A Comprehensive Study of Road Traffic Accidents in the UK: Hotspot Analysis and Severity Prediction Using Machine Learning

Harshita Vidapanakal
CSE Department
PES University
Bengaluru, India
harshita.vidapanakal@gmail.com

Utkarsh Gupta
CSE Department
PES University
Bengaluru, India
utkarsh348@gmail.com

Varun MK
CSE Department
PES University
Bengaluru, India
mkvarun2001@gmail.com

***Abstract*— It is vital to study road traffic accident data as it is recorded due to the fact that decisions taken due to these studies have quite real life or death impacts. It is also important to get the full picture of the situation and truly understand what are the real pain points in our system and infrastructure and where do we need to focus our efforts so as to make the lives of the public safer, We use KDE plots to analyse hotspots of accident-prone areas weighed against severity over years to understand the evolution of these dangerous zones. Furthermore, we use machine learning algorithms to predict the accident severity given certain parameters and to understand what factors influence the severity of the accident. Finally, we perform time series analysis on the data to understand the trend of the accidents and to predict the number of accidents in the near future.**

***Keywords*—** *hotspots, KDE, Severity prediction, time series analysis*

Introduction

A steep increase in population and motorisation has led to a rising trend in road accidents. Road traffic injury is the eighth leading cause of death of the global level and the leading cause of death for young people aged 15-29 years. One of the primary causes being the increase in road transport in comparison to less progress in other types of transportation systems and insufficient infrastructures in Iran, has significantly increased the urban pollution, road users wasted time and above all the damages caused by traffic accidents. [1][2]

Efforts taken to prevent road accidents face the imperative problem of answering the question of where to implement safety precautionary measures. We answer this question by identifying “Hot spots”. “Hot Spots” or “Black Spots” are locations identified by a high accident occurrence compared with the other locations. [3]

This is why it is important for us to analyse where and when these accidents occur frequently. The locations, which are identified by a high accident occurrence compared with the other locations, are known as hotspots or black spots [3]. Occurrences of accidents are not random in space and time. They depend on factors such as geometric design, severe weather conditions, time of the day, etc. Plotting, spotting and cataloguing these hotspots helps us locate the most accident-prone areas and where the focused efforts, energy and resources of the concerned authorities are required to diagnose the core issues and help counter the said issues. Simply plotting the individual crash sites on a map does not work either as this does show a strong concentration of incidents at certain locations but is vague in the information it provides, giving no density information.

KDE(Kernel Density Estimation) plots are employed and weighed against a severity index to understand the precarious nature of certain locations. They are also plotted over time to understand the evolution of these hotspots.

A number of factors affecting the severity of an accident are interrelated. Various statistical ML models have been employed to predict this severity with the factors taken into consideration being vehicle and road conditions. The relationship between these parameters not only help predict severity but also give us information as to what exactly influences the severity of these accidents the most.

The models used on the data are Naive-Bayes, Logistic regression, AdaBoost, XGBoost and Random Forest classifiers to classify the data based on different levels of prediction.

Previous Work

A study of past work has shown that many traditional statistical methods have been used to detect hotspots[3][4]. In these models, spatial characteristics of hotspots were

modelled as a constant for a given period of time which is not true[5]. This dataset provides a means to analyse the surrounding environment, severe weather conditions which can further be used to calculate severity index or simply take these factors into account during model building which shows a significant improvement in ranking/detection of hotspots.[4][5]. GIS-based statistical analytic techniques like KDE(Kernel Density Estimation) to analyse hotspots according to time intervals and seasons with and without taking SI(Severity Index) into consideration. This paper was able to conclude that SI is strongly related to seasons, time of the day. Comap was used to visualise all these various results. It is another example of how time and space are taken into consideration. Their study concluded that traffic accidents occurred frequently twice in a day and they were able to pinpoint the intersections[6]. Customised spatial weight matrices have been used to detect hotspots too. In this procedure, an Inverse

Network Distance-Band Spatial Weights Matrix of Intersections(INDSWMI) is built to form the network of roads by taking road network constraints into consideration. Then the k-nearest Distance Band Spatial Weights Matrix Between Crash and Intersection (KDSWMCI) using the adjacency crashes for each intersection. These are then used to perform Intersection Hotspot Analysis(IHA) with the measure of statistic being Getis-Ord Gi* statistics and the accuracy was measured using Intersection Prediction Accuracy Index(IPAI). [7] KDE has also been compared with other methods in some studies. In one study, KDE and Kriging were used for hotspot detection with the measure of accuracy being Prediction Accuracy Index(PAI). It was found that the list of hotspots identified by the two methods were moderately different. It relates crashes as a function of potential variables to make its predictions.[8] Other unsupervised machine learning methods like k-means have also successfully profiled vehicle accident hotspots. [9]

Statistical models which predict the severity of the accident exploit the relationship between the different parameters in question. Artificial Neural networks have also been used in this regard. As with any statistical model, they have a set of assumptions and will perform poorly in case these assumptions are violated. Bagheri et al[10] performed a feature analysis and showed that the average speed of vehicles and average traffic volume are the most contributing factors to accidents. They used ANN and log regression models for their prediction. ANN was also used by Cansiz to estimate the number of fatalities[11].
Regression has also been employed to estimate quantities related to accidents. Injuries, fatalities and number of accidents. They also used Genetic Algorithm models for the same and concluded that ANN worked the best for their data.[12][13] O'Donnell considered data from New South Wales, Australia uses ordered logit model and ordered probit models to estimate the probability of injury and death given conditions like seating position, blood alcohol level, vehicle type, make of collision, etc. But these models, with the exception of ANN, have assumptions which do not hold in real life. This is why the Bayesian network model was implemented since BNs are capable of making predictions without the need for pre assumptions and can also be used to make graphical representations of complex relationships. The factors which affected their classification the most were accident type, lighting and number of injuries[14][15]. Assi[16] uses 15 crash-related parameters in an FCM(Fuzzy C-Means) clustering algorithm to improve the power of the SVM model to predict crash severity given vehicle and road condition attributes. It also explored 3 other models: FNN(Feed-Forward Neural Networks), FNN-FCM and SVM(Support Vector Machines).

Data description

The region of interest is the United Kingdom. The study is based on road accidents that occurred in 2005-2017 in the UK. The data shows us a lot of interesting information about the accident that we can choose to visualize such as the number of accidents, the conditions and status of the environment at the time, severity of these accidents etc.

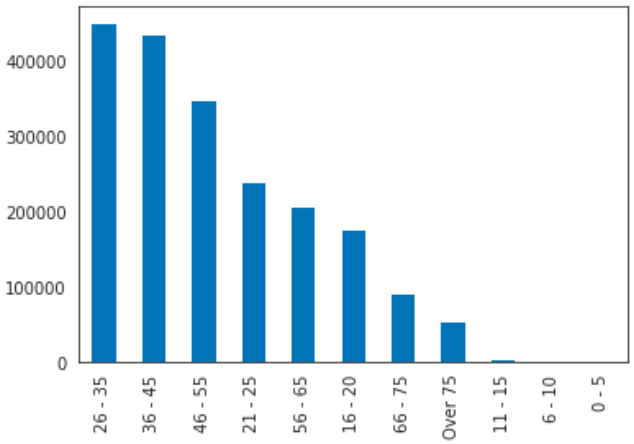

Figure 1: Number of accidents involving each age group

One such observation is the decrease in the number of road accidents in the UK between 2005-2017[Fig.1], with most accidents including drivers in the age range of 26-35(little more than 20%). [Fig.1]

While not so surprising was that the number of fatal accidents formed a smaller portion of the total number of accidents[Fig.2], what was still interesting to observe was that the number of fatal accidents in the

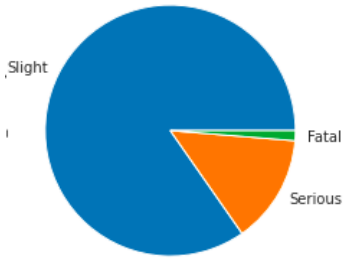

Figure 2: Proportions of severity in accidents

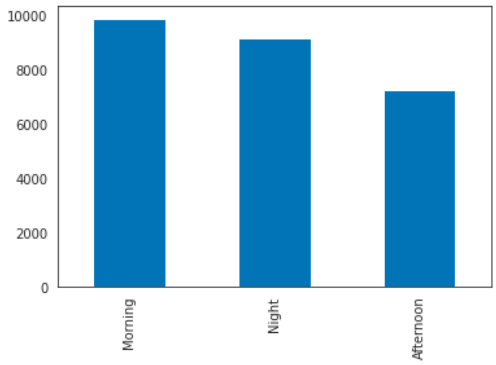

Figure 3: Number of fatal accidents at each time of the day

morning and night were more than the ones that occurred in the afternoon, arguably at peak traffic. Fig[3]

Kernel Density Estimation

KDE is a non-parametric spatial statistical tool that estimates probabilities, probability density functions for random variables. It is a solution to a smoothening problem to get inferences from the data. We aim to analyse and plot the hotspots of all such accidents so as to pinpoint what are the points of concern in the highway network. In the dataset, all road accidents are classified under three levels of severity: Slight, Serious and Fatal. It is vital that we do not treat all records the same but give more weight to the locations that proved to be more dangerous overall by forming a severity index[6]. This shortcoming is accentuated by the fact that the more severe accidents are vastly outnumbered by the minor, slight accidents. Here we choose to allot quartic weights to each severity level.

This can skew our understanding and analysis of the map when our focus is looking for hazardous locations. Another observation that we have noted is that there is an overall decrease in the total number of accidents every year.

This is good but it also means that plotting all accidents for a certain location can only give a part of the information. It becomes quite important to then use KDE to plot accident hotspots for certain periods of time to not just record all hazardous locations but also to take notice of how these hotspots evolved over time.

This temporal evolution of hotspots can be interpreted in three different ways corresponding to how they change over time. [17]

- Type 1 - There is a dissipation in hotspots at certain locations either due to change in infrastructure or applications of new safety protocols.
- Type 2 - Indicates hotspots that have shown no change in their characteristics over time and continue to identify dangerous locations.
- Type 3 - Appearance of new hotspots or increase in density of ones present already, suggesting worsening conditions.

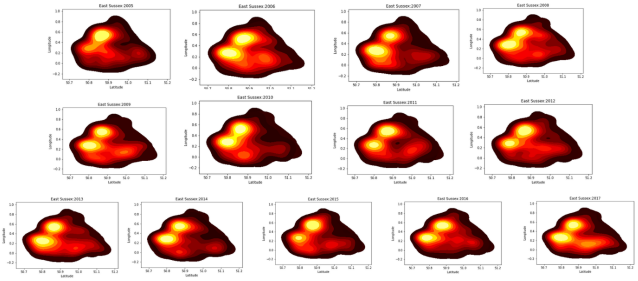

Figure 4: KDE Plot of East Sussex(2005-2017)

A. Case Study 1: East Sussex

As we can see in Fig[4], the weighted graph of accidents in 2005 gives us a slightly different result compared to the regular KDE plot wherein the hotspot at +0.6 longitude, +50.9 latitude is given more importance than the hotspot at +0.3 longitude, +50.8 latitude.

There are two important observations to make of the East Sussex study:

- There is an appearance of a prominent hotspot at the coordinates +0.3 longitude, +50.8 latitude, suggesting some dangerous change over time during the decade that needs to be dealt with.
- Another matter of concern is the worsening conditions at the coordinates +0.1 longitude, +51 latitude. This depicts a location that could prove to be more hazardous in the future if not handled well.

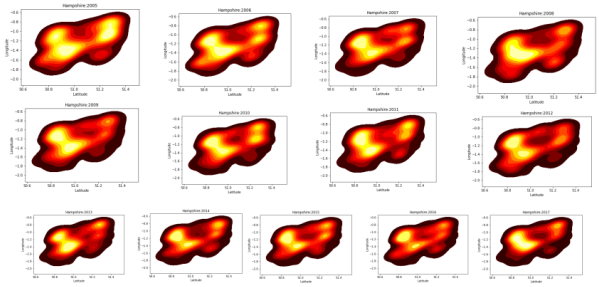

Figure 5: KDE Plot of Hampshire(2005-2017)

B. *Case Study 2: Hampshire*

Again, there are two important observations worth noting in Fig[5]

The bigger hotspot centered around -1.3 longitude, +50.9 latitude remains more or less the same size with the same intensity over the years. It is also the location with visibly the highest severity of accidents in the general area. This means that not much is being done about this issue or the tactics tried thus far have proved to be unsuccessful

The hotspot centered around -1.0 longitude, +51.3 latitude has slowly dissipated over time from 2005 to 2017. This is good because if we need models to study and implement in other similar locations, it is helpful to understand where exactly did certain measures work and why so.

Accident severity predictiction

The next step in understanding this dataset is to predict the severity of the accidents taking place based on the parameters recorded. The datasets separates different records into different levels of severity - Slight, Serious, Fatal. We use classification models - a whole division of machine learning algorithms to sort different accident records into these levels and essentially predict the severity, leading to the final step which is to analyse which attributes or parameters in the dataset seem to have the most influence on damage. Here, we try five separate models - Naive-Bayes classifier, Logistic Regression, AdaBoost, XGBoost and Random Forest Classifier.

A. *Data Preprocessing*

A huge majority of the parameters that are a part of the dataset are categorical attributes which requires us to encode them. We use the ordinal encoder to encode each of these attributes. Another point of concern is that this dataset is extremely imbalanced i.e.as expected, the number of “Fatal” and “Serious” accidents are much less in number when compared to “Slight” accidents. To tackle this issue, we use an imbalanced learning sampling method SMOTE(Synthetic Minority Oversampling Technique). This is a data augmentation technique wherein instead of creating copies of minority classes to oversample, we synthesize new records(being relatively similar) from the data. The drawback of this issue is that due to oversampling this way, training times are increased by a lot which makes it difficult to tune hyperparameters.

B. *Naive Bayes Classifier*

Categorical NB classifies discrete features which assume categorical distribution for each feature where the features have to be encoded using label encoding techniques such that each category would be mapped to a unique number. In Multinomial Naive Bayes, to estimate weights, only a single class c is used. On the other hand, in Complement Naive Bayes Algorithm, all training data of the classes except c class is used.

C. *Logistic Regression*

Logistic regression is a classification model built around the logistic function to predict the probability of the occurrence of a particular class. The model works best when the classes are linearly separable and particularly distinct in their characteristics. Similar to linear regression, the model is fit on the data using different kinds of optimisation algorithms, the decision for which can be made by considering the nature and size of the data. The model is implemented with cross validation to automatically select better hyperparameters along with the saga solver considering the presence of more than two classes and the size of the final dataset.

D. *Adaboost and XGBoost*

AdaBoost or “Adaptive Boosting” is a boosting ensemble machine learning algorithm that trains multiple weak classifiers sequentially. It is an iterative method to chain these weak classifiers and correct the misclassifications made by the previous model. Weak classifiers are better than random classification but not accurate enough to be useful. Decision trees are usually the default stumps or weak classifiers at each step and the cumulative result is one of a much more accurate and strong model than any one individual model used. Weights dependent on the error rate are given to each classification based on if it was classified correctly or not, to mark these misclassifications for the next classifier to work on. We train xgboost with default parameters using gbtree as the booster.[18]

E. *Random Forest*

The Random Forest classifier model is also an ensemble machine learning model wherein instead of boosting, there are individual trees in this “forest”, each acting as an independent predictor. A majority voting scheme is then applied to all the individual classifiers and based on this vote is the final classification allotted. An important point to note about the Random Forest model and its functionality is that each of these individual trees work independently and are not correlated in their working i.e, each tree is different and is fit on a unique set of attributes from the dataset

Results

Each of the models were tried with stratified values and with SMOTE to tackle the extreme imbalance in the data. The Naive-Bayes and Logistic regression models did not perform well on this dataset. Naive-Bayes, presumably, due to it assuming the independence of features in the dataset. Logistic regression delivered 86% without and close to 49% with SMOTE while Naive Bayes gave a mean accuracy of 57%. Boosting models, AdaBoost specifically gave a decent accuracy on the stratified split but performed poorly while trying to classify the minority classes, namely “Serious” and “Fatal”. After being fit on the SMOTE balanced dataset, the average accuracy dropped but its performance on the more hazardous classes increased. We infer that the accuracy of

the XGBoost model is 70% which is the same as our Adaboost implementation.

With a balanced dataset, we see a good improvement in accuracy for the minority classes as well as the overall accuracy, giving a final mean accuracy of 92%.
Feature Importance is a technique that is used to determine how important an input attribute is at determining/predicting the target attribute. Each input attribute is given a score based on their importance.We plot the feature importance graph for our Random Forest model which gives us the feature importance on how likely it is for an input attribute to affect the accident severity.

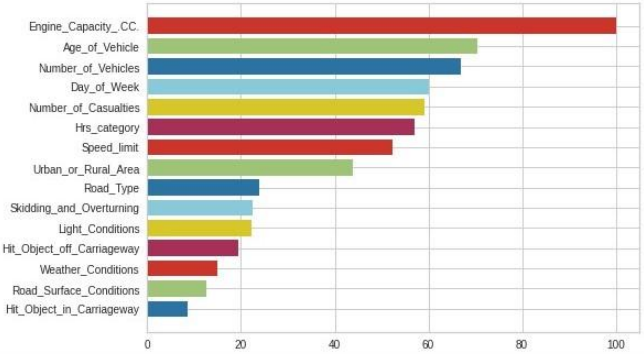

Figure 6: Feature Importance Results

It is clear from Fig[6] that the engine capacity is what affects the accident severity the most.

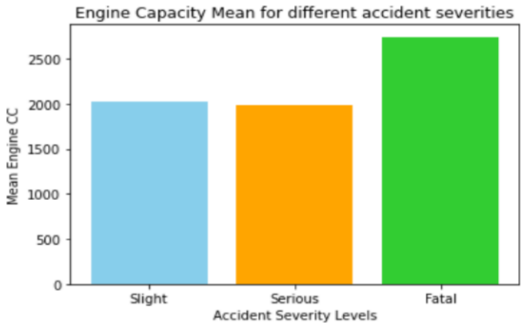

Figure 7: Mean Engine CC vs accident severities

Looking at Fig[7] graph it is clear that most fatal accidents on average have a higher Engine CC than slight or serious accidents.

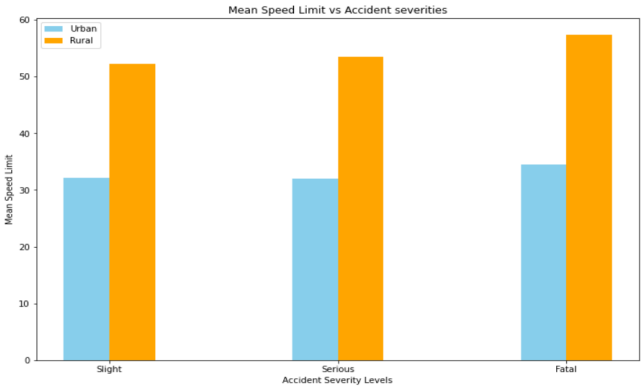

Figure 8: Mean max speed limit vs accident severities

From Fig[8] it is clear that the average of max speed limit of the roads where accidents took place varies a lot for urban and rural areas.

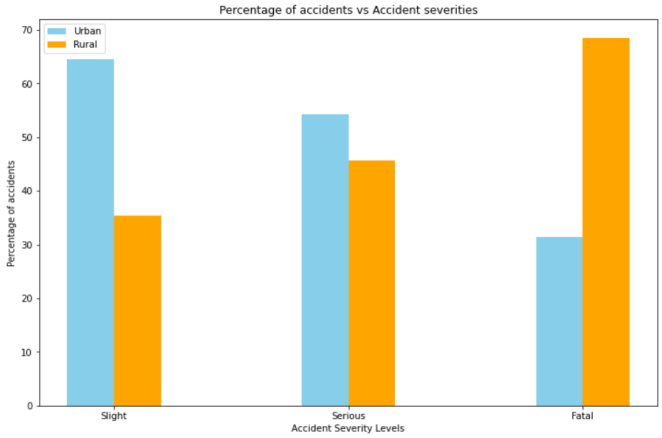

Figure 9: Percentage of accidents vs accident severities

Fig[9] is a plot of the percentage of accidents in rural/urban areas for different accident severity.

Time Series Analysis

We extract from the data, the total number of accidents that have taken place in Liverpool every month for all the years recorded in the dataset to form a time series dataset, giving us the opportunity to predict the number of accidents that can take place in the future given the current trend. ARIMA stands for Auto Regressive Integrated Moving Average. An ARIMA model is used for forecasting. To cope with seasonal behavior Seasonal ARIMA(SARIMA) was introduced.

Any time series can be split into 4 components: Trend, Seasonality, Level, and Error/Noise.

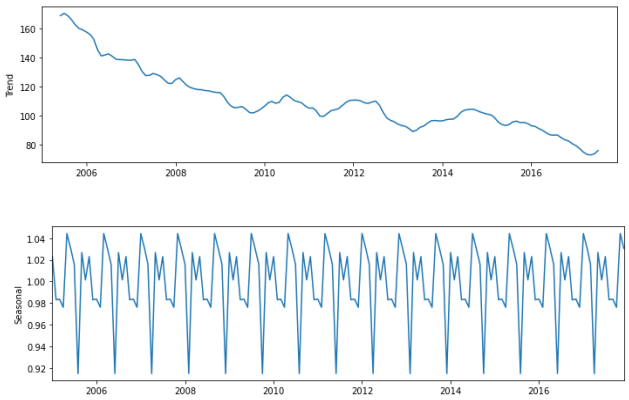

Figure 10: Seasonal Decomposition(Trend and Seasonal)

Looking at the seasonal decomposition[Fig 10] we can see that we have a consistent repeating seasonal pattern and a downwards trend pattern.

Estimation of the p,d, and q value for ARIMA is done using the auto_arima that automates the process of choosing (p,d,q) value by fitting different combinations of values and comparing their AIC(Akaike Information Criterion) values, and choosing a model with the highest log-likelihood and lowest AIC.[19]
auto_arima estimated the p,d,q values to be ARIMA(0,1,2) and SARIMAX(0,1,2)(0,1,2).

We ran tests on an ARIMA model and a SARIMA model using the statsmodels module.

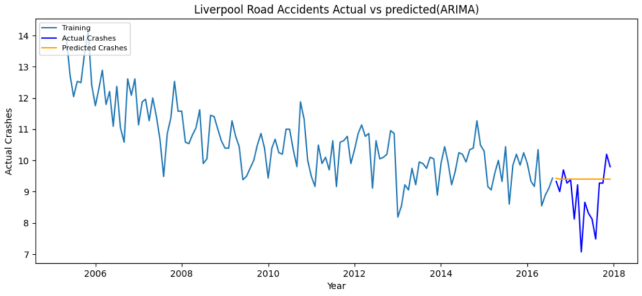

Figure 11: ARIMA Model Actual vs Predicted Crashes(For Liverpool)

```
ARIMA Model Stats
MSE: 0.9531314941719268
MAE: 0.7005260022915876
RMSE: 0.9762845354567113
MAPE: 0.08654080719531453
```

Figure 12: ARIMA Model Statistics

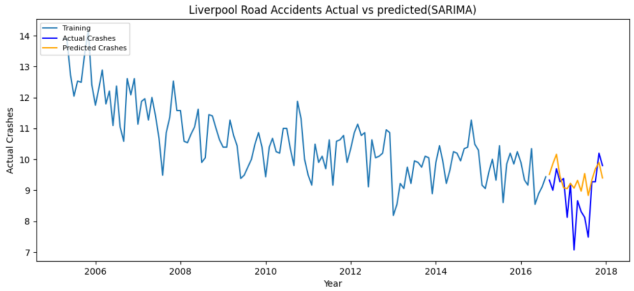

Figure 13: SARIMA Model Actual vs Predicted Crashes(For Liverpool)

```
SARIMA Model Stats
MSE: 0.6923629492031005
MAE: 0.6360298362847876
RMSE: 0.8320834989369159
MAPE: 0.07770107951080232
```

Figure 14: SARIMA Model Statistics

Looking at both the results we can say that our SARIMA model [Fig 14] did a better job on the test data with MAPE being 0.077 which means predicted values differ from the actual values by 7%.

Future Work and Conclusion

To improve upon this work, we can carry out further hyperparameter tuning on each of the models and utilise better pre-processing techniques suited to each model to achieve better accuracy. A tool can be made around KDE plots which would make it easier for authorised personnel to make use of such techniques.

Utkarsh aided the data pre-processing tasks and trained classification machine learning models on the dataset to predict accident severity. In addition to this, he worked on the hotspot analysis pipeline with KDE and inferred results for the same.

Varun aided the time series analysis tasks and trained xgboost machine learning model on the dataset to predict severity. He also worked on feature importance and plotting correlation of severity with other important attributes to draw inferences.

Harshita aided with the EDA and literature survey which helped us gauge what to work on and form our goals for this project, hotspot detection using KDE. She, in addition to this, contributed to some of the machine learning based models used for severity prediction.

Acknowledgment

We would like to thank our professor Dr. Gowri Srinivasan for giving us the opportunity to pursue this project and for her guidance through the execution of the same and express our gratitude to the Data Analytics team for guiding us throughout. We would also like to express our gratitude to the user “Thanasis” and the UK Department of Transport on Kaggle for making the extensive data of road accidents publicly available for study and analysis.

References

[1] Zimmerman K, Jinadasa D, Maegga B, Guerrero A (2015) Road traffic injury on rural roads in Tanzania: measuring the effectiveness of a road safety program. Traffic Inj Prev 16:456–460

[2] Bargegol I, Gilani VNM, Ghasedi M, Ghorbanzadeh M (2016) Delay modeling of un-signalized roundabouts using neural network and regression. Comput Res Prog Appl Sci Eng (CRPASE) 2:28–34

[3] Isabelle Thomas, Spatial data aggregation: Exploratory analysis of road accidents, Accident Analysis & Prevention, Volume 28, Issue 2, 1996, Pages 251-264, ISSN 0001-4575, https://doi.org/10.1016/0001-4575(95)00067-4.

[4] Choudhary, Jayvant & Ohri, Anurag & Kumar, Brind. (2015). Spatial and statistical analysis of road accidents hot spots using GIS

[5] V. Prasannakumar, H. Vijith, R. Charutha, N. Geetha, Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment, Procedia - Social and Behavioral Sciences, Volume 21, 2011, Pages 317-325, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2011.07.020.

[6] Khanh Giang Le, Pei Liu & Liang-Tay Lin (2020) Determining the road traffic accident hotspots using GIS-based temporal-spatial statistical analytic techniques in Hanoi, Vietnam, Geo-spatial Information Science, 23:2, 153-164, DOI: 10.1080/10095020.2019.1683437

[7] Zhang, Zhonggui & Ming, Yi & Song, Gangbing. (2020). A New Approach to Identifying Crash Hotspot Intersections (CHIs) Using Spatial Weights Matrices. Applied Sciences. 10. 1625. 10.3390/app10051625.

[8] Thakali, L., Kwon, T.J. & Fu, L. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. J. Mod. Transport. 23, 93–106 (2015). https://doi.org/10.1007/s40534-015-0068-0

[9] Sinclair, C and Das, S (2021) Traffic accidents analytics in UK urban areas using k-means clustering for geospatial mapping. https://doi.org/10.1109/sefet48154.2021.9375817

[10] Cansız OF (2011) Improvements in estimating a fatal accidents model formed by an artificial neural network. SIMULATION 87:512–522. https://doi.org/10.1177/0037549710370842

[11] Fatemeh BK, Abdolreza S, Abbas M (2012) Variable efficiency appraisal in freeway accidents using artificial neural networks—case study, cictp, multimodal transportation systems-convenient, safe, cost-effective, efficient, pp. 2657–2664. https://doi.org/10.1061/9780784412442.269

[12] Akgungor A, Dogan E (2008) Estimating road accidents of turkey based on regression analysis and artificial neural network approach, data and information technology; highways; planning and forecasting; safety and human factors. https://trid.trb. org/view/873702 29.

[13] Akgungor A, Dogan E (2009) An artificial intelligent approach to traffic accident estimation: model development and application. Transport 24:135–142. https://doi. org/10.3846/1648-4142.2009.24.135-142

[14] De Oña, J.; Mujalli, R.O.; Calvo-Poyo, F. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accid. Anal. Prev. 2011, 43, 402–411.

[15] Simoncic, M. A Bayesian network model of two-car accidents. J. Transp. Statistics 2004, 7, 13–25.

[16] Assi, Khaled & Rahman, Syed Masiur & Mansoor, Umer & Ratrout, Nedal. (2020). Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. International Journal of Environmental Research and Public Health. 17. 5497. 10.3390/ijerph17155497.

[17] Michal Bíl, Richard Andrášik, Jiří Sedoník, A detailed spatiotemporal analysis of traffic crash hotspots, Applied Geography, Volume 107, 2019, Pages 82-90, ISSN 0143-6228, https://doi.org/10.1016/j.apgeog.2019.04.008.

[18] A Gentle Introduction to XGBoost for Applied Machine Learning (machinelearningmastery.com)

[19] What is auto-arima? by Eswara Prasad(https://medium.com/featurepreneur/what-is-auto-arima-b8025c6d732d)