11

Applying Big Data Analytics on Motor Vehicle Collision Predictions in New York City

Dhanushka Abeyratne and Malka N. Halgamuge

Department of Electrical and Electronic Engineering, Yellowfin (HQ), The University of Melbourne, Australia

11.1 Introduction

11.1.1 Overview of Big Data Analytics on Motor Vehicle Collision Predictions

Due to population growth, there are more traffic accidents, which have become a global concern. In the past, researchers have conducted studies to find out the common cause for motor vehicle collisions. Log-linear models, data mining, logical formulation, and fuzzy ART are some of the methods widely used to perform research [1]. Even with the use of these methods, data analysis is a complicated process. However, with the improvements in technology, data mining is defined to be a highly accurate method for the analysis of big data.

With the use of big data application, few researchers have focused mainly on understanding the significance of vehicle collisions. A research performed by Shi et al. [2] explains the significance of identifying the traffic flow movements on highways to minimize the impact on vehicle collisions. This research has used a time series data approach under clustering analysis to comprehend traffic flow movements. It was converted using the cell transformation method.

A similar study by Yu et al. [3] considers traffic data mining to be a big data approach. The author further carries out the study using past traffic data with the application of common rules via data mining, which is based on a cloud computing technique. Traffic trend prediction and accident detection within a MAP-reduce framework are used. However, the data sets have many missing pieces and were redundant with values that cannot be normalized [1].

Data mining is a widely used computing technique adapted to determine unfamiliar patterns in a large data set [4]. For prospective use, data mining is a comprehensive field that presents meaningful patterns. These techniques are classified into three categories; classification, prediction, and data clustering, in order to trace seasonal trends and patterns [4]. Classification is a commonly used method that forecasts unknown values acknowledged to generated mode [5].

This chapter proposes a framework that uses a classifying technique for collision predictions that has been undertaken by Python programming language. Python is an effective

Intelligent Data Analysis: From Data Gathering to Data Comprehension,
First Edition. Edited by Deepak Gupta, Siddhartha Bhattacharyya, Ashish Khanna, and Kalpna Sagar.

programming language that is used in the theoretical and mathematical analysis for large data sets [6]. Python programming language supports various data mining techniques and algorithms that mainly clusters and classifies. As this option has many beneficial features, it is one of the most suitable tools for making scalable applications. Thus, it can be utilized for the framework of big data analysis in wide motor vehicle collision data sets to obtain reliable results.

A specific data set has been obtained for this chapter about vehicle collision in a large city in the United States (New York City). A data mining technique is exercised to perform further analysis on the data set. Based on recent news reports, New York City (NYC) roads are believed to have an increase in motor vehicle collisions. The National Safety Council has conducted a preliminary survey that confirms that the year 2016 had the deadliest accidents on NYC roads over several decades [7]. Thus, there is a need to predict and assess the association of vehicles involved in a collision and their attributes found in the data set.

11.2 Materials and Methods

This chapter uses an analytical approach to data mining, which forecasts relevant attributes corresponding to the source of other related attributes. Analysis of variance (ANOVA) table generated from *k*-means clustering, *k*-nearest neighbor (*k*NN), naïve Bayes, and random forest classification algorithms are used in this chapter to understand the association of statistical data collected.

11.2.1 Collection of Raw Data

In this chapter the data is collected from an open data website of the New York Police Department (NYPD) for the period 2012–2017. The data set contains information about motor vehicle collisions occurred in NYC with all collision-related attributes.

11.2.2 Data Inclusion Criteria

This raw data set consists of 26 attributes and 1 048 575 cases filtered down using vehicle type code. All the unknown vehicle type codes and blank values have been removed. Filtered data includes 14 attributes. Date attribute was further expanded to generate four attributes (day of the week, day, month, and year). As an example, 1/1/2017; day of the week = Sunday, day = 1, month = Jan, and year = 2017. Finally, the selected collisions data set contains 998 193 and 17 attributes of data (Table 11.1).

11.2.3 Data Preprocessing

The study of Sharma and Bhagat [8] recognizes that the raw data gathered contains noisy data, missing values, data dependency, and multiple sources of data. Thus, the preprocess method identified as the initial process of data mining is recommended to make raw data more reliable. It has three stages: data cleaning, data integration, and data transformation. Primarily, data cleaning excludes missing values and errors while recognizing attributes of

Table 11.1 Illustration of data set attributes.

Attributes	Description
Day of the week	Day of the week collision occurs (Mon–Sun)
Day	Day of the month collision occurs (1–31)
Month	Relevant month collision occurs (Jan–Dec)
Year	Relevant year collision occurs (2012–2017)
Time	Relevant time of the day collision occurs
Borough	Location where the collision happened.
Latitude	Geographical location
Longitude	Geographical location
Number of persons injured	Number of people were injured in the collision.
Number of persons killed	Number of people were killed in the collision.
Number of pedestrians injured	Number of pedestrians were injured in the collision.
Number of pedestrians killed	Number of pedestrians were killed in the collision.
Number of cyclists injured	Number of cyclists were injured in the collision.
Number of cyclists killed	Number of cyclists were killed in the collision.
Number of motorists injured	Number of motorists were injured in the collision.
Number of motorists killed	Number of motorists were killed in the collision.
Vehicle_Group	Group of vehicles which caused the collision

Table 11.2 Categorized vehicle groups.

Vehicle group	Vehicle type-codes
Large_Vehicle	Bus, fire truck, large commercial vehicle
Medium_Vehicle	Pick-up truck, small commercial vehicle, van, livery vehicle
Small_Vehicle	Passenger vehicle, sort-utility/wagon, taxi, pedicab
Very_Small_Vehicle	Motorcycle, scooter, bicycle

raw data. Subsequently, data integration links the data into a reliable structure. In the last part of preprocessing, data is converted into acceptable forms for data mining [8].

During the data preprocessing, the vehicle type-code is further categorized into four groups, depending on vehicle passenger capacity and size of the vehicle. This developed attribute is then added for analysis into the data set. Table 11.2 illustrates categorized groups.

11.2.4 Data Analysis

In this chapter, data has been analyzed using the Python programming language. Python programming language presents significant support for the process of experimental

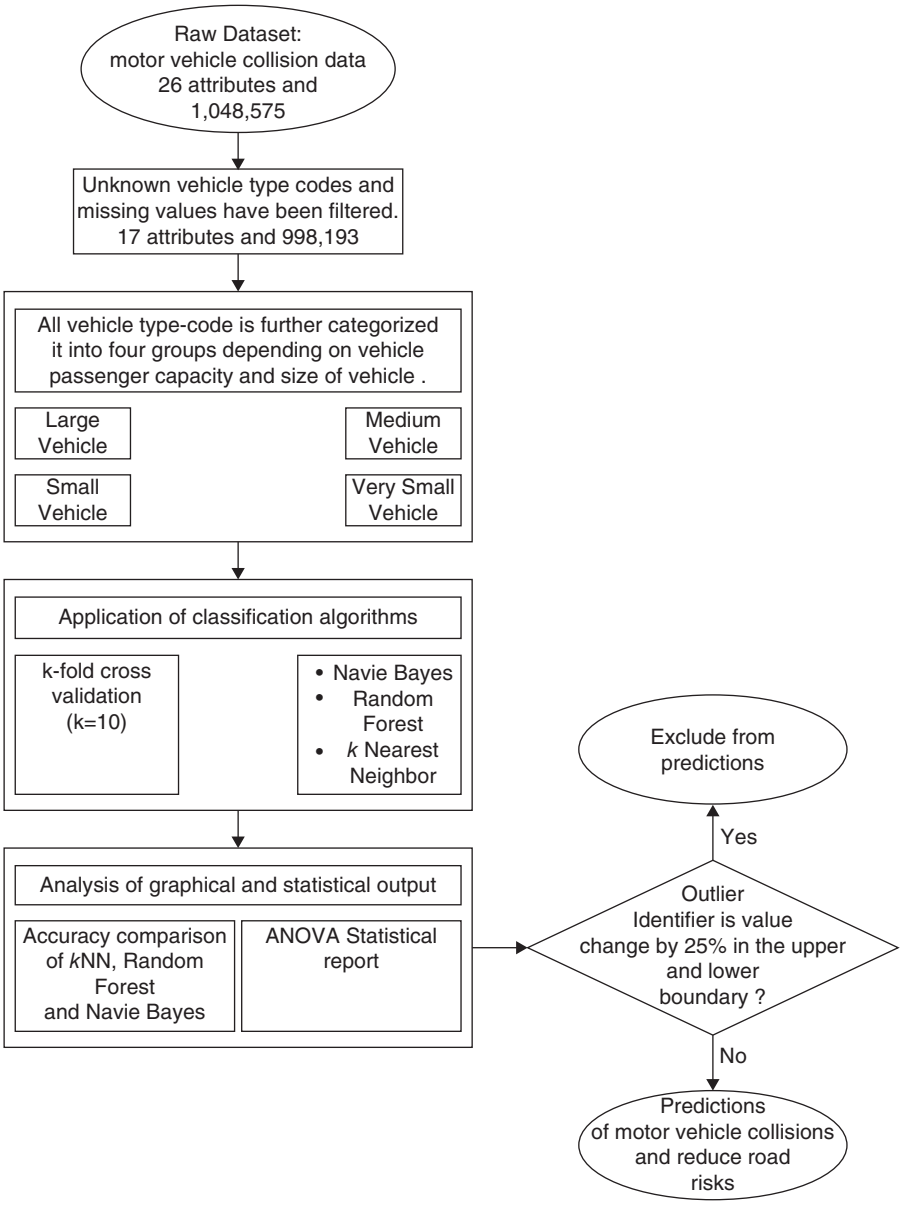

Figure 11.1 Overall methodology of data analysis process.

data mining, combined with categorizing data inputs, for measuring statistical learning schemes. Relevant statistical data is generated rapidly and is precise. Further, it visualizes the input data and learning outcome for a large set of data that becomes clearer [6]. Figure 11.1 illustrates the overall data analysis process of this chapter.

Table 11.3 Description of classification algorithms and functionalities.

Classifiers	Description
Naive Bayes	Naive Bayes is a probabilistic classifier that uses a statistical approach for classification [9].
*k*NN	*k* nearest neighbor is a simple, widely known and efficient algorithm for pattern recognition. It classifies samples considering the class of nearest neighbor [10].
Random Forest	Random forest is a combination of tree-structured predictors via tree classification algorithm. It is identified as an automatic compensation mechanism with the advantages of speed, accuracy, and stability [11].

11.3 Classification Algorithms and K-Fold Validation Using Data Set Obtained from NYPD (2012–2017)

11.3.1 Classification Algorithms

Table 11.3 demonstrates the description of the applied classification algorithms and its functionalities.

11.3.1.1 k-Fold Cross-Validation

The k-fold cross-validation is identified as a well-known statistical experimental technique where k disjointed blocks of objects are produced by the database with the random division. Later, the data mining algorithm is formulated using k−1 blocks and the remaining block is applied to test the function of algorithm; this is a repetitive process by k times [12]. K-fold (K = 10) cross-validation method is used to analyze data in this chapter.

This chapter follows three main steps for data analysis. The three steps are as follows:

1. Data selection
2. Data preprocessing
3. Data analysis using classification methods

All three algorithms are systematically carried out for data analysis using K-fold validation.

Algorithm 11.1 Data Selection

```
LOAD NYPD_collision_data
STRING [] VTC= SELECT Vehicle_Type_Code
READ VTC
IF VTC = " ", "Unkown", "Other"
     DELETE
ELSE
```

```
     ADD
END IF
SAVE filtered_collision_data [n=998,193]
THEN
LOAD filtered_collision_data
STRING [] Date= SELECT Date
SEPARATE Date = "Day_of_ the_ week", "Day", "Month", "Year" manually
SAVE Seperated_Filtered_vollision_Data [n=998,193]
```

Algorithm 11.2 Data Preprocessing

```
LOAD Seperated_Filtered_collision_Data
STRING [] VTC= SELECT Vehicle Type Code
READ VTC
FOR n = 1 to 998,193
IF VTC= "Bus, Fire Truck, Large Commercial Vehicle" THEN
SAVE VTC as "Large Vehicle"
END IF
IF VTC= "Pick-up Truck, Small Commercial Vehicle, Van, Livery Vehicle" THEN
SAVE VTC as "Medium Vehicle"
END IF
IF VTC= "Passenger Vehicle, Sort-Utility/Wagon, Taxi, Pedicab" THEN
SAVE VTC as "Small Vehicle"
END IF
IF VTC= "Motorcycle, Scooter, Bicycle" THEN
SAVE VTC as "Small Vehicle"
END IF
SAVE Grouped_Filtered_ Collision_Data [n=998,193]
```

Algorithm 11.3 Classification Analysis Using k-Fold Cross-Validation

```
LOAD Grouped_Filtered_ Collision_Data
INT [] k= 10 (number of tests)
Step1:
SELECT k-NEAREST NEIGHBOUR
FOR k= 1 to 10 DO
Learn k nearest Neighbour based on predictions
END FOR
RETURN k = Accuracy% (k1), Accuracy%(k2), Accuracy% (k3),....., Accuracy% (k10)
Step2:
SELECT RANDOMFOREST
FOR k= 1 to 10 DO
Learn RandomForest based on predictions
END FOR
RETURN k = Accuracy% (k1), Accuracy%(k2), Accuracy% (k3,....., Accuracy% (k10)
```

```
Step3:
SELECT NAVIEBAYES
FOR k= 1 to 10 DO
Learn NavieBayes based on predictions
END FOR
RETURN k = Accuracy% (k1), Accuracy% (k2), Accuracy% (k3) ,....., Accuracy% (k10)
```

11.3.2 Statistical Analysis

Statistically, the data is analyzed by using a Python one-way ANOVA table. In fact, Python programming allows users to understand significance value (*p*-value) in a data set variable. In Python, the scipy library inherits functionality to carrying out one-way ANOVA tests called scipy.stats.f_oneway. One-way ANOVA is a statistical implication test that allows the comparing multiple groups simultaneously [13]. This helps to find out that each vehicle group mean value in a given period differs from each other depending on the time period.

11.4 Results

11.4.1 Measured Processing Time and Accuracy of Each Classifier

This chapter has classified using algorithms using k-fold across-validation test option and k-means clustering are the main techniques used to achieve results. Accuracy and processing times were computed by using *k*NN, random forest, and Naive Bayes classification. Table 11.4 displays a comparison of each classifier results.

The high accuracy recorded 95.03% and 94.93% in random forest and *k*NN classifiers. Subsequently, Naive Bayes indicates the accuracy of 70.13% in contrast to *k*NN and random forest classifiers that comprise maximum inaccurate instances of 29.87%.

The maximum processing time is consumed by *k*NN with 682.5523 to construct the model subsequently with random forest of 507.98 seconds. Naive Bayes has determined the minimum time to build the model = 5.7938 seconds.

Figure 11.2 demonstrates a graphical representation of *k*NN, random forest, and Naive Bayes classifiers accuracy output comparing to k-fold value. The result explains the average accuracy of data set as constant. However, *k*NN and random forest outperformed Naive

Table 11.4 Comparison of classifier results.

Classifier	Processing time	Accuracy %
Random forest	507.98316	95.03
*k*NN	682.5523	94.93
Naive Bayes	5.7938	70.13

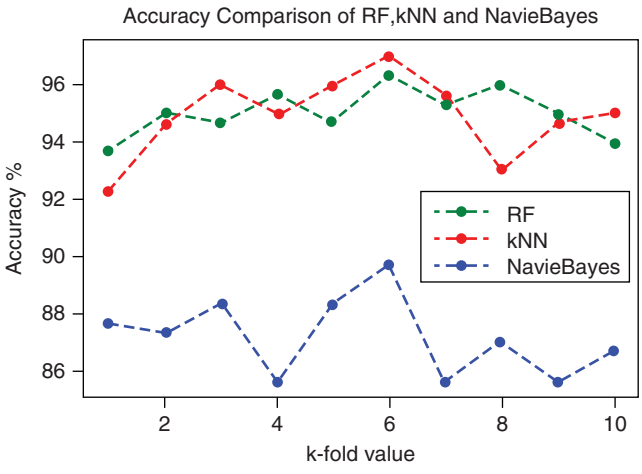

Figure 11.2 Accuracy comparison of RF and *k*NN.

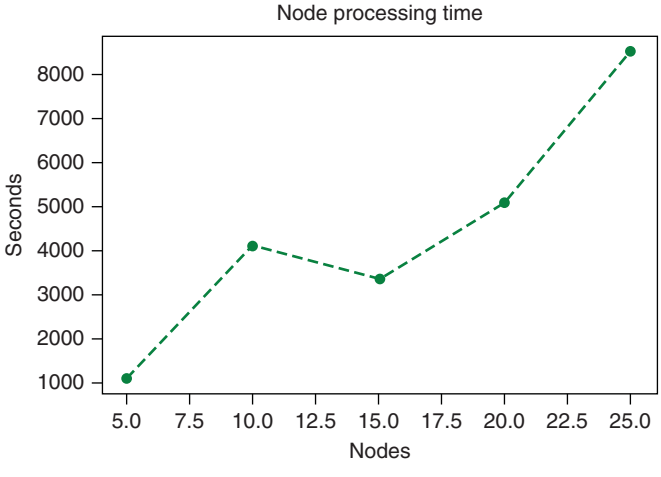

Figure 11.3 Random forest node processing time.

Bayes accuracy. The highest accuracy has been recorded k6 = 97% in *k*NN while the lowest is recorded in k7 = 85.66% in Naive Bayes. Further, Figure 11.1 indicates k6 instance as the highest accuracy for all three classifiers. Nonetheless, comparing all above results, it is evident that random forest and *k*NN prediction related to vehicle groups will be highly accurate.

Figure 11.3 explains that the total processing time of total nodes is in a linear growth where there is evidence of the accurate frequency of data set in a random forest classifier. However, 10–15 node processing time has decreased in the classifier.

Figure 11.4 illustrates the constant accuracy of 95.033% data comparing to each total number of nodes. It is further evident constant high accuracy data prediction in random forest is spread among each node.

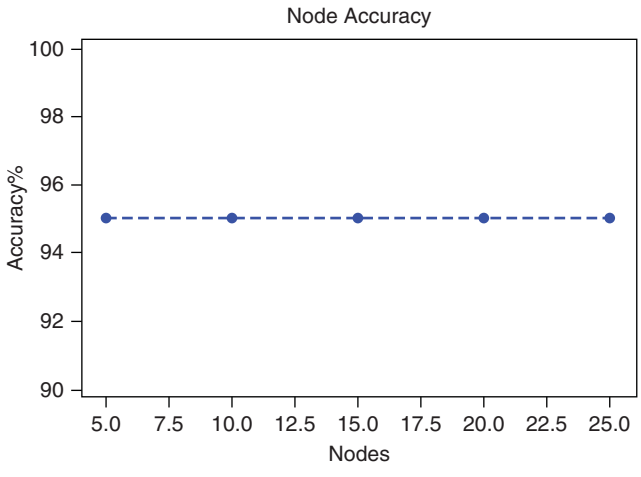

Figure 11.4 Random forest node accuracy.

11.4.2 Measured *p*-Value in each Vehicle Group Using K-Means Clustering/One-Way ANOVA

Python one-way ANOVA is carried out considering three groups of time periods in the data set. The relevant python algorithm of one-way ANOVA process is shown.

Algorithm 11.4 Data Preparation of Vehicle Group

```
LOAD Grouped_Filtered_collision_Data
STRING [] VG= SELECT Vehicle_Group
READ VG
FOR n = 1 to 998193
IF VG= "Large Vehicle" THEN
SAVE VG as "Test 1"
END IF
IF VG= "Medium Vehicle" THEN
SAVE VG as "Test 2"
END IF
IF VG = "Small Vehicle" THEN
SAVE VG as "Test 3"
END IF
IF VG = "Very Small Vehicle" THEN
SAVE VG as "Test4"
END IF
SAVE pvaluetest_ Collision_Data [n=998193]
```

Algorithm 11.5 Data Preparation of Year Group

```
LOAD pvaluetest_Colision_Data
STRING [] Period= SELECT Year
SEPARATE Period = "2016-2017", "2014-2015", "2012-2013" manually
SAVE pvaluegrouptest_Collision_Data [n=998193]
```

Algorithm 11.6 *p*-Value Test

```
LOAD pvaluegrouptest_Collision_Data
INT [] VG = Select Vehicle Group
STRING [] Period= SELECT Year group
Step1:
SELECT ONE-WAY ANOVA
FOR VG = "Test 1" DO
Learn pvalue based on mean
END FOR
Step2:
SELECT ONE-WAY ANOVA
FOR VG = "Test 2" DO
Learn pvalue based on mean
END FOR
Step3:
SELECT ONE-WAY ANOVA
FOR VG = "Test 3" DO
Learn pvalue based on mean
END FOR
Step4:
SELECT ONE-WAY ANOVA
FOR VG = "Test 4" DO
Learn pvalue based on mean
END FOR
RETURN INT [] p-value = (p1, p2, p3, p4)

SAVE p1, p2, p3, p4
```

According to the above analysis of algorithms, it is observed that different vehicle groups are compared against a period of years. In the following four tests *p*-value <0.05 considered as statistically significant (Table 11.5).

The tests output yields every *p*-value of $p < 0.001$. Its justified *p*-value is significant at a 99.99% confidence level while indicating a high significance between different means of each vehicle group in given periods.

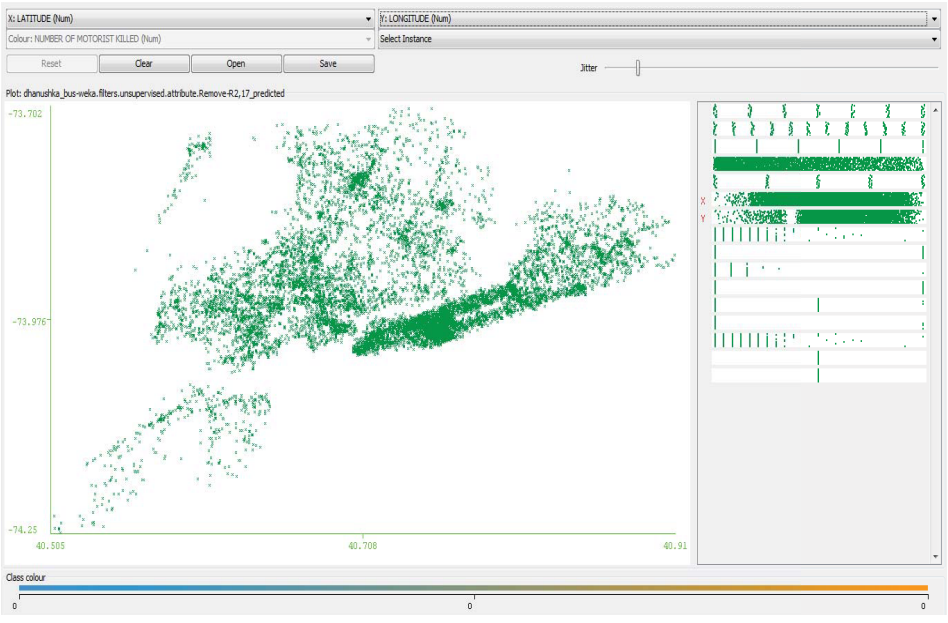

Figure 11.5 Heat map of large vehicle collisions.

11.4.3 Identified High Collision Concentration Locations of Each Vehicle Group

Visualization of collision heat maps for each vehicle group are generated using random forest classifier. It represents the locations that have more collision occurrence. The heat maps illustrate a meaningful pattern that precisely confirms the association between number of collisions and location for each vehicle group. Latitude and longitude attributes are used to represent this information on the heat maps (Figures 11.5 and 11.6).

11.4.4 Measured Different Criteria for Further Analysis of NYPD Data Set (2012–2017)

Figure 11.7 trends indicate between the years 2013 and 2016 the numbers of collisions and number of persons injured has been increasing year-wise while killed persons were stable.

Table 11.5 Analyzed *p*-value test results.

Test no	Vehicle_Group	p-Value
Test 1	Large vehicle	1.777e-159
Test 2	Medium vehicle	0.001
Test 3	Small vehicle	0.001
Test 4	Very small vehicle	4.798e-98

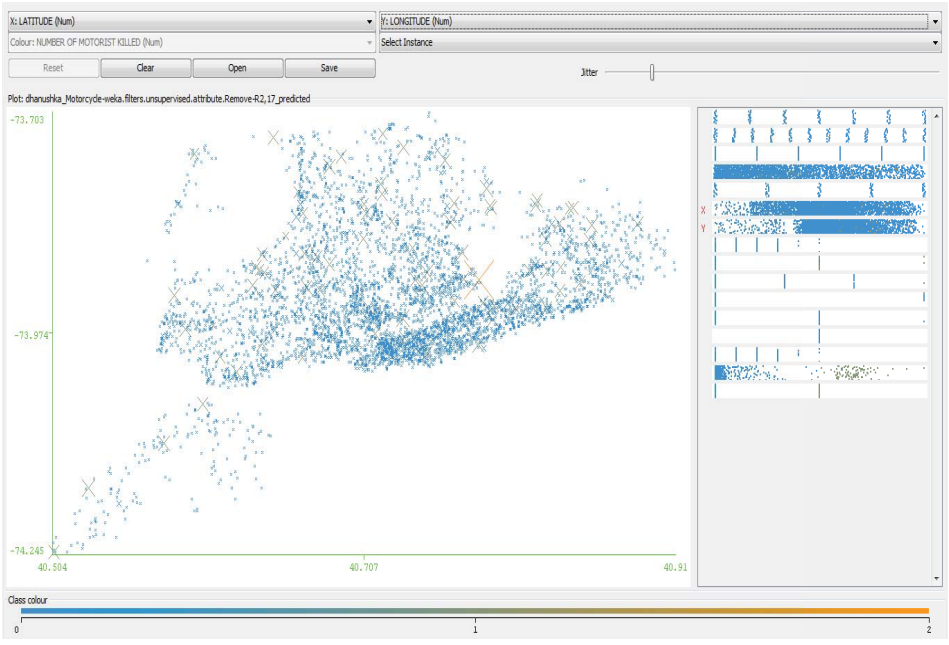

Figure 11.6 Heat map of very-small vehicle collisions.

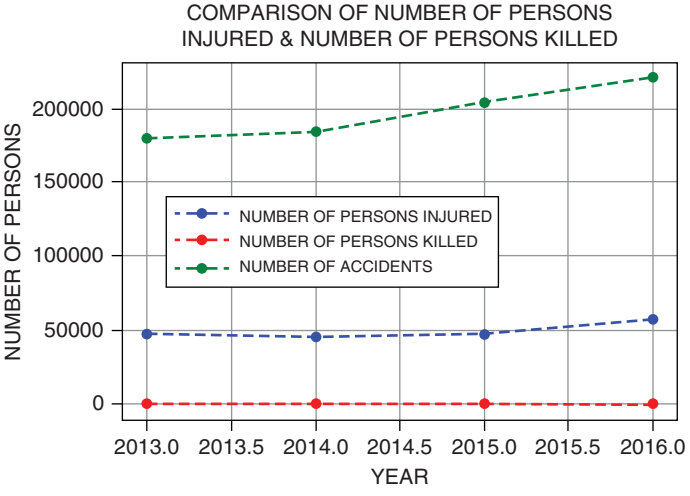

Figure 11.7 Comparison of number of collisions, persons injured, and persons killed year-wise.

Further, this is evident by numbers recorded for 2017 until the month of April. Within four months of 2017, the number of collisions, injured, and killed persons recorded were 122 245, 31 057, and 105, respectively, which is an approximately 50% growth comparing to the previous year. Therefore, the year 2017 was more fatal than 2016 and could be recorded as the deadliest year for NYC roads in decades.

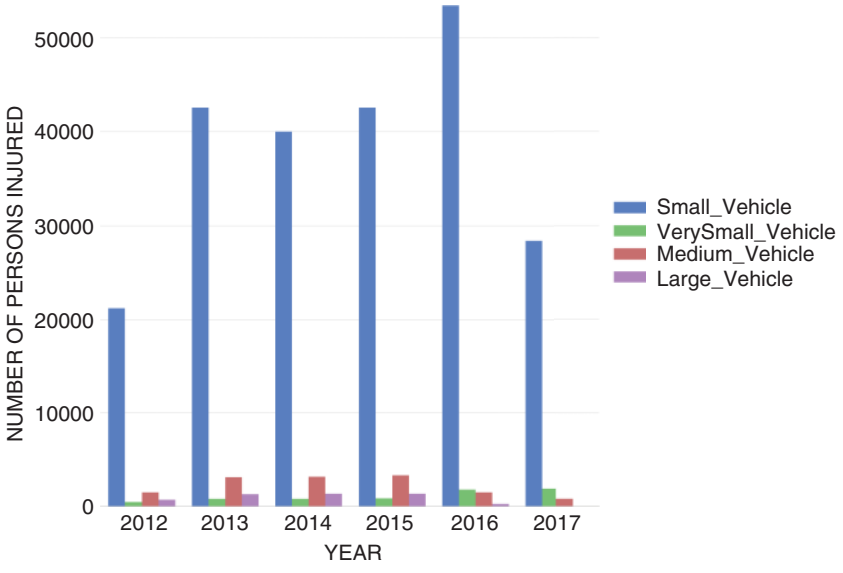

Figure 11.8 Number of persons injured based on vehicle groups.

Figures 11.8 and 11.9 show that most of the fatal collisions are recorded from medium-sized vehicles. The reason for this could be that the increasing number collisions occurred by passenger vehicles in NYC. In 2017, it recorded rapid growth compared to other vehicle types. Further Figure 11.8 shows the number of persons killed in very small vehicles is considerably higher than other groups of vehicles. This could be due to less safety in very small vehicles.

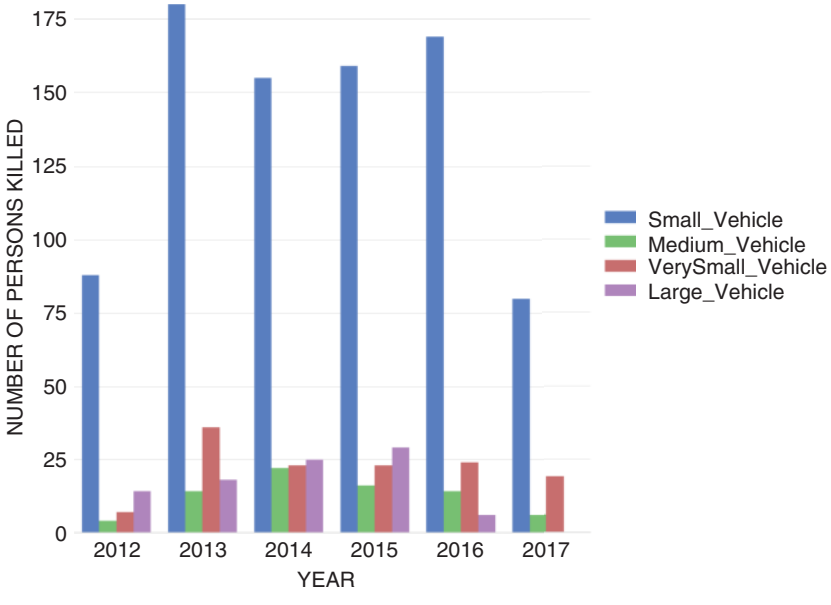

Figure 11.9 Number of persons killed based on vehicle groups.

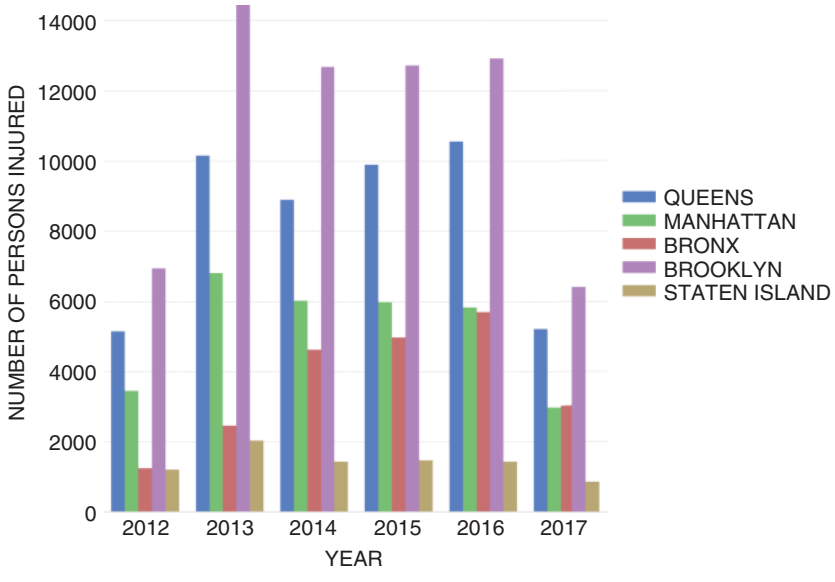

Figure 11.10 Number of persons injured based on borough.

Figures 11.10 and 11.11 shows the number of persons injured and killed in each borough. Queens and Brooklyn can be identified as the areas where many fatal crashes occur. This could be due to decreased road safety in both of these areas as well as high traffic conditions.

After analyzing 998 193 motor vehicle collisions (reported during 2012–2017) in NYC, there is a probability of extreme cases. Therefore, distribution of collision severity of

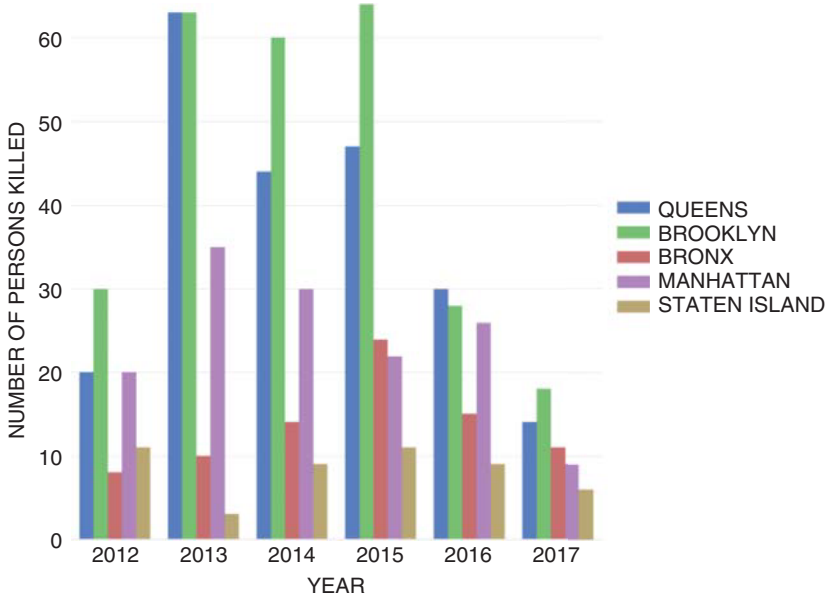

Figure 11.11 Number of persons killed based on borough.

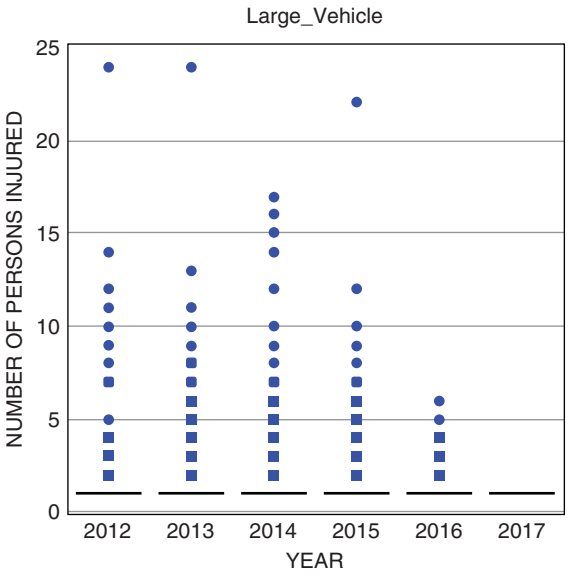

Figure 11.12 Number of persons injured in medium vehicles over N-68802 collisions. Outliers are marked in blue squares.

each vehicle group requires testing the normality of their distribution. The full data set (n = 998 193) were used. Mean value is calculated for "injured persons" and "killed persons" attributes in data set and outliers were identified.

The outlier is a critical representation of the spread of data, as value change by 25% in the upper and lower boundary, which do not affect any prediction. However, if data is from a normal distribution, the outlier is considered inefficient compared to the standard deviation [14]. With the purpose of estimating outliers, one person injured, and another killed one have been used. Therefore, any value below or above one is defined as the outlier. Figures 11.12–11.19 show the outliers in each year injured and killed people are recorded based on the vehicle group.

In Figures 11.12–11.19, significant numbers of outliers for the severity of collisions in each vehicle group can be observed. Most of the outliers are visible in the injured numbers. This could be due to the growth of injured numbers over the reported killed. However, the forecasts were carried out by excluding the extremely severe cases, which were found during the outlier analysis process. In these predictions, taking into consideration the outliers of severe collision cases is significant and very critical.

11.5 Discussion

In recent years, the growth of motor vehicle collisions has become critical. The number of road incidents has increased leading to more injuries, disabilities, and mortality on a global level. On a daily basis, more people will experience collision as a result of traffic congestion, which causes delays for vehicles passing through areas with lane closures.

The outcome of this chapter will predict emerging patterns and trends in motor vehicle collisions that may reduce the road risks. In fact, it will help to predict patterns of collision

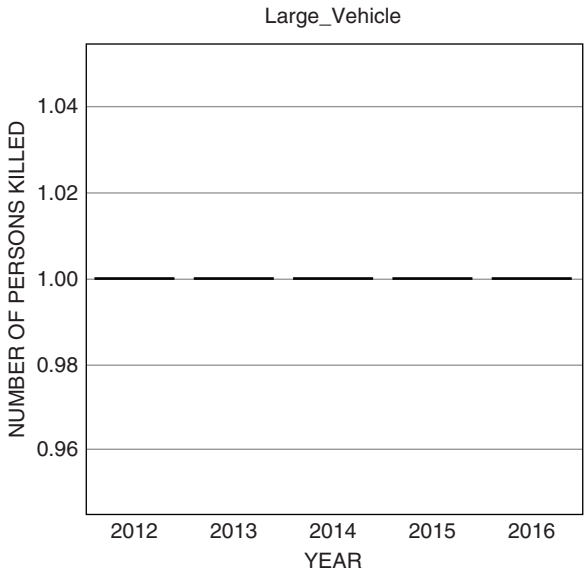

Figure 11.13 Number of persons killed in medium vehicles over N-68802 collisions. Outliers are marked in blue squares.

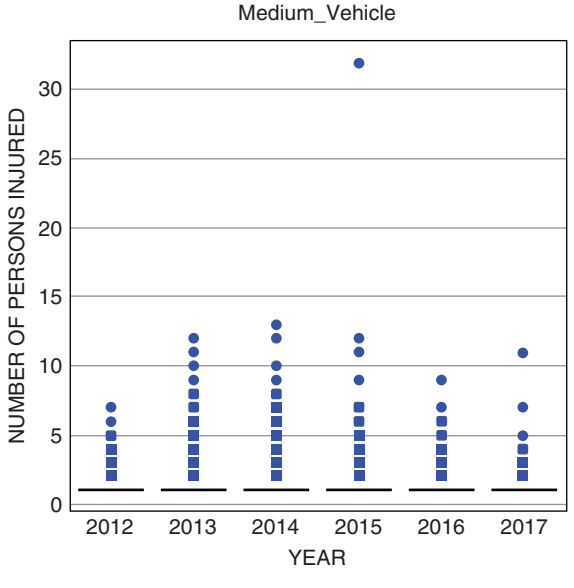

Figure 11.14 Number of persons injured in large vehicles over N-27508 collisions. Outliers are marked in blue squares.

and severity engaged with each type of vehicle. This chapter has used 998 193 large genuine data sets from NYPD as the source for data analysis. Therefore, the analyzed patterns were very reliable for overcoming road risks. The results of this chapter can even be used by NYPD to identify and prevent road risk on NYC roads. This chapter has used machine learning classification algorithms of *k*-NN, random forest, and Naive Bayes, as these received good results in our previous research [15, 16]. Using these three accuracy classifiers can predict the different vehicle groups and identify particular risk groups.

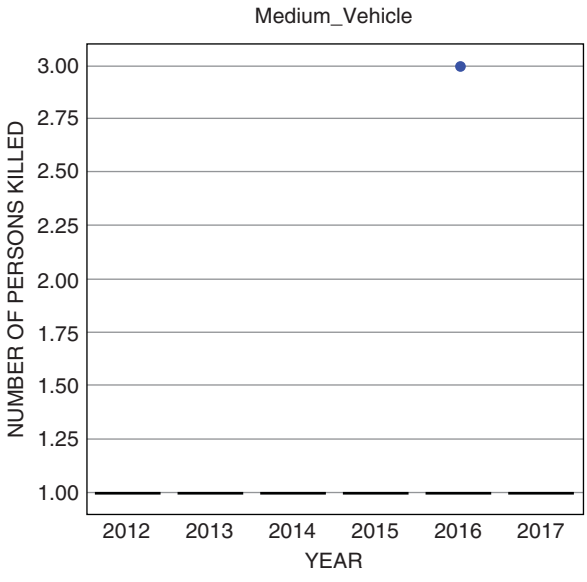

Figure 11.15 Number of persons killed in large vehicles over N-27508 collisions. Outliers are marked in blue squares.

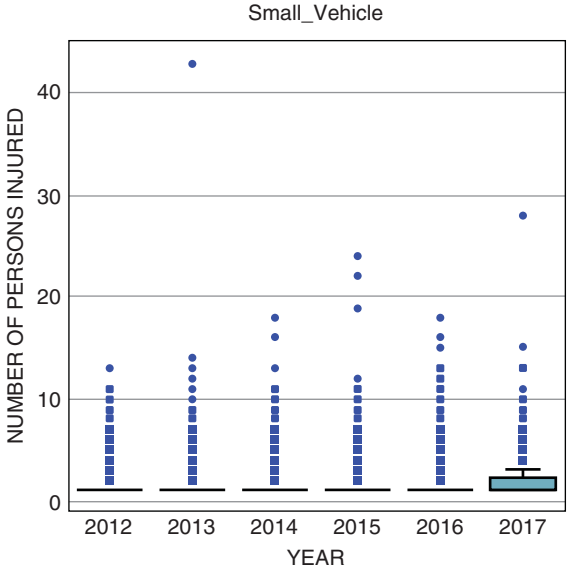

Figure 11.16 Number of persons injured in small vehicles over N-892174 collisions. Outliers are marked in blue squares.

Among three classifiers' data sets, the random forest generated the highest data prediction accuracy results. Random forest is a tree-structured algorithm that is used for pattern recognition [11]. The main reasons for using random forest as a classification technique is due to its nature of producing accurate and consistent predictions [9]. The random forest algorithm previously used in several studies demonstrates predictions. For instance, random forest has been used for a data-driven model for crossing safety by predicting collisions in railway and roadway crossings [17].

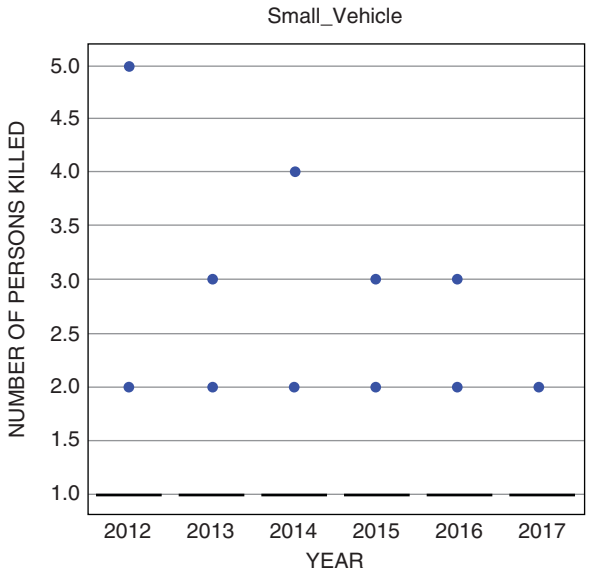

Figure 11.17 Number of persons killed in small vehicles over N-892174 collisions. Outliers are marked in blue squares.

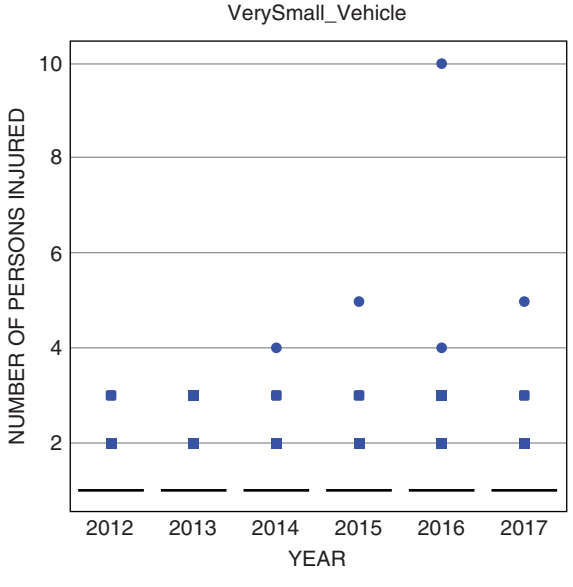

Figure 11.18 Number of persons injured in very small vehicles over N-9705 collisions. Outliers are marked in blue squares.

On the other hand, studies carried by several researchers [18, 19] suggested Naive Bayes classification as the accurate classification technique for collision analysis. These studies have used only numerical inputs for relevant prediction analysis. Therefore, Naive Bayes as a statistical pattern recognizer has produced high accuracy than other classifiers. Nevertheless, according to this chapter, random forest provides the most accurate prediction over the selected data set.

Additionally, the data set was statistically analyzed using one-way ANOVA. It shows high significance between different means of each vehicle group in the given periods. This is

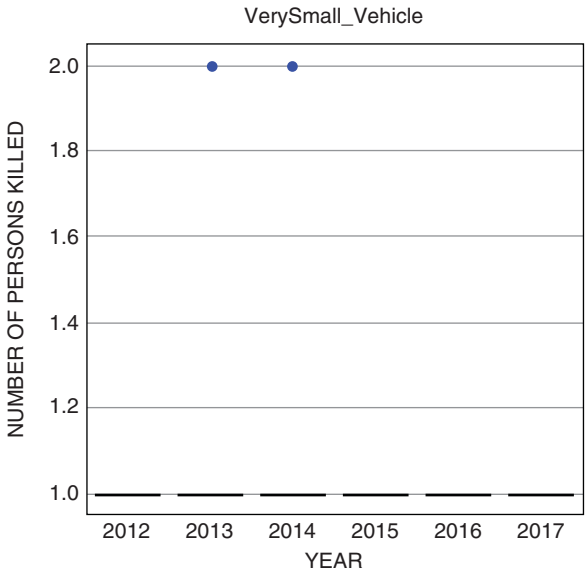

Figure 11.19 Number of persons killed in very small vehicles over N-9705 collisions. Outliers are marked in blue squares.

evident that vehicle groups are highly significant for collision patterns. Therefore, collision severity has been analyzed comparing to the vehicle group. It is evident that small vehicles were the reason for the high collision severity. This chapter reveals that between the years 2012–2017 motor vehicle collisions have increased in NYC with severity. However, it is observed that some extreme values are present in the data. Hence, this chapter excludes extreme values using outlier analysis. Further graphical representation of location using latitude and longitude confirmed the pattern of collision for each vehicle group. Brooklyn and Queens boroughs are identified as locations with the highest severe collisions.

The main limitation of this chapter occurred in the data analysis phase. During the classification carried out using Python programming, Naïve Baiyes did not capture the data in $k = 4$ and $k = 7$ k-fold instances. However, this did not impact the overall result since Naïve Bayes generated lesser value for other k-fold instances comparing to *k*NN and random forest.

The results obtained from this chapter confirmed the significance of vehicle groups in motor vehicle collisions and road risks. Identified vehicle groups could accurately predict the location and severity of the collisions. Therefore, further studies can consider these identified vehicle groups for future road accident–related research. This academic chapter acknowledged patterns through vehicle collision analysis and converted this knowledge for relevant road safety authorities and law enforcement officers to minimize motor vehicle collisions.

11.6 Conclusion

The data mining classification techniques are widely used for data analysis to obtain valuable findings. In this chapter, there are three main classifiers that have been used to generate

accurate prediction over motor vehicle collision data. In total 998 193 data collisions were tested among these three classifiers. The analysis in this chapter shows that the random forest classifier has the highest data prediction accuracy algorithm with 95.03%, following *k*NN at 94.93%, and naïve Bayes at 70.13%. Additionally, the analysis of random forest node processing time and accuracy further confirms it is the most suitable prediction classifier of this collision data. The findings of this chapter show that there has been an increase in motor vehicle collisions on NYC roads during 2012–2017. Considering only the recent years 2016–2017, there has been a growth of approximately 50% in vehicle collisions. Among these collisions, the small vehicle group recorded the highest. It is evident that the identification of motor vehicle groups has a significant impact on the severity of collisions on NYC roads. Brooklyn and Queens boroughs are identified as locations with the highest and most severe collision rates. However, this chapter generates results for collisions in NYC only; it cannot be universally applied to collisions occurring in other parts of the world. Further, these results can be used to improve road safety and minimize any potential road risks. Finally, it highlights valuable information for road authorities and law enforcing bodies to manage inner-city traffic in an efficient manner.

Author Contribution

D.A. and M.N.H. conceived the study idea and developed the analysis plan. D.A. analyzed the data and wrote the initial paper. M.N.H. helped preparing the figures and tables, and in finalizing the manuscript. All authors read the manuscript.

References

1 Abdullah, E. and Emam, A. (2015). Traffic accidents analyzer using big data. In: *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, 392. Las Vegas: IEEE https://doi.org/10.1109/CSCI.2015.187.

2 Shi, A., Tao, Z., Xinming, Z., and Jian, W. (2014). Unrecorded accidents detection on highways based on temporal data mining. *Mathematical Problems in Engineering* 2014: 1–7. https://doi.org/10.1155/2014/852495.

3 Yu, J., Jiang, F., and Zhu, T. (2013). RTIC-C: a big data system for massive traffic information mining. In: *2013 International Conference on Cloud Computing and Big Data*, 395–402. IEEE https://doi.org/10.1109/CLOUDCOM-ASIA.2013.91.

4 Sharma, S. and Sabitha, A.S. (2016). Flight crash investigation using data mining techniques. In: *2016 1st India International Conference on Information Processing (IICIP)*, 1–7. IEEE https://doi.org/10.1109/IICIP.2016.7975390.

5 Chauhan, D. and Jaiswal, V. (2016). An efficient data mining classification approach for detecting lung cancer disease. In: *2016 International Conference on Communication and Electronics Systems (ICCES)*, 1–8. IEEE https://doi.org/10.1109/CESYS.2016.7889872.

6 Ince, R.A.A., Petersen, R., Swan, D., and Panzeri, S. (2009). Python for information theoretic analysis of neural data. *Frontiers in Neuroinformatics* 3 https://doi.org/10.3389/neuro.11.004.2009.

7 Korosec, K., 2017. 2016 Was the Deadliest Year on American Roads in Nearly a Decade [WWW Document]. Fortune. http://fortune.com/2017/02/15/traffic-deadliest-year (accessed 10.8.17).

8 Sharma, S. and Bhagat, A. (2016). Data preprocessing algorithm for web structure mining. In: *2016 Fifth International Conference on Eco-Friendly Computing and Communication Systems (ICECCS)*, 94–98. IEEE https://doi.org/10.1109/Eco-friendly.2016.7893249.

9 Witten, I., Frank, E., and Hall, M. (2011). *Data Mining Practical Machine Learning Tools and Techniques*, 3e. Boston: Morgan Kaufmann.

10 Shi, A., Tao, Z., Xinming, Z., and Jian, W. (2014). Evolution of traffic flow analysis under accidents on highways using temporal data mining. In: *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*, 454–457. IEEE https://doi.org/10.1109/ISDEA.2014.109.

11 Chen, T., Cao, Y., Zhang, Y. et al. (2013). Random Forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-based Complementary and Alternative Medicine* 2013: 1–11. https://doi.org/10.1155/2013/298183.

12 Salvithal, N. and Kulkarni, R. (2013). Evaluating performance of data mining classification algorithm in Weka. *International Journal of Application or Innovation in Engineering and Management* 2: 273–281.

13 Rao, J., Xu, J., Wu, L., and Liu, Y. (2017). Empirical chapter on the difference of Teachers' ICT usage in subjects, grades and ICT training. In: *2017 International Symposium on Educational Technology (ISET)*, 58–61. IEEE https://doi.org/10.1109/ISET.2017.21.

14 Halgamuge, M.N. and Nirmalathas, A. (2017). Analysis of large flood events: based on flood data during 1985–2016 in Australia and India. *International Journal of Disaster Risk Reduction* 24: 1–11. https://doi.org/10.1016/j.ijdrr.2017.05.011.

15 Halgamuge, M.N., Guru, S.M., and Jennings, A. (2005). *Centralised Strategies for Cluster Formation in Sensor Networks, in Classification and Clustering for Knowledge Discovery*, 315–334. Cambridge, UK: Springer-Verlag.

16 Wanigasooriya, C., Halgamuge, M.N., and Mohamad, A. (2017). The analyzes of anticancer drug sensitivity of lung cancer cell lines by using machine learning clustering techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)* 8 (9): 1–12.

17 Trudel, E., Yang, C., and Liu, Y. (2016). Data-driven modeling method for analyzing grade crossing safety. In: *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 145–151. IEEE https://doi.org/10.1109/CSCWD.2016.7565979.

18 Al-Turaiki, I., Aloumi, M., Aloumi, N., and Alghamdi, K. (2016). Modeling traffic accidents in Saudi Arabia using classification techniques. In: *4th Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT)*. Riyadh: IEEE https://doi.org/10.1109/KACSTIT.2016.7756072.

19 Li, L., Shrestha, S., and Hu, G. (2017). *Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques*, 363–370. IEEE https://doi.org/10.1109/SERA.2017.7965753.