Mining Traffic Accident Data for Hazard Causality Analysis

Dimitrios Tasios
The Data Mining and Analytics Research Group
School of Science & Technology
International Hellenic University
Thessaloniki, Greece
d.tasios@ihu.edu.gr

Christos Tjortjis
The Data Mining and Analytics Research Group
School of Science & Technology
International Hellenic University
Thessaloniki, Greece
c.tjortjis@ihu.edu.gr

Andreas Gregoriades
Department of Hotel and Tourism Management
Cyprus University of Technology
Limassol, Cyprus
andreas.gregoriades@cut.ac.cy

***Abstract*— Over 1.25 million people are killed, and 20-50 million people are seriously injured by traffic accidents every year globally, according to the World Bank. This paper aims to identify patterns in traffic accident data, collected by Cyprus Police between 2007 and 2014. The dataset that was used includes information regarding 3 groups of accident properties: human, vehicle and general environmental or infrastructural information. Data mining techniques were used, and several patterns were identified. Five classifiers were evaluated using a preprocessed dataset, to extract accident patterns. Preliminary results indicate some of the main issues with regards to accident causalities in Cyprus that could be used for real time accident warnings.**

***Keywords*— Classification, Artificial Intelligence and Applications, Data mining, Traffic accidents**

I. INTRODUCTION

Traffic accidents constitute a major problem in modern societies that cause property damage, injuries and loss of human life. According to the World Health Organisation and the World Bank, over 1.25 million people die every year from car accidents and 20-50 million people are seriously impacted by road traffic injuries [1]. It is important to mention that more than 90% of the road fatalities occur in countries with low to medium per capita income. with African and Middle East countries having the highest rates in road deaths. Despite the efforts made by governments to ease this problem, accidents remain unpredictable and factors that caused them often undetermined. Among all types of accidents in human transportation (airplane, train, boat, road) car accidents remain the deadliest type [2].

Causes of traffic accidents differ in each case, with most influencing categories being a) environmental factors, such as weather, light conditions and temperature b) infrastructural factors, such as road characteristics: width, number of lanes. and c) human factors that are linked to human activity or inactivity, such as not indicating while turning, speeding, traffic lights violations and alcohol consumption [3].

With the evolution of information technology big data can be obtained and analyzed in an efficient and effective manner. Governments have been collecting accident data for some time now, with large databases like the European CARE containing large volume of information [4]. This data availability encouraged researchers to perform analyses to identify patterns that could inform authorities on how to reduce accidents.

Data mining techniques enable the identification of patterns which can explore the causes of car accidents. These include techniques such as classification, clustering and association rules [5], [9].

The aim of this paper is the identification of useful information that could be used by local authorities for the reduction of accidents occurrences. The remaining of this paper is structured as follows. Section II reviews related work and section III presents the proposed classification methodology. Section IV discusses the dataset and the required preprocessing performed prior to classification. Section V presents the selected data mining techniques and discusses their results. Finally, section VI provides conclusions and section VII directions for future work.

II. PREVIOUS WORK

Researchers have extensively investigated traffic accidents to find patterns that could explain the reasons that lead to accidents [6]. Classification has been used to classify the fatality of accidents (fatal or severe/slight injuries incurring, property loss etc.) based on some input. Work by Geetha et al. [2] evaluated the performance of different classifiers, namely: decision trees, Naïve Bayes, K-nearest neighbor and hybrid decision tree, using the same hybrid learning algorithms as for Artificial Neural Networks [5]. The classification label options where: “Fatal”, “Severe injury”, “Slight injury” and “Property loss”. The first three classifiers achieved accuracy of 80.64%, 79.87% and 81.23% respectively. One of the most valuable variables of the initial dataset for the classification was the collision type, and had 7 categories: not collision, rear end, head on, rear to rear, angle, sideswipe same direction, and sideswipe opposite direction. From these 7 categories the category with the biggest proportion of fatal injuries was the head-on collision with 1.54% and only these records were used for the hybrid decision tree in their work [5]. For the hybrid decision tree classification, the label names are: “no injury”, “possible injury”, “non-incapacitating-injury”, “incapacitating injury” and “fatal injury”. For this classification approach they chose one of the four labels and represented it as 1 and all the others as 0, which is called “once against all” approach. They trained the classifiers with different random splits of the initial dataset. Then they used hybrid decision trees. Several numbers of hidden neurons were used. The best classification accuracy result for the “no injury” class was 82.95% and generalisation accuracy 63.49%, with 95 hidden neurons. For “possible injury” class classification accuracy was 73.89% and generalisation accuracy 69.10%, with 95 neurons. The “non-incapacitating injury” class had a classification accuracy of 70.68% and generalisation accuracy 61.78%, with 109

hidden neurons. Finally, for the "fatal injury" class classification accuracy was 92.43% and generalisation accuracy 90%, with 76 hidden neurons. As a result, it was clear that the most accurate algorithm for non-incapacitating injury, incapacitating injury and fatal accidents was the hybrid approach [8].

In a similar work by Miao et al. [3] used decision trees and neural networks on an accident dataset from the National Automotive Sampling System, called General Estimates System. These data were a sample probability from the initial 6.4 million police accident reports in the USA from 1995 to 2000. A part of the initial dataset containing 417640 cases with different label variables about the driver, the road, the car and the accident type characteristics were chosen for investigation. The dataset was narrowed down to head on collision accidents only, because of its high fatality rate. The head on crash have 3 subcategories which are the front, front right corner and front left corner. The dataset was narrowed down even more to include only front impact accidents. As a result, the number of instances used was reduced to 10247. The value of "travel speed at the time of impact" was missing in 67.68% of the cases, so the column was not used for classification, even knowing this is considered a critical feature. Again, the one label against all method was used. There were five labels for the severity of passenger injury. The Neural Network was trained using Back Propagation of 100 epochs and learning rate 0.01. Also, the Conjugate Gradient descent of 500 epochs was used for the minimization of the mean square error. On the other hand, the decision tree was trained using the Gini Index. Prior class probabilities were set as equal and the minimum number per node was 5. The maximum number of nodes was 1000 and the maximum level of the tree was 32. From the results it was observed that for the classification of every single label, the accuracy of the Decision tree was always better than that of neural networks. Especially the biggest difference was observed in fatal injury error, with a 14% difference in the accuracy of the two classifiers, whilst the smallest difference in accuracy was 4% for the non-incapacitating injury label.

Krishnaveni et al. took the probability sample accident dataset from the department of Transport, in Hong Kong, which was intended to be the a nationally representative from the annual accidents' reports. [5]. The initial dataset contained 6.4 million instances, while the produced dataset had only 34575 instances. 14576 of these instances referred to accident information, 9628 to vehicle and the rest to casualty information. The dataset contains only driver information. Five different classifiers were used and Genetic algorithms for feature selection. They used Naïve Bayes, J48, AdaBoostM1, Partial decision tree and Random Forest for classification. Random Forest was the most accurate classifier. The same process was applied to the other two datasets; again, Random Forest was the most accurate classifier.

Work by Mahajan et al. used a dataset from the National highway of India [12]. The core of their approach was the application of enhanced decision tree algorithms to provide simple and efficient classification models in contrast with existing algorithms. The attributes of the dataset contained information about the road, the pedestrian facilities, light conditions, weather conditions and the location. They used WEKA's J48 [13] and their conclusion was that the algorithm is efficient in large datasets.

III. Problem Definition and Approach

The aim of this work is to identify specific patterns in accidents that occurred in Cyprus between 2007 and 2011, as well as from 2012 to 2014. In order to achieve this, the following 5 classifiers were implemented in Python and applied to the dataset.

1. Decision tree [10]
2. Random Forest
3. Gradient boosting [7]
4. Multi-layer perceptron (MLPC)
5. Voting classifier

In the first step, all classifiers were implemented with the default settings. Some customization of the decision tree classifier was attempted to increase accuracy. Especially the decision tree classifier is applied to the dataset with different maximum depths in order to identify a depth that avoids overfitting and results in acceptable accuracy.

Decision trees were visualized using the "Graphviz" library [11]. From the visualization of the decision trees we seek to identify patterns that met some criteria. For instance, a pattern is strong if it contains at least 10% of the initial samples. However, in some specific cases less than 10% may be acceptable because the dataset may involve imbalanced classes. Another criterion is that of at least 85% purity at the leaf node.

The dataset is organized in three different Comma Separated files (.csv). The first file contains general information about the circumstances under which each accident took place. The second file contains information about people involved in the accident and the third one contains information about the vehicles involved. The general accident data file contains information for every single accident that happened during 2007-11 and 2012-14. This file included 58 columns, corresponding to the features of the dataset. For the first period, there were 9862 records and 3918 records for the second period. The second .csv file contained information about every person involved in an accident for the aforementioned periods. There were 15 variables for both periods. The first period contains 9529 records and the second 9322 records. The last .csv file referred to vehicles involved in accidents. For both periods there were 19 columns that refer to 19 different features. For the first period there were 18589 records and for the second 7273.

IV. Dataset Preprocessing

As mentioned previously, the data were split across three .csv files. For pattern extraction every file was processed separately by classifying several critical attributes. In these files there were records with missing values. For the current work a specific value was assigned to all missing values. Table I contains the most representative variables from the initial input dataset.

A. Vehilce related data

Some of the dataset variables were categorized prior to data mining. The goal was to decrease the number of classes for multi-class variables. Some of these class values appeared in a very small number of instances, as a result, we

kept those class values with the most instances and the rest were merged into a single class value.

First, we addressed the driver's age category. In our datasets, there are many different values ranging from 0 to 99 years old. In order to have an efficient classification we created the following age categories: i) "Wrong or illegal" which contains drivers age less than the eligible, i.e. less than 17 years old. We concatenated on the wrong records with illegal age because we are unable to know which category they belong to. ii) "New drivers", with drivers having 3 years or less of driving experience. iii) "20-30", iv) "30-40", v) "40-50", vi) "50-65", vii) "65-75" and viii) "75-99". These are depicted in Fig. 1, for both periods, the left graph is for 2007-11 and the right for 2012-14.

TABLE I: INPUT VARIABLES

GENERAL RELATED	***VEHICLE RELATED***	***HUMAN RELATED***
AREA_CODE	ACCIDENT_DATE	POSITION_IN_VEHICLE
ACCIDENT_TYPE	DRIVER_AGE	PROTECTIVE_MEASUR
DISTRICT_ACCIDENT_	DRIVER_GENDER	EJECTION
ACCIDENT_DATE	DRIVER_LICENCE_TYP	NATIONALITY
NO_VEHICLES	VEHICLE_TYPE	AGE
NO_INJURED	MANUFACTURER	GENDER
ABANDON_IND	CAPACITY_CC	ALCOHOL
STRIKE_LEAVE_IND	MANUFACTURE_YEAR	ROLE_IN_ACCIDENT
MAIN_ROAD	APPROPRIATE_IND	INJURY_SEVERITY
RESIDENCE_AREA	DAMAGE	TRANSFER_TO_HOSPIT
KM	SECOND_EVENT	HOSPITAL
TRAFFIC_CONTROL	ACTION_BEFORE_ACC	ACCIDENT_DATE
ROAD_WIDTH		
DIRECTION		
BREAK_LANE_WIDTH		
BARRIER		
CONSTRICTION		
BREAK_LANE		
PEDESTRIAN_CROSSIN		
WEATHER		

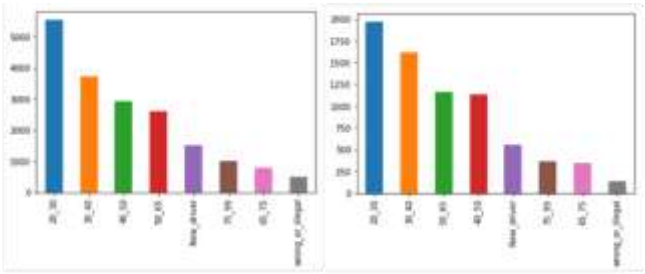

Figure 1: Accident Frequency per age group

Observing Fig. 1 we see that age category "20-30" contributed the most to accidents. The category in second place in both figures is "30-40". The first difference we notice is in the third place of accidents contribution which in the left part of Fig. 1 we have in the third place the age category "30-40" and in the fourth place the age category "40-50". On the other hand, the next period in the third place we have "50-65" and in the fourth place with small difference we have "40-50" years old.

Another variable we categorized was the age of the car. We splitted the data in six categories. The first one is "brand-new" car, i.e. less than a year old. The second is "new" car, aged between 1-5 years old, the next for cars between "5-10" years old, "10-15" years old and "15-20" years old. The last class is for cars older than 20 years. These are depicted in Fig. 2, for both periods, the left graph is for 2007-11 and the right for 2012-14.

Observing Fig. 2, we notice that the age category of cars with the biggest contribution to accidents was that of 5-10 years old and the one with the least contributions was brand new. Also, in both periods, 10-15 years old cars are in second place. The only difference between the two periods is in the third and fourth place where cars 1-5 years old had bigger contribution in addition to 15-20 years old, in contrast with the second period where the opposite happens.

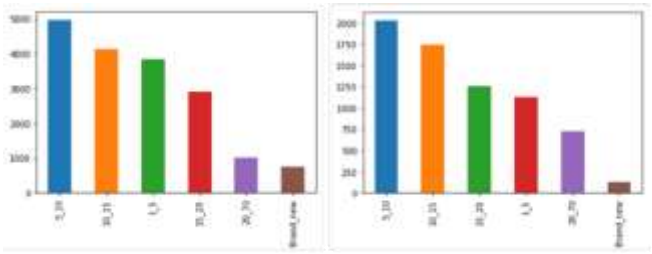

Figure 2: Accident frequency per vehicle age group

B. Human related data

This part of the dataset includes information for every person that was involved in the accident. As a result, it was difficult to try to further categorize the existed variables in order to achieve better accuracy, like before. The only change in the initial dataset that occurred was the isolation of the year and the day of the month as separate features from the datetime column.

C. General accident data

Four new features were extracted from the variable accident-date: the year of the accident, the month, the day of the week and if it was weekend or not. Also, from the variable "time", we kept only the hour that the accident happened. Furthermore, from the visualization of the time that when accidents happened, one more variable was created, called time-teams. Accidents were grouped into clusters referred as teams according to the time they occur (e.g. rush hour).

Moreover, since Cyprus is a popular tourist destination, three more binary variables were created: The "Months of high tourism", "Month of low tourism" and the "Months of regular tourism". Depending on the time of the year that accidents occurred, these variables were assigned the value of 1 or 0 if they satisfy the condition.

V. DATA MINING

We used classification by selecting one attribute as the class attribute, and the remaining attributes as predictive ones for each experiment. Our goal was to find out which variables from the dataset had impact to the classification variable in every approach.

A. Vehicle related data

The first attribute selected as a class attribute was gender. Classification on the driver's gender aimed at investigating specific habits that may lead to car accidents and differ depending on the gender. Results showed that men took part in the most accidents by far. Gender classifier accuracy is shown in Table II.

TABLE II: GENDER RELATED RESULTS

Classifier	**2007-11**	**2012-14**
Decision Tree	70.71%	70.95%
Random Forest	75.63%	75.63%

Gradient Boosting	78.67%	78.67%
MLPC	0.053%	0.053%
Voting classifier	77.89%	77.89%

From Table II one can observe that the classifiers had an acceptable accuracy, motivating further investigation. The next step was to visualize the decision tree in order to evaluate which are the most significant features for the differentiation of gender.

The next class attribute used was driver's license type. The goal was to further investigate accident patterns related to the license type. The results shown in both periods that almost 80% of the accidents were caused by drivers with regular driving license. Table III contains classifier accuracy results.

TABLE II: DRIVER'S LICENSE TYPE RESULTS

Classifier	2007-11	2012-14
Decision Tree	78.45%	68.72%
Random Forest	85.71%	75.12%
Gradient Boosting	86.12%	77.80%
MLPC	80.55%	72.50%
Voting classifier	81.44%	73.12%

The findings from different classifiers shows that there is an average difference of 10% in the classification accuracy between the two periods.

The third attribute chosen for classification was the driver's age categories. From the dataset 8 age categories were created. The first category is the "Wrong or Illegal" which refers to ages less than 17 which are illegal for driving in Cyprus. However, the category was named and as wrong because there is an instance with driver's age 4 years old. In this case it was impossible to specify if there was a driver 4 years old or it was a mistake in the recording process. Also, the "New drivers' category" was created which contains ages 17-20. The next age categories are "20-30", "30-40", "40-50", "50-65", "65-75" and "75-99". With the use of a decision tree classifier we calculated the accuracy per driver's age category, as shown in Table IV.

TABLE IV: DRIVER'S AGE CATEGORIES CLASSIFICATION RESULTS

Age category	2007-11	2012-14
Wrong or illegal	97.31%	97.86%
New driver	91,68%	91.44%
20-30	70.41%	72.30%
30-40	79.28%	75.94%
40-50	83.75%	83.36%
50-65	85.12%	81.92%
65-75	94.91%	95.27%
75-99	97.55%	97.45%

From Table IV it is obvious that for both periods the accuracy of the age categories "Wrong or illegal", "New driver", "65-75", "75-99" is very high. Also, for the remaining age categories we have a decent accuracy.

Another crucial attribute was vehicle type, with no missing values. The classification approach would help to further understand the circumstances under which the accidents happened according to vehicle type. Results are shown in Table V.

For the classification on vehicle type we used the default parameters except from the decision tree where different max. depths were used until achieving the one with the best accuracy. The optimum max. depth was found to be 8.

TABLE V: VEHICLE'S TYPE CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	87.54%	86.80%
Random Forest	85.31%	84.60%
Gradient Boosting	88.24%	89.14%
MLPC	67.75%	70.85%
Voting classifier	72.88%	74.64%

B. Human related data

The first classification approach of that part of the dataset included the position of passengers in the vehicle. The goal of that approach was to correlate several factors from the existing dataset with this position. The position in vehicle variable takes 12 different values. Value "1" illustrates the driver's position. Values 2-10 illustrate the seating passengers' position and value "11" the standing passenger's position. All other types of passengers, as well as when the position was unknown are illustrated by "12".

Table VI contains the accuracy of specific classifiers for classification on the position of the passengers. For better results all the seating passengers' values were set to "2". The classification approach's goal was the discrimination of the position of the passenger, especially if they were the driver, seating passengers, standing passengers or their position was unknown. Table VI contains the accuracy of every classifier for the two periods.

TABLE VI: PASSENGER POSITION CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	99.05%	98.98%
Random Forest	96.05%	97.90%
Gradient Boosting	96.32%	99.67%
MLPC	62.90%	73,.56%
Voting classifier	93.07%	96.30%

It is obvious from Table VI, that the accuracy of all classifiers was over 95%, except from the MLPC. The recorded accuracy of the decision tree was with max. depth = 8, which was the best accuracy after several trials with different max. depths.

Furthermore, the protective measures classification approach aims at the identification of specific patterns for human behavior and accidents impact to them, according to the protective measures that were used during the accident. The respective attribute illustrates 5 different types of safety equipment that were used by the person involved. With value 1 referring to no restraint used, 2 for seat belt, 3 for child restraint, 4 for helmet and 5 for unknown protective measures. Results are shown in Table VII.

TABLE VII: PROTECTIVE MEASURES CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	78.88%	79.27%
Random Forest	78.85%	78.60%
Gradient Boosting	18.94%	72.27%
MLPC	48.74%	59.78%
Voting classifier	78.80%	79.19%

From Table VII it is obvious that the accuracy is satisfactory except from Gradient boosting and MLPC with accuracy less than 50% in the first period.

Next we used the ejection attribute aiming at the identification of the circumstances under which specific passengers were ejected from the vehicle. The ejection attribute can take values from 1 to 4. Value 1 refers to passengers who were not ejected, 2 to these who were partially ejected, 3 to these who were ejected and finally 4 to passengers that is unknown if they were ejected or not. Results are shown in Table VIII.

The accuracy of the classifiers for the ejection of a passenger is satisfactory. It is close to 80% for the first period and almost 90% for the second period of accidents.

TABLE VIII: EJECTION CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	78.22%	89.86%
Random Forest	78.12%	89.81%
Gradient Boosting	78.64%	89.43%
MLPC	59.28%	76.56%
Voting classifier	77.96%	90.24%

C. General data

The first classification approach in this sub-section aimed at the identification of patterns correlated with the month that the accident happened. Table IX shows the accuracy of the classifiers used per period.

TABLE IX: MONTH CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	90.47%	76.40%
Random Forest	27.36%	29.46%
Gradient Boosting	95.28%	96.55%
MLPC	0.08%	0.08%
Voting classifier	21.33%	13.39%

In Table IX we see that the accuracy of the classifiers is good enough to extract strong patterns. All classifiers were implemented with the default settings except from the decision tree which was implemented with different max. depths until the most accurate was found

The next classification was for the accident type attribute, which illustrates accident fatality. There were four different classes: “fatal” indicates they were deaths, “Serious injury” indicates that people were injured seriously, “Slight injury” indicates mild injuries and “Damage” where no one was injured except from damages to the vehicles. Results are shown in Table X.

TABLE X: ACCIDENT TYPE CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	77.70%	75.25%
Random Forest	68.80%	73.33%
Gradient Boosting	79.52%	80.73%
MLPC	46.12%	61.09%
Voting classifier	63.30%	61.09%

Finally, the classification of the number of vehicles participating to car accidents could play a leading role to strong pattern extraction. The identification of the reason why many vehicles take part to one accident. Results are shown in Table XI.

From Table XI, we observe that the accuracies of the classifiers are good enough for pattern extraction, particularly in the second period.

TABLE XI: NUMBER OF VEHICLES CLASSIFICATION RESULTS

Classifier	2007-11	2012-14
Decision Tree	86.67%	89.92%
Random Forest	84.63%	88.13%
Gradient Boosting	86.46%	89.54%
MLPC	65.73%	45.05%
Voting classifier	82.05%	87.50%

VI. CONCLUSIONS

This study aimed to analyse historical accidents occurred in Cyprus during 2007-11 and 2012-14 as recorded by Cyprus Police. The initial data was preprocessed, and subsequently 5 classification techniques were used for the 3 main groups of factors that could lead to accidents. The knowledge that was extracted and the classifiers that were developed could be used as part of the specification of a prospective mobile application that could be used for real time accident warnings using as input static and dynamic information regarding the driver, the environment, the infrastructure and traffic conditions, that can be obtained from onboard sensors, V2V & V2C protocols and historical records. A summary of the knowledge that was extracted from the data mining is presented below.

The first conclusions came from the visualizations of the dataset. In both periods the biggest percentage of drivers who contribute to accidents are between 19 and 40 years old. In both periods, the age category with the highest contribution is 20-30 years old. In addition, 75% percent of the drivers were male in both periods. From all the drivers who contributed to the accidents, 79% and 82%, respectively for the two periods, had a regular driving license. The vehicle type in both periods with the biggest contribution to accidents was “saloon car” and the vehicle manufacturer with the biggest contribution was “166”.

Following the visualization of the dataset, several variables were selected for further analysis, using 5 classifiers. The application of these classifiers to the dataset resulted in the creation of decision trees. With the help of Graph Viz library [11] these trees were visualized; an example is given in Fig. 3. These visualizations contained information on how many records belong to each branch of the tree, from which strong patterns were extracted.

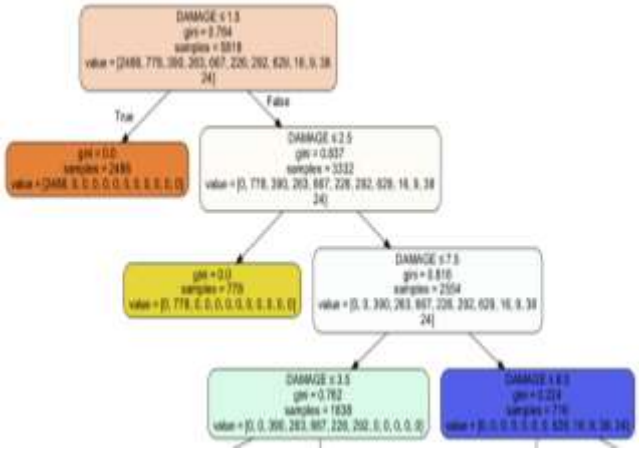

Figure 3: Part of decision tree created with Graph viz

According to the classification on gender, the patterns that were extracted refer only to men. In the first period, men below 30 years old or above 40 years old. were involved in accidents with bicycles and motorcycles up to 50cc. Men less than 75 years old contributed to accidents with motorcycles between 125cc and 2008cc and all these accidents occurred in specific territories. Also, the vehicles with which men contributed to accidents were split in two categories: those over 12 years old with cc between 1.513 and 2008 and those with cc between 2008 and 2.773. In the next period the commercial vehicles were more than 20 years old and the cc was less than 1809. Also, taxis and motorbikes were between 1.809 cc and 2.148 cc.

Another variable where the classification yielded strong patterns was driving license. In the first period, drivers without license were involved in accidents with motorcycles up to 50cc and age over 75 years old or less than 65 years. Also, drivers with regular driving license who were involved in accidents were over 18 years old.

Following that, age classification revealed one pattern. For drivers younger than 17, which had the legal right to drive that was called as wrong ages recordings or illegal, a pattern was extracted only for the period 2007-2012. Those drivers were driving without driving license or the information was not recorded, and they were involved in accidents with mopeds up to 50cc.

From the classification of the passengers' position in the vehicle, two strong patterns where extracted.

1. If age<16.5 and position in vehicle is passenger and not the driver then they were slightly injured or not injured.
2. If age >17,5 position is equal to driver or passenger then they were using seat belts and did not use drugs.

Moreover, from the classification of the variable of protective measures, strong patterns were extracted. The classification approach showed for the first period, that passengers aged 24-84 involved in accidents while they were riding a motorcycle and were wearing a helmet. In these accidents the number of involved vehicles were more than 2. Also, infants less than 3,5 years old, wearing seat belts, were not injured fatally.

Additionally, strong patterns were extracted from the accident's type classification; however only for the period 2007-2011. For fatal accidents it was extracted that when the police officers' grade was not "Constable" and the accident did not happen to a specific point (ZZ527) then the ambulance was reaching the accident's location in less than 25 minutes.

VII. Future Work

This work aimed at extracting strong patterns on the cause of accidents. These patterns were extracted from the visualization of the decision tree classifier with the help of "Graph Viz" library [11].

Fig. 3, shows part of the decision tree constructed using Gini index as the cost function to evaluate the splits in the dataset. Entropy and Information gain could have been used alternatively [8]. Colors in the leaf nodes represent values of the class variable. Starting from the root and for every internal node one attribute is selected for a split.

Principal component analysis could be used in the future for improving the classifier. Principal Components Analysis (PCA) is already being used in real life problems. There are more than 50 features in the dataset; PCA could decrease its dimensionality and improve the speed and accuracy of calculations.

Another approach for future work could be different data preprocessing. Our approach involved filling in missing values with a specific value for unknown data. However, there are other approaches, such as filling the average value and the imputation in which there is a prediction of the missing values before using the classifier. Moreover, the dataset could be merged for the two periods and implement the same or new classifiers in order to compare existing findings with more generic ones.

Generally, all the findings of the dataset could be used for prediction. Insurance companies could use these data for customizing the cost of insurance according to the characteristics of the car, the driver and the places they drive. Also, applications such as google maps could give real time warnings to users, especially for tourists who are unfamiliar with the roads based on real time data feeds into the classification models.

VIII. References

[1] World Bank (2017). The High Toll of Traffic Injuries: Unacceptable and Preventable. https://openknowledge.worldbank.org/handle/10986/29129.

[2] K. Geetha, C. Vaishnavi (2015), Analysis on Traffic Accident Injury Level Using Classification. *Int'l Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 5, No. 2, pp. 953-956,

[3] M.M. Chong, A. Abraham, M. Paprzycki (2004), Traffic Accident Analysis Using Decision trees and Neural Networks, *IADIS Int'l Conf. Applied Computing*, Vol. 2, pp. 39-42.

[4] V. Saurbah (2018), "CARE database common accident data set".

[5] S. Krishnaveni, M. Hemalatha (2011), A Perspective Analysis of Traffic Accident using Data Mining Techniques, *Int'l Journal of Computer Application,s* 23(7).

[6] A. Gregoriades, A. Christodoulides (2017), Traffic Accidents Analysis using Self-Organizing Maps and Association Rules for Improved Tourist Safety. ICEIS (1), 452-459.

[7] Y. Chen, Z. Jia (2013), A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index, *Computational and Mathematical Methods in Medicine*, vol. 2013.

[8] Tzirakis P. and Tjortjis C. (2017), "T3C: Improving a Decision Tree Classification Algorithm's Interval Splits on Continuous Attributes", *Advances in Data Analysis and Classification*, Vol. 11, No. 2, pp. 353-370.

[9] K. Polat, S. Gunes (2007), Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform, *Applied Mathematics and Computation*, Vol. 187, no. 2, pp. 1017-1026.

[10] W. Du, Z. Zhan (2002), Building decision tree classifier on private data, *Proc. IEEE Int'l Conf. Privacy, security and data mining*, Vol. 14, pp. 1-8.

[11] Gansner E., North S.(1997). An opern graph visualization system and its applications to software engineering, *Software Practice and Experience* 30(11), pp. 1-5.

[12] N. Mahajan, B.P. Kaur (2016), Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm, *Int'l Journal of Computer Applications*, Vol. 135, No. 6, pp.1-3.

[13] I. Witten, E. Frank, M. Hall, and C.J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Ed., Morgan Kaufmann, 2016.