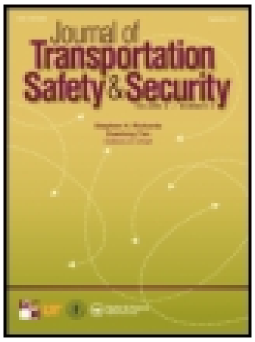

Journal of Transportation Safety & Security

ISSN: 1943-9962 (Print) 1943-9970 (Online) Journal homepage: http://www.tandfonline.com/loi/utss20

Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates

Madhar Taamneh, Sharaf Alkheder & Salah Taamneh

To cite this article: Madhar Taamneh, Sharaf Alkheder & Salah Taamneh (2016): Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates, Journal of Transportation Safety & Security, DOI: 10.1080/19439962.2016.1152338

To link to this article: http://dx.doi.org/10.1080/19439962.2016.1152338

Accepted author version posted online: 27 Apr 2016.

Data Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates

Madhar Taamneh[1], Sharaf Alkheder[1], Salah Taamneh[2]

[1]Department of Civil Engineering
Hijjawi Faculty for Engineering Technology
Yarmouk University, P.O. Box 566, Irbid 21163, Jordan
sharafalkehder@yu.edu.jo; sharafalkeder@hotmail.com

[2]Department of Computer Science
Collage of Natural Science and Mathematics
University of Houston, 4800 Calhoun Road, Houston, TX 77004, United States

Abstract:

Road traffic accidents are among the leading causes of death and injury worldwide. In Abu Dhabi, in 2014, 971 traffic accidents were recorded, which contributed to 121 fatalities and 135 severe injuries**.** Several factors contribute to injury severity, including driver-related factors, road-related factors, and accident-related factors. In this paper, data-mining techniques were employed to establish models (classifiers) to predict the injury severity of any new accident with reasonable accuracy, based on 5973 traffic accident records in Abu Dhabi over a six-year period

from 2008 to 2013. Additionally, the research aimed to establish a set of rules that can be used by the United Arab Emirates Traffic Agencies to identify the main factors that contribute to accident severity. Using WEKA (Waikato Environment for Knowledge Analysis) data mining software, four well-known classification algorithms were employed to model the severity of injury. These algorithms included: Decision Tree (J48), Rule Induction (PART), Naive Bayes, and Multilayer Perceptron. The effectiveness of each method in predicting accident severity was evaluated in three different ways. First, the entire dataset was used as a training set for the algorithm. Second, accuracy was evaluated using cross-validation with 10-fold. Third, to overcome the problems that resulted from the imbalanced distribution of accident severity in our dataset, the dataset was resampled to bias the accident severity distribution toward a uniform distribution, and then cross-validation with 10-foldwas used again to evaluate the performance. Furthermore, to establish the main contributing factors for road accidents severity, rules generated by the Decision Tree (J48) algorithm were further explored. The results showed that the overall accuracy of the Decision Tree (J48) classifier, the Rules Induction (PART) classifier, and the Multilayer Perceptron classifier in predicting the severity of injury resulting from traffic accidents, using 10-foldcross-validation, were similar. The Naive Bayes classifier exhibited less accuracy. Additionally, the prediction accuracy of the classifiers was enhanced after resampling the training set. The results indicated that the most important factors associated with fatal severity were age, gender, nationality, year of accident, casualty status, and collision type.

18-30 year olds were the most vulnerable age group to traffic accidents. There was a clear trend in accident reduction over the period of the study. Drivers were involved more frequently in traffic accidents than passengers and pedestrians. Male drivers were involved more frequently in

traffic accidents than female drivers. UAE, Asian, and Arab nationalities had the highest traffic accident frequency; Gulf and other nationalities had lower traffic accident frequency. The highest number of traffic accidents occurred at right angles. Pedestrian-vehicle type collisions had the next highest number of traffic accidents, followed by rear-end collisions and sideswipe collisions.

Keywords: Traffic Accidents, Data Mining, WEKA, Prediction, Modeling

1. Introduction

Road accidents are a major threat worldwide that continue to cause casualties, injuries and fatalities on roadways on a daily basis, resulting in huge losses both at the economicand social levels. According to World Health Organization records, road accidents have become one of the major causes of death worldwide. Accidents cause approximately 1.27 million annual deaths and between 20 and 50 million injuries (de Oña, López, & Abellán, 2013). This global problem needs more attention to reduce the severity and the frequency of accident occurrence. The historical data about previous accidents represents a formidable opportunity for researchers to identify the most influential factors in such accidents, which in turn play a key role in finding appropriate solutions to mitigate this problem in the future. It is, however, a very challenging task to extract knowledge from these data as they are typically huge and high dimensional. In recent years, several data mining techniques have been effectively used to extract useful knowledge from large data sets containing information about traffic accidents. Classification methods are among the most commonly used techniques in mining road traffic accidents, where the goal is building classifiers that are capable of predicting the severity of new accidents. These classifiers are built using training sets of data in which the severity of the accidents is known. Classification methods are categorized into three types: mathematical, probabilistic, and rule-based classifiers. The accuracy of a classifier depends heavily on the characteristic of the collected data, a fact that makes it almost impossible to recommend a certain classifier for a particular kind of problems. As a result, researchers tend to test multiple classifiers before deciding which one to use. In this work, we investigate the performance of several classification algorithms in predicting the injury severity of new accidents (i.e., death, severe, moderate, or minor) occur in Abu Dhabi, based on 5973 traffic accident records collected from the same city over a six-year period from 2008 to 2013. Additionally, the work aimed to establish a set of rules that can be used by the United Arab Emirates Traffic Agencies to identify the main factors that contribute to accident severity. Using WEKA (Waikato Environment for Knowledge Analysis) data mining software, four well-known classification algorithms were employed to model the severity of injury. These algorithms were: two rule-based classification algorithms (Decision Tree (J48), and Rules Induction (PART)), one mathematical classification algorithm (Multilayer Perceptron), and one probabilistic classification algorithm (Naive Bayes). The effectiveness of each method in predicting the accident severity was evaluated in three different ways. More details are presented in the subsequent sections.

2. Related Work

Accident data extracted from traffic police records represent a major source of data that can be used in an extensive analysis to investigate traffic accident severity as a function of more than 100 different contributing parameters (Delen, Sharda, & Bessonov, 2006). Some studies tried to extract the most important factors that contribute to traffic accident occurrence (Chang & Wang, 2006; Chen & Jovanis, 2000; Kopelias, Papadimitriou, Papandreou, & Prevedouros, 2007; Xie, Zhang, & Liang, 2009).

Many studies in the literature focus on studying accident severity (Kashani & Mohaymany, 2011; Newgard, Lewis, & Jolly, 2002). Furthermore, numerous data analysis techniques were used to analyze the collected accidents database, focusing on analyzing different accident data attributes. These techniques range from advanced regression analysis methodologies (Milton, Shankar, & Mannering, 2008; Yamamoto & Shankar, 2004; Yau, Lo, & Fung, 2006) to spatiotemporal methods and data mining techniques.

Our work focuses on using data mining techniques to predict accident severity in Abu Dhabi, UAE. There is significant research in the literature focused on using data mining techniques in accident data analysis. Kashani, Rabieyan, & Besharati (2014) investigated the main factors influencing crash severity of motorcycle passengers using a Classification and Regression Trees (CART) method. The CART method has also been used to analyze the injury severity of traffic crashes (Chang & Chien, 2013). The predictive accuracy of the model built with a total of 16 variables was 74%. Area type, land use, and injured part of the body (e.g., head, neck) were the most influential factors affecting the fatality of motorcycle passengers. Chang and Wang (2006) developed a classification and regression tree (CART) model to determine the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. They used CART because it does not require a pre-defined underlying relationship between the dependent variable and independent variables and has been shown to be a powerful tool, particularly for prediction and classification problems. The results indicated that the most important variable associated with crash severity is vehicle type. Pedestrians, motorcycle and bicycle riders were identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents.

Mujalli and de Oña (2011) presented a method to decrease the number of variables used in modeling accident severity using Bayesian Networks (BN). Several data mining variable selection algorithms were used to select subsets of variables on which the BNs were built. The performance of the BNs with select subsets of variables was compared with the original BN (with all variables) using five indicators. The process was repeated to reduce the number of variables further. The results indicated that it is possible to reduce the number of variables used to model traffic accident injury severity through BNs without reducing the performance of the model. Gregoriades (2007) also used BNs to model traffic accidents without considering traffic accidents as a deterministic assessment problem. De Oña et al (2011) used Bayesian Networks (BNs) to classify traffic accidents according to injury severity. A total of 1536 accidents on rural highways in Spain were used, and 18 variables representing the contributing factors were used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured.

Abellan et al. (2013) introduced an effective method for extracting rules from decision trees. The study focused on traffic accident data from rural roads in Granada (Spain) from 2003 to 2009. More than 70 relevant rules were obtained using the new method, whereas with one decision tree they extracted only five relevant rules from the same dataset.

Kwon et al. (2015) inspected 25 fields that are most relevant to car accidents, drawn from traffic accident reports between 2004 and 2010 from the California Highway Patrol (CHP). Using the Naive Bayes classifier and the decision tree classifier, the relative importance of the data fields, i.e., risk factors, was revealed with respect to the resulting severity level. The performances of the classifiers were compared, and a binary logistic regression model was used as the basis for the comparisons. Some of the high-ranking risk factors were strongly dependent on each other, and their incremental gains on estimating or modeling severity level were evaluated quantitatively. The analysis showed that only a handful of the risk factors in the data dominated the severity level and that dependency among the top risk factors was an imperative trait to consider for an accurate analysis.

3. Methodology

This study aims to explore the performance of several data mining techniques in predicting the severity of the accidents that occurred in Abu Dhabi from 2008 to 2013. In addition, the aim is to identify the main factors that contribute to the severity of such accidents. This section explains the methods used in the study, including collecting the data, building the classification models, and extracting the required knowledge. These steps can be divided into four phases: data collection, data preprocessing, building and validating classification models, and knowledge extraction.

3.1 Data Sources and Description

The dataset used in this study was obtained from the Emirate of Abu Dhabi, UAE in the form of an Excel spreadsheet. This dataset contains information about 5973 road accidents that occurred over a period of 6 years from 2008 to 2013. Each accident is described using 48 pieces of information (i.e., attributes) that were recorded at the time of the accident. These attributes can be classified into three categories: driver-related attributes, road-related attributes, and accident-related attributes. The accidents were classified into four levels based on severity: death, severe, moderate, and minor. The percentage of each category was 3%, 7%, 31%, and 59%, respectively. The number of fatal and severe accidents only accounts for a relatively small proportion of the total accidents.

3.2Data Preprocessing

Throughout this paper, the term “target variable” is used to refer to the accident severity attribute that the models are seeking to predict, and the term “input variables” refers to the rest of the attributes. Before performing data mining, noisy and unreliable data were removed from the dataset. Additionally, some input variables containing irrelevant and redundant information were removed. Finally, for input variables with too many values, categorization made the data more suitable for classification.

The dataset was carefully screened for the problems mentioned above, and the following changes were made: removing the invariant attributes (e.g., the name of the emirate), removing the descriptive and wordy attributes (e.g., detailed description of the accident), removing the irrelevant attributes (e.g., accident ID), removing the records with unknown values in the target attribute, categorizing the attributes with too many values (e.g., accident reason), and removing the redundant information (e.g., Date, Age). The final list of the attributes is presented in Table 1.

Table 1: Class Labels of the Selected 16 Attributes with Their Data Type and Description (Accident, Driver, and Road)

Attributes Name		Data Type	Description
Accident Related Attributes	Year	Numeric	The year of accident
	Day	Nominal	The day of accident
	Time	Nominal	The accident occurred at what time of the day.
	Reason	Nominal	Reason of the accident

	Accident Type	Nominal	Type of the accident
Driver Related Attributes	Gender	Nominal	Gender of the driver
	Nationality	Nominal	Nationality of the driver
	Age Rank	Numeric	Age of the driver
	Seat Belt	Nominal	The usage of a seatbelt during driving
	Casualty Status	Nominal	Whether the casualty is driver, passenger, or pedestrian
	Degree of Injury	Nominal	Death, severe, moderate, minor.
Road Related Attributes	Lighting	Nominal	Lighting condition of the road at the time of accident
	Road Surface	Nominal	Whether the surface of the road was dry, wet, sandy, or oily
	Speed Limit	Numeric	The road speed limit
	Lane Numbers	Numeric	The number of road lanes
	Weather	Nominal	The weather conditions

3.3 Building the Classification Models

Using WEKA data mining software, four well-known classification algorithms were employed to classify the severity of road accidents in Abu Dhabi, UAE from 2008 to 2013. These methods are: decision tree, Rules Induction, Naive Bayes, and Multilayer Perceptron. The effectiveness of each method in predicting accident severity was evaluated in three different ways. First, the whole dataset was used as a training set for the algorithm, and the accuracy of the classifier was determined based on how well it predicted the class of each accident. Second, accuracy was evaluated using cross-validation with 10-fold.The whole dataset was randomly partitioned into 10 subsets. Of the 10 subsets, a single subset was used as the testing data, and the remaining 9 subsets were used as training data. This process was then repeated 10 times, with each of the 10 subsets used exactly once as the testing data, resulting in the whole dataset being used for validation. Finally, the overall performance is calculated by averaging the 10 results from the folds. Third, to overcome the problems that result from the imbalanced distribution of accident severity in the dataset, the dataset was resampled to bias the accident severity distribution toward a uniform distribution, and then cross-validation with 10-foldwas used again to evaluate its performance. The WEKA tool was used to perform the sampling without replacement. In the following sections, a brief description of each algorithm used in this study is presented.

3.3.1 Decision Tree (J48)

Decision tree learning is widely used in data mining. The output of this method is a classification model that predicts the value of a target attribute based on the input attributes. The decision tree constructs classification models in the form of trees. Each interior node in these trees represents one of the input variables, and it has a number of branches equal to the number of possible values of that input variable. Each leaf node holds a value of the target attribute. The leaf node represents the decision made based on the values of the input variables from the root to the leaf. The decision tree algorithm was used to understand existing data and to predict the severity of the new accidents.

3.3.2 Rules Induction (PART)

Rule induction is one of the most important tools in data mining. It is an iterative process that follows a divide-and-conquer approach. At each iteration, a subset of the training set is used to generate rules using one of the decision tree algorithms. PART is a rule induction algorithm that creates a partial C4.5 decision tree in each iteration and chooses the best leaf to be a rule. The PART algorithm is used to create a classifier and to generate rules about the accidents.

3.3.3 Naïve Bayes (NB)

Naïve Bayes is a classification algorithm based on Bayesian theorem, with the naïve assumption that each pair of input variables is independent. Although this assumption is oversimplified, this algorithm has effectively been used in many complicated real-world problems especially document classification and spam filtering. Moreover, Naïve Bayes has proven to be very fast in learning and classifying data. In this paper, the Naïve Bayes algorithm is applied to the accidents dataset to gain insight into its performance in predicting the severity of accidents.

3.3.4 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a supervised learning technique that is used for classification and regression. MLP creates a feed-forward artificial neural network that consists of multiple nodes organized in three or more layers (the input layer, the output layer, and one or more hidden layers in between). The input variables are mapped onto the output variables using one or more hidden layers in between. As a result of using a backpropagation algorithm in training the generated networks, MLP has been successfully used to solve many difficult problems. MLP has the capability of separating data that are not linearly separable. In this study, the MLP technique was used to generate a classifier to accurately predict the severity of a future accident.

3.3.5 Knowledge Extraction

To understand the main factors that have an impact on road accident severity, the rules generated by the decision tree (J48) and PART algorithms were extracted. These rules were generated using the whole dataset as a training set for both algorithms. Then, the rules were organized based on the severity class into four groups (i.e., death, severe, moderate, and minor). For each group, the study only presents the rules with the least number of misclassified instances because presenting the full spectrum of rules is not feasible. More details about these rules and the information they provide are given in the results and discussion section.

4. Results and Discussion

A total of 5973 traffic accident records in Abu Dhabi over a six-year period from 2008 to 2013 were examined in this study. Sixteen (16) predictor variables were used with the class variable of severity of injury to build models (classifiers) to predict the degree of injury severity in traffic accidents. Additionally, the work aimsto identify the most important variables that contribute to traffic accidents. The software used to build the prediction models (classifiers) was WEKA. Four different classification techniques (algorithms) were applied to the data. The classification techniques were: Decision Tree (J48), Rules Induction (PART), Naive Bayes, and Multilayer Perceptron. A number of hyper-parameter settings were evaluated for each model, and the setting that yields the best performing model was chosen. Table 2 presents the selected setting for each model. The effectiveness of each algorithm in predicting the accident severity was evaluated in three different ways. First, the whole dataset was used as a training set. Second, the accuracy was evaluated using cross-validation with 10-folds. Third, to overcome the problems that result from the imbalanced distribution of accident severity in the dataset, the dataset was resampled to bias the accident severity distribution toward a uniform distribution, and cross-validation with 10-folds was used to evaluate the prediction performance. Furthermore, to establish the main contributing factors to road accidents severity, the rules generated by the Decision Tree (J48) algorithm were further explored. The accuracy of these classifiers is presented and discussed in the following sections.

Table 2: Hyper-parameters settings for all classifiers

Classifier	**Parameter**	**Description**	**Values**
J48	Binary Splits	Whether to use binary splits on nominal attributes when building the trees.	F
	Min Num Obj	Minimum number of instance per leaf	2
	Num Folds	Determines the amount of data used for reduced-error pruning	3
	Confidence Factor	The confidence factor used for pruning	0.25
	Unpruned	Whether pruning is performed	F
PART	Binary Splits	Whether to use binary splits on nominal attributes when building the trees.	F
	Min Num Obj	Minimum number of instance per leaf	2
	Num Folds	Determines the amount of data used for reduced-error pruning	3
	Confidence Factor	The confidence factor used for pruning	0.25
	Unpruned	Whether pruning is performed	F
MLP	Hidden Layers	The number of hidden layers	a (i.e , one hiddernlayer with 10 nodes)
	Learning Rate	The amount the weights are updated.	0.3
	Momentum	Momentum applied to the weights during updating.	0.2
	Normalize Attributes	This will normalize the attributes	True
	reset	This will allow the network to reset with a lower learning rate	True

NB	Use Kernel Estimator	Use a kernel estimator for numeric attributes rather than a normal distribution.	False
	Use Supervised Discretization	Use supervised discretization to convert numeric attributes to nominal ones.	Flase

4.1 Classifier Accuracy

The performance of a classifier model is defined from a matrix, known as confusion matrix, which shows the correctly and incorrectly classified instances for each class. Table 3 shows the 2×2 confusion matrix for a binary classifier that has only two classes-positive and negative (in our case it becomes 4×4 as we have 4 classes). The TP, TN, FP, FN can be described as follow:

- True Positive (TP): instances that are positive and classified as positive.
- True Negative (TN): instances that are positive but classified as negative
- False Positive (FP): instances that are negative but classified as positive
- False Negative (FN): instances that are negative and classified as negative

Table 3: Confusion Matrix

	Predicated Class	
True Class	Positive	Negative

Positive	TP	FN
Negative	FP	TN

The measures that are used to evaluate the performance of a classifier are computed from the generated confusion matrix. The most widely used evaluation measure is the accuracy rate, which shows the percentage of correctly classified instances and calculated as follow:

$$Accuracy \ = \frac{TP + \ Tn}{TP + FP + FN + TN}$$

Another evaluation measures that are commonly used to evaluate the effectiveness of a classifier for each class are: the True Positive Rate (TPR), the False Positive Rate (FPR), and Receiver Operating Characteristic (ROC) curve. The following part explains how to calculate these measures for class Positive in Table 3.

The True Positive Rate or Recall is the proportion of instances which were classified as Positive, among the all instances belong to the class Positive. Note that the overall accuracy of a classifier can also be calculated by taking the weighted average of all recall values.

$$Recall \ = \frac{TP}{TP + FN}$$

The False Positive Rate or (1- specificity) is the proportion of instances which were classified as class Positive, but belong to a different class, among all instances which are not of class Positive.

$$FPS \ = \frac{FP}{TN + FP}$$

Finally, ROC curve is a plot of the true positive rate (i.e., Recall) against the false positive rate at various threshold settings. Which shows the trade-offs between true positive (benefits) and false positive (costs).

The prediction results for the Decision Tree (J48) classifier are presented in Table 4. Three approaches were used to analyze the data and to develop the prediction models: analysis based on the whole data set as a training set, analysis using 10-foldcross-validation method, and finally, analysis based on resampling the training set. Crash severity was divided into four classes: death, severe, moderate, and minor. Decision Tree (J48) prediction accuracy based on the training data set for minor, moderate, severe, and death crash severity accidents was 91.36%, 75.85%, 38.28%, and 26.51%, respectively. These results are logical because the proportion of data available for model training for minor, moderate, severe, and death crash severity accidents was 59%, 31%, 7%, and 3%, respectively. For the J48 models, the overall prediction accuracy based on the training data was approximately 81%. Based on the 10-foldcross-validation,the prediction accuracy for minor, moderate, severe and death crash severity accidents was 86.95%, 66.27%, 20.57%, and 6.63%, respectively. The overall prediction accuracy for the 10-foldcross-validation was approximately 73.62%.Because the crash severity is imbalanced (i.e., the highest instances are for minor class, followed by moderate class, then severe and death classes), resampling was used to balance the class attributes to increase the prediction performance of the classifiers. The Decision Tree (J48) prediction accuracy after resampling for minor, moderate, severe, and death crash severity accidents was 79%, 80.93%, 92.84%, and 99.34%, respectively. For J48 models after resampling, the overall prediction accuracy for the training data was approximately 88.08%.

From these results, an enhancement in the prediction accuracy was observed after resampling the training set.

Table 4: Prediction results of the J48 model

Algorithm	Sample	Observed Injury	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy (Recall)	AUCs	Time (sec.)
J48	Using training set	Death	44	122	26.51%	0.927	0.09
		Severe	147	237	38.28%	0.914	
		Moderate	1275	406	75.85%	0.909	
		Minor	3004	284	91.36%	0.915	
		Overall	4470	1049	80.99%	0.913	
	Cross-validation (10-fold)	Death	11	155	6.63%	0.684	0.14
		Severe	79	305	20.57%	0.787	
		Moderate	1114	567	66.27%	0.823	
		Minor	2859	429	86.95%	0.839	
		Overall	4063	1456	73.62%	0.826	
	Resampled training set	Death	1355	9	99.34%	0.988	0.05
		Severe	1324	102	92.84%	0.967	
		Moderate	1087	256	80.93%	0.886	
		Minor	1095	291	79.00%	0.892	
		Overall	4861	658	88.08%	0.934	

The prediction results for the Rule Induction (PART) classifier are presented in Table 5. The PART prediction accuracy based on the training data set for minor, moderate, severe, and death crash severity accidents was 92.19%, 76.38%, 35.94%, and 33.13%, respectively. The overall prediction accuracy based on the training data was approximately 81%. Based on the 10-foldcross-validation the prediction accuracy for minor, moderate, severe, and death crash severity accidents was 82.35%, 68.76%, 14.32%, and 9.04%, respectively. The overall prediction accuracy was approximately 71.22%.The prediction accuracy after resamplingthe training datasetfor minor, moderate, severe, and death crash severity accidents was 71.79%, 65.89%, 91.23%, and 99.19%, respectively. The overall prediction accuracy after resampling was approximately 82.18%. From these results, an enhancement was observed in the prediction accuracy after resampling the training data.

Table 5: Prediction results of the PART model

Algorithm	**Sample**	**Observed Injury**	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Accuracy**	**AUCs**	**Time (sec.)**
PART	Using training set	Death	55	111	33.13%	0.96	0.51
		Severe	138	246	35.94%	0.934	
		Moderate	1284	397	76.38%	0.924	
		Minor	3023	256	92.19%	0.928	
		Overall	4500	1010	81.67%	0.928	
	Cross-validation (10-fold)	Death	15	151	9.04%	0.61	1.45
		Severe	55	329	14.32%	0.666	
		Moderate	1156	525	68.76%	0.768	
		Minor	2683	575	82.35%	0.797	
		Overall	3909	1580	71.22%	0.773	
	Resampled training set	Death	1353	11	99.19%	0.984	0.77
		Severe	1301	125	91.23%	0.959	
		Moderate	885	458	65.89%	0.841	
		Minor	995	391	71.79%	0.85	
		Overall	4534	985	82.15%	0.909	

The prediction results for the Multilayer Perceptron classifier are presented in Table 6. The Multilayer Perceptron prediction accuracy based on the training data set for minor, moderate, severe, and death crash severity accidents was 97.29%, 88.52%, 36.98%, and 16.87%, respectively. The overall prediction accuracy based on the training data was approximately 88.01%. Based on 10-foldcross-validation, the prediction accuracy for minor, moderate, severe, and death crash severity accidents was 84.15%, 70.49%, 23.87%, and 8.43%, respectively. The overall prediction accuracy was approximately 73.2%. The prediction accuracy after resampling of the training dataset for minor, moderate, severe, and death crash severity accidents was 74.09%, 72.75%, 89.90%, and 97.72%, respectively. The overall prediction accuracy after resampling was approximately 83.69%. From these results, an enhancement was observed in the prediction accuracy after resampling the training data.

Table 6: Prediction results of the Multilayer Perceptron model

Algorithm	Sample	Observed Injury	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy, (Recall)	AUCs	Time (sec.)
Multilayer Perceptron	Using training set	Death	28	138	16.87%	0.738	252.55
		Severe	142	242	36.98%	0.777	
		Moderate	1488	193	88.52%	0.955	
		Minor	3199	89	97.29%	0.952	
		Overall	4857	662	88.01%	0.934	
	Cross-validation (10-foldOverall)	Death	14	152	8.43%	0.597	302.69
		Severe	74	310	23.87%	0.619	
		Moderate	1185	496	70.49%	0.735	
		Minor	2767	521	84.15%	0.749	
		Overall	4040	1479	73.20%	0.731	
	Resampled training set	Death	1333	31	97.72%	0.984	306.59
		Severe	1282	144	89.90%	0.943	
		Moderate	977	366	72.75%	0.831	
		Minor	1027	359	74.09%	0.857	
		Overall	4619	900	83.69%	0.904	

The prediction results for the Naïve Bayes classifier are presented in Table 7. The Naïve Bayes prediction accuracy based on the training data set for minor, moderate, severe, and death crash severity accidents was 88.23%, 21.36%, 1.30%, and 1.81%, respectively. The overall prediction accuracy based on the training data was approximately 59.21%. Based on the10-foldcross-validation, the prediction accuracy for minor, moderate, severe, and death crash severity accidents was 88.23%, 19.33%, 0.78%, and 0.0%, respectively. The overall prediction accuracy was approximately 58.51%. The prediction accuracy after resampling of training dataset for minor, moderate, severe, and death crash severity accidents was 60.53%, 61.13%, 51.19%, and

59.31%, respectively. The overall prediction accuracy after resampling was approximately 57.05%. From these results, an enhancement in the prediction accuracy was observed after resampling the training data.

The Multilayer Perceptron classifier takes a longer time to build than the other classifiers. It takes approximately 300 seconds, whereas the Decision Tree (J48) classifier, Rules Induction (PART), and Naive Bayes take 0.09, 0.91, and 0.04 seconds, respectively. Figure 1 shows that the overall accuracy of the Decision Tree (J48) classifier, the Rules Induction (PART) classifier, and the Multilayer Perceptron classifier in predicting the severityof injury resulting from traffic accidents, using 10-foldcross-validation,is similar, whereas less accuracy was observed for Naive Bayes. Additionally,the prediction accuracy of the classifiers was enhanced after resampling the training set.

Table 7: Prediction results of the Naïve Bayes model

Algorithm	**Sample**	**Observed Injury**	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Accuracy**	**AUCs**	**Time (sec.)**
Naïve Bayes	Using training set	Death	3	163	1.81%	0.721	0.03
		Severe	5	379	1.30%	0.661	
		Moderate	359	1322	21.36%	0.612	
		Minor	2901	387	88.23%	0.627	
		Overall	3268	2251	59.21%	0.627	
	Cross-validation (10-fold)	Death	0	166	0.00%	0.627	0.02
		Severe	3	381	0.78%	0.605	
		Moderate	325	1356	19.33%	0.586	
		Minor	2901	387	88.23%	0.597	
		Overall	3229	2290	58.51%	0.599	
	Resampled training set	Death	809	555	59.31 %	0.717	0.06
		Severe	730	696	51.19 %	0.638	
		Moderate	821	522	61.13 %	0.596	

		Minor	789	597	60.53 %	0.641	
		Overall	3149	2370	57.05 %	0.648	

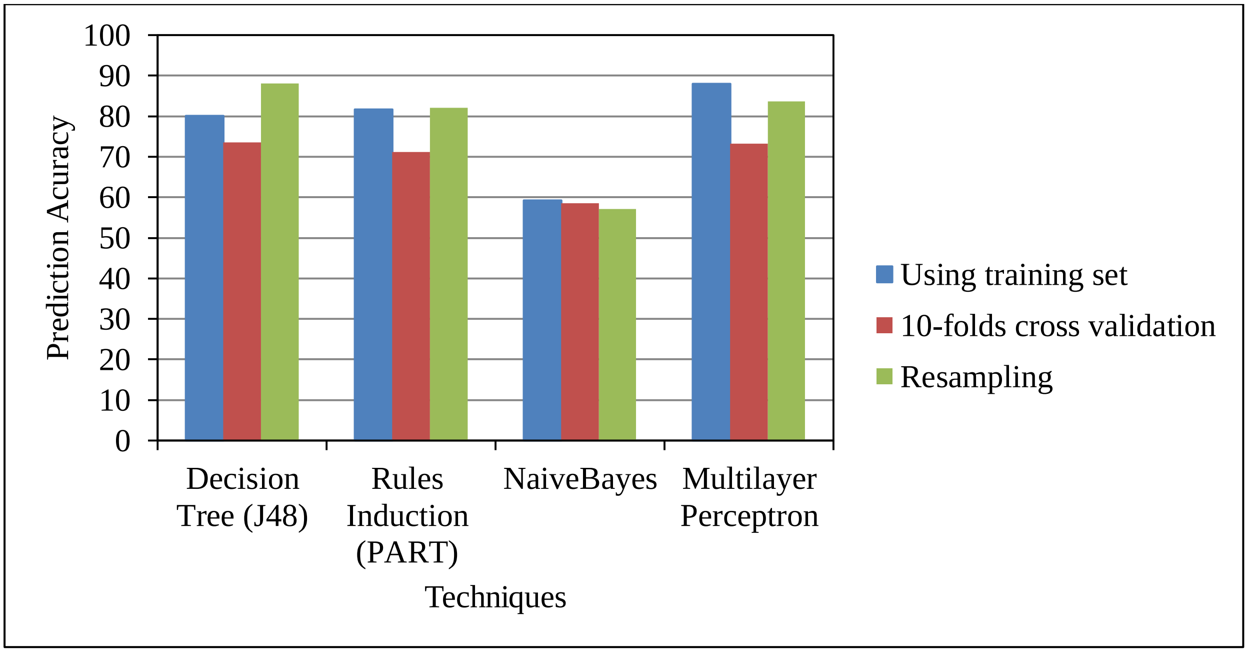

Figure 1: Overall prediction performance using different techniques.

To further investigate the prediction efficiency of the used techniques under each scenario, the Receiver Operating Characteristics (ROC) curve, also known as the relative operating characteristic curve, was studied. The area under the ROC curve (AUC) quantifies the overall discriminative ability of a test. An entirely random test has an AUC of 0.5, whereas a perfect test has an AUC of 1.00. Figure 2 shows the ROC curves for all classes using each classifier under various scenarios (i.e., using training set, cross-validation, cross-validation with resampling). The figure indicates that the ROC curves in the first scenario (i.e., using training set) are better than the corresponding ones in the second scenario. However, using the same instances for both testing and validation of a classification model can increase the possibility of overfitting, resulting in that model being only good for the used data set. The figure also shows that before resampling that data, the ROC curves are better for Minor and Moderate classes as opposed to Death and Severe classes. This is obviously due to the fact that Death and Severs classes are underrepresented classes as they represent only 10 percent of total accidents. After resampling the data, The ROC curves show that better accuracy can be achieved (i.e., true positive) at low cost (i.e., false positive) for all classes especially for Death and Severe accidents.

Within each scenario, J48 always outperforms the other classifiers, while Naïve Bayes always gives the worse results. Also, PART algorithm seems to be relatively better than MLP algorithm in all scenarios.

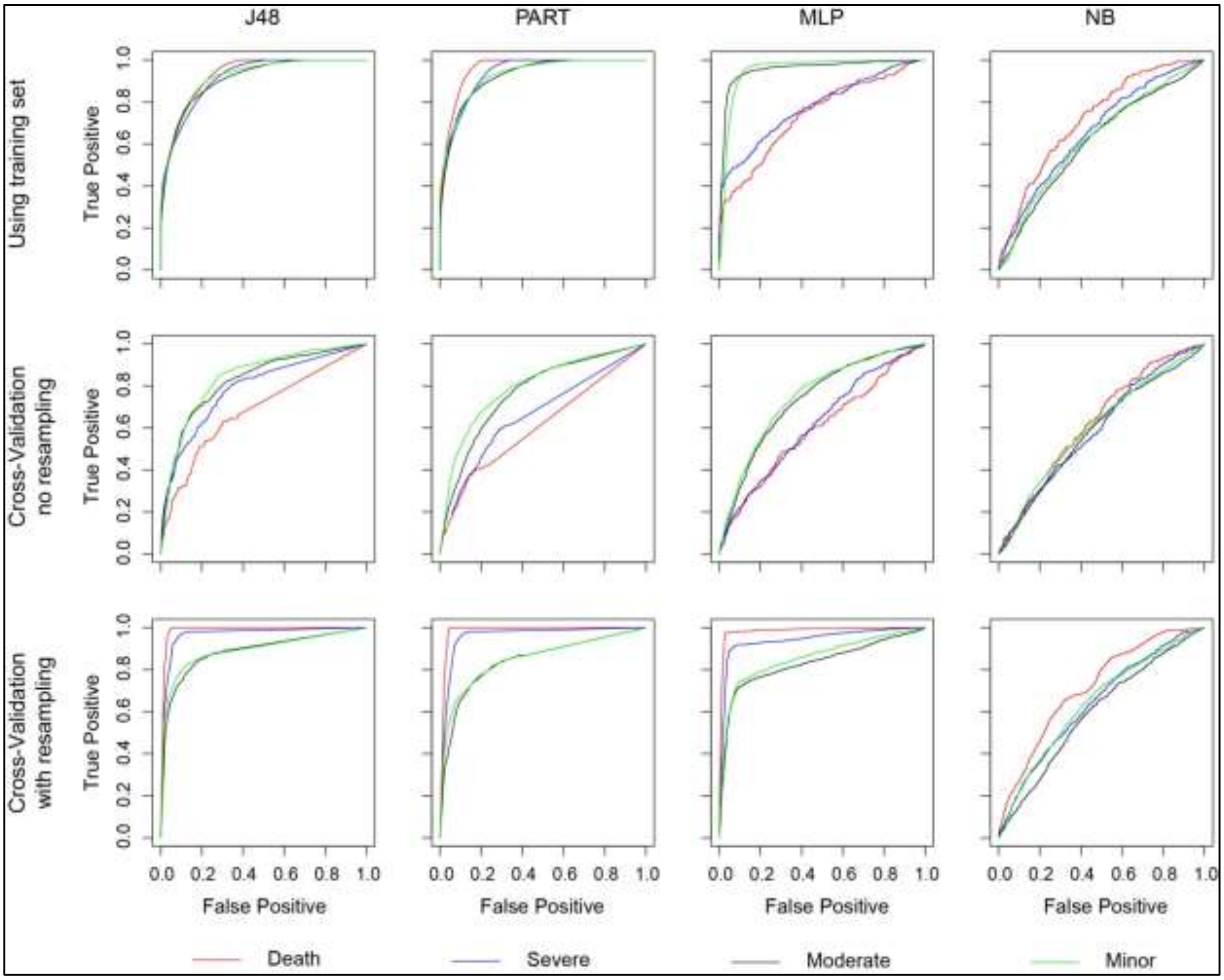

Figure 2: Receiver operating characteristic (ROC) curves for all classes using each classifier under three scenarios (training set, cross-validation, and cross validation with resampling)

4.2 Decision Rules Extraction

The significance of this study is in its development of new insights related to road accidents in Abu Dhabi Emirate. These insights provide valuable help in developing methods to improve road safety, particularly in the phase of choosing appropriate means and budget allocations of resources. Table 8 presents a partial output of the Decision Tree (J48) rules. The number of generated rules for death, severe, moderate and minor crash severity accidents is 10, 20, 94, and 145, respectively. Because of space constraints, Table 8 presents only the most significant rules identified in this research.. The frequency of each input attribute in the death, severe, moderate, and minor class rules is illustrated in Figures 3 through 6. As shown in Figure 3, the year of the accident, Age, casualty status, gender, nationality, and collision type are the primary splitters in the classification tree. This implies that these variables are critical in classifying injury severity in the traffic accidents. An age between 18-30 years is the most vulnerable age group to traffic accidents. This is logical because this age group tends to drive fast, has lack of experience, lacks respect for road users and is less responsible. Actions should be taken to educate this age group about road safety issues, and regulations should be updated to better enforce those regulations on road users to reduce traffic accidents. Year of accident is an important input variable (attribute) because it appears in almost all rules. Figure 7 shows a clear trend in accident reduction over time. This reduction could be attributed to the new regulations set forth in 2008. As shown in Figure 8, drivers are involved more frequently in traffic accidents than passengers and pedestrians. Figure 9 shows that males are involved more frequently in traffic accidents than females for all categories of causality status (drivers, pedestrians, and passengers). As can be seen from Figure 10, there is a clear reduction in accidents number for both genders over the years. This could be attributed to the unified traffic law implemented in 2009. Also, it can be noticed that the accident decrease rate for males over years is higher than that for females.

Figure 11 presents the relationship between the number of traffic accidents and the victim’s nationality. UAE, Asian, and Arab nationalities have the highest traffic accident frequency. Gulf and the other nationalities have lower traffic accident frequency. The traffic accident involvement over years based on nationality is depicted in Figure 12 it can be seen that for all nationalities the accident number is decreasing with years. The most dominant nationalities are Arab, UAE, and Asia, respectively.

Figure 13shows the number of traffic accidents for each collision type. The highest number of traffic accidents occurred at a right angle. Pedestrian-vehicle type collisions were the next highest, followed by rear-end collisions and sideswipe collisions. Knowing the most important factors that are directly linked to the severity of injury will help UAE authorities to effectively allocate resources to improve road safety.

Table 8: Partial output of the decision tree (J48) rules

Class Attribute	No. of Rules	Generated Rules	Total number of instances / misclassified instances
Death	10	Year = 2012 AND Casualty status = Passenger AND Gender = Female AND Nationality = UAE AND Age Rank = 31-45	8.0
Severe	20	Year = 2009AND Age Rank = >60	20.0 / 2.0
		Year = 2010 AND Seat belt = Y AND Casualty status = Passenger AND Gender = Male AND Nationality = Asia AND Age Rank = 31-45	16.0 / 4.0
		Year = 2008 AND Accident type = Head-on collision AND Casualty status = Driver AND Gender = Male AND Nationality = Asia AND Age Rank = 31-45	11.0 / 1.0
Moderate	94	Year = 2009 AND Casualty status = Pedestrian AND Age Rank < 18	42.0 / 8.0
		Year = 2013 AND Age Rank < 18	48.0
		Year = 2008AND Age Rank = >60	23.0
		Year = 2012 AND Nationality = Arab AND Age Rank = 18-30	36.0 / 5.0
		Year = 2010 AND Casualty status = Driver AND Nationality =	54.0 / 7.0

		Asia AND Age Rank = 18-30	
		Year = 2009 AND Gender = Female AND Nationality = UAE AND Age Rank = 18-30	36.0
		Year = 2008 AND Casualty status = Passenger AND Gender = Female AND Nationality = Asia AND Age Rank = 31-45	24.0
		Year = 2011 AND Casualty status = Driver AND Age Rank = 31-45	121.0 / 41.0
Minor	145	Year = 2008 AND Age Rank < 18	167.0 / 32.0
		Year = 2009 ANDRoad Surface = Dry AND Casualty status = Passenger AND Age Rank < 18	64.0 / 1.0
		Year = 2010 AND Nationality = Asia AND Age Rank < 18	22.0 / 5.0
		Year = 2010 AND Nationality = UAE AND Age Rank < 18	32.0 / 2.0
		Year = 2011 AND Age Rank < 18	98.0
		Year = 2012 AND Age Rank < 18	108.0 / 6.0
		Year = 2010 AND Age Rank >60	38.0 / 7.0
		Year = 2008 ANDNationality = Arab AND Age Rank = 18-30	165.0 / 49.0
		Year = 2009 ANDCasualty status = Pedestrian AND Nationality = Arab AND Age Rank = 18-30	21.0
		Year = 2010 ANDNationality = Arab AND Age Rank = 18-30	99.0
		Year = 2011 AND Gender = Male ANDNationality = Arab AND Age Rank = 18-30	41.0 / 2.0

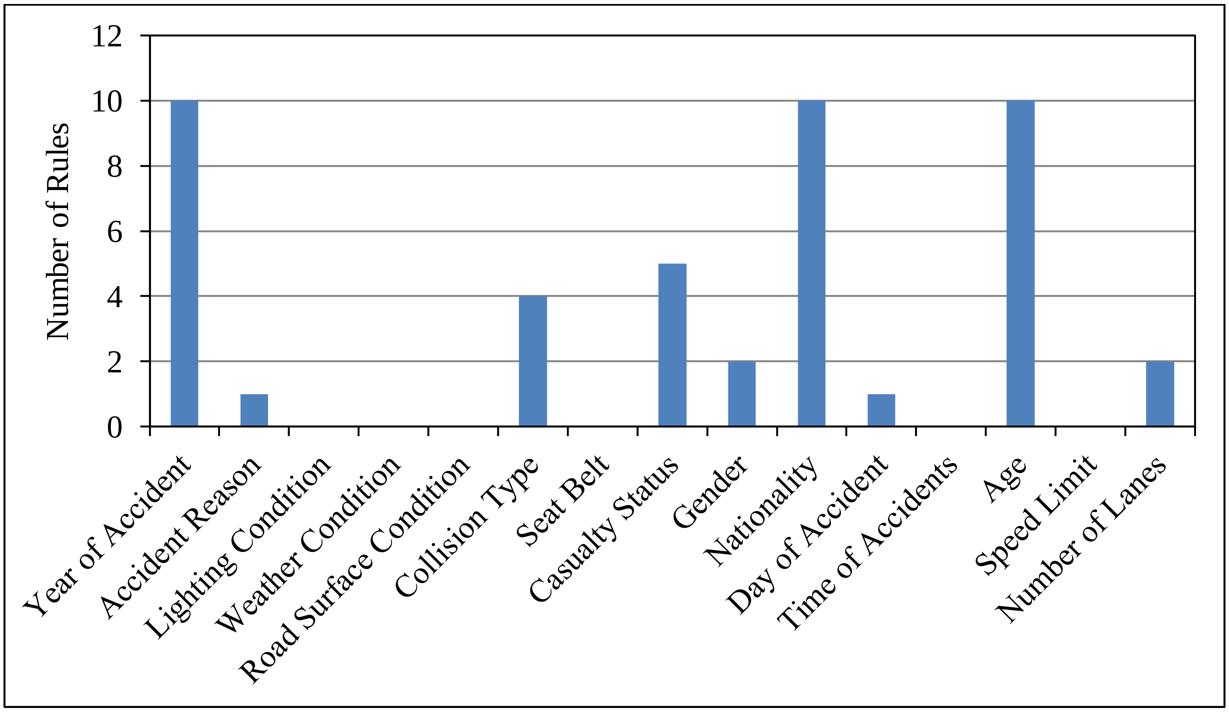

Figure 3: Distribution of Death Crash Severity Accidents

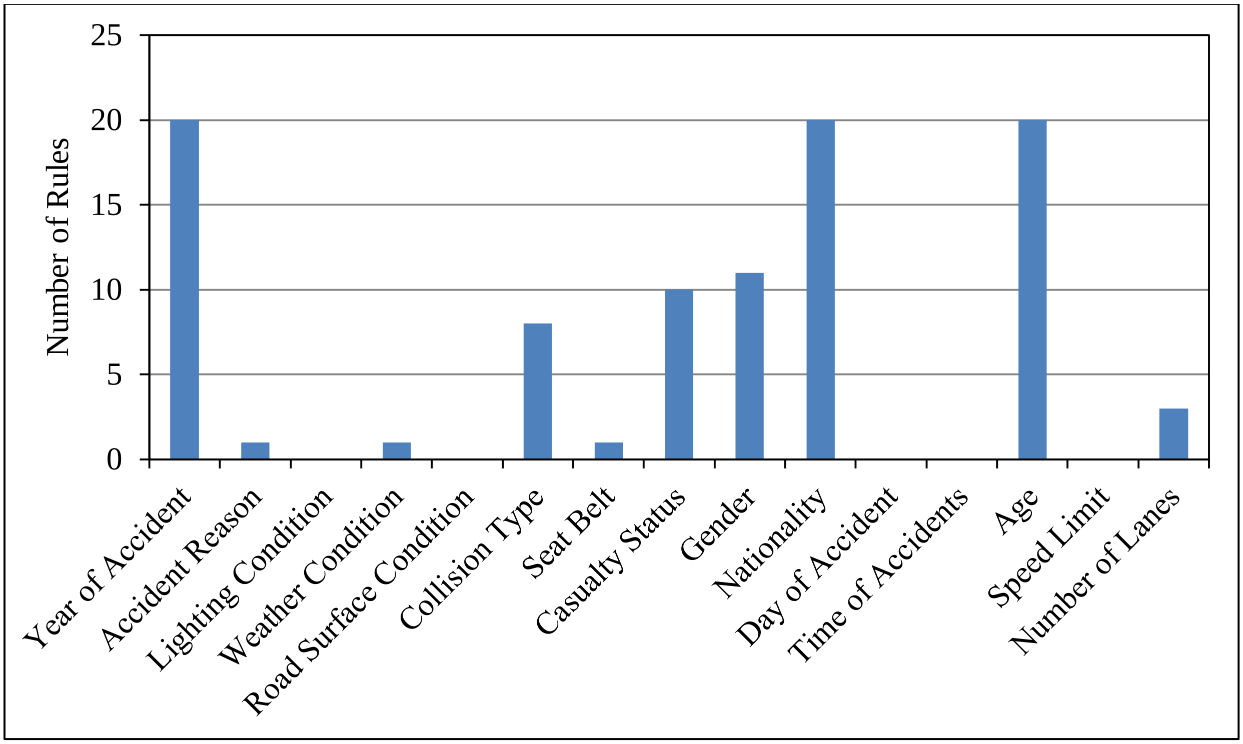

Figure 4: Distribution of Severe Crash Severity Accidents

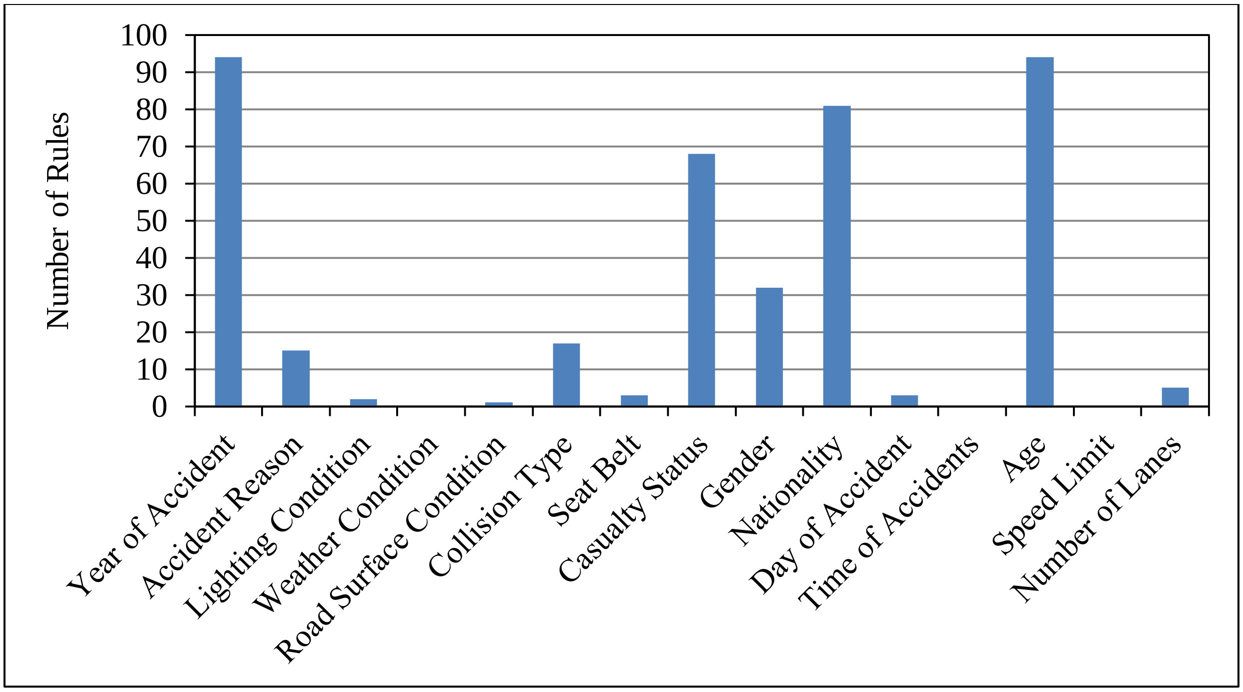

Figure 5: Distribution of Moderate Crash Severity Accidents

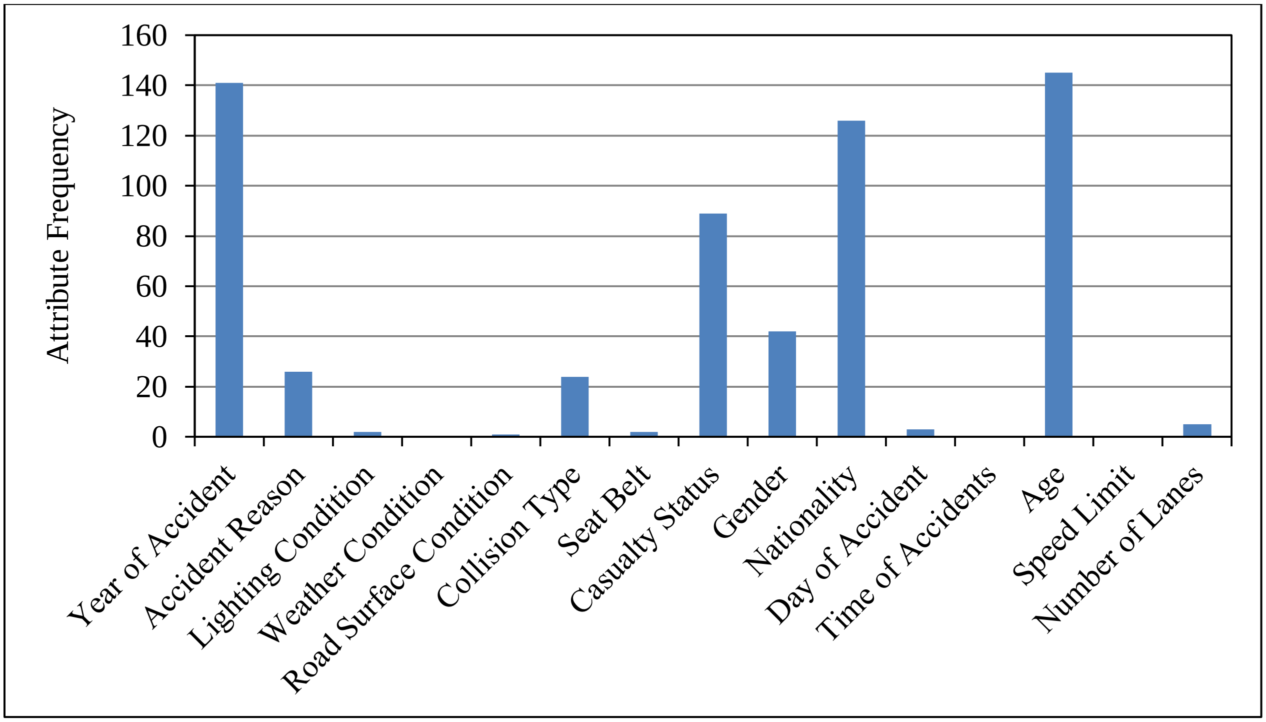

Figure 6: Distribution of Minor Crash Severity Accidents

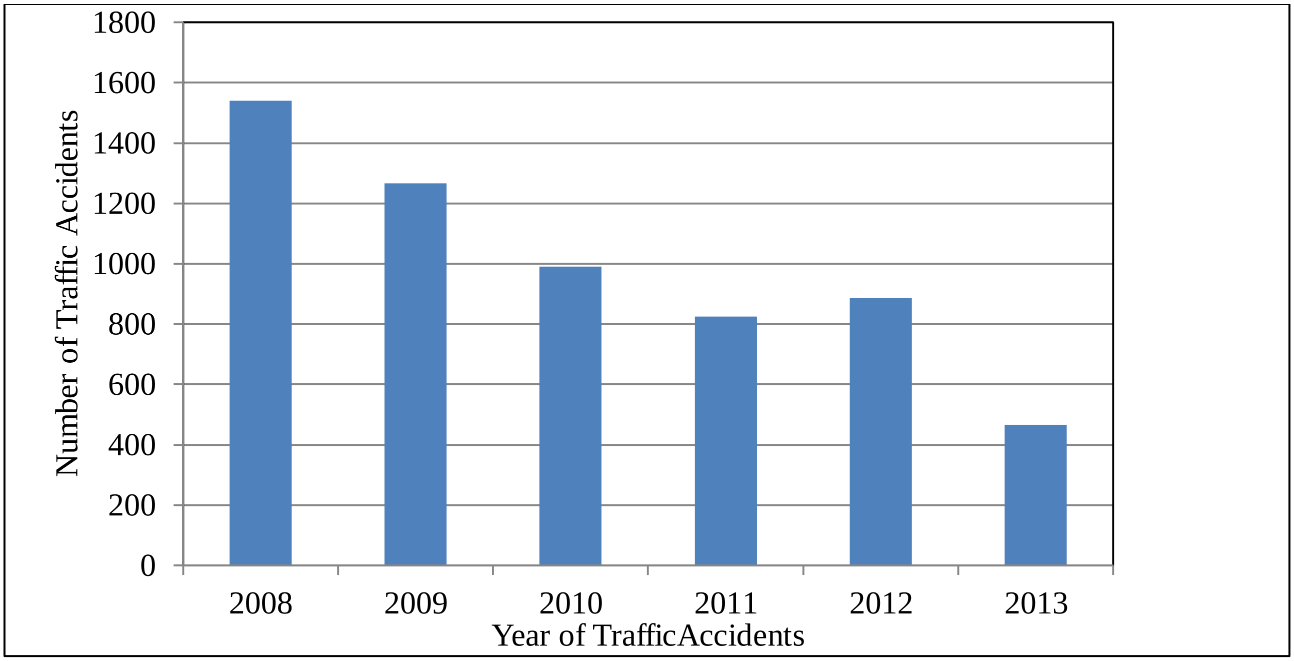

Figure 7: Number of traffic accidents over a six-year period

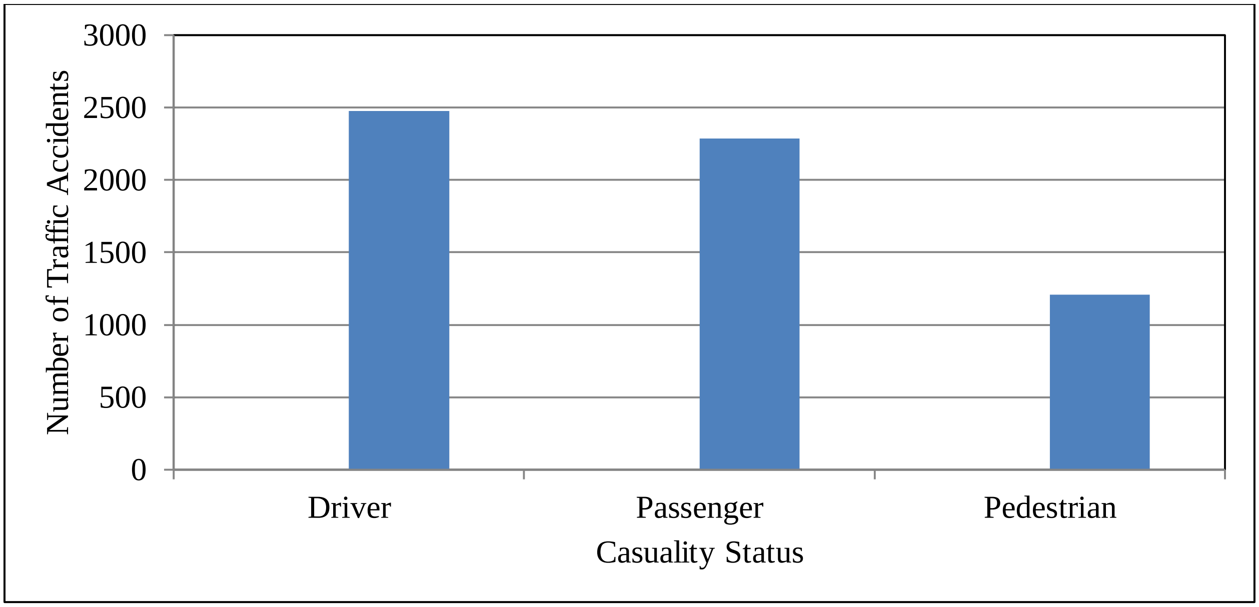

Figure 8: Number of traffic accidents versus casualty status

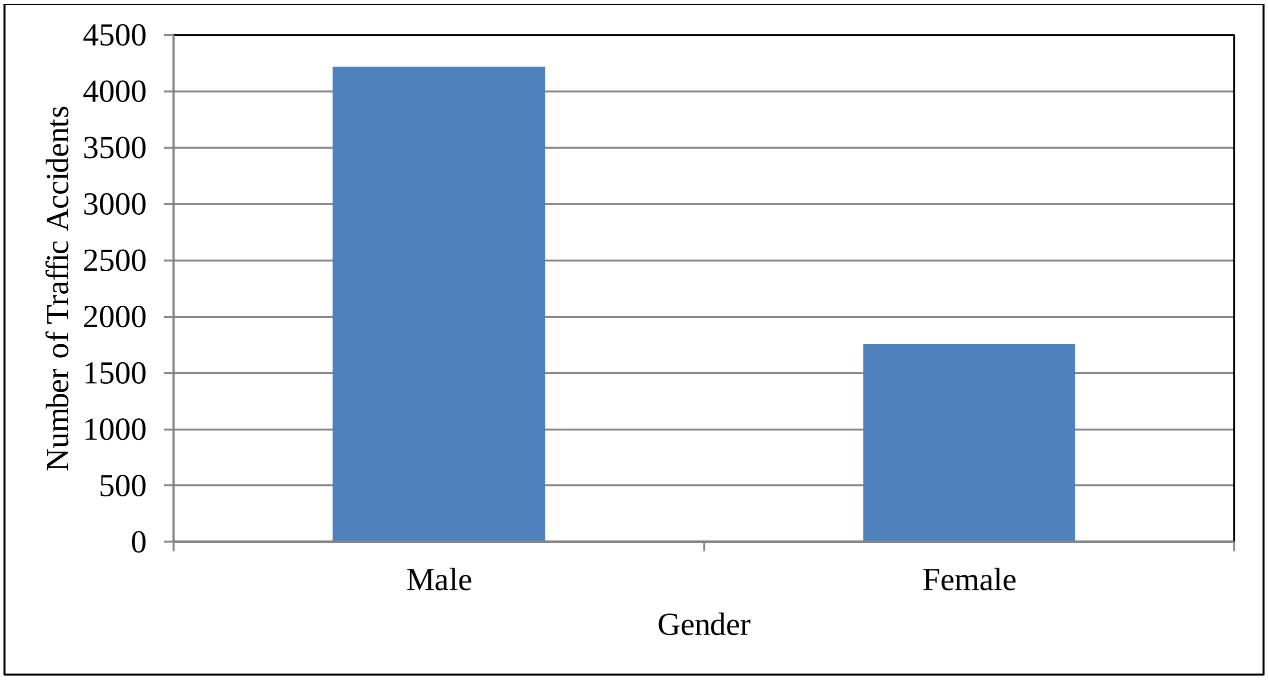

Figure 9: Number of traffic accidents versus Gender

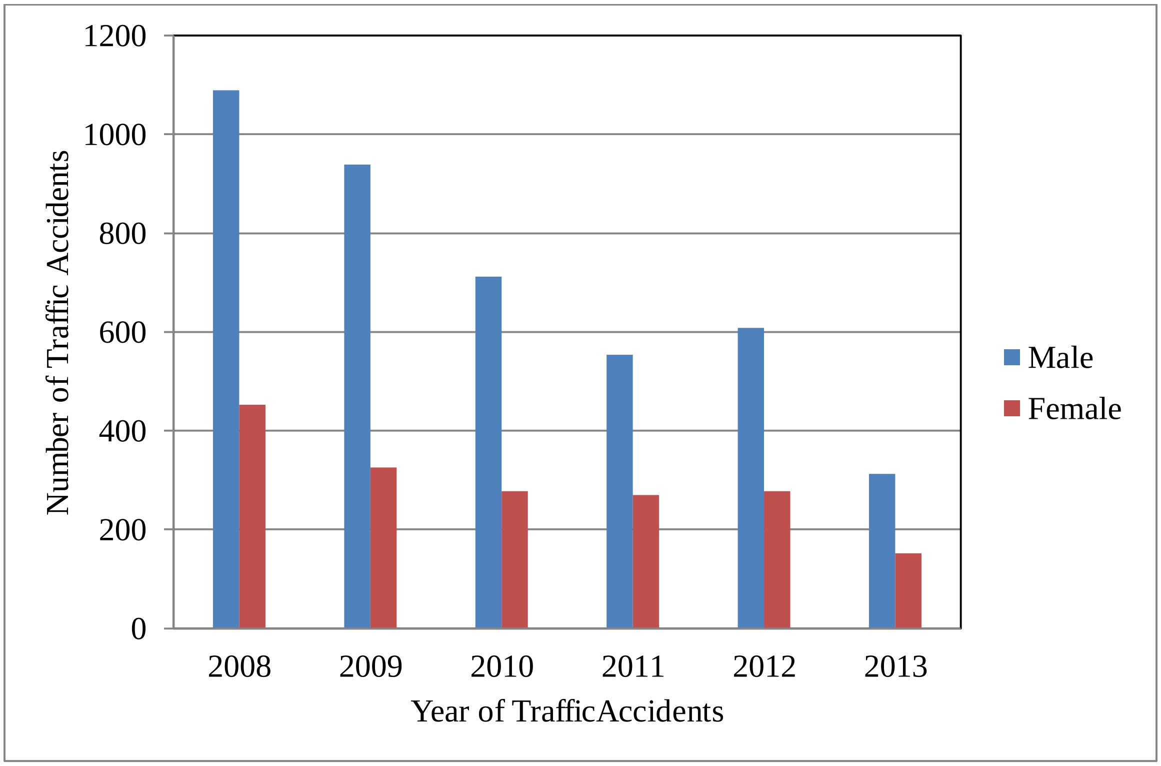

Figure 10: Number of traffic accidents versus Gender over years

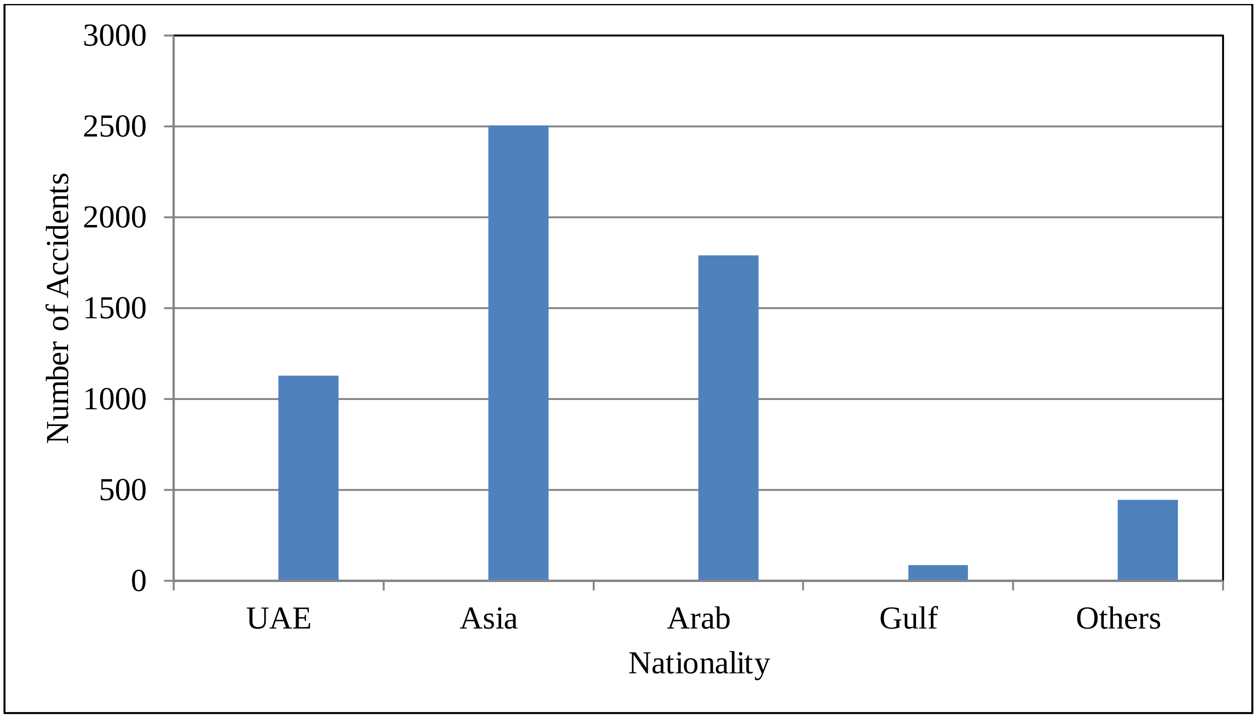

Figure 11: Number of traffic accidents versus victim nationality

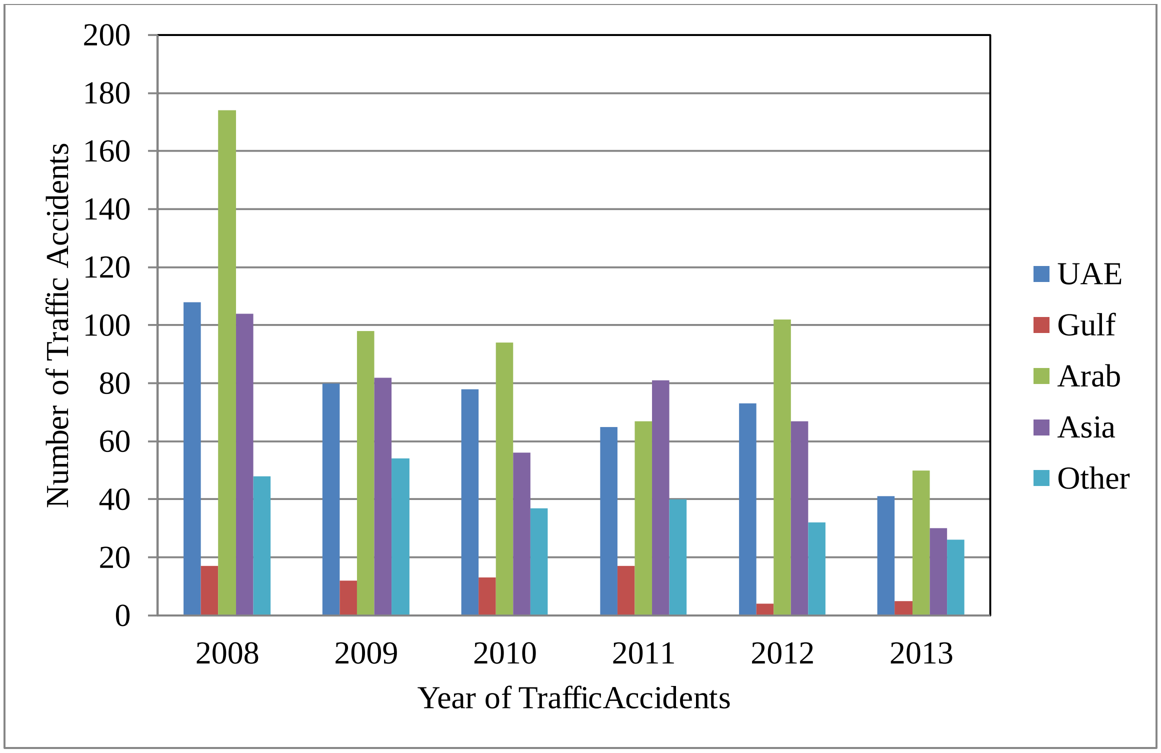

Figure 12: Number of traffic accidents versus victim nationality over years

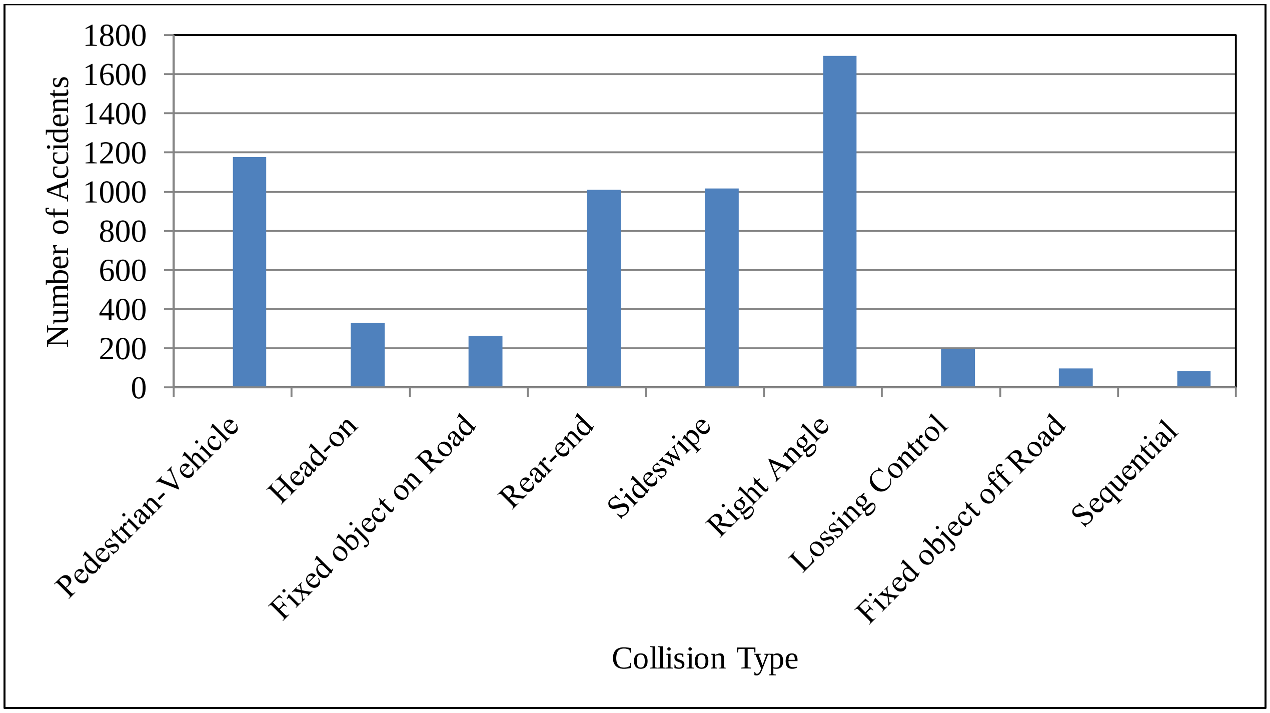

Figure 13: Number of traffic accidents versus collision type

To further support our findings regarding the most influential input variables in predicting the severity of accidents, a well-known attribute evaluator, called InfoGainAttributeEval, was used to evaluate the worth of input variables by measuring the information gain with respect to the target variable. The relative importance that this evaluator assign for each attribute is presented in Figure 14. As shown in the figure, the variables that were observed to have higher impact in determining the accident severity have the highest information gain among the other except for the gender variable. The gender variable, however, was assigned value relatively close to the value given to the seventh variable in last (i.e., speed Limit), which make it not far from being among the most influential variables.

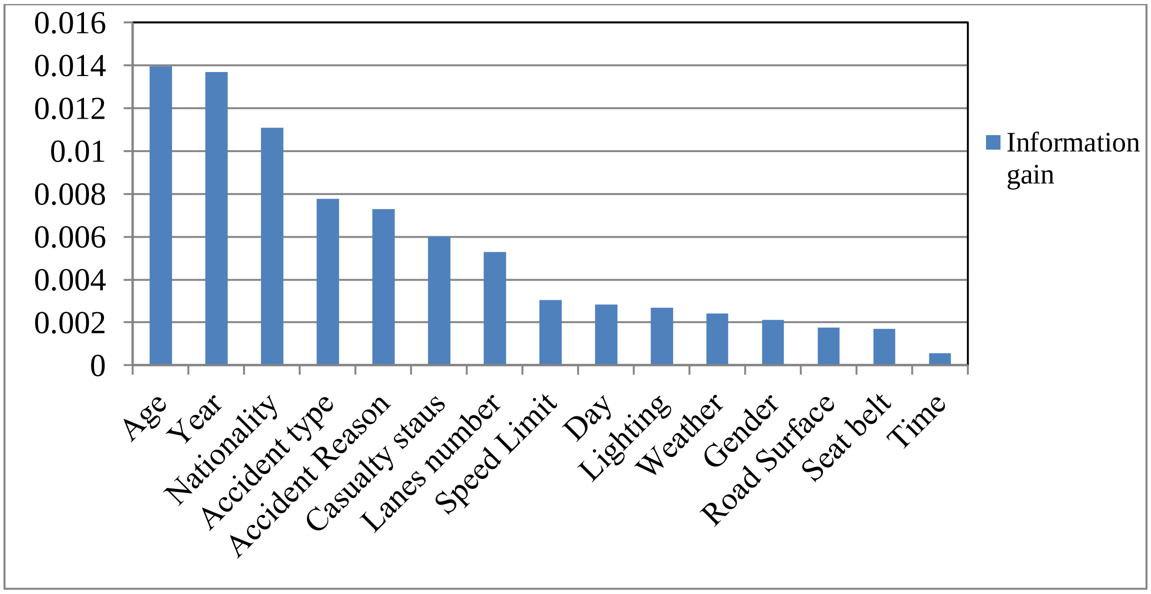

Figure 14: The impact of input variables in predicting the accidents severity

5. Conclusions

The contributions of this paper are two-fold: first, to establish models (classifiers) to predict the injury severity of any new accident with reasonable accuracy; second, to establish a set of rules that could be used by safety analysts to identify the main factors that contribute to injury severity. This paper uses WEKA software to analyze traffic accident data to validate the ability of this data-mining technique to classify traffic accidents according to their injury severity and to identify the main factors that have an impact on accident severity. In this paper, Decision Tree (J48), Rules Induction (PART), Naïve Bayes, and Multilayer Perceptron classifiers are applied to predict the severity of injury of traffic accidents based on 5973 traffic accident records from Abu Dhabi over a six-year period from 2008 to 2013. For the Decision Tree (J48) model, the overall prediction accuracy for the whole dataset used as a training set, with 10-foldcross-validation, and after resampling the training set was 81%, 73.62%, and 88.08%, respectively. For the Rule Induction (PART) model, the overall prediction accuracy for the whole dataset used as a training set, with 10-foldcross-validation, and after resampling the training set was 81%, 71.22%, and 82.18%, respectively. For the Multilayer Perceptron model, the overall prediction accuracy for the whole dataset used as a training set, with 10-foldcross-validation, and after resampling the training set was 88.01%, 73.20%, and 83.69%, respectively. For the Naïve Bayes model, the overall prediction accuracy for the whole dataset used as a training set, with 10-foldcross-validation, and after resampling the training set was 59.21%, 58.51%, and 57.05%, respectively. Based on the aforementioned, results indicate that the overall accuracy of the Decision Tree (J48) classifier, the Rules Induction (PART) classifier, and the Multilayer Perceptron classifier in predicting the severity of injury resulting from traffic accidents, using 10-foldcross-validation, is similar. Lower accuracy was observed for Naive Bayes. Additionally, the prediction accuracy of the classifiers was enhanced after resampling the training set.The most important factors associated with fatal severity for injured persons are age, gender, nationality, year of accident, casualty status, and collision type. An age of 18-30 years is the most vulnerable age group to traffic accidents for all different type of causality status. There is a clear trend of accident reduction over time. Injuries sustained in traffic accidents more frequently happen to drivers than passengers or pedestrians..UAE, Asian, and Arab nationalities have the highest traffic accident frequency. Gulf and other nationalities have lower traffic accident frequency. The highest number of traffic accidents occurred at right angles. Pedestrian-vehicle type collisions were the next most frequent, followed by rear-end collisions and sideswipe collisions.

6. References

Abellan, J., L◌́opez, G., and De Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40, 6047–6054.

Chang, L. -Y., & Chien, J. -T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science, 51(1), 17–22.

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

Chen, W. H., & Jovanis, P. P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. Transportation Research Record, 1717, 1–9.

De Oña, J., Mujalli, R. O., and Calvo, F. J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention*, 43, 402–411.

De Oña, J., López, G., & Abellán, J. (2013). Extracting decision rules from police accident reports through decision trees. *Accident Analysis & Prevention*, 50, 1151–1160.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention, 38, 434–444.

Gregoriades, A., 2007. Towards a user-centred road safety management method based on road traffic simulation. In: Proceedings of the 39th Conference on Winter Simulation: 40 years! The Best is Yet to come, Washington, DC, pp. 1905–1914.

Kashani, A. T., & Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two lane, two-way rural roads based on classification tree models. Safety Science, 49(10), 1314–1320.

Kashani, A. T., Rabieyan, R., & Besharati, M. M. (2014). A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. Journal of Safety Research, 51, 93-98.

Kopelias, P., Papadimitriou, F., Papandreou, K., & Prevedouros, P. (2007). Urban Freeway Crash Analysis. Transportation Research Record: Journal Transportation Research Board, 2015, 123–131.

Kwon, O.H., Rhee, W. and Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. Accident Analysis and Prevention, 75, 1–15.

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. Accident; Analysis and Prevention, 40, 260–266.

Mujalli, M. O., and de Oña, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research, 42, 317–326

Newgard, C. D., Lewis, R. J., & Jolly, B. T. (2002). Use of out-of-hospital variables to predict severity of injury in pediatric patients involved in motor vehicle crashes. Annals of Emergency Medicine, 39(5), 481–491.

Xie, Y., Zhang, Y., & Liang, F. (2009). Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. Journal of Transportation Engineering ASCE, 135(1), 18–25.

Yamamoto, T., & Shankar, V. N. (2004). Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident; Analysis and Prevention, 36, 869–876.

Yau, K. K. W., Lo, H. P., & Fung, S. H. H. (2006). Multiple-vehicle traffic accidents in Hong Kong. Accident; Analysis and Prevention, 38, 1157–1161.