AUT Journal of Civil Engineering

AUT J. Civil Eng., 5(3) (2021) 421-432
DOI: 10.22060/ajce.2021.19386.5728

A Data Mining Framework to Identify Important Factors of Fatigue and Drowsiness Accidents

A. Tavakoli Kashani*, M. Rakhshani Moghadam, S. Amirifar

Civil Engineering Department, Iran University of Science and Technology, Tehran, Iran.

ABSTRACT: Fatigue and drowsiness are the major factors contributing to accidents worldwide. According to statistics, 20 to 40 percent of traffic accidents in Iran are due to drivers' fatigue. This study aims to identify the most important variables affecting the occurrence of fatigue and drowsiness accidents based on the classification and regression tree (CART) method. At first, 859, 378 police crash data of provinces Tehran, Fars, and Mazandaran during seven years (2011-2018) were segmented into homogeneous groups using the two-step clustering algorithm. Next, an oversampling technique is applied to deal with the crash data imbalance problem. Finally, the classification and regression tree combined with the boosting algorithm increases the accuracy of the models. The results of the classification tree showed that the main variables affecting the occurrence of fatigue and drowsiness accidents are: road type, time of day, road traffic direction, local land use, shoulder type, vehicle type, control type, and collision type. Moreover, the road type variable was the only significant factor in residential suburban areas of Mazandaran and Fars provinces. Also, the common variable in residential urban areas of all three provinces was the time of day. It was concluded that the combination of the CART algorithm with oversampling and boosting increased the accuracy of the models. Identifying influential factors in fatigue and drowsiness accidents in the three mentioned provinces could improve the engineering and executive interactions and appropriate educational programs.

Review History:

Received: Dec. 12, 2020
Revised: Sep. 24, 2021
Accepted: Nov. 24,2021
Available Online: Dec. 03, 2021

Keywords:

Fatigue and drowsiness
Clustering analysis
Imbalance data
Classification and regression tree
Boosting

1- Introduction

The human factor accounts for 90% of traffic accidents in Iran [1, 2]. Drowsiness is the second most important cause of accidents after alcohol consumption [3], which has caused safety problems worldwide [4]. It was estimated that fatigue and drowsiness-related accidents account for 20% of all accidents in the world [5]. A nationwide Canadian survey found that more than half of drivers have driven under fatigue conditions, and one in five Canadian drivers claimed they had fallen asleep while driving at least once in the past 12 months [6]. Besides, 35,000 (9%) French drivers have reported that they are forced to stop driving at least once a month due to drowsiness [7].

Numerous studies have been conducted on fatigue and drowsiness worldwide. For example, 4-year traffic accidents data in china (2006-2010) indicated that male drivers, truck drivers have a high risk of driving fatigue behavior. Midnight to dawn and during busy hours, driving on streets without night lighting increases the incidence of fatigue-related accidents [8]. Since fatigue and drowsiness accidents have been demonstrated that are common on high-speed motorway driving [9-11], several studies have shown that fatigue-related accidents in remote areas and rural areas have a high risk [12, 13]. Moreover, many drivers have reported experiencing drowsiness accidents on low-speed roads, regardless of speed zone, any drowsiness accidents occur when commuting from/to work [14]. Furthermore, a study by Filtness et al. [15] indicated that driver drowsiness is not limited to high-speed and highway driving and that low-speed drowsiness accidents are frequent. For this reason, in this study, the crash data are divided into homogeneous clusters based on urban, suburban, residential, and non-residential areas and driver gender by using two-step clustering.

The regression analyses such as Logit [16, 17] and Probit [18, 19] models have been extensively used in traffic crash data. In regression modeling, the relationship between the dependent variable and the independent variables should be defined before the modeling, and if the hypotheses are rejected, the model estimation will face a significant error [20]. The classification and regression tree (CART) technique, which is a non-parametric algorithm, has been used in recent years to analyze traffic safety. The CART model is represented by a tree structure where each branch of the tree can be composed of different combinations of variables [21]. Also, the CART model can identify the most important variables and remove the least important ones. Since, Breiman [22] suggested the combination of the CART model with boosting algorithm

*Corresponding author's email: alitavakoli@iust.ac.ir

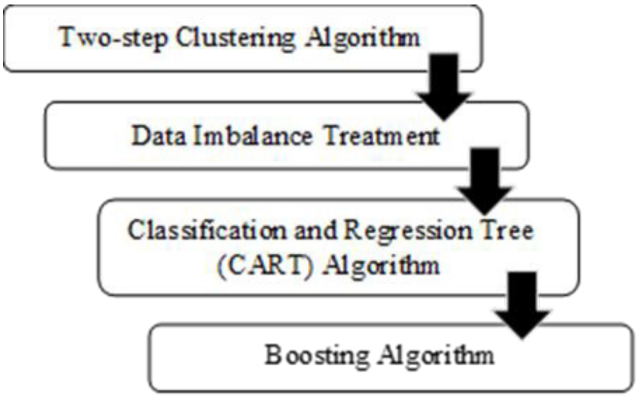
Two-step Clustering Algorithm
Data Imbalance Treatment
Classification and Regression Tree (CART) Algorithm
Boosting Algorithm

Fig. 1. A data mining framework.

increases the accuracy of the models, this combination is used in the current study. The classification algorithms only work well if the data are balanced. That is to say, the number of instances of different classes is the same. Traffic crash data are usually imbalanced. If the modeling is done with imbalanced data, it makes the accuracy of the minority class model difficult to compare with the accuracy of the majority class model, and also if this model is not correct in reality [23, 24]. In this study, the resampling technique is used to deal with data imbalance problem.

A review of the literature reveals that various factors, including human, vehicle, road, and environment can affect occurring of fatigue and drowsiness accidents. Therefore, the current study aimed to use the classification and regression tree (CART) method in combination with cluster analysis, imbalance dataset and boosting algorithm to investigate the important factors affecting the occurrence of fatigue and drowsiness accidents in three provinces of Iran. Since similar crashes are grouped into separate clusters by their levels of contributory factors, priority safety countermeasures, educational programs, and enforcement measures can be tailored to each cluster. The rest of this paper is organized as follows: The two-step, Oversampling, CART, and Boosting methods are explained in Section 2. Section 3 introduces the case study, and Section 4 conclusions are presented.

2- Method

In this study, classification and regression trees were used to identify the most important variables affecting the occurrence of fatigue and drowsiness accidents in three provinces of Iran. This section provides a summary of this model. A data mining framework is presented in Fig. 1.

As shown in this figure, first of all, the two-step clustering was used to divide data into homogeneous clusters. Then, the over-sampling method was used to treat the imbalanced structure of the dataset and investigate whether the accuracy of CART models could be improved using the amplification algorithm.

2- 1- Two-step clustering

The first step is to divide the crash data into homogeneous groups based on traffic parameters. In this study, the two-step clustering method was selected [25]. Which could be used for very large databases. It is the only algorithm that allows the examination of continuous and discrete variables. In addition, this method allows determining the number of clusters by setting the max and min amount. The two-step clustering algorithm consists of two steps. The first step is pre-clustering all the identified similar records as primary clusters. The primary clusters in the first step are used as input to the second step. In the second step, a hierarchical algorithm based on the log-likelihood distance measure is used to cluster the primary clusters that are the output of the previous step and place similar observations in the same cluster. To calculate the log-likelihood value, it is assumed that all variables are independent and that continuous variables follow a normal distribution and the stratified variables follow a polynomial distribution. One feature of the hierarchical clustering method is that it generates several solutions with a different number of clusters in a single run, which allows the researcher to discover a wide range of solutions with a different number of clusters [26]. The Bayesian Information Criterion (BIC) is calculated for any potential number of clusters.

$$BIC(J) = -2LL(K) + r(k)\log(N) \qquad (1)$$

Where LL(k) is a log-likelihood function with K clusters, r(k) is the number of independent parameters, and N is the total number of records in the database.

Using the "change ratio" in the BIC, an initial estimate of the number of suitable clusters is made for each merge compared to the first merge. It is assumed that dBIC(J) be the difference between two models with j and (j+1) clusters (dBIC(J) = BIC(J) - BIC(J+1)). Then, the "change ratio" for model j will be equal to:

$$R_1(J) = \frac{dBIC(J)}{dBIC(1)} \tag{2}$$

If dBIC(1) <0, then the number of clusters is considered equal to 1 and the next step is omitted. Otherwise, the initial estimate for the number of clusters k is equal to the lowest number for which R_1(J) <0.04.

In the second step, the largest relative increase in the distance between two close clusters is detected for different solutions, and it is used to accurately make the initial estimate of the number of clusters. This is expressed as follows:

$$R_2(k) = \frac{R_1(k)}{R_1(k+1)} \tag{3}$$

Among the R_2 values, two higher values are determined. If the maximum value is more than 1.15 times the second value, then the model with the highest R_2 value is selected as the optimal number of clusters. Otherwise, among the two models with the highest R_2 value, the model with the highest number of clusters is selected as the optimal model [27].

2- 2- Data Imbalance Treatment

An imbalanced database is one where there is a significant difference between the numbers of samples belonging to different classes. If the modeling is done with unbalanced data, the accuracy of the minority class model can hardly be compared with that of the majority class model and this model is not correct [23, 24]. Resampling is one of the most widely used methods to solve the class imbalance problem [28]. It changes the class distribution of samples so that minority classes appear more prominently in the training dataset. The data balancing technique applies to the training data (70% of data).

The oversampling method was used to solve the data imbalance problem because the crash database used in this study has imbalanced classes and the percentage of accidents caused by fatigue and drowsiness are 8%, 2.9%, and 2.35% and the percentage of driver fatalities is 0.1%, 0.57% and 0.32% from Tehran, Fars, and Mazandaran provinces. The oversampling method acts in a way that if the number of classes were C and sort the classes in ascending order according to the sample size (e.g., $n_1 \leq n_2 \leq$ □ $< n_C$) where n_C is the sample size of the majority class, in the oversampling method, the class 1 to C-1 samples randomly reach the number of majority class (n_C) samples by the substitution. It should be noted that during the sampling, different ratios can be tested because the number of samples in each class need not be the same.

2- 3- Classification and Regression Tree (CART)

In the current study, a common method in data mining called classification and regression tree (CART) was used to identify the factors affecting the occurrence of fatigue and drowsiness accidents in each cluster. This method was developed by Breiman et al. (1984) [29]. Forming a decision tree with binary divisions is also designed for quantitative and qualitative variables. Before the development of the CART model, as Breiman et al. (1984) [29] suggested, part of the data should be randomly assigned to training (70%) and another part to testing (30%).

The CART model starts by examining a variable in the first step by separating the data set at the root node. The training data is divided into two child nodes and each child node is then divided into two grandchild nodes, respectively. Also, the CART algorithm tries to divide each parent node into two homogeneous child nodes in which all the existing records belong to a specific category. The data in each sub-node is more homogeneous than the previous node. If the nodes created are homogeneous enough, the branching will stop and they will be the final node or leaf.

To separate each node into two subgroups, the Gini index is used, which is one of the most common indices, and is defined as follows:

$$\begin{aligned} P(j \mid m) &= \frac{p(j,m)}{p(m)}, \\ p(j,m) &= \frac{\pi(j) N_j(m)}{N_j} \\ p(m) &= \sum_{j=1}^{J} p(j,m) \end{aligned} \tag{4}$$

$$Gini(m) = 1 - \sum_{j=1}^{J} p^2(j \mid m) \tag{5}$$

Where J is the number of target variables, π(j) is the initial probability of category J, N_j(m) is the number of observations of category J in node m, N_j is the total number of observations in class J of the root node. p(j|m) is the probability that the observations of category J are in class m [30]. If all observations in a node are of the same category, Gini(m) is zero and represents the highest purity in the node. However, the maximum value of Gini(m) is obtained when all the observations are present in the same proportion in the node. The Gini index in each node is calculated for all variables and the variable is selected as the splitting variable when the minimum Gini value is obtained for the variable.

An important feature of this algorithm is that it tries to minimize the tree size while improving the quality of the decision. For this purpose, after constructing the tree, the tree pruning operation is performed using the misclassification cost method, which is calculated according to the following equation:

$$\text{Misclassification cost} = \sum_{t=1}^{T} p(t)[1-\sum_{j=1}^{J} p^2(j\,|\,t)] \tag{6}$$

Where p(t) is the share of observations in the final node t of the total observations and T is the number of final nodes [30]. One of the most important advantages of the decision and regression tree is determining the importance of variables.

The importance of variable X with h levels that intervene in the model is defined by the following equation:

$$VIMX = \sum_{i=1}^{h} \frac{nx_i}{n}(I(C\,|\,X = x_i) - (c)) \tag{7}$$

Where C is the class variable (human factor), nx_i is the number of cases that $X=x_i$, n is the number of total cases, and I is the Gini index [31].

2- 4- Boosting

In recent years, much research has been proposed on the combination of classifiers to create a final classification. The classifier combination methods are called collective learning methods, which are usually more accurate than primary classifiers.

This study used the combination of CART with boosting. Boosting is a hybrid meta-algorithm in the field of machine learning that is used to reduce imbalances and variances. In particular, one of the problems of trees is the high variance. The main reason for this high variance is the hierarchical nature of the process, where an error in the top-down division is propagated in all of the bottom divisions. Therefore, combining decision trees as a weak classifier with boosting results is one of the best classifiers [32].

3- Case study

3- 1- Crash data

The data used in the current study were collected from the traffic police accident database of provinces Tehran, Fars, and Mazandaran of Iran for a period of 7 years (2011-2018). The data was recorded by a traffic police officer at the scene of the accidents.

This study aimed to identify the factors affecting the occurrence of fatigue and drowsiness accident. In the current study, the dependent variable is a human factor, which is a binary variable (fatigue and drowsiness, not fatigue and drowsiness). Finally, after clearing the database, 647,185 crash data of Tehran province, 137,455 crash data of Fars province, and 74,738 crash data of Mazandaran province remained. Based on the objective of this study, 19 independent variables were extracted that include general characteristics about each accident (such as type of collision, type of vehicle, type of road, the direction of road traffic, local land use, type of traffic control, weather conditions, time of day, etc.) and information about the drivers involved (such as driver age and gender, type of license, the status of a seat belt or helmet use (restraint use), and severity of driver injury). Table 1 presents the variables and subsets of each variable used in the study by each province.

3- 2- Two-step Clustering

In the first step, the accident clustering was performed using all the variables presented in Table 1. The optimal cluster number of Tehran, Fars, and Mazandaran provinces was k = 4 with the silhouette coefficient of 0.7. If the silhouette coefficient for a clustering analysis ranges from 0.51 to 0.7, it can be stated that the algorithm was able to discover a "good" cluster structure among the data [33]. The clusters were characterized by driver gender, area type, and land use. The frequency ratio of each subcategory of variables in all four clusters by province can be seen in Table 2. In this regard, the clusters were named as follows:

The clustering of Tehran province in clusters 1 and 2 is similar to clusters 1 and 2 of Fars and Mazandaran provinces. However, in cluster 3 of Tehran province, all accidents of male drivers occur on suburban roads. Therefore, this cluster was designated as "accidents with male drivers on suburban roads". Also, in cluster 4 of Tehran province, 100% of accidents occurred in non-residential land uses, and 100% of accidents also occurred in urban areas, in addition, the gender of 100% of drivers is male. Therefore, this cluster was designated as "accidents with male drivers in non-residential urban areas".

3- 3- Data Imbalance Treatment

After grouping the training set into four homogeneous clusters, the data balancing was performed using the oversampling method. In the first step, the accidents were balanced with the human factor of fatigue and drowsiness, and in the next step, the drivers' injuries were also balanced.

3- 4- CART method

Finally, after grouping and balancing the training set, the most important independent variables of each cluster were identified using the CART model. Fig. 2 shows the percentage of fatigue and drowsiness accidents in

The clusters with the same title are in the provinces. As can be seen, the percentage of fatigue and drowsiness accidents in female drivers cluster and also in the cluster of accidents with male drivers in residential urban areas in Tehran province is higher than Fars province that is higher than Mazandaran province (Tehran> Fars> Mazandaran). Besides, the percentage of fatigue and drowsiness accidents in the cluster

Table 1. Variable description.(Continude)

Variable	Levels	Frequency %		
		Tehran	Fars	Mazandaran
Traffic control	Other device	11.18	4.42	0.64
	No control	21.81	27.91	50.8
Collision type	Unknown	29.88	42.21	21.66
	Fixed object collision	3.81	6.31	6.22
	Collision with motorcycle	21.52	28.22	21.13
	Two vehicle collision	73.31	57.76	67.7
	Running off	2.8	2.8	1.63
	Overturning	4.91	4.91	3.31
Lighting condition	Day light	70.97	66.9	66.52
	Dark	26.19	30.16	29.02
	Dusk/dawn	2.84	2.94	4.46
Time-of-the-day	24-02	4.11	4.09	3.85
	02-04	1.68	1.77	2.12
	04-06	1.09	1.03	1.21
	06-08	5.11	4.73	4.06
	08-10	8.87	8.45	7.58
	10-12	11.22	12.3	10.9
	12-14	12.71	13.65	14.02
	14-16	13.85	11.07	11.96
	16-18	13.25	11.26	13.15
	18-20	12.08	12.32	13.52
	20-22	8.89	10.76	10.07
	22-24	7.42	8.55	7.55
Day-of-the-week	Saturday	14.34	14.71	14.17
	Sunday	14.39	14.32	13.34
	Monday	14.44	13.96	13.38
	Tuesday	14.51	14.13	13.47
	Wednesday	14.92	14.59	14.81
	Thursday	15.21	15.52	16.46
	Friday	12.19	12.78	14.36
Month-of-year	April	6.32	8.79	10.12
	May	8.04	9.02	9.1
	June	8.43	8.7	9.34
	July	8.51	8.6	9.06
	August	8.9	9.12	8.88
	September	8.89	9.57	9.44
	October	8.73	8.83	7.85
	November	8.55	8.1	7.69
	December	8.29	7.59	7.53
	January	8.3	7.83	7.3
	February	8.5	7.43	7.28
	March	8.54	6.43	6.4

Table 1. Variable description.

Variable	Levels	Frequency %		
		Tehran	Fars	Mazandaran
Drivers' Gender	Male	88.29	91.26	92.84
	Female	11.71	8.74	7.16
Drivers' Age	<25	14.71	19.06	15.27
	25-44	60.48	59.79	61.4
	>44	24.81	21.15	23.33
License type	Unlicensed	99.36	95.85	97.51
	Licensed	0.64	4.15	2.49
Vehicle type	Auto	74.24	71.27	70.05
	Pick	6.42	7.43	8.94
	Truck/Light Truck	8.39	8.58	10.69
	Motorcycle	10.95	12.72	10.22
Restraint Use	Used	11	8.13	9.64
	Not used	4.49	7.62	5.25
	Unknown	84.51	84.25	85.11
Weather condition	Cloudy	1.55	0.55	6.4
	Rain	2.86	2.08	8.22
	Snow	0.51	0.17	0.44
	Clear	94.87	94.02	84.21
	Fog	0.21	0.19	0.63
Terrain	Rolling	0.35	1.21	0.7
	Level	98.81	97.87	92.73
	Mountainous	0.84	0.92	6.56
Road type	Freeway	1.7	1.2	0.8
	Highway	28.1	3.7	2.5
	Major Road	7.7	18.2	37.8
	Minor road	1	5.5	10.01
	Major street	55.2	65.6	38.8
	Minor street	6.1	4.3	6.56
	Rural road	0.1	0.1	1.97
	Direct road	0.1	0.5	0.55
Shoulder type	Paved	2.33	9.01	10.04
	Stabilized gravel	2.3	10.86	29.27
	None	95.37	80.13	60.69
Road Configuration	Two-Way, Not Divided	20.84	34.91	39.14
	Two-Way, Divided	59.01	49.41	52.42
	One-Way	20.16	15.68	8.44
Land use	Non-Residential	21.98	22.07	33.17
	Residential	78.02	77.93	66.84
Area type	Suburban	7.09	23.57	46.96
	Urban	92.91	76.43	53.4
Traffic control	Police	20.48	18.85	14.93
	Stop sign	0.99	1.7	3.82
	Right-of-way sign	0.48	0.69	4.46
	Traffic signal	15.19	4.22	3.7

Table 2. Summary of unilabiate distributions for the variables in each cluster

Variable	Level	Cluster-1			Cluster-2			Cluster-3			Cluster-4		
		TEH[1]	FRS[2]	MZN[3]	TEH	FRS	MZN	TEH	FRS	MZN	TEH	FRS	MZN
Driver gender	Female	0	0	0	0	0	100	100	0	0	0	100	0
	Male	100	100	100	100	100	0	0	100	100	100	0	100
Area type	Urban	0	100	40	100	25.7	58.3	96.6	0	0	100	88.2	100
	Rural	100	0	60	0	74.3	41.7	3.4	100	100	0	11.8	0
Land use	Residential	24.7	100	0	0	0	68.2	83.6	100	100	100	85.2	100
	Non-Residential	75.3	0	100	100	100	31.8	16.4	0	0	0	14.8	0

[1] THE = Tehran, [2] FRS = Fars, [3] MZN = Mazandaran.

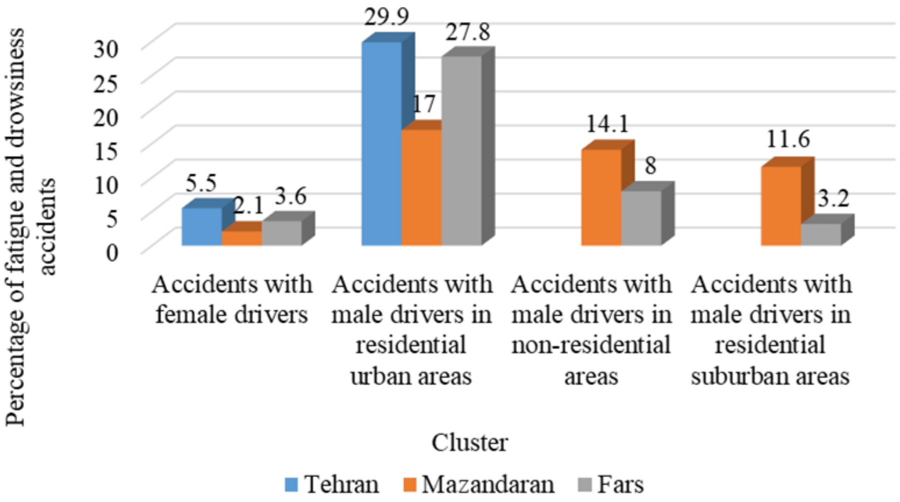

Fig. 2. The percentage of fatigue and drowsiness accidents in the clusters with the same title in the provinces.

of accidents with male drivers in non-residential areas and residential suburban areas cluster in Mazandaran province is higher than Fars province. In Mazandaran and Fars provinces, the frequency percentage of fatigue and drowsiness accidents in residential areas of urban and suburban areas is higher than the frequency percentage of fatigue and drowsiness accidents in non-residential areas. It can be concluded that the incidence of fatigue and drowsiness accidents in residential areas are higher than that in non-residential areas. As shown in Table 3 several important variables increased the incidence of fatigue and drowsiness accidents in each cluster of provinces and the first two variables are more important. First, the most important independent variables are described based on the decision trees of each cluster.

Cluster 1 (accidents with female drivers):

In Tehran province, more than 90% of fatigue and drowsiness accidents have occurred on freeways.

Also, the probability of fatigue and drowsiness accidents on other roads increases between 12 midnight and 6 am. In Fars province, more than 90% of fatigue and drowsiness accidents have occurred on freeways, highways, and side roads, and the probability of fatigue and drowsiness accidents on other roads without shoulders is low. In Mazandaran province, the occurrence of fatigue and drowsiness accidents has intensified on one-way roads. Also, on the roads with two-way divided and two-way undivided directions, the fatigue and drowsiness accidents in residential land use are revealed to be more likely than non-residential land use.

Table 3. The relative importance of the most important variables in Tehran province.

Cluster No.	Cluster description	Most important variables in Mazandaran	Most important variables in Fars
1	Accidents with female drivers	Road configuration = 0.19 Land use = 0.19 Traffic control = 0.13	Road type = 0.44 Shoulder type = 0.27
2	Accidents with male drivers in residential urban areas	Time of day= 0.36 Road type = 0.12	Time of day = 0.24 Road type = 0.22 Vehicle type = 0.18 Traffic of day
3	Accidents with male drivers in non-residential areas	Traffic control device = 0.32	Road Configuration = 0.22 Type of accident = 0.14 Time of day = 0.12
4	Accidents with male drivers in residential suburban areas	Road type = 0.3	Road type = 0.45

Table 4. The relative importance of the most important variables

Cluster No.	Cluster description	Most important variables in Tehran
1	Accidents with female drivers	Road type = 0.3 Time of day = 0.24 Month = 0.22 Type of accident=0.19
2	Accidents with male drivers in residential urban areas	Time of day = 0.77
3	Accidents with male drivers in non-residential urban areas	Time of day = 0.51 Vehicle type = 0.35
4	Accidents with male drivers on suburban roads	Road type = 0.56 Time of day = 0.31

Cluster 2 (accidents with male drivers in residential urban areas):

In Tehran and Mazandaran provinces, the probability of fatigue and drowsiness accidents is high between 12 midnight to 8 am. In Mazandaran province, at other times of the day, the fatigue and drowsiness accidents on freeways, main roads, and rural roads are more likely than other roads. Also, in Fars province, more than 90% of fatigue and drowsiness accidents have occurred in 4-6 am, and in other hours of the day, the probability of fatigue and drowsiness accidents on freeways, highways, side roads, and straight roads is 88.2%.

Cluster 3 (accidents with male drivers in residential suburban areas):

Road type is the most important variable in Mazandaran and Fars provinces. In Mazandaran province, the probability of fatigue and drowsiness accidents on freeways, highways, and straight roads is 90%. Also, in Fars province, the probability of fatigue and drowsiness accidents on freeways is over 90%. In the suburban roads of Tehran province, the probability of fatigue and drowsiness accidents on freeways and straight roads is 71.3%, and on other roads, the occurrence of fatigue and drowsiness accidents in the period from 10 pm to 10 am is more likely than other hours of the day.

Cluster 4 (accidents with male drivers in non-residential areas):

In Mazandaran province, the probability of fatigue and drowsiness accidents on roads controlled by right-of-way and stop signs is 78%. Also, in Fars province, the occurrence of fatigue and drowsiness accidents on the roads with one-way traffic is more probable than the roads with two-way divided traffic and two-way undivided traffic. Moreover, on the roads with two-way divided and two-way undivided traffic, the occurrence of fatigue and drowsiness accidents in a single-vehicle collision (vehicle with a fixed object, run-off-road, overturning, and falling) is more likely than the vehicle collision with motorcycle and with another vehicle. Besides, in the accidents with male drivers in non-residential urban areas of Tehran province, the probability of fatigue and drowsiness accidents in motorcyclists is low, and the incidence of fatigue and drowsiness accidents also increase between 12 midnight and 10 am.

The variables of road type and time of day were common

Table 5. Prediction accuracy of models by treatments.

Province	Male drivers in non-residential urban areas	Male drivers in residential urban areas	Male drivers in residential suburban areas
Tehran	12 PM- 10 AM	12 PM- 8 AM	No common variable
Mazandaran	No common variable	Freeway, Major road, Rural road	Freeway, Highway, Direct road
Fars	No common variable	Freeway, Highway, Minor road, Direct road	Freeway

Table 6. The difference between the variables of road type and time of day in the common cluster in-province.

Cluster description	Tehran	Fars	Mazandaran
Accidents with female drivers	Freeway	Freeway, Highway, Minor road	No common variable
Accidents with male drivers in residential urban areas	10 PM-6 AM	4-6 AM	10 PM-8 AM
Accidents with male drivers in residential suburban areas	No common variable	Freeway	Freeway, Highway, Direct road

in different clusters of provinces. The difference between the two variables in the common clusters inter-province and in-province can be seen in Table 5 and 6.

Driving near the dawn has been identified as an important factor of fatigue and drowsiness accidents [9, 34]. The current study confirmed this finding and also showed different times based on the province, land use, and area type. For instance, in the urban of Tehran province, fatigue and drowsiness accidents have increased at 10 pm -6 am in the residential areas, but these kinds of crashes increased from midnight to 10 am in the non-residential areas. Whereas, in the cluster of urban residential areas, fatigue and drowsiness accidents have increased at 4-6 am in Fars province but 10 pm to 8 am in Mazandaran province. Since the pattern of the time of occurrence fatigue and drowsiness accidents were different among clusters in each province, future studies might be considered time-of-day effects in urban, rural, residential, and non-residential areas among several provinces. Moreover, the results of the current study showed that fatigue and drowsiness accidents are reduced on the roads without shoulders, which may be because the risk of fatigue and drowsiness accidents increases when the driver feels more comfortable driving. The current work, like others studies [8-10, 35], demonstrated that the probability of fatigue and drowsiness is high in the freeways. However, the roads such as main road, straight road, side road, and rural road in some clusters were among the important factors of fatigue and drowsiness accidents. Furthermore, the results showed that the fatigue and drowsiness accidents in residential areas are higher than that in non-residential areas. All these findings indicated that fatigue and drowsiness accidents are very common in low-speed roads and residential areas. This indicates that in addition to freeways, motorways, and non-residential areas, low-speed roads and residential areas need special future attention in search of identifying factors that play a vital role in causing fatigue and drowsiness accidents.

3- 5- Boosting

The combination of the CART model with boosting algorithm and oversampling is applied in each cluster of three provinces. According to Table7 , in addition to the high overall accuracy of the models, the majority class (not-fatigue and drowsiness accidents) and the minority class (fatigue and drowsiness accidents) have high accuracy when combining the CART model with the oversampling method. Also, combining the boosting algorithm with the oversampling method produces better results. This study shows that the boosting method can be effective in handling unbalanced data when combined with the oversampling method.

4- Conclusion

The present study used a four-step data mining framework. First, the crash data were divided into homogeneous groups using the clustering analysis based on three variables: land use, type of area, and driver gender. Next, the balancing technique was used to balance the data of accidents caused

Table 7. The difference between the variables of road type and time of day in the common cluster inter-province.

Province	Cluster description	Over-sampling			Over-sampling + Boosting		
		Accuracy %			Accuracy %		
		Overall	majority class	minority class	Overall	majority class	minority class
Mazandaran	Accidents with female drivers	78.26	83.84	69.6	93.74	96.52	89.41
	Accidents with male drivers in residential urban areas	74.35	75.7	73.24	84.21	82.23	86.27
	Accidents with male drivers in residential suburban areas	77.96	73.12	85.8	86.5	85.55	87.53
	Accidents with male drivers in non-residential areas	79.14	77.3	81.4	88.94	89.61	88.17
Fars	Accidents with female drivers	90.75	91.2	90.2	98.24	99.36	96.84
	Accidents with male drivers in residential urban areas	76.14	71.8	84.3	82.38	79.72	86.13
	Accidents with male drivers in residential suburban areas	84.29	79.54	90.8	87.62	87.37	87.86
	Accidents with male drivers in non-residential areas	76.28	76.3	76.3	78.59	78.14	79.19
Tehran	Accidents with female drivers	81.06	78.5	84.5	81.22	82.11	80.31
	Accidents with male drivers in residential urban areas	75.9	70.4	85.7	76.65	74.03	80
	Accidents with male drivers in non-residential urban areas	75	74.76	85.5	81.65	82.1	81.2
	Accidents with male drivers on suburban roads	76.39	83.3	71.6	83.07	82.1	84.15

by fatigue and drowsiness and the severity of driver injuries. Finally, the classification and regression tree algorithm was employed to identify the factors that cause fatigue and drowsiness accidents. In addition, the boosting algorithm was utilized to increase the modeling accuracy.

Some of the most important factors were common among the clusters. However, some other important factors were specific to each cluster of provinces. This may indicate the different nature of fatigue and drowsiness accidents in each cluster of provinces. The most important factors affecting the fatigue and drowsiness accidents based on the CART tree in each cluster are as follows:

The common variable in residential urban areas in all three provinces was the time of day. In Tehran province, the fatigue and drowsiness accidents increase only in the period from 12 midnight to 8 am, but in Mazandaran province, in addition to the period from 10 pm to 8 am, the freeway, main road, and rural road increase the incidence of fatigue and drowsiness accidents. However, in Fars province, in 4-6 am and freeway, highway, minor road, and straight road increase the incidence of fatigue and drowsiness accidents.

Road type was the only important variable in the residential suburban areas of Mazandaran and Fars provinces. Freeway, highway and straight road in Mazandaran province and freeway in Fars province increase the incidence of fatigue and drowsiness accidents. Also, in the suburban roads of Tehran province, in addition to the freeway and straight road, the period from 10 pm to 10 increase the incidence of fatigue and drowsiness accidents.

In the non-residential areas of Mazandaran province, the incidence of fatigue and drowsiness accidents increases only by controlling the right-of-way and stop signs, but in Fars province, it increases on one-way roads and in a vehicle collision with a fixed object, run-off-road, overturning, and fall accidents. Also, in the non-residential urban areas of Tehran province, the period from 12 midnight to 10 am in all types of vehicles (except motorcycles) increases the incidence of fatigue and drowsiness accidents.

The common variable in accidents with female drivers in Tehran and Fars provinces was road type. Freeway in Tehran province and freeway, highway, and minor road in Fars province increase the incidence of fatigue and drowsiness accidents. In addition, the incidence of fatigue and drowsiness accidents increases in Tehran province between 12 midnight and 6 am, but it increases in Fars province on roads with shoulders. In Mazandaran province, however, the one-way roads and residential land use increase the incidence of fatigue and drowsiness accidents.

In addition, the incidence of fatigue and drowsiness accidents in residential areas was higher than in non-residential areas in Fars and Mazandaran provinces.

The results of this study indicate that the combined use of clustering, balancing, CART algorithm, and boosting can be useful to obtain a holistic view over the fatigue and drowsiness accidents and to prioritize the safety countermeasures.

References

[1] H. Zakeri, K. Kadkhodazadeh, Review of Contributing Factors in Road Traffic Accidents in Iran, 284060423X, 2015.

[2] M. Rad, A.L. Martiniuk, A. Ansari-Moghaddam, M. Mohammadi, F. Rashedi, A. Ghasemi, The pattern of road traffic crashes in South East Iran, Global journal of health science, 8(9) (2016) 149.

[3] W.B. Verwey, D.M. Zaidel, Preventing drowsiness accidents by an alertness maintenance device, Accident Analysis & Prevention, 31(3) (1999) 199-211.

[4] T. Akerstedt, Consensus statement: fatigue and accidents in transport operations, Journal of sleep research, 9(4) (2000) 395-395.

[5] A.W. MacLean, D.R. Davies, K. Thiele, The hazards and prevention of driving while sleepy, Sleep medicine reviews, 7(6) (2003) 507-521.

[6] D.J. Beirness, H.M. Simpson, K. Desmond, The road safety monitor 2004: Drowsy driving, 2005.

[7] P. Philip, P. Sagaspe, E. Lagarde, D. Leger, M.M. Ohayon, B. Bioulac, J. Boussuge, J. Taillard, Sleep disorders and accidental risk in a large group of regular registered highway drivers, Sleep medicine, 11(10) (2010) 973-979.

[8] G. Zhang, K.K. Yau, X. Zhang, Y. Li, Traffic accidents involving fatigue driving and their extent of casualties, Accident Analysis & Prevention, 87 (2016) 34-42.

[9] J. Connor, R. Norton, S. Ameratunga, E. Robinson, I. Civil, R. Dunn, J. Bailey, R. Jackson, Driver sleepiness and risk of serious injury to car occupants: population-based case-control study, Bmj, 324(7346) (2002) 1125.

[10] P. Philip, C. Chaufton, L. Orriols, E. Lagarde, E. Amoros, B. Laumon, T. Akerstedt, J. Taillard, P. Sagaspe, Complaints of poor sleep and risk of traffic accidents: a population-based case-control study, PloS one, 9(12) (2014) e114102.

[11] S. Niu, G. Li, Fatigue driving prediction on commercial dangerous goods truck using location data: the relationship between fatigue driving and driving environment, Journal of advanced transportation, 2020 (2020).

[12] N. Haworth, G. Rechnitzer, Description of fatal crashes involving various causal variables, 1993.

[13] G. Ryan, J. Spittle, Truck crashes in Western Australia, in: Road Safety Research and Enforcement Conference, Melbourne, November 1995.

[14] K. Armstrong, A.J. Filtness, C.N. Watling, P. Barraclough, N. Haworth, Efficacy of proxy definitions for identification of fatigue/sleep-related crashes: An Australian evaluation, Transportation Research Part F: Traffic Psychology and Behaviour, 21 (2013) 242-252.

[15] A.J. Filtness, K.A. Armstrong, A. Watson, S.S. Smith, Sleep-related vehicle crashes on low speed roads, Accident Analysis & Prevention, 99 (2017) 279-286.

[16] C. Gnardellis, G. Tzamalouka, M. Papadakaki, J.E. Chliaoutakis, An investigation of the effect of sleepiness, drowsy driving, and lifestyle on vehicle crashes, Transportation research part F: traffic psychology and behavior, 11(4) (2008) 270-281.

[17] X. Zhang, X. Wang, X. Yang, C. Xu, X. Zhu, J. Wei, Driver drowsiness detection using mixed-effect ordered logit model considering time cumulative effect, Analytic methods in accident research, 26 (2020) 100114.

[18] S. Soares, T. Monteiro, A. Lobo, A. Couto, L. Cunha, S. Ferreira, Analyzing driver drowsiness: From causes to effects, Sustainability, 12(5) (2020) 1971.

[19] M. Yu, C. Zheng, C. Ma, J. Shen, The temporal stability of factors affecting driver injury severity in run-off-road crashes: A random parameters ordered probit model with heterogeneity in the means approach, Accident Analysis & Prevention, 144 (2020) 105677.

[20] L.-Y. Chang, H.-W. Wang, Analysis of traffic injury severity: An application of non-parametric classification tree techniques, Accident Analysis & Prevention, 38(5) (2006) 1019-1027.

[21] H. Sharma, S. Kumar, A survey on decision tree algorithms of classification in data mining, International Journal of Science and Research (IJSR), 5(4) (2016) 2094-2097.

[22] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees. Chapman & Hall/CRC, (1998).

[23] H. Jeong, Y. Jang, P.J. Bowman, N. Masoud, Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data, Accident Analysis & Prevention, 120 (2018) 250-261.

[24] R.O. Mujalli, G. López, L. Garach, Bayes classifiers for imbalanced traffic accidents datasets, Accident Analysis & Prevention, 88 (2016) 37-51.

[25] T. Chiu, D. Fang, J. Chen, Y. Wang, C. Jeris, A robust and scalable clustering algorithm for mixed type attributes in large database environment, in: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, 2001, pp. 263-268.

[26] M. Norusis, P. Statistics, Advanced statistical procedures companion, p. 152ff, (2005).

[27] I.R.A. Hamid, J.H. Abawajy, An approach for profiling phishing activities, Computers & Security, 45 (2014) 27-41.

[28] D. Thammasiri, D. Delen, P. Meesad, N. Kasap, A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, Expert Systems with Applications, 41(2) (2014) 321-330

[29] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees–crc press, Boca Raton, Florida, (1984).

[30] Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. Data mining: A knowledge discovery approach. New York: Springer Verlag, 2007.

[31] A. Pande, M. Abdel-Aty, Assessment of freeway traffic parameters leading to lane-change related collisions, Accident Analysis & Prevention, 38(5) (2006) 936-948.

[32] J.H. Friedman, The elements of statistical learning: Data mining, inference, and prediction, springer open, 2017.

[33] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20 (1987) 53-65.

[34] J. Horne, L. Reyner, Driver sleepiness, Journal of sleep research, 4 (1995) 23-29.

[35] P.-H. Ting, J.-R. Hwang, J.-L. Doong, M.-C. Jeng, Driver fatigue and highway driving: A simulator study, Physiology & behavior, 94(3) (2008) 448-453.

HOW TO CITE THIS ARTICLE

A. Tavakoli Kashani, M. Rakhshani Moghadam, S. Amirifar, A Data Mining Framework to Identify Important Factors of Fatigue and Drowsiness Accidents, AUT J. Civil Eng., 5(3) (2021) 421-432.

DOI: 10.22060/ajce.2021.19386.5728