Prediction of Traffic-Violation Using Data Mining Techniques

Md Amiruzzaman

Kent State University, Kent, OH 44242, USA
mamiruzz@kent.edu

Abstract. This paper presents the prediction of traffic-violations using data mining techniques, more specifically, when most likely a traffic-violation may happen. Also, what are the contributing factors that may cause more damages (e.g., personal injury, property damage, etc.) are discussed in this paper. The national database for traffic-violation was considered for the mining and analyzed results indicated that a few specific times are probable for traffic-violations. Moreover, most accidents happened on specific days and times. The findings of this work could help prevent some traffic-violations or reduce the chance of occurrence. These results can be used to increase cautions and traffic-safety tips.

Keywords: traffic, prediction, crime, violations, data mining

1 Introduction

According to [1] approximate population of US is 326,200,000, and there are 196,000,000 licensed drivers [2]. However, based on the data presented in [2],every day in average of 112,000 tickets are issued for different types of traffic-violations (mainly speeding). Altogether, approximately 41,000,000 tickets are issued every year (see Table 1.). The statistics provides an overview of the traffic-violations in the US, and there are number of reasons that causes traffic-violations. As the number of vehicles are increasing every day, so does the chance of traffic-violations [3]-[4]. Often, traffic-violations lead to road accidents and injuries (Chen et al., 2004; Nath, 2006).

Chen et al. in [3] classified different types of crime at different law-enforcement level. Such as, sex crime in law-enforcement level two, and theft (e.g., robbery, burglary, larceny, etc.) in law-enforcement level three. In their classification, traffic-violation is one of the common local crimes [3]. In general, bad weather, unskilled drivers, drunk drivers, and drivers who pay less attention while driving may cause traffic-violations, as well as road accidents. However, there may be some other contributing reasons that may lead to traffic-violations and road accidents. For example, speeding, reckless driving, driving under influence of drugs or alcohol, hit-and-run, road rage, etc. The research [3] mainly focused on crimes and who is committing them, rather than traffic-violations.

Table 1. Traffic-violation statistics

Driving Citation Statistics	
Average number of people per day that receive a speeding ticket	112,000
Total annual number of people who receive speeding tickets	41,000,000
Total percentage of drivers that will get a speeding ticket this year	20.6 %

Solomon et al. (2006) analyzed traffic-violation data to develop traffic safety program [4]. Their research focused on identifying places where traffic-violations occurred and how to better monitor those places. Solomon et al. (2006) proposed to use more camera/surveillance to monitor those identified high traffic-violation places and use those surveillance footages to identify responsible parties [4]. This research [4] helped to improve traffic-safety programs.

In a separate study, Saran and Sreelekha (2015) found correlations between drunk driver, careless driving, over the speed limit and road accidents [5]. However, these findings are not something new to the law-enforcement agencies and research communities. Moreover, [5] mainly focused on statistical analysis (i.e., correlation analysis) and surveillance. In their paper, Saran and Sreelekha (2015) [5] used Artificial Neural Network (ANN) for vehicle detection. They also focused on Intelligent Transport System (ITS), which incorporate latest computer technologies and computer vision [5]. Saran and Sreelekha (2015) indicated that ANN is superior in classificying moving vehicles than Support Vector Machine (SVN) and k-nearest neighbor (k-nn) algorithms. Note that, SVN and k-nn are two most popular algorithms that are widely used in data mining.

Gupta, Mohammad, Syed and Halgamuge (2016) found a correlation between crime rates and accidents from Denver city of Colorado state [6]. Note that traffic-violations may lead to violent crimes as well. For example, drunk driver may cause some property damage or injury to others. From their mining research, Gupta, et al. (2016) were able to predict that in the months of January and February, most crimes are likely to occur. These findings were helpful to the law-enforcement agencies (Gupta et al., 2016). The major drawback of [6] research is that authors only focused on one specific city of a state. Analyzing national database is necessary to understand how traffic-violations occurring in the US.

Nath (2006) indicated that most criminals along with other crimes, committed traffic-violation crimes as well [7]. One of the interesting findings from Nath (2008) was to claim that 10% criminals commits 50% of the crimes. Chen et al. (2004) mentioned that a traffic-violation is a primary concern for city, county, and state level law-enforcement agencies. In [7], authors mainly focused on where and how many Closed-Circuit Television (CCTV) would be helpful to find responsible parties.

The purpose of this study is to predict traffic-violations based on previous incidents. The national database for traffic violations is to be examined to determine any factors that contributed to previous traffic-violations and developed the prediction. Also, what time and days are most violations occur will be determined using the mining as well.

The rest of this paper is organized as follows: Section 2 describes existing literatures. Section 3 describes the method used in this study and Section 4 summarizes the experimental results. Section 5 presents discussion about the experimental results and Section 6 concludes the paper with implications and future works.

2 Literature review

Chen et al. (2004) studied different types of crime, such as traffic-violations, sex crime, theft, fraud, arson, gang/drug offenses, violent crime, and cybercrime [3]. Also, they classified these crime types to different law-enforcement levels (e.g., level one, level two, etc.). Chen et al. (2004) identified traffic-violations as level one crime and one of the common local crimes [3]. They mentioned that speeding, reckless driving, causing property damage or personal injury in a collision, driving under influence of drugs or alcohol, hit-and-run, and road rage are common reasons for traffic-violations [3]. According to Chen et al. (2004), traffic-violations mostly considered as less harmful crime, however, sometimes this type of crime could cause severe bodily injury or property damage [3]. Even though, Chen et al. (2004) [3] discussed about traffic-violation and other crimes, but their work actually did not focus on traffic-violation analysis. Rather, their work focused on other types of crime analysis and prediction of those crimes to help law-enforcement agencies.

Solomon, Nguyen, Liebowitz and Agresti (2006) demonstrated how to use data mining (DM) and evaluate cameras that monitor red-light-signals in traffic intersections [4]. Based on their findings they proposed some techniques to improve traffic safety programs. In their work, they used different modeling techniques, such as decision trees, neural networks, market-basket analysis, and k-means. Solomon et al. (2006) focused on identifying places where red-light-signal violations occurred and how to better monitor those places. The red-light violation is known as red light running (RLR), and according to the Federal Highway Administration (FHWA), approximately 1,000 Americans were killed and 176,000 were injured in 2003 because of RLR.

To describe the severity of RLR and its damage on the economy, Solomon et al. (2006) in [4] wrote, “The California Highway Patrol estimates that each RLR fatality costs the United States $2,600,000 and other RLR crashes cost between $2,000 and $183,000, depending on severity (California State Auditor, 2002)” (p. 621). As for the recommendation, they proposed to use more camera/surveillance to monitor those identified high traffic-violation places and use those surveillance footages to identify responsible parties. As for their data, they used traffic-violation data from Washington, DC area; the data was collected between the year 2000 and 2003 (Solomon et al., 2006). In terms of findings, their [4] work helped law-enforcement agencies to find responsible parties using the red light camera (RLC). However, placing RLCs in a right place is not an easy task. Data mining technique can be helpful to determine the high accident zone and place RLCs in appropriate locations.

In a separate study [5], Saran and Sreelekha (2015) found correlations between drunk driver, careless driving, over the speed limit and road accidents. However, these findings are not something new to the law-enforcement agencies and to the research communities [5]. Their work [5] was more of a classification than data mining. They used videos obtained from closed circuit television (CCTV) cameras placed in roadsides or driveways are used for the surveillance. They used artificial neural networks (ANN) to detect different types of vehicles [5]. While detecting different types of vehicles are important and interesting work, however, the need for traffic-violation data mining remain the unsolved. In their work [5], Saran and Sreelekha (2015) mainly focused on road safety and surveillance system.

Gupta, Mohammad, Syed and Halgamuge (2016) found a correlation between crime rates and accidents from Denver city of Colorado state. Note that traffic-violations may lead to violent crimes as well [6]. For example, drunk driver may cause some property damage or injury to others. To describe the phenomenon, they said in [6] “The major cause of road accidents is drink driving, over speed[ing], carelessness, and the violation of traffic rules” (p. 374). From their mining research, Gupta, et al. (2016) were able to predict that in the months of January and February, most crimes are likely to occur. These findings were helpful to the law-enforcement agencies (Gupta et al., 2016). They used data from the National Incident-Based Reporting System (NIBRS), The dataset contained 15 attributes and 372,392 instances [6]. While, Gupta, et al. (2016) in [6] presented interesting findings based on their data mining research, however, their work is mainly focused on a specific city of a specific state. It is important that a research study focus on the entire US and try to generalize the findings mentioned in [6].

Nath (2006) in [7] indicated that most criminals along with other crimes, committed traffic-violation crimes as well. One of the interesting findings from Nath (2008) was to claim that 10% criminals commits 50% of the crimes. Chen et al. (2004) mentioned that a traffic-violation is a primary concern for city, county, and state level law-enforcement agencies. They also added that traffic-violations and other criminal activities may be related, and information obtained from traffic-violations can be further used to find criminals. They focused on getting contact information from the Department of Motor Vehicles (DMV).

This paper, will provide an overview of traffic-violation data mining as well as some interesting findings that can be helpful to maintain cautions and prevent unwanted traffic-violations. The proposed data mining predicts where and what time of the day the incidents (traffic-violations) will occur based on National database. Also, what combinations of factors contribute to traffic-violations.

3 Method

Several data mining algorithms were used to analyze the data. For example, Naïve Bayes, J48 decision tree, Decision Table, and Support Vector Machine. Also, a few statistical analysis, such as, linear regression analysis, correlation analysis, and reliability analysis were considered to analyze the final data. Multiple tools were used to process and analyze the data. For example, SPSS (i.e., Statistics is a software package developed by IBM company) tests helped to determine which attributes should be considered for data mining. Also, WEKA[1] (i.e., Waikato Environment for Knowledge Analysis) tool was used to perform data mining algorithms [8] on the research dataset.

3.1 Data

The data was downloaded from the national database for public data[2]. The original database consists of 36 attributes. However, there were lots of attributes that did not show any variations. For example, the accident attribute only had ”No” as a value. Attributes like that does not contribute to data analysis, so, those attributes were deleted before the final analysis. The database consisted over one million records. Of course, some of the rows had some missing values or wrong values (e.g., human errors). Missing values and wrong values seemed to be due to user errors. The database included demographic information, such as, gender of vehicle drivers, and place of incidents, driver state, driver city, etc.

[1] https://www.cs.waikato.ac.nz/~ml/weka/downloading.html

[2] https://catalog.data.gov/dataset

3.2 Preprocessing

The initial task for the preprocessing was to identify which attribute to keep and which attributes to discard. Of course, the database included overwhelming amount of data. However, for the data mining, only the most important and relevant attributes were considered for final analysis. The preprocessing process included deleting missing data, deleting irrelevant attributes, modifying records to meaningful format, etc.

- SPSS tests helped to determine which attribute could to be deleted or not included for data mining as well as final analysis (see Table 2.).
- Missing and repeating attributes were discarded as well. Also, wrong entries were discarded from final selection of data analysis.
- The dataset was divided into training set and testing set. The training set consisted 67% of the data, whereas testing test consisted of 33% of the total number of records. Holdout method was used to determine the training set and testing set.

Table 2. Inter-Item correlation matrix

	Personal Injury	Property Damage	Alcohol	Contributed To Accident
Personal Injury	1.000	-0.016	0.013	0.346
Property Damage	-0.016	1.000	0.019	0.368
Alcohol	0.013	0.019	1.000	0.014
Contributed To Accident	0.346	0.368	0.014	1.000

Initial processing After the determining the training set and testing set, and deciding to keep some candidate attribute. Again, SPSS tests were executed to determine which attribute should be deleted to further increase the accuracy of the result. Mainly the test helped to determine which item should be deleted is ”items-deleted” to increase the reliability value. For example, SPSS tests indicated time of the incident should be deleted to increase the reliability of the results.

Initial results Initial processing suggested that most traffic-violations happened in Maryland (DC), more specifically in Washington, DC area. Also, after modifying the date of incident to weekdays (e.g., Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday), it was noticed that most traffic-violations happend on Tuesday and Wednesday (see Figure 1.). This is maybe because people are more anxious on mid-week (i.e., we call it mid-week effect).

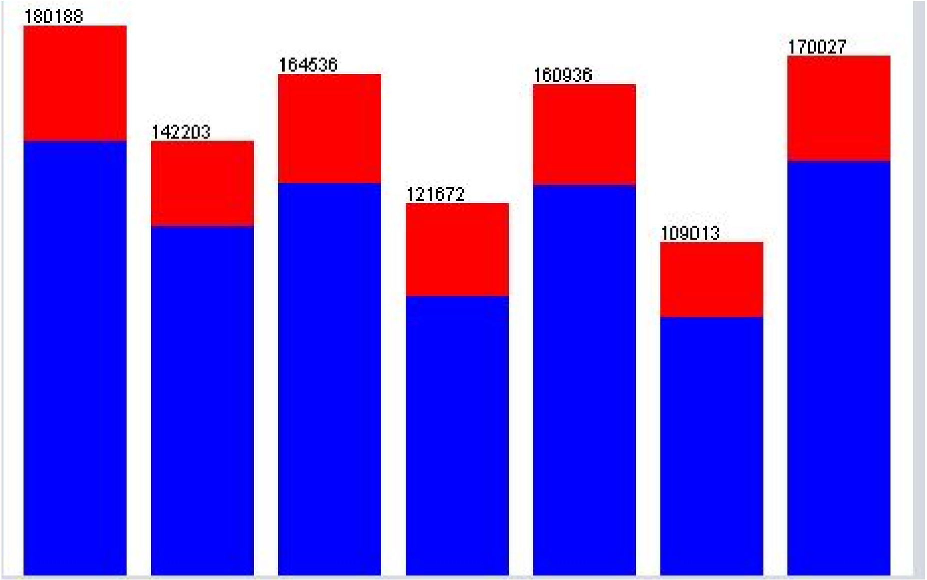

Fig. 1. Number of incidents in days. (x-axis is days–Sunday (starting from left), and end with Saturday (on the right); y-axis is the number of incidents.)

4 Results

4.1 SPSS

Correlation analysis helped to determine that property damage and alcohol were correlated (17%). Similarly, contributed to accident and property damage were correlated (34%); contributed to accident and personal injury were correlated (37%). The correlation values were calculated using the following equation (see Eq. 1)

$$r_{xy} = \frac{\sum_n^{i=0}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_n^{i=0}(x_i - \bar{x})^2 \sum_n^{i=0}(y_i - \bar{y})^2} \qquad (1)$$

where,
r_{xy} is the correlation value between variables,
x and y, $\sum$ is the symbol for ”sum up”,
x_i is the individual value of variable x,
$\bar{x}$ is the mean of variable x.
Similarly, y_i is individual value of variable y,
$\bar{y}$ is the mean of variable y.

In this analysis linear regression was used to verify some of the prediction made by the WEKA software. The regression equation can be expressed as (see Eq. 2)

$$y_i = a + bx_i + c \tag{2}$$

where,
Y is the dependent variable that the equation tries to predict,
X is the independent variable that is being used to predict Y,
$x_i \in X$, and $i = 1, 2, 3, ..., n$,
$y_i \in Y$, and $i = 1, 2, 3, ..., n$,
a is the Y-intercept of the line,
b is the slope,
and c is a value called the regression residual, which can be calculated by $|\hat{y_i} - y_i|$, where $\hat{y_i}$ is the expected value of y.
The values of a and b are selected so that the square of the regression residuals is minimized.

More detail about regression eqation and example of regression can be found online[3]. The results obtained from linear regression analysis is presented in Table 3.

Table 3. Linear regression analysis

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	0.404	0.163	0.163	0.125

Reliability values were calculated using equation below (see Eq. 3)

$$\alpha = \frac{N \times \bar{c}}{\bar{v} + (N - 1) \times \bar{c}} \tag{3}$$

where,
N is the number of items $\bar{c}$ is average iter-item covariance
and $\bar{v}$ is average variance.
The reliability of four attributes (i.e., personal injury, property damage, alcohol, and contributed to the accident) was 0.435 (see Table 4.)

4.2 Naïve Bayes

The Naïve Bayes classifier is one of the most popular classifiers in data mining. To describe the strength of Naïve Bayes [9] wrote " The naïve Bayes classifier

[3] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

Table 4. Reliability statistics

Cronbach's α	Cronbach's α based on standarized items	N of items
0.435	0.362	4

computes the likelihood that a program is malicious given the features that are contained in the program. This method used both strings and bytesequence data to compute a probability of a binary's maliciousness given its features" (p. 6). Results obtained from Naïve Bayes is prsented in Table 5.

Table 5. Comparisions of different methods

Method name	Correctly classified (%)	Incorrectly classified (%)	Kappa statistics	Root Mean Square Error (RMSE)	precision	recall
J48 decision tree	97.67	2.32	0.24	0.14	0.98	0.99
Naïve Bayes	97.60	2.39	0.06	0.13	0.97	0.99
Support Vector Machine (SVM)	97.61	2.38	0.00	0.15	0.97	1.00
Decision table	97.64	2.35	0.24	0.13	0.98	0.99

Following the mathematical definition will help to explain how the Naïve Bayes classifier works.
Let,
the dataset be d,
and set of classes $C = c_1, c_2, ..., c_n$, and predicted class $c \in C$.
The Naïve Bayes classification can be expressed as (see Eq. 4),

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{4}$$

Over 500,000 instances were analyzed using Naïve Bayes (Weka could not return any results over 0.5 million records).
67% of them as training set and 33% of them as testing set.

The confusion matrix helped to compute the accuracy of classifying algorithms. Therefore, the accuracy of a classifying algorithm can be defined as (see

Eq. 5),

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{5}$$

here, TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative

With 97.6% accuracy Naïve Bayes algorithm was able to classify traffic violations—personal injury, property damage, and the presence of alcohol. The confusion matrix of Naïve Bayes has shown that only 297 records were classified as ”True Negative” (see Table 6)

Table 6. Confusion matrix (Naïve Bayes)

		Predicted class	
		No	Yes
Actual class	No	True positive = 327107	False negative = 331
	Yes	False positive = 7715	True negative = 297

In the database different types of vehicle was reported. For example, motorcycle, automobile, station wagon, limousine, etc. Naïve Bayes algorithm was able to classifiy traffic-violations based on vehicle type with accuracy of 87.444%. Also, Naïve Bayes algorithm reported that automobile had the highest incident records.

4.3 J48

The J48 decision tree algorithm was used to visualize and determine how prediction was made. In fact, J48 algorithm uses a mathematical model to determine information gain can help to determine which variable fits better in terms of target variable prediction. There are other data mining research, such as [10] used J48 decision tree to predict their outcome variables as well.

Following the mathematical definition will help to explain how SVN classifier works.
Let,
the dataset be d,
The dependent variable is Y (i.e., the target variable that the algorithm is trying to classify).
The dataset d is consists of vector x, which is composed of the features, $x_1, x_2, x_3, \ldots$ etc. that are used to make the classification or the decision tree. Then, the decision tree algorithm can be expressed as (see Eq. 6)

$$(x, Y) = (x_1, x_2, x_3, \ldots, x_k, Y) \tag{6}$$

where, k is number of features in vector x.

Around 5:00 pm, the traffic-violation happened did not involve alcohol, which make sense as most people leave their work at that time. However, perhaps the rush to go home may cause those traffic-violations at that time. On the other hand, most traffic-violations between 12:00 am and 1:00 am involved alcohol, which indicates that those occurred by drunk drivers. Perhaps, law-enforcement agencies should look into those incidents and maintain more cautions. The J48 algorithm classified with the accuracy of 97.6% correct classification. The confusion matrix of J48 has shown that only 1290 records were classified as "True Negative" (see Table 7)

Table 7. Confusion matrix (J48)

		Predicted class	
		No	Yes
Actual class	No	True positive = 326350	False negative = 1088
	Yes	False positive = 6722	True negative = 1290

In addition, the J48 algorithm was able to classify traffic-violations based on vehicle type with accuracy of 87.433%. Also, J48 algorithm reported that automobile had the highest incident records.

4.4 Support Vector Machine (SVM)

Support vector machine (SVM) is one of the powerful data classification tools. The SVM was invented at ATT Bell Laboratories by Cortes and Vapnik in 1997 [11]. To describe the strength of SVM classification algorithm Kim, Pang, Je, Kim, Bang and Yang (2003) in [11] wrote, " The SVM learns a separating hyperplane to maximize the margin and to produce a good generalization ability" (p. 2757).

Witten and Frank (2009) in [12] mentioned, "Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible" (p. 188)

Following the mathematical definition will help to explain how SVN classifier works.
Let,
the dataset be d,
and set of classes $C = c_1, c_2, ..., c_n$,
and predicted class $c \in C$.

Also, the input set $X = x_1, x_2, ..., x_n$ and $x \in X$.
Here, X is input and C is output.
Now, if we want to classify $c = f(x, \alpha)$,
where,
α are the parameters of the function, then SVN can be expressed as (see Eq. 7)

$$f(x, \{w, b\}) = sign(w \times x + b) \tag{7}$$

where, w is weight and b is bias.

SVN algorithm was able to classify traffic-violations based on vehicle type with accuracy of 87.433%. Also, reported that automobile had the highest incident records. The confusion matrix shows the accuracy of SVM classifier (see Table 8)

Table 8. Confusion matrix (SVM)

		Predicted class	
		No	Yes
Actual class	No	True positive = 327438	False negative = 0
	Yes	False positive = 8012	True negative = 0

4.5 Decision Table

The Decision Table (DT) is a rule based classification model is ”Decision table”. This type of method generates rules of associations from the data and groups the data or classifies the data. The decision table uses best-first search and cross-validation for evaluation [12].

here, the symbol ”$\overset{def}{=}$” represents defining relationship. Let, $f(x) \overset{def}{=} x + 1$ definies the ralationship of x with function f. In terms of predicting relationship using DT can be defined as (see Eq. 8)

$$R(x, y) \overset{def}{=} y = x \tag{8}$$

where, R is relationship function between x and y. Which indicates that some y helps to predict x

DT algorithm was able to classify traffic-violations based on vehicle type with accuracy of 87.451%. The DT analysis reported that automobile had the highest incident records. The confusion matrix shows the accuracy of SVM classifier (see Table 9)

Table 9. Confusion matrix (Decision table)

		Predicted class	
		No	Yes
Actual class	No	True positive = 326203	False negative = 1235
	Yes	False positive = 6664	True negative = 1348

5 Discussion

5.1 Learning from the data processing

The original data was download as comma-separated values (CSV) file. However, I was important that csv file should be converted to WEKA supported file format. A Java program was written to csv file to Attribute-Relation File Format (arff) file format. During the conversion process, it was discovered that arff file is sensitive to date format. What format is used in the file should be explicitly mentioned in the original arff file, otherwise WEKA software cannot recognize the data type.

During the data processing and analyzing from visualization tool provided by WEKA, it was discovered that WEKA support csv file as input as well.

In order to make sense of time of incident, time attribute was discretized to nearest hour value. So, all time was discretized to 24-hour format, excel function was used to accomplish this task (e.g., MROUND(B2, "1:00")). Also, during the presentation and feedback from experts, it was suggested to include date of the incident. However, date was not much informative. So, date was converted to day; built-in excel function was used to convert date to day number (e.g., WEEKDAY(A2), and then format was changed to dddd to get the day).

During the analysis κ value was calculated; κ value measures relative improvement over random predictor. The κ statistics was computed using following equation (see Eq. 9)

$$\kappa = \frac{D_{observed} - D_{random}}{D_{perfect} - D_{random}} \tag{9}$$

In terms of success, precusion and recall values were calculated as well. For precision Eq. 10 was used.

$$precision = \frac{TP}{TP + FP} \tag{10}$$

where,
number of true positive is TP,
and number of false positive is FP.

Comparisions of different algorithm in terms of precision is shown in Table 10.

Table 10. Precision comparision

Naïve Bayes	J48	SVM	Decision Table
0.977	0.980	0.976	0.980

For recall value Eq. 11 was used.

$$recall = \frac{TP}{TP + FN} \tag{11}$$

where,
number of true positive is TP,
and number of false negative is FN.

Comparisons of different algorithm in terms of recall is shown in Table 11.

Table 11. Recall comparision

Naïve Bayes	J48	SVM	Decision Table
0.999	0.997	1.000	0.996

After obtaining *precision* and *recall* values, $F - statistics$ was computed (see Eq. 12).

$$F - statistics = \frac{2 \times recall \times precision}{recall + precision} \tag{12}$$

Comparisons of different algorithm in terms of $F - statistics$ is shown in Table 12. All algorithsm provided same $F - statistics$ value.

Table 12. F-measure comparision

Naïve Bayes	J48	SVM	Decision Table
0.988	0.988	0.988	0.988

To evaluate the prediction accuracy, root mean-squared error ($RMSErrors$) was computed (see Eq. 13).

$$RMSErrors = \sqrt{\sum_{n}^{i=1} (\hat{y}_i - y_i)^2} \tag{13}$$

where, y_i is the observed value for the ith observation
and $\hat{y}_i$ is the predicted value.

Comparisons of different algorithm in terms of root mean square error is shown in Table 13.

Table 13. Root mean-squared error ($RMSErrors$) comparision

Naïve Bayes	J48	SVM	Decision Table
0.132	0.143	0.152	0.131

6 Conclusion

Obtained results from data mining and statistical analysis suggested that personal injury was a must, if driver is drunk. Also, around 1:00 am was the most dangerous time to go out (see Figure 2.); most property damage and personal injury happened because of drunk drivers between 11:00 pm to 1:00 am. This was the time when most incidents occurred as well. Among all the cities, DC area seemed to be more consistent with these results. Therefore, if you are in the DC area during this specified times, then try not to hang out in the DC area at that time.

Perhaps, analyzing more data and latest database from law-enforcement agencies could help us to find more interesting information. Also, use different data mining algorithms could help to understand the data better as well. Having a domain expert could be beneficial to interpret the findings and add more implications.

As for the future study, visualization technique can be used to visualize the intensity of traffic violations over gegraphic locations, and accident prone areas. Moreover, deep learning can be applied to identify or classify areas based on their voilation probability as well.

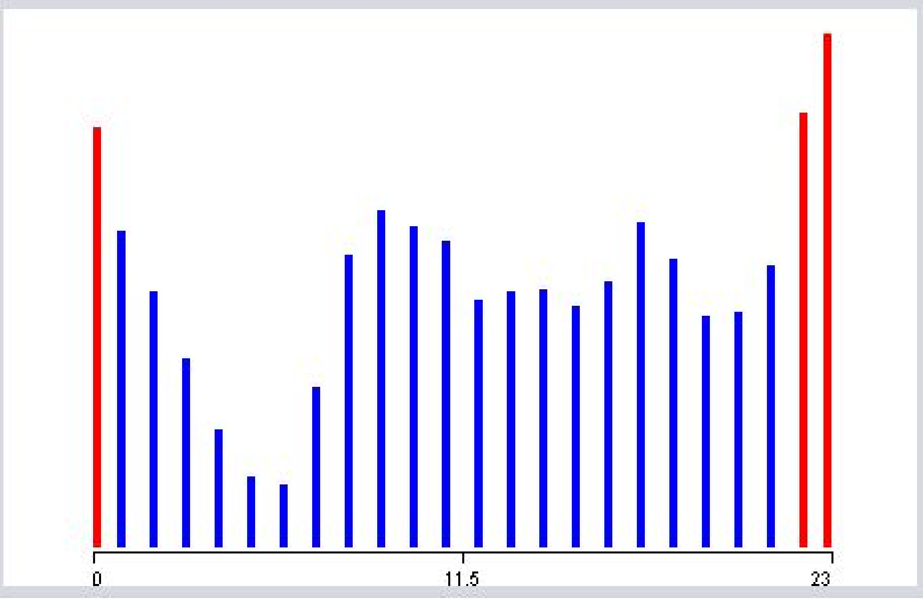

Fig. 2. Number of traffic-violations in 24 hours. (x-axis is hours–0 or 24 (starting from left), then 1, 2, and end 23 (right); y-axis is number of incidents.)

Acknowledgment

The author would like to thank to open data website[4] for making the dataset available for research and analysis. A special thank you to those who participated in the initial presentation and provided valuable feedback (part of this paper was presented and was submitted as a class project). Also, thank to Dr. Kambiz Ghazinour for helping me to think further about the data and analysis process.

References

1. Estimates, A.P.: U.S. and world population clock (2017) Accessed: 11-19-2017.
2. Brain, S.: Driving citation statistics (2016) Accessed: 11-20-2017.
3. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples. computer **37**(4) (2004) 50–56
4. Solomon, S., Nguyen, H., Liebowitz, J., Agresti, W.: Using data mining to improve traffic safety programs. Industrial Management & Data Systems **106**(5) (2006) 621–643
5. Saran, K.B., Sreelekha, G.: Traffic video surveillance: Vehicle detection and classification. In: 2015 International Conference on Control Communication and Computing India (ICCC). (2015)

[4] https://catalog.data.gov/dataset

6. Gupta, A., Mohammad, A., Syed, A., Halgamuge, M.N.: A comparative study of classification algorithms using data mining: Crime and accidents in denver city the usa. Education **7**(7) (2016)
7. Nath, S.V.: Crime pattern detection using data mining. In: Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on. (2006) 41–44
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11**(1) (November 2009) 10–18
9. Schultz, M.G., Eskin, E., Zadok, F., Stolfo, S.J.: Data mining methods for detection of new malicious executables. In: Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on, IEEE (2001) 38–49
10. Olson, D.L., Delen, D., Meng, Y.: Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems **52**(2) (2012) 464–473
11. Kim, H.C., Pang, S., Je, H.M., Kim, D., Bang, S.Y.: Constructing support vector machine ensemble. Pattern recognition **36**(12) (2003) 2757–2767
12. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Second edn. Elsevier Inc (2005)