pasted-image.tiffpasted-image.tiff

Road-traffic accident prediction:

Predicting the Number of Casualties

Jemal Ahmed Andeta

Supervisor: Joe Steinhauer

Examiner: Göran Falkman

Master Degree Project in Informatics

with a specialization in Data Science

Spring term 2021

# Abstract

*Efficient and effective road traffic prediction and management techniques are crucial in Intelligent Transportation Systems. It can positively influence road advancement, safety enhancement, regulation formulation, and route planning to rescue living things in advance. This thesis considers road safety by predicting the number of casualties if an accident occurs using multiple traffic accident attributes. It helps individuals (drivers) or traffic offices to adjust and control their contributions for the occurrence of an accident before emerging it. To conduct the thesis, three candidate algorithms from three regression fit patterns: bagging, linear and non-linear, are proposed and evaluated. The gradient boosting machines (GBoost), Linear support vector regression (LinearSVR), and extreme learning machines (ELM) are the selected algorithm. The models are evalued using RMSE and MAE performance evaluation metrics. The GBoost achieved a better performance than the other two with a low error rate and minimum prediction interval value for 95% prediction interval. Each model was interpreted using the Shap model interpretation technique.*

**Keywords**: *Road traffic accident; Accident Casualties; Gradient Boosting Machines*; *Extreme Learning Machines; Support Vector Regression*

Table of Contents

[Abstract II](#_Toc72361010)

[List of Figures III](#_Toc72361011)

[List of Tables IV](#_Toc72361012)

[1. Introduction 5](#_Toc72361013)

[2. Background 7](#_Toc72361014)

[2.1. Road Traffic Accident 7](#_Toc72361015)

[2.1.1. Traffic Accident 7](#_Toc72361016)

[2.2. Machine Learning 8](#_Toc72361017)

[2.2.1. Gradient boosting machines (GBoost) 9](#_Toc72361018)

[2.2.2. Support Vector Regression (SVR) 9](#_Toc72361019)

[2.2.3. Extreme Learning Machines (ELM) 10](#_Toc72361020)

[2.3. Related Works 10](#_Toc72361021)

[3. Problem Definition 12](#_Toc72361022)

[3.1. Problem statement 12](#_Toc72361023)

[3.2. Research aim 12](#_Toc72361024)

[3.3. Study Contributions 13](#_Toc72361025)

[3.4. Application Area 13](#_Toc72361026)

[4. Methods and Approach 14](#_Toc72361027)

[4.1. Data Preprocessing 14](#_Toc72361028)

[4.2. Feature Selection 14](#_Toc72361029)

[4.3. Approaches 15](#_Toc72361030)

[4.4. Evaluation 15](#_Toc72361031)

[A. Root Mean Squared Error (RMSE) 15](#_Toc72361032)

[B. Mean Absolute Error (MAE) 16](#_Toc72361033)

[4.5. Performance Analysis 16](#_Toc72361034)

[A. Prediction Interval 16](#_Toc72361035)

[B. Model Interpretation 17](#_Toc72361036)

[4.6. Strategy 17](#_Toc72361037)

[4.7. Ethical Aspects 18](#_Toc72361038)

[5. Implementation 19](#_Toc72361039)

[5.1. Data Preprocessing and Balancing 19](#_Toc72361040)

[6. Results 22](#_Toc72361041)

[6.1. Feature Selection 22](#_Toc72361042)

[7. Discussion 28](#_Toc72361043)

[8. Conclusion and Future Work 30](#_Toc72361044)

[8.1. Conclusion 30](#_Toc72361045)

[8.2. Future Work 30](#_Toc72361046)

[References 32](#_Toc72361047)

[Appendix 35](#_Toc72361048)

[A. Definitions of the input and target features 35](#_Toc72361049)

[B. Hyperparameter tuning configurations 39](#_Toc72361050)

# List of Figures

[Figure 1: The interaction amongst the traffic accidence, congestion, and emergency service 8](#_Toc72360500)

[Figure 2: Sequential ensemble approach 9](#_Toc72360501)

[Figure 3: A general process of the proposed prediction model 19](#_Toc72360502)

[Figure 4: Variables with their attributes of the dataset 20](#_Toc72360503)

[Figure 5: hourly accident frequency 20](#_Toc72360504)

[Figure 6: Occurrence of accidents per response (number of casualties) value 21](#_Toc72360505)

[Figure 7: The balanced occurrence of accidents per response (number of casualties) value 21](#_Toc72360506)

[Figure 8: The performance of feature importances (ELM approach) 22](#_Toc72360507)

[Figure 9: The overall performance of feature importances 23](#_Toc72360508)

[Figure 10: A sorted version of an average line of Figure 9 23](#_Toc72360509)

[Figure 11: The performance analysis of a group of features 24](#_Toc72360510)

[Figure 12: The performance of the prediction models 24](#_Toc72360511)

[Figure 13: Analysis of the performance of models using a prediction Interval technique 25](#_Toc72360512)

[Figure 14: A Shap values of ELM model 26](#_Toc72360513)

[Figure 15: A Shap values of GBoost model 26](#_Toc72360514)

[Figure 16: A Shap values of LinearSVR model 27](#_Toc72360515)

# List of Tables

[Table 1: Definitions of the input and target features 35](#_Toc72360482)

[Table 2: Hyperparameter tuning for the GBM approach 39](#_Toc72360483)

[Table 3: Hyperparameter tuning for the ELM approach 39](#_Toc72360484)

# Introduction

A road traffic accident is a severe threat to human life and the safety of the living location. According to the World Health Organization (WHO) report, road traffic accident results in about 1.35 million deaths annually and leave between 20 to 50 million people with non-fatal injuries (WHO, 2021). In addition to the deaths and injuries, road traffic accidents also acquire an economic burden on victims and families, injuries treatment costs, loss of productivity due to disabled or killed, and loss of resources.

Numerous studies contributed to the solutions to the analysis, classification, and prediction of road traffic accidents. The contributions can be categorized into two: pre-accident and post-accident. The post-accident works are based on hand accident cases to analyze, classify, predict losses due to accidents. For example, classify severity level into slight, serious, or fatal and predict accident duration. On the other hand, the pre-accident works are based on hand road traffic attributes to predict losses of an accident in advance. The pre-accident prediction also can be work on the identification of hazard location or condition tasks. Pre-accident solutions are essential for sustainable human life and living environment than post-accident results.

Road traffic accidents may occur due to the number of traffic attributes. The attributes like weather condition, pedestrian, driver’s experience (including age and sex), vehicle status (type, number of the wheel, size, age, and speed), travel time, travel day, traffic flow, road status, traffic rule, light condition, and area (urban, rural, or junction). The impact of these attributes on the accident will be different, especially for the pre-accident prediction tasks. Road traffic accidents may be initiated by various factors mentioned as traffic attributes. The driver is the primary actor who can control or manage the occurrence of an accident. As a driver assistant technique or tool, making a situation analysis and alerting about consequences (traffic accident) may support the driver shortly will be better than nothing. Traffic office as a secondary actor can also track vehicles at high risk for traffic accidents in a real-time phenomenon. Thus, using such technologies could rescue people (travelers, drivers, roadside workers, or pedestrians) at risk.

This thesis considers a pre-accident prediction task. As a solution, a traffic accident prediction model was proposed to support individuals or groups such as drivers and traffic office by predicting the number of casualties based on various human and natural road traffic factors. For the comparative analysis of the solution, three regression algorithms, namely gradient boosting machines (GBoost), extreme learning machine (ELM), and support vector regression (SVR), were implemented. Root\_mean\_squered\_error (RMSE) and mean\_absolute\_error (MAE) performance evaluation metrics were used with a prediction interval analysis technique to handle the uncertainty aspect of the predicted value for a new observation.

The remaining part of the thesis divides into seven sections. The background information, problem statements, methods and approaches, implementation, results, discussion and conclusion, and feature works are the report's main sections.

# 

# Background

This section discusses the background information in the context of the problem and prior solutions to this thesis, such as road traffic accident, machine learning, gradient boosting machine (GBoost), support vector regression (SVR), extreme learning machine (ELM), and finally related works.

## Road Traffic Accident

A road traffic accident is defined as “*an accident which is a random, and multi-factor event always preceded by a situation in which one or more road users have failed to cope with the road environment*”(Gebru, 2017). Road traffic accidents can be expressed as events or accidents like a car accident, vehicle collision, vehicle crash, pedestrian, road debris, animal or other obstruction like a pole, building, or tree. Mostly, consequences of accidents are quantifiable in losses like fatality, injury, resource in number, or money. Human life is the primary matter, and we need to work for safe rather than resources.

Researchers have proposed solutions for both scenarios: safe and resource management and control. Prior studies showed what researchers suggested solutions for safe strategy (as pre-accident) and resource control (as post-accident) are explained in the following subsection.

## Traffic Accident

Traffic post-accident solutions reflect how traffic accidents can be managed and controlled efficiently and effectively based on an existing accident. For example, classifying or evaluating the level of seriousness of a traffic accident, predicting or analyzing the response time of an emergency service regarding an accident, and analyzing the level of traffic congestion due to traffic accidents. The interactions of the three situations are shown in *Figure 1* below (Haynes *et al.*, 2019).

Figure 1: The interaction amongst the traffic accidence, congestion, and emergency service

Traffic Accidence

(Level of seousness)

Traffic Congestion

(Level of traffic state)

Emergency Service

(Response time)

Classifying road traffic accidents is also valuable for identifying causes of the accident to manage or solve them in advance. Based on traffic attributes like accident type, light condition, road type, and road characteristics, traffic accidents are classified into five categories: occurred near or inside curve road with one injury, straight road with one injury, straight road with two injuries, mostly non-highway roads with two injuries, and straight or slightly curve road with mostly one injury (Nandurge and Dharwadkar, 2017). The categories will help to investigate situations to be considered and make solutions before occurring accidents.

## Machine Learning

Machine learning is an area of discipline within AI that concerns the study of computer algorithms that improve automatically through experience. Formally definition of Machine Learning by Tom M. Mitchell reads as follows (Mitchell, 1997).

“*A computer program is said to learn from experience* ***E*** *with respect to some class of task* ***T*** *and performance measure* ***P****, if its performance at tasks in* ***T****, as measured by* ***P****, improves with experience* ***E****.*” - Tom M. Mitchell

For this thesis, experience **E** is from the road traffic accident historical data collected for ten (2005-2014) years in Great Britain. Task **T** predicts the number of casualties from accident variables that return a continuous output variable (projected number of casualties). The performance measure **P** is related to the prediction error.

Machine learning classifies into three types: supervised, unsupervised, and reinforcement learning. Supervised learning generates a model to predict an output based on labeled historical observations of that output. Furthermore, based on the type of problem, i.e., the output value (discrete or continuous), supervised learning classifies into a regression or classification problem, gives a continuous or discrete (commonly known as a class) output value, respectively. Regression works by fitting the input data, whereas classification separates the input data into the most promising class. Unsupervised learning is different from supervised one. It takes an unlabeled input feature and returns newly organized data by grouping, clustering, or organizing based on similarity (pattern) measures. Reinforcement learning is again different from both supervised and unsupervised learning. It learns by behavior-driven a continually updates its state and reward.

In this thesis, the problem is a regression task. The final learning model should predict the number of causalities based on accident input features. We have focused on the regression side of learning algorithms. Thus, three regression algorithms are selected and described below: GBoost, SVR, and ELM.

### Gradient boosting machines (GBoost)

A GBoost is a popular machine learning algorithm that builds an ensemble of shallow trees sequentially (Greenwell, 2021). Each current tree learns and improves on the previous one, where each new tree in the order tries to fix up the error made on the last one, shown in *Figure 2* (Greenwell, 2021).

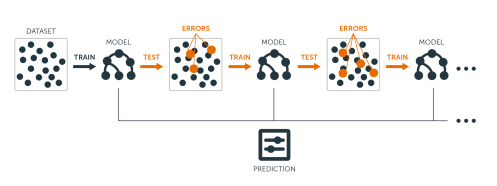


Figure 2: Sequential ensemble approach

### Support Vector Regression (SVR)

SVR is another version of the support vector machine (SVM) to use as a regression algorithm. SVR has achieved a better performance than an artificial neural network (Ahmad *et al.*, 2020). SVR may not fit with more than 10000 observations. Instead, a LinearSVR version can handle a prediction task with a large dataset or observations (scikit-learn developers, 2021).

### Extreme Learning Machines (ELM)

ELM is another and the third selected algorithm to be implemented and evaluated with the two baseline approaches. ELM is a training method whose training speediness is very fast, in a single hidden feed-forward artificial neural network (SLFN)(Ertuğrul, 2020; Huang, Zhu and Siew, 2006).

## Related Works

In this subsection, significant related works are expressed with two contexts: the problem and the implementation approach.

In the context of the problem, Li, Zhao and Liu (2020) have proposed an accident prediction model(Li, Zhao and Liu, 2020). The model can predict the number of accidents to occur. They conceived on evaluating the weightiness of each feature on accident and then a prediction process for multiple interrelated variables using five different approaches, including the one proposed by the authors: grey model, multivariate grey model, linear regression, backpropagation neural network, and the optimal multivariate grey model. The proposed optimal multivariate grey model approach outperforms better or low MAPE, MAE, and RMSE performance values using two traffic accident statistical data (Zhe Jiang province and Chong Qing city in China). Similarly, Jiber *et al.*(2020) also proposed (Jiber *et al.*, 2020)

Another important traffic accident classification model to predict the level of crash severity was proposed (Assi *et al.*, 2020) using limited traffic attributes (vehicle and road conditions) from a crash dataset of Great Britain and two machine learning approaches (support vector machine and feed-forward neural networks) with and without fuzzy c-means clustering. The model can predict either a severe or non-severe crash for the primary purposes of giving initial information about the effect and prepare for handling for the victims. The support vector machine with fuzzy c-means approach outperforms better results than the other three.

In another context, the implementation approach, Li *et al.* (2016) has been proposed a combined approach using an error correction model and ELM to predict wind power generation (Li *et al.*, 2016). It handles the impact on the unit commitment due to the variation of the wind generation. The ELM approach achieved promising results in terms of accuracy via the normalized RMSE.

To conclude the section, the prediction of the number of casualties is not addressed. The ELM approach may inherit a better prediction result.

# Problem Definition

## Problem statement

Based on the articulations provided in the previous section, traffic accidents can be managed and controlled differently for different purposes like traffic congestion, response time estimation, traffic accident duration, and level of accident seriousness. As a pre-accident solution, some studies have been proposed solutions: for example, to predict the number of accidents in a time-series fashion (Jiber *et al.*, 2020) and predict the occurrence of traffic accidents (Li, Zhao and Liu, 2020). However, such solutions can not be state or assured the number of losses in terms of the number of injuries and fatalities.

In this thesis, the prediction of losses regarding the number of casualties of any road traffic accident task is considered. The evaluation of multiple traffic accident attributes for a better prediction performance is another crucial aspect of the study.

## Research aim

This thesis aims to predict the number of casualties in a traffic accident based on traffic attributes. It is a pre-accident task and needs to investigate the impact of each feature on the prediction model.

Thus, to address the intended aim, the following two research questions are formulated:

1. *Is it possible to predict the number of casualties of a traffic accident?*
2. *Are all traffic input attributes relevant and essential to predict the number of casualties of a traffic accident?*

Analyzing and evaluating each attribute’s weight for accident loss predictions is one vital task to answer both research questions. It helps to characterize any accident and consider weighty features at the time of model development.

Practically, to address the two research questions, the following objectives are framed:

* *To examine and make a critical decision on data sampling*
* *To rank, the traffic attributes based on their impact on the output attribute*
* *To model and evaluate the intended prediction task*
* *To analyze the performance of the model*

## Study Contributions

Contributions of the study can be expressed with three points: regression algorithms, dataset balancing, and novelty of the prediction task. From the point of regression algorithms, the SVR, GBoost, and ELM are considered to analyze the performance of the intended prediction model. The impact of these algorithms during the fit and testing model is examined.

Often, data balancing is mainly applicable for datasets with classification tasks. The balancing task promotes the number of observations from minor classes and de- promotes the number of observations from major classes. It helps to represent features from each class in a balanced fashion during the implementation of a classification model. In this study, in regression task, response values are considered as classes and applied the promotion and de-promotion of the number of observations per unique response value.

An aspect of the prediction task, predicting the number of casualties from road traffic accident attributes, is unique and examined as a new task.

## Application Area

Nowadays, the production or use of self-driving cars is increasing. Such prediction models can be embedded in the patterns, minimizing crashes and supporting the device (car). Drivers also can get support by using as a support system. Traffic office can track individuals who are at risk in real-time situations.

# Methods and Approach

In this chapter, the requirements and techniques that are used to execute the expected solution are identified with the alignment of the objectives of the thesis. Multiple algorithms from different patterns help to evaluate the main goals of the study. Three separate regression algorisms are applied to predict consequences in terms of the number of casualties from a traffic accident. The chapter has five subsections: data preprocessing, feature selection, implementation approach, evaluation metrics, and model interpretation with validity analysis scenario. Each subsection has an aim to achieve at least one objective of the study.

## Data Preprocessing

A road accident historical data collected in Great Britain (GB) from 2005 to 2014 (Fedit, 2021) contains three files. As shown in *Figure 6*, 29 unique response values from the x-axis are stated, and the response values are not in an approximate occurrence. It can affect the representation of the response values (number of casualties) during training and testing models.

These 29 unique values can be considered as classes and can balance them using appropriate techniques. The oversampling for minor classes and undersampling for major classes have been achieved promising results in classification tasks (Hacibeyoglu and Ibrahim, 2018; Hilario *et al.*, 2018). Thus, both oversampling and undersampling techniques are considered for the minor and major response values. Based on *Figure 6,* the number of casualties 1 and 2 are major response values, whereas the remaining are under minor values. The oversampling technique can be achieved by extracting synthetic observations from neighborhood observations. On another side, the undersampling approach can be achieved by taking a subset from observations.

## Feature Selection

All accident cause-related features will not crucial for better performance. In this case, the essential features for the response feature are selected. In analyzing the relationship between input features and response features, feature importance analysis is considered by weightiness the impact of each feature on a prediction model. It helps investigate the effect and contribution of each feature to predict the number of casualties from traffic accident attributes.

For this task, a permutation importance evaluation technique was used. The method checks the performance of each feature by taking the difference of two errors that came from two prediction models. One model is with the feature itself, and another is without the feature. In this case, the higher the error rate decreases, the feature will be essential than all others. This permutation technique evaluates the contribution of each feature with all other features. For example, if there are five features, each feature will have five different importance weights. The final importance weight can be an average of the weights.

Thus, the second objective of the thesis was addressed using the permutation feature importance technique. It gives a rank of features based on individual average permutation importance weight value. Another significant decision is selecting algorism for the prediction model during the impact analysis. This evaluation task was conducted with three different regression algorithms mentioned in the following subsections to get a better generalizable impact analysis.

## Approaches

As described in the problem statement section above, road traffic information collects from diverse sources and needs an efficient method for processing and extracting useful information, especially in real-time applications. The intended road traffic accident prediction requires real-time processing and prediction. The LinearSVR, ELM, and GBoost algorithms are selected based on three nature of regression algorithms: linear, non-linear, and bagging techniques, respectively. The decision will help to examine the prediction model from different perspectives.

## Evaluation

Based on prior works stated in the related work section, MAPE, MAE, and RMSE evaluation metrics were used. In this work, the two standard prediction evaluation techniques: MAE and MSE, are used to address a comparative analysis of the results. Furthermore, the prediction results are also analyzed using the prediction interval technique to manage the uncertainty aspect of the predicted value from a new observation.

### Root Mean Squared Error (RMSE)

RMSE is the standard deviation of the residuals. It is the average of the squares of the difference between the actual (true) value and the model prediction value. The mathematical formula for RMSE is expressed below, y denotes the actual value, and ŷ denotes the predicted value at each observation k and the number of observations n.

|  |  |
| --- | --- |
|  | ( ) |

More significant errors have an unreasonably large effect due to RMSE being proportional to the squared prediction error, making RMSE sensitive to outliers (Chai and Draxler, 2014).

### Mean Absolute Error (MAE)

MAE is an arithmetic average of the absolute errors, the difference between the model prediction and the actual value. The mathematical formula for MAE is expressed below, y denotes the actual value, and ŷ denotes the estimated value at each observation k and the number of observations n.

|  |  |
| --- | --- |
|  | ( ) |

## Performance Analysis

Regression models predict values for a new observation. Such models can not give clues regarding the certainty of predicted values and the impact of observation values on the model. The last objective of the study requires model analysis to address the uncertainty of predicted values and the effect of values from observations on the model during the training phase.

### Prediction Interval

A prediction interval is a range of values that is likely to contain the predicted value of a new observation given specified settings of the predictors. For example, for a 90% prediction interval of [15 26], we can be 90% confident that the new observation will fall a certain probability in the range [15 26]. The prediction interval analysis task will handle the uncertainty aspect of the predicted value for a new observation. The mathematical formula for prediction interval (The Pennsylvania State University, 2021) is expressed below.

|  |  |
| --- | --- |
|  | ( 3 ) |

Where ŷi is the predicted value for a new observation xi, the t(1-⍺/2,n-2) is t-multiplier with an n-2 degree of freedom due to MSE denominator n-2, and is the standard error of the fit.

### Model Interpretation

The prediction models used by this thesis are kind of complex models that can not be easy to interpret. A Shapley Additive exPlanations (Shap) model interpretation technique is a valuable method to analyze and interpret a prediction model (Lundberg and Lee, 2017). The Shap can diagnose the contribution of the values of each feature on the model and sort out the importance of each feature in expressive orders. Therefore, the Shap technique is accepted for the interpretation task of the models.

## Strategy

Analyzing and justifying likely validity threats is a critical process in enquiring the validity of the results of an experiment (Feldt and Magazinius, 2010). Validity is acceptable if it is possible to generalize the results to a broader group of interest or valid for the group that conducted the study. Though, this does not imply that adequate validity is the most general validity. Suppose an organization made a study and it was designed to only answer questions relevant to them. In that case, the results only need to be appropriate for them (Wohlin *et al.*, 2012).

The low statistical power, fishing and error rate, reliability of measures, and conclusion validity threats (Wohlin *et al.*, 2012) are considered to mitigate them by taking explicit attention during overall processes. A low statistical power threat was alleviated by applying down-sampling major response values and over-sampling for the minor response values. A fishing and error rates threat were helped by comparing the prediction performance using three prediction models. A hyperparameter tuning technique was applied for ELM and GBoost with limited parameters to enhance the performance. Reliability of measures threat was helped by guaranteeing consistent data splits into test and train throughout the implementation of the model. Lastly, the conclusion threat was handled by analyzing the performance of each model using prediction intervals and model interpretation using the Shap.

## Ethical Aspects

The only data source for implementing the prediction models is a public dataset found from a public repository (Fedit, 2021). means that no need to consent for research purposes.

# Implementation

A general process of the proposed prediction model is shown in *Figure 3* below. The subsections of this section elaborate on the main functions of *Figure 3*.

Data Preprocessing, Balancing and Sampling

Feature Selection

Ranked Features

Model and Evaluation

Prediction Interval Analysis

Model Interpretation

Figure 3: A general process of the proposed prediction model

## Data Preprocessing and Balancing

The dataset contains three separate files with a specified number of observations on each: accidents (1,640,597), causalities (2,216,720), and vehicles (3,004,425). The detail of the files is shown in *Figure 4*.

Most of the variables initially were in categorical form and transformed into a numerical format (with a data dictionary). The dataset does not have traffic density-related information. The thesis intends to predict the number of causalities based on input variables. The input variables should be cause-related variables from each variable. Thus, among all 67 attributes of the dataset, 29 are selected as predictor (input) features and an additional one response feature (number of casualties), as described in the

Appendix section in *Table 1*. The remaining variables related to post-accident characteristics will not be significant for the prediction task.



Figure 4: Variables with their attributes of the dataset

Among 29 features, the time feature is in the form of (hh: mm). The time feature on each observation is processed and transformed into a 0 or 1 value based on the boundary or range from the hourly average frequency of accidents, taken transformation experience from a prior study (Xia, Nan and Xu, 2017). As shown in *Figure 5* below, Value 0 means that if the observation is below the average time, otherwise 1. Hours between 8 and 20 from the weekday category and hours between 10 and 20 from the weekend side, both inclusive are tended to more risky time intervals. Thus, the observations whose time feature under the specified intervals transformed into 1 otherwise 0.

|  |  |
| --- | --- |
|  |  |

Figure 5: hourly accident frequency

The dataset has another issue, i.e., an imbalanced dataset. As shown in *Figure 6*, among 29 unique response values, the accident frequency is very high at 1 (covers about 76.5% of the observations), moderate at 2 (about 16.2 %), and openly more minor the remaining 27 response values.

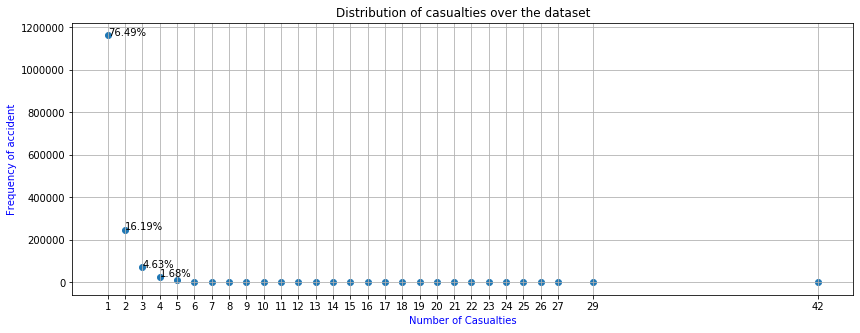


Figure 6: Occurrence of accidents per response (number of casualties) value

The (Hacibeyoglu and Ibrahim, 2018) and the oversampling and under-sampling are characterized (Hilario *et al.*, 2018) and these

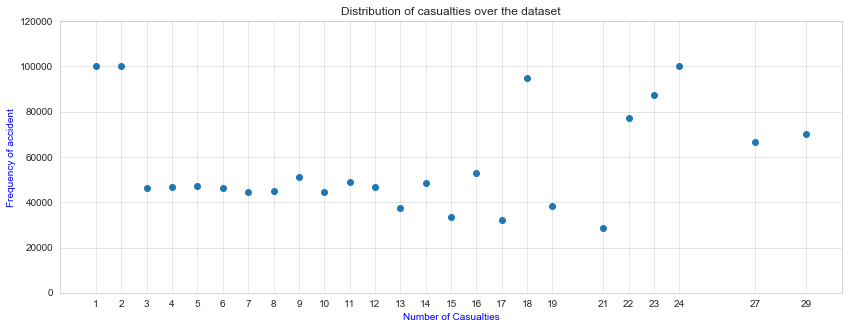


Figure 7: The balanced occurrence of accidents per response (number of casualties) value

The remaining processes of *Figure 1* are expressed in the next section.

# Results

## Feature Selection

To evaluate the performance of each feature, a permutation importance measurement technique was implemented for each model and obtained results. *Figure 8* below shows the importance score for the ELM model. The variation of importance score offers a good performance that is showing small outliers.

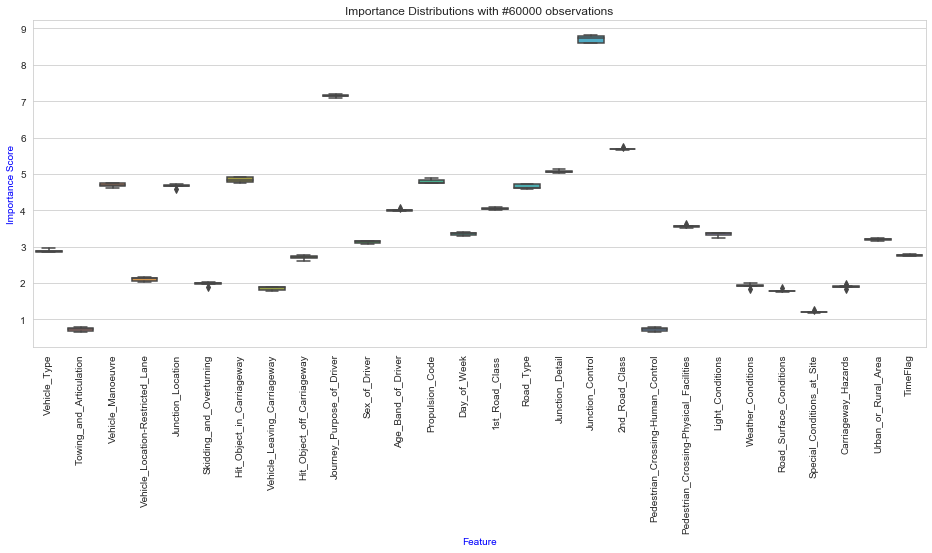


Figure 8: The performance of feature importances (ELM approach)

Similarly, the permutation importance technique was applied for the two remaining models. A summary from each model and an average of them is calculated, as shown in *Figure 9*.

A sorted version of the average of the three models is depicted in *Figure 10*. The figure shows sorted features based on individual importance scores.

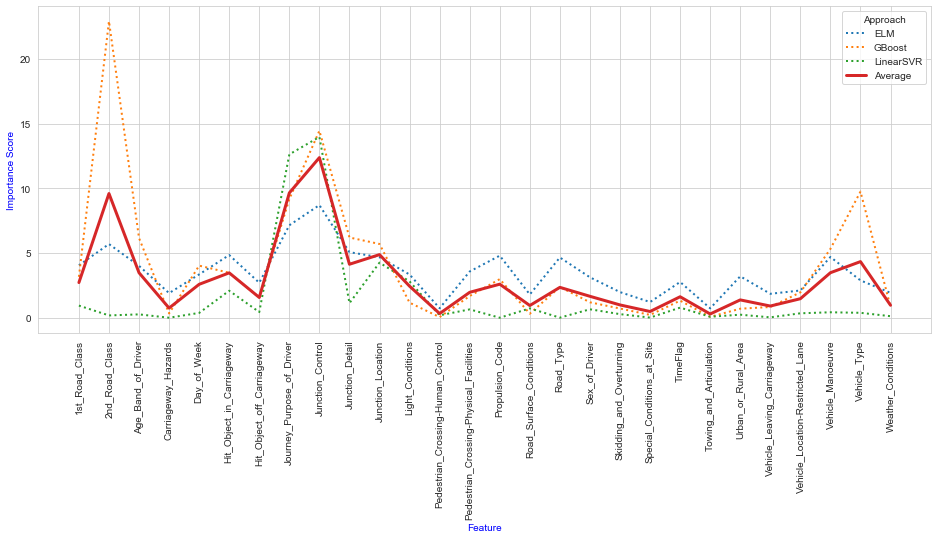


Figure 9: The overall performance of feature importances

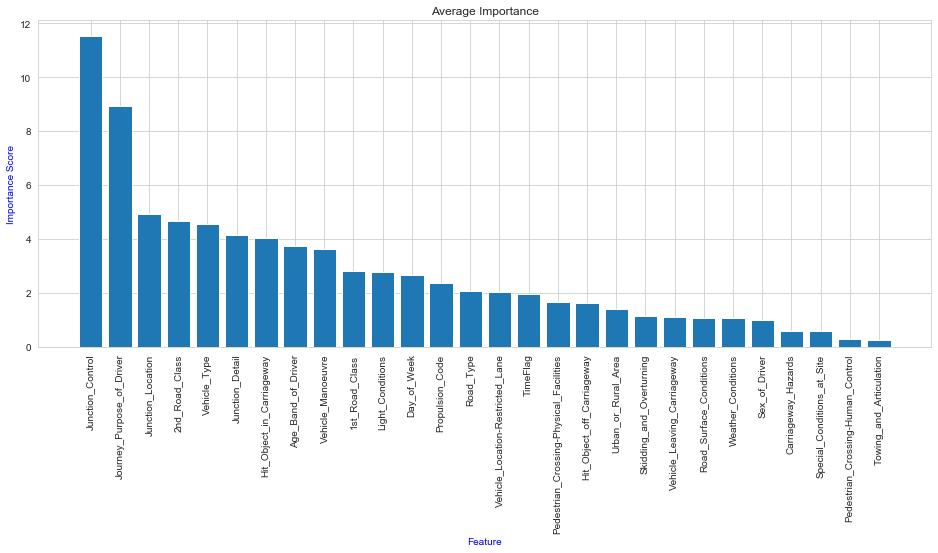


Figure 10: A sorted version of an average line of Figure 9

*Figure 10* shown above can not decide which parameters are essential or not for a better prediction performance. Therefore, the analysis of groups of features from left to right is implemented, shown in *Figure 11*. The figure shows a clue that the prediction models can give better results or minimum error rates with RMSE, MAE, and data version. So, features from Junction\_Control to Urban\_Rural\_Area are relatively more important. Because adding extra features Urban\_Rural\_Area increasing the error rate in most of the combinations of metrics smoothly.

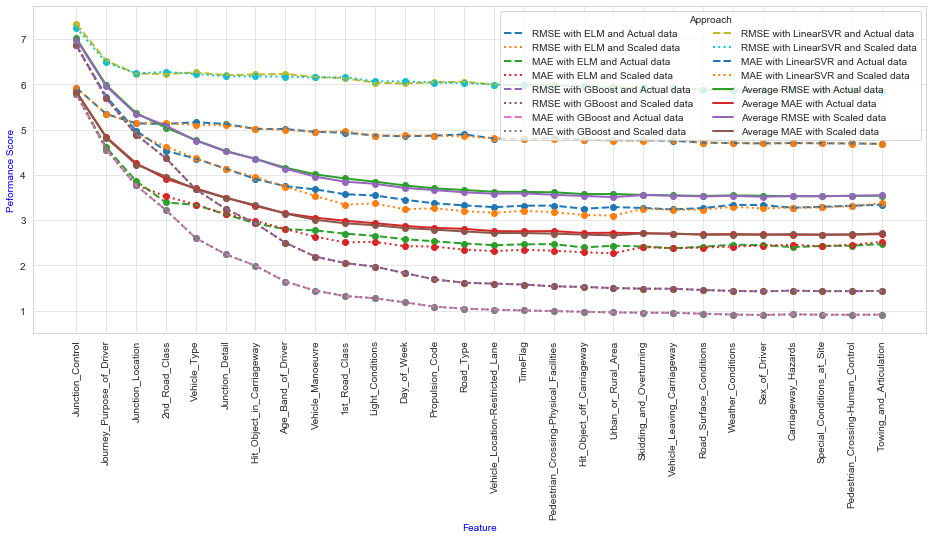


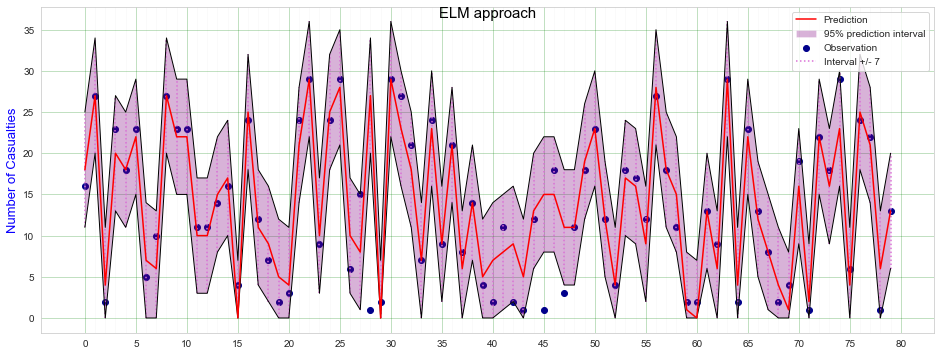
Figure 11: The performance analysis of a group of features

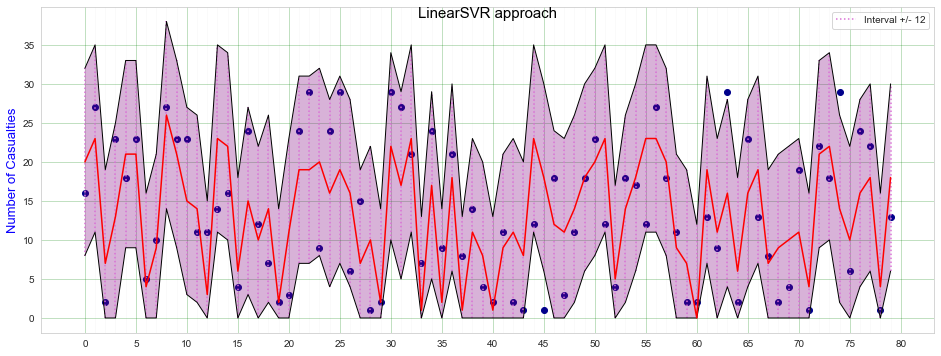
The essential features are identified, 19 features from Junction\_Control to Urban\_Rural\_Area, as depicted in *Figure 11* above. Using these selected features, the performance of prediction models in terms of error rate and time complexity examined, as showing in *Figure 12* below.



Figure 12: The performance of the prediction models

*Figure 12* shows the performance of the models. But it can not be helpful in terms of the uncertainty of a predicted value for a new observation and interpret the impact of features values on the models. *Figure 13* depicted below shows how far a predicted value from the expected value with up or downrange.





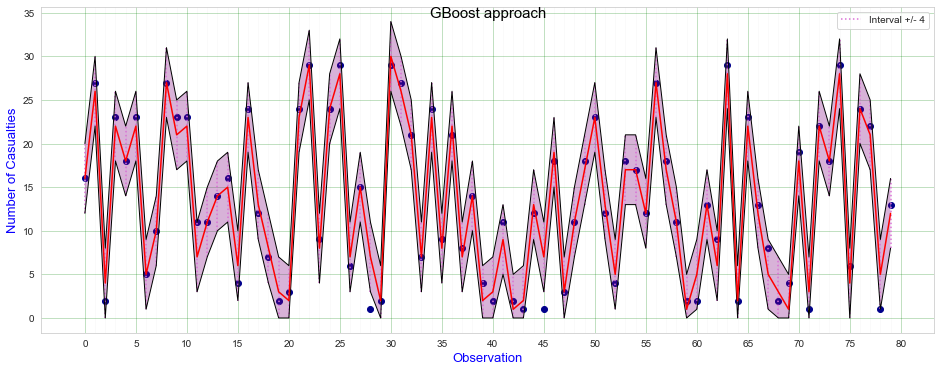


Figure 13: Analysis of the performance of models using a prediction Interval technique

Model interpretation is another essential task that can show contributions of values from each feature on the models. The Shap values of each model are shown in *Figure 14*, *Figure 15*, and *Figure 16* in descending order of the importance of features.

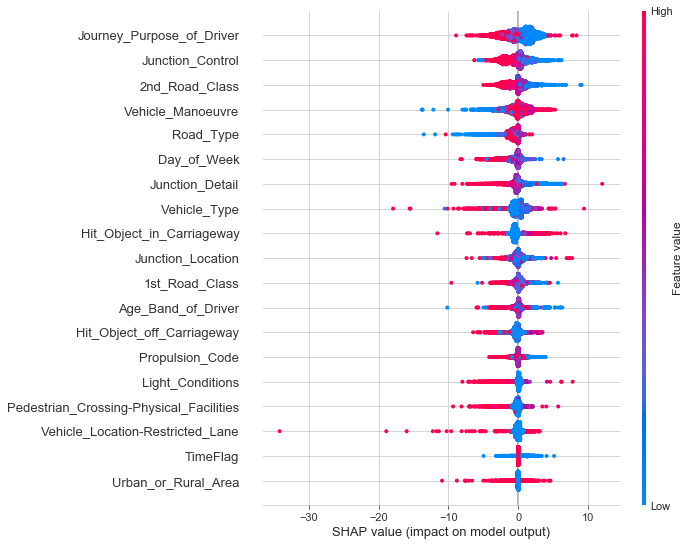


Figure 14: A Shap values of ELM model

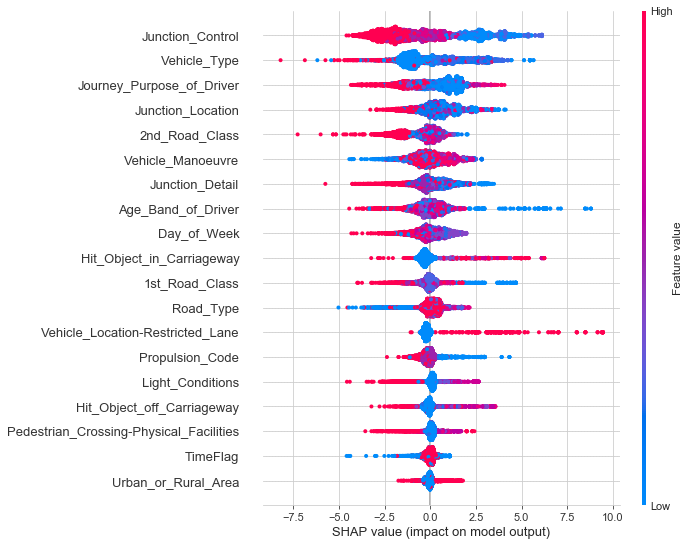


Figure 15: A Shap values of GBoost model

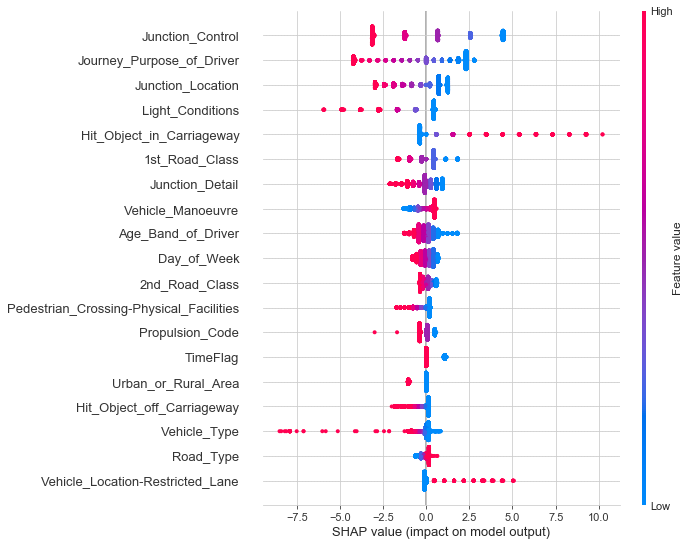


Figure 16: A Shap values of LinearSVR model

# Discussion

This study focused on two primary aspects: predict the number of casualties and evaluate the impact of road traffic accident attributes on the prediction models. Four objectives support these two primary aspects: to analyze the dataset, rank the relevant traffic accident attributes, model and evaluations, and perform performance analysis.

The dataset was analyzed and transformed into a balanced version (*Figure 7*) from imbalanced observations (*Figure 6*) per response value. Another critical decision made on the missing values, which primarily denoted by -1, and features which contain -1 may affect the performance of the interpretation of the model. Thus, observations that have -1 on any features are ignored. In this case, some known important features are ignored from relevant feature lists by the threshold feature, i.e., Urban\_Rural\_Area, as shown in *Figure 11*. For example, weather\_condition might be an essential factor for a traffic accident.

The second objective of the study is feature ranking. The features are ranked based on their permutation performance on the prediction model, as shown in *Figure 10*. A relatively clear boundary for selecting essential features based on different settings is obtained, as depicted in *Figure 11*. Based on the average performance shown in *Figure 11*, the Junction-Control feature is an essential feature over others. About 19 features are selected as relevant features. The remaining eight features are contributing inversely in terms of error rate. The more input features will burden the performance of the intended models. This feature elimination also contributes to the aspect of fit and prediction complexity.

The third objective of the study is addressed by applying three regression algorithms with the selected features. The performance of the models is shown in *Figure 12*. It indicates the error rate and time complexity. The GBoost approach achieved the worst fit-time and a better error rate. This will help to say bagging techniques, for example, GBoost may perform the worst in terms of fit-time than others like linear and non-linear approaches. In this scenario, LinearSVM is from linear and ELM also from non-linear sides. The fit complexity may not important for the overall performance of a model. For example, the GBoost approach achieved better performance and can be chosen as the best candidate. Because model fit is a one-time task, and only prediction complexity will be necessary and be considered. Thus, GBoost is a better approach than the others two.

The final objective is model interpretation and analysis. Based on the prediction interval analysis shown in *Figure 13*, the GBoost approach achieved a minimum prediction interval value,i.e., 4. In this case, a 95% prediction interval, a predicted value will be up or down by four (4) from an expected value for a new observation. It achieved the uncertainty of the predicted value with a found interval value. For the remaining two candidates are 7 and 12, which showing the performance of the approaches is low.

Besides, *Figure 14*, *Figure 15*, and *Figure 16* show the interpretation of the model with sorted features based on individual contribution to the model. For example, the first feature, Junction\_Control, in *Figure 16* is showing that the feature has five unique values, as tabulated in *Table 1*. The first value contributed more (the red color), but the last or fourth value contributes less than all others.

Generally, the answer to the first research question will be yes. Predict the number of casualties is achieved by examining the series objectives mentioned above to evaluate and analyze the effects with candidate models. The answer to the second question will be no. Among 28 features, 19 of them are selected as relevant features. The remaining nine features are not appropriate for the intended prediction task.

# Conclusion and Future Work

## Conclusion

Examining and evaluating the performance of input features on the prediction model is an essential task to eliminate non-significant features and then get better performance. Besides model evaluation in terms of performance, analysis of fit and prediction time is also important to decide which approach can be better in large-scale usage. For example, an extensive fit time can not be a big issue for prediction usage because it is a one-time operation, unlike the prediction operation.

The GBoost took a maximum fit time than ELM and LinearSVR approach. Once the GBoost model fit, the prediction complexity gives a promising performance in terms of error rate, prediction time, and a less 95% prediction interval (i.e., 4). Based on the overall performance of the candidate models, the bagging (GBoost) technique is the most expensive one in fit-time than linear (LinearSVR) and non-linear (ELM) approaches. But the error rate is contrary that GBoost achieved a better performance than ELM, and ELM achieved a better than LinearSVR.

The impact of traffic accident attributes on the prediction model is shown visibly that the variability of effects. Means that all attribute are not relevant for the regression algorithms to predict the number of the casualties. Yes, the prediction of the number of casualties can be achieved using road traffic accident attributes to address the first research question.

## Future Work

The study contributed to the three points: data balancing, a novel prediction task, and evaluation of three regression algorithms from different fit patterns. The open issue of the study can be summarized with three perspectives: dataset itself, traffic accident attributes, and the hyperparameter tuning. The minimum response value is 1, and the models can not detect an accident with 0 number of casualties. Thus, the prediction model can be enhanced using a dataset containing 0 number of casualties and accident-free observations. After evaluating feature importance, some crucial features were ignored, for example, weather\_condition, which can have a high role in traffic accidents. If the weather condition in numeric format rather than categorical form might be an essential feature group. This will helps to improve the performance of the prediction and the exhaustiveness of the study of feature importance. Thirdly, a hyperparameter tuning task for the GBoost and ELM algorithms was applied using a grid-search algorithm with a maximum of three parameters, as tabulated in *Table 2* and *Table 3*. The mentioned regression algorism has more parameters that can be tuned for better performance.

# References

Ahmad, M. S. *et al.* (2020) ‘A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens’, *Construction and Building Materials*, 248, p. 118475. doi: 10.1016/j.conbuildmat.2020.118475.

Assi, K. *et al.* (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, p. 5497. doi: 10.3390/ijerph17155497.

Chai, T. and Draxler, R. R. (2014) ‘Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature’, *Geoscientific Model Development*, 7(3), pp. 1247–1250. doi: 10.5194/gmd-7-1247-2014.

Ertuğrul, Ö. F. (2020) ‘A novel randomized machine learning approach: Reservoir computing extreme learning machine’, *Applied Soft Computing*, 94, p. 106433. doi: 10.1016/j.asoc.2020.106433.

Fedit, B. (2021) *UK Accidents 10 years history with many variables | Kaggle*. Available at: https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables (Accessed: 7 April 2021).

Feldt, R. and Magazinius, A. (2010) *Validity Threats in Empirical Software Engineering Research - An Initial Survey.*, p. 379.

Gebru, M. K. (2017) ‘Road traffic accident: Human security perspective’, *International Journal of Peace and Development Studies*, 8(2), pp. 15–24. doi: 10.5897/IJPDS2016.0289.

Greenwell, B. B. & B. (2021) *Chapter 12 Gradient Boosting | Hands-On Machine Learning with R*. Available at: https://bradleyboehmke.github.io/HOML/gbm.html (Accessed: 7 April 2021).

Hacibeyoglu, M. and Ibrahim, andMohammed H. (2018) ‘The Effect of Over-sampling and Undersampling Techniques in Medical Datasets’, in *International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES’18)*. *International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES’18)*, Safranbolu, Turkey. Available at: http://indexive.com/Paper/Download/96/the-effect-of-over-sampling-and-under-sampling-techniques-in-medical-datasets (Accessed: 25 April 2021).

Haynes, S. *et al.* (2019) ‘Data Analytics: Factors of Traffic Accidents in the UK’, in *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, pp. 120–126. doi: 10.1109/DESSERT.2019.8770021.

Hilario, A. F. *et al.* (2018) *Learning from Imbalanced Data Sets*. Springer International Publishing. doi: 10.1007/978-3-319-98074-4.

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. (2006) ‘Extreme learning machine: Theory and applications’, *Neurocomputing*, 70(1), pp. 489–501. doi: 10.1016/j.neucom.2005.12.126.

Jiber, M. *et al.* (2020) ‘Road Traffic Prediction Model Using Extreme Learning Machine: The Case Study of Tangier, Morocco’, *Information (Switzerland)*, 11. doi: 10.3390/info11120542.

Li, Z. *et al.* (2016) ‘Short-term wind power prediction based on extreme learning machine with error correction’, *Protection and Control of Modern Power Systems*, 1. doi: 10.1186/s41601-016-0016-y.

Nandurge, P. A. and Dharwadkar, N. V. (2017) ‘Analyzing road accident data using machine learning paradigms’, in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 604–610. doi: 10.1109/I-SMAC.2017.8058251.

Pham, B. T. *et al.* (2020) ‘Extreme Learning Machine Based Prediction of Soil Shear Strength: A Sensitivity Analysis Using Monte Carlo Simulations and Feature Backward Elimination’, *Sustainability*, 12(6), p. 2339. doi: 10.3390/su12062339.

scikit-learn developers (2021) *sklearn.svm.SVR — scikit-learn 0.24.2 documentation*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html?highlight=svr#sklearn.svm.SVR (Accessed: 1 May 2021).

The Pennsylvania State University (2021) *3.3 - Prediction Interval for a New Response | STAT 501*, *PennState: Statistics Online Courses*. Available at: https://online.stat.psu.edu/stat501/lesson/3/3.3 (Accessed: 7 April 2021).

WHO, W. (2021) *Road safety*. Available at: https://www.who.int/westernpacific/health-topics/road-safety (Accessed: 9 February 2021).

Wohlin, C. *et al.* (2012) ‘Experimentation in Software Engineering’, in, pp. 123–151. doi: 10.1007/978-3-642-29044-2\_10.

Xia, X., Nan, B. and Xu, C. (2017) ‘Real-Time Traffic Accident Severity Prediction Using Data Mining Technologies’, in *2017 International Conference on Network and Information Systems for Computers (ICNISC)*. *2017 International Conference on Network and Information Systems for Computers (ICNISC)*, pp. 242–245. doi: 10.1109/ICNISC.2017.00059.

# 

# Appendix

## Definitions of the input and target features

Table 1: Definitions of the input and target features

| Feature | Type | Numerical Value/Label | Role |
| --- | --- | --- | --- |
| Number\_of\_Casualties | Integer | 1 to 42/ Continuous | Target |
| Journey\_Purpose  of\_Driver | Categorical | 1/Journey as part of work  2/Commuting to/from work  3/Taking pupil to/from school  4/Pupil riding to/from school  5/Other  6/Not known  15/Other/Not known (2005-10) | Input |
| Propulsion\_Code | Categorical | 1/Petrol  2/Heavy oil  3/Electric  4/Steam  5/Gas  6/Petrol/Gas (LPG)  7/Gas/Bi-fuel  8/Hybrid electric  9/Gas Diesel  10/New fuel technology  11/Fuel cells  12/Electric diesel | Input |
| Junction\_Control | Categorical | 0/Not at a junction or within 20 meters  1/Authorised person  2/Auto traffic signal  3/Stop sign  4/Give way or uncontrolled | Input |
| Junction\_Location | Categorical | 0/Not at or within 20 metres of junction  1/Approaching junction or waiting/parked at junction approach  2/Cleared junction or waiting/parked at junction exit  3/Leaving roundabout  4/Entering roundabout  5/Leaving main road  6/Entering main road  7/Entering from slip road  8/Mid Junction - on roundabout or on main road | Input |
| 2nd\_Road\_Class | Categorical | 0/Not at a junction or within 20 meters  1/Motorway  2/A(M)  3/A  4/B  5/C  6/Unclassified | Input |
| Vehicle\_Type | Categorical | 1/Pedal cycle  2/Motorcycle 50cc and under  3/Motorcycle 125cc and under  4/Motorcycle over 125cc and up to 500cc  5/Motorcycle over 500cc  8/Taxi/Private hire car  9/Car  10/Minibus (8 - 16 passenger seats)  11/Bus or coach (17 or more pass seats)  16/Ridden horse  17/Agricultural vehicle  18/Tram  19/Van / Goods 3.5 tonnes mgw or under  20/Goods over 3.5t. and under 7.5t  21/Goods 7.5 tonnes mgw and over  22/Mobility scooter  23/Electric motorcycle  90/Other vehicles  97/Motorcycle - unknown cc  98/Goods vehicle - unknown weight | Input |
| Junction\_Detail | Categorical | 0/Not at a junction or within 20 meters  1/Roundabout  2/Mini-roundabout  3/T or staggered junction  5/Slip road  6/Crossroads  7/More than 4 arms (not roundabout)  8/Private drive or entrance  9/Other junction | Input |
| Hit\_Object\_  in\_Carriageway | Categorical | 0/None  1/Previous accident  2/Road works  4/Parked vehicle  5/Bridge (roof)  6/Bridge (side)  7/Bollard or refuge  8/Open door of the vehicle  9/Central island of the roundabout  10/Kerb  11/Other objects  12/Any animal (except ridden horse) | Input |
| Hit\_Object\_  off\_Carriageway | Categorical | 0/None  1/Road sign or traffic signal  2/Lamp post  3/Telegraph or electricity pole  4/Tree  5/Bus stop or bus shelter  6/Central crash barrier  7/Near/Offside crash barrier  8/Submerged in water  9/Entered ditch  10/Other permanent object  11/Wall or fence | Input |
| Age\_Band\_of\_Driver | Categorical | 1/0 - 5  2/6 - 10  3/11 - 15  4/16 - 20  5/21 - 25  6/26 - 35  7/36 - 45  8/46 - 55  9/56 - 65  10/66 - 75  11/Over 75 | Input |
| Vehicle\_Manoeuvre | Categorical | 1/Reversing  2/Parked  3/Waiting to go - held up  4/Slowing or stopping  5/Moving off  6/U-turn  7/Turning left  8/Waiting to turn left  9/Turning right  10/Waiting to turn right  11/Changing lane to left  12/Changing lane to the right  13/Overtaking moving vehicle - offside  14/Overtaking static vehicle - offside  15/Overtaking - nearside  16/Going ahead left-hand bend  17/Going ahead right-hand bend  18/Going ahead other | Input |
| 1st\_Road\_Class | Categorical | 1/Motorway  2/A(M)  3/A  4/B  5/C  6/Unclassified | Input |
| Light\_Conditions | Categorical | 1/Daylight  4/Darkness - lights lit  5/Darkness - lights unlit  6/Darkness - no lighting  7/Darkness - lighting unknown | Input |
| Day\_of\_Week | Categorical | 1/Sunday  2/Monday  3/Tuesday  4/Wednesday  5/Thursday  6/Friday  7/Saturday | Input |
| Road\_Type | Categorical | 1/Roundabout  2/One way street  3/Dual carriageway  6/Single carriageway  7/Slip road  9/Unknown  12/One-way street/Slip road | Input |
| Vehicle\_Location-Restricted\_Lane | Categorical | 0/On main c'way - not in restricted lane  1/Tram/Light rail track  2/Bus lane  3/Busway (including guided busway)  4/Cycle lane (on main carriageway)  5/Cycleway or shared use footway (not part of main carriageway)  6/On lay-by or hard shoulder  7/Entering lay-by or hard shoulder  8/Leaving lay-by or hard shoulder  9/Footway (pavement)  10/Not on carriageway | Input |
| TimeFlag | Categorical | 0/Not risky  1/Risky time | Input |
| Pedestrian\_Crossing-Physical\_Facilities | Categorical | 0/No physical crossing facilities within 50 metres  1/Zebra  4/Pelican, puffin, toucan or similar non-junction pedestrian light crossing  5/Pedestrian phase at traffic signal junction  7/Footbridge or subway  8/Central refuge | Input |
| Urban\_or\_Rural\_Area | Categorical | 1/Urban  2/Rural  3/Unallocated | Input |

## Hyperparameter tuning configurations

Table 2: Hyperparameter tuning for the GBM approach

|  |  |  |
| --- | --- | --- |
| Parameter | Configured inputs | outcome |
| num\_estimators | [50, 100,500, 1000, 1500] | 1000 |
| learn\_rates | [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] | 0.6 |
| loss | ['ls', 'lad', 'huber','quantile'] | ‘ls’ |
| criterion | ['friedman\_mse', 'mse', 'mae'] | 'friedman\_mse' |

Table 3: Hyperparameter tuning for the ELM approach

|  |  |  |
| --- | --- | --- |
| Parameter | Configured inputs | outcome |
| n\_neurons | [(100, 50), (200, 100), (200, 200), (300, 200), (300, 300), (400, 300), (400, 400), (500, 400),(500, 500)] | (300, 300) |
| ufunc | ['sigm', 'relu', 'tanh', 'lin'] | 'tanh' |
| pairwise\_metric | [*'cosine', 'euclidean', 'cityblock'*] | *'euclidean'* |