SAJJAD SHOKOUHYAR, Ph.D.[1]
E-mail: s_shokouhyar@sbu.ac.ir
EHSAN TAATI, M.Sc.[1]
(Corresponding author)
E-mail: ehsantaati@yahoo.com
SARA ZOLFAGHARY, M.Sc.[1]
E-mail: sa.zolfaghary@gmail.com
[1] Shahid Beheshti University
Information Technology Management Department
Faculty of Management and Accounting, Tehran, Iran

Safety and Security in Traffic
Original Scientific Paper
Submitted: 8 Dec. 2016
Accepted: 3 July 2017

THE EFFECT OF DRIVERS' DEMOGRAPHIC CHARACTERISTICS ON ROAD ACCIDENTS IN DIFFERENT SEASONS USING DATA MINING

ABSTRACT

According to World Health Organization, each year, over 1.2 million people die on roads, and between 20 and 50 million suffer non-fatal injuries. Based on international reports, Iran has a high death rate caused by road accidents. The objective of this study was to extract implicit knowledge from road accident data sets on roads of Iran through data mining. In this regard, three useful data mining techniques were combined: clustering, classification and rule extraction. Following the preparation stage, data were segmented via three clustering algorithms; Kohonen, K-Means and Two-step. Two-step cluster analysis is a one-pass-through data approach which generates a fairly large number of pre-clusters. Next, the optimized algorithm and cluster were identified, after which, in the classification level and by adding the drivers' demographic features through C5.0, a classification algorithm was employed so as to make the decision tree. Ultimately, the effects of these demographic features were investigated on road accidents. The characteristics such as age, job, driving license duration and gender proved to be more important factors in accident analysis. Certain rules of accidents were then extracted in each season of the year.

KEY WORDS

traffic accidents; demographic features; data mining; season of the year;

1. INTRODUCTION

In recent decades, human ability to generate and collect data has increased. Accordingly, analysing, interpreting and making the maximum use of data is difficult and resource-demanding due to the exponential growth of many business, governmental and scientific databases. According to [1] the data mining technique enables organizations to properly utilize their capital data and promote decision-making. Road safety is a major concern of a country's transportation industry. Authors of study [2] believe that in the effort to alleviate the issue of vehicle accidents, it is crucial to identify factors leading to accidents through developing a capacity to design and implement an effective traffic information system that can provide timely and accurate traffic information. On the basis of World Health Organization report [3] more than 1.2 million people die each year on the world's roads, while from 20 to 50 million suffer non-fatal injuries. Iran is experiencing the highest rate of such accidents resulting in fatalities and various levels of injuries (*Table 1*) whose costs, more often than not, entail a great impact on the socio-economic development of a society.

According to World Health Organization, road injuries annually impose six billion dollars on Iran's economy which is approximately 8 percent of the country's GDP. It, therefore, goes without saying that it is of utmost significance for researchers in traffic safety to understand the circumstances under which the drivers and passengers are more bound to be killed or severely injured in an automobile crash. In Iran, little research has been done on data mining techniques for identifying factors contributing to accidents; hence, the incentive to conduct the present study.

The objective of this research was to identify the impact of drivers' demographic features on road accidents through data mining techniques. In addition, at the end of this paper, certain accident rules pertaining to each season of the year are extracted from the implicit knowledge of the data. Clustering road accident data using data mining algorithms allows one to discover rules that can be applied to the improvement of road safety by the traffic police and corresponding organizations such as road, transportation and highway engineers. The share of each demographic variable effective in accidents is also specified based on the classification results.

Table 1 – Road traffic mortality rate in Eastern Mediterranean countries (per 100,000 population) [3]

Country	Mortality rate	Country	Mortality rate
Libya	73.4	Yemen	21.5
Iran (Islamic Republic of)	32.1	Morocco	20.8
Saudi Arabia	27.4	Iraq	20.2
Jordan	26.3	Syrian Arab Republic	20.0
Somalia	25.4	Kuwait	18.7
Oman	25.4	Afghanistan	15.5
Djibouti	24.7	Qatar	15.2
Tunisia	24.4	Pakistan	14.2
Sudan	24.3	Egypt	12.8
Lebanon	22.6	United Arab Emirates	10.9

2. LITERATURE REVIEW

A considerable amount of studies have been carried out to identify the factors most important in increasing the level of injuries or crash severity. The objective is to reduce the number of people killed and/or injured in traffic accidents through eliminating or controlling these factors. For this purpose, each research employs a certain type of data mining technique. Generally, problems related to data mining are classified into two groups: supervised learning and unsupervised learning. In summary, supervised learning is synonymous with classification and comes from the labelled examples in training datasets. Unsupervised learning is essentially synonymous with clustering, a process where the input example is not class labelled [4]. Some studies belong to the first category, some to the second one and some of them make use of both supervised and unsupervised learning techniques. The studies associated with supervised learning were primarily considered. The authors of study [5] conducted non-parametric techniques of classification tree for the analysis of severity of injuries and determining the relationship between the severity of accidents, driver characteristics and environmental features. The results showed that pedestrians, motorcyclists and cyclists are more involved in severe injuries compared with car drivers. Fortin M et al. [6] applied a multivariate logistic regression to specify the independent contribution of drivers, crash, and vehicle characteristics to drivers' fatality risk. It was found that increasing the seatbelt use, reducing speed, and reducing the number and severity of driver's side impacts might preclude fatalities. Other researchers [7] used decision-tree modelling in order to study the relationship between the severity of accidents and motorist features. In this study, Clementine software was employed so as to build the models to predict and identify the relationship between the variables that affect the driver's responsibility, such as age, education, type of license, driving experience and other environmental factors. The results indicated that to create a predictive model, the decision tree is a good algorithm and the foregoing variables influence the severity of the injuries. Chen-SH [8] conducted a data mining research focusing on building tree-based models to analyse freeway accident frequency. These authors developed classification and regression tree (CART) and negative binomial regression models to establish the empirical relationship between traffic accidents, highway geometric variables, traffic characteristics and environmental factors. Based on their findings, the average daily traffic volume and precipitation variables were the key determinants of freeway accident frequency. A logistic regression tree and LR algorithms were employed by Pakgohar et al. [9] to determine the factors influencing road crashes. They focused on the environmental and human factors to conduct their study. They found out that weather, lighting conditions, time of day, driving license, not buckling up and using alcohol entail major car accidents. Some of them have made use of unsupervised learning techniques expressed in the following: authors of study [10] used Kohonen algorithm to investigate data mining techniques relating to the recorded characteristics and severity of road accidents with accident severity in Ethiopia. They concluded that road factors (type of road surface, weather conditions, lighting conditions and road direction) are factors determining accident occurrence. Xu C et al. [11] via k-means modelling, dealt with non-behavioural factors including the geometric characteristics of freeway, traffic factors such as traffic volume during the day and such environmental factors as annual raining, all of which played significant parts in road accidents. The third group of studies makes use of both techniques: Ng K-S et al. [12] used a combination of cluster analysis, regression analysis and geographic information system (GIS) to classify homogeneous accident data, estimate the number of traffic accidents and assess RTA risk in Hong Kong. Their resulting algorithm indicated an enhancement in accident risk estimation compared to estimates merely based on historical accident records. These authors claimed that the proposed algorithm

could be used to help authorities identify effectively the areas with high accident risks, and serve as reference for town planners as far as road safety is considered. Authors of study [13] combined data mining and statistical regression methods to identify the main factors associated with the levels of pedestrian injury. Based on their research, it was found that pedestrian age, location type, driver age, vehicle type, driver alcohol involvement, lighting conditions, and several built environment characteristics influence the likelihood of fatal crashes. Olutayo and Eludire [1] analysed Traffic Accident Using Decision Trees and a two-step algorithm by using WEKA software, where the three most important causes of accidents were proven to be tyre burst, loss of control and overspeeding.

In the present research, which belongs to the last group, the clustering method was used which is unsupervised learning and classification technique pertaining to the supervised learning methods. It should be noted that, unlike this work, none of the foregoing studies employed classification algorithms in their determination of rules.

3. DATA ANALYSIS AND EXPERIMENTAL RESULTS

3.1 Conceptual framework

In the first stage, after collecting data from the Traffic Police database, accident data were collected from two datasets related to accidents and operating vehicle information and an accident database was established in order to continue the data mining stages. In the second stage, data pre-processing was conducted and in addition to data clearing, human factors were identified and separated from the original data for the clustering stage. The reason for the separation was to eliminate the effect of drivers' demographic data on the clustering stage. The data would be employed in the classification stage. The third stage was dedicated to data modelling and rule extraction. Clustering and classification are the two models used in this paper. Following the clustering stage, the drivers' demographic data were added to the output of the clustering and classification model so as to identify the rules of accidents in a different season of the year and the importance degree of drivers' demographic variables in stage four. (*Figure 1*)

3.2 Data collection

The data were collected from road accidents on the roads leading to the east of Tehran from 2011 to 2014. Tehran is the capital and the largest city in Iran with a large population. In this paper, 8,950 records related to road accidents were studied. Data were extracted from Iran's driving police database, the reference for all accidents in Iran.

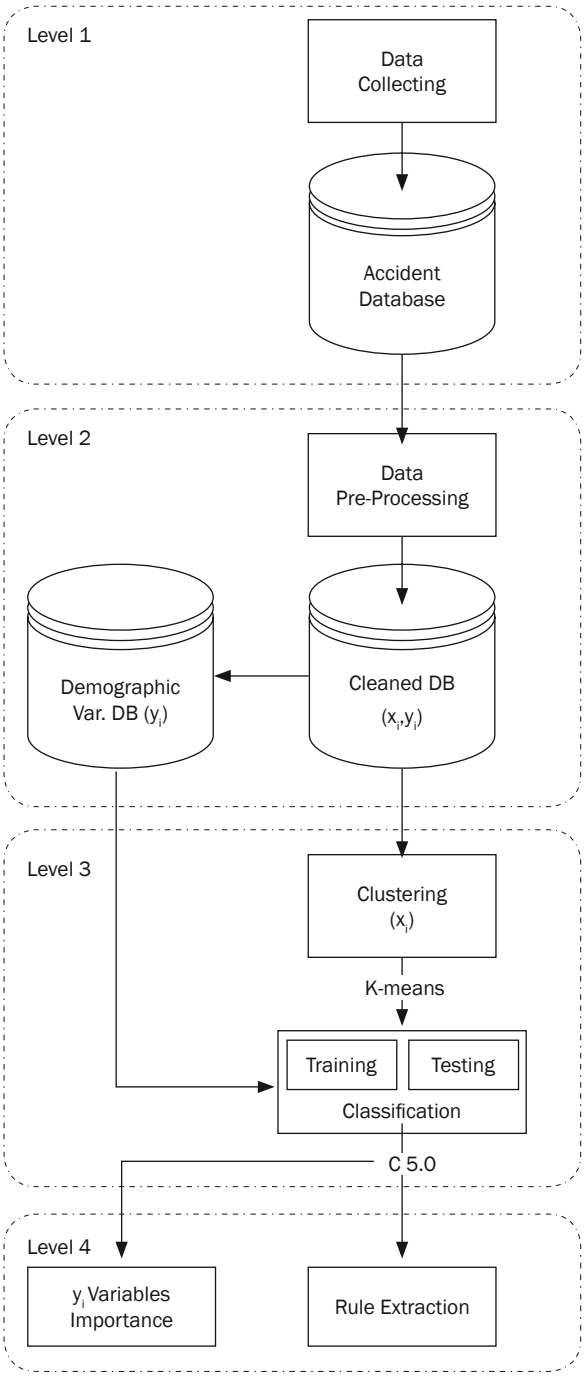

Figure 1 – The conceptual framework of the research

Most of the accidents (about 87%) involved no injuries. Approximately one out of ten accidents entailed human injury and around 3 percent led to mortality. In about 5% of the fatal accidents, the drivers were young women (20-30 years of age) while in the rest of the accidents, men (20-65 years old) were responsible. In Iran, as in most countries, the legal age to have

a driving license is 18. Eighteen percent of the total drivers contributing to this study had less than one year of driving experience, while, based on the traffic rules in Iran, the drivers with less than one year of driving experience are not allowed to drive on suburban roads. Needless to say, more attention must be paid by the government to such younger drivers.

The data were recorded in COM-114 forms by expert police officers of accidents and kept in accidents' information database. This forms record information on accidents and their causes in five major parts. The data included information regarding the conditions of the accident such as the profile of the vehicle involved in the accident, the road profile, the driver profile and the state of driving laws [14]. The accident data were completed by police officers and updated at the end of each month regarding the health status of the injured people (for instance if the rate of mortality augmented in the days following the accident).

Variables in this study were selected from previous studies and the opinions of road accident expert. A total of 17 variables were selected from previous studies and 3 variables were chosen based on expert opinion; the database was formed with 20 data fields (*Table 2*).

Table 2 – Variables used in the analysis of accident data

Label	variable	Description	Researcher
-	ID	Accident Identification Number	-
y_1	Accident date	Date of accident	-
y_2	Driver age	Quantities	[15]
x_1	Collision manner	Front to a fixed object, front to the right side, front to the left side, behind to a fixed object, behind to the right side, behind to the left side, front to front, front to behind, right side to a fixed object, behind to an object on the left	[16]
y_3	Driver education	Uneducated, elementary school passed, high school passed, BS, MS, PhD	Proposed by expert
y_4	Driving license date	Driver's license date	Proposed by expert
y_5	Driver occupation	Driver's job	Proposed by expert
y_6	Gender	Male, female	[10]
x_2	Human factors	Drug abuse, violation of traffic rules, over-agedness, unfamiliarity with roads, alcohol abuse, fatigue and sleepiness, hurrying, loss of control, overtaking, overspeeding	[17]
x_3	Lane lines	Broken white lines, single continuous white line	[18]
x_4	Lighting condition	Daylight, dark, dusk/dawn	[19]
x_5	Road defects	Lack of safety barriers along the road, poor lighting on roads, road erosion, surface defects, bumps, inadequate traffic road signs	[20]
x_6	Road geometric characteristic	Uphill, straight, downhill, flat	[21]
x_7	Road surface condition	Dry, wet, icy, gravel/sand, slush/mud, standing oil, other	[22]
x_8	Safety equipment	No special safety equipment, air bag, ABS break	[20]
x_9	Road direction	One-way, separated two-way, unseparated two-way	[19]
x_{10}	General cause of accident	Not looking ahead, sudden opening of the car door, exceeding the speed limit, swerve to the left or right, abrupt change in direction	[18]
x_{11}	Type of collision	Collision with motorcycle/bicycle, two-vehicle collision, multi-vehicle collision, collision with pedestrian, collision with animal, fixed object collision, overturn, fire/explosion	[21]
x_{12}	Type of region	Mountain, plain, foothill	[23]
x_{13}	Type of shoulder	No shoulder, asphalt	[10]
x_{14}	Vehicle type	Mini bus, bus, pickup, light truck, truck, ambulance, truck with trailer, motorcycle, bicycle, agricultural vehicles, highway const. equipment, fire truck, police car, other	[20]
x_{15}	Weather conditions	Clear, fog, rain, snow, storm, cloud, dust	[22]

For more convenience in later references, we applied a symbolic notation to the variables. These variables (labelled x_i) were used in the clustering stage, while others (labelled y_i) were employed in the classification stage.

3.3 Data pre-processing

Real world data generally suffer from problems such as noise, bias, extreme changes in dynamic range and sampling. If employed as they are in data mining projects, such data will not be able to yield acceptable results or they will pose problems in the data mining process. Data pre-processing includes all processes done on raw data, rendering them simpler and more effective for subsequent processes such as classification, clustering or other models.

In this paper, certain pre-processing techniques were used to prepare data for modelling and analysing (*Table 3*).

After pre-processing, the data were collected in a database named "Cleaned DB" (*Figure 1*). The Cleaned DB was later used in the data analysis step. This database includes x_i and y_i variables as shown in *Figure 2*.

3.4 Data analysis

The third stage was devoted to data analysis. Two modelling approaches were carried out on the data. First, through clustering, similar records were identified; subsequently, via applying rule classification technique on the demographic data, the similarity was extracted in the form of several rules. Each stage was further discussed.

3.4.1 Clustering

Clustering includes a set of records, each with a set of attributes. In cluster analysis, records with the most similarity (according to the likeness criterion defined earlier) are placed in a cluster. According to Han [4], there exist four types of clustering methods: partitioning, hierarchical, density-based, and grid-based methods (*Table 4*).

Three algorithms of the partitioning methods were employed: K-Means, Kohonen and Two-step, all of which belong to the unsupervised learning category in data mining. K-Means is a prototype-based, simple partitional clustering algorithm which finds K non-overlapping clusters. Such clusters are presented

ID	Accident date	Driver Age	Collision manner	Driver education	Driving license date	Driver occupation	Gender	Human factors	Lane lines
	y1	y2	x1	y3	y4	y5	y6	x2	x3
4247	06/03/2012	45	front to right side	uneducated	14/03/2011	self employed	male	Fatigue and sleepiness	broken white lines
4249	13/03/2012	47	front to behind	high school passed	18/07/2010	employee	male	violation of traffic rules	broken white lines
4250	13/03/2012	54	front to behind	BS	26/03/2011	military	male	Fatigue and sleepiness	broken white lines
4251	07/03/2012	59	right side to fixed object	uneducated	08/08/2010	household	female	unfamiliarity with the roads	broken white lines
4252	06/03/2012	26	right side to fixed object	high school passed	11/04/2001	self employed	male	Fatigue and sleepiness	broken white lines
4298	06/06/2012	43	front to behind	uneducated	12/08/2001	employee	male	violation of traffic rules	broken white lines
4299	19/06/2012	33	front to behind	BS	23/04/2009	employee	male	wrong overtaking	broken white lines
4300	11/07/2012	32	right side to fixed object	high school passed	17/08/2007	self employed	male	violation of traffic rules	broken white lines
4439	29/05/2012	62	right side to fixed object	uneducated	13/06/2011	self employed	male	violation of traffic rules	broken white lines
4470	16/06/2012	30	front to behind	BS	03/08/2009	self employed	male	wrong overtaking	broken white lines
4473	28/06/2012	41	behind to right side	uneducated	20/09/2000	driver	male	violation of traffic rules	broken white lines
4474	28/06/2012	40	front to right side	BS	01/05/2011	self employed	male	alcohol drinking	broken white lines
4485	14/06/2012	34	front to right side	high school passed	14/08/1997	employee	male	violation of traffic rules	single continuous whi
4486	14/06/2012	32	front to front	elementary school passed	13/12/2011	self employed	male	violation of traffic rules	broken white lines
4487	16/06/2012	40	front to behind	BS	05/06/2007	lawyer	male	alcohol drinking	single continuous whi
4488	16/06/2012	51	front to fixed object	BS	26/02/2011	unemployed	male	wrong overtaking	broken white lines
4489	16/06/2012	29	front to behind	high school passed	17/12/1999	household	female	wrong overtaking	broken white lines

Figure 2 - A view of Cleaned DB

Table 3 – Data pre-processing techniques employed in data preparation

Technique	Usage	Result
Removing duplicate data	Find and remove duplicate records in data set	Data volume was reduced to 8,950 records.
Create a new feature	Convert and transform a variable and make a new one	Date pertaining to driving license and accident were transformed to driving license duration and season.
Discrete optimization	Convert continuous variables to discrete variables via binning method	Driver age and driver license duration were converted to discrete by granulation operation with a Bin value of 10.
Management of missing values	Find and remove or fill missing values such as the data unknown, not properly collected, wrongly entered or not entered at all (null)	C&RT algorithm (a decision tree algorithm managing missing data through defining null values as allowed values [24]) was randomly used for the management of missing values.

Table 4 – Overview of clustering methods

Methods	General characteristics
Partitioning Methods	Find mutually exclusive clusters of spiral shapes Distance-based May use mean or medoid to represent cluster centre Effective for small to medium size data sets
Hierarchical Methods	Clustering is a hierarchical decomposition Cannot correct erroneous merge or split
Density-based Methods	Can find arbitral shaped clusters Clusters are dense regions in space that are separated by low density regions Cluster density: Each point must have a minimum number of neighbouring points. May filter out the outliers
Grid-based Methods	Use a multi-resolution grid data structure Fast processing

by their centroid (a cluster centroid typically refers to the points in that cluster) [25]. The Self-organizing Map (SOM), commonly known as Kohonen network, is a computational method for the visualization and analysis of high-dimensional data, in particular, experimentally acquired information [26]. The network contains input and output layers. The neurons connect the two layers and map the input layer on the output layer with a two-dimensional discrete graphics. Such type of algorithm is used in clustering a subset of data into distinct sections when the groups are primarily ambiguous. The two-step cluster method is a scalable cluster analysis algorithm designed to handle very large datasets, able to tackle both continuous and categorical variables and attributes. This method includes two steps: The cases (or records) are firstly pre-clustered into many small sub-clusters; secondly, the sub-clusters resulting from the pre-clustering step are clustered into a desired number of clusters. It can also automatically select the number of clusters [27].

As mentioned earlier, the demographic variables (y_i variables) were excluded from Cleaned DB and were not used as inputs for the clustering stage. In the clustering stage, the relevant records of the datasets were used (x_i variables), namely, collision manner, human factors, lane lines, lighting condition, road defects, road geometric characteristics, road surface condition, safety equipment, separated or non-separated direction of road, general cause of accident, type of collision, type of region, type of shoulder, vehicle type, weather condition.

Three criteria were used to identify the optimized algorithm: Runtime, silhouette index and the number of output clusters. The optimized algorithm was selected with indices of low runtime, high silhouette index and expert opinion on the number of clusters. In this paper, K-Means was selected as the optimized algorithm (*Table 5*).

Table 5 – Comparison of the clustering algorithms

	Silhouette	Runtime [s]	Number of clusters
Kohonen	0.1	18	12
Two-step	0.1	10	2
K-Means	0.2	8	5

Owing to its high runtime, Kohonen algorithm was considered as weak, hence not selected. Between the two other algorithms, K-Means was chosen as the optimized algorithm because of its lower runtime and higher silhouette index. The results of K-Means clustering are shown in *Figure 3*. The number of clusters was identified through several executions of each algorithm and in consultation with experts.

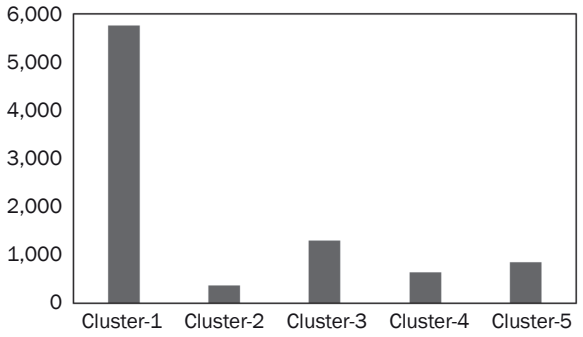

Figure 3 – The volume of clusters created by K-Means

According to the obtained results, the highest share (64.5%, 5,775 records) belongs to Cluster 1. Cluster 3, with nearly 1,300 records, occupies the second position. Other clusters have less than 1,000 records.

Variable distribution among clusters are shown in *Figure 4*, which indicates how x_i variables were allocated to each cluster. Cluster 1 is the biggest and has the best variable distribution, hence selected as the optimized cluster on which further work was done.

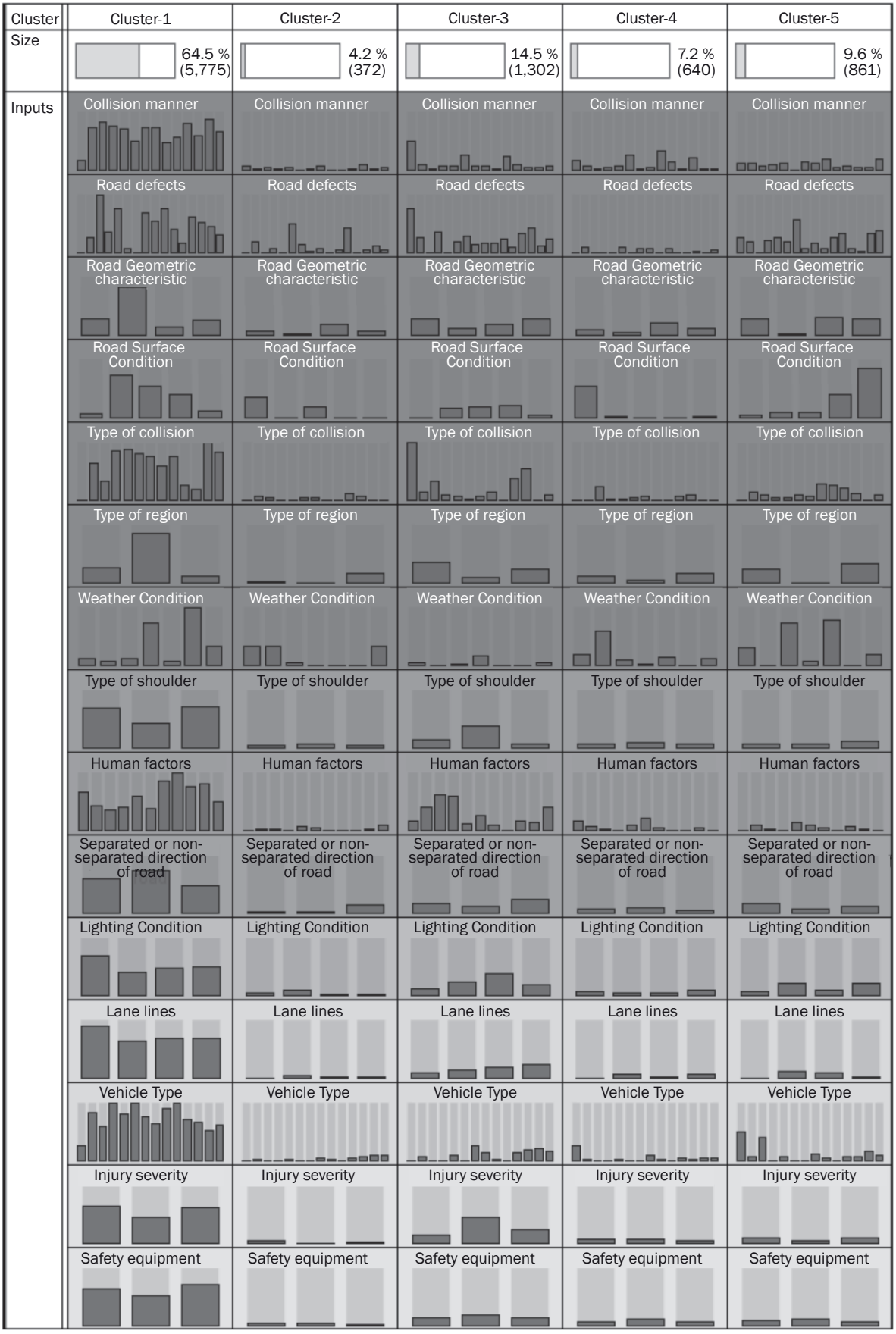

Figure 4 – x_i variable distribution in clustering by K-Means

Predictor importance is an evaluation factor provided by IBMSPSS Modeler14.0. This factor renders it possible to know the importance degree of each contributed variable in order to drop or ignore those that matter least and keep those that matter most. This process is done by indicating the relative importance of each predictor or variable in the estimation model. Predictor importance indicates how well the variable can differentiate various clusters. If the values are relative, their sum for all variables is displayed as 1. [28]

As seen in *Table 6*, weather condition, road defect, type of collision, collision manner, road geometric characteristics, and road surface conditions have the highest effect on clustering, meaning that such variables (predictors) are totally related to the predication with clustering technique in this paper. Other such variables as type of shoulder, human factors, separated or non-separated direction of road, lighting conditions, lane lines, vehicle type, injury severity and safety equipment have less contribution to clustering and the general cause of accident has no contribution.

3.4.2 Classification and Rule Extraction

As mentioned in the previous section, Cluster 1, obtained from the modelling operation of algorithm K-Means, was selected as the optimized cluster. In the next stage of modelling, classification was used in order to specify the drivers' demographic features.

Given that clustering based on similarity identifies data similar to one another, it becomes clear what the reason for the similarities is and what rules identify them in different seasons of the year. Due to the availability of the date of accident there exists a new field called Accident Occurrence Season. This field was added to the set of demographic variables of drivers (*Table 7*) of vehicles excluded from the dataset prior

Table 6 – variable importance of Cluster 1 in K-Means clustering

Label	Variable	Description	Input Importance Degree
x_1	Collision manner	front to fixed object, front to the right side, front to the left side, behind to fixed object, behind to the right side ,behind to the left side, Front to front, front to behind, right side to fixed object, behind to objects on the left	1
x_5	Road defects	lack of safety barriers along the road, poor lighting on the road, road erosion, surface defects, bump, inadequate traffic road signs	1
x_6	Road geometric characteristics	uphill, straight ,downhill ,flat	1
x_7	Road surface condition	dry, wet, icy, gravel/sand, slush/mud, standing oil, other	1
x_{11}	Type of collision	collision with motorcycle/bicycle, two-vehicle collision, multi-vehicle collision, collision with pedestrian, collision with animal, fixed object collision, overturn, fire/explosion	1
x_{12}	Type of region	mountain , plain , foothill	1
x_{15}	Weather condition	clear, fog, rain, snow, storm, cloud, dust	1
x_{13}	Type of shoulder	no shoulder, asphalt	0.76
x_2	Human factors	drug abuse ,violation of traffic rules, unfamiliarity with the roads, alcohol abuse, fatigue and sleepiness, hurrying ,loss of control, overtaking, overspending	0.75
x_9	Direction of the road	one-way, separated two-way, unseparated two-way	0.70
x_4	Lighting condition	daylight, dark, dusk/dawn	0.40
x_3	Lane lines	broken white lines, single continuous white line	0.12
x_{14}	Vehicle type	bus, middle bus , pickup , light truck, truck, ambulance, truck with trailer, agricultural vehicles, highway const. equipment, fire truck, police car, other	0.12
x_8	Safety equipment	no special safety equipment, air bag, ABS break	0.01
x_{10}	General cause of accident	not looking ahead, sudden opening of the car door, exceeding the speed limit, car deviation swerve to the left or right, abrupt change in direction	0

Table 7 – The drivers' demographic variables

Label	Variable	Description	Data Type
y_2	Driver Age	The age of the driver riding the vehicle	continuous
y_7	Driving License Duration	The duration of the driver's license up to the date of accident	continuous
y_5	Driver Occupation	The job of the driver	nominal
y_3	Driver Education	Uneducated, elementary school passed, high school passed, BS,MS,PHD	nominal
y_6	Gender	Male, female	flag
y_8	Accident Occurrence Season	Spring, summer, fall, winter	nominal

to clustering and kept in the Demographic Var. DB (*Figure 1*). According to *Table 7*, y_7 y_8 are new variables, obtained from y_4 and y_1 (*Table 2*), respectively.

Figure 5 demonstrates the dispersion of road accidents in different seasons as regards Driving License Duration and Driver Age. Each black spot in the rectangles represents an accident. As expected, in all seasons of the year, most accidents occurred because of the younger and less experienced drivers.

According to *Table 7*, data type column, and based on *Figure 5*, y_7 and y_2 have continuous values. In order to perform the classification operation, discrete values are needed; therefore, we employed binning granule in software IBM SPSS Molder where the data were converted into discrete values with 10 granules.

Variables y_i were combined with regards to Cluster 1 data on the basis of ID field of the Cleaned DB (*Figure 2*) and used as classification input with algorithm C5.0 which is an extension of C4.5 algorithm, itself an extension of ID3[29]. The C.50 algorithm first grows on an overfit tree, subsequently pruning it back to create a more stable model [24]. Variable season (y_8) was selected as the category feature, while other fields were selected as inputs. To maximize the interpretability, C5.0 classifiers were expressed as decision trees or sets of if-then rules, forms generally easier to fathom compared with neural networks [30]. In C5.0, several new techniques are introduced [31]:

- Boosting: several decision trees are generated and combined to improve the predictions.
- Variable misclassification costs: which makes it possible to avoid harmful errors.
- New attributes: dates, times, timestamps, ordered discrete attributes.
- Values can be marked as missing or not applicable for particular cases.
- Supports sampling and cross-validation.

Because of model assessment, it is required to partition the data with separate testing and training data. The testing data were not used in building the model, which is explained later in the model evaluation section. The data were analysed with C5.0 (*Figure 6*). Note that the Season field is selected as the target field.

As explained in Section 3.4.1, with predictor importance indicator one is able to rank the contribution degree of each value in data analysis. *Figure 6* clarifies how y_i variables take part in the classification stage. Nearly 40 percent of the variables are related to Driver Education, the highest share among variables present in C.50 analysis. Driving License Duration, with about thirty percent contribution in the classification,

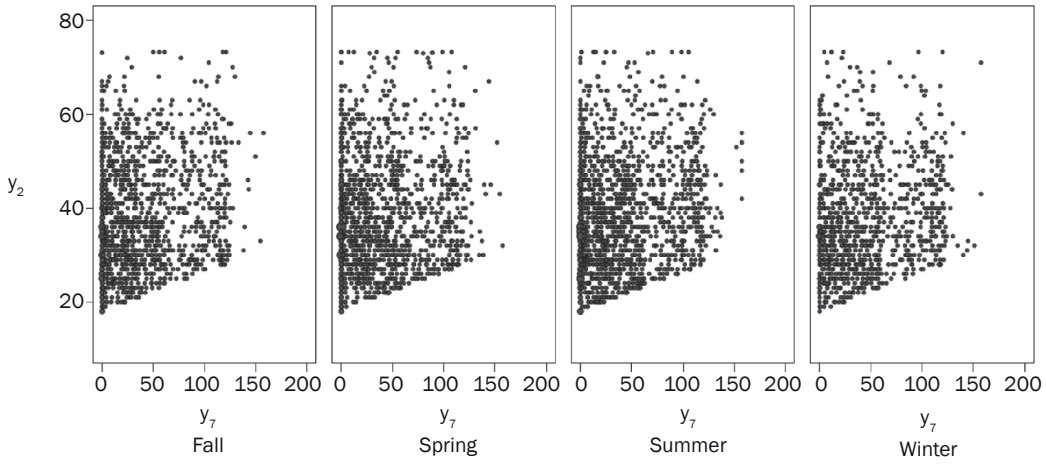

Figure 5 – Statistics of road accidents in different seasons based on y_2 (Driver Age) and y_7 (Driving License Duration)

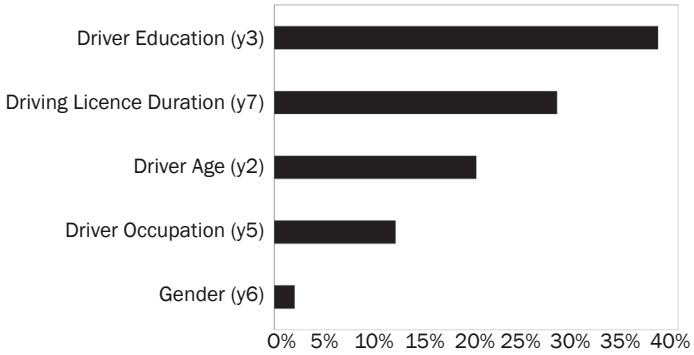

Figure 6 – The percentage of y_i variable importance in the classification stage done by C5.0

occupies the second rank. Next comes the Driver age just 20 percent share. Comparisons to other variables, Occupation (with less than 15% contribution) does not seem to be playing any significant roles in accidents. The gender has the lowest influence on accidents by merely 2 percent.

Overall, *Table 8* shows that among the demographic features, education and experience entail the highest impact (approximately 70 percent).

Table 8 – y_i variable importance in the classification with algorithm C5.0

Label	Variable	Predictor Importance [%]
y_3	Driver Education	38
y_7	Driving License Duration	28
y_2	Driver Age	20
y_5	Driver Occupation	12
y_6	Gender	02

As mentioned before, one of the usages of C.50 is to interpret the relationship between variables using rule sets. The rules extracted from C5.0 analysis are illustrated in *Table 9*. Thirty-four rules were extracted from Cluster 1. There are two indexes in *Table 9*: Incidence and Confidence, the former identifying the frequency of the rule in a given data set and the latter representing the portion of records within the same correctly classified set [32]. For example in *Table 9* there are 470 records that support rule 2 with an accuracy of 0.311. Note that if the confidence of a rule is one, it means that the rule is perfectly (100%) accurate. Via such indexes, the quality of any rule can be appraised.

Take rule 1 for instance, if the driving license duration of a driver is ranged between 12 and 67 months and the driver occupation is military- or labour-related, there exists a 33 percent chance that the accident will happen in fall. This rule was repeated 45 times in the dataset. As far as rule 2 is concerned, if the driver education is "high school passed" and the driver age is from 31 to 39 or 62 to 69 and the driving license duration is in the range of 12 to 67 months the accident can happen with 31% chance in spring. This rule was repeated 470 times in the dataset.

Table 9 – The rules extracted from C.50 algorithm

ID	Rule Description	Incidence	Confidence
1	if DL_Duration_BIN in [1] and Driver Occupation in ["Military Labor"] then Fall	45	0.333
2	if Driver Education in ["High School passed"] and Driver Age_BIN in [4 8] and DL_Duration_BIN in[1] and Driver Occupation in ["Self-employed"] then Spring	470	0.311
3	if Driver Education in ["Elementary School passed"] and DL_Duration_BIN in [1] and Driver Occupation in ["Self-employed"] then Summer	292	0.322
4	if Driver Education in ["BS"] and DL_Duration_BIN in [2] and Driver Occupation in ["Public sector employee"] then Winter	82	0.354
. . .	. . .	. . .	. . .

4. MODEL EVALUATION

To predict the accuracy of the model, the Assessment Matrix and two more new fields were used, namely, *$C-Season and $CC-*Season (*Figure 7*). These fields predict the value of each record and the confidence value for the rule. For C5.0 rule sets, the prefixes are $C- for the target field and $CC- for the confidence field. Season was considered as the target field. Through these fields and using the training and testing data, we were able to evaluate the model. The

	SERIALNO	Partition	$C-Season	$CC-Season
1	2001	1_Training	Fall	0.293
2	4550	1_Training	Spring	0.335
3	12341	2_Testing	Spring	0.335
4	52802	2_Testing	Winter	0.354
5	52821	1_Training	Fall	0.750
6	52867	1_Training	Summer	0.346
7	52879	2_Testing	Spring	0.311
8	52893	1_Training	Winter	0.354
9	52907	1_Training	Winter	0.354
10	52914	1_Training	Summer	0.371
11	52933	2_Testing	Summer	0.346
12	52935	2_Testing	Summer	0.308
13	52946	2_Testing	Spring	0.500
14	52957	1_Training	Spring	0.319
15	53289	2_Testing	Spring	0.319

Figure 7 – Prediction ($C-Season) and confidence ($CC-Season) values in C5.0 rule induction model

Partition field, as already mentioned, represents that the records have been used in training or testing stages by the model. It is possible to measure the accuracy of the model in testing and training stages.

Table 10 demonstrates the Model Assessment Matrix by the training values, where the operation accuracy of the model is specified by its rule set or model. For example, a model predicting about 76 percent of summer data, and around 73 percent of spring data is considered as accurate.

Table 10 shows the accuracy of the appropriate season if they have been used in another season. For example, the model correctly predicts about 70 percent of the fall if employed in summer.

The Model Assessment Matrix by the testing values is shown in *Table 11*. As observed, the model accurately predicts about 62 percent of summer, and about 50 percent of spring. The results of the testing values indicate how well the model performs with the new values.

5. DISCUSSION AND CONCLUSION

Road accidents are ineluctable factors as regards human mortality. Over the past few decades, the possibility of accidents occurring has increased; on the other hand, there exists much more information and reports as to the nature of accidents compared with the past. Owing to the large amount of information and the paucity of structure in data, traditional statistical methods to analyse data cannot work properly. Currently, data mining is one of the best ways to face and surmount such challenges and researchers can become informed of accidents through the related techniques and tools, which ultimately results in effective solutions to reducing accidents.

In this paper, road accidents in Iran were analysed using two competent and important techniques of data mining, and clustering and classification. The analysis sample was accidents happening on roads leading to

Table 10 – The Model Assessment Matrix by the training values

Season		Fall	Spring	Summer	Winter
Fall	Count	144	209	291	27
	Row %	64.81	55.24	70.22	18.50
Spring	Count	85	296	316	22
	Row %	31.09	73.23	71.12	14.12
Summer	Count	99	217	458	40
	Row %	27.31	52.32	76.32	13.40
Winter	Count	73	162	260	43
	Row %	22.65	62.11	68.21	22.31

Table 11 – The Model Assessment Matrix by the testing values

Season		Fall	Spring	Summer	Winter
Fall	Count	99	234	354	29
	Row %	42.31	40.20	63.12	13.67
Spring	Count	108	215	358	24
	Row %	27.45	50.31	67.12	10.92
Summer	Count	131	291	374	36
	Row %	27.20	50.69	62.22	11.73
Winter	Count	72	164	269	37
	Row %	22.42	60.82	59.22	15.31

Tehran. After collecting and clearing the data by three important algorithms in software IBM SPSS MODELER, they were analysed in two stages.

In the clustering stage, two major findings were obtained. First, among the three clustering algorithms, Kohonen, K-Means and Two-step, K-Means was found to be more acceptable because of the Silhouette index, runtime and number of clusters. Second, based on the variable importance index which identifies the relative importance of each contributed variable in the clustering stage, the following variables were regarded as the most relative factors in accidents: type of region, type of collision, weather condition, road geometric characteristics, road surface condition, and road defects. This is in line with the findings of Xu C et al. [11] study, and in total contrast to Mohamed MG et al. [13] where none of the foregoing variables were mentioned as effective factors in accident.

According to the classification stage carried out by C.50 algorithm, among the demographic features of the drivers, gender brought the least impact on accidents while education and experience played the most significant parts as Regassa Z [7] emphasized.

Furthermore, 34 rules were extracted from the clustered data by K-Means which is considered as a novel finding compared to previous studies. According to the classification done by season as the target field, certain distinct rules were identified for each season of the year. The incidence and confidence were the key indexes in the evaluation of the rules following the patterns existing for each season. The following rules had the highest points in as far as incidence and confidence:

Spring: "Driver_Age_BIN in [4 8] (488; 0.35)", meaning it was repeated 488 times with a 35 percent confidence. Most accidents were caused by drivers aged 22 to 30 or 53 to 60 years.

Summer: "Driver_Education in "Elementary school passed"(292; 0.32)". This means that most accidents in summer were caused by drivers with low education.

Fall: "Driver_Age_BIN in [2 7] (167; 0.29)": Most accidents happened because of drivers between 22 and 30 or 53 and 60 years of age.

Winter: "Driver_Occupation in "Military" (85; .032), most accidents in this season occurred by drivers with military-related occupations.

Factors effective on road accidents, which resulted from clustering and classification processes in this paper, conduce to precluding and reducing the accident rates. In addition, based on the rules of the classification stage, drivers who may have an accident in a specific season can be identified and further notified by the police. The demographic factors were also rated in the present study which might be conducive to future related studies.

Among the limitations of this study is the lack of information as to the environmental circumstances of the roads, neither were we able to get access to the drivers or their survivors to find out more relevant and effective information, such as where they were coming from or what they were doing (eating or using the cell phone) at the time of accident, the driving duration or distance before the accident and so forth. Furthermore, the results of data mining are sensitive to the quality of input data. Certain values went missing due to human error in the COM-114 forms completed by the police officers or when the data were entered to the computers by the operators.

Further research has to be done employing other types of data mining algorithms for both clustering and classification, such as SOM (Self Organization Map) or fuzzy clustering algorithms such as FC-Means. It is also useful to repeat this work using more variables such as the time of accident, driving duration prior to the accident and the driving speed.

سجاد شکوهیار، دکتری تخصصی[1]
E-mail:s_shokouhyar@sbu.ac.ir
احسان طاعتی، کارشناسی ارشد[1]
E-mail:ehsantaati@yahoo.com
سارا ذوالفقاری، کارشناسی ارشد[1]
E-mail: sa.zolfaghary@gmail.com

1 دانشگاه شهید بهشتی تهران، دانشکده مدیریت و حسابداری، تهران، ایران

چکیده:
مطابق آمار سازمان بهداشت جهانی، هر ساله 1.2 میلیون انسان در جاده ها کشته شده و حدود 20 تا 50 میلیون نفر از صدمات آن رنج می برند. بر اساس گزارشات بین المللی، ایران نرخ بالایی در تصادفات جاده ای دارد. هدف از این مقاله استخراج دانش پنهان از مجموعه داده های تصادفات جاده ای ایران از طریق داده کاوی بوده است. به همین منظور سه تکنیک پرکاربرد داده کاوی شامل خوشه بندی، طبقه بندی و استخراج قواعد با یکدیگر ترکیب شده اند. ابتدا داده‌ها توسط سه الگوریتم خوشه بندی Two-Step وKohonen, K-Means بخش‌بندی شده‌اند. سپس الگوریتم و خوشه بهینه شناسایی شده و پس از آن در مرحله طبقه بندی با بکارگیری الگوریتم C.50 و اضافه کردن خصوصیات دمگرافیک رانندگان درخت تصمیم ساخته شده است. در نهایت تاثیر ویژگیهای دمگرافیک رانندگان بر تصادفات جاده ای مورد بررسی قرار گرفته است. مشخصه هایی نظیر سن، شغل، مدت زمان گواهینامه، و جنسیت بیشترین تاثیر را در تصادفات جاده ای داشته اند. الگوهای مشخصی در خصوص تصادفات جاده ای در هر فصل سال شناسایی و ارایه شده است.

کلمات کلیدی:
تصادفات جاده ای، ویژگیهای دمگرافیک، داده کاوی، فصل های سال

REFERENCES

[1] Olutayo V, Eludire A. Traffic accident analysis using decision trees and neural networks. International Journal of Information Technology and Computer Science (IJITCS). 2014;6(2): 22-8.
[2] Ossenbruggen PJ, Pendharkar J, Ivan J. Roadway safety

in rural and small urbanized areas. Accident Analysis & Prevention. 2001;33(4): 485-98.
[3] WHO. Global status report on road safety: time for action 2016. Available from: http://www.who.int/gho/publications/world_health_statistics/2016/whs2016_AnnexA_RoadTraffic.pdf?ua=1.
[4] Han J, Kamber M, Pei J. Data mining: concepts and techniques. Elsevier; 2011.
[5] Chang L-Y, Wang H-W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention. 2006;38(5): 1019-27.
[6] Fortin M, Bédard S, DeBlois J, Meunier S. Predicting individual tree mortality in northern hardwood stands under uneven-aged management in southern Québec, Canada. Annals of Forest Science. 2008;65(2): 12 p.
[7] Regassa Z. Determining the degree of driver's responsibility for car accident: the case of Addis Ababa traffic office. Unpublished Master's Thesis. Addis Ababa University; 2009.
[8] Chen SH. Mining patterns and factors contributing to crash severity on road curves. Queensland University of Technology; 2010.
[9] Pakgohar A, Tabrizi RS, Khalili M, Esmaeili A. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science. 2011;3: 764-9.
[10] Beshah T, Hill S, editors. Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. 2010 AAAI Spring Symposium: Artificial Intelligence for Development, 22-24-Mar. 2010, Stanford, CA, USA; 2010.
[11] Xu C, Liu P, Wang W, Li Z. Evaluation of the impacts of traffic states on crash risks on freeways. Accident Analysis & Prevention. 2012;47: 162-71.
[12] Ng K-s, Hung W-t, Wong W-g. An algorithm for assessing the risk of traffic accident. Journal of Safety Research. 2002;33(3): 387-410.
[13] Mohamed MG, Saunier N, Miranda-Moreno LF, Ukkusuri SV. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety Science. 2013;54: 27-37.
[14] Khosravi Shadmani F, Soori H, Karmi M, Zayeri F, Mehmandar M. Estimating of Population Attributable Fraction of Unauthorized Speeding and Overtaking on Rural Roads of Iran. Iranian Journal of Epidemiology. 2013;8(4): 9-14.
[15] Alizadeh SS, Mortazavi SB, Sepehri MM. Prediction of vehicle traffic accidents using Bayesian networks. Scientific Journal of Pure and Applied Sciences. 2014;3(6): 356-62.
[16] Carrasco CE, Godinho M, de Azevedo Barros MB, Rizoli S, Fraga GP. Fatal motorcycle crashes: a serious public health problem in Brazil. World Journal of Emergency Surgery. 2012;7(Suppl 1).
[17] Yang J, Li F, Zhou J, Zhang L, Huang L, Bi J. A survey on hazardous materials accidents during road transport in China from 2000 to 2008. Journal of Hazardous Materials. 2010;184(1): 647-53.
[18] Shanthi S, Ramani RG. Classification of vehicle collision patterns in road accidents using data mining algorithms. International Journal of Computer Applications. 2011;35(12): 30-7.
[19] Fogue M, Garrido P, Martinez FJ, Cano J-C, Calafate CT, Manzoni P. A system for automatic notification and severity estimation of automotive accidents. IEEE Transactions on Mobile Computing. 2014;13(5): 948-63.
[20] Chong MM, Abraham A, Paprzycki M. Traffic accident analysis using decision trees and neural networks. arXiv preprint cs/0405050. 2004.
[21] Martín L, Baena L, Garach L, López G, de Oña J. Using data mining techniques to road safety improvement in Spanish roads. Procedia-Social and Behavioral Sciences. 2014;160: 607-14.
[22] Williams K, Idowu AP, Olonade E. Online Road Traffic Accident Monitoring System for Nigeria. Transactions on Networks and Communications. 2015;3(1): 10-21.
[23] Malgundkar T, Rao M, Mantha S. GIS driven urban traffic analysis based on ontology. International Journal of Managing Information Technology. 2012;4(1): 15-23.
[24] Linoff GS, Berry MJ. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons; 2011.
[25] Wu J. Advances in K-means clustering: a data mining thinking: Springer Science & Business Media; 2012.
[26] Kohonen T, Honkela T. Kohonen network. Scholarpedia. 2007;2(1):1568.
[27] IBM_Corporation. TWOSTEP CLUSTER Algorithms 2013. Available from: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.algorithms/alg_twostep.htm.
[28] Center IK. Predictor Importance 2012. Available from: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/model_nugget_variableimportance.htm.
[29] Brijain RP, Kushik KR, editors. A survey on decision tree algorithm for classification. International Journal of Engineering Development and Research; 2014;2(1): 5 p.
[30] Rulequest Research. Data Mining Tools See5 and C5.0 2015. Available from: http://www.rulequest.com/see5-info.html.
[31] Bujlow T, Riaz T, Pedersen JM, editors. A method for classification of network traffic based on C5. 0 Machine Learning Algorithm. Proceedings of 2002 International Conference on Computing, Networking and Communications (ICNC), 30 Jan.-2 Feb. 2012, Maui, HI, USA. IEEE; 2012. p. 237-41.
[32] IBM_Corp. Predictive Modeling with IBM SPSS Modeler 2010. Available from: https://www.scribd.com/document/141849191/MELJUN-CORTES-Predictive-Modeling-With-IBM-SPSS-Modeler.