# Traffic Road Accident Severity Prediction in Real-Time Using Artificial Intelligence

Rafael Torrecilla Rubio
Dublin City University,
Insight Centre for Data Analytics
*Dublin, Ireland*
rafael.torrecillarubio2@mail.dcu.ie

Parth Khare
Dublin City University
*Dublin, Ireland*
parth.khare2@mail.dcu.ie

Sayli Dixit
Dublin City University
*Dublin, Ireland*
sayli.dixit2@mail.dcu.ie

Apurva Saswadkar
Dublin City University
*Dublin, Ireland*
apurva.saswadkar2@mail.dcu.ie

Aaditi Pikale
Dublin City University
*Dublin, Ireland*
aaditi.pikale2@mail.dcu.ie

*Abstract*—**Traffic Road Accidents cause several severe injuries or even deaths every year among the United States of America. Deaths and injury damage can be reduced if the injured patient is treated in the least time possible. However, Emergency Medical Systems are limited and information of patient condition at the time of accident notification is usually not available. Thus, miss-interpretation of severity will lead to death or injure severity. The proposed research shows a working algorithm for Real-Time classification of severe accidents using little easy automatically collectable information such as Temperature and Humidity. Research methodology is explained for transparency and reproducibility of results as well. Final model was built with K-Nearest Neighbors (KNN) showing high precision and recall scores and could be deployed for States such as California, Texas and Florida. Developed model can be combined with on-car severity prediction technologies or used alone if not available. Model extension to other areas as well as improvement of already achieved results for previously mentioned states can easily be performed with KNN. Extension to further countries is possible although country development state should be taken into consideration as it affects the model.**

***Keywords—Traffic Road Accident, Severity Prediction, Real-Time, Artificial Intelligence***

## I. Introduction

Traffic road accidents (TRA) are a relevant form of unnatural death among the United States of America (USA). According to the data provided by the Word Health Organisation (WHO), 34064 deaths occurred during 2013 in the US [1], being 63% of the total death caused to occupants of 4-wheeled vehicles [2]. Compared to other developed countries such as Japan, Australia, Germany, France and other European States, the USA presents a higher number of deaths per 100.000 inhabitants than the previous countries [3]. Thus, an improvement on traffic road security, health services and transport policies will lead to a death rate reduction.

This research article focuses on the development of a Real-Time Artificial Intelligence (AI) based model for severity grade prediction at the time that the traffic road accident is notified to the competent authorities. The algorithm will help the determination of one among four severity grades (minor, moderate, severe and fatal) in order to better distribute police and emergency medical systems (EMS). Prior identification of severe and fatal accidents can lead to life saving. As stated in [4], one of the most common types of death produced in TRA is the loss of oxygin that occurs in less than 4 minutes. Moreover, the article explains that health treatment during the first hour after an accident, the so-called “Golden Hour”, significantly reduces death probability as well as health damage to the involved patient(s). Additional research conducted in [5] and [6] shows that death rate increases the more delayed the EMS arrives and presents an automatic crash notification system on vehicles. However, not all vehicles may have incorporated the system, in which case the proposed algorithm will help determine the severity. Moreover, the information provided by the automatic crash notification system could be combined with the model proposed in this paper. Thus, this research presents new methodologies to reduce EMS arrival time that can be complemented as well with state of the art technologies, providing an additional source for live saving.

Previous research on traffic road accident prediction indicates that Machine Learning (ML) and Artificial Neural Networks (ANN) approaches can be implemented for TRA prediction. The research conducted by Mohammed et al. [7] reviews models that can have been applied for this aim, identifying advantages and drawbacks. Complementary, research in [8], data mining techniques are presented for the determination or variables affecting the severity grade in TRA. Successful implemented models for a specific road type (toll road) in Indonesia [9] and general roads in Switzerland [10] use datasets that include road geometrical characteristics which are then given as inputs to the model. However, models vary on measured variables on datasets, implemented approaches and, as stated in [11], the development grade of the country also has an influence on how the model design. The last paper explains this issue as developing countries possess less quality roads and less safe vehicles than developed countries.

Moreover, additional factors other than road geometrical characteristics affect the number of traffic road accidents as well as their severity. The WHO [12] explains that human factors such as age or gender among other variables have an influence as well, being young males at greater risk of suffering a fatal traffic road accident. Models have been implemented with the aforementioned human factors and those variables have been included for crash prediction as in [13] and [14]. These articles present successfully implemented models for traffic road accident detection as well as the corresponding severity. Nonetheless, as determined in [14], severe and fatal accidents were predicted for the training set but not for the test set, as occurrence is lower than for minor and moderate severities. This adds a challenge to overcome to the proposed model in this paper, as severe and fatal severities need the more sanitary attention.

All previous models try to predict TRA as occurrences. The models aim to reduce the number of traffic road accidents and severity by providing a software tool. This software will be used in future infrastructure and transport management by the competent authorities. However, these models have been trained with some measured variables obtained by exhaustive research time after the accident has occurred. These models would find, therefore, difficulties in predicting an accident severity by the time it is notified as no prior information on gender or age, for instance, is provided. Thus, the proposed algorithm should be based only on variables that can easily be measured at the time of the accident.

Another difficulty lies in the difference among datasets. As explained before, mentioned models differ on the measured variables included. The dataset for this research has been obtained from [15], provided by the research conducted in [16] and [17]. For instance, when compared to the model in [14], both models group severities in four categories. However, minor severity accidents represent the most significant category while being the least recorded on the dataset used for this research. Data mining techniques are to be applied to deal with theseissues.

## II. Data Management and methodology

### A. *Data Management*

To provide transparent and reproducible research to help future research on the area, data have been treated following the guidelines found in the Cross Industry Standard Process for Data Mining (CRISP-DM) [18] data mining methodology.

The research project has been divided into three groups among involved researchers: Project Management, Data Mining and Model Development. Each group performed work on the assigned tasks and once completed, results were shared among all groups to decide following steps and to assure retro-alimentation among phases specified in the CRISP-DM methodology.

The section below, *Methodology*, details the steps followed during the research process. This includes not only the right steps but also challenges found and techniques implemented to overcome them. This will provide references for future projects to avoid issues at different project stages. Feedback is an important part of the process performed during the whole research and among all phases. Finally, the working model is presented.

CRISP-DM phases are written in italic to better guide readers through the research process.

### B. *Methodology*

First of all, *Business Understanding* is needed to address specific challenges to be improved or solved in the field. Research on the area led to a gap in knowledge: Whether a Real-Time prediction of the severity of TRA at the accident notification time is possible. Providing a positive solution, can lead to life saving and injure reduccion for severe and fatal accidents.

The previous approach implies that post-TRA data such as speed, age or gender can't be used for model development. Thus, weather conditions and geometrical characteristics of the road are the only variables that can be used as inputs. Furthermore, implemented variables should easily and quickly be obtained and, if possible, automatically by the algorithm.

Onced that the main research question has been determined, the next step is the *Data Understanding*. The dataset used possesses almost 3 million records (2,974,335) with 49 different acquired variables during the period of 2016-2019 in the USA, requiring 1 GB of disk memory. Variables include some road characteristics (e.g. traffic signal, give way or junction), weather condition (e.g. precipitation, wind speed or visibility), location (State, zip code, street) or time data (e.g. start time or end time). No Human Factors, such as age or sex, or Traffic Flow are recorded in the data-set. Severity Grades are given in four categories: Severity I-Low, Severity II-Minor, Severity III-Moderate and Severity IV-Serious.

Despite a high number of disponible variables, coming back to the research question, the more variables needed to be gathered for the model, the more complex it will be and the more difficult to implement in Real-Time. Thus, dimensionality reduction is needed. Dimensionality reduction can be implemented by application of a correlation matrix to detect multi-colinearity. However, no strong multi-colinearity was detected with this technique. Further dimensionality reduction is conducted by applying Principal Component Analysis (PCA) to find which variables contribute the most to the Severity Grade. Although Autoencoders can be used for this process as well, PCA is preferred, as most significant variables are known and selected as inputs before implementing the model. Results of implementing PCA on the data-set is presented in Fig. 1.

PCA analysis showed that two variables are responsible for almost all contribution to the Severity Grade, being the remaining variable contribution neglectable.

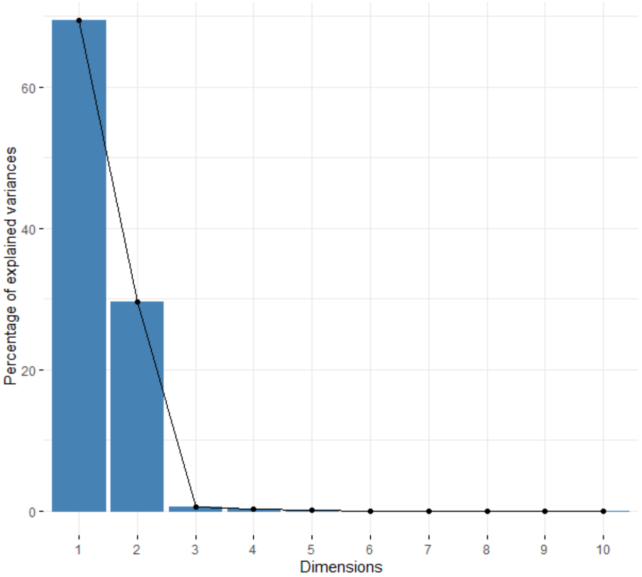

*Fig. 1: Principal Component Analysis for data-set. Inputs 1, Temperature, and 2, Humidity show the most contribution to the system. Remaining variables contribution is neglectable in comparison . Visibility and pressure account as the third and fourth position respectively*

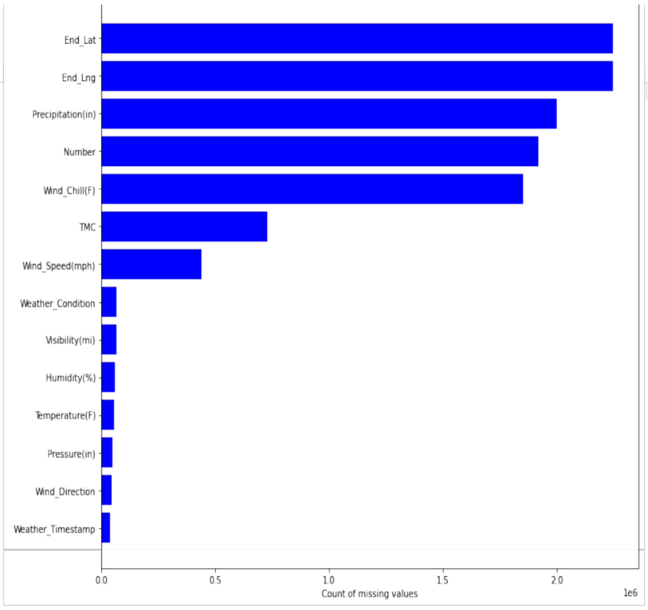

*Fig. 2: Count of missing values for variables in the data-set. Variables contributing the most to changes in Severity (Temperature and Humidity) show little missing values compared to the total size of the data-set( < 1%)*

Missing values are present along the data-set. The number of missing values was determined, as seen in Fig. 2. Results showed that Temperature and Humidity presented few blank values. Thus, eliminating samples with missing Temperature or Humidity values was the strategy implemented as part of the *Data Preparation*. Moreover, most Severe accidents are the most relevant to this research, being these samples out of the mean and average range values, acting more as outliers. Thus, replacing blank values with statistical analysis has been rejected.

Another issue found along the data-set is the presence of unbalances. Data is highly unbalanced, with little the least samples corresponding to Severity I, Low, followed by Severity IV, Severe, Severity III and II. Details are illustrated in Table I. Small percentage of sample is expected for Severity IV, as they are rare occurrences. However, Severity I shows few sample as well. This may be caused due to the measuring scale in Severity, non-notification of low impact TRA or lost of data.

TABLE I. SAMPLES COUNT PER SEVERITY GRADE

| **Sample Count** | **Severity Grade** | | | |
|---|---|---|---|---|
| | ***I***<br>***Low*** | ***II***<br>***minor*** | ***III***<br>***moderate*** | ***IV***<br>***severe*** |
| **Number of samples** | 968 | 199341 | 88762 | 92337 |
| **Percentage (%)** | 0.03 | 67.02 | 29.84 | 3.10 |

Data plots can offer a hint on which model is expected to work better. Due to the fact that two variables contribute the most to the variation in the Severity Grade, a Temperature vs Humidity plot for all four severities, as in Fig. 3, can be used to discover any patterns before selecting a model. Nonetheless, Severities no patterns could be recognised as data for all severities seems to be equally distributed, being some outliers shown as well. This presents a difficult challenge as Severity Grades can easily be confused and remaining variables in the data-set don't contribute enough and thus, incorporating them is not expected to greatly increase results.

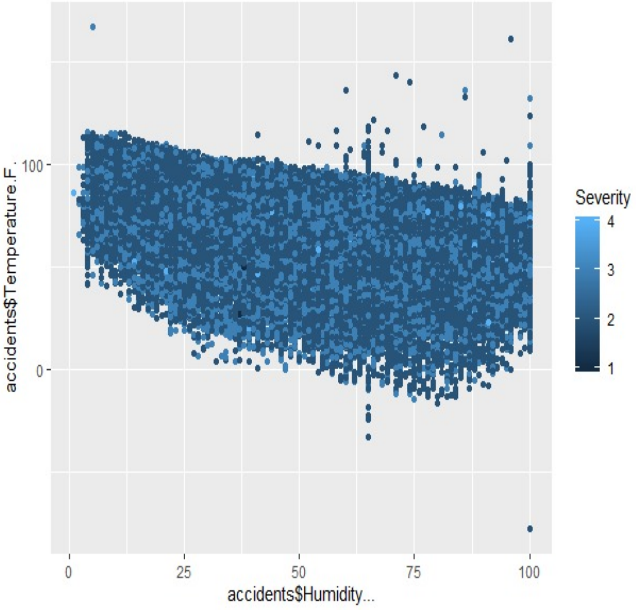

*Fig. 3: Humidity vs Temperature plot for all samples in the data-set and all four Severity Grades. No visible clusters or patterns can be observed at first sight*

Clusters in Fig. 3 are difficult to visualize due to the amount of data, Humidity vs Temperature. Thus, plots per each individual were created. Patterns were not easily visible, although pairs of Severities presented more data in similar areas. Specifically, pairs Severity I-IV and Severity II-III plots illustrated similar results.

After understanding the data, the *Data preparation* was conducted by eliminating missing values and randomly splitting the data in Training (70%), Validation (20%) and Test (10%). Randomly splitting was selected to avoid any bias on the model. Loaded inputs were Temperature and Humidity with One-Hot encoded labels for the corresponding Severity Grade.

*Modelling* predictions on Severity Grade being Temperature and Humidity the inputs with different techniques such as Multi-Layer perceptron (MLP), Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN) and Random Forest. For the case of MLP, Adam and Stochastic Gradient Descent (SGD) optimizers were used. Additional hyperparameters introduced for SGD were decay and momentum. Previous models showed low accuracy for the training and validation data. *Evaluation* confirmed that algorithms were trying to always predict Severity II, as this severity is the most representative sample group. Additional analysis was performed with four inputs, including the next two most significant variables: Visibility and Pressure, which concluded in similar results with continuous prediction of Severity II. Despite of algorithm's failure, these results were expected and this step was realized to find out useful information for future models.

During the *Data Understanding* phase, PCA was performed on each of the Severity Grades. PCA showed that even though Temperature and Humidity are the most significant inputs for the system, the presented vector is not the same. This means that patterns can be detected with AI. However, vector orientation for Severity I and IV is similar and thus, they will be confused. Moreover, individual plots Temperature vs Humidity for each Severity Grade corroborate PCA vector analysis as clusters and patterns were easier to identificate.

Distinguish Severity IV cases from the rest of Grades is critical for the research question. Thus, a solution is to be found to the challenge of being unable to distinguish between Severity I and IV. Coming back to the *Business Understanding*, low impact TRA are usually notified by the involved affected participants. However, this is not usually the case for severe TRA. Thus, Severity I accidents can be more accurately being classified by the competent authorities at the time of notification. This factor leads the research to the development of a new classification algorithm for the classification of Severity Grades II, III and IV.

Eliminating Severity I during the *Data Preparation* leads to better balancing possibilities as there are more samples for Severity IV than for Severity I. The aim is to avoid results shown previously during the M*odelling* phase, where only one class was predicted due to imbalances in data-set. Thus, samples for Severity II and III were randomly reduced to the number of samples equal to the ones for Severity IV for training, validation and test data-sets.

*Modelling* with different techniques and balanced data to classify three Severity Grades, excluding Severity I, showed that the applied models were performing similarly than previous implementations. At this stage, further *Business Understanding* was required to discover any missing information to overcome this issue. Research led to the conclusion that there were hidden unbalances in the data-set among States, as found in [19]. Most states contained only a few samples. Three States accumulated most of the samples: California, Texas and Florida.

*Data Preparation* with samples for those specific States was implemented and then, used for modelling with different techniques. Among previously applied algorithms, KNN was determined to perform most accurately for the three aforementioned States. Examination of the confusion matrix indicated that classes were effectively being detected, although some miss-classifications were still presented. The most miss-classifications were found for class II and III, which was expected as a result of the previous PCA vector analysis and plots. Models were validated with the validation data to contrast results obtained with the training samples.

During the *Evalutation* phase, the final models for the three States were implemented on the test data. Overall, results obtained were less accurate but still valid. Thus, algorithm performed well on previously unseen data.

## III. Results

A working model was achieved for Real-Time prediction of TRA Severity Grade at the time of accident notification. Unbalances in the data-set were overcome, showing that Severity Grades can accurately be predicted for three states: California, Texas and Florida. Remaining states present few sample data and predictions were not reliable.

Severity I, low impact, was eliminated from the classification model as it can be determined by competent authorities at the notification time. Among the remaining Severity Grades, II and III showed a strong similarity and were frequently miss-classified. Nonetheless, Severity IV, severe, is the main focus of this article, as it is the type of TRA that needs the quickest EMS response. Results are presented in Table II for the training set and Table III for the test set.

TABLE II. Prediction values for severe accidents on training data

| **State** | **Metrics – Training Data** | | |
|---|---|---|---|
| | ***Accuracy avg.*** | ***Precision Severity IV*** | ***Recall Severity IV*** |
| **California** | 0.71 | 0.98 | 0.73 |
| **Texas** | 0.81 | 0.99 | 0.81 |
| **Florida** | 0.65 | 0.90 | 0.68 |

TABLE III. Prediction values for severe accidents on training data

| **State** | **Metrics – Test Data** | | |
|---|---|---|---|
| | ***Accuracy avg.*** | ***Precision Severity IV*** | ***Recall Severity IV*** |
| **California** | 0.57 | 1.00 | 0.81 |
| **Texas** | 0.76 | 1.00 | 0.68 |
| **Florida** | 0.59 | 0.93 | 0.64 |

Regarding results obtained for model training in Table II, predictions average accuracy for Severity II, III an IV range between 65 and 81%. As the research focus lies on the identification of Severity IV TRA, average accuracy values represent a metric for the overall performance of the model, although precision and recall for Severity IV are the metrics of interest, as they show how well it is (miss-) classified.

Severity IV precision for all states is over 90%, being even 99% for California and Texas. On the other hand, Recall values remain lower within a range of 70 and 79% but still providing acceptable scores. Overall, Texas is the best predicted state.

Model *Evalutation* results, presented in Table III, show a considerable decreasse in model accuracy, ranging between 57% and 76%. The most significat decreasse was shown in California, reducing model accuracy from 71% to 57%. However, overall accuracy is not the most interesting metric to be taken into account for the research aim. Severity IV recognition is more relevant, being Precision and Recall more important. For the case of California, Precision and Recall scores performed better for the test data. Thus, even though overall accuracy is lower, model performs better on classification of Severity IV TRA. Precision scores were better for all States. Nonetheless, lower Recall scores were shown for Texas and Florida.

Model *Deployment* is then possible for the mentioned states and it can easily be implemented as only two variables are required for Severity classification, Temperature and Humidity, which could automatically be acquired in Real-Time. *Deployment* will lead to live saving and injure reduction and to the collection of additional samples to further algorithm learning. Model extension to other states can be implemented if samples are increased and the data-set becomes more balanced through the remaining states in the USA.

## IV. Conclusion and future work

A functional model for prediction of Severe Traffic Road Accidents in Real-Time has successfully been implemented using Artificial Intelligence methods with high Precision and Recall values for severe accidents. Model can help competent authorities to better deploy Emergency Medical Systems determining the Severity Grade at the time of notification.

Early recognition of severe accidents and quick sanitary response will lead to live saving and injury minimation.

Data-set is highly unbalanced with few samples for Low Impact and severe Traffic Road Accidents. Furthermore, it is geographically unbalanced, being the most representative States California, Texas and Florida. It was not possible to implement a working solution to the remaining States as more samples are needed.

Future work includes the collection of additional samples for the extension of the model through other States among the USA. Moreover, further gathered samples for states where the model is deployed can increase learning of current algorithm. Acquisition of complementary data-sets is highly advised for model improvement.

Extension to other countries can be conducted as well. However, country development status is to be considered as it has an impact in predictions due to differences in road quality, vehicle type or traffic policies.

Ehtical issues should be consired before and during *deployment* phase as proposed solution has an impact on human beings.

## References

[1] Road safety, estimated number of road traffic deaths, 2013, gamapserver.who.int, http://gamapserver.who.int/gho/interactive_charts/road_safety/road_traffic_deaths/atlas.html (accessed on 17th April, 2020).

[2] Road safety, Number of road traffic deaths and distribution by type of road user, 2013, gamapserver.who.int, http://gamapserver.who.int/gho/interactive_charts/road_safety/road_traffic_deaths3/atlas.html (accessed on 17th April, 2020).

[3] Road safety, Estimated road traffic death rate (per 100 000 population), 2016, gamapserver.who.int, http://gamapserver.who.int/gho/interactive_charts/road_safety/road_traffic_deaths2/atlas.html (accessed on 17th April, 2020).

[4] S. Gopalakrishnan, “A Public Health Perspective of Road Traffic Accidents”, 2012, J Family Med Prim Care. 2012 Jul-Dec; 1(2): 144–150. doi: 10.4103/2249-4863.104987: 10.4103/2249-4863.104987.

[5] D. E. Clark, B. M. Cushing, “Predicted effect of automatic crash notification on traffic mortality”, Accident Analysis and Prevention 34 (2002) 507– 513.

[6] M. et all.“Automatic Accident Detection: Assistance Through Communication Technologies and Vehicles. IEEE Vehicular Technology Magazine.”, 2012, 7(3):90-100. doi:10.1109/MVT.2012.2203877.

[7] A. Mohammed et al. “Classification of traffic accident prediction models: A review paper.”, 2018, International Journal of Advances in Science Engineering and Technology, ISSN(p): 2321 –8991, ISSN(e): 2321 –9009, Volume-6, Issue-2, Apr.-2018, http://iraj.in

[8] L. Li, S. Shrestha, G. Hu, “Analysis of Road Traffic Fatal Accidents Using Data Mining Techniques”, SERA 2017, June 7-9, 2017, London, UK

[9] A. Irfan, R. Al Rasyid, S. Handayani, “Data mining applied for accident prediction model in Indonesia toll road”, 2018, AIP Conference Proceedings 1977, 060001, https://doi.org/10.1063/1.5043013 (accessed on 5th April, 2020).

[10] B. García de Soto, A. Bumbacher, M. Deublein, B. T. Adey, “Predicting road traffic accidents using artificial neural network models.”, 2018, Infrastructure Asset Management 5(4): 132–144, https://doi.org/10.1680/jinam.17.00028 (accessed on 27th March, 2020).

[11] F. N. Ogwueleka et al., “An Artificial Neural Network Model for Road Accident Prediction: A Case Study of a Developing Country”, 2014, Acta Polytechnica Hungarica, Vol. 11, No. 5.

[12] Road traffic injuries, who.int, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed on 17th April, 2020).

[13] C. Dong, K. Xie, X. Sun, M. Lyu, H. Yue, “Roadway traffic crash prediction using a state-space model based support vector regression approach.”, 2019, PLoS ONE 14(4): e0214866. https://doi.org/10.1371/journal.pone.0214866 (accessed on 5th April, 2020).

[14] S. Alkheder, M. Taamneh, S. Taamneh, “Severity Prediction of Traffic Accident Using an Artificial Neural Network”, 2016, Journal of Forecasting, J. Forecast. DOI: 10.1002/for.2425.

[15] US Accidents (3.0 million records), kaggle.com, https://www.kaggle.com/sobhanmoosavi/us-accidents (accessed on 17th April, 2020).

[16] Moosavi et all. “A Countrywide Traffic Accident Dataset.”, 2019.

[17] Moosavi et all. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[18] CRISP-DM, Data Mining Phases, crisp-dm.eu, http://crisp-dm.eu/reference-model/ (accessed on 17th April, 2020).

[19] Sobhan Moosavi, US-Accidents: A Countrywide Traffic Accident Dataset, December 2019, smoosavi.org, https://smoosavi.org/datasets/us_accidents (accessed on 17th April, 2020).