**Student: Daniel Pereira Peres**

**Student number: 19215101**

**Email: daniel.peres2@mail.dcu.ie**

**Program of Study: MCM1 - Artificial Intelligence**

**Discipline: CA683 - Data Analytics and Data Mining**

**Submission date: 19/04/2020**

Predicting accident severity using ML and features we know beforehand

**Abstract**

Understanding traffic accidents is a critical topic, necessary to reduce one of the significant causes of death and injuries worldwide. Many studies have been made trying to map accidents to its root causes in order to take cautions to reduce the number of fatalities. Differently of other studies, the aim is being able to predict how severe could be an accident given the characteristics a driver could know beforehand, given the driver, vehicle profile and weather conditions only.

**Introduction**

Traffic accidents are a major concern of governments world-wide because it is one of the lead causes of death and injuries, especially for young people. There are many studies on this subject, and they were used as a guide for this work. Applying machine learning to an UK dataset with data of 1.6MI accidents recorded during ten years, it was possible to predict the accident severity on a 97% accuracy for fatal, 81% accuracy for serious and 85% accuracy for slight ones, with f-1 of 0.98, 0.82 and 0.82 respectively. We got an even better result using an extended number of features, as the road type, if there were carriageway hazards on the road, etc., which is still valid but is not the answer for our main question because we do not know them in advance.

The final solution consists of using SMOTE algorithm to create synthetic data to solve the imbalanced classes problem and Random Forests as a regressor.

**Literature Review**

Many studies have been made to predict road traffic accidents (RTA) and its severity, focusing on predict or establish different factors that may influence injury severity, either using machine learning or not.

The first mention is a paper from Western Michigan University College of Civil Engineering [1], which was very inspiring for this work. They pointed out that the vehicle type, weather and road conditions are essential to predict the accident severity. However, the goal of their studies are to measure the performance of different machine learning algorithms to predict accident severity rather than its causes. They used 271 thousand rows data from 2010 to 2016 car crashes occurred on Michigan, combined to over-sampling SMOTE strategy to create new synthetic data to handle the imbalance between the slight and serious\fatal accidents. It was tested many classifiers, like Random Forests, Logistic Regression, Naïve Bayesian Classifier and Ada Boost classification tree. The best accuracy of 75% was achieved with random forests and F1 score of 0.707.

Another interesting paper [2], it was used a deep learning model using Recurrent Neural Network (RNN) to predict injuries caused by car accidents. The accuracy achieved by their best model was 71%, and they concluded that vehicle failure and objects on the road had a major influence on severity.

On the paper from Diponegoro University [3], they used features as the day of the week, type of the road, type of vehicle, weather, road surface among others, to predict the accident severity with Naïve Bayes probability. They conclude that at noon, is the period where there is a higher probability (54%) of fatal accidents. If the vehicle is a truck, the probability of major injury is 55%, and the fatal accident is 72%. If the sex of the driver is male, the probability of major injury or fatal is almost 100%. However, the accuracy of their model is only 39.49%.

The last of the machine learning references [4], they used artificial neural network (ANN) to predict the injury severity of traffic accidents in Abu Dhabi, over a six year period, from 2008 to 2013. They had 48 features at the time of the accident. They extracted the best 16 features to work with. Their accuracy was of 74.6%.

Other literature than machine learning related was used, in order to understand the patterns found on the data and to prove their accuracy. For example, that males are more reckless while driving than females [5], stress and fatigue has a high correlation with car accidents [6], there is a clear correlation between alcohol and car accidents [7], the rural areas accidents are almost three times more likely to end up with a fatal accident than the urban areas [8], and lastly, an RSA leaflet [9] that suggests many people think is acceptable use smartphones on slow-moving traffic and traffic lights.

They are all properly referenced on the EDA section.

**Data Mining Methodology**

The model adopted is CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining.

The steps are split in this document as Data understanding, Data cleaning and preparation, Explanatory data analysis (on the main topics), Modelling and Conclusions.

**Data Understanding**

The UK government keep a record of road accidents that are reported on its territory using a form called Stats19. Once it comes from police reports, it does not include minor accidents. This data is publicly available on data.gov.uk, or in Kaggle.com.

The dataset contains information from 2005 to 2014 for about 1.6MI accidents and more than 3MI vehicles that were involved. This data is originally on two separate files which can be joined using the Accident\_Index identifier and a lookup excel file. It is necessary to do the lookup on the excel file because the categorical features on the main files have an integer encoding, which is not suitable for our analysis and machine learning. This lookup was made using an external tool and exported again, so our input files already have the categorical values – no need for lookup again.

The accident severity, which is our goal to predict, is very imbalanced. We have 38083 entries for fatal accidents, 373281 for serious accidents and most of them are slight, with 2593061 records.

Among the 53 variables, we have information that was split on some categories:

* Location (Latitude, Longitude, Location OSGR)
* Driver (Age, Journey Purpose, Sex)
* Date and time
* Kind of manoeuvre the vehicle was executing
* Local (Urban or Rural, Area type, …)
* Road (Speed limit, Road Type, Road Surface Conditions, …)
* Vehicle (Type, Propulsion, Age of vehicle)
* Weather conditions (Light conditions, raining, snowing, ...)

Notice that not all variables were used for the machine learning process. It is excluded all variables that happened after the accident and would bias our results, such as: Did police office attend to the scene, number of casualties, number of vehicles involved and others.

**Data Cleaning and Preparation**

Data range check

The first step was to look for the data range. It was found some outliers present, such as an accident which had 67 vehicles involved and 93 casualties, one vehicle with 111 years, one driver with 100 years and drivers with less than 15 years old. These facts were analysed and, the drivers under 15 years were not driving a car, but another type of vehicle, as pedal cycle. All this information’s are accurate and were kept. Some variables have an out-of-range value, like the age of driver and vehicle as -1. After reading the dataset documentation, it was found that when they did not have the data, it was set to -1. These records were dropped, once they were not a significative number of rows and we could not infer this data from the other variables. It will be discussed in more details on the data imputation section.

Missing values check

The second step was to look for missing values. Some variables had over 30% of missing values, including second road class and junction control. The second road class feature was dropped, once we do not have enough information to infer its value and was shown not very important. The junction control, we could have made an imputation using a regression over the junction detail, but it would cause big collinearity on these variables, so this feature was also dropped.

Data imputation

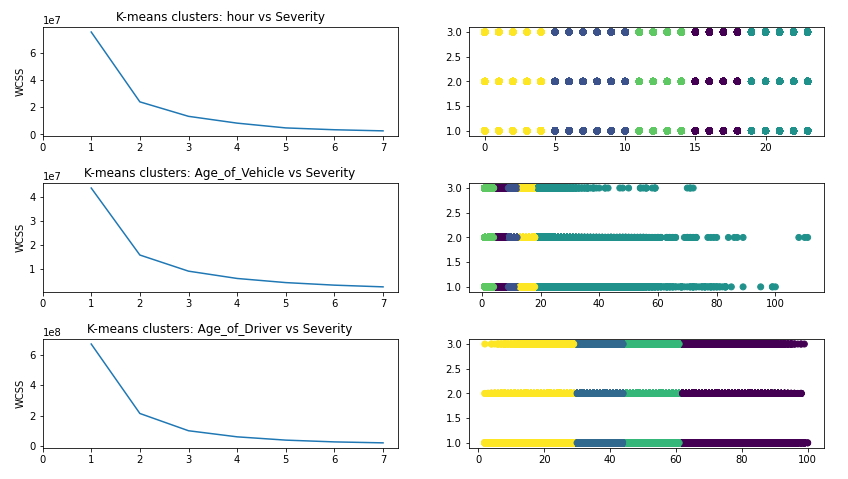
The engine capacity is an essential feature and could not be dropped, even with has 26% of values missing. It was made a data imputation using the vehicle type mean as value.

Similarly, the propulsion type has almost 25% of missing value, and it is crucial to be dropped. It was made a regression model using the engine capacity, age of vehicle and vehicle type to input this information with 90% of accuracy, but it could not be used because when the propulsion type is missing, the engine capacity and age of vehicle are also missing. Another attempt was made using only the vehicle type, but the accuracy was only around 60%. Besides our efforts to input this information, we had to drop the rows with missing values.

Feature Engineering

Analysing the data correlation was identified that the OSGR Locations are highly collinear with latitude and longitude, so they were dropped.

Some variables were very spread, as vehicle age, driver age and engine capacity. It was used K-Means clustering using the feature data and the accident severity to suggest us the best way to split the information into bins.



As suggested, the vehicle age was split into five bins. We added a new bin to the driver’s age, from 1-16, because it is the legal age to drive on the UK. Similarly, we applied the same logic to the hours of the day: 0-5, 10 to 14, 14-18 and 18-23 hours. Despite the fact we could have split the engine capacity into bins, we decided to let it as the original continuous value.

It was added a feature for the day of the week in the model, for the weekends and the weekends night, which is the time when we have more severe accidents. The weekends night feature was created using the period from Friday 6 p.m to Saturday at 6 a.m, and Saturday from 6.pm to Monday at 6 a.m.

The propulsion type was normalized, once we had thirteen possible values, and eleven of them had a small parcel of less than 0.1% of the dataset. So we changed in a way we have three possible values: Petrol, Heavy Oil, and others.

Explanatory Data Analysis (main features)

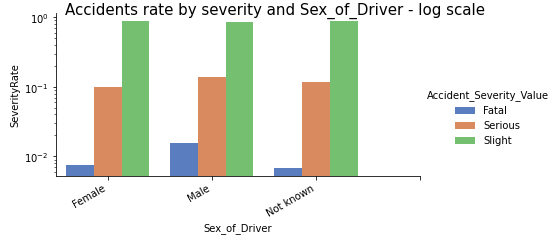
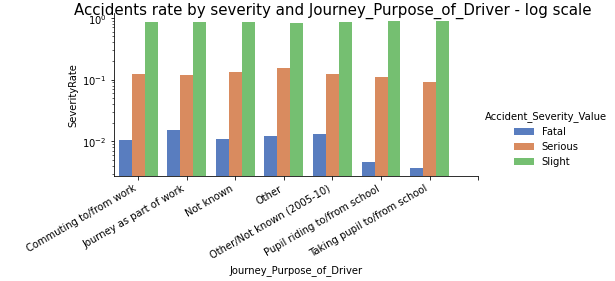
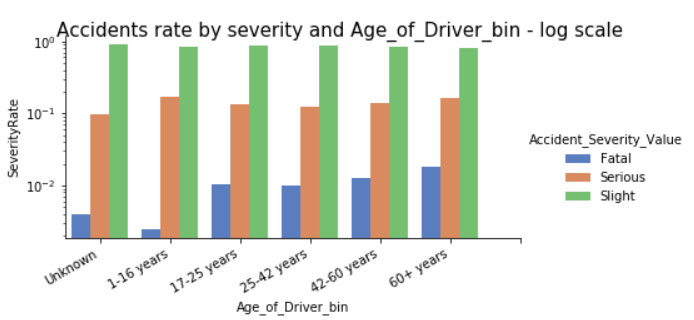
Using the correlation matrix for the continuous values, we can conclude there is an evident correlation between the speed limit and the number of casualties \ vehicles involved in the accident.

To study the accident’s severity offenders, we plotted every feature against the three levels of severity to compare their effects on each one. Once our severity is very imbalanced (only 0.01% are fatal), we are not using absolute numbers, but theirs rates on that feature. Additionally, our plots are on the logarithmical scale. It is the best way to visualize and compare the rates of the fatal class.

On regards to severity, the driver’s age follows a small crescent trend after 17 years old, the age when they start to drive vehicles. We can conclude that the older the driver is, more significant the changes to be involved in a fatal accident. It could be related to the fact that as older a person is, it became more vulnerable to injuries.

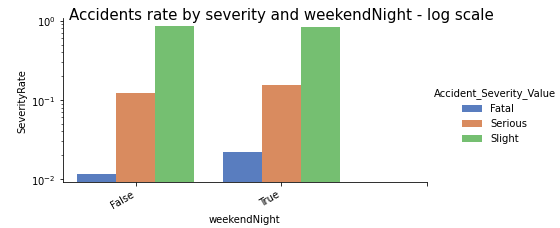
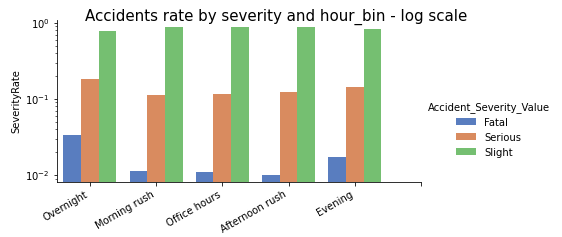
The Journey Purpose is quite impressive, the severity is higher when people are working than when they are riding to school or getting kids from school. It could be related to the different stress this situation cause **[6]**, and the fact people drive with more caution when kids are on board.

The last driver variable is sex, which has a high correlation with the accident severity. Males have the double of the chances to get involved in a fatal accident. There are plenty of studies showing that males are more reckless than females while driving **[5]**.

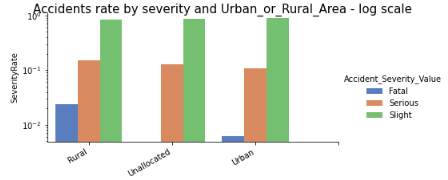


About the datetime features, we can highlight the hour, the weekday and the weekends night. As we can see in the first plot, overnight accidents as much more serious and fatal than any other period. This could be caused either by lack of illumination or fatigue, which is known as one of the main causes of accidents. **[6]**

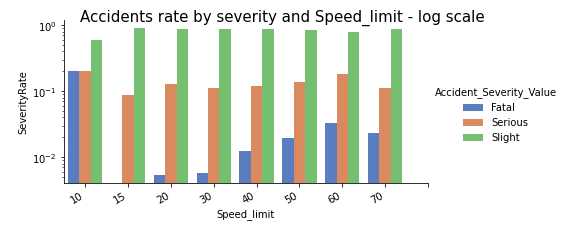
During the weekends we have fewer accidents, but they are more severe, especially during the weekend’s night: the 12 hours period of Friday and Saturday after 6 p.m. Our data has no information about alcohol consumption, but we can infer that during this period, there is a higher consumption of this substance, which is another widely known cause of accidents. **[7]**



The Urban areas correspond to 65% of all accidents reported, but in general, accidents on rural areas tend to result in much more severe injuries. This is supported by many studies realized. **[8]**



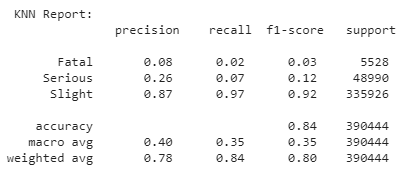
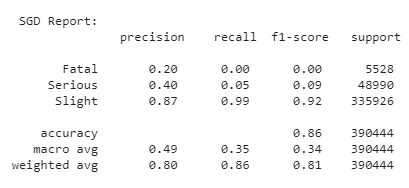
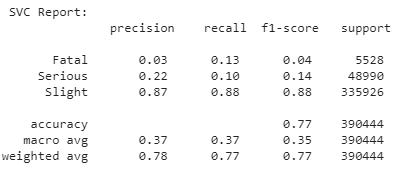
On the many features related to road conditions, as expected, the speed limit has considerable influence. Curiously, the accidents occurred at 10m.p.h has the same proportion of fatal and serious injuries, and it is also as big as the slight ones. This could be caused because, at this speed limits, people feel more comfortable to use their cell phones **[9]**, which is another widely known cause of accidents.



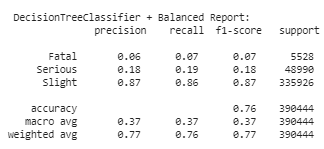
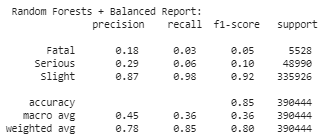
Modelling

Our target variable is the accident severity, which is labelled as Fatal, Serious and Slight. The variables that occurred after the accident were dropped to not bias our results, such as the number of casualties and the number of vehicles involved. On the categorical variables, it was applied one-hot-encoding. To make the prediction, we tried many regressors available on Scikit-learn and compared their results. To compare the results was measured their accuracy (number of right predictions) and the F-1 score, which takes on account the number of true positives and false positives. The PCA (Principal Component Analysis) usage was tested, but we did not go forward due to bad results, and the reduced number of variables on the final approach.

The SVC (Support Vector Classification), SGD (Stochastic Gradient Descent) and KNN (K-nearest neighbours) predictions performed very bad. They were more tended to classify the severity as Slight, probably due to the bias introduced for the imbalanced data classes.



It was also tried decision trees and random forest. On Scikit-learn, these methods have a particular property to deal with imbalanced data, the ‘class-weight’ parameter. It tries to adjust the weights inversely proportional to the class frequencies. It was run tests either with and without this option. The results were a little better than the previous methods but still unsatisfactory. Additionally, there was no benefit of using the ‘class-weight’ parameter.

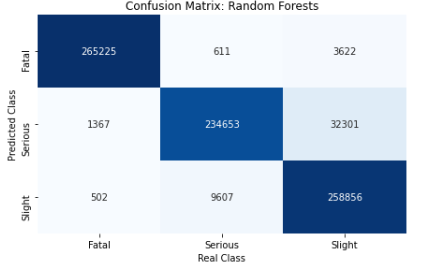
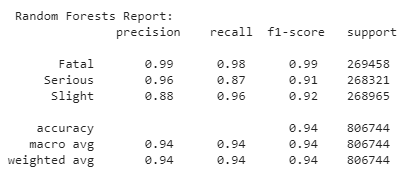


To overcome this problem, encouraged by the over-sampling used on a referenced paper **[1]**,it was used an over-sampling technique called SMOTE (Synthetic Minority Over-sampling Technique). **[10]**. SMOTE generates synthetic data for the classes with fewer rows, based on the nearest neighbours of existing data for those classes.

After the over-sampling, we run the decision trees and random forests methods again, because they had better performance on the previous attempt.

The results now are very good, with an accuracy of 99% and f1-score of 0.99 for the best class, and an accuracy of 88% and f1-score of 0.92 for the worst class. As expected, the random forest method had a better performance.

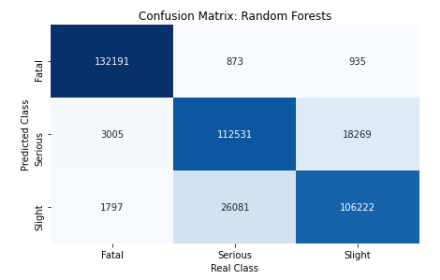
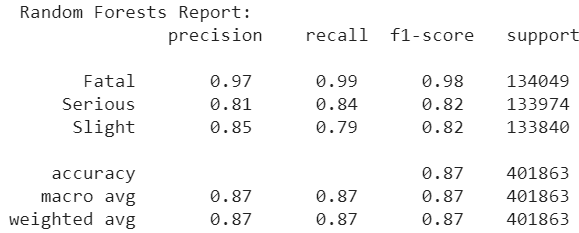
The most important features are engine capacity, urban or rural area, hour, age of the driver, sex of driver and speed limit.



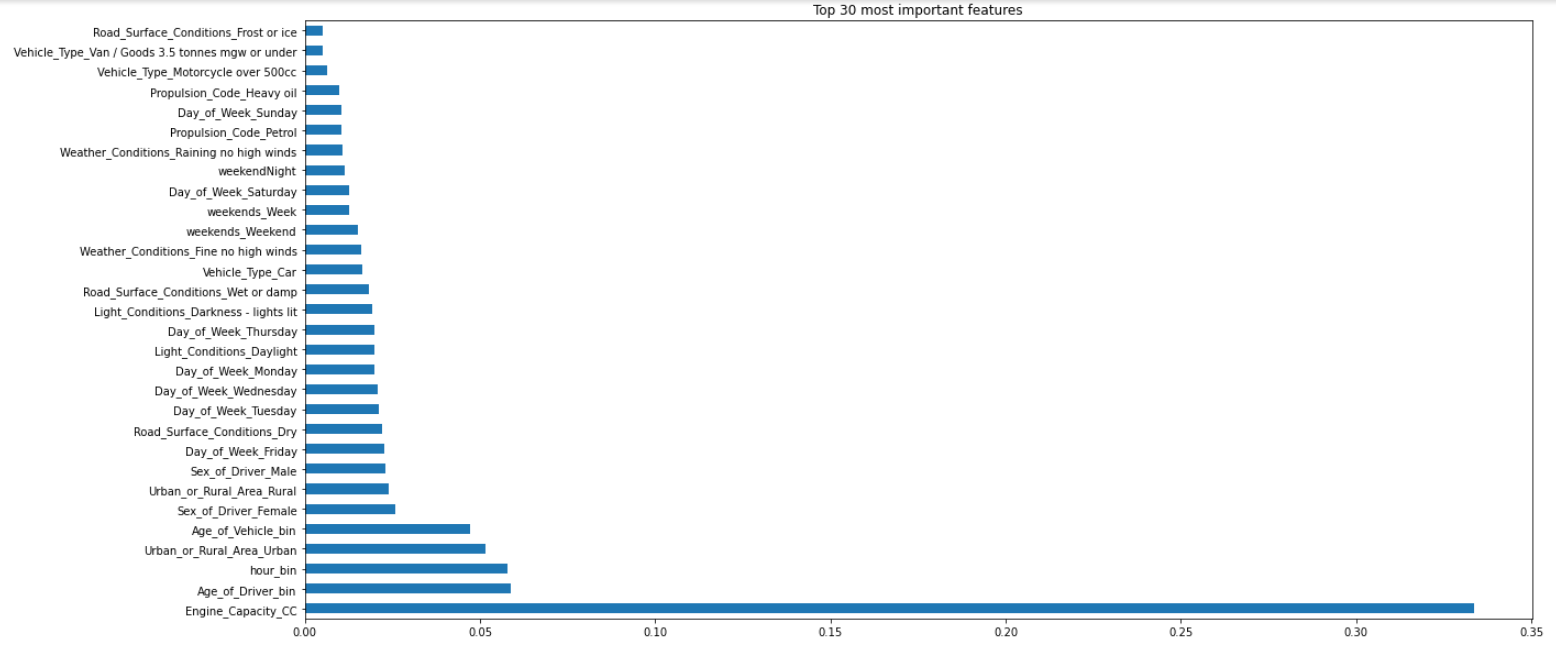
On all the previous tests, it was used the following features: *Features used: Speed limit, 1st Road Class, Road Type, Junction Detail, Pedestrian Crossing Human Control, Pedestrian Crossing Physical Facilities, Light Conditions, Urban or Rural Area, Day of Week, Weather Conditions, Road Surface Conditions, Carriageway Hazards, Special Conditions at Site, Engine Capacity CC, Vehicle Location Restricted Lane, Vehicle Leaving Carriageway, Vehicle Type, Vehicle Manoeuvre, Journey Purpose of Driver, Sex of Driver, Propulsion Code, Was Vehicle Left Hand Drive, Junction Location, Towing and Articulation, Age of Driver, Age of Vehicle, weekends, weekendNight, hour*

To make a fair comparison with other studies, and to proper address the answer to our aim, we restricted, even more, the features, to match the features they had, and the ones we can know in advance. The results are still pretty good:

Features used: *“Light Conditions, Urban or Rural Area, Day of Week, Weather Conditions, Road Surface Conditions, Engine Capacity CC, Vehicle Type, Sex of Driver, Propulsion Code, Age of Driver, Age of Vehicle, weekends, weekendNight, hour”*



Top 30 most important features



**Conclusions**

With this study, using 14 variables that we can now in advance before driving around, we could predict the severity of an accident we might be involved. We can predict the risk of driving on certain conditions.

Additionally, we can conclude that the combination of SMOTE and Random Forests proved to be an excellent way to make this prediction.

Finally, the most essential features, except for the engine capacity, are the same as exposed during the exploration analysis, even when it was done previous any prediction.

**Future Work**

One major problem we faced in this work was limited computing power. Even using google colab, the model execution crashed many times due to the size of the dataset, especially after SMOTE, turning very difficult to test it with different hyperparameters and trying other models. There is space for improvement in this aspect.

**Links**

GoogleDrive\Colab (recommended)

* **CarAccidentsUK.ipynb (main notebook)**
* **CarAccidentsFunctionsUK.ipynb (helper functions notebook)**
* **Data (folder with the dataset)**

[**https://drive.google.com/drive/u/3/folders/1yeexTbmNkmtMEwbb6PRtXYta4cqQt2p7**](https://drive.google.com/drive/u/3/folders/1yeexTbmNkmtMEwbb6PRtXYta4cqQt2p7)

**GitHub (notebooks and zipped dataset)**

[**https://github.com/developdaniels/CA683\_CarAccidentsSeverity**](https://github.com/developdaniels/CA683_CarAccidentsSeverity)

**References**

**[1] M. Sameen and B. Pradhan, "Severity Prediction of Traffic Accidents with Recurrent Neural Networks", *Applied Sciences*, vol. 7, no. 6, p. 476, 2017. Available: 10.3390/app7060476.**

**[2] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 272-276.**

**[3] W. Budiawan, S. Saptadi, Sriyanto, C. Tjioe and T. Phommachak, "Traffic Accident Severity Prediction Using Naive Bayes Algorithm - A Case Study of Semarang Toll Road", *IOP Conference Series: Materials Science and Engineering*, vol. 598, p. 012089, 2019. Available: 10.1088/1757-899x/598/1/012089.**

**[4]** **Alkheder, S., Taamneh, M., and Taamneh, S. ( 2017) Severity Prediction of Traffic Accident Using an Artificial Neural Network. *J. Forecast.*, 36: 100– 108. doi:**[**10.1002/for.2425**](https://doi.org/10.1002/for.2425)**.**

**[5] O. TAUBMAN - BEN-ARI and L. FINDLER, "RECKLESS DRIVING AND GENDER: AN EXAMINATION OF A TERROR MANAGEMENT THEORY EXPLANATION", *Death Studies*, vol. 27, no. 7, pp. 603-618, 2003. Available: 10.1080/07481180302898.**

**[6] A. Bener, E. Yildirim, T. Özkan and T. Lajunen, "Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study", *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 4, no. 5, pp. 496-502, 2017. Available: 10.1016/j.jtte.2017.07.005.**

**[7] N. Jacobs, M. Harris and D. Trew, "Interrelation between Alcohol and Accidents", *Journal of the Royal Society of Medicine*, vol. 82, no. 7, pp. 447-447, 1989. Available: 10.1177/014107688908200736.**

**[8] C. Zwerling, "Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors", *Injury Prevention*, vol. 11, no. 1, pp. 24-28, 2005. Available: 10.1136/ip.2004.005959.**

**[9]** **"Background to Driver Distraction / Mobile Phones & Driving", *Rsa.ie*, 2015. [Online]. Available: https://www.rsa.ie/PageFiles/1879/Background%20to%20Driver%20Distraction%20(Mobile%20Phones%20and%20Driving).pdf. [Accessed: 15- Apr- 2020].**

**[10] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. Available: 10.1613/jair.953.**