120	97	161	104	28	14	58	29	32	39	89	56	37
Total number of Dead individuals	50	131	106	187	121	39	16	76	41	38	50	110	67	42
Average	1.19	1.1	1.1	1.16	1.17	1.44	1.23	1.33	1.46	1.22	1.31	1.25	1.21	1.16

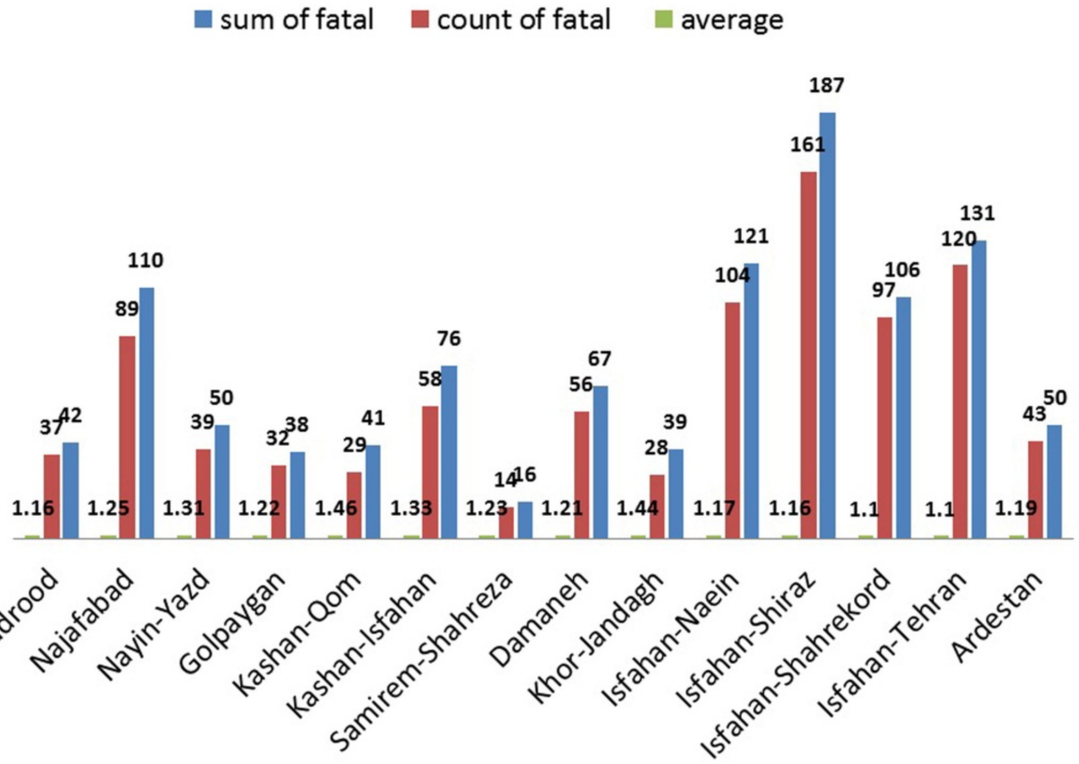


Fig. (8). Details of fatal accidents.

Table 3. Comparing the results in the decision tree.

Partition	C&R Tree		C5 Tree		CHAID Tree		QUEST Tree	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Correct	45.53%	43.98%	69.09%	70.18%	63.32%	61.65%	46.58%	46.72%

The Second Method: The C 5.0 Tree

The C5.0 algorithm is a type of single-variable decision tree which is the enhanced form of C4.5. This algorithm is able to be applied in the form of decision tree of a set of rules. It needs to be highlighted that in majority of applied programs there is a preference over the set of rules due to the fact that perceiving the set of rules is easier than that of decision tree.

Results: 69.9% of the training data is confirmed in this model while in 70.18% of the testing data, the accident type can be accurately diagnosed by this model.

The Third Method: The CHAID Tree

This algorithm is one of the oldest decision tree algorithms. In order to identify the enhanced divisions, the division tree is produced and Chi Square statistical methods are employed.

Results**:** 63.32% accuracy on the training data and 61.65% accuracy on the testing data are confirmed.

The Fourth Method: The Quest Tree

This is a binary classification method for making decision tree that has been designed for reducing the processing time needed for analyzing the large C&RT trees. One of the objectives of this algorithm is to lower the tendency of the tree classification methods as regards the predictors which produce many branches. In this method, the predictive characteristics can be categorical and numerical but the target characteristic must be categorical. The entire branches in this algorithm are binary.

Results**:** this model shows 46.58% accuracy on the training data and 46.72% accuracy on the testing data.

It needs to be asserted that in all the models, we can reach 100% accuracy on the training data which is achievable by expanding the tree's branches but we will face over fitting which means very weak results of such models on the testing data. To eradicate such a problem, all the models were restricted by pruning methods so that the accuracy will be maintained in addition to having appropriate generality as well as generalizability and use with different conditions.

Contrary to other methods, the C5.0 tree finally obtained better results on both the training data and testing data with high accuracy and appropriate generality and the final rules were extracted from this model. By implementing the algorithm of the C5.0 tree on the dataset and considering the "type of accident" variable as the target variable, the rules governing the accidents of Isfahan Province for years 2011-

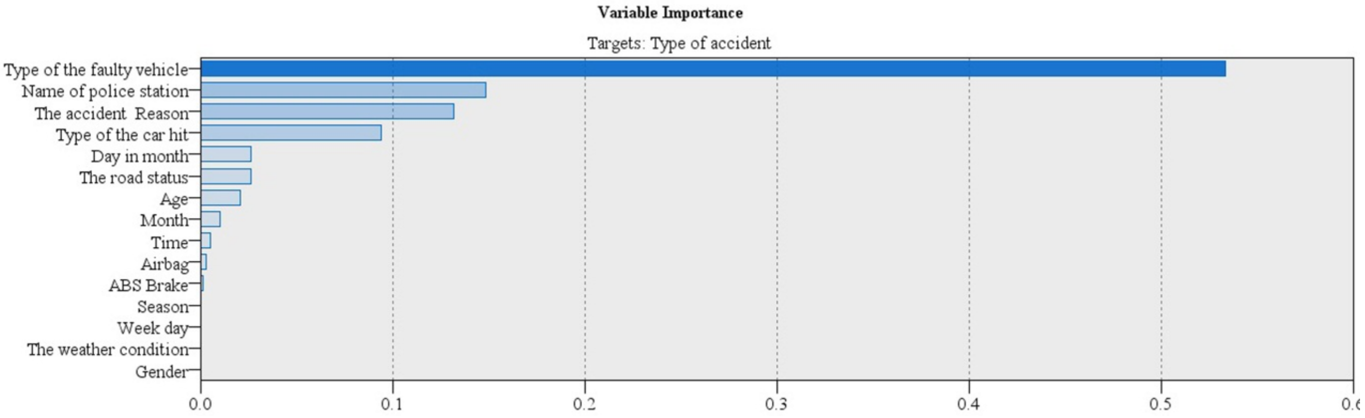


Fig. (9). The importance of characteristics in C5.0 model.

2013 were extracted and were placed separately in three categories of fatal, casualty, and car crash. These rules cannot be generalized for all the accidents but there is a probability for their occurrences in these conditions considering the past occurrences.

Important characteristics in predicting the accident type were discovered by using the C5.0 decision tree model as displayed in Fig. (**9**).

The importance of the input characteristics is classified on the basis of the results of the characteristic selection. As seen in the given table, the characteristics being known as important stayed at a higher level than the others. It is observed that type of the faulty vehicle in the intensity of accident type (Casualty/fatal/car crash) exerted the largest impact.

Rules for Fatal Accidents

- **If** the *reason for accident* **is** ["an abrupt change of path"] **and** the *type of the faulty vehicle* is ["car"] **and** the *weather is* ["foggy "], **then** <u>the accidents will be fatal.</u>
- **If** the *reason for accident* **is** ["Weak driving" or " unallowable overtaking "], **then** <u>the accidents will be fatal.</u>
- **If** the *reason for accident* **is** ["swerving left or right"] **and** *type of the car hit* is ["front to left side "] **and** the *type of the faulty vehicle* is ["bus"] **then** <u>the accidents will be fatal.</u>
- **If** the *reason for accident* **is** ["abrupt change of path"] **and** *type of the faulty vehicle* is ["motorcycle"] **and** *time* **is** ["12am to 3am" "3am to 6am " "6am to 9am" "9am to 12pm" "12pm to 15pm"] and *road status* **is** ["single carriageway"] **then** <u>the accidents will be fatal</u>

Rules for Injury Accidents

- **If** the *reason for accident* **is** ["exceeding the speed limit"] **and** the *weather* is ["Dusty "or "cloudy "] **then** injury <u>accidents happen</u>
- **If** the *reason for accident* **is** ["abrupt change of path"] **and** the *type of the faulty vehicle* is ["car"] **and** the *weather* **is** ["clear "] and the *police station zone* **is** ["isfahan- tehran"] **and** the *driver is* female, **then** the <u>injury accidents happen.</u>
- **If** the *reason for accident* **is** ["abrupt change of path"] **and** the *type of the faulty vehicle* is ["heavy vehicles"] and the *road status* **is** ["single carriageway"] **and** the *Type of the car hit* is ["front to rear"] **then the** <u>injury accidents happen</u>

Rules for Car Crash Accidents

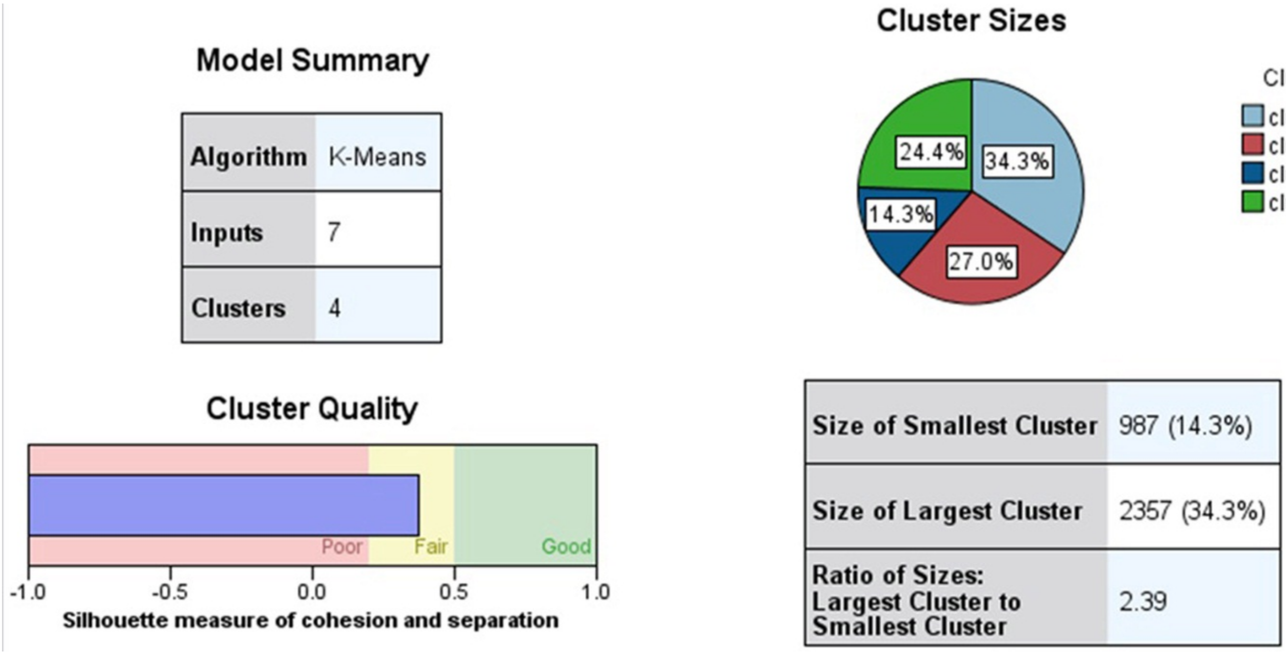
- **If** the *reason for accident* **is** ["exceeding the speed limit "] **and** the *weather* **is** ["rainy "] **then the** <u>car crash accidents happen</u>

By considering the "reason for accident" as the target variable, some other accurate rules were extracted such as:

- **If** the *Police station zone* **is** ["Khur-Jandagh" "Nayin-Yazd" "Isfahan-Tehran"], **then** <u>inability to control the vehicle is due to fatigue and sleeplessness</u>
- **If** the *vehicle systems* **in** ["Peugeot 206" and " Peugeot pars], **then** the <u>inability to control the vehicle is due to over-speed</u>

5.1. Clustering

The K-means algorithm has the best performance while implementing the clustering algorithms on the dataset but the results obtained from clustering are not very favorable. Fig. (**9**) represents graphs which indicate the analysis of implementing the K-means algorithm on the dataset. Moreover, clustering modeling was done with K-means, Kohonen, and TowStep algorithms and the clusters' quality was appraised with the Silhouette Criterion. Each algorithm was investigated with a different number of clusters (from 3 to 12 clusters) and at the end the K-means algorithm with four clusters had the highest criterion so that it was used accordingly. It needs to be highlighted that the Silhouette Criterion is equal to 0.4 for the final model, revealing that the data has a structure; yet, the coverage and distribution of the data of this structure are not clearly indicated. It seems that each road follows its own particular pattern and distinguishing the data in terms of time, vehicle, and the road status is not generalizable to all of the roads. Because one objective in this research was to introduce an appropriate model of data type for the road accidents of Iran, we made



Model Summary

Algorithm	K-Means
Inputs	7
Clusters	4

Size of Smallest Cluster	987 (14.3%)
Size of Largest Cluster	2357 (34.3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.39

Fig. (10). Implementing the K-means algorithm.

use of clustering as well for analysis and interpretation. Nonetheless, it was realized that no significant results were obtained through implementing the clustering. Fig. (**10**) shows the results of K-means clustering.

6. SUGGESTIONS

One of the items to be used in future research can be the region code in the plate number of the vehicles that have caused the accidents along with scrutinizing the transport of non-local vehicles of Isfahan Province, especially during the new year and summer vacations. In case of analyzing and data mining in this regard, the national accident dataset can be utilized in addition to the variables used in this study.

CONCLUSION

The objective in this research was to analyze and monitor the road traffic accidents using the data mining techniques in suburban roads and this research was conducted as a case study investigating the roads of Isfahan Province. The obtained results in this study are interesting and significant which can be considered by authorities as invaluable information to be used for decreasing the road accidents. Furthermore, five algorithms existing in data mining was used in this study for knowledge discovery of the accident dataset of Isfahan Province and it was found out that out of the mentioned five algorithms, the C5.0 decision tree algorithm proved to generate the best results and performance. Later in this research clustering of the data was also performed but did not result in separation of clusters with a specific meaning. Considering the clustering results, it can be concluded that each route follows its own particular pattern and differentiating the data with reference to time, vehicle, and the road status is not generalizable to all of the routes. In determining the accident type as Casualty, fatal, and car crash, the most important characteristic was the type of vehicle. On the other hand, other characteristics such as gender, the weather condition, time and date of the accidents had insignificant effects.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We wholeheartedly appreciate the Social Department of the Police Force of Isfahan Province and Road Police of Isfahan Province and Department of Road and urban planning of Isfahan Province for helping us to collect the data and all the other individuals who assisted us in performing this research.

REFERENCES

[1] Transportation Research Institute, "Iran's Comprehensive Road Safety Studies", Study Report, Road and Transportation Ministry, 2006.

[2] H.-H. Tsai, "Global data mining: An empirical study of current trends, future forecasts and technology diffusions," *Expert Syst. Appl.,* vol. 39, pp. 8172-8181, 2012.

[3] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transport Res. A-Pol,* vol. 44, pp. 291-305, 2010.

[4] C. G. Prato, S. Bekhor, A. Galtzur, D. Mahalel, and J. Prashker, "Exploring the potential of data mining techniques for the analysis of accident patterns," In: *Proceedings of the 12th WCTR Conference,* Technical University of Denmark, 2010.

[5] X. Jianfeng, C. Xiaodong, W. Shuangwei, and Z. Tao, "Accident cause analysis method based on traffic accident information system," In: *International Conference on Computer Application and System Modeling (ICCASM),* 2010, pp. 13-230.

[6] G. Gecchele, R. Rossi, M. Gastaldi, and A. Caprini, "Data Mining Methods for Traffic Monitoring Data Analysis: A case study," *Proc. Soc. Behav. Sci.,* vol. 20, pp. 455-464, 2011.

[7] T. Rui, Y. Zhaosheng, and Z. Maolei, "Method of Road Traffic Accidents Causes Analysis Based on Data Mining," In:

International Conference on Computational Intelligence and Software Engineering (CiSE), 2010, pp. 1-4.

[8] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Anal. Prev.,* vol. 38, pp. 1019-1027, 2006.

[9] T. Beshah, D. Ejigu, A. Abraham, V. Snasel, and P. Kromer, "Pattern recognition and knowledge discovery from road traffic accident data in ethiopia: Implications for improving road safety," In: *World Congress on Information and Communication Technologies (WICT)*, 2011, pp. 1241-1246.

[10] L. Ruimin, Z. Xiaoqiang, Y. Xinxin, L. Junwei, C. Nan, and Z. Jie, "Incident duration model on urban freeways using three different algorithms of decision tree," In: *International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2010, pp. 526-528.

[11] B. Depaire, G. Wets, and K. Vanhoof, "Traffic accident segmentation by means of latent class clustering," *Accident Anal. Prev.,* vol. 40, pp. 1257-1266, 2008.

[12] J. de Ona, G. Lopez, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks," *Accident Anal. Prev.,* vol. 51, pp. 1-10, 2013.

[13] S. Y. Sohn and S. H. Lee, "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea," *Safety Sci.,* vol. 41, pp. 1-14, 2003.

[14] E. Bayam, J. Liebowitz, and W. Agresti, "Older drivers and accidents: A meta analysis and data mining application on traffic accident data," *Expert Syst. Appl.,* vol. 29, pp. 598-629, 2005.

[15] V. A. Olutayo, and A. A. Eludire, "Traffic Accident Analysis Using Decision Trees and Neural Networks," *Int. J. Inf. Tech. Comput. Sci.,* vol. 6, no. 2 pp. 22-28, 2014.

[16] X-F. Zhang, and L. Fan. "A decision tree approach for traffic accident analysis of saskatchewan highways," *Electrical and Computer Engineering (CCECE), 26th Annual IEEE Canadian Conference on. IEEE*, 2013.

[17] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.,* vol. 23, pp. 3-13, 2000.

[18] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Morgan Kaufmann: US 2011.

Received: April 7, 2014 Revised: September 23, 2014 Accepted: September 25, 2014

© Mansouri and Kargar; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.