Mining Traffic Data

Dimitrios Tasios

SID: 3308170024

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data science

DECEMBER 2018

THESSALONIKI – GREECE

Mining Traffic Data

Dimitrios Tasios

SID: 3308170024

Supervisor:	Prof. Christos Tjortjis
Supervising Committee Members:	Dr. Christos Berberidis
	Dr. Agamemnon Baltagiannis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data science

December 2018

THESSALONIKI – GREECE

Abstract

Over 1.25 million people are killed, and 20-50 million people are seriously impacted by road traffic injuries on earth every year according to the world bank. This dissertation aims to the identification of traffic accident patterns in Cyprus, according to data collected by the local Police from 2007 to 2014. The dataset contains general, human based and vehicle-based information about the accidents. With the help of data mining, several patterns are extracted. Several classifiers were applied to the dataset in order to extract patterns related to the human factor, the car factor and the general data for every single accident. Findings from classification could be used by local authorities for accident prevention and by insurance companies for risk analysis.

Dimitris Tasios

12/12/2018

Acknowledgements

I would like to express the deepest appreciation to Professor Christos Tjortjis for the guidance of the whole dissertation through these months. I would also like to thank Assistant Prof. Andreas Gregoriades for providing the data.

Contents

List of Figures:

List of Tables:

1 Introduction

Traffic accidents have negative effects to our society. They cause property damage, injuries and even human life losses. According to the report that produced by the World bank and funded by the Bloomberg philanthropies over 1.25 million people die every year from car accidents and 20-50 million people are seriously impacted by road traffic injuries. It is remarkable that more than 90% of the road deaths happen in low-income and middle-income countries. Also, the road death rate remains highest in Africa and Middle East. [30]

Hence it is an obligation for governments to try to reduce this phenomenon. However, until now a quite large number of accidents are unpredictable and the factors that caused them still undetermined. There are different traffic accidents in the field of transportation. There are airplane accidents and car accidents. Every type of consists of different factors that need further scientific investigation. The deadliest type of accidents are car accidents [4].

Με σχόλια [CT1]: Check references

In every single accident caused, there were different circumstances. These circumstances can be categorized in three categories. The first one is the natural elements during the accident. These elements could be the temperature and the climate in the specific area. The second category is the road's specific characteristics. For example, the width of the road and the number of traffic lanes. The third category contains the human factor. It includes all the factors that are linked to human activity, such as the violation of speed limits or traffic lights and alcohol consumption. [5]

In the last decades the evolution of technology gave us the ability to process big amounts of data faster. As a result, in order to reduce accidents, the governments started collecting as much accident data as possible. Nowadays we have large databases like the European CARE that contains a lot of information for further process [29]. So, with the use of the modern IT capabilities we can produce strong accident analysis.

Data mining is the method that can handle these amounts of data and extract strong patterns which can justify the reason that many of these car accidents have happened. Data mining contains several techniques, such as data preprocessing for better data manipulation, the ability to classify by establishing a factor as the class attribute. Also, one can use clustering and association rules. [6]

The main aim of this dissertation is the extraction of useful information that could be used by local authorities for the reduction of accidents. The structure of this dissertation follows. The second chapter reviews related work and presents the proposed classification methodology. The third chapter discusses the dataset and the preprocessing necessary prior to classification. Following that, in the fourth chapter the selected data mining techniques are presented and results are provided. Finally, the fifth chapter provides conclusions and directions for future work. [7]

Με σχόλια [CT2]: What is this about?

2 Background

Data mining is an interdisciplinary subject that can be explained in different ways. The core of data mining is the discovery of knowledge. Strictly speaking it is the discovery of strong patterns from large amounts of data [18].

This process of knowledge discovery from data contains specific essential steps. These are the following:

- Data cleaning: The removal of inconsistency and noise from data.
- Data integration: The right combination of data from different sources for their better manipulation.
- Data selection: Selection of the most relevant data for analysis.
- Data transformation: The transformation of data into appropriate forms for mining.
- Data mining: The process where efficient algorithms are applied to the data with focus on pattern extraction.
- Pattern evaluation: Identification of interesting patterns with the use of interestingness measures.
- Knowledge presentation: The final step where the whole information gained is being visualized.

2.1 Data storages that can be mined

Data mining can be applied almost to all kinds of data, although the main forms of data for mining applications can be found in databases, data warehouses and transactions. Data mining can also be applied to data streams, graph data, text data and other forms of data. [24]

A database management system (DBMS) consists of collections of interrelated data. The software programs provided for database structures gives the user the ability to manage and access data. Particularly the user can manage concurrent, shared and dis-

tributed data. Also, it ensures security and consistency of the stored data from situations like system failures and unauthorized access.

Data warehouses are informational repositories which have been collected from several sources and are being stored under a unified schema. Their creation contains data cleaning, data integration, data transformation, data loading and a periodic refreshment of data. The core of use of data warehouses is the decision-making process in large companies. As a result, the stored data provide information from a historical perspective and are summarized. [23]

Finally, transactional databases collect all the data that describe transactions being made. In most cases there is a unique key (ID) and the items that have been purchased. More than one table commonly exist to describe analytically the features of purchased items for example. [11]

2.2 Data mining fundamentals

The fundamental of data mining is exploratory data analysis, frequent pattern discovery, classification and clustering [12]. The core of exploratory data analysis is the exploration of numeric and categorical attributes individually or jointly with the aim of extraction of statistic characteristics from variables. Some statistical information is the spread of a variable, centrality and dispersion. The visualization of these characteristics can give to the user the ability to gain more insights about the variables. [13]

Frequent pattern analysis mining aims to the extraction of strong patterns from huge datasets. Generally, a pattern is the co-occurrence of attribute values which are called itemsets or more complex patterns, such as sequences of relationships. The goal in the whole procedure is the recognition of hidden trends and behaviors in data. [14]

Clustering is the process which partitions the points into groups called clusters. The partition criterion is the similarity of points within a cluster and the dissimilarity between two points of two different clusters. There are different types of clustering such as density based, graph based, spectral based, hierarchical and representative based. The type of clustering that the user chooses depends on the data and the characteristics of the desired cluster. [15]

Classification is the task that predicts the label of an unlabeled given point. In order to build a classifier model several points correctly classified, called the training set, are required. By the end of training the classifier is ready to predict the label of any new

point. However, the accuracy of classification depends strongly on the training data. There are different types of classification such as decision trees, probabilistic classifiers, support vector machines and so on. [16]

2.3 Previous work

Researchers have extensively investigated traffic accidents. They aimed at mining available information in order to analyze it and find patterns expected for the explanation of the reasons that lead to accidents. [17].

The purpose of classification in this section is to classify the fatality of the accident. In order to achieve that kind of classification Geetha et al. built with the help of WEKA a J48 decision tree, a Naïve Bayes classifier, K-nearest neighbor classifier and a hybrid decision tree where they used the same hybrid learning algorithms as for Artificial neural networks [4]. The classification label options where: “Fatal”, “Severe injury”, “Slight injury” and “Property loss”. The first three classifiers had accuracy 80.641, 79.867 and 81.231 respectively. Then the dataset was cleaned from outliers. Then the was a new separation in the labels. Every time they chose one of the four labels and the representing it as 1 and all the others as 0. They trained the classifiers in different random splits of the initial dataset. Then they used the hybrid decision tree. Several numbers of hidden neurons used for every approach. The best results for no injury class were training performance 82.95% and 63.49% testing performance with 95 hidden neurons. For possible injury class the training accuracy was 73.89% and 69.10% was the testing with 95 neurons. The non-incapacitating injury class had a training accuracy of 70.68% and testing accuracy 61.78% with 109 hidden neurons. Finally, for the fatal injury class the training accuracy was 92.43% and 90% for testing with 76 hidden neurons. As a result, from the observations it was clear that the most accurate algorithm for non-incapacitating injury, incapacitating injury and fatal was the hybrid approach. [1]

Miao et al. applied 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. The used part of the initial dataset contained 417.640 cases with different label variables about the driver, the road, the car and the accident type characteristics. Because the head on collision had the biggest fatality of injuries records, the dataset narrowed down to the head on collision only. Moreover, the dataset narrowed even

more only to the front impact accidents. As a result, the number of instances used was 10.247. Also, the variable for travel speed at the time of impact was missing in 67.68% of the cases, so the column was not used for the classification even knowing that is 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 trained using Back Propagation of 100 epochs and learning rate 0.01. Also, the Conjugate Gradient descent of 500 epochs used for the minimization of the mean square error. On the other hand, the decision tree was trained with the help of Gini. The prior class probabilities were set as equal and the minimum number per node were 5. The maximum number of nodes was 1000 and the maximum level of the tree was 32. Finally, from the results it was observed that for the classification of every single label the accuracy of the Decision tree was always better than the neural networks. Especially the biggest difference was observed in the fatal injury error with a 14% difference in the two classifiers accuracy. While the smallest difference in accuracy was 4% in the non-incapacitating injury label. [8]

Krishnaveni et al. took the probability sample accident dataset from the Transport department of the government of Hong Kong. The initial dataset was a parted from 6.4 million instances while the produced dataset has only 34.575 instances. 14576 of these instances belong to the accident information, 9628 belongs to vehicle information and the rest belong to casualty. The dataset has only information for the drivers, not for passengers. He used five different classifiers for the classification process and the Genetic algorithm for Feature selection. Especially for the classification problem he used Naïve Bayes classifier, J48, AdaBoostM1 classifier, Partial decision tree classifier and the Random forest tree classifier. For every attribute of the accident instances used the given classifiers. Then the genetic algorithm used in order to have feature reduction. Random forest was the most accurate classifier. The same process was applied in the two other datasets and again Random forest was the most accurate classifier. [9]

Mahajan et al. used the dataset from the National highway of India. It contains records from Mukerian to Jalandhar and Punjab. The core of his scientific approach was the application of enhanced decision tree algorithms to a dataset in order to provide simple and efficient classification models in contrast with the existing algorithms. The attributes of the dataset contained information about the road, the pedestrian facilities, light conditions, weather conditions and the location. The algorithm that applied was

C4.5 which is the enhancement of ID3, using the concept of entropy. The algorithm applied with the help of WEKA. The conclusion of this approach was that the algorithm is efficient in large datasets. [10]

2.4 Selected classifiers

After the study of the previous related work the following classifiers were selected because there were already tested in the same manner in the past:

1. Decision Tree
2. Random Forest
3. Gradient boosting
4. Multi-layer perceptron
5. Voting classifier

For every single classification all the classifiers are implemented. However, the decision tree classifier is the one from which strong patterns are extracted.

Decision tree is a supervised learning technique that is being used in data mining. The aim is to construct a model that is able to predict the value(class) of a target variable according to several other variables. Every single interior node corresponds to one of the input variables. There are edges to children for every possible value of that input variable. Every leaf represents a value of the target variable given the values of the input variable represented by the path from the root to the leaf node. [15]

For every decision tree there is a single target feature that is being called the “Classification”. Every element of the domain of the classification is being called a “Class”. In a decision tree each internal node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes. [19]

The construction of a decision tree is available only from class-labeled training tuples. In a decision tree every internal node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. The topmost node in a tree is the root node. For the construction of a decision tree algorithms usually works from the top to down, by choosing every time a variable that splits best the set of

remaining items. The criteria for this splitting are the metrics. The metrics measure the homogeneity of the target variable within the subsets. Two main metrics are the Gini impurity and the information gain. [20]

Gini impurity is a metric that very often is being used from classification and regression trees. It corresponds how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The summing of probability P_i if an item with label i being chosen times the probability of a mistake in categorizing that item. It reaches its minimum when all the cases in a node turn into a target category. The computation of a Gini impurity metric for a set of items with J classes and if P_i is the fraction of items labeled with the class s ι in the set.:

$$\mathrm{I}_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

On the other hand, information gain is the metric that in each step choose the split that results in the purest daughter on nodes. This purity measurement it is called information and is measured in bits. For every node of a tree the information gain represents the expected amount of information that is being demanded to declare for a new instance if should be classified or not. The calculations of entropy and information gain are being computed from the following equations:

$$\mathrm{H}(T) = \mathrm{I}_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$$

$$\overbrace{IG(T,a)}^{\text{Irformatior Gair}} = \overbrace{\mathrm{H}(T)}^{\text{Entropy (parent)}} - \overbrace{\mathrm{H}(T|a)}^{\text{Weighted Sum of Entropy (Children)}}$$

$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a} p(a) \sum_{i=1}^{J} -\Pr(i|a) \log_2 \Pr(i|a)$$

3 Data and problem definition

The selection of the dataset and understanding the problem domain is crucial for every data mining approach. The dataset must be in a form that it would be effective for further investigation and implementation in data mining techniques such as classification and clustering. Also, the understanding of the problem domain is crucial because in every implementation it must be clear the aim of mining and the already existing supplying data.

3.1 Dataset

The dataset to be mined illustrates information from traffic accidents that took part in Cyprus during 2007-2011 and 2012-2014. The data were collected by Cyprus police. The dataset is organized in three different comma separated files. The first file contains general information about the circumstances under which every single accident took part. The second file contains information about every person involved in the accident and the third one contains information about the vehicles that took part in each accident.

The general accident data file contains information for every single accident that happened during the periods 2007 -2011 and 2012-2014. For both periods there are 58 columns which illustrate the features of the dataset. For the first period there are 9862 records which illustrate the circumstances under every single recorded accident that happened. For the second period there are again 58 columns and 3918 instances.

On the other hand, the file individual contains information about every single person that was involved in an accident in 2007-2011 or 2012-2014. There are 15 variables for both periods. The first period contains 9529 records and the second 9322 records.

The last file refers to all the information for every vehicle that was involved in the accident in the two periods. For both periods there 19 columns that refer to 19 different features. For the first period there are 18589 instances and for the second 7273 records. Further details on the dataset can be found in the appendix.

3.2 Problem definition

We aim to identify specific patterns that exist in the accidents tha happend in Cyprus from 2007 to 2011 and from 2012 to 2014. In order to achieve that goal 5 selected classifiers are implemented and applied to the dataset with the use of Python programming language. These classifiers are the following:

1. Decision tree
2. Random forest clasifier
3. Gradient boosting classifier
4. Multi layer perceptron
5. Voting clasifier

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

Also, with the use of the “Graph viz” library we try to visualise decision trees. From the visualazation of the decision trees we seek to identify patterns that meet the conditions set. These conditions are that a pattern can be assumed as 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 unbalanced classes. Another criterion is that of with at least 85% purity of the leaf node.

4 Mining accident data

The dataset is arranged in three different comma separated (.csv) files. Every file is processed separately by classifying several critical attributes.

4.1 Vehicle related data mining

In the beginning we identify the variables that the csv files contain. So, we trace all the variables and their values to check if there are missing values. The following variables contain no missing values:

1. Accident account identity
2. District accidents number
3. Drivers age
4. Drivers gender
5. Driver's license type
6. The date of the accident

On the other hand, the variables with the most missing values were:

1. Driver's license expiry day
2. Manufacturer
3. Insurance issue date
4. Insurance expiry date

Also, there were some variables with less than 20missing values:

1. Cars capacity in CC
2. Driver's license indicator
3. Insurance company
4. Appropriate indicator
5. Damage
6. Second event
7. Action before accident
8. Manufacturer year

4.1.1 Variable categorization

Our next step in order to find specific patterns is the categorization of our variables. We can achieve that goal by decreasing the number of classes in the classification approach in variables that have a lot of classes. Some of these classes are outliers, as a result we keep the most common ones and the rest formulate one class.

First, we address the driver's age category. In our datasets there are a lot of different values ranging from 0 to 99 years old. In order to have an efficient classification we create the following age categories. Our first category is the "Wrong or illegal" which contains drivers age which less than the eligible (less than 17 years old). We concatenate the wrong records with the illegal because we are unable to know in which category they belong to. Our next category is the "New drivers" where we have drivers with 3 years of experience and less. The following categories are the "20-30", "30-40", "40-50", "50-65", "65-75" and the last one is "75-99".

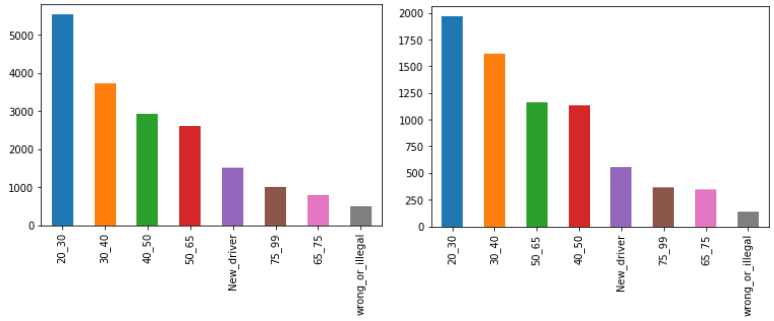

Figure 4-1: The number of drivers involved in accidents and the age category they belong to (2007-2011 to the left and 2012-2014 to the right)

From the above figures we realize that the age category 20-30 years old contributed the most to accidents. The category in second place in both figures is 30-40 years old. The first difference we notice is in the third place of accidents contribution which in the first figure we have in the third place the age category 30-40 and in the fourth place the age category 40-50 years old. On the other hand, the next period in the third place we have 50-65 years old and in the fourth place with small difference we have 40-50 years old.

Another variable we must categorize is the age of the car. We split the data in six categories. The first one is the brand-new car, i.e. less than a year old. The second one is

new cars, aged between 1-5 years old. The following classes are 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.

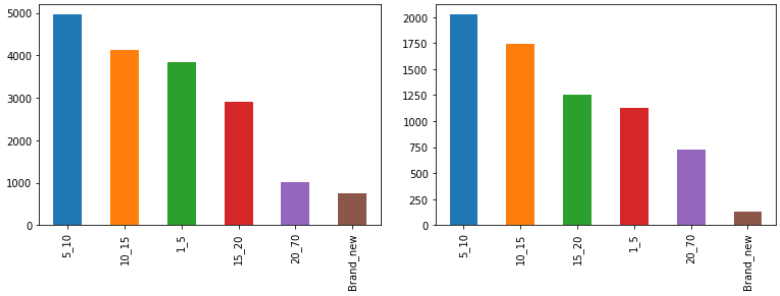

Figure 4-2: The age of cars involved in accidents (2007-2011 to the left and 2012-2014 to the right).

From the figures above, we notice that the cars age category with the biggest contribution to accidents is that of 5-10 years old and the cars age category with the least contributions are brand new ones. Also, in both periods 10-15 years old cars are in the 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.

4.1.2 Classifications

The following classifiers were used for classification :

1. Decision Tree
2. Random Forest
3. Gradient boosting
4. Multilayer perceptron
5. Voting

The first variable selected for classification was gender, after the visualization of the driver's gender in the two periods (Figure 4). The results showed that men took part in the most accidents by far. The classification of the driver's gender had as goal the investigation of specific habits that may differ depending on gender and lead to more car accidents. Table 1 contains classifier accuracy results.

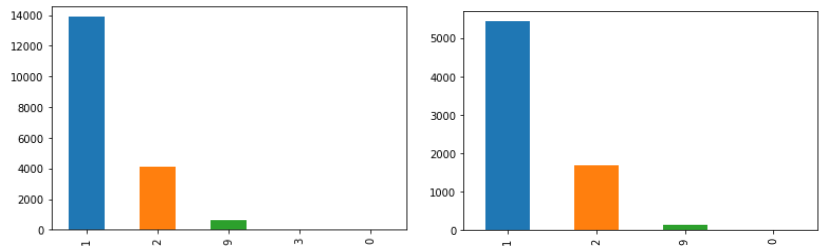

Figure 4-3: Drivers gender contribution in accidents. (2007-2011 to the left and 2012-2014 to the right)

Table 1:Classifiers accuracy of driver's gender classification.

Classifier	**2007-2011**	**2012-2014**
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 the above we can understand that the classifiers had a neutral accuracy capable for further investigation. The next step was the visualization of decision tree in order to evaluate which are the most significant features for the differential of the gender.

The first approach of classification of driver's gender was applied without any tree depth limitations, in order to investigate the way the algorithm behaves in this specific dataset. However, it was obvious that the created trees had a depth of over 10 and that resulted in overfitting. As in every dataset there are specific outliers and the algorithm was trying to create a sub tree that could cover their situation. The next step was the application of the decision tree with specific max depths. After several tries it was specified that the most accurate one was with max depth=8.

The decision tree with max depth=8 created some specific rules for every period. For the first period between 2007-2011 the initial split of the tree was on Driver's license type. From the initial 14871 instances, 14096 instances ended up at the left side of

the split where the driver's license was Learners, Regular or no license. At the right part of the split where all the instances where there was no information about the driver's license type. On that right part of the tree, two sub trees were created according to the age of the driver. Drivers aged between 75-99 ended up in the left subtree and everyone else in the left subtree.

From figure 5 it is obvious that this subtree refers to a small part of the dataset, with only 206 instances where most of the drivers are male. Since it was already known that all these instances belong to the unknown class, we can suppose that either there were misclassified, or their characteristics were really close to the male characteristic and the algorithm found their gender on its own according to the rest of the variables.

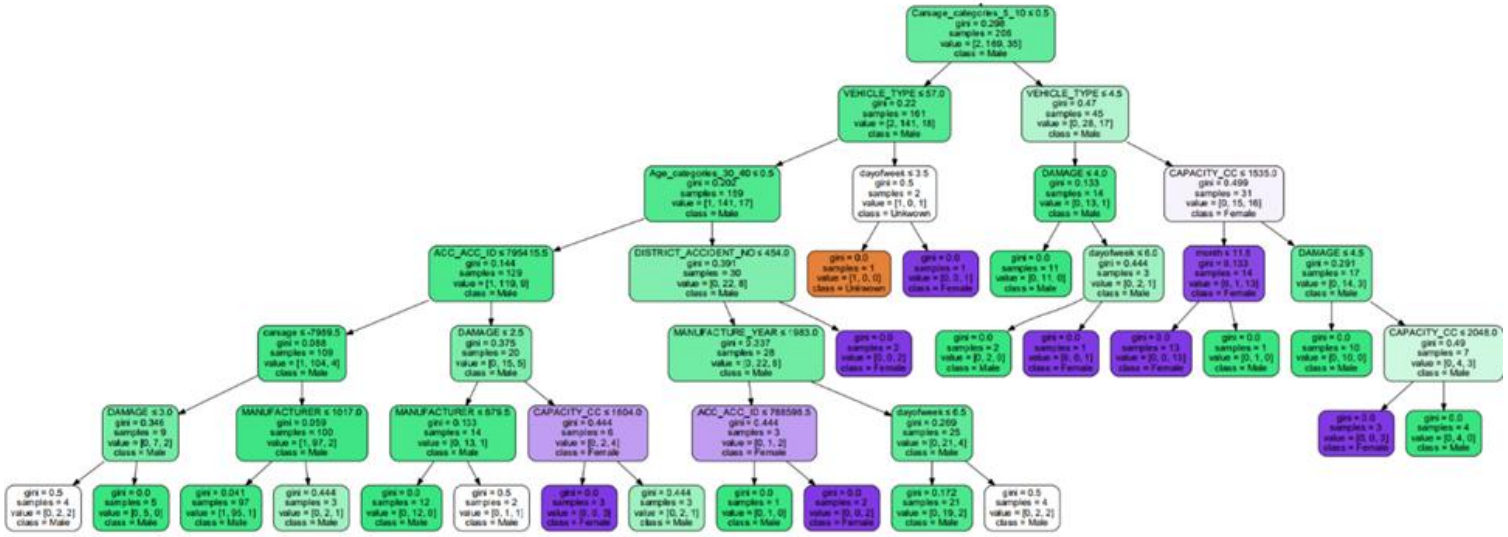

Figure 4-4: Drivers without information about their license type who belong to the age category 75-99 in 2007-2012

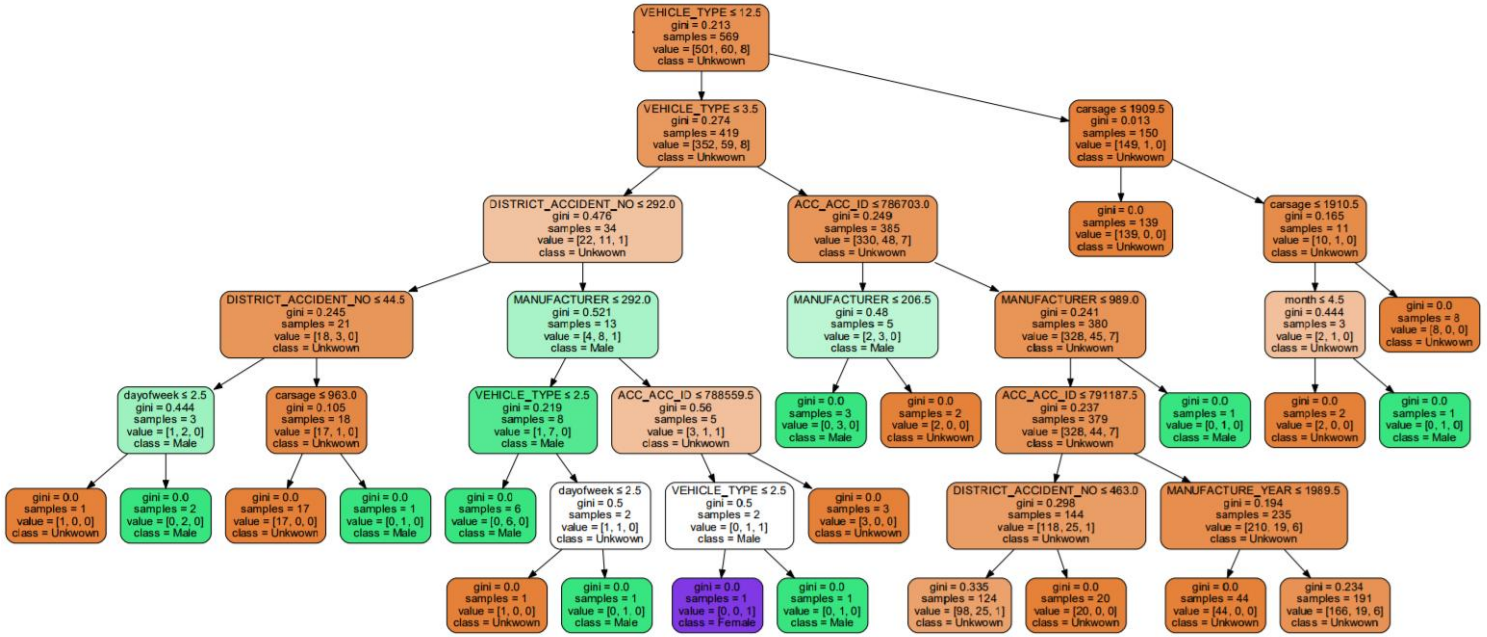

Figure 4-5 : Drivers without information about their license type who do not belong to the age category 75-99 in 2007-2012

In contrast with the previous subtree, the one illustrated in figure 6 most of the 569 initial records are correctly classified as Unknown, except from some misclassification of male as female. On the other hand, from the initial split of the tree on drivers with known license, the next split was on the cc capacity of the car. Next split is on capacity over 2008.5 cc with 3022 records and 11074 instances with less than 2008.5 cc with. Splitting continues on several other features.

By observing the whole visualization of the decision tree created for the period 2007-2011 strong specific patterns were noted. These patterns apply to at least 3% of the whole data with a good accuracy. Examples include the following patterns:

1. Driver License type = "Learners" or "Regular" or "No license" → Capacity_CC≤2008.5 →Vehicle Type≤5.5 →Vehicle Type ≤3.5 → Vehicle Type≤ 2.5 → Age category 30-40 = "Not" →Vehicle Type → Drivers gender= "Male" with 95,43% accuracy.
2. Driver License type = "Learners" or "Regular" or "No license" → Capacity_CC≤2008.5 →Vehicle Type≤5.5 →Vehicle Type ≤3.5 → Vehicle Type> 2.5 →Capacity_CC≥125.5 →Age category 75-99= "Not"→District accident Number ≤715→ Drivers gender= "Male" with 99,07% accuracy.
3. Driver License type = "Learners" or "Regular" or "No license" → Capacity_CC≤2008.5 →Vehicle Type>5.5 →Cars age≤12.5 → Capacity_CC>1395.5 → Capacity_CC>1607.5 →License Indicator≤1.5→Vehicle type ≤6.5 → Drivers gender= "Male" with 72,42% accuracy.
4. Driver License type = "Learners" or "Regular" or "No license" → Capacity_CC≤2008.5 →Vehicle Type>5.5 → Cars age≥12.5 →Insurance company >1.5 →Capacity_CC≤1513.5 → Cars age>15.5 → Manufacturer year>1932 → Drivers gender= "Male" with 80% accuracy.
5. Driver License type = "Learners" or "Regular" or "No license" → Capacity_CC≤2008.5 →Vehicle Type>5.5 → Cars age≥12.5 →Insurance company >1.5 → Capacity_CC>1513.5 → Manufacturer ≤ 955.5→ District accident number > 175.5 → Drivers gender= "Male" with 87% accuracy.
6. Drivers License type≤6.5 → Capacity_CC>2008.5 → Vehicle type > 6.5 → Vehicle type≤58.5 → Capacity_CC ≤2775.5 →Manufacturer≤998.5 → Capacity_CC ≤2773 → Manufacturer > 23.5→ Drivers gender= "Male" with 95.3% accuracy.

On the other hand, for the data of the next period we extracted the following patterns:

1. Driver's License type = "Learners" or "Regular" or "No license" → Capacity_CC≤1809.5 →Vehicle Type>4.5 → Cars age>15.5 → Manufacturer year ≤ 1992.5→ Account accident ID≤800911.5→ District accident number≤489 → Second event ≤13.5 → Drivers gender= "Male" with 92% accuracy.
2. Driver's License type = "Learners" or "Regular" or "No license" → Capacity_CC>1809.5 → Vehicle type ≤ 6.5 → Capacity_CC ≤2148.5 → License in-

dicator≤1.5 → Account accident id> 797622 → Accoun accident ID> 797918 → Insurance Company ≤52.5 → Drivers gender= "Male" with 77.77% accuracy.

3. Driver's License type = "Learners" or "Regular" or "No license" → Capacity_CC > 1809.5 → Vehicle type ≤6.5 → Capacity_CC > 2148.5 → District accaccident number ≤498 → Accident account id >797611→ Insurance company >23 → District accident number>45.5 → Drivers gender= "Male" with 95.37% accuracy.
4. Driver's License type = "Learners" or "Regular" or "No license"→ Capacity _CC>1809.5 → Vehicle type >6.5 → MAnufacturer≤1127 → Insurance company ≤51.5 → Manufacturer >44.5 → Account accident ID≤8016767→ Drivers gender= "Male" with 97.77% accuracy.

All extracted strong patterns from both periods refer to the male class. This is a result of the imbalance of the class attribute. The following variable classification is on driver's license type. In this classification approach the goal is to further investigate accident patterns related to the license type. The results shown in both periods (Figure 7) that almost 80% of the accidents are caused by drivers with regular driving license. Table 2 contains classifier accuracy results.

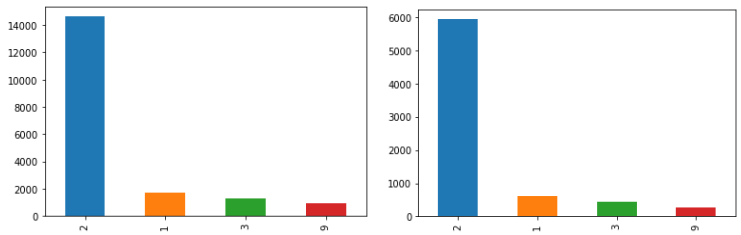

Figure 4-6: Drivers contribution to accidents according to their driving license (2007-2011 to the left and 2012-2014 to the right)

Table 2 Classifiers accuracy on driving license classification.

Classifier	2007-2011	2012-2014
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 the different classifier's accuracy shows that there is 10% difference in the classification accuracy between the two periods. Following that, the decision tree was visualized in order to observe the different patterns that may existed.

From the visualization of the decision tree it was obvious that there were some specific outlier values and overfitting. The first approach was without any limitation in the max tree depth. Following that the classifier was applied several times with different max depths in order to achieve a good accuracy and avoid overfitting. This goal was achieved with max depth 8 and an accuracy of 85.82% for the first period and 87.56% for the second period. Then the decision tree of both periods with max depth was visualized.

The first split of the decision tree for the period 2007-2011 occurred in the Drivers Gender, where in the left branch there are instances with unknown driver's gender (507 samples). The following splits for the unknown gender are the age categories 20-30 and 40-50, where only three samples belong to them and the rest do not. Finally, for the drivers whose cars manufacturer number was less than 545, there were 487 samples which were classified as unknown. The rest was also classified as unknown with manufacturer number bigger than 595, except from one sample.

In contrast, the other branch of the tree where all the genders are well known, the first split was by the insurance company attribute. Specifically, it splits cars in two categories: Those who do not have insurance (1824 samples) and all the others where the insurance company is known or there is no information about it (12540 samples).

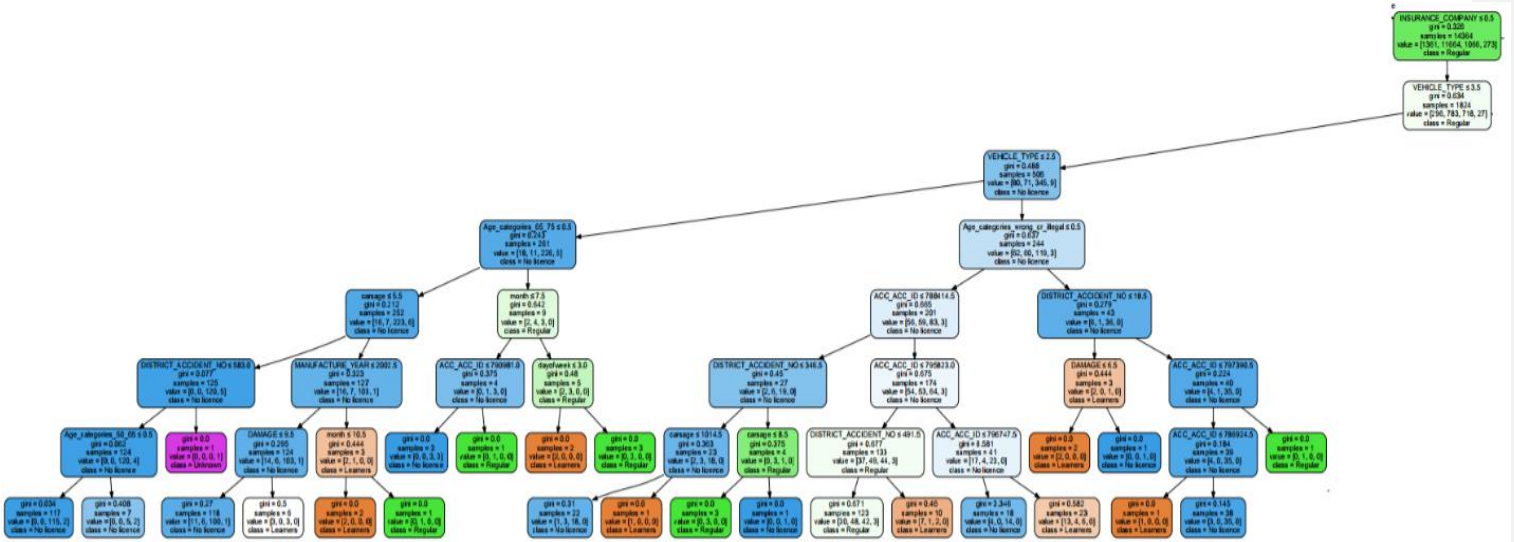

Figure 4-7: Driver's license type classification branch with bicycles and motorcycles without insurance (2007-2011)

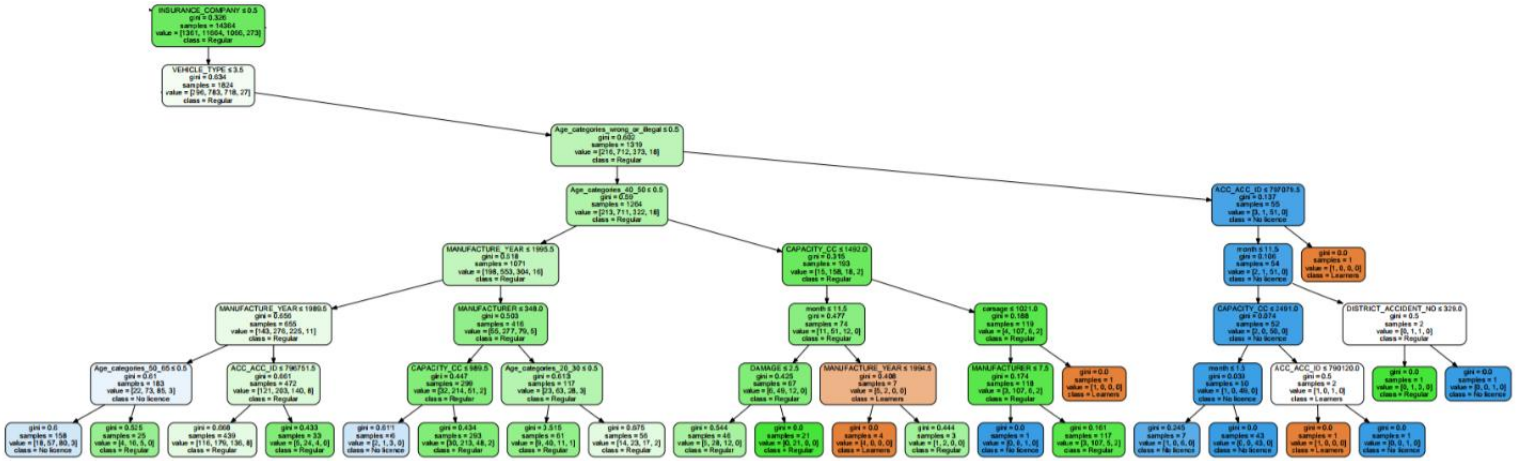

Figure 4-8: Driver's license type classification branch for all vehicles without insurance except bicycles and motorcycles (2007-2011)

For those without any insurance the next split happens according to their vehicle type. Moreover, it split the data in the first cluster where the bicycles and motorcycles belong (Figure 8) and the other cluster where all the other kind of vehicles belong (Figure 9). Following that splits in both branches it uses the age category split and the manufacturer year and the district number.

Returning to the branch where the insurance company is well known or unknown, the split took part according to the vehicle type. Especially from the 12540 samples two new branches created. The left branch contains all the motorbikes and the bicycles (1478 samples) and the second branch contains all the other vehicle types.

For the motorbikes and bicycles the next split depends on the age category of the driver. If the drivers do not belong to Wrong or illegal category, then the next split variable is the Insurance company. The other branch with the wrong or illegal ages of drives\(141 samples), is being spitted to 10 leaf nodes after 3 layers of dividing. Most of the samples are Learners.

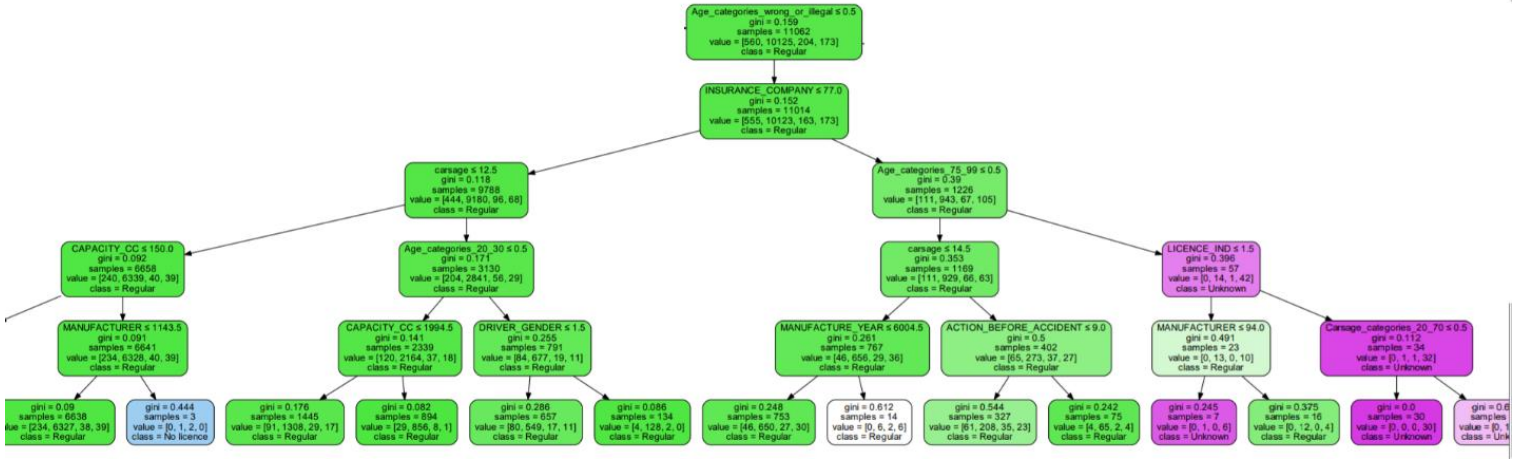

Figure 4-9: Drivers without wrong or illegal age license type classification branch for all vehicles except bicycles and motorcycles (2007-2011).

Turning back to the vehicle type split, all the vehicle types except motorbikes and bicycles (11062 samples) again are divided by the age category wrong or illegal. From these samples only the 44 belong to this age category and the most of them are without license. All the other samples are now being divided by the insurance company. Where the 9788 of the samples had insurance company number ≤77 and almost all of them had Regular driving license. The same happens for the other insurance's companies except 36 samples where the class is Unknown.

In contrast with the period 2007 -2011, in the next period the split variables change dramatically in the manner of hierarchy. To begin with, the first split depends in the vehicle type. If the vehicle type is motorbike or bicycle, then the next split is according to the insurance company. On the other hand, for all the other vehicles types the next split depends on the driver's gender and then the tree examines if the ages of the drivers are illegal or wrong

From all the above observation the following patterns were extracted for the period 2007-2011:

1. Drivers= "Male" or "Female" → Insurance company = "Unknown" → Vehicle type<3.5 → Vehicle type ≤2.5 → Age category 65-75= "Not" → Class = "No license" with 86.59% accuracy.
2. Drivers= "Male" or "Female" → Insurance company ≥1 → Vehicle type>3.5 → Age category wrong or illegal= "Not" → Class = "Regular" with 91% accuracy.

On the other hand, for the data of the next period the following patterns were extracted:

1. Vehicle type > 3.5 → Drivers gender = "Male" or "Female" → Age category wrong or illegal = "Not" → Insurance company≥1 → Class = "Regular" with 93.38% accuracy.

The next chosen variable for classification was the drivers age categories. From all the available ages 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 drivers age 4 years old. In this case was impossible to specify if there was a drivers 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". In the figures below, it is obvious that almost the same pattern of age categories contributed to accidents was noticed. Except the change of position between 40-50 and 50-65.

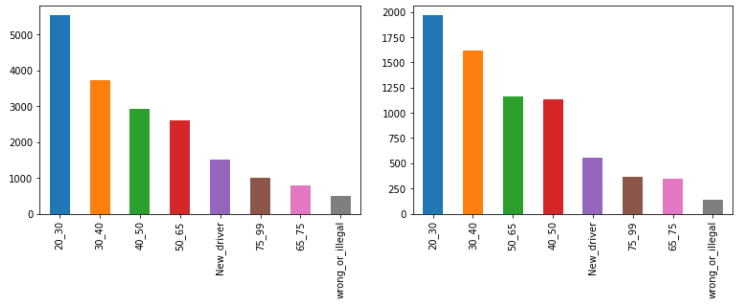

Figure 4-10: Drivers contribution to accidents according to their age category (2007-2011 to the left and 2012-2014 to the right).

Again, the visualization of the decision tree classifier was obvious that there were some specific outliers' values and the tree was overfitting. The first approach was without any limitation of the max depth of the tree. Following that the classifier applied several times with different max depths in order to achieve a good accuracy and avoid overfitting. This goal achieved with max depth 8. In this case dummy variables were created for the age categories. So, for every specific category there is a variable were the value is "0" or "1". The following table contains the results of the decision tree accuracy for every specific age category for the two periods.

Table 3 Classifiers accuracy of driver's age category classification.

Age category	**2007-2011**	**2012-2014**
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 the findings above it is obvious that for both periods the accuracy of the age categories “Wrong or illegal”, “New driver”, “65-75”, “75-99” is exceptional. Also, for the rest age categories we have a decent accuracy.

The creation of dummy variables had as a result the creation of an attribute for every age category. For every classification approach there were only two possible results “0” or “1”. That helped in the decision tree visualization.

The created decision tree for the “Wrong or illegal” category for the first period had as first splitting point the Capacity of CC<74.5. From the initial 14871 samples only the 803 satisfied that rule. On the other hand, the rest 14068 samples split on the Driver license type <2,5. The 12522 samples satisfy that rule and the next split was on even smaller number of driving license. And the splitting continues. The following figure shows the branch with the most samples.

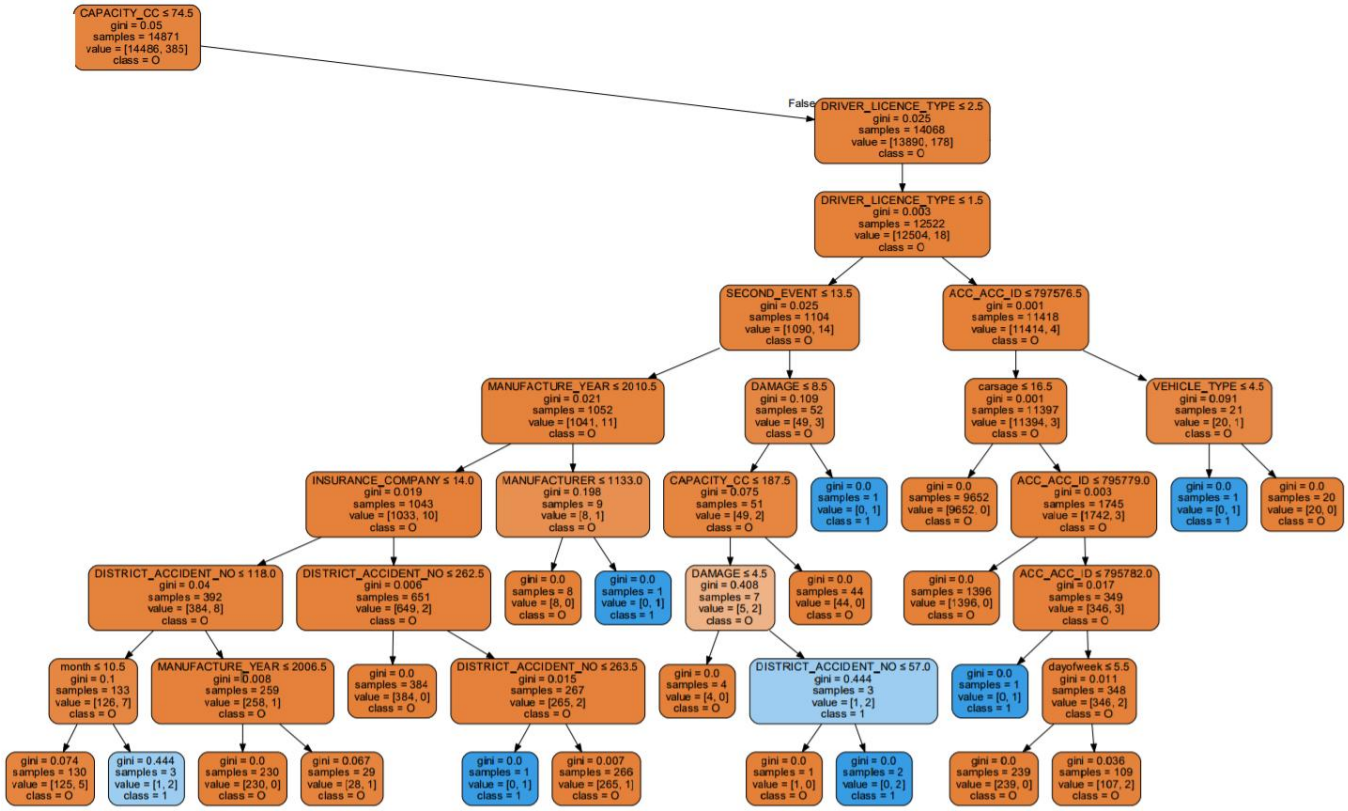

Figure 4-11: Driver's "Wrong or illegal" age category classification. The branch with the most samples (2007-2011).

The same procedure was followed also with the period 2012-2014. Where the most splits differed from the previous period. For the first period of accidents the number of drivers whose age was in that age category was 483 and the second period 136. According to that number and by assuming that a strong pattern has at least 10% no patterns found for the first period. In addition, the second period the following pattern was extracted:

1. Vehicle type ≤2.5 → Driver's license type >2.5 → Drivers license type >1.5 → District accident number >82 → District accident number ≤214 → Account accident id ≤801449.5 → Driver gender = "Male" or "Unknown" → Damage ≤2

The next decision tree that was created was for the age category "New Driver" which contains the ages between 17 and 20 years old. The decision tree of the first period starts with 14871. The first splits took place according to the Cars capacity and then two main branches created with 9555 samples and 5316 respectively. The following figure represents the branch with the most samples.

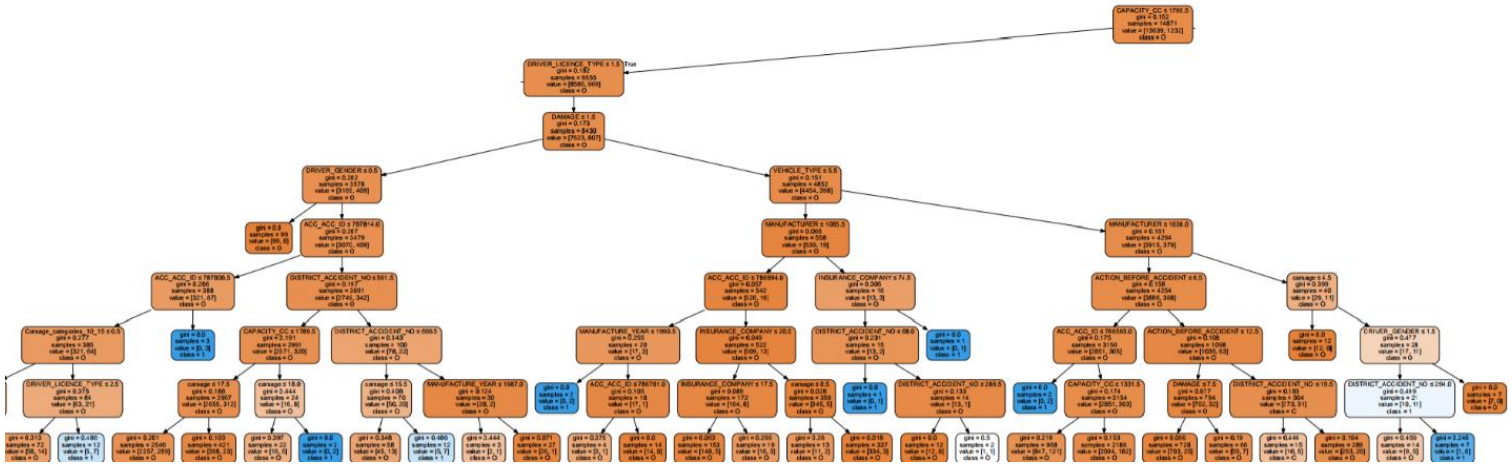

Figure 4-12: Driver's "New drivers" age category classification. The branch with the most samples (2007-2011).

The same procedure was followed also with the period 2012-2014. Where the most splits differed from the previous period. For the first period of accidents the number of drivers whose age was in that age category was 1518 and the second period 552. According to that number and by assuming that a strong pattern has at least 10% no patterns found for both periods.

The next decision tree that was created was for the age category “20-30”. The decision tree of the first period starts with 14871. The first splits took place according to the Cars capacity and then two main branches created with 11754 samples and 3117 respectively. The following figure represents the branch with the most samples.

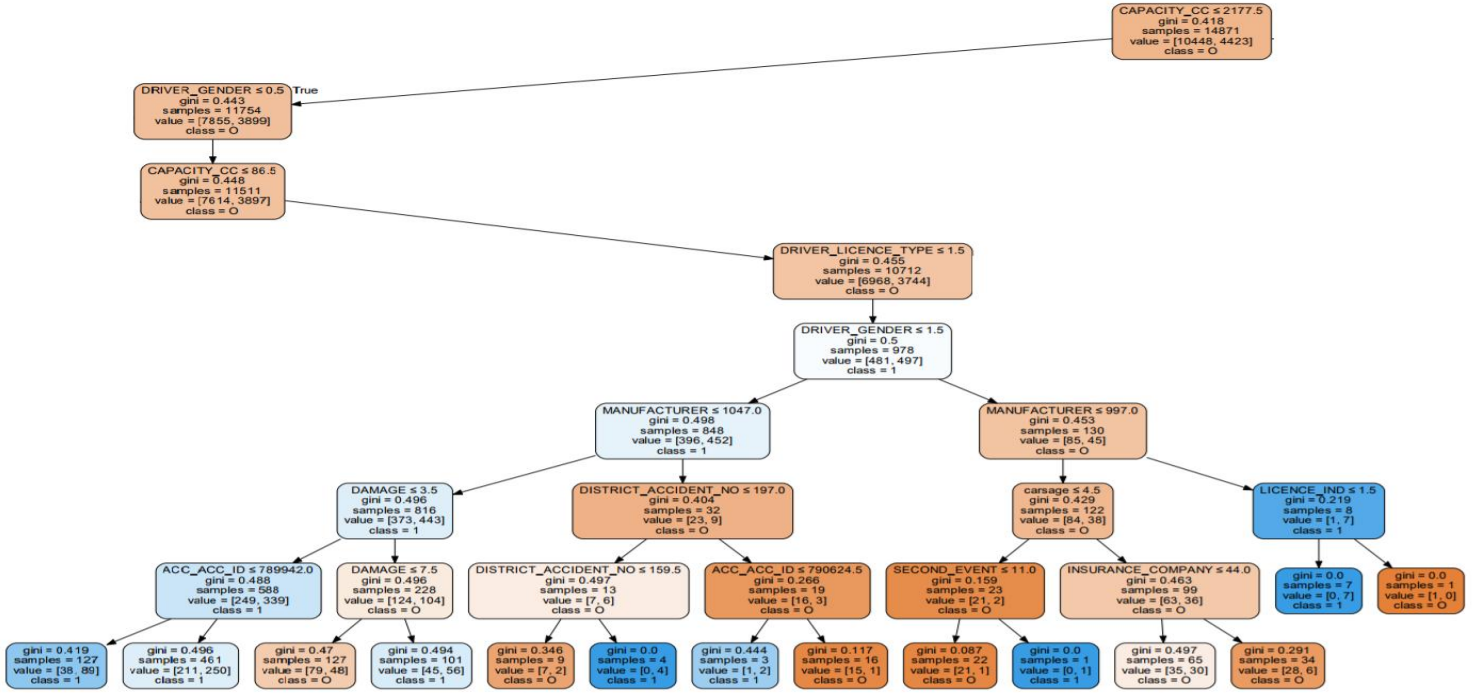

Figure 4-13: Driver's "20-30" age category classification. The branch with the most samples (2007-2011).

The same procedure was followed also for the period 2012-2014. Where the most splits differed from the previous period. For the first period of accidents the number of drivers whose age was in that age category was 5534 and the second period 1963. According to that number and by assuming that a strong pattern has at least 10% no patterns found for both periods. Finally, after the following the same procedure for the rest of the age categories, there were no strong patterns to be extracted.

Another crucial variable of the dataset was the vehicles' type, which was one of the variables with no missing values. The classification approach would help further understanding the circumstances under which the accidents happened according to the vehicle type. The following figure represents the number of accidents per vehicle type during 2007-2011 and 2012-2014.

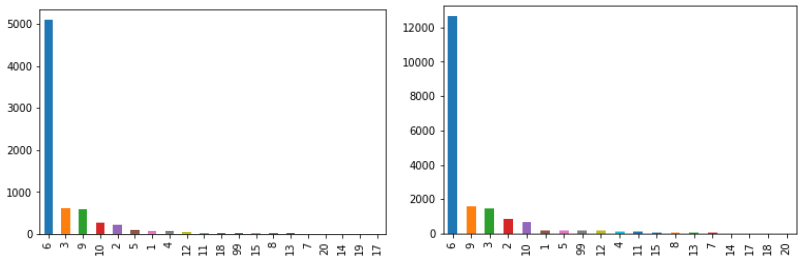

Figure 4-14: Contribution to accidents according to vehicle type (2007-2011 to the left and 2012-2014 to the right)

From the figures it is easily noticeable that the vehicle type with the biggest contribution to accidents in both periods is type "Other car". The following four type categories which contribute to an acceptable number of accidents have small changes in the sequence in the two periods. It is obvious that vehicle type 3 takes second place from vehicle type 9 in the period 2012-2014 and the same happens with vehicle type 10 and 2 respectively.

Table 4 Classifiers accuracy of vehicles type classification.

Classifier	2007-2011	2012-2014
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%

For the classifications approach of the vehicle type we used the default parameters except from the decision tree where different max depths were used until reaching to the one with the best accuracy. The optimum max depth was 8.

One more time the decision tree was visualized for further investigation of the classification procedure. Our main goal was the extraction of strong patterns according to the initial samples and the number of every class instances. It was obvious from the previous figure that the classes of the classification were unbalanced by far.

4.2 Human related data mining

The part of the dataset which was oriented to the data of the persons who involved in the accidents had the following features:

1. Accidents account identity
2. Vehicle consecutive
3. Position in vehicle
4. Protective measures
5. Ejection
6. Nationality
7. Age
8. Gender
9. Corps
10. Alcohol
11. Role in accident
12. Injury severity
13. Transfer to hospital
14. Hospital

15. Accident date

In the first period of accidents (2007-2011) there were no missing values in any feature. In the second period (2012-2014) there was insignificant number of missing values: 2 and 5 missing values out of 9323 in the transfer to hospital and hospital variables respectively.

4.2.1 Feature creation

In this specific part of the dataset there is information for every single person that took part in the accident. As a result, it was difficult to try to categorize even more the existed variables in order to achieve better accuracy such as in the previous part of the dataset. 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.

Also, in order to investigate the effect of the financial crisis in the car accidents in Cyprus these two periods financial data were merged with the initial dataset. The main economic indicators for Cyprus for 2005-2021 can be found at the website of the Ministry of finance of Cyprus, including predictions for 2018-2021. The data available are the following:

1. GDP at constant market prices 2005(%change)
2. Employment (persons, % change)
3. Unemployment rate Labor force survey
4. Harmonized Index of Consumer prices(%change)
5. Budget balance (% of GDP)
6. Public Dept (% of GDP)

For the classification approach it was decided to use only GDP at constant market prices % change and the Unemployment rate % change.

4.2.2 Classification approaches

The first classification approach of that part of the dataset was including the position of the individual passenger inside the vehicle. The goal of that approach was the correlation of several factors from the existing dataset with the position of the passenger inside the vehicle. As it was mentioned earlier, from the column which describes the date of the accident were extracted the year and the day of the week

that the accident happened. Then the initial date column was deleted. Also, the two financial columns were added.

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 the other types of passengers and when the position was unknown are illustrated by value "12".

The following table contains the accuracy of specific classifiers for the classification of the position of the passengers. For better classification results all the seating passengers' values were settled into "2". The classification approach's goal was the discrimination of the position of the passenger and especially if the passenger was driver, seating passenger, standing passenger or their position was unknown. The following table contains the accuracy of every single classifier as applied for the two periods of accidents.

Table 5 Classifiers accuracy of passenger position in vehicle classification.

Classifier	**2007-2011**	**2012-2014**
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 the results above, that the accuracy from all the classifiers is exceptional except from the Multi-layer perceptron which had not so good results in comparison with the other classifiers, whose accuracy was over 95%. The recorded accuracy of the decision tree was with max depth = 8, which was the best accuracy after several trials with different max depths. From the visualization of the decision tree with the Graph viz library, the following patterns were extracted for the period 2007-2011:

1. Role in accident≤ 17.5 → Role in accident >1.5 → Role in accident ≤16.5 → Role in accident ≤6.5 → Role in accident ≤5.5 → Role in accident ≤4.5 → Role accident >3.5 → Class= "Driver" with 100% accuracy (606 samples out of 7623).

2. Role in accident≤ 17.5 → Role in accident >1.5 → Role in accident ≤16.5 → Role in accident ≤6.5 → Role in accident >5.5 → Class= "Driver" with 100% accuracy (1039 samples out of 7623).
3. Role in accident≤ 17.5 → Role in accident >1.5 → Role in accident >16.5 → Ejection >0.5→ Class = "Driver" with 100% accuracy (2507samples out of 7623).
4. Role in accident >17.5 → Role in accident ≤18.5 → Ejection≤1.5 → Age ≤96.5 → Age≤16.5 → Injury severity>2.5 → Class= "Seated passenger" with 100% accuracy (214samples out of 7623).
5. Role in accident >17.5 → Role in accident ≤18.5 → Ejection≤1.5 → Age ≤96.5 → Age>16.5→ Class= "Seated passenger" with 100% accuracy (932 samples out of 7623).

For the next period of accidents (2012-2014) the patterns extracted are the following:

1. Alcohol≤4.5 → Protective measures>0.5 → Role in accident ≤17.5 → Role in accident ≤16.5→ Role in accident ≤15.5 → Role in accident ≤6.5 → Age >17.5 → Class= "Driver" with 100% accuracy (596 samples out of 7457).
2. Alcohol≤4.5 → Protective measures>0.5 → Role in accident ≤17.5 → Role in accident >16.5 → Class= "Driver" with 100% accuracy (3437 samples out of 7457).
3. Alcohol >4.5 → Protective measures >0.5 → Role in accident >17.5 → Role in accident ≤20.5 → Role in accident ≤19 → Class= "Seating passenger" with 100% accuracy (921samples out of 7457).

The classification approach of the number of vehicle consecutive is critical for future accidents. The goal of this classification is to identify specific patterns that lead to the human factors that cause accidents with more than two consecutive cars. Eventually the identification of the human factors that cause multiple vehicle collisions is a big advantage.

In that feature the variable can take values from 1 to 98 which illustrates the number of vehicles that took part in the accident. The following table contains the accuracy of every single classifier applied for the two periods of accidents.

Table 6 Classifiers accuracy of number of vehicle consecutive classification.

Classifier	2007-2011	2012-2014
Decision Tree	65.21%	64.71%
Random Forest	63.90%	61.44%
Gradient Boosting	66.78%	52.11%
MLPC	37.88%	39.57%
Voting classifier	65.16%	60.48%

From the table above, it is noticed that the accuracy for all classifiers is below 70% and, in some cases, such as the multi-layer perceptron, the accuracy is lower than 40%. For that reason, we assume that our classifiers are not strong, and we are unable to extract patterns from them.

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 protective measure variable illustrates 5 different types of safety equipment that were used by the person involved. With value 1 referring to no restraint used, value 2 using seat belt, value 3 child restraint, 4 using helmet and 5 unknown protective measure.

Table 7 Classifiers accuracy of using protective measures classification.

Classifier	2007-2011	2012-2014
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 the accuracy table it is obvious that the accuracy is satisfactory except from the Gradient boosting and Multilayer perceptron with accuracy less than 50% in the first period and not so good accuracy in contrast with other classifiers. From the visualiza-

tion of the decision tree with the Graph viz library the following patterns were extracted for 2007-2011:

1. Role in accident ≤7.5 → Role in accident>1.5 → Role in accident >3.5 → Age > 23.5 → Vehicle seq>1.5 → Age ≤84→ ACC_ACC_ID> 786569.5 → Injury severity >1.5 → Class= "Helmet" with 84.81% accuracy (620 samples out of 7623).
2. Role in accident <7.5 → Ejection≤1.5 → Role in accident ≤25.5 → Age ≤3.5 → Vehicle seq≤1.5 → Injury severity >1.5 → Position in vehicle ≤7 → ACC_ACC_ID≤788681 → Class = "Seat belt" with 80.96% accuracy (268 samples out of 7623).
3. Role in accident <7.5 → Ejection≤1.5 → Role in accident ≤25.5 → Age ≤3.5 → Vehicle seq≤1.5 → Injury severity >1.5 → Position in vehicle ≤7 → ACC_ACC_ID>788681 → Class = "Seat belt" with 77.55% accuracy (1113 samples out of 7623).
4. Role in accident <7.5 → Ejection≤1.5 → Role in accident ≤25.5 → Age ≤3.5 → Vehicle seq>1.5 → Ejection >0.5 → ACC_ACC_ID≤794955.5 → Nationality≤2.5 → class= "Seat belt" with 89.43% accuracy (1253 samples out of 7623).
5. Role in accident <7.5 → Ejection≤1.5 → Role in accident ≤25.5 → Age ≤3.5 → Vehicle seq>1.5 → Ejection >0.5 → ACC_ACC_ID>794955.5 → ACC_ACC_ID> 794957 → Class= "Seat belt" with 81.84% accuracy (573 samples out of 7623).

For the next period of accidents (2012-2014) the patterns extracted are the following:

1. Position in vehicle >0.5 → Role in accident >8.5 → Age≤89.5 → Ejection ≤1.5 → Age>5.5 →Role in accident ≤28 → Injury severity >1.5 → Alcohol≤7 → Class= "Seat belt" with 82.08% accuracy (4320 samples out of 7457).

On the other hand, the classification of ejection variable aimed at the identification of the circumstances under which specific passengers were ejected from the vehicle during the accident.

The ejection variable which illustrates if a passenger was ejected from the vehicle can take values from 1 to 4. Value 1 refers to passengers who were not ejected, value 2

to these who were partially ejected, value 3 to these who were ejected and finally value 4 to passengers that is unknown if they were ejected or not.

Table 8 Classifiers accuracy of passenger ejection classification.

Classifier	**2007-2011**	**2012-2014**
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%

The accuracy of the classifiers for the ejection of a passenger or not from the vehicle during the accident is satisfactory. It is close to 80% for the first period of accidents and almost reached the 90% for the second period of accidents. The decision tree was visualized and the extracted patterns for the first period of accidents is the following:

1. Role in accident ≤7.5 → Position in vehicle >0.5 → Injury severity ≤2.5 → Age≤29.5 → Transfer to hospital ≤1.5 → Protective measures ≤2.5 → Role in accident >4.5 → ACC_ACC_ID >787369.5→ class= "Ejected" with 86% accuracy (148 samples out of 7623).
2. Role in accident >7.5 → Protective measures ≤1.5 → Injury severity >2.5 → ACC_ACC_ID → >789161 → Position in vehicle >0.5 → Role in accident ≤32.5 → ACC_ACC_ID → 797167.5 →Role in accident ≤26.5 → Class= "Not ejected" with 91.37% accuracy (290 samples out of 7623).
3. Role in accident >7.5 → Protective measures > 1.5 → Protective measures ≤4.5 → ACC_ACC_ID≤794613.5 → ACC_ACC_ID≤794607 → Transfer to hospital ≤3.5 → Position in vehicle ≤7 → Day of week >1.5 → Class= "Not ejected" with 89.64% accuracy (1265 samples out of 7623).
4. Role in accident >7.5 → Protective measures > 1.5 → Protective measures ≤4.5 → ACC_ACC_ID≤794613.5 → ACC_ACC_ID≤794607 → Transfer to hospital >3.5 → Day of week ≤5.5 → ACC_ACC_ID≤792888.5 → Class= "Not ejected" with 90.4% accuracy (452 samples out of 7623).

For the next period of accidents (2012-2014) the patterns extracted are the following:

1. Protective measures>0.5 → Role in accident>8.5 → Nationality ≤6 → Injury severity≤2.5 → Protective measures >1.5 → Role in accident ≤32 → ACC_ACC_ID≤800810.5 → ACC_ACC_ID ≤800328.5→ Class= "Not ejected" with 99.16% accuracy (238 samples out of 7457).
2. Protective measures>0.5 → Role in accident>8.5 → Nationality ≤6 → Injury severity>2.5 → Role in accident ≤32.5 → ACC_ACC_ID≤800533.5 → CORPS≤7.5 → ACC_ACC_ID≤798106.5 → Class= "Not ejected" with 96.57% accuracy (620 samples out of 7457).
3. Protective measures>0.5 → Role in accident>8.5 → Nationality ≤6 → Injury severity>2.5 → Role in accident ≤32.5 → ACC_ACC_ID≤800533.5 → CORPS≤7.5 → ACC_ACC_ID>798106.5 → Class= "Not ejected" with 99.11% accuracy (3040 samples out of 7457).
4. Protective measures>0.5 → Role in accident>8.5 → Nationality ≤6 → Injury severity>2.5 → Role in accident ≤32.5 → ACC_ACC_ID>800533.5 → ACC_ACC_ID> 800536.5 → Protective measures>1.5 → Class = "Not ejected" 96.65% accuracy (1271 out of 7457).

Furthermore, the classification of Corps variable aims to identify if the involved person in the accident belongs to a Corp or not. The crop variable takes values from 1 to 6. Value 1 refers to the police, value 2 refers to the national guard, value 3 refers British bases, value 4 refer to the UNFICYP, value 5 refer to ELDYK and value 6 refers to every other corp.

Table 9 Classifiers accuracy of CORPS classification.

Classifier	**2007-2011**	**2012-2014**
Decision Tree	94.85%	94.63%
Random Forest	95.90%	95.22%
Gradient Boosting	94.22%	94.95%
MLPC	94.91%	93.83%
Voting classifier	95.59%	95.06%

The classifiers accuracy of the CORPS classification approach was remarkable for of the classifiers that were used. The recorded accuracy of the decision tree was with max depth = 8, which was the best accuracy after several trials with different max depths. From the visualization of the decision tree with the Graph viz library the only

patterns extracted from both periods were for people who do not belong to Corps. That's because most people do not belong to Corps and the classes number of instances is unbalanced. As a result, we have an exceptional classification accuracy, however that does not mean that our classifier is good for all the classes.

Another crucial classification was that of alcohol and drugs used by passengers. That classification could identify the pattern of people's behavior after using alcohol or drugs and even specify specific limit that passengers cause less damage.

The variable takes values from 1 to 6. Value "1" refers to passengers who were not involved neither with alcohol nor drugs. Value "2" refers to passengers positive to alcohol. Value "3" refers to passengers who failed to provide samples. Value "4" refers to passengers who are positive to drugs. Value "5" refers to passengers who were not demanded to take the test and the value "6" to those that is unknown what happened.

Table 10 Classifiers accuracy of driver's alcohol and drugs test classification.

Classifier	**2007-2011**	**2012-2014**
Decision Tree	69.56%	83.05%
Random Forest	71.09%	82.52%
Gradient Boosting	68.15%	82.30%
MLPC	56.55%	61.93%
Voting classifier	67.94%	81.50%

The table above illustrates the accuracy of the classifiers applied to the dataset for the classification of alcohol consumption per person inside the vehicle. The first period of accidents the accuracy of all the classifier is not so good, it is near 70% and in a specific case like the multi-layer perceptron classifier it is lower than 60%. On the other hand, the next period of accidents the classification accuracy is really improved and overcame 80% in most cases except from the multi-layer perceptron which just overcame 60%. For the first period of accidents we were unable to extract strong patterns. However, on the second period of accidents the following patterns were extracted:

1. Position in vehicle ≤1.5 → Protective measures > 0.5 → Age ≤89.5 → Vehicle seq≤1.5 → Nationality ≤1.5 → Injury severity >1.5 → Day of week >1.5 → Day of week ≤6.5→ Class= "Under the limit" with 84.82% accuracy (1235 samples out of 7457).

2. Position in vehicle ≤1.5 → Protective measures > 0.5 → Age ≤89.5 → Vehicle seq>1.5 → ACC_ACC_ID≤ 801479 → Role in accident >5 →Injury severity >1.5 → Month>11.5 → Class= "Under the limit" with 93.27% accuracy (1679 samples out of 7457).
3. Position in vehicle >1.5 → ACC_ACC_ID>797792 → Role in accident≤29 → Month ≤9.5 → Role in accident >9 → GDP>-4.5 → ACC_ACC_ID≤801158.5 → ACC_ACC_ID >798142.5 →Class = "Test not demanded" with 99.47% accuracy (381 samples out of 7457).

The classification of the role accident illustrates the role accident for every single person that was involved in the accident. The variable takes values from 1 to 36. All values describe the role of the passenger. i.e. if they were inside a vehicle or pedestrians. If inside a vehicle, it describes if it was the driver or a passenger and the kind of vehicle.

Table 11 Classifiers accuracy of person's role in accident classification.

Classifier	**2007-2011**	**2012-2014**
Decision Tree	76.81%	77.69%
Random Forest	76.86%	75.81%
Gradient Boosting	11.54%	0.05%
MLPC	32.21%	54.53%
Voting classifier	69.72%	69.97%

The accuracy of the classifiers for the role in accident is moderate in both periods. It is assumed that the accuracy is satisfactory for extracting patterns for further investigation. The decision tree was visualized and the extracted patterns for the first period of accidents is the following:

1. Position in vehicle≤1.5 → Position in vehicle > 0.5 → Ejection≤1.5 → Protective measures≤3 → Gender ≤1.5 → Age≤41.5 → Protective measures >1.5 → Age ≤27.5 → Class= "Car driver" with 88.93% accuracy (458samples out of 7623).
2. Position in vehicle≤1.5 → Position in vehicle > 0.5 → Ejection≤1.5 → Protective measures≤3 → Gender ≤1.5 → Age≤41.5 → Protective measures >1.5 → Age >27.5 → Class= "Car driver" with 75.67 % accuracy (308 samples out of 7623).

3. Position in vehicle≤1.5 → Position in vehicle > 0.5 → Ejection≤1.5 → Protective measures≤3 → Gender >1.5 → ACC_ACC_ID>786722.5 →Alcohol ≤7 →Age > 15.5 → Class= "Car driver" with 95.46% accuracy (821 samples out of 7623).
4. Position in vehicle≤1.5 → Position in vehicle > 0.5 → Ejection≤1.5 → Protective measures>3 → Protective measures ≤8 → Age≤65 → Age>18.5 → Protective measures≤5 → Class= "Driver of motorcycle" with 74% accuracy (101samples out of 7623).
5. Position in vehicle≤1.5 → Position in vehicle > 0.5 → Ejection>1.5→ Protective measures>3 → Age>18.5 → Age≤52.5 → Protective measures ≤5 → Nationality≤1.5 → Class = "Driver of motorcycle" with 88% accuracy (413samples out of 7623).
6. Position in vehicle>1.5 → Ejection ≤1.5 → Age≤33.5 → Corps ≤1.5 → Protective measures>1.5 → Gender ≤1.5 → Age≤25.5→ Hospital ≤2.5 → Class = "Car passenger" with 88.88% accuracy (272 samples out of 7623).
7. Position in vehicle>1.5 → Ejection ≤1.5 → Age≤33.5 → Corps ≤1.5 → Protective measures>1.5 →Gender > 1.5 → Nationality≤1.5 → ACC_ACC_ID > 789814 → Class= "Car passenger" with 98.67% accuracy (223 samples out of 7623).
8. Position in vehicle>1.5 → Ejection >1.5 → Protective measures >3.5 → Protective measures ≤5 → Age >20.5 → Unemployment >3.8 →Month≤11.5 → Age>21.5 → Class= "Passenger motorcycle" with 96.96% accuracy (32 samples out of 7623).
9. Position in vehicle≤1.5 → Position in vehicle ≤0.5 → Vehicle seq≤1 → Class= "Pedestrian" with 100% accuracy (834 samples out of 7623).

For the next period of accidents (2012-2014) the patterns extracted are the following:

1. Position inn vehicle ≤1.5 → Vehicle seq ≤0.5 → Class = "Pedestrian" with 100% accuracy (371 samples out of 7457).

4.3 General Data classifications

The part of the dataset that contains the general data about the circumstances that the accident happened has only three attributes with many missing values which are:

1. Ambulance called

2. Ambulance arrived
3. Ambulance time

All the values refer to the time variables which illustrate the times around the ambulance. Also, there are four more variables with a few missing values which are:

1. Police called time
2. Police arrived time
3. Police time
4. Ambulance called by

4.3.1 Feature creation

It is crucial for the classification approach to create new features. Four new features were extracted from the accident day variable for further processing. The year of the accident, the month, the day of the week and if it was weekend or not. Also, from the time variable the hour of the day that the accident happened was extracted. Furthermore, from the visualization of the time accidents happened one more variable was created, called time teams. The accidents according to the amount of accidents happened in the time of the 24 hours were classified into groups.

Moreover, since Cyprus is an island that in specific parts of the year there are a lot of tourists, three more variables were created in a dummy variable manner. The "Months of high tourism", "Month of low tourism" and the "Months of regular tourism" were created. According to the part of the year that accident took part was classified to one of these variables with value "1" where it belongs and value "0" where it does not belongi.

4.3.2 General Data classification

After the new features that were created in the csv file the first classification approach was about the month that the accident happened. That classification has as an aim the identification of patterns correlated with the month that the accident happened. In the following table there are the accuracies of the used classifiers per period of accidents.

Table 12: Classifiers accuracy of the month that the accident happened classification.

Classifier	2007-2011	2012-2014
Decision Tree(Max depth=12)	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%

From the classifications above we see that the accuracy of the classifiers is good enough to extract strong patterns for the classes of the classifier. All the classifiers were implemented with the default settings except from the decision tree which was implemented with different max depths until the most accurate found.

Another classification that was implemented was for the accident type variable, which illustrates the accident fatality. There were four different classes. The "fatal" where they were deaths, "Serious injury" where people were injured seriously, the "Slight injury", where people where soft injuries and the "Damage" where no one was injured except some damages in the vehicles.

Table 13: Classifiers accuracy of the accident's type classification

Classifier	2007-2011	2012-2014
Decision Tree(Max depth=5)	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%

The table above illustrates the accuracy of the classifiers that was applied for the classification of accident type. The accuracies are accepted for further extraction of patterns. For the period 2012 -2014 no patterns extracted in contrast with the period 2007-2011 where the following patterns were extracted:

1. No_Injured>0.5→Photos_Ind≤1.5→Police_officer= not AA → Point_AZZ527= "Not" →Ambulance_time ≤24.5 Class= "Fatal".

2. No_Injured>0.5→Photos_Ind>1.5→Ambulance_Time>1.5 → Police_District>53 → First_event≤10.5→ Class= “Serious Injury”

Moreover, the classification of the weekend variable is crucial. The aim of this classifications is the identification of any impact of the number of accidents or the fatality of them according to the weekend. Variable. The variable takes values “0” if the accident happened from Monday to Friday and value “1” if the accident happened in the weekend.

Table 14: Classifiers accuracy of the Weekend variable classification.

Classifier	**2007-2011**	**2012-2014**
Decision Tree(Max depth=5)	57.57%	57.14%
Random Forest	55.04%	54.97%
Gradient Boosting	59.60%	57.25%
MLPC	53.01%	55.35%
Voting classifier	55.34%	47.70%

The classifiers used for the classification of the variable had not good accuracy. Furthermore, because the accuracy is less than 70% is unacceptable for pattern extraction.

On the other hand, the classification of the number of vehicles contributed to cars could play leading role to strong pattern extraction. The identification of the reason why many vehicles contributes to the same accident.

Table 15: Classifiers accuracy of the number of cars contributed to the accident classification

Classifier	**2007-2011**	**2012-2014**
Decision Tree(Max depth=5)	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%

From the table above, we observe that the accuracies of the classifiers are good enough for pattern extraction. The extracted patterns for both periods had as basics splitting features those who illustrates the area where the accident took part.

5 Conclusions and recommendations

5.1 Conclusions

In this study the basic aim was the analysis of the accident dataset from Cyprus during 2007-2011 and 2012-2014. The first conclusions came from the visualizations of the dataset. In both periods the biggest percentage of drivers who contribute to accidents are between 19-40 years old. In both periods the age category with the biggest contribution is 20-30 years old. Also, 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 classifiers were implemented. From the decision tree classifier and with the help of the Graph viz library of Python the tree was visualized in order to extract strong patterns.

According to the gender classification, the patterns that were extracted refer only to men. In the first period men’s age who were involved in accidents with bicycles and motorcycles up to 50cc were below 30 years old or above 40 years old. Men who contributed to accidents with motorcycles between 125 cc and 2008 cc were less than 75 years old and all these accidents were caused 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 1513 and 2008 and those with cc between 2008 and 2773. 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 1809 cc. and 2148 cc.

Another classification approach from which strong patterns were extracted was the classification of Driving license. In the first period, drivers without license involved in accidents with motos up to 50cc and age over 75 years old or less than 65 years. Also,

the drivers with regular driving licenses who were involved in accidents were over 18 years old.

Following that, from the age categories classification there were no patterns extracted except from one. For drivers younger than 17, which had the legal right to drive that was called as wrong ages recordings or illegal, one 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 moped up to 50cc.

From the classification of position in vehicle 4, two strong patterns where extracted for every period. In 2007-2011, car passengers whose age was less than 16.5 were slightly injured or not injured at all. On the other hand, in the next period the passengers and drivers of bicycles and motorcycles whose age was over 17.5 years old did not use drugs. Also, car drivers were using seat belts and did not use drugs.

Another classification that gave patterns was that of protective measures. The classification approach showed for the first period, that passengers aged 24-84 involved in accidents with more than two vehicles while they were riding a motorcycle, were wearing a helmet. Also, infants less than 3.5 years old, wearing seat belts, involved in cars accidents, 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 "AA" 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.

5.2 Future work

This dissertation's main aim was the extraction of 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.

Principal component analysis could be used in the future for the improvement of the classifier. Principal components analysis is already being used in real life problems [25]. There are more than 50 features in the dataset; principal component analysis could decrease its dimensionality and improve the speed and accuracy of calculations.

Another approach for future work could be different preprocessing of the dataset. Our approach involved filling in missing values with a specific value for unknown data.

However, there are other approaches such as filling with 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 converged 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 that drive. Also, applications such as google maps could give real time warnings to the users, especially for tourists who are unfamiliar with the roads of the island, according to the places, the average speed and the general circumstances that the accidents happened.

References

Με σχόλια [CT3]: Homogenise format and double check

[1] Tzirakis P. and Tjortjis C., “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, 2017

[2] Tjortjis C., Saraee M., Theodoulidis B., Keane J.A, “Using T3, an Improved Decision Tree Classifier, for Mining Stroke Related Medical Data”, Methods of Information in Medicine, 46:5, October 2007, pp. 523-529

[3] Tjortjis C. and Keane J.A., 'T3: an Improved Classification Algorithm for Data Mining' in Lecture Notes Computer Science Vol. 2412, pp. 50-55, 2002, Springer-Verlag.

[4] K. Geetha, C. Vaishnavi (2015), Analysis on Traffic Accident Injury Level Using Classification. International Journal of Advanced Research in Computer Science and Software Engineering

[5] Miao M. Chong, Ajith Abraham, Marcin Paprzycki, Traffic Accident Analysis Using Decision trees and Neural Networks

[6] S. Krishnaveni, 1M. Hemalatha(2011), A Perspective Analysis of Traffic Accident using Data Mining Techniques

[7] Naina Mahajan, Bikram Pal Kaur (2016), Analysis of Factors of Road Traffic Accidents using Enhanced Decision Tree Algorithm

[8] S.R. Safavian, D. Landgrebe(1991) A survey of decision tree classifier methodology

[9] Kemal Polat, Salih Gunes (2007), Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform

[10] Wenliang Du, Zhijun Zhan (2002), Building decision tree classifier on private data

[11] Yifei Chen, Zhenyu Jia (2013) A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index

[12]T Chen, C Guestrin (2016), XGBoost: A scalable tree boosting system

[13]G Ke, Q Meng, T Finley, T wang, W Chen, Lightgbm: A highly efficient gradient boosting decision tree

[14] VF Rodriguez-Galiano, B Ghimire, J Rogan, (2012), An assessment of the effectiveness of a random forest classifier for land-cover classification

[15] L Fraiwan, K Lweesy, N Khasawneh, H Wenz (2012), Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier

[16] Y Zhang, H Zhang, J Cai, B Yang, A weighted voting classifier based on differential evolution

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

[18] Tiwari, Prayag, Huy Dao, and Gia Nhu Nguyen. "Performance Evaluation of Lazy, Decision Tree Classifier and Multilayer Perceptron on Traffic Accident Analysis." Informatica 41.1 (2017).

[19] Tiwari, Prayag, Sachin Kumar, and Denis Kalitin. "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques", International Conference on Computational Intelligence, Communications, and Business Analytics. Springer, Singapore, 2017.

[20] Kumar, Sachin, Prayag Tiwari and Kalitin Vladimirovich Denis. "Augmenting Classifiers Performance through Clustering: A Comparative Study on Road Accident Data." IJIRR 8.1 (2018): 57-68. Web. 17 Nov. 2017.

[21] Tiwari P., Nguyen G.N., Prasad M., Pratama M., Ashour A.S., Dey N., Analysis of Airplane crash by utilizing Text Mining Techniques- Accepted by Acta Informatica.

[22] K Sachin, Shemwal V.B., Solanki V., Tiwari P., K Denis. “A Conjoint Analysis of Road Accident Data using K-modes clustering and Bayesian Networks,” Annals of Computer Science and Information System, Volume 10, 53-56.

[23] Tiwari, P., Mishra, B. K., Kumar, S., & Kumar, V. (2017). Implementation of n-gram Methodology for Rotten Tomatoes Review Dataset Sentiment Analysis. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 7(1), 30-41.

[24] Tiwari P., Nguyen G.N., Kumar S., Yadav P., Ashour A.S., Dey N., Sentiment Analysis based on Russian and English Review Dataset, accepted in Statistical Analysis and Data Mining: The ASA Data Science Journal.

[25]. Prayag Tiwari. Comparative Analysis of Big Data. International Journal of Computer Applications 140(7):24-29, April 2016. Published by Foundation of Computer Science (FCS), NY, USA.

[26] P. Tiwari, "Improvement of ETL through integration of query cache and scripting method," 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, 2016, pp. 1-5.

[27] P. Tiwari, "Advanced ETL (AETL) by integration of PERL and scripting method," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2016, pp.1-5.

[28] P. Tiwari, S. Kumar, A. C. Mishra, V. Kumar and B. Terfa, "Improved performance of data warehouse," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2017, pp.94-104.

[29] Verma Saurbah, “CARE database common accident data set”,29, November,2018.

[30] “World Bank. 2017. The High Toll of Traffic Injuries: Unacceptable and Preventable. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/29129 License: CC BY 3.0 IGO.”

Appendix

CARD NO. 1: GENERAL ACCIDENT DATA

Field No.	ENGLISH FIELD NAME	FIELD DESCRIPTION	VALUES	DESCRIPTION OF VALUES
1	AREA_CODE	CODE FOR ACCIDENT LOCATION (URBAN OR RURAL)	T	TOWN
			R	RURAL
2	ACCIDENT_ TYPE	ACCIDENT SEVERITY	1	FATAL
			2	SERIOUS INJURY
			3	SLIGHT INJURY
			4	DAMAGE
3	POLICE_ DISTRICT	CODE NUMBER FOR TOWN OR DISTRICT WHERE THE ACCIDENT OCCURED	10	NICOSIA-RURAL
			11	NICOSIA-TOWN
			20	FAMAGUSTA-RURAL
			30	LIMASSOL-RURAL
			33	LIMASSOL-TOWN
			40	LARNAKA-RURAL
			44	LARNAKA-TOWN
			60	PAFOS-RURAL
			66	PAFOS-TOWN
			70	MORFOU-RURAL
4	PO-LICE_STATION	CODE NUMBER OF POLICE STATION WHICH INVESTIGATED THE ACCIDENT	2112	NICOSIA DIVISION TRAFFIC BRANCH
			2150	AGIOS DOMETIOS
			2151	LICAVITOS
			2152	OMORFITA
			2153	PILI PAFOU
			2154	STROVOLOS
			2160	DEFTERA
			2161	KLIROU
			2162	PALECHORI
			2163	PERA CHORIO
			2164	PERISTERONA
				KOKKINOTRIMITHIA

			2165	LAKATAMIA
			2170	FAMAGUSTA DIVISION TRAFFIC BRANCH
			2512	AGIA NAPA
				AVGOROU
			2560	DERYNIA
			2561	XILOTYMPOU
			2562	XILOFAGOU
			2563	PARALIMNI
			2564	LIMASSSOL DIVISION TRAFFIC BRANCH
			2565	LIMASSOL -CENTRAL STATION
			2312	AGIOS IOANNIS
				AGIOS NIKOLAOS
			2350	AGROS
			2352	AVDIMOU
			23502	GERMASOGIA
			2360	EPISKOPI
			2361	KALO CHORIO
			2351	LANIA
			2362	MONI
			2363	PACHNA
			2364	PLATRES
			2365	TROODOS
			2366	PISSOURI
			2367	POLEMIDIA
			2368	LARNACA DIVISION TRAFFIC BRANCH
			2370	ATHIENOU
			2376	ARADIPPOU
			2212	KALAVASOS
				KITI
			2260	KOFINOU
			2261	LEFKARA
			2262	OROKLINI
			2263	**ZIGI**
			2264	PAFOS DIVISION TRAFFIC BRANCH
			2265	KOUKLIA
			2266	PANAGIA
			2277	POLI CHRISOCHOUS
			2412	STROUMPI
			2460	PEGIA
			2461	KELOKEDARA
			2462	MORFOU DIVISION TRAFFIC BRANCH
			2464	ASTROMERITIS

			2465 2466 2612 2660 2661 2662 2663 2664 **2665**	EVRICHOU KAKOPETRIA KAMPOS PEDOULAS PIRGOS
5	**(AR) DIS-TRICT_ ACCI-DENT_ NO	CONSECUTIVE NUMBER OF ACCIDENT, ON DISTRICT REGISTER	00001-99999	
6	**(AR) ACCI-DENT_ DATE	DATE OF ACCIDENT	DATE (BRITISH) 9999999999 FOR UN-KNOWN	
7	**(AR) ACCI-DENT_ DAY	DAY OF ACCIDENT	1 2 3 4 5 6 7	SUNDAY MONDAY TUESDAY WEDNESDAY THURSDAY FRIDAY SATURDAY
8	**(AR) ACCI-DENT_ TIME	TIME OF ACCIDENT	TIME (HOUR-MINUTES) 9999 FOR UNKNOWN	
9	**(AR) NO_ VEHICLES	NUMBER OF VEHICLES INVOLVED IN ACCIDENT	01-99	
10	**(AR) NO_ INJURED	NUMBER OF CASUALTIES INVOLVED IN ACCIDENT		
11	**(AR) NAMES_ EX-CHANGED_IND	EXCHANGE OF NAMES/ ADDRESSES BETWEEN IN-VOLVED PERSONS	1 2	YES NO
12	**(AR) POLICE_ IND	POLICE VISITED THE ACCIDENT SCENE	1 2	YES NO

13	**(AR) ABAN-DON_ IND	INVOLVED PERSONS LEFT SCENE	1 2	YES NO
14	**(AR) PHO-TOS_IND	PHOTOS OF ACCIDENT SCE-NE TAKEN	1 2	YES NO
15	**(AR) STRIKE_LEAVE _IND	HIT & RUN ACCIDENT	1 2	YES NO
16	**(AR) POLICE_ STATION_ ACCIDENT_ NO	CONSECUTIVE NUMBER OF ACCIDENT ON POLICE STA-TION REGISTER	00001-99999	
17	**(AR) FAC-TOR_ A	APPARENT CONTRIBU-TING FACTOR 1	 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16	**HUMAN** ALCOHOL INVOLVEMENT DRUGS (ILLEGAL) PRESCRIPTION MEDICATION SUDDEN ILLNESS LOST CONSIOUSNESS FELL ASLEEP PHYSICAL DISABILITY DRIVER INEXPERIENCE UNSAFE SPEED FAILURE TO KEEP TO NEAR SIDE FAILURE TO KEEP TO PROPER TRAFFIC LANE LANE CHANGING (IMPROPERLY) OVERTAKING IMPROPERLY ON NEAR SIDE OVERTAKING IMPROPERLY ON OFF-SIDE CUTTING IN FAILURE TO STOP/ALLOW PEDESTRIAN CROSSING FAILURE TO GIVE RIGHT-OF-WAY TURNING LEFT WITHOUT CARE TURNING RIGHT WITHOUT CARE MAKING U TURN BACKING UNSAFELY

				TRAFFIC SIGN DISREGARDED
			17	TRAFFIC SIGNALS DISREGARDED
				POLICE SIGNAL DISREGARDED
			18	CROSSING WITHOUT CARE AT UNCONTROLLED JUNCTION
			19	FAILURE TO SIGNAL PROPERLY
			20	PULLING OUT FROM NEAR SIDE
				PULLING OUT FROM OFF-SIDE
			21	DRIVER INATTENTION/ DRIVING WITHOUT CARE
			22	FOLLOWING TOO CLOSELY
			23	STOPPING SUDDENLY
			24	SWERVING/ RUNNING OFF THE ROAD OUT OF CONTROL
				DAZZLED BY LIGHTS OF OTHER VEHICLE
			25	DRIVER OPENING SIDEDOOR
				OTHER ERROR ON BEHALF OF DRIVER
			26	DRIVER HAMPERED BY PASSENGER, ANIMAL, OR LUGGAGE
			27	PASSENGER OPENING SIDEDOOR
			28	BOARDING OR ALIGHTING BUS WITHOUT CARE
			29	OTHER ERROR ON BEHALF OF PASSENGER
				PEDESTRIAN CROSSING WITHOUT DUE CARE
				PEDESTRIAN IMPROPERLY USING PEDESTRIAN CROSSING
			30	OTHER ERROR ON BEHALF OF PEDESTRIAN
				VEHICLE
			31	BRAKES DEFECTIVE
			32	HEADLIGHTS DEFECTIVE
			33	REAR LIGHTS DEFECTIVE
				OTHER LIGHTING DEFECTIVE
			34	STEERING FAILURE
				TYRE/WHEEL FAILURE
			35	TOW HITCH DEFECTIVE
			36	OVERSIZED VEHICLE
			37	OVERLOADED VEHICLE
			38	OTHER VEHICULAR FACTOR
				ENVIRONMENTAL
			39	LANE MARKING IMPRORER / INADEQUATE

			40	TRAFFIC SIGNS IMPROPER /INADEQUATE
			41	TRAFFIC SIGNLAS IMPROPER/ NOT-WORKING
				OBSTRUCTIONS /DEBRIS ON ROAD
			42	PAVEMENT DEFECTIVE
			43	PAVEMENT SLIPPERY (CONSTRUCTION)
				SHOULDERS DEFECTIVE
			44	GLARE (ROAD SURFACE)
			45	VIEW OBSTRUCTED /LIMITED
			46	PAVEMENT SLIPPERY (WEATHER)
			47	STRONG WIND
			48	SUN GLARING
			49	ANIMAL ACTION
				OTHER ENVIROMENTAL FACTOR
			50	
			51	
			52	
			53	
			54	
			55	
			56	
			57	
			58	
			59	

			60	
			61	
			62	
			63	
			64	
			65	
			66	
18	**(AR) PEDES-TRIAN_ AC-TION	PEDESTRIAN ACTION	01	CROSSING ROAD MASKED BY STATIONARY VEHICLE
			02	CROSSING ROAD NOT MASKED BY STATIONARY VE-HICLE CROSSING ROAD MASKED BY MOVING VEHICLE
			03	CROSSING ON PEDESTRIAN CROSSING WALKING ON ROAD, FACING TRAFFIC - NO FOOTPATH
			04	WALKING ON ROAD, FACING TRAFFIC WITH FOOT-PATH
			05	WALKING ON ROAD, BACK TO TRAFFIC- NO FOOT-PATH
			06	WALKING ON ROAD, BACK TO TRAFFIC- WITH FOOT-PATH
			07	STANDING OR PLAYING ON ROAD ON FOOTPATH, REFUGE OR OTHER OFF-ROAD PLACE
			08	OTHER ACTION IN ROADWAY
			09	
			10	
			11	
19	**(AR) MAIN_ROAD	MAIN ROAD NUMBER		e. g: A0001
20	**(AR) RESI-DENCE_ AREA	BUILT-UP AREA	1	YES
			2	NO

21	**(AR) FAC-TOR_B	APPARENT CONTRIBU-TING FACTOR 2	1-66	SEE FIELD NO.17
22	**(AR) KM	DISTANCE FROM THE BE-GINNING OF ROAD IN KMS, OR FROM 1ST LOCATION (FIELD NO. 27)	1-998 999	DISTANCE FROM BEGINNING OF ROAD IN KMS UNKNOWN
23	**(AR) MTR	DISTANCE IN METRES FROM PREVIOUS KILOMETRE POST, IN METRES	1-999	DISTANCE FROM PREVIOUS KILOMETRE POST, IN METRES
24	**(AR) FAC-TOR_C	APPARENT CONTRIBU-TING FACTOR 3	1-66	SEE FIELD NO.17
25	**(AR) TRAF-FIC_ CONTROL	MEANS OF TRAFFIC CONTROL	1 2 3 4 5 6 7 8 9	NONE POLICE STOP SIGN GIVE WAY SIGN ROUNDABOUT TRAFFIC SIGNALS TRAFFIC SIGNALS & POLICE FLASHING TRAFFIC SIGNALS TRAFFIC SIGNALS OUT OF ORDER
26	**(AR) ROAD_WITDH	ROAD WIDTH		e. g : 07.50 metres
27	**(AR) POINT_ A	1ST LOCATION CODE		e. g: M0104
28	**(AR) POINT_ B	2ND LOCATION CODE		e. g: N0105
29	**(AR) DIREC-TION	DIRECTION OF TRAVEL (ON MOTORWAYS)	0,1,2,9	0=NOT APPLICABLE 1=DIRECTION 1 2=DIRECTION 2 9=UNKNOWN

30	**(AR) BREAK_ LANE_WIDTH	WIDTH OF FOOTWAY/ SHOULDER		e. g: 02.50 metres
31	**(AR) DIA-GRAM_ CODE	COLLISION DIAGRAM	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17	NOSE TO TAIL OVERTAKING FROM RIGHT SIDE OVERTAKING FROM LEFT SIDE FRONTAL (FROM OPPOSITE DIRECTIONS) SIDE COLLISION (FROM OPPOSSITE DIRECTIONS) ON STATIONARY VEHICLE REVERSING ON STATIONARY VEHICLE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE FROM SIDE

			18	FROM SIDE
			19	FROM SIDE
			20	FROM SIDE
			21	FROM SIDE
			22	RUN-OFF TO LEFT
			23	RUN-OFF TO RIGHT
			24	ON FIXED OBJECT
			25	OTHER
			26	WALKING OR STANDING ON ROAD
			27	CROSSING FROM LEFT SIDE CROSSING FROM RIGHT SIDE
			28	CROSSING FROM LEFT SIDE BEHIND PARKED VEHICLE
			29	CROSSING FROM RIGHT SIDE BEHIND PARKED VEHI-CLE
			30	CROSSING AT ROAD JUNCTION
			31	

			32 33 34 35 36 37 38 39 40	CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING AT ROAD JUNCTION CROSSING DIAGONALLY OTHER
32	**(AR) FAC-TOR_D	APPARENT CONTRIBU-TING FACTOR 4	1-66	SEE FIELD NO. 17
33	**(AR) CON-JUCTION_ TYPE	JUNCTION TYPE	1 2	INTERSECTION OF 2 OR MORE ROADS ′ T′ JUNCTION

			3	STAGGERED JUNCTION
			4	′Y′ JUNCTION
			5	ROUNDABOUT
			6	MOTORWAY SLIP ROAD
			7	OTHER
			8	NOT ON JUNCTION
34	**(AR) ROUTE_ PERMITTED	ALLOWED TRAFFIC MOVE-MENTS	1	SINGLE (1-WAY)
			2	DOUBLE (2-WAY)
			3	BOTH ABOVE (ONLY AT JUNCTIONS)
35	**(AR) BARRI-ER	TYPE OF ROAD SEPARATION	1	NONE
			2	BROKEN SINGLE LINE
			3	CONTINUOUS LINE
			4	DOUBLE CONTINUOUS LINE
			5	GHOST ISLAND
			6	ISLAND PHYSICAL BARRIER
			7	ISLAND WITHOUT PHYSICAL BARRIER
			8	COMBINATION OF MORE THAN 1 OF THE ABOVE
36	**(AR) CON-STRICTION	ROAD NARROWING	1	NONE
			2	ONE-WAY BRIDGE
			3	TWO-WAY BRIDGE
			4	OTHER

37	**(AR) PAVE-MENT_ TYPE	ROAD SURFACE TYPE	1 2 3 4	ASPHALT PAVEMENT STONE DIRT OTHER
38	**(AR) BREAK_ LANE	TYPE OF SHOULDER	1 2 3 4	FOOTWAY PAVED SHOULDER UNPAVED SHOULDER OTHER
39	**(AR) SPEED_ LIMIT	SPEED LIMIT		
40	**(AR) ROAD_ WORK	ROAD WORKS	1 2	YES NO
41	**(AR) BUS_ STOP	BUS STOP	1 2	YES NO
42	**(AR) PEDES-TRIAN_ CROSS-ING	PEDESTRIAN CROSSING FACILITIES	1 2 3 4 5 6 7	NONE ZEBRA CROSSING PEDESTRIAN TRAFFIC SIGNAL CROSSING PEDESTRIAN PUSH BUTTON PEDESTRIAN PELICAN CROSSING POLICE CONTROLLED CROSSING OTHER
43	**(AR) LIGHT-ING	LIGHT CONDITIONS	1 2 3 4 5 6	DAYLIGHT DAWN DUSK NIGHT-STREET LIT NIGHT-STREET UNLIT UNKNOWN

44	**(AR) FIRST_EVENT_ PLACE	LOCATION OF 1ST EVENT	1	ON ROAD
			2	OFF ROAD
			3	UNKNOWN
45	**(AR) ROAD_DESCR	ROAD DESCRIPTION	1	STRAIGHT & FLAT
			2	STRAIGHT & GRADE
			3	STRAIGHT & HILL CREST
			4	CURVED & FLAT
			5	CURVE & GRADE
			6	CURVED & HILL CREST
46	**(AR) PAVE- MENT_ STATUS	ROAD SURFACE CONDITION	1	DRY
			2	WET
			3	MUDDY
			4	SNOW/ICE
			5	SLUSH
			6	OTHER
47	**(AR) WEATHER	WEATHER	1	CLEAR/FINE
			2	RAIN/HAIL
			3	FOG
			4	SNOW
			5	OTHER

48	**(AR) FIRST_EVENT	TYPE OF ACCIDENT (1ST EVENT/ COLLISION)		**COLLISION WITH OTHER VEHICLE**
			1	NOSE TO TAIL
			2	SIDE TO SIDE
			3	HEAD ON
			4	ANGLE
			5	STATIONARY MOTOR VEHICLE
			6	BICYCLE
				COLLISION WITH OTHER MOVING OBJECT
			7	PEDESTRIAN
			8	ANIMAL
			9	OTHER OBJECT (NOT FIXED)
				COLLISION WITH FIXED OBJECT
			10	LIGHT SUPPORT/UTILITY POLE
			11	GUARD RAIL
			12	MEDIAN/ BARRIER
			13	TRAFFIC ISLAND
			14	PAVEMENT (KERBING)
			15	SIGN POST
			16	BRIDGE STRUCTURE
			17	CULVERT/ HEAD WALL
			18	EMBANKMENT /DITCH
			19	CRASH CUSHION
			20	BUILDING /WALL
			21	TREE
			22	OTHER FIXED

				OBJECT
				NO COLLISION
			23	OVERTURNED (IN ROAD)
			24	RAN OFF ROADWAY ONLY
			25	CROSSED MEDIAN
			26	OTHER
49	**(AR) POLICE_ OFFICER_ GRADE	POLICE INVESTI-GATOR (RANK)	**AA**	CONSTABLE
			EE	SERGEANT
			GG	OFFICER
50	**(AR) POLICE_ OFFICER_NO	POLICE INVESTI-GATOR NUMBER	1-9999 **(0000)***	
51	**(AR) POLICE_ CALLED	POLICE NOTIFICATION TIME	0000-2359	HOUR AND MINUTES
52	**(AR) POLICE_ ARRIVED	POLICE ARRIVAL TIME	0000-2359	HOUR AND MINUTES
53	**(AR) PO-LICE_TIME	TIME FOR POLICE TO AR-RIVE	0001-9959	HOURS AND MINUTES
54	**(AR) AMBU-LANCE_ CALLED_BY	NOTIFIED BY	(-)	NOT NOTIFIED
			1	PERSON INVOLVED
			2	PASSER - BY
			3	POLICEMAN
			9	UNKNOWN
55	**(AR) AMBU-LANCE_ CALLED	AMBULANCE NOTIFICATION TIME	0000-2359	HOUR AND MINUTES
56	**(AR) AMBU-LANCE_ AR-RIVED	AMBULANCE ARRIVAL TIME	0000-2359	HOUR AND MINUTES

57	**(AR) AMBU-LANCE_ TIME	TIME FOR AMBULANCE TO ARRIVE	0001-9959 BLANC 9999	HOURS AND MINUTES NO AMBULANCE AT SCENE UNKNOWN

CARD NO. 2: VEHICLE DATA

No.	FIELD NAME	FIELD DESCRIPTION	VALUES	DESCRIPTION OF VALUES
1	AREA_CODE	CODE FOR ACCIDENT LOCATION (URBAN OR RURAL)	T R	TOWN RURAL
2	ACCIDENT_TYPE	ACCIDENT SEVERITY	1 2 3 4	FATAL SERIOUS INJURY SLIGHT INJURY DAMAGE
3	POLICE_ DISTRICT	CODE NUMBER FOR TOWN OR DISTRICT WHERE THE ACCIDENT OCCURED	SEE FIELD NO.3 CARD NO.1	SEE FIELD NO. 3 CARD NO.1
4	POLICE_STATION	CODE NUMBER OF POLICE STATION WHICH INVESTIGATED THE ACCIDENT	SEE FIELD NO.4 CARD NO.1	SEE FIELD NO.4 CARD NO.1
5	DISTRICT_ ACCIDENT_NO	CONSECUTIVE NUMBER OF ACCIDENT, ON DISTRICT REGISTER	SEE FIELD NO.5 CARD NO.1	SEE FIELD NO.5 CARD NO.1
6	**(EI) VEHICLE_SEQ	VEHICLE CONSECUTIVE NO.	01-98 99	CONSECUTIVE NUMBER UNKNOWN
7	**(EV) DRIVER_ID_ NO	DRIVER IDENTITY CARD NO.	000001-999998 *(0000000000) 999999	IDENTITY CARD NO. UNKNOWN
8	**(EI) DRIVER_AGE	DRIVER AGE	01-98 99	AGE UNKNOWN
9	**(EI) DRIVER_GENDER	DRIVER SEX	1 2	MALE FEMALE
10	(EV) DRIVER_LICENCE_TYPE	TYPE OF DRIVER´S LICENCE	1 2 3 9	LEARNER´S REGULAR NO LICENCE UNKNOWN

11	**(EV) DRIVER_LICENCE_NO	DRIVER´S LICENCE NUMBER	000001-999998 *(0000000000) 999999	LICENCE NO. UNKNOWN
12	**(EV) DRIVER_LICENCE_EXP	EXPIRY DATE OF LICENCE	BRITISH DATE	
13	**(EV) INSURANCE_ COMPANY	INSURANCE COMPANY	01-32 49 50 --- 99	e.g.: AEGIS INSUR.COMP. REPUBLIC OF CYPRUS OTHER NO INSURANCE UNKNOWN
14	**(EV) INSURANCE_NO	INSURANCE CERTIFICATE NO.	CHARACTERS AND NUMBERS*(0000000000)	
15	**(EV) INSURANCE_ ISSUE_DATE	DATE OF ISSUE OF INSURANCE CERTIFICATE	BRITISH DATE	
16	**(EV) INSURANCE_ EXPIRY_DATE	DATE OF EXPIRY OF INSURANCE CERTIFICATE	BRITISH DATE	
17	**(EV) MANUFACTURER	VEHICLE MAKE	0000 0001-9999	UNKNOWN MAKE CODE NUMBER e.g.: 0211 FOR AUDI
18	**(EV) MANUFACTURE_ YEAR	VEHICLE YEAR OF CONSTRUCTION	01-9999 0000	YEAR UNKNOWN
19	**(EV) CAPACITY_CC	VEHICLE CAPACITY CC	00001-99998 99999	CAPACITY UNKNOWN
20	**(EV) REGISTRATION_ NO	VEHICLE REGISTRATION NO.	e.g.: AAA111 *(0000000000000000)	
21	**(EV) VEHICLE_TYPE	VEHICLE TYPE	01 02 03	BICYCLE MOPED UP TO 49CC MOTORCYCLE (50 CC AND OVER)

			04	TAXI
			05	RENTAL CAR
			06	OTHER CAR
			07	MINI-BUS
			08	BUS
			09	LIGHT COMMERCIAL VEHICLE UP TO 2 TONS (SINGLE REAR TYRES)
			10	VAN UP TO 2 TONS (SINGLE REAR TYRES)
			11	MEDIUM COMMERCIAL VEHICLE – 2 AXLES (DOUBLE REAR TYRES)
			12	HEAVY COMMERCIAL VEHICLE – MORE THAN 2 AXLES
			13	ARTICULATED HEAVY COMMERCIAL VEHICLE
			14	AGRICULTURAL TRACTOR
			15	OTHER MOTOR VEHICLE
			16	ANIMAL OR CARRIAGE
			17	UNKNOWN
22	**(EV) DAMAGE	POSITION OF 1ST IMPACT ON VEHICLE	01	FRONT
			02	RIGHT FRONT WING
			03	RIGHT DOOR
			04	RIGHT BACK WING
			05	REAR
			06	LEFT REAR WING
			07	LEFT DOOR/S
			08	LEFT FRONT WING
			09	ROOF
			10	UNDERSIDE
			11	NONE
23	**(EV) SECOND_ EVENT	TYPE OF ACCIDENT (2nd EVENT)	01-26	SEE FIELD NO.48 CARD NO.1
24	**(EV) LICENCE_ IND	CIRCULATION LICENCE NUMBER	**1**	YES
			2	NO
			9	**UNKNOWN**

25	**(EV) APPROPRIATE_IND	ROAD WORTHINES CERTIFICATE	1 2 9	YES NO UNKNOWN
26	**(EV) ACTION_ BEFORE_ ACCIDENT	PRE-ACCIDENT VEHICLE ACTION	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19	GOING STRAIGHT AHEAD MAKING RIGHT TURN MAKING LEFT TURN MAKING U TURN STARTING FROM PARKING STARTING IN TRAFFIC SLOWING OR STOPPING STOPPED IN TRAFFIC ENTERING PARKED POSITION PARKED AVOIDING OBJECT/POTHOLE IN ROAD AVOIDING PEDESTRIAN IN ROAD AVOIDING VEHICLE IN ROAD CHANGING LANES OVERTAKING MERGING IN MOTORWAY (ACCELERATION LANE) DIVERGING IN MOTORWAY (DELERATION LANE) BACKING OTHER
27	**(AR) ACCIDENT_ YEAR	YEAR OF ACCIDENT	e.g. 2004	

CARD NO. 3 PERSONS INVOLVED DATA

No.	FIELD NAME	FIELD DESCRIPTION	VALUES	DESCRIPTION OF VALUES
1	AREA_CODE	CODE FOR ACCIDENT LOCATION (URBAN OR RURAL)	T R	TOWN RURAL
2	ACCIDENT_ TYPE	ACCIDENT SEVERITY	1 2 3 4	FATAL SERIOUS INJURY SLIGHT INJURY DAMAGE
3	POLICE_ DISTRICT	CODE NUMBER FOR TOWN OR DISTRICT WHERE THE ACCIDENT OCCURED		SEE FIELD NO. 3 CARD NO. 1
4	POLICE_ STATION	CODE NUMBER OF POLICE STATION WHICH INVESTIGATED THE ACCIDENT		SEE FIELD NO. 4 CARD NO. 1
5	DISTRICT_ ACCIDENT_NO	CONSE-CUTIVE NUMBER OF ACCIDENT, ON DISTRICT REGISTER		SEE FIELD NO. 5 CARD NO. 1
6	**(EI) VEHICLE_SEQ	VEHICLE OCCUPIED BY PERSON INVOLVED	1 2 0	VEHICLE NO.1 VEHICLE NO.2 NO VEHICLE
7	**(EI) POSITION_IN_ VEHICLE	POSITION IN/ON VEHICLE OF PERSON INVOLVED	1 2-10 11 12	DRIVER SEATED PASSENGER STANDING PASSENGERS UNKNOWN
8	**(EI) PROTECTIVE_MEASURES	SAFETY EQUIPMENT USED BY PERSON INVOLVED EJECTION FROM VEHICLE OF PERSON INVOLVED	1 2 3 4 5	NO RESTRAINT USED SEAT BELT CHILD RESTRAINT HELMET UNKNOWN

9	**(EI) EJEC-TION		1	NOT EJECTED
			2	PARTIALLY EJECTED
			3	EJECTED
			4	UNKNOWN
10	**(EI) NATION-ALITTY	NATIONALITY OF PERSON INVOLVED	1	CYPRIOT
			2	TOURIST
			3	OTHER
11	** (EI) AGE	AGE OF PERSON INVOLVED		
12	** (EI) GENDER	SEX OF PERSON INVOLVED	1	MALE
			2	FEMALE
			3	UNKNOWN
13	** (EI) CORPS	SERVICE PERSONELL	1	POLICE
			2	NATIONAL GUARD
			3	BRITISH BASES
			4	UNFICYP
			5	ELDYK
			6	OTHER
14	**(EI) ALCO-HOL	ALCOHOL AND DRUGS	1	NEITHER INVOLVED
			2	ALCOHOL POSITIVE
			3	FAILED TO PROVIDE SAMPLE
			4	DRUGS POSITIVE
			5	TEST NOT DEMANDED
			6	UNKNOWN
15	** (EI) ROLE_ IN_ ACCIDENT	TYPE OF PERSON INVOLVED	01	PEDESTRIAN
			02	PEDAL CYCLIST
			03	PASSENGER OF PEDAL CYCLE
			04	MOPED RIDER (UNDER 50CC)
			05	MOPED PASSENGER
				DRIVER OF MOTORCYCLE OR TRICYCLE

			06	PASSENGER M/CYCLE & OR TRICYCLE **DRIVER OF VEHICLE WITH UP TO 8 PASSENGERS**
			07	**PASSENGER OF VEHICLE WITH UP TO 8 PASSEN-GERS**
			08	DRIVER (OTHER VEHICLES) PASSENGER (OTHER)
			09	ANIMAL RIDER
			10	TAXI DRIVER TAXI PASSENGER
			11	RENTAL CAR DRIVER
			12	RENTAL CAR PASSENGER
			13	CAR DRIVER
			14	CAR PASSENGER
			15	MINIBUS DRIVER
			16	MINIBUS PASSENGER
			17	BUS DRIVER
			18	BUS PASSENGER
			19	LIGHT COMMECIAL VEHICLE UP TO 2 TONS DRIVER
			20	LIGHT COMMECIAL VEHICLE UP TO 2 TONS PASSEN-GER
			21	VAN DRIVER
			22	VAN PASSENGER
			23	MEDIUM COMMERCIAL VEHICLE (2 AXLES) DRIVER MEDIUM COMMERCIAL VEHICLE (2 AXLES) PASSEN-GER
			24	DRIVER- HEAVY COMMERCIAL VEHICLE (MORE THAN 2 AXLES)
			25	PASSENGER- HEAVY COMMERCIAL VEHICLE (MORE THAN 2 AXLES)
			26	DRIVER- ARTICULATED HEAVY COMMERCIAL VEHI-CLE
			27	PASSENGER- ARTICULATED HEAVY COMMERCIAL VEHICLE
			28	AGRICLTURAL TRACTOR DRIVER AGRICLTURAL TRACTOR PASSENGER CARRIAGE DRIVER
			29	CARRIAGE PASSENGER

				FATAL
			30	SERIOUS INJURY
				SLIGHT INJURY
				NO INJURY
			31	MINISTRY OF HEALTH AMBULANCE
				FIRE BRIGADE AMBULANCE
			32	PRIVATE AMBULANCE
				OTHER AMBULANCE
			33	POLICE VEHICLE
				OTHER VEHICLE
			34	UNKNOWN
			35	
				GOVERNMENT/ PUBLIC
			36	PRIVATE CLINIC
16	** (EI) INJURY_ SEVERITY	INJURY TYPE	1	NONE
			2	
			3	
			4	
17	**(EI) TRANS-FER_ TO_ HOS-PITAL	MEANS OF CONVEY-ANCE TO…	1	
			2	
			3	
			4	
			5	
			6	
			7	
18	**(EI) HOSPI-TAL	HOSPITAL	1	
			2	
			3	