IJTE, Vol. 11, No. 1, Summer 2023, pp. 1285-1300

Research Paper

A Data Mining Approach to Gain Insight into Traffic Violations of Young Drivers in Developing Countries

Vahid Khalifeh[1,*], Navid Nadimi[2], Amir Mohammadian Amiri[3]

Received: 2022/03/22 *Accepted: 2022/08/30*

Abstract

In developing countries, the population of younger adults is relatively higher. In addition, the frequency of traffic violations, committed by young drivers, is considerable. Consequently, annually a large portion of road crashes is recorded among this age group. This paper aims to study the traffic rule violations of young drivers in Iran. Focusing on the behavior of young drivers and understanding the mechanisms that affect the occurrence of violations among this group of drivers can be helpful to promote traffic safety. For this purpose, 567 drivers in the range of 18 to 40 years old have been studied. Then, different data mining approaches such as descriptive analysis, correlation analysis, multinomial logistic regression (MLR), and Random Forest (RF) were used. The main contribution of this paper is using different analytic methods to gain insight into traffic violations of young drivers and to propose potential countermeasures to decrease this issue. Results indicated that driving over speed limits, red-light running, and angry driving are the most frequent violations. The frequency of using mobile phone while driving, as a source of distraction, has been found to be highly correlated with other violations. As the frequency of previous traffic fines, the number of days with access to private cars, and the frequency of previous crashes increase, more diverse types of violations with high frequencies are expected in the future. In addition, the frequency of risky violations was found to be higher among men and those with lower education levels.

Keywords: Safety, Violation, Young driver, Crash, Data mining

[*] Corresponding author. E-mail: vahid.khalifeh@sirjantech.ac.ir

[1] Assistant Professor, Department of Civil Engineering, Sirjan University of Technology, Sirjan, Iran

[2] Assistant Professor, Faculty of Engineering, Shahid Bahonar University, Kerman, Iran

[3] Postdoctoral Researcher, McMaster Institute for Transportation & Logistics (MITL), McMaster University, Hamilton, Canada

1. Introduction

Road crashes annually kill many people throughout the world. According to the World Health Organization (WHO) report, 1.35 million people are annually killed in road crashes [WHO, 2018] More than 90 percent of mortalities caused by road crashes were from low- or middle-income countries, although they own less than 60 percent of the vehicles [Kerwin & Bushman, 2020] Studies indicate that traffic crashes have been one of the three leading causes of death among younger adults [Trivedi & Rawal, 2011]. Alver et al. [2014] surveyed crashes among young drivers aged 18 to 19. They found that almost 24 percent of the drivers have had at least one crash in the last three years [Alver et al., 2014]. De Melo et al. [2017] reported that the rate of hospitalization among young people (15-29 years old) is more than twice higher than other age groups, due to severe injuries [de Melo et al., 2017]. Rahman et al. [2021] estimated that young drivers (15-24 years) constitute more than 20 percent of Louisiana's fatal traffic crashes in 2018 [Rahman et al., 2021]. According to Iranian Legal Medicine Organization statistics, about half of the road crash fatalities are people younger than 40 years old [Alimohammadi et al., 2020].

Aberrant driving behaviors can be one of the leading causes of crashes [Reason et al., 1990]. Aberrant behaviors generally consist of slips, lapses, mistakes, unintended violations, and deliberate violations [Reason et al., 1990]. Traffic violations are those deviations from necessary behaviors for safe driving and avoiding dangerous situations [Reason et al., 1990]. Deliberate violations are accompanied by more high-risk situations than other aberrant behaviors. Various violations can cause different degrees of risk to road users [Reason et al., 1990]. Traffic violations with a high risk of crash occurrence include but are not limited to red-light running, lack of seat belt use, speeding, mobile phone use, impaired driving, the wrong way entry, unofficial racing, showing hostility to other drivers, using vehicles with technical problems and illegal backing-up [Alver et al., 2014; Asadianfam et al., 2020; Scott-Parker & Oviedo-Trespalacios, 2017].

In Eskişehir, Turkey, according to the police reports, 5.4 percent of crashes were due to red-light violations [Karacasu & Er, 2011]. Shaaban and Pande [2018] stated that the red-light running had been correlated with the frequency and severity of traffic crashes [Shaaban & Pande, 2018]. Some studies have shown that seat belt violation has negative impact on the severity of crashes. On the other hand, fastening the seat belt reduces the risk of death up to 48 percent [Hatfield et al., 2014]. Numerous studies have been conducted on speed limit violations. The results of a survey showed that driving at high speeds increases the risk of crash occurrence by an average of 60 percent [Williams et al., 2006]. Castillo-Manzano et al. [2019] studied the impact of driving speed limits on traffic crash fatalities [Castillo-Manzano et al., 2019]. Rossi et al. [2020] showed that speed reduction not only reduces traffic crashes and severity but also has very significant impact on human health [Rossi et al., 2020]. Mobile phone use and impaired driving are other violations that play a serious role in both the occurrence and severity of traffic crashes. According to previous studies, use of alcohol and drugs before driving can cause health problems and thus increase improper driving performance [Font-Ribera et al., 2013; Simonsen et al., 2018]. Sullman and Baas [2004] showed that using mobile phone while driving could increase the risk of accidents up to 9 times higher in New Zealand [Sullman & Baas, 2004]. Zhang et al. [2019] indicated that anger driving could also increase the risk of crash occurrence [T. Zhang et al., 2019]. The interaction between traffic accidents and risky driving of young drivers (16–25 years old) was studied using the hierarchical segmentation analysis via decision trees [Scott-Parker & Oviedo-Trespalacios, 2017]. It was found that mobile phone use and drinking alcohol had a greater impact on traffic crashes than other factors.

Some researchers have shown that the frequency of high-risk driving behaviors, such as speeding, ignoring traffic signs, illegal overtaking, distraction, not fastening the seat belt, and driving under the influence of alcohol is considerable among young drivers [Carter et al., 2014; Dunlop & Romer, 2010; Klauer et al., 2014]. Besides, crashes caused by impaired driving or mobile phone use are higher among young people [Erin Goodman et al., 2020; Lipovac et al., 2017]. Bener et al. [2006] stated that, compared to older adults, young drivers have more tendency to use mobile phone while driving [Bener et al., 2006]. Red-light running annually causes 165,000 injuries and 800 deaths in the United States, and most red-light runners are among young drivers (under the age of 30 years) [Mane & Pulugurtha, 2018].

Some studies have focused on the causes of traffic violations among different drivers. Emotions, exploration, inexperience, and recklessness are the most contributing elements among younger drivers in breaking the laws [Cestac et al., 2011]. Zhang et al. [2014] investigated driver, vehicle, road, and environmental risk factors to evaluate two traffic violations in China, speeding and drunk driving. They found that several factors have a more prominent role in these two violations, consisting of gender, private vehicle ownership, lack of street lighting at night, and poor visibility [G. Zhang et al., 2014]. Precht et al. [2017] showed that passenger presence, anger, and individual differences were the main variables contributing to traffic violations. The authors found that traffic violations increase due to emotional disorders like anger, frustration, surprise, and sadness. Moreover, some distractions, such as backing up for a lost object or touching the screen of a mobile phone can lead to more driving errors. In contrast, cognitive distractions did not have any effect on driving performance [Precht et al., 2017]. Alavi et al. [2017] represented the influence of personality characteristics, driving behavior, and mental illnesses on driving violation using logistic regression. They applied the Manchester Driving Behavior Questionnaire (MDBQ) for evaluating driving behaviors. The Big Five Personality Test and the Semi-Structural Interview (SADS) were utilized to investigate personality characteristics and to identify any mental diseases among drivers, respectively. They found that young drivers revealed more risky driving behaviors than other age groups. Moreover, their results showed that driver’s education did not have any impact on traffic violations. Besides, the results revealed that traffic violations could be directly related to depression and anxiety disorders [Alavi et al., 2017]. Ambo et al. [2021] assessed two major factors influencing the traffic violation using multinomial logit model in China, including congested driving and weather conditions. They found that taking these risk factors into account would decrease the frequency of traffic violations and enhance traffic safety [Ambo et al., 2021]. Hadji Hosseinlou et al. [2018] introduced the peripheral landscapes, number of interchanges, number of passing lanes, average speed as the factors, affecting traffic violations in freeways [Hadji Hosseinlou et al., 2018]. Rosenbloom and Perlman [2016] found that those drivers who travel alone commit more violations, despite their gender and age [Rosenbloom & Perlman, 2016].

The present paper intends to assess the traffic violation behavior of young drivers based on a self-reported survey in a developing country. For this purpose, a data mining approach is used to provide a complete view of the violations committed by young drivers. Based on the literature review, the main contributions of this paper consist of:

This study focuses on young drivers in developing countries and the main objective is to propose countermeasures to reduce dangerous traffic violations of this group to promote traffic safety.

In this study, risky traffic rule violations are regarded simultaneously. In addition, the relationships between their frequencies are compared.

Impact of previous negative experiences of young drivers such as traffic crashes and traffic fines on violation commitment are assessed.

Based on the frequency and diversity of committed violations a new indicator is developed and considered as the target variable. To the best of the authors' knowledge, no study has previously investigated the violation behavior of young drivers from different viewpoints.

Following, the utilized dataset and methodological approach are explained comprehensively. Subsequently, the outcomes are discussed in more detail, which is followed by a discussion and conclusions.

2. Methodology

Data mining is a relatively new field of science, which has attracted a great deal of attention in recent decades. Data mining tries to analyze data from different aspects and summarize a huge database with different attributes in the form of useful information. Different methods such as statistics, machine learning, artificial intelligence and database system can be used in a data mining process for anomaly detection, association rule learning, clustering, classification, regression, summarization, and sequential pattern mining [Fayyad et al., 1996]. In this paper, it is intended to use summarization, and classification to answer the questions outlined in Table 1.

Accessing to the violation reports is not an easy task, since a great number of violations have not been recorded in developing countries. Thus, this paper has relied on the self-reported violations of young drivers (in the range of 18 to 40 years old) provided by a questionnaire. Young drivers related to Tehran in Iran. Tehran is the capital of Iran a big city with more than 9 million population. Passenger cars are top-rated in Tehran, and approximately 40 percent of trips are done by them.

Table 1. Research questions

Number	Question
1	What is the frequency of committing each violation among young drivers?
2	What is the relationship between drivers' characteristics and traffic violations?
3	What is the relationship between the frequency of previous crashes and traffic violations?
4	What is the relationship between the frequency of previous fines and traffic violations?
5	What is the relationship between the frequencies of different violations?
6	What is the impact of drivers' characteristics, previous crashes, and fines on the diversity and frequency of drivers' violations?

The sample size to be surveyed is determined by Equation 1 (Johnson & Wichern, 2002).

$$n >= N\left[1+\frac{N-1}{pq}\left(\frac{d}{Z_{\frac{\alpha}{2}}}\right)^2\right]^{-1} \qquad (1)$$

Where;

n: Sample size, number of young drivers for data collection

N: Population size, number of total young drivers in Iran (about 30 million),

Z: 1.96 for 95% confidence level,

p,q: The quality characteristics which are to be measured. Where no previous experience exists then the value of p is taken as 0.5 and q=1-p=0.5,

d: the desired level of precision and is considered 0.5%.

Based on Equation 1, at least 385 young drivers must be surveyed. In this study, 567 questionnaires were filled, among which 515 cases were acceptable based on the controls considered during the survey. Data collection was conducted from March 2019 to November

2019 in Tehran (Iran) (before the COVID-19 outbreak). The questionnaire has been sent to different people via Email, Twitter, Instagram, WhatsApp, and other virtual methods. In the questionnaire, several questions have been designed to control if it has been filled with enough precision and concentration. In addition, the duration of filling the questionnaire has been recorded. Those responses, which have been done in a time less than the standard, were omitted. Meanwhile, it was tried to assure the respondents about keeping the results confidential, and that there will be no enforcement because of the self-reported traffic violations. Some rewards were also considered for completing this survey to increase the motivation to answer the questionnaires, precisely. Table 2 presents the distribution of the respondents' characteristics. In the questionnaire, first, drivers' characteristics consist of their age, gender, driving experience, education, income, and number of days with access to a private car during a week were asked.

Table 2. Driver's characteristics and related categories

Variable	Acronym	Categories	Acronym	Frequency
Gender	GE	Female	GE1	41.8
		Male	GE2	58.2
Age	AG	18-23	AG1	51.6
		23-28	AG2	24.1
		28-35	AG3	14
		35-40	AG4	10.3
Driving experience (years)	DE	<1	DE1	13
		1-5	DE2	48
		5-10	DE3	22
		>10	DE4	17
Education	EDU	B.Sc. (or B.Sc. student)	EDU1	65
		M.Sc. (or M.Sc. student)	EDU2	27
		Ph.D. (or Ph.D. student)	EDU3	8
Income (USD)	INC	<50	INC1	39
		50-100	INC2	21
		100-150	INC3	9
		150-200	INC4	9
		200-250	INC5	5
		>250	INC6	17
Number of days with private car availability during a week	PCA	1	PCA	4
		2		7
		3		11
		4		14
		5		18
		6		21
		7		25

It was also asked about the frequency and severity of traffic crashes (CF) in recent year as well as any traffic fines (TF) received during the past five years. At last, young drivers reported their monthly frequency of committing each violation (VF). Different methods are used to answer each research question. Each research question needs a specific tool.

Summarization is used to answer question 1 (in Table 1), and results are plotted in the form of bar charts for each of the violations. Violations such as driving over speed limit (SLV), driving without seat belt (SBV), disregarding the red-light (RLV), impaired driving (IDV), unofficial races (URV), driving with vehicle defects (VDV), using mobile phone while driving (MDV), showing hostility to other drivers (HDV), deliberately driving the wrong way down a deserted one-way street (WRV) and illegal backing-up (IBV) were considered in this study.

In questions 2 to 4 (in Table 1), it is intended to assess the impact of different attributes such as drivers' characteristics, TF, and CF on VF. The frequency of each violation has been reported by a 5-point rating scale (Never, Rarely, Occasionally, Frequently, and Very often), which can be considered as classification problem. Different methods can be used for classification problems, out of which multiple logistic regression (MLR) was selected in this study. Logistic regression is a statistical technique to show the effect of quantitative or qualitative variables on a multivariate dependent variable. Logistic regression analysis is similar to linear regression analysis, except that in contrast to linear regression, the dependent variable is a qualitative variable with two dimensions or more. If the dependent variable is multidimensional (three variables and more), it is called MLR [Peng et al., 2018]. In MLR models, the target variable is the commitment of each violation, and the input variables are driver's characteristics and previous background of drivers in relation to crashes and fines. For each violation, a separate MLR model is needed, thus ten unique models are made.

To answer question 5 (in Table 1), the correlation between the frequencies of occurrence of each pair of violations is calculated. The correlation between ordinal variables can be determined by the Kendall rank correlation coefficient. This is a statistic to measure the correlation between ordinal variables. This correlation indicator ranges between +1 and -1. The closer the value is to zero, the weaker the correlation is. Positive signs mean that as one variable increases the other one also increases and vice versa.

Question 6 (in Table 1) is considered to have an in-depth evaluation in relation to the frequency and diversity of the violation behavior of drivers. For this purpose, two new variables as the diversity of committed violations (DOV) and average frequency of violations (AFV) are introduced. By multiplying DOV and AFV, a new indicator is produced to specify the diversity and frequency of total violations for each driver. This variable was named diversity and average frequency of violations (DAFV). There are 10 dangerous violations in the questionnaire. Whenever the frequency of commitment of each violation is reported something other than never (based on 5-point scale), one type of violation commitment has been considered for that driver. Therefore, for a driver who has stated the commitment of violations as in Table 3, the DOV is five. For each scale also a number is considered, thus we have never=0, rarely=1, occasionally=2, frequently=3, very often=4). Finally, for the example in Table 3, AFV is 1 and DAVF is 5.

Table 3. A sample to show the calculation process for DAVF

Violation	SLV	SBV	RLV	IDV	URV	VDV	MDV	HDV	WRV	IBV
Frequency based on a 5-point scale	Rarely	Frequently	Occasionally	Never	Never	Never	Occasionally	Never	Occasionally	Never
Frequency by a number	1	3	2	0	0	0	2	0	2	0

The minimum and maximum values for DAFV are 0 and 40 (ten type of violations multiplied by 4 as very often for all of them), respectively. We have divided this range into three classes as low, moderate and high. Now the target variable is considered to be the DAFV class. Drivers' characteristics (Table 2), CF, and TF are the input variables. Random forest (RF) is used to understand the relationship between input attributes and the target. RF is a prediction method that works by generating many classifiers and finally aggregating the results [Liaw, A., & Wiener, 2002]. RF can be used for classification problems, and it consists of several tree-structures classifiers. The result of each tree is considered as a vote. The final class for the input attributes in RF is determined based on the most popular votes, regarding the combination of trees [Breiman, 2001]. In RF, each node is split using the best among a subset of predictors randomly chosen at that node. This method has performed very well compared to many other commonly-used classifiers and is more robust against overfitting [Breiman, 2001]. Implementing this method requires the determination of the models' parameters, including the number of trees to grow and the number of variables randomly sampled as candidates at each split. The reasons for selecting RF for this step, returns to its simplicity, reducing overfitting in decision trees, higher accuracy than decision trees, being compatible with categorical variables, and giving the importance of features.

Finally, based on the analyses done for Questions 1 to 6, a discussion is presented about the violation behavior of young drivers. Then, several countermeasures are suggested to control this aberrant behavior among young drivers in Iran.

3. Results

In this section, the results are presented based on the order displayed in Table 2.

3.1. Relative Frequencies

The relative frequency of committing each violation is depicted in Figure 1. The horizontal axis indicates the type of violation and the vertical axis is the relative frequency based on the previously introduced 5-point scale.

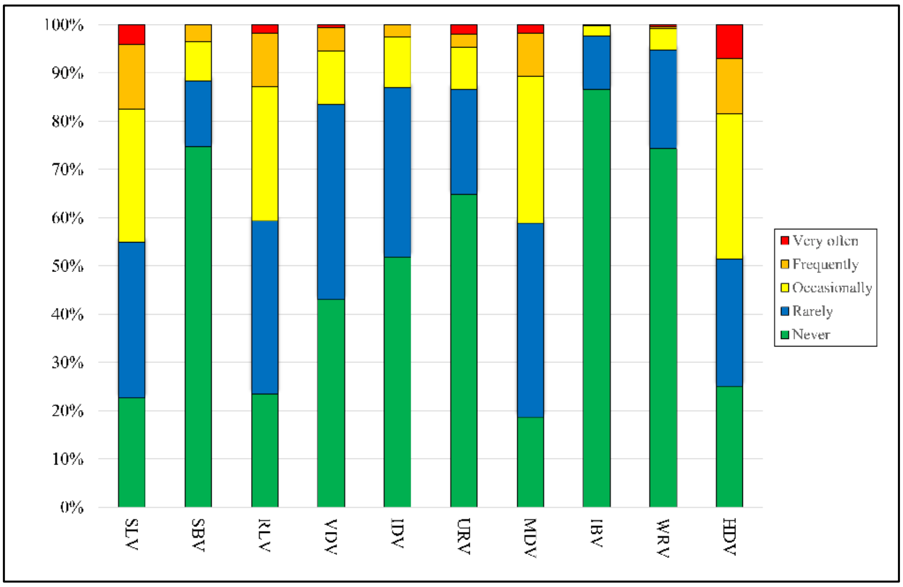

Figure 1. Relative frequencies of different violations

3.2. MLR Outputs

Table 4 indicates the MLR model fitting, model accuracy details and estimates of parameters for each violation.

Table 4. MLR model outputs

Violation	Model fitting information		Accuracy (percentage)					Overall	Parameter estimate									
	Chi-square	R^2	Class1	Class2	Class3	Class4	Class5		Class1		Class2		Class3		Class4		Class5	
									variable	B	variable	B	variable	B	variable	B	variable	B
SLV	248.46	0.41	0.54	0.48	0.39	0.29	0.14	0.43	PCA	-0.33	PCA	-0.29	-	-	-	-	Ref	
									CF	-0.65	-		-	-	-	-		
									TF	-1.65	TF	-0.67	-	-	-	-		
									GE1	2.17	GE1	1.55	-	-	-	-		
									EDU1	-17.27	EDU1	-18.13	EDU1	-17.99	EDU1	-17.96		
									EDU2	-17.42	EDU2	-17.60	EDU2	-17.22	-	-		
									INC4	2.22	INC2	1.94	INC2	2.21	INC2	2.32		
SBV	71.14	0.16	0.987	0.057	0.000	0.000	-	0.745	EDU1	-18.40	edu1	-17.57	edu1	-17.78	Ref		-	
									EDU2	-18.78	edu2	-18.95	PCA	0.28				
RLV	141.07	0.26	0.47	0.53	0.31	0.18	0.11	0.41	PCA	-0.42	GE1	1.66	-	-	-	-	Ref	
									EDU1	-15.34	EDU1	-16.26	EDU1	-16.84	EDU1	-17.29		
									EDU2	-13.90	EDU2	-14.35	EDU2	-14.57	-	-		
IDV	158.27	0.30	0.81	0.46	0.11	0.00	-	0.59	PCA	-0.88	PCA	-0.85	PCA	-939.00	Ref		-	
									TF	-0.63	-	-	-	-				
									EDU1	-15.33	EDU1	-16.04	EDU1	-15.95				
									EDU2	-14.48	EDU2	-14.68	-	-				
URV	185.98	0.35	0.94	0.18	0.04	0.21	0.20	0.66	TF	-1.10	TF	-0.58	-	-	-	-	Ref	
									DE3	4.24	-	-	DE3	4.70	-	-		
									INC4	17.67	INC4	18.44	INC4	16.23	-	-		
VDV	89.64	0.18	0.68	0.51	0.02	0.16	0.00	0.51	DE2	-17.01	DE2	-17.37	DE2	-17.44	DE2	-16.93	Ref	
									DE3	-16.92	DE3	-17.19	DE3	-17.40	-	-		
									EDU1	-16.51	EDU1	-17.24	EDU1	-17.64	-	-		
									INC2	18.49	INC2	19.39	INC2	18.88				
									INC3	18.11	INC3	19.06	INC3	18.70				
									INC4	17.56	INC4	18.01	INC4	17.59	-	-		
MDV	210.61	0.36	0.35	0.67	0.40	0.07	0.33	0.47	CF	-1.21	CF	-1.11	CF	-0.85	-	-	Ref	
									TF	-1.53	TF	-1.16	TF	-0.76	TF	-0.65		
									EDU1	-18.15	EDU1	-19.02	EDU1	-19.41	EDU1	-18.42		

									EDU2	-16.75	EDU2	-17.22	EDU2	-16.76	-		
HDV	132.25	0.24	0.40	0.41	0.55	0.10	0.14	0.39	CF	-0.94	CF	-0.73	CF	-0.63	CF	-0.56	Ref
									TF	-0.05	-	-	DE1	-2.12	-	-	
									GE1	1.55	GE1	1.93	GE1	1.53	-	-	
WRV	149.622	0.330	0.969	0.057	0.261	0.500	1.000	0.749	DE3	14.24	DE3	14.06	-	-	-	-	Ref
									EDU1	-10.42	EDU1	-10.96	-	-	-	-	
IBV	101.69	0.297	0.989	0.123	0.364	-	1.000	0.879	-	-	-	-	-	-	-	-	Ref

Note: Ref means reference category

3.3. Violations' Relationships

The relationships between the frequencies of committing each pair of violations are in Table 5. Only those coefficients, which were significant at 0.05 level, are presented.

Table 5. Kendall rank correlation coefficient

	SLV	SBV	RLV	IDV	URV	VDV	MDV	HDV	WRV	IBV
SLV	1	0.071	0.268**	0.300**	0.494**	0.222**	0.372**	0.281**	0.293**	0.235**
SBV	0.071	1	-0.019	0.059	0.154**	0.048	0.125**	0.112**	0.104*	0.222**
RLV	0.268**	-0.019	1	0.241**	0.270**	0.295**	0.258**	0.277**	0.207**	0.214**
IDV	0.300**	0.059	0.241**	1	0.316**	0.290**	0.365**	0.294**	0.257**	0.268**
URV	0.494**	0.154**	0.270**	0.316**	1	0.268**	0.378**	0.275**	0.274**	0.336**
VDV	0.222**	0.048	0.295**	0.290**	0.268**	1	0.264**	0.204**	0.292**	0.198**
MDV	0.372**	0.125**	0.258**	0.365**	0.378**	0.264**	1	0.277**	0.311**	0.233**
HDV	0.281**	0.112**	0.277**	0.294**	0.275**	0.204**	0.277**	1	0.270**	0.246**
WRV	0.293**	0.104*	0.207**	0.257**	0.274**	0.292**	0.311**	0.270**	1	0.293**
IBV	0.235**	0.222**	0.214**	0.268**	0.336**	0.198**	0.233**	0.246**	0.293**	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

3.4. RF Outputs

Figure 2 indicates the correlation between input variables and DAFV class. There are strong correlations between some input variables. However, RF is fairly insensitive to the multicollinearity problem. All of the correlations are positive, this means that each pair of variables move in the same direction.

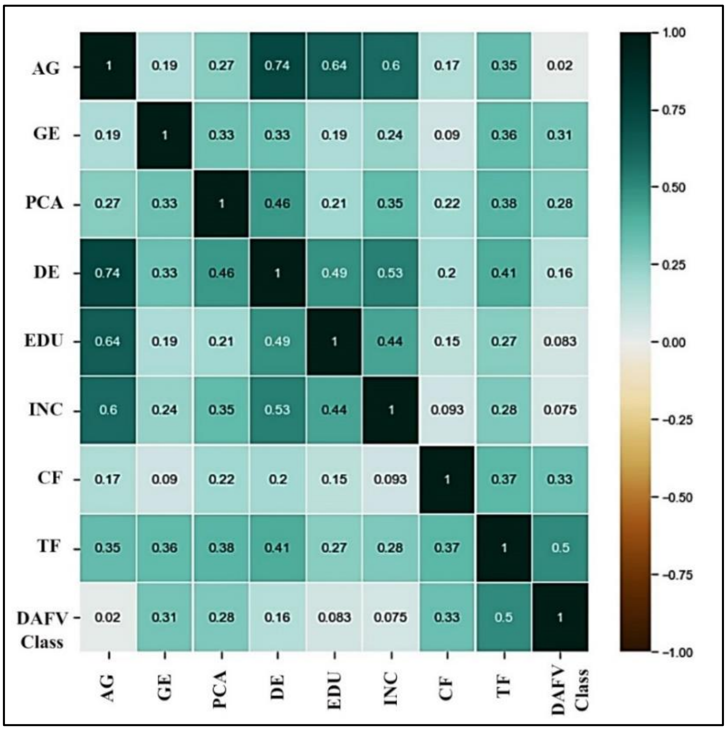

Figure 2. Correlation between input variables in the RF model

Figure 3 shows the features importance. Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. The feature importance in RF indicates which variables are more effective in decreasing the weighted impurity. In each node, a feature is used to decide about dividing the data set into separate sets. For each feature, it was calculated how on average, it decreases the impurity of the split.

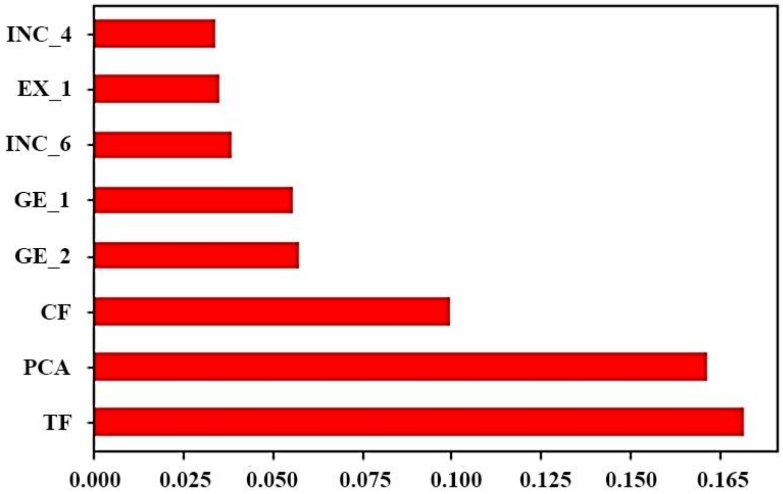

Figure 3. Feature importance in RF

To assess the RF model three criteria, consist of precision, recall and F1-score are calculated in Table 6. The precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. The F1 score is the harmonic mean of precision and recall. The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is zero, if either the precision or the recall is zero.

Table 6. Precision, recall, and F1-score

Criteria Category	Precision	Recall	F1-Score
Low	1.000	0.333	0.500
Medium	0.620	0.574	0.596
High	0.645	0.742	0.690

The results of precision, recall and F1-score are satisfactory.

4. Discussion

The violations, which have the highest frequencies, are HDV, SLV, RLV, and MDV, respectively (Figure 2). Previously, as reviewed in the introduction it had been proved that HDV, SLV, RLV, and MDV could cause serious crashes. Therefore, to reduce crash risk, it is essential to reduce these violations among young drivers in Iran. Enforcement by cameras can be proposed as one of the easiest and most efficient countermeasures for SLV, MDV and RLV reduction. Moreover, Stanojevic et al. [2018] proved that police presence and enforcement is a reliable approach to reduce angry and aggressive driving) [Stanojević et al., 2018]. The lack of cameras for SLV, MDV, and RLV enforcement can be seen in some cities and streets in Iran. In addition, because of the budget limitations and the economic embarogo, maintenance and repair operations are not done sequentially. Thus, there are some cameras, which do not work precisely and cannot record violations. This consequence has had a negative impact on young drivers' behavior and has reduced the influence of such enforcements. Police presence in the streets also has decreased in recent years, again because of the budget constraints for providing enough human resources. Nevertheless, crashes have more direct and indirect costs and it would be beneficial to devote a budget for cameras and police presence to reduce risky violations.

On the other hand, IBV, WRV, SBV, and IDV are those violations with the least frequencies. Recently, new and efficient cameras have been installed in the of-ramps to record the IBV in the urban arterials. This has been useful to reduce this violation. WRV can block a path and cause serious problems in relation to traffic flow, thus drivers try to avoid this violation. In addition, the urban network design, especially in Tehran (Capital of Iran) is in the form that there are different routes to a destination, thus drivers do not perceive the necessity for WRV. About ten years ago, strict regulations were established about SBV. After that, using seat belt has become a prerequisite for driving; now, drivers use it as a habit. This also proves the role of strict enforcement to reduce traffic violations, especially among young and novice drivers. Therefore, SBV frequency has also been reported low. Because of religious beliefs, using alcoholic drinks is not prevalent in Iran. It seems that it was the main reason that IDV has not been reported high among young drivers.

The MLR model outputs indicate that when PCA increases, the odds of SLV, RLV, and IDV violation commitment also increase (Table 4). This means that those drivers, who have higher access to private cars during a week and also have more driving experience (Figure 2), are more susceptible to SLV, RLV, and IDV with high frequencies.

Male drivers have more potential to commit SLV and HDV than women. These violations were among the highest frequent violations. This can necessitate the difference between driving education and training programs held for young men than women.

The higher the education level of drivers is, the lower the frequencies of SLV, SBV, RLV, IDV, VDV, MDV, WRV commitments are. This can be an important result, which shows the role of education in driving behavior. Violations can increase the crash risk and disarrangements. Those with higher education levels have understood this fact and try to avoid them. Because of the considerable correlation between EDU and AG, it can be stated that the age of young drivers also has an indirect impact on reducing dangerous violations.

Among those drivers who have received more traffic fines, the frequency of SLV, IDV, URV, VDV, and MDV are expected to be higher. This result indicates that those drivers with a previous negative background in relation to traffic violations repeat these violations frequently. This relates to the fact that traffic fines are not preventive and their amount and type cannot discourage drivers from violating driving rules. [Sagberg & Ingebrigtsen, 2018] evaluated two hypotheses about the impact of previous penalty points of drivers on the probability of receiving new penalty points. Their analyses indicated that the relationship between the number of penalty points in a year and the number of received penalty points in the subsequent year is in the form of an inverse U. This means that those who have committed traffic violations previously have the potential to do new violations; however, when they are at risk of losing their driving license or a sever danger, they would commit fewer violations than the previous year.

The higher the frequency of previous crashes is, the more SLV, MDV, and HDV frequencies are expected. This result again approves the fact that these violations have a great role in crash risk. Furthermore, the necessity for reducing these violations has been highlighted.

The Kendalls Tae results (Table 5) indicate most violations have a relatively high correlation with MDV. Using mobile phone, while driving can cause driving distraction (Oviedo-Trespalacios et al., 2016). This can be the main reason for the high correlation between MDV and other violations. Distracted drivers are more susceptible to other violations. Previously, strict controls and enforcements existed in Iran about using mobile phone during driving. Nevertheless, in the recent years, although mobile phones have become more popular, especially among young drivers, less strict regulations are held on using such devices while driving.

The most important variables in the RF model for predicting the frequency and diversity of dangerous violations' commitments were TF, PCA, CF, and GE2, respectively (Figure 4). These variables also have the most correlation with DAFV class, which indicates their impact on the frequency and diversity of violations committed by young drivers (Figure 3).

The number of previous traffic fines is the most suitable variable to identify the violation behavior of young drivers. Here, violation behavior means the frequency and diversity of violations. The higher the frequency of TF is, the higher the frequency of different violations by a young driver is expected to be accordingly. This shows the necessity for an overlook on the amount and type of traffic fines in Iran, to highlight their intervention role. Among the drivers who use their private cars more frequently during a week, the frequency and diversity of violations are higher. This can be managed by pricing methods in big cities, which would be useful for traffic efficiency at the same time. In addition, the modification of traffic fines can also decrease the violations among those drivers who use their cars more frequently in a typical week. Drivers with high crash previous crash frequencies, also do various dangerous violations more frequently. Male drivers also commit various violations with high frequencies. Therefore, the need for a revision in driving education and training programs of young drivers has been proved again using the RF model.

5. Conclusion

Traffic crashes are still one of the leading causes of death in developing countries. Deliberate violations are the main source of traffic crashes. The population of young drivers in developing countries is high, and a considerable portion of violations relate to this age group. Thus, this paper focused on studying the traffic violations of young drivers. For this purpose, Iran as a country with a high frequency of road crash fatalities and a huge number of young drivers, was selected as the case study. It was tried to provide an in-depth insight into traffic violations of young drivers in this developing country. For this purpose, data mining techniques were applied to the data collected by a self-reported survey. In the data mining process, first, the relative frequency of violations committed by young drivers was determined. Then it was tried to determine the impact of drivers’ characteristics and previous negative backgrounds of drivers (e.g., traffic fines and crashes) on the frequency and diversity of committed violations. The correlation of the frequency of committing each violation with other violations also was calculated. Results indicated that HDV, SLV, RLV, and MDV are the most frequent violations. It is necessary to highlight the previously successful role of enforcement with the help of cameras and police presence in reducing SBV in Iran. Out of all input variables, those drivers who have previously received more traffic fines, male drivers, drivers who drive more frequently showed the highest potential for doing dangerous violations in the future. A revision on the type and the number of traffic fines, discriminating driving education and training programs based on gender, and pricing strategies are suggested as potential countermeasures for reducing the frequency and diversity of risky violations. The frequency of using mobile phone while driving also had a considerable correlation with several dangerous violations. However, the control and enforcement on this violation have not been taken seriously. At last, the relationship between the frequency of previous crashes with frequency and diversity of dangerous violations can prove the role of violations on traffic crashes of young drivers.

6. References

- Alvaro, P. K., Burnett, N. M., Kennedy, G. A., Min, W. Y. X., McMahon, M., Barnes, M., Jackson, M., & Howard, M. E. (2018). "Driver education: Enhancing knowledge of sleep, fatigue and risky behaviour to improve decision making in young drivers". Accident Analysis and Prevention, Vol. 112, pp. 77–83. https://doi.org/10.1016/j.aap.2017.12.017

- Amiri, A. M., Sadri, A., Nadimi, N., Shams, M., (2020). "A comparison between Artificial Neural Network and Hybrid Intelligent Genetic Algorithm in predicting the severity of fixed object crashes among elderly drivers". Accident Analysis and Prevention, Vol. 138, pp. 105468. https://doi.org/10.1016/j.aap.2020.105468

- Asadamraji, M., Saffarzadeh, M., Borujerdian, A., & Ferdosi, T. (2018). "Hazard detection prediction model for rural roads based on hazard and environment properties". Promet - Traffic - Traffico, Vol. 30, No. 6, pp. 683–692. https://doi.org/10.7307/ptt.v30i6.2638

- Asadamraji, M., Saffarzadeh, M., Ross, V., Borujerdian, A., Ferdosi, T., & Sheikholeslami, S. (2019). "A novel driver hazard perception sensitivity model based on drivers' characteristics: A simulator study". Traffic Injury Prevention, Vol. 20, No. 5, pp. 492–497. https://doi.org/10.1080/15389588.2019.1607971

- Asadamraji, M., Saffarzadeh, M., & Mirzaee Tayeghani, M. (2017). "Modeling Driver's Hazard Perception using Driver's Personality Characteristics". International Journal of Transportation Engineering Vol. 5, No. 2, pp. 167-182.
https://doi.org/10.22119/IJTE.2017.46520

- Beanland, V., Goode, N., Salmon, P. M., & Lenné, M. G. (2013). "Is there a case for driver training? A review of the efficacy of pre- and post-licence driver training". Safety Science. Vol. 51, No. 1, pp. 127–137. https://doi.org/10.1016/j.ssci.2012.06.021

- Behnood, H. R., Rajabpour, M., Rassafi, A. A., & Hermans, E. (2019). "Efficiency Analysis of Road Safety Pillars by Applying the Results of a Structural Equations Model in Data Envelopment Analysis Efficiency". International Journal of Transportation Engineering. Vol. 7, No. 3, pp. 315-327. https://doi.org/10.22119/IJTE.2019.141484.1423

- Brijs, K., Cuenen, A., Brijs, T., Ruiter, R. A. C., & Wets, G. (2014). "Evaluating the effectiveness of a post-license education program for young novice drivers in Belgium". Accident Analysis and Prevention, Vol. 66, pp. 62–71.
https://doi.org/10.1016/j.aap.2014.01.015

- Christie, R. (2001). "The effectiveness of driver training as a road safety measure: an international review of the literature". Road Safety, Research, Policing and Education Conference : Proceedings : Regain the Momentum : Hilton on the Park.

- Clinton, K.M. and Lonero, L. (2006). "Evaluating driver education programs". Department of Violence & Injury Prevention & Disability, World Health Organization, D. of V. I. P. D. (VIP). (2018). Global Status Report on Road Safety (WHO). www.WHO.int/violence_injury_prevention

- Eboli, L., & Mazzulla, G. (2012). "Structural Equation Modelling for Analysing Passengers' Perceptions about Railway Services". Procedia

- Social and Behavioral Sciences, Vol. 54, pp. 96–106.
https://doi.org/10.1016/j.sbspro.2012.09.729

- Elvik, R., & Vaa, T. (2009). "Handbook of Road Safety Measures". Elsevier Science.

- Golob, T. F. (2003). "Structural equation modeling for travel behavior research". Transportation Research Part B: Methodological. Vol. 37, No. 1, pp. 1–25. https://doi.org/10.1016/S01912615(01)00046-7

- Haworth, N., Kowadlo, N., Tingvall, C. (2000). "Evaluation of Pre-Driver Education Program". Monash University Accident Research Centre Reports, Vol. 167, pp. 76.

- Hirsch, P., Maag, U., & Laberge-Nadeau, C. (2006). "The role of driver education in the licensing process in Quebec". Traffic Injury Prevention, Vol. 7, No. 2, pp. 130–142. https://doi.org/10.1080/15389580500517644

- Johnson, R., & Wichern, D. (2002). "Applied Multivariate Statistical Analysis (6th ed.)". Vol. 5, No. 8, Prentice hall, Upper Saddle River.

- Ker, K., Roberts, I., Collier, T., Beyer, F., Bunn, F., & Frost, C. (2005). "Post-licence driver education for the prevention of road traffic crashes: A systematic review of randomised controlled trials". Accident Analysis and Prevention, Vol. 37, No. 2, pp. 305–313. https://doi.org/10.1016/j.aap.2004.09.004

- Kumfer, W., Liu, H., Wu, D., Wei, D., & Sama, S. (2017). "Development of a supplementary driver education tool for teenage drivers on rural roads". Safety Science, Vol. 98, pp. 136–144. https://doi.org/10.1016/j.ssci.2017.05.014

- L. Arbuckle, J. (2007). "Amos User's Guide". SPSS Inc.

- Lonero, L. P. (2008). "Trends in Driver Education and Training". American Journal of Preventive Medicine. Vol. 35, No. 3, pp. S316–S323. https://doi.org/10.1016/j.amepre.2008.06.023

- Mayhew, D. R., & Simpson, H. M. (2002). "The safety value of driver education and training". Injury Prevention, Vol. 8 (SUPPL. 2), pp. ii3-ii8. https://doi.org/10.1136/ip.8.suppl_2.ii3

- Mayhew, Daniel R. (2007). "Driver education and graduated licensing in North America: Past, present, and future". Journal of Safety Research, Vol. 38, No. 2, pp. 229–235. https://doi.org/10.1016/j.jsr.2007.03.001

- Nadimi, N., Sangdeh, A. K., & Amiri, A. M. (2020). "Deciding about the effective factors on improving public transit popularity among women in developing countries". Transportation Letters. pp. 1-9. https://doi.org/10.1080/19427867.2020.1801022

- Petzoldt, T., Weiß, T., Franke, T., Krems, J. F., & Bannert, M. (2013). "Can driver education be improved by computer based training of cognitive skills?". Accident Analysis and Prevention, Vol. 50, pp. 1185–1192. https://doi.org/10.1016/j.aap.2012.09.016

- Rodwell, D., Hawkins, A., Haworth, N., Larue, G. S., Bates, L., & Filtness, A. (2018). "A mixed-methods study of driver education informed by the Goals for Driver Education: Do young drivers and educators agree on what was taught?". Safety Science, Vol. 108, pp. 140–148. https://doi.org/10.1016/j.ssci.2018.04.017

- Sadia, R., Bekhor, S., & Polus, A. (2018). "Structural equations modelling of drivers' speed selection using environmental, driver, and risk factors". Accident Analysis and Prevention, Vol. 116, pp. 21–29. https://doi.org/10.1016/j.aap.2017.08.034

- Shell, D. F., Newman, I. M., Córdova-Cazar, A. L., & Heese, J. M. (2015). "Driver education and teen crashes and traffic violations in the first two years of driving in a graduated licensing system". Accident Analysis and Prevention, Vol. 82, pp. 45–52. https://doi.org/10.1016/j.aap.2015.05.011

- Thomas, J. R. V., & Jones, S. J. (2014). "Injuries to 15-19-Year olds in road traffic crashes: A cross sectional analysis of police crash data". Journal of Public Health (Germany), Vol. 22, No. 3, pp. 245–255. https://doi.org/10.1007/s10389-014-0617-8

- Zhao, X., Xu, W., Ma, J., Li, H., & Chen, Y.

(2019). "An analysis of the relationship between driver characteristics and driving safety using structural equation models". Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 62, pp. 529–545. https://doi.org/10.1016/j.trf.2019.02.004

- Zong, F., Yu, P., Tang, J., & Sun, X. (2019). "Understanding parking decisions with structural equation modeling". Physica A: Statistical Mechanics and Its Applications, Vol. 523, pp. 408–417. https://doi.org/10.1016/j.physa.2019.02.038