Comparison of Classification Models for Road Accident High Severity Prediction– GB 2019

1st Susana Reche Rodríguez
National College of Ireland
Dublin, Ireland
x17165628@student.ncirl.ie

Tutor: John Bohan
National College of Ireland
Dublin, Ireland
john.bohan@ncirl.ie

***Abstract*—Great Britain has suffered an increase in road traffic accidents in recent years. Last 2019 26.30% of the accidents caused deaths or seriously injured casualties. It is important to reduce the number of accidents but also their severity. The current study has tried to identify the most influential environmental factors causing a higher accident severity in 2019. After a first exploratory analysis, associations and correlations tests, and feature selection methods applied to the dataset the most relevant variables resulted to be: junctiondetail, speedlimit, roadtype, road1class, road, crossing, urbanrural, pedestrian, junctioncontrol, weather, special. Later on, different supervised classification algorithms were compared (classification and regression tree (CART), Random Forest (RF), K Nearest Neighbor (KNN), Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM)). Recall and precision were the performance evaluation measures adopted to compare the different models. Accuracy and F-score were also used as secondary measures. SVM with polynomial kernel and LR offered the higher recall. SVM confirmed roadtype, speedlimit, junctiondetail, crossing as most important variables and LR refined the information, pointing to roads with no crossing around 50m or conflictive roundabouts as the most dangerous. The recall for LR was 67.42% using just 5 variables.**

Keywords—classification, machine learning, car, accident, severity, accident severity, unbalanced data, binary classification, environmental factors

I. Introduction

Road traffic accidents were the 8th leading cause of death and caused 1.35 million deaths worldwide last 2018 [1].

Even though Europe is the continent with less deaths caused by road accidents "Fig.1", Great Britain has increased in recent years the number of deaths "Fig.2" especially from 2018 to 2019 "Fig.3".

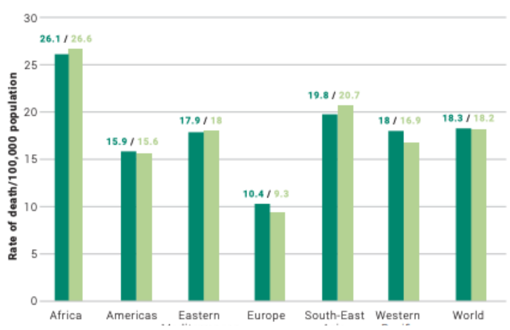


Figure 1: Rates of road traffic death per 100,000 population by WHO regions (2013-2016) [1]

Great Britain has in place a national road safety strategy [Appendix I] which has decreased significantly the number of deaths since 2006. Last 2019, road accidents caused 1,752 deaths, 30,144 seriously injured casualties and 121,262 slightly injured casualties [3], based on the Collision Reporting and Sharing (CRASH) Reporting System [Appendix II]

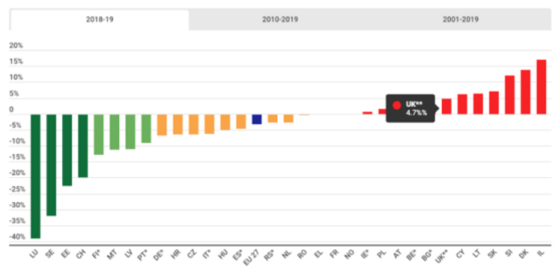


Figure 2: Relative change in road deaths (%) 2010-2019 [2]

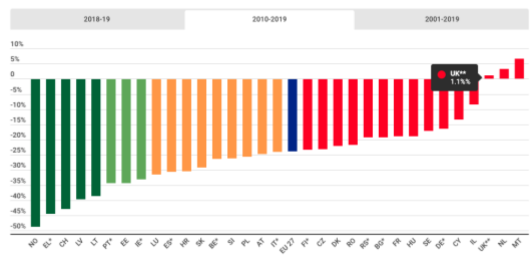


Figure 3: Relative change in road deaths (%) 2018-2019 [2]

High severity accidents are the ones taking lives or causing lifetime damage, and also higher government and insurance cost. Human or vehicle factors are harder to control by the government, for this reason the aim of the present study is to identify the major environmental factors contributing to a higher severity of the accidents for 2019. The latest Great Britain Road Safety data related to 2019 [4] was used. The raw dataset contained 117,536 records and 32 variables [Appendix III]. Stratified sampling was applied to reduce the number of instances to improve computing performance and the SMOTE technique to solve the unbalance class issue. Associations and correlations were studied and different feature selection techniques (Chi-Squared, Boruta, Least Absolute Shrinkage and Selection Operator (LASSO) Regression) were applied to identify the most relevant features.

The recall, precision, together with accuracy and F-score, of different machine learning classification algorithms were compared, to find the best predictor of high severity accidents. The main factors contributing to the prediction of the model were identified.

The project has been enterally done with R, R Studio and different R packages.

I. Objectives of the Study

1) To identify the environmental features causing a higher severity in the accidents occurred last 2019.

2) To discover the most reliable model predicting the higher severity accidents.

II. Related work

A. Machine Learning Application on Accident Injury Severity and Accident Severity Prediction

Machine learning (ML) has been applied in different ways to understand road accidents:

- identification of major factors causing the accidents [5],
- vehicle collision prediction [6] [7] [8],
- accident frequency prediction [9],
- accident / accident injury severity prediction [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24], among others.

The main algorithms applied in the previous accident prediction studies are:

- SVM
- CART, AdaBoost classification tree, J48 decision tree, ID3 decision tree, extremely randomized trees
- NB
- KNN
- RF
- LR, multinomial logistic regression (MLR), binary quantile regression, multivariate adaptative regression splines (MARS)
- Instance-Based learning with parameter k (IBk)

Also, in some cases neural networks (NN) were applied.

- artificial neural Networks (ANN)
- feed-forward neural networks (FNN),
- fuzzy C-means clustering based feed-forward neural networks (FNN-FCM)

Among the traditional model's RF was the one bringing better results in previous studies, so was included in the current study. When compared with other classification models, NN outperformed the other models. As NN are a black box which can't provide enough information about the main factors helping with the prediction, were not included in the current study.

Most previous studies trying to predict accident severity use vehicle, individual and environmental factors, as all the factors together can bring a higher accuracy. The current study tries to use only environmental factors as are the ones easier to modify to reduce the severity of the accidents. Only a couple of studies were found using only environmental factors. [24] study uses only environmental factors and conceives the research as multi-class classification, achieving a recall for some of the classes of more than 90%, however most of them are around 40% or below. Even though multiclass classification could be possible as the raw data contains 3 classes, the current study will use a binary classification in order to simplify the models.

[25] work uses a LR to identify the main environmental factors affecting road accident severity in the United States. LR is a model which brings many insights about the most relevant features, for this reason was included in the current study.

Some of those previous studies have the focus on Great Britain road accidents [11] [18] [19]. All of them used a sample of data extracted from different years (from 3 to 9 years). As the environment conditions, including the traffic signs, can change so much over time, and the focus is on analyzing the information related to 2019, the current study focused only on the latest data from 2019.

B. Feature Engineering

Understanding the most influential factors causing high accident severity allows reducing the number of variables used on the predictive models. By doing so computational resources can be reduced and the performance of the models can be improved.

A wide range of feature reduction techniques were used on previous studies CART [20] [10], RF [6] [26] or LR [10]. In [12] work, 6 different techniques are tested Pearson Correlation, Chi-2, estimated coefficients of the Logistics Regression, Recursive Feature Elimination (RFE) with Logistic Regression, feature importance in Random Forest, Gradient Boost Decision Tree (GBDT). A more novel approach is taken in [15] where particle swarm optimization (PSO) is used. In [16] frequency and relevance analysis techniques are used. [27] uses Voting Algorithm for Aggregated Feature Selection (VAAFS).

Even if not related to car accidents, [28] analyses pros and cons of hierarchical cluster analysis (HCA) and categorical principal component analysis (CATPCA), used as feature selection methods, being HCA the easier to interpret. [29] successfully uses multiple correspondence analysis (MCA). MCA seems a very robust method to be used with categorical data, and for this reason was tested in this study.

In conclusion, as per previous studies, the best approach is using a variety of feature selection techniques to understand the most important features.

C. Unbalanced dataset

As the number of casualties with high severity are normally much lower than the casualties with slighter severity, road accidents severity prediction carries out the complexity of dealing with an unbalanced dataset. [6] work uses Balanced Random Forest, a version of RF which deals with unbalanced datasets.

[21] work uses different oversampling: Synthetic Minority Over-sampling Technique (SMOTE), borderline SMOTE (BLSMOTE), Majority Weighted Minority Oversampling Technique (MWMOTE), and k-means SMOTE (KMSMOTE). The KMSMOTE was the best performing method. [8] also uses SMOTE combined with under sampling the majority class with stratified maximum dissimilarity sampling.

As SMOTE seems a very robust and easy to apply method, the current study incorporated this approach.

D. *Performance Evaluation Metrics*

The performance metrics used on previous accident severity prediction studies focus on accuracy, recall, specificity and F1-score [19] [10] [12] [13] [18] [24]. In some other cases, average recall F1-score, geometric mean [21], ROC curve [22] [13] [23] and Area under the Receiver Operating Characteristic Curve (AUC) [23] were used.

Overall performance metrics are important to consider but, as the current study focused on the prediction of high accident severity, recall and precision were the main metrics considered.

III. METHODOLOGY

A. *Research Design*

"Fig. 4" shows the steps followed in the current study.

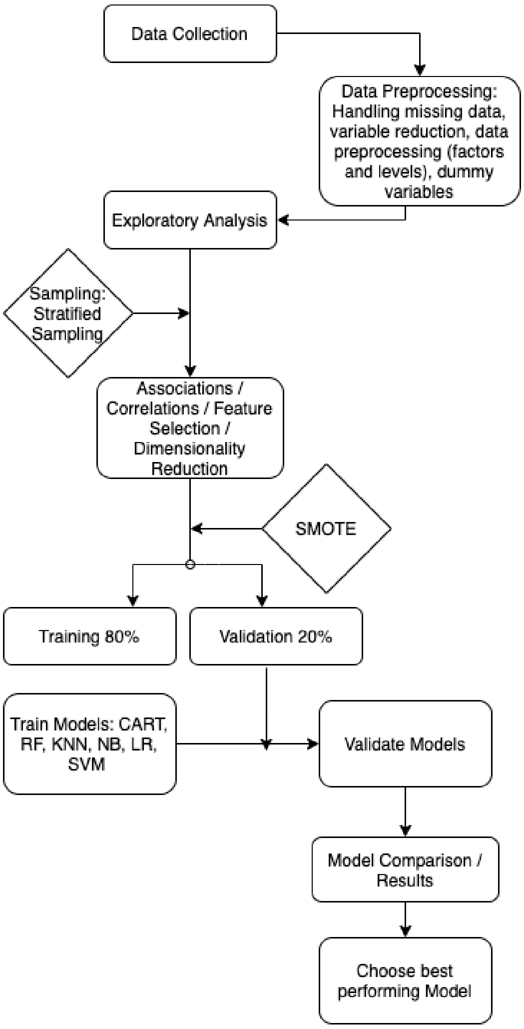


Figure 4: Research Design Chart

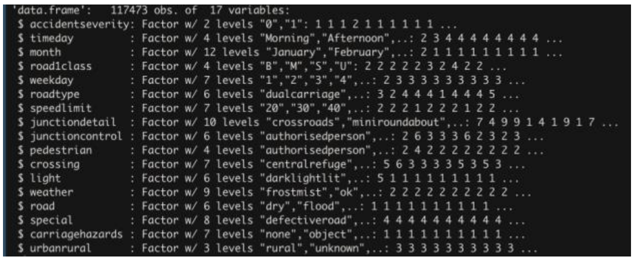
B. *Hardware specifications*

All data processing and model fitting has been done in R using a MacBook Pro with a 2.2 GHz 6-Core Intel Core i7 processor and 16 GB 2400 MHz DDR4 memory.

C. *Dataset and Initial Data Pre-Processing*

Current and relevant data is the most adequate to understand which factors are causing a higher accident severity at the moment. For this reason, the Great Britain Road Safety data related to 2019 [4] was used in this study. The dataset contained 117,536 records and 32 variables [Appendix III]. 14 irrelevant variables were dropped, 63 empty values were deleted, and the variables were transformed into factors. The levels of the factors were modified to facilitate the understanding of the data set.

The selected target variable was "accident severity". The variable contained 3 levels but was transformed into a binary variable grouping the most severe accidents together.



```
'data.frame':   117473 obs. of  17 variables:
 $ accidentseverity: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
 $ timeday         : Factor w/ 4 levels "Morning","Afternoon",..: 2 3 4 4 4 4 4 4 4 4 ...
 $ month           : Factor w/ 12 levels "January","February",..: 2 1 1 1 1 1 1 1 1 1 ...
 $ road1class      : Factor w/ 4 levels "B","M","S","U": 2 2 2 2 2 3 2 4 2 2 ...
 $ weekday         : Factor w/ 7 levels "1","2","3","4",..: 2 3 3 3 3 3 3 3 3 3 ...
 $ roadtype        : Factor w/ 6 levels "dualcarriage",..: 3 2 4 4 4 1 4 4 4 5 ...
 $ speedlimit      : Factor w/ 7 levels "20","30","40",..: 2 2 2 1 2 2 2 1 2 2 ...
 $ junctiondetail  : Factor w/ 10 levels "crossroads","miniroundabout",..: 7 4 9 9 1 4 1 9 1 7 ...
 $ junctioncontrol : Factor w/ 6 levels "authorisedperson",..: 2 6 3 3 3 6 2 3 2 3 ...
 $ pedestrian      : Factor w/ 4 levels "authorisedperson",..: 2 4 2 2 2 2 2 2 2 2 ...
 $ crossing        : Factor w/ 7 levels "centralrefuge",..: 5 6 3 3 3 3 5 3 5 3 ...
 $ light           : Factor w/ 6 levels "darklightlit",..: 5 1 1 1 1 1 1 1 1 1 ...
 $ weather         : Factor w/ 9 levels "frostmist","ok",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ road            : Factor w/ 6 levels "dry","flood",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ special         : Factor w/ 8 levels "defectiveroad",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ carriagehazards : Factor w/ 7 levels "none","object",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ urbanrural      : Factor w/ 3 levels "rural","unknown",..: 3 3 3 3 3 3 3 3 3 3 ...
```

Figure 5: Dataset structure after first round of transformation

D. *Initial Exploratory Analysis*

Road accidents occurred mostly in urban areas, in medium size roads with a maximum speed of 30 mile per hour. The main problems appeared in junctions with more than 4 arms with no roundabout, tor staggered junctions or slip roads.

The higher number of accidents occurred on Saturdays and afternoons and during July and November.

The accidents with higher severity follow the same patterns as the accidents with less severity. It is hard to define which environmental features increase the severity by looking at "Fig.6", "Fig.7", "Fig.8" and "Fig.9".

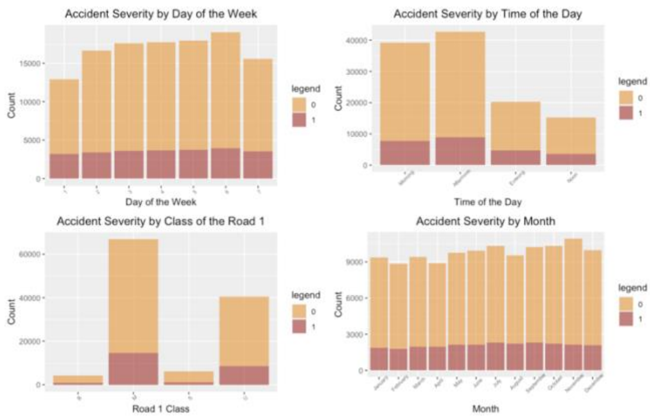


Figure 6: Distribution of Accident Severity in relation to day of the week, time of the day, class of the road 1, month

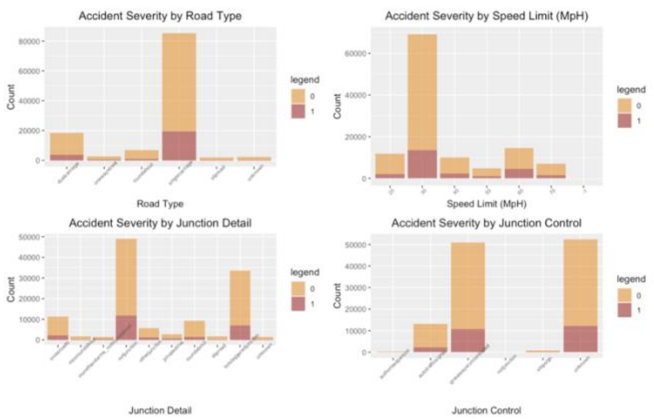


Figure 7: Distribution of Accident Severity in relation to road type, speed limit, junction detail and junction control.

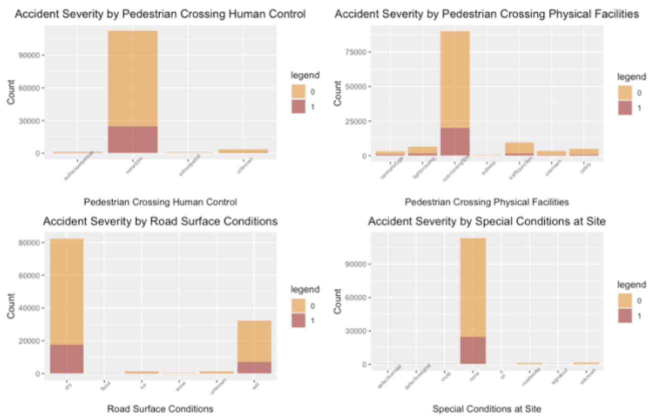


Figure 8: Distribution of Accident Severity in relation to pedestrian crossing human control, pedestrian crossing physical facilities, road surface conditions and special conditions at site

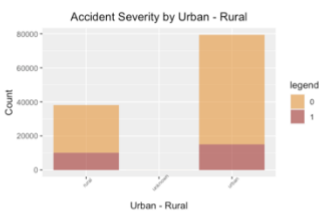


Figure 9: Distribution of Accident Severity in relation to Urban-Rural

E. Improve Performance

In order to reduce computational resources but still have a statistically significant sample, stratified sampling was applied to the raw dataset [30]. The data was divided into homogenous subgroups. Then a sample from each subgroup was extracted in a way that the resulting final sample has the same distribution as the raw data as per [31]. The number of rows was reduced from 117,536 to 30,000.

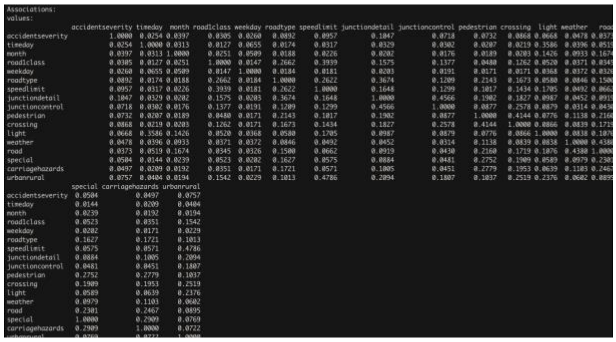
F. Dummy Variables

The categorical variables needed to be transformed in numerical, to understand the correlation and to be used in some of the models. The method applied was transforming each of the categories of a variable into 0 (absence) or 1 (presence), also known as hot encoding or dummy variables.

G. Associations / Correlations

To detect if any feature could be dropped for being independent from accident severity a Chi-Squared test was performed. The test revealed that the null hypothesis of independence couldn't be rejected for any of them.

When dealing with categorical variables statistical measures are more trustable than distance metrics [32] [33] [34] , for this reason the preferred method to understand associations between the variables was the Cramer's V test [35], using the GoodmanKruskal [36] [pag. 14 , Appendix IV]



```
Associations:
values:
                  accidentseverity timeday  month roadlclass weekday roadtype speedlimit junctiondetail junctioncontrol pedestrian crossing  light weather   road
accidentseverity            1.0000  0.0254 0.0397     0.0305  0.0260   0.0892     0.0957         0.1047          0.0718     0.0732   0.0868 0.0668  0.0478 0.0373
timeday                     0.0254  1.0000 0.0313     0.0127  0.0655   0.0174     0.0317         0.0329          0.0302     0.0207   0.0219 0.3586  0.0396 0.0519
month                       0.0397  0.0313 1.0000     0.0251  0.0509   0.0188     0.0226         0.0202          0.0176     0.0189   0.0203 0.1426  0.0933 0.1674
roadlclass                  0.0305  0.0127 0.0251     1.0000  0.0147   0.2662     0.3939         0.1575          0.1377     0.0480   0.1262 0.0520  0.0371 0.0345
weekday                     0.0260  0.0655 0.0509     0.0147  1.0000   0.0184     0.0181         0.0203          0.0191     0.0171   0.0171 0.0368  0.0372 0.0326
roadtype                    0.0892  0.0174 0.0188     0.2662  0.0184   1.0000     0.2622         0.3674          0.1209     0.2143   0.1673 0.0580  0.0846 0.1500
speedlimit                  0.0957  0.0317 0.0226     0.3939  0.0181   0.2622     1.0000         0.1648          0.1299     0.1017   0.1434 0.1705  0.0492 0.0662
junctiondetail              0.1047  0.0329 0.0202     0.1575  0.0203   0.3674     0.1648         1.0000          0.4566     0.1902   0.1827 0.0987  0.0452 0.0919
junctioncontrol             0.0718  0.0302 0.0176     0.1377  0.0191   0.1209     0.1299         0.4566          1.0000     0.0877   0.2578 0.0879  0.0314 0.0430
pedestrian                  0.0732  0.0207 0.0189     0.0480  0.0171   0.2143     0.1017         0.1902          0.0877     1.0000   0.4144 0.0776  0.1138 0.2160
crossing                    0.0868  0.0219 0.0203     0.1262  0.0171   0.1673     0.1434         0.1827          0.2578     0.4144   1.0000 0.0866  0.0839 0.1719
light                       0.0668  0.3586 0.1426     0.0520  0.0368   0.0580     0.1705         0.0987          0.0879     0.0776   0.0866 1.0000  0.0838 0.1076
weather                     0.0478  0.0396 0.0933     0.0371  0.0372   0.0846     0.0492         0.0452          0.0314     0.1138   0.0839 0.0838  1.0000 0.4380
road                        0.0373  0.0519 0.1674     0.0345  0.0326   0.1500     0.0662         0.0919          0.0430     0.2160   0.1719 0.1076  0.4380 1.0000
special                     0.0504  0.0144 0.0239     0.0523  0.0202   0.1627     0.0575         0.0884          0.0481     0.2752   0.1909 0.0589  0.0979 0.2301
carriagehazards             0.0497  0.0209 0.0192     0.0351  0.0171   0.1721     0.0571         0.1005          0.0451     0.2779   0.1953 0.0639  0.1103 0.2467
urbanrural                  0.0757  0.0404 0.0194     0.1542  0.0229   0.1013     0.4786         0.2094          0.1807     0.1037   0.2519 0.2376  0.0602 0.0895
                  special carriagehazards urbanrural
accidentseverity   0.0504          0.0497     0.0757
timeday            0.0144          0.0209     0.0404
month              0.0239          0.0192     0.0194
roadlclass         0.0523          0.0351     0.1542
weekday            0.0202          0.0171     0.0229
roadtype           0.1627          0.1721     0.1013
speedlimit         0.0575          0.0571     0.4786
junctiondetail     0.0884          0.1005     0.2094
junctioncontrol    0.0481          0.0451     0.1807
pedestrian         0.2752          0.2779     0.1037
crossing           0.1909          0.1953     0.2519
light              0.0589          0.0639     0.2376
weather            0.0979          0.1103     0.0602
road               0.2301          0.2467     0.0895
special            1.0000          0.2909     0.0769
carriagehazards    0.2909          1.0000     0.0722
urbanrural         0.0769          0.0722     1.0000
```

Figure 10: Cramer's V Test – Associations – Greybox package [33]

and greybox package "Fig.10". Also, the correlations between the variables were checked using the dummy variables and the corplot package [pag.15, Appendix V].

The finding was that none of the dependent variables was highly associated or correlated with the target variable. In the following tables appear the variables most associated "Table 1" and correlated "Table 2" with accident severity.

Associations	
Target Variable	*Features*
accidentseverity	roadtype, speedlimit, junctiondetail, crossing and urbanrural

Table 1: Associations Target Variable and Features

Correlations	
Target Variable	*Features - Categories*
accidentseverity	roadtype_singlecarriage, speedlimit_60, junctiondetail_notjunction, pedestrian_unknown

Table 2: Correlations Target Variable and Feature Categories

Also, low associations and correlations appeared when testing the relations between the dependent variables. Some high correlations were found between categories of the same variables, but never reaching 1 or -1. This fact ensures the absence of multicollinearity, an assumption required to perform some of the machine learning algorithms.

H. Feature Selection

Apart from the previous statistical tests, three different feature selection methods were used to understand the most relevant factors affecting the severity of the accident:

- The Fselector package [37] uses the Chi-Squared test to find the weight of the different variables.
- The Boruta package [38] [39] is a wrapper feature selection algorithm which uses RF. However, it is important to highlight that RF is biased towards variables with high cardinality (too many unique values) [40] [41].
- The glmnet package [42] allows performing a LASSO regression [39] [43], which selects the most relevant variables and drops the non-relevant ones by assigning them a coefficient of zero.

The resulting variables “Table 3”, were really similar to the ones determined by the associations and correlations.

Feature Selection	
Feature Selection	*Most Relevant Variables*
Chi-Squared	junctiondetail, speedlimit, roadtype, crossing, urbanrural, pedestrian, junctioncontrol
Boruta – Random Forest	speedlimit, roadtype, road1class, road, urbanrural, weather, pedestrian
LASSO Regression	pedestrian, roadtype, speedlimit, special, urbanrural, crossing, junstioncontrol

Table 3: Feature Selection - Most Important Variables

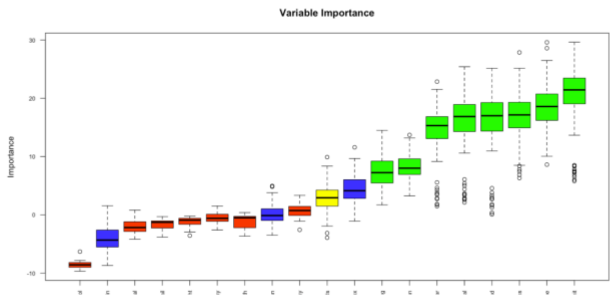


Figure 11: meanImp from Boruta package

I. Dimensionality Reduction

Multiple Correspondence Analysis [MCA] [44] [45] was performed using the FactoMineR package, however the eigenvalues of the resulting dimensions were very low. The new dimensions were not used in the study.

J. Balance the Dataset

An unbalanced dataset contains many more records for one of the classes. When feeding a classification algorithm with an unbalanced dataset, the algorithm gives priority to the class with more instances (overfitting of the majority class) [46].

From the 30,000 records result of the stratified sampling only 23.06% belonged to class 1 (high severity). To overcome the issue, the current study used a data augmentation technique. In concrete, the technique used was the Synthetic Minority Over-sampling (SMOTE) [47] [48] [49], which uses bootstrapping and KNN to increase the number of records of the class with less instances. In this case also the number of instances of the class with more records was reduced by half.

After applying SMOTE the classes were closed to 50% each. The total number of records suffered a small reduction from 30,000 to 27,668 records.

K. Models - Binary Classification

The prediction of accident severity was stated as a binary classification problem, where 1 is high severity and 0 low severity. The most important goal was predicting high severity accidents.

Based on the exploratory analysis, associations, correlations and feature selection information 4 different groups of variables were created to feed the models: Group 0 “Table 4”, Group 1 “Table 5”, Group 2 “Table 6” and Group 3 “Table 7”.

Group 0
Variables (16 variables)
Timeday, month, road1class, weekday, roadtype, speedlimit, junctiondetail, junctioncontrol, pedestrian, crossing, light, weather, road, special, carriagehazards, urbanrural

Table 4: Variables in Group 0

Group 1
Variables (8 variables)
roadtype, speedlimit, junctiondetail, junctioncontrol, pedestrian, crossing, light, urbanrural

Table 5: Variables in Group 1

Group 2
Variables (8 variables)
roadtype, speedlimit, junctiondetail, pedestrian, crossing, light, weather, urbanrural

Table 6: Variables in Group 2

Group 3
Variables (5 variables)
roadtype, speedlimit, junctiondetail, crossing, urbanrural

Table 7: Variables in Group 3

Each of those groups contained the dataset with categorical variables and also the transformed numerical dataset.

The categorical data was used in CART, RF, NB and SVM. SVM converts the data to dummy by itself.

Before using it in the models, the variables (categories of variables) with variance zero (with a unique value) were removed from the numerical variable’s dataset, as those variables contained not valuable information. The cleaned numerical dataset was then used in KNN and LR.

The different groups were split into training (80%) and validation (20%). The training dataset was used to train the models and the validation to test the models.

Different machine learning algorithms were used to solve the binary classification problem:

1) *Classification and Regression Tree:* Divides the data into smaller portions using recursive partitioning to identify patterns which can be then used for the prediction of the class. Selects the variables that bring more information gain and less entropy [50]. Can be biased towards variables with greater number of categories [51]. Dummy variables can degrade the performance, as the higher value is considered [52] [53] [54], for this reason categorical variables were used.

2) *Random Forest:* Uses decision trees and bagging (bootstrap aggregation), to generate many trees that will give the prediction of the class [55].

3) *K-Nearest Neighbors:* Assigns a class to the record based on the class of the k records which are nearby. Choosing the optimal k is important to don’t over fit or underfit the model. Normally the distance is measured by the Euclidean distance, for this reason the variables need to be numerical.

4) *Naïve Bayes:* Use probability of each class based on each of the variables, to predict the class in the validation data set. This model has the assumption that the independent features are independent from each other (correlation zero), which is normally not happening [56]. This fact makes the model very rigid making logistic regression a better performer [57].

5) *Logistic Regression:* Works similar to the linear regression but runs the output through a logistic or sigmoid function. The dependent variable needs to be binary, the observations need to be independent from each other and there needs to be none or not much collinearity between the dependent variables. [58] [59] [60] [61].

6) *Support Vector Machine:* Represent the instances in a higher dimensional space, when is not linearly separable, to be able to place the record under the right class (in case of a classification problem). Requires numerical data and can be linear or nonlinear. The linear model can be used when the data can be separated by a straight line. The current study used radial, polynomial and sigmoid kernels (type of function) [62].

IV. Results

A. Performance Evaluation Metrics

The most important metrics for the study were recall and precision, as the study is centered around predicting high severity accidents. Also, the accuracy was taken into consideration, to see how accurate the model was predicting both classes.

On the other hand, to enable comparing the overall performance of the model among the different models the F-score was calculated as well.

Recall (Sensitivity) is the percentage of correctly predicted positive values among the total amount of real positives.

$$\frac{TP}{(TP+FN)}$$

Precision (Pos Pred Value) is the percentage of correctly predicted positive values, among the total of predicted positives.

$$\frac{TP}{(TP+FP)}$$

Accuracy, total number of matches among all the classes.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

F-score combines sensitivity and specificity and allows comparing the effectiveness of different models.

$$\text{F-score} = \frac{(2*Precission*Recall)}{(Precission+Recall)} = \frac{(2*TP)}{(2*TP+FP+FN)}$$

B. Results

Overall, the recall was very low for most of the models. The models performing the best were LR, RF and SVM (Polynomial) [63].

When performing KNN in Groups 1 to 2, the k needed to be reduced to 1 to avoid "too many ties" [64] [65]. In Group 3 reducing the k to 1, setting the parameter "use. All" to FALSE was tested, but still "too many ties" appeared. By setting k to 1, there is more noise and there is a high risk of misclassification. For this reason, KNN results could not be considered accurate.

Performance Group 0 (16 variables)				
Model	*Recall (Sensitivity)*	*Precision (Pos Pred Value)*	*Accuracy*	*F-score*
CART	0.3503	0.8761	0.6504	0.5005
RF	0.5965	0.7732	0.7108	0.6735
KNN	0.4295	0.7714	0.6511	0.5518
NB	0.4573	0.8554	0.6564	0.5960
LR	0.7283	0.4971	0.6558	0.5909
SVM (Radial)	0.4291	0.8312	0.671	0.5660
SVM (Polynomial)	0.4693	0.7494	0.6562	0.5772
SVM (Sigmoid)	0.4150	0.8159	0.6607	0.5502
SVM (Linear)	0.4577	0.7753	0.6625	0.5756

Table 8: Performance Group 0

Performance Group 1 (8 variables)				
Model	*Recall*	*Precision*	*Accuracy*	*F-score*
CART	0.5589	0.6905	0.6542	0.6178
RF	0.4505	0.8599	0.6885	0.5912
KNN	0.4432	0.8055	0.6681	0.5718
NB	0.4830	0.6844	0.6302	0.5663
LR	0.6733	0.5029	0.6294	0.5758
SVM (Radial)	0.4291	0.7856	0.656	0.5550
SVM (Polynomial)	0.4526	0.7421	0.6477	0.5623
SVM (Sigmoid)	0.4038	0.7467	0.6334	0.5242
SVM (Linear)	0.4017	0.7517	0.6345	0.5236

Table 9: Performance Group 1

Performance Group 2 (8 variables)				
Model	*Recall*	*Precision*	*Accuracy*	*F-score*
CART	0.3847	0.7605	0.6318	0.5109
RF	0.4722	0.7849	0.6714	0.5897
KNN	0.4823	0.7366	0.6549	0.5829
NB	0.4450	0.7149	0.6338	0.5485
LR	0.6757	0.5145	0.6338	0.5842
SVM (Radial)	0.4226	0.7547	0.6426	0.5418
SVM (Polynomial)	0.5192	0.6713	0.6325	0.5855
SVM (Sigmoid)	0.4002	0.7245	0.624	0.5156
SVM (Linear)	0.4118	0.7075	0.6208	0.5206

Table 10: Performance Group 2

Performance Group 3 (5 variables)				
Model	*Recall*	*Precision*	*Accuracy*	*F-score*
CART	0.2523	0.8300	0.6003	0.3870
RF	0.3836	0.7756	0.6363	0.5133
KNN				
NB	0.4512	0.6674	0.6132	0.5384
LR	**0.6742**	**0.4219**	**0.609**	**0.5190**
SVM (Radial)	0.4067	0.7138	0.6218	0.5182
SVM (Polynomial)	**0.5821**	**0.6124**	**0.6068**	**0.5969**
SVM (Sigmoid)	0.4378	0.6599	0.6061	0.5264
SVM (Linear)	0.4378	0.6603	0.6063	0.5265

Table 11: Performance Group 3

Following Occam's razor principle [66], the models with higher recall using the fewer variables were LR and SVM (Polynomial). To discover the most influential variables in SVM different new subsets were created. The most relevant variables result of the SVM (Polynomial) model appeared to be the ones determined by the Cramer's V test (see v4 and v5 in "Table 12" and the related results in "Table 13"). In both winning SVM models the class 1 was correctly predicted in around 60% of the cases with an accuracy of around 60%.

Group 3 Versions	
Version	*Variables*
v1	speedlimit, junctiondetail, crossing, urbanrural
v2	roadtype, junctiondetail, crossing, urbanrural
v3	roadtype, speedlimit, crossing, urbanrural
v4	**roadtype, speedlimit,junctiondetail, urbanrural**
v5	**roadtype, speedlimit,junctiondetail, crossing**

Table 12: Group 3 Versions

Performance Group 3 Versions				
SVM (Polynomial)	*Recall*	*Precision*	*Accuracy*	*F-score*
Original Group 3	**0.5821**	**0.6124**	**0.6068**	**0.5969**
Group 3 v.1	0.5221	0.5938	0.5824	0.5556
Group 3 v.2	0.2802	0.8184	0.6090	0.4175
Group 3 v.3	0.4845	0.6533	0.6137	0.5564
Group 3 v.4	**0.6189**	**0.5661**	**0.5723**	**0.5913**
Group 3 v.5	**0.5362**	**0.6179**	**0.6023**	**0.5742**

Table 13: SVM Polynomial – Performance Group 3 Versions

Even though the precision is only 42% for the LR model, it can be considered the winner, as predicted the class 1 in nearly 70% of the records.

At the LR model some categories appeared irrelevant (when the confidence interval crosses 1 as per "Fig.13" [67]): roadtype_roundabout, speedlimit_30, speedlimit_20, speedlimit_40 and crossing zebra.

The categories with higher significance as per "Fig. 12" were: roadtype_unknown, speedlimit_60, speedlimit_70, junctiondetail_crossroads, junctiondetail_roundabout, crossing_notcrossing50m, crossing_unknown, crossing_trafficjunction. From those, crossing_notcrossing50m and junctiondetail_roundabout were the factors affecting the most severity.

```
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -10.92989   84.47678  -0.129 0.897054
roadtype_singlecarriage              0.18656    0.07244   2.575 0.010015 *
roadtype_dualcarriage                0.17699    0.07831   2.260 0.023819 *
roadtype_unknown                     1.87237    0.09847  19.015  < 2e-16 ***
roadtype_roundabout                  0.11833    0.10064   1.176 0.239695
speedlimit_30                        0.05765    0.07342   0.785 0.432323
speedlimit_20                        0.06408    0.08221   0.779 0.435700
speedlimit_60                        0.37168    0.07941   4.680 2.86e-06 ***
speedlimit_40                        0.03838    0.08286   0.463 0.643230
speedlimit_70                        0.45209    0.08767   5.157 2.51e-07 ***
junctiondetail_notjunction           0.34260    0.04520   7.579 3.48e-14 ***
junctiondetail_torstaggeredjunction  0.09551    0.04789   1.994 0.046117 *
junctiondetail_crossroads            0.34563    0.05708   6.055 1.41e-09 ***
junctiondetail_roundabout           -0.31178    0.07556  -4.126 3.69e-05 ***
crossing_notcrossing50m             -0.28989    0.08118  -3.571 0.000356 ***
crossing_unknown                     0.97940    0.09729  10.066  < 2e-16 ***
crossing_trafficjunction            -0.47932    0.09340  -5.132 2.86e-07 ***
crossing_lightcrossing              -0.21068    0.09674  -2.178 0.029431 *
crossing_zebra                      -0.17051    0.10042  -1.698 0.089509 .
urbanrural_urban                    10.49068   84.47668   0.124 0.901169
urbanrural_rural                    10.57993   84.47668   0.125 0.900333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 12: Logistic Regression Summary - Group 3

```
                                              OR       2.5 %     97.5 %
(Intercept)                         1.791477e-05          NA 10.2866232
roadtype_singlecarriage             1.205097e+00 1.04602255  1.3896511
roadtype_dualcarriage               1.193621e+00 1.02410048  1.3921851
roadtype_unknown                    6.503693e+00 5.36846471  7.8980766
roadtype_roundabout                 1.125611e+00 0.92429311  1.3714078
speedlimit_30                       1.059349e+00 0.91753337  1.2236314
speedlimit_20                       1.066182e+00 0.90762190  1.2528263
speedlimit_60                       1.450172e+00 1.24138559  1.6948177
speedlimit_40                       1.039127e+00 0.88345581  1.2225759
speedlimit_70                       1.571591e+00 1.32375668  1.8667410
junctiondetail_notjunction          1.408612e+00 1.28928712  1.5392554
junctiondetail_torstaggeredjunction 1.100215e+00 1.00170999  1.2085763
junctiondetail_crossroads           1.412875e+00 1.26341454  1.5802771
junctiondetail_roundabout           7.321459e-01 0.63116608  0.8488029
crossing_notcrossing50m             7.483451e-01 0.63825392  0.8775089
crossing_unknown                    2.662860e+00 2.20121680  3.2235664
crossing_trafficjunction            6.192026e-01 0.51557440  0.7435837
crossing_lightcrossing              8.100346e-01 0.67007462  0.9791792
crossing_zebra                      8.432371e-01 0.69254668  1.0266872
urbanrural_urban                    3.597853e+04 0.06248663         NA
urbanrural_rural                    3.933744e+04 0.06830921         NA
```

Figure 13: Odds and Confidence Interval - Group 3

V. CONCLUSIONS AND FUTURE WORK

In conclusion, the study determined that the main environmental features causing higher severity were roadtype, speedlimit, junctiondetail, crossing and urban rural. The two most important features among those were roads without crossing and roundabouts. And the best model predicting higher severity accidents was LR.

As the main factors causing a higher severity accident determined by the LR model are the roads without crossing or roundabouts, the full data set can be filtered out looking for those types of accidents occurred last 2019. Once the geographical points affected have been identified some possible actions are:

- Analyze if on a road with a speed of 60/70m/h with no crossing in 50m some kind of new road signs, traffic lights or new crossings can be incorporated.
- Analyze if the roundabouts design can be improved together with the traffic lights.

Possible future work in order to keep improving the insights obtained with the dataset, in order to reduce the high accident severity, can be done in many areas:

1) *Dealing with big data sets:* Other sampling techniques could be used, as well as parallel computing to being able to use the full dataset.
2) *Dealing with unbalanced dataset:* Apart from other oversampling techniques it would be good trying to use a penalized version of the algorithms which could deal with unbalanced dataset.
3) *Categorical variables into numerical variables.* Among all the different ways to transform categorical variables into numerical variables [68] an interesting encoding to test is embedding, which uses NN to transform the variables into vector spaces [69] [70].
4) *Feature engineering:* Some appealing techniques to try could be hierarchical clustering [71] or variable clustering [72]. Also, many other groups of features could be tested in the models in order to find the best performing ones.
5) *Multiclass classification:* The study can be approached as multiclass instead of a binary classification.

6) *Training/Validation:* k-fold cross validation [73] [74] could be employed to maximize the outcome from the training stage.
7) *KNN:* perform a deeper research on how it would be possible to deal with ties.

REFERENCES

[1] World Health Organization, "Global status report on road safety 2018," 2020. [Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/. [Accessed 30 October 2020].

[2] European Transport Safety Control, "Road deaths in the European Union – latest data," 2019. [Online]. Available: https://etsc.eu/euroadsafetydata/. [Accessed 30 October 2020].

[3] Department for Transport, "Reported road casualties in Great Britain: 2019 annual report," 30 September 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/922717/reported-road-casualties-annual-report-2019.pdf. [Accessed 30 October 2020].

[4] Department for Transport, "Road Safety Data," 30 September 2020. [Online]. Available: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data. [Accessed 30 October 2020].

[5] A. T. K. &. R. R. Mohammad Bagher Anvari, "Identifying the Most Important Factors in the At-Fault Probability of Motorcyclists by Data Mining, Based on Classification Tree Models," 4 April 2017. [Online]. Available: https://link.springer.com/article/10.1007/s40999-017-0180-0. [Accessed 14 October 2020].

[6] T. G. T. G. B. J. Antoine Hebert, "High-Resolution Road Vehicle Collision Prediction for the City of Montreal," 11 November 2019. [Online]. Available: https://arxiv.org/pdf/1905.08770.pdf. [Accessed 13 October 2020].

[7] K. B. A. A. M. M. D. S. ALI AHMED MOHAMMED, "CLASSIFICATION OF TRAFFIC ACCIDENT PREDICTION MODELS: A REVIEW PAPER," April 2018. [Online]. Available: https://www.researchgate.net/publication/321462497_Classification_of_Traffic_Accident_Prediction_Models_A_Review_Paper. [Accessed 13 October 2020].

[8] R. S. G. L. M. M. Matthias Schlögl, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset," June 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457518307760?. [Accessed 14 October 2020].

[9] H.-c. C. D.-j. L. P. L. Li-yen CHANG, "Analysis of freeway accident frequency using multivariate adaptive regression splines," 2012. [Online]. Available: https://core.ac.uk/download/pdf/82324321.pdf. [Accessed 11 October 2020].

[10] M.-M. C. a. M.-C. Chen, "Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest," 18 May 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/5/270/pdf. [Accessed 31 October 2020].

[11] D. A. Shibani, "ANALYSIS OF TRAFFIC ACCIDENT SEVERITY ON GREAT BRITAIN ROADWAYS AND JUNCTION," January 2016. [Online]. Available: https://www.researchgate.net/publication/303903666_Analysis_of_Traffic_Accident_Severity_on_Great_Britain_Roadways_and_Junctions. [Accessed 13 October 2020].

[12] M. L. ,. M. D. a. A. S. Natalia Selini Hadjidimitriou, "Machine Learning for Severity Classification of Accidents Involving Powered Two Wheelers," 10 October 2020. [Online]. Available: https://ezproxy.ncirl.ie:2102/stamp/stamp.jsp?tp=&arnumber=8842628. [Accessed 13 October 2020].

[13] S. K. &. D. Toshniwal, "Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India," 1 May 2017. [Online]. Available: https://link.springer.com/article/10.1007%2Fs12544-017-0242-z. [Accessed 14 October 2020].

[14] F. Fan and Y.-B. Qian, "Analysis of Factors Affecting the Severity of Car Accidents at iNtersections Based on Cumulative Logistic Model," 25 December 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8446691. [Accessed 15 October 2020].

[15] C. Qiu, X. Zuo and F. Xiang, "A Particle Swarm Optimization based Feature Selection Method for Accident Severity Analysis," 2 August 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9066708. [Accessed 15 October 2020].

[16] T. K. Bahiru, D. K. Singh and E. A. Tessfaw, "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity," 21 April 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8473265. [Accessed 14 October 2020].

[17] M. Kyi Pyar Hlaing University of Information Technology, N. T. T. Aung, S. Z. Hlaing and K. Ochimizu, "Analysis of accident severity factor in Road Accident of Yangon using FRAM and Classification Technique," 7 November 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8921119. [Accessed 15 October 2020].

[18] S. a. E. U. o. S. G. M. U. Charith Silva School of Computing and M. Saraee, "Predicting Road Traffic Accident Severity using Decision Trees and Time-Series Calendar Heatmaps," 9 November 2019. [Online]. Available: https://ieeexplore.ieee.org/document/9214709. [Accessed 15 October 2020].

[19] S. M. R. U. M. a. N. R. Khaled Assi, "Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol," 30 July 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/15/5497/pdf. [Accessed 13 October 2020].

[20] M. F. K. M. A.-O. A. A. Aljarrah, "Investigating Key Factors Influencing the Severity of Drivers Injuries in Car Crashes Using Supervised Machine Learning Techniques.," 2019. [Online]. Available: http://eds.a.ebscohost.com/eds/detail/detail?vid=13&sid=ad524b5e-31e1-4996-8faa-38790b408a1e%40sessionmgr4006&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=139207272&db=a9h. [Accessed 13 October 2020].

[21] A. P. M. T. J. M. P. D. G. R. P. Sobhan Sarkar Ph. D., "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," May 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753520300138?. [Accessed 13 October 2020].

[22] R. R. M. M. B. Ali Tavakoli Kashani, "A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers," December 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022437514000929?via%3Dihub. [Accessed 14 October 2020].

[23] W. M. U. K. M. Rabia Emhamed AlMamlook Ph.D. Candidate, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity," 11 April 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8717393. [Accessed 15 October 2020].

[24] H. J. Lukuman Wahab, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," 4 April 2019. [Online]. Available: http://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=10&sid=136dec24-7dcb-4057-84c9-af40b5b06ca7%40pdc-v-sessmgr02. [Accessed 14 October 2020].

[25] W. Z. Yubian Wang, "Analysis of Roadway and Environmental Factors Affecting Traffic Crash Severities," 10-15 July 2016. [Online]. Available: https://reader.elsevier.com/reader/sd/pii/S2352146517307147?token=E626527F1A5288689D51487325DC504D4E66DE3FB6703A677F0346666762153705D42A5F3A6AAE828EBA46DE7D969DD6. [Accessed 1 December 2020].

[26] S. C. Ifeoma Oduntan, "Comparative Analysis of Classification Accuracy of SixMachine Learning Algorithms on New York City Dataset forCrime Prediction," April 2020. [Online]. Available: https://www.researchgate.net/publication/340863124_Comparative_Analysis_of_Classification_Accuracy_of_Six_Machine_Learning_

Algorithms_on_New_York_City_Dataset_for_Crime_Prediction. [Accessed 1 November 2020].

[27] S. S. R. GEETHA RAMANI, "A Pragmatic Approach for Refined Feature Selection for the Prediction of Road Accident Severity," 2014. [Online]. Available: https://pdfs.semanticscholar.org/73fe/869b78b589d27a16386ea44cf1b1d490f8f8.pdf. [Accessed 22 October 2020].

[28] H. Ř. ZDENĚK ŠULC, "DIMENSIONALITY REDUCTION OF CATEGORICAL DATA: COMPARISON OF HCA AND CATPCA APPROACHES," 6 September 2015. [Online]. Available: http://amse-conference.eu/history/amse2015/doc/Sulc_Rezankova.pdf. [Accessed 20 October 2020].

[29] F. I. U. M. F. U. Sheng Guan School of Computing and Information Sciences, M. Chen, H.-Y. Ha, S.-C. Chen, M.-L. Shyu and C. Zhang, "Deep Learning with MCA-based Instance Selection and Bootstrapping for Imbalanced Data Classification," 30 October 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7423094. [Accessed 1 December 2020].

[30] G. Malato, "Stratified sampling and how to perform it in R," 7 May 2019. [Online]. Available: https://towardsdatascience.com/stratified-sampling-and-how-to-perform-it-in-r-8b753efde1ef. [Accessed 28 October 2020].

[31] Wikipedia, "Stratified sampling," [Online]. Available: https://en.wikipedia.org/wiki/Stratified_sampling. [Accessed 31 October 2020].

[32] S. Zychlinski, "The Search for Categorical Correlation," 24 February 2018. [Online]. Available: https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9. [Accessed 2 November 2020].

[33] I. Svetunkov, "Marketing analytics with greybox," 20 October 2020. [Online]. Available: https://cran.r-project.org/web/packages/greybox/vignettes/maUsingGreybox.html. [Accessed 2 November 2020].

[34] Outside Two Standard Deviations, "An overview of correlation measures between categorical and continuous variables," 13 September 2018. [Online]. Available: https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365. [Accessed 3 November 2020].

[35] Wikipedia, "Cramér's V," [Online]. Available: https://en.wikipedia.org/wiki/Cram%C3%A9r%27s_V. [Accessed 3 November 2020].

[36] R. Pearson, "The GoodmanKruskal package: Measuring association between categorical variables," 18 March 2020. [Online]. Available: https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html. [Accessed 2 November 2020].

[37] L. K. Piotr Romanski, "Package "Fselector"," 16 May 2018. [Online]. Available: https://cran.r-project.org/web/packages/FSelector/FSelector.pdf. [Accessed 11 November 2020].

[38] M. B. K. a. W. R. Rudnicki, "Package "Boruta"," 21 May 2020. [Online]. Available: https://cran.r-project.org/web/packages/Boruta/Boruta.pdf. [Accessed 12 November 2020].

[39] S. Prabhakaran, "Feature Selection – Ten Effective Techniques with Examples," May 2018. [Online]. Available: https://www.machinelearningplus.com/machine-learning/feature-selection/. [Accessed 12 November 2020].

[40] C. Scheidel, "Understanding Bias in RF Variable Importance Metrics," 20 June 2018. [Online]. Available: https://blog.methodsconsultants.com/posts/be-aware-of-bias-in-rf-variable-importance-metrics/. [Accessed 31 October 2020].

[41] Mawazo, "Combating High Cardinality Features in Supervised Machine Learning," 9 October 2017. [Online]. Available: https://pkghosh.wordpress.com/2017/10/09/combating-high-cardinality-features-in-supervised-machine-learning/. [Accessed 31 October 2020].

[42] T. H. [. c. R. T. B. N. K. T. N. S. J. Q. Jerome Friedman [aut], "Package "glmnet"," 13 June 2020. [Online]. Available: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf. [Accessed 12 November 2020].

[43] T. H. a. J. Qian, "Glmnet Vignette," 26 June 2014. [Online]. Available: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html. [Accessed 12 November 2020].

[44] F. Husson, "Multiple Correspondence Analysis with FactoMineR," 30 November 2016. [Online]. Available: https://www.youtube.com/watch?v=Sl1-UYD6iac. [Accessed 5 November 2020].

[45] S. H. Lan Huong Nguyen, "Ten quick tips for effective dimensionality reduction," 20 June 2019. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907. [Accessed 5 November 2020].

[46] G. Lahera, "Unbalanced Datasets & What To Do About Them," 22 January 2020. [Online]. Available: https://medium.com/strands-tech-corner/unbalanced-datasets-what-to-do-144e0552d9cd. [Accessed 31 October 2020].

[47] M. Amunategui, "SMOTE - Supersampling Rare Events in R," [Online]. Available: http://amunategui.github.io/smote/. [Accessed 25 October 2020].

[48] J. Browniee, "SMOTE for Imbalanced Classification with Python," 21 August 2020. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. [Accessed 31 October 2020].

[49] R. Kunert, "SMOTE explained for noobs - Synthetic Minority Over-sampling TEchnique line by line," 6 November 2017. [Online]. Available: https://rikunert.com/SMOTE_explained. [Accessed 1 December 2020].

[50] N. S. Chauhan, "Decision Tree Algorithm, Explained," January 2020. [Online]. Available: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html. [Accessed 1 November 2020].

[51] S. Rawale, "Understanding Decision Tree, Algorithm, Drawbacks and Advantages.," 30 May 2018. [Online]. Available: https://medium.com/@sagar.rawale3/understanding-decision-tree-algorithm-drawbacks-and-advantages-4486efa6b8c3. [Accessed 1 November 2020].

[52] G. Altay, "Categorical Variables in Decision Trees," March 2020. [Online]. Available: https://www.kaggle.com/gabrielaltay/categorical-variables-in-decision-trees. [Accessed 25 October 2020].

[53] @rakshithvasudev, "What is One Hot Encoding? Why And When do you have to use it?," [Online]. Available: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f. [Accessed 25 October 2020].

[54] "Categorical Data," 6 January 2018. [Online]. Available: https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63. [Accessed 25 October 2020].

[55] T. Yu, "Understanding Random Forest," 12 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed 1 November 2020].

[56] The Stanford Natural Language Processing Group, "Properties of Naive Bayes," 7 April 2009. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html. [Accessed 1 November 2020].

[57] V. Kotu, "Naïve Bayes," 2019. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/naive-bayes. [Accessed 1 November 2020].

[58] Statistics Solutions, "Assumptions of Logistic Regression," [Online]. Available: https://www.statisticssolutions.com/assumptions-of-logistic-regression/. [Accessed 1 November 2020].

[59] J. Le, "Logistic Regression in R Tutorial," 10 April 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/logistic-regression-R. [Accessed 8 November 2020].

[60] "Logistic Regression in R: A Classification Technique to Predict Credit Card Default," 4 November 2019. [Online]. Available: https://blog.datasciencedojo.com/logistic-regression-in-r-tutorial/. [Accessed 8 November 2020].

[61] Basel R Bootcamp, May 2019. [Online]. Available: https://therbootcamp.github.io/ML_2019May/_sessions/Fitting/Fitting.html#13. [Accessed 15 November 2020].

[62] Data Flair, "Kernel Functions-Introduction to SVM Kernel & Examples," [Online]. Available: https://data-

flair.training/blogs/svm-kernel-functions. [Accessed 28 November 2020].

[63] AIQCAR, "12 Support Vector Machine(SVM) Polynomial Kernel Detail Explanation," 3 April 2019. [Online]. Available: https://www.youtube.com/watch?v=Xoz3LeOWOGU. [Accessed 12 December 2020].

[64] Prune, "what is the fundamental solution for Error in knn: too many ties in knn," 26 October 2015. [Online]. Available: https://stackoverflow.com/questions/33344872/what-is-the-fundamental-solution-for-error-in-knn-too-many-ties-in-knn. [Accessed 6 December 2020].

[65] V. Lavrenko, "kNN.11 Breaking ties between nearest neighbors," 15 September 2015. [Online]. Available: https://www.youtube.com/watch?v=dPg-35JQ7Ew. [Accessed 12 December 2020].

[66] Wikipedia, "Occam's razor," [Online]. Available: https://simple.wikipedia.org/wiki/Occam%27s_razor. [Accessed 7 December 2020].

[67] T. Hicks, "A beginner's guide to interpreting odds ratios, confidence intervals and p-values," 13 August 2013. [Online]. Available: https://s4be.cochrane.org/blog/2013/08/13/a-beginners-guide-to-interpreting-odds-ratios-confidence-intervals-and-p-values-the-nuts-and-bolts-20-minute-tutorial/. [Accessed 30 November 2020].

[68] M. Barlaz, "Contrast coding in R," [Online]. Available: https://marissabarlaz.github.io/portfolio/contrastcoding/. [Accessed 2 November 2020].

[69] Wikipedia, "Categorical variable," [Online]. Available: https://en.wikipedia.org/wiki/Categorical_variable#Categorical_variables_and_regression. [Accessed 2 November 2020].

[70] P. Brokmeier, "An Overview of Categorical Input Handling for Neural Networks," 15 January 2019. [Online]. Available: https://towardsdatascience.com/an-overview-of-categorical-input-handling-for-neural-networks-c172ba552dee. [Accessed 2 November 2020].

[71] University of Cincinnati, "Hierarchical Cluster Analysis," [Online]. Available: https://uc-r.github.io/hc_clustering. [Accessed 1 December 2020].

[72] M. Gebeyaw, "Using MCA and variable clustering in R for insights in customer attrition," 21 April 2017. [Online]. Available: https://datascienceplus.com/using-mca-and-variable-clustering-in-r-for-insights-in-customer-attrition/. [Accessed 1 December 2020].

[73] kassambara, "Cross-Validation Essentials in R," 11 March 2018. [Online]. Available: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/#k-fold-cross-validation. [Accessed 1 November 2020].

[74] J. Browniee, "A Gentle Introduction to k-fold Cross-Validation," 3 August 2020. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed 1 November 2020].

[75] GOGIRL Car Insurance, "What's the Difference Between a Dual Carriageway and a Motorway?," [Online]. Available: https://gogirl.co.uk/news-and-advice/dual-carriageway-motorway-difference/. [Accessed 16 October 2020].

[76] Wikipedia, "Great Britain road numbering scheme," [Online]. Available: https://en.wikipedia.org/wiki/Great_Britain_road_numbering_scheme. [Accessed 16 October 2020].

[77] DATA.GOV.UK, "Road Safety Data," 30 September 2020. [Online]. Available: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data. [Accessed 16 October 2020].

[78] Wikipedia, "Multicollinearity," [Online]. Available: https://en.wikipedia.org/wiki/Multicollinearity. [Accessed 7 December 2020].

[79] B. I. D. K. Z. M. J. K. Enis Gegic, "Car Price Prediction using Machine Learning Techniques," 2019. [Online]. Available: http://eds.a.ebscohost.com/eds/pdfviewer/pdfviewer?vid=12&sid=ad524b5e-31e1-4996-8faa-38790b408a1e%40sessionmgr4006. [Accessed 13 October 2020].

[80] M. Cavaioni, "Machine Learning: Unsupervised Learning — Feature selection," 6 February 2017. [Online]. Available: https://medium.com/machine-learning-bites/machine-learning-unsupervised-learning-feature-selection-a9bdccd70f95. [Accessed 14 October 2020].

[81] M. A. A. D. F. M. Alfonso Montella, "Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery," November 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000145751100114X?via%3Dihub. [Accessed 14 October 2020].

[82] Outside Two Standard Deviations, "An overview of correlation measures between categorical and continuous variables," 13 September 2018. [Online]. Available: https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365. [Accessed 15 October 2020].

[83] GOV.UK, "Accident and casualty costs (RAS60)," 2020. [Online]. Available: https://www.gov.uk/government/statistical-data-sets/ras60-average-value-of-preventing-road-accidents. [Accessed 16 October 2020].

[84] S. C. Ifeoma Oduntan, "Comparative Analysis of Classification Accuracy of Six Machine Learning Algorithms on New York City Dataset for Crime Prediction," April 2020. [Online]. Available: https://www.researchgate.net/publication/340863124_Comparative_Analysis_of_Classification_Accuracy_of_Six_Machine_Learning_Algorithms_on_New_York_City_Dataset_for_Crime_Prediction. [Accessed 18 October 2020].

[85] S. K. Gajawada, "Chi-Square Test for Feature Selection in Machine learning," 4 October 2019. [Online]. Available: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223. [Accessed 18 October 2020].

[86] STHDA Statistical tools for high-throughput data analysis, "MCA - Multiple Correspondence Analysis in R: Essentials," 24 September 2017. [Online]. Available: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/. [Accessed 20 October 2020].

[87] IBM, "Multiple Correspondance Analysis," [Online]. Available: https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/categories/idh_mcan.html. [Accessed 20 October 2020].

[88] G. Sanchez, "Multiple Correspondence Analysis in R," [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/2120_cfbb7161b23e494ea47707f5a12032f1.html. [Accessed 20 October 2020].

[89] Wikipedia, "Multiple correspondence analysis," [Online]. Available: https://en.wikipedia.org/wiki/Multiple_correspondence_analysis. [Accessed 20 October 2020].

[90] P. Vivek Yadav, "Building a student intervention system: MCA for dimensionality reduction," 24 June 2016. [Online]. Available: http://vxy10.github.io/2016/06/24/si-mca/. [Accessed 20 October 2020].

[91] A. Reusova, "Hierarchical Clustering on Categorical Data in R," 1 April 2018. [Online]. Available: https://towardsdatascience.com/hierarchical-clustering-on-categorical-data-in-r-a27e578f2995. [Accessed 22 October 2020].

[92] A. Rawat, "SVM Model for GermanCredit dataset," 2017. [Online]. Available: https://rpubs.com/arpitr/svm. [Accessed 23 October 2020].

[93] Data Flair, "e1071 Package – Perfect Guide on SVM Training & Testing Models in R," [Online]. Available: https://data-flair.training/blogs/e1071-in-r/. [Accessed 23 October 2020].

[94] AFIT Data Science Lab R Programming Guide, "Support Vector Machine," [Online]. Available: https://afit-r.github.io/svm. [Accessed 23 October 2020].

[95] R Documentation, "ksvm{kernlab} Support Vector Machines," [Online]. Available: http://finzi.psych.upenn.edu/R/library/kernlab/html/ksvm.html. [Accessed 23 October 2020].

[96] DATA BLOG WORLD, "Support Vector Machines (SVM) in R (package 'kernlab')," [Online]. Available: http://dataworldblog.blogspot.com/2017/08/support-vector-machines-svm-in-r.html. [Accessed 23 October 2020].

[97] Wikipedia, "Feature selection," [Online]. Available: https://en.wikipedia.org/wiki/Feature_selection. [Accessed 25 October 2020].

[98] R. Agarwal, "The 5 Feature Selection Algorithms every Data Scientist should know," 27 July 2019. [Online]. Available: https://towardsdatascience.com/the-5-feature-selection-algorithms-

every-data-scientist-need-to-know-3a6b566efd2. [Accessed 25 October 2020].

[99] H. M. S. S. U. R. a. M. W. Syed Asim Ali Shash, "A Comparative Study of Feature Selection Approaches: 2016-2020," February 2020. [Online]. Available: https://www.researchgate.net/publication/339474097_A_Comparative_Study_of_Feature_Selection_Approaches_2016-2020. [Accessed 25 October 2020].

[100] J. J. H. H. T. F. a. D. L. Yiyou Guo, "SIP-FS: a novel feature selection for data representation," December 2018. [Online]. Available: https://www.researchgate.net/publication/323285330_SIP-FS_a_novel_feature_selection_for_data_representation. [Accessed 25 October 2020].

[101] "Feature Selection with the Caret R Package," 22 August 2019. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/. [Accessed 25 October 2020].

[102] S. Haimoura, "Resampling Methods for Unbalanced Datasets — Fraudulent Transactions," 14 November 2019. [Online]. Available: https://towardsdatascience.com/https-towardsdatascience-com-resampling-methods-for-unbalanced-datasets-5b565d0a247d. [Accessed 25 October 2020].

[103] Analytics Vidhya, "Practical Guide to deal with Imbalanced Classification Problems in R," 28 March 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/. [Accessed 25 October 2020].

[104] dalpozz, "unbalanced," 4 September 2017. [Online]. Available: https://github.com/dalpozz/unbalanced. [Accessed 25 October 2020].

[105] "R package s: unbalanced," 3 March 2014. [Online]. Available: https://www.imsbio.co.jp/RGM/R_function_list?package=unbalanced&init=true. [Accessed 25 October 2020].

[106] S. H. Lan Huong Nguyen, "Ten quick tips for effective dimensionality reduction," 20 June 2019. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907#sec004. [Accessed 26 October 2020].

[107] C. W. B. Z. J. Q. J. F. S. K. K. B.-R. F. B. a. H. B. Nadia Sourial, "Correspondence analysis is a useful tool to uncover the relationships among categorical variables," 22 July 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718710/. [Accessed 27 October 2020].

[108] C. WANG, "How to Tune Support Vector Machine Parameters?," 16 August 2019. [Online]. Available: https://www.rpubs.com/CHENW05/520528. [Accessed 27 October 2020].

[109] T. Bock, "Correspondence Analysis Versus Multiple Correspondence Analysis: Which to Use and When?," [Online]. Available: https://www.displayr.com/correspondence-analysis-versus-multiple-correspondence-analysis-use/. [Accessed 27 October 2020].

[110] Datacamp, "Multiple Correspondence Analysis," [Online]. Available: https://campus.datacamp.com/courses/helsinki-open-data-science/dimensionality-reduction-techniques?ex=10. [Accessed 27 October 2020].

[111] A. Naviani, "Neural Network Models in R," 9 December 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/neural-network-models-r. [Accessed 30 October 2020].

[112] Wikipedia, "Sensitivity and specificity," [Online]. Available: https://en.wikipedia.org/wiki/Sensitivity_and_specificity. [Accessed 31 October 2020].

[113] M. K. a. K. Johnson, "Feature Engineering and Selection: A Practical Approach for Predictive Models," 21 June 2019. [Online]. Available: https://bookdown.org/max/FES/. [Accessed 5 November 2020].

[114] M. Kuhn, "The caret Package," 27 March 2019. [Online]. Available: https://topepo.github.io/caret/index.html. [Accessed 6 November 2020].

[115] G. B. Nongsiej, "What is the meaning of the error message "too many ties in KNN" in R?," 11 September 2014. [Online]. Available: https://www.quora.com/What-is-the-meaning-of-the-error-message-too-many-ties-in-KNN-in-R#:~:text=1.,integers%20in%20your%20feature%20representation.&text=2.-,KNN%20doesn't%20know%20how%20to%20resolve,between%20the%20different%20classification%20labels.. [Accessed 8 November 2020].

[116] J. D. Rosenblatt, "R (BGU course)," 10 October 2019. [Online]. Available: http://www.john-ros.com/Rcourse/parallel.html. [Accessed 8 November 2020].

[117] missinglink.ai, "Classification with Neural Networks: Is it the Right Choice?," [Online]. Available: https://missinglink.ai/guides/neural-network-concepts/classification-neural-networks-neural-network-right-choice/. [Accessed 8 November 2020].

[118] M. Pagels, "Introducing One of the Best Hacks in Machine Learning: the Hashing Trick," 30 December 2017. [Online]. Available: https://medium.com/value-stream-design/introducing-one-of-the-best-hacks-in-machine-learning-the-hashing-trick-bf6a9c8af18f. [Accessed 8 November 2020].

[119] Zhubarb, "Encoding of categorical variables with high cardinality," 6 June 2019. [Online]. Available: https://stats.stackexchange.com/questions/411767/encoding-of-categorical-variables-with-high-cardinality. [Accessed 8 November 2020].

[120] H. Quinlan, "Word Embeddings," [Online]. Available: https://cbail.github.io/textasdata/word2vec/rmarkdown/word2vec.html. [Accessed 8 November 2020].

[121] Google, "Machine Learning Crash Course," [Online]. Available: https://developers.google.com/machine-learning/crash-course/embeddings/obtaining-embeddings. [Accessed 8 November 2020].

[122] J. Brownlee, "A Gentle Introduction to Model Selection for Machine Learning," 2 December 2019. [Online]. Available: https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/. [Accessed 14 November 2020].

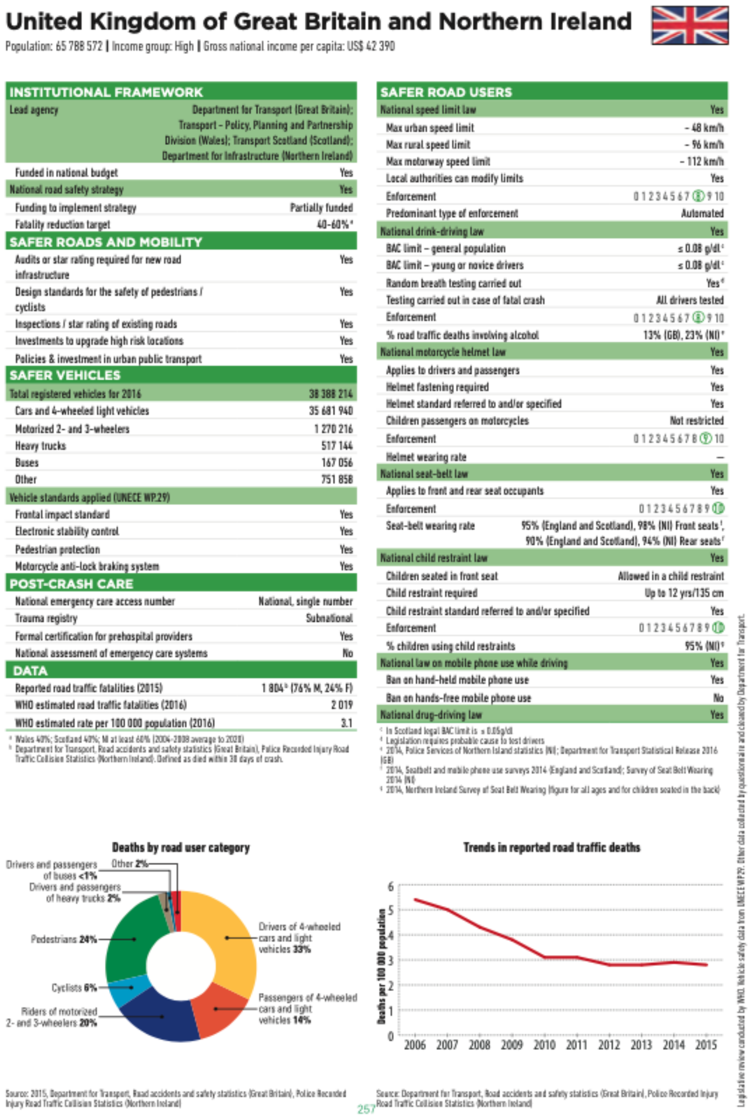
[123] M. J. S. J. C. v. B. E. F. S. M. C. S. a. J. A. K. Zubair Afzal, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," 2 March 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3602667/. [Accessed 15 November 2020].

[124] R. L. P. W. Z. C. Chuanqi Wang, "Partition cost-sensitive CART based on customer value for Telecom customer churn prediction," July 2017. [Online]. Available: https://www.researchgate.net/publication/334129451_Partition_cost-sensitive_CART_based_on_customer_value_for_Telecom_customer_churn_prediction. [Accessed 15 November 2020].

[125] B. B. &. B. Greenwell, "Hands-On Machine Learning with R," 1 February 2020. [Online]. Available: https://bradleyboehmke.github.io/HOML/DT.html. [Accessed 2020 November 2020].

[126] J. Brownlee, "Cost-Sensitive Learning for Imbalanced Classification," 7 February 2020. [Online]. Available: https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/#:~:text=from%20false%20negatives.-,Cost%2Dsensitive%20learning%20is%20a%20subfield%20of%20machine%20learning%20that,algorithm%20modifications%2C%20and%20ensemble%20m. [Accessed 15 November 2020].

APPENDIX I – UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND COUNTRY / AREA PROFILE



United Kingdom of Great Britain and Northern Ireland

Population: 65 788 572 | Income group: High | Gross national income per capita: US$ 42 390

INSTITUTIONAL FRAMEWORK	
Lead agency	Department for Transport (Great Britain); Transport – Policy, Planning and Partnership Division (Wales); Transport Scotland (Scotland); Department for Infrastructure (Northern Ireland)
Funded in national budget	Yes
National road safety strategy	Yes
Funding to implement strategy	Partially funded
Fatality reduction target	40-60% [a]
SAFER ROADS AND MOBILITY	
Audits or star rating required for new road infrastructure	Yes
Design standards for the safety of pedestrians / cyclists	Yes
Inspections / star rating of existing roads	Yes
Investments to upgrade high risk locations	Yes
Policies & investment in urban public transport	Yes
SAFER VEHICLES	
Total registered vehicles for 2016	38 388 214
Cars and 4-wheeled light vehicles	35 681 940
Motorized 2- and 3-wheelers	1 270 216
Heavy trucks	517 144
Buses	167 056
Other	751 858
Vehicle standards applied (UNECE WP.29)	
Frontal impact standard	Yes
Electronic stability control	Yes
Pedestrian protection	Yes
Motorcycle anti-lock braking system	Yes
POST-CRASH CARE	
National emergency care access number	National, single number
Trauma registry	Subnational
Formal certification for prehospital providers	Yes
National assessment of emergency care systems	No
DATA	
Reported road traffic fatalities (2015)	1 804 [b] (76% M, 24% F)
WHO estimated road traffic fatalities (2016)	2 019
WHO estimated rate per 100 000 population (2016)	3.1

[a] Wales 40%; Scotland 40%; NI at least 60% (2004-2008 average to 2020)
[b] Department for Transport, Road accidents and safety statistics (Great Britain), Police Recorded Injury Road Traffic Collision Statistics (Northern Ireland). Defined as died within 30 days of crash.

SAFER ROAD USERS	
National speed limit law	Yes
Max urban speed limit	– 48 km/h
Max rural speed limit	– 96 km/h
Max motorway speed limit	– 112 km/h
Local authorities can modify limits	Yes
Enforcement	0 1 2 3 4 5 6 7 8 9 10
Predominant type of enforcement	Automated
National drink-driving law	Yes
BAC limit – general population	≤ 0.08 g/dl [c]
BAC limit – young or novice drivers	≤ 0.08 g/dl [c]
Random breath testing carried out	Yes [d]
Testing carried out in case of fatal crash	All drivers tested
Enforcement	0 1 2 3 4 5 6 7 8 9 10
% road traffic deaths involving alcohol	13% (GB), 23% (NI) [e]
National motorcycle helmet law	Yes
Applies to drivers and passengers	Yes
Helmet fastening required	Yes
Helmet standard referred to and/or specified	Yes
Children passengers on motorcycles	Not restricted
Enforcement	0 1 2 3 4 5 6 7 8 9 10
Helmet wearing rate	—
National seat-belt law	Yes
Applies to front and rear seat occupants	Yes
Enforcement	0 1 2 3 4 5 6 7 8 9 10
Seat-belt wearing rate	95% (England and Scotland), 98% (NI) Front seats [f], 90% (England and Scotland), 94% (NI) Rear seats [f]
National child restraint law	Yes
Children seated in front seat	Allowed in a child restraint
Child restraint required	Up to 12 yrs/135 cm
Child restraint standard referred to and/or specified	Yes
Enforcement	0 1 2 3 4 5 6 7 8 9 10
% children using child restraints	95% (NI) [g]
National law on mobile phone use while driving	Yes
Ban on hand-held mobile phone use	Yes
Ban on hands-free mobile phone use	No
National drug-driving law	Yes

[c] In Scotland legal BAC limit is ≤ 0.05g/dl
[d] Legislation requires probable cause to test drivers
[e] 2014, Police Services of Northern Island statistics (NI); Department for Transport Statistical Release 2016 (GB)
[f] 2014, Seatbelt and mobile phone use surveys 2014 (England and Scotland); Survey of Seat Belt Wearing 2014 (NI)
[g] 2014, Northern Ireland Survey of Seat Belt Wearing (figure for all ages and for children seated in the back)

Source: 2015, Department for Transport, Road accidents and safety statistics (Great Britain), Police Recorded Injury Road Traffic Collision Statistics (Northern Ireland)

Source: Department for Transport, Road accidents and safety statistics (Great Britain), Police Recorded Injury Road Traffic Collision Statistics (Northern Ireland)

Legislative review conducted by WHO. Vehicle safety data from UNECE WP.29. Other data collected by questionnaire and cleared by Department for Transport.

257

Figure 14: United Kingdom of Great Britain and Northern Ireland Country / Area Profile [1]

APPENDIX II – CLASSIFICATION OF INJURY SEVERITY USING THE CRASH REPORTING SYSTEM

Injury in CRASH	Detailed severity	Severity classification
Deceased	Killed	Killed
Broken neck or back	Very Serious	Serious
Severe head injury, unconscious	Very Serious	Serious
Severe chest injury, any difficulty breathing	Very Serious	Serious
Internal injuries	Very Serious	Serious
Multiple severe injuries, unconscious	Very Serious	Serious
Loss of arm or leg (or part)	Moderately Serious	Serious
Fractured pelvis or upper leg	Moderately Serious	Serious
Other chest injury (not bruising)	Moderately Serious	Serious
Deep penetrating wound	Moderately Serious	Serious
Multiple severe injuries, conscious	Moderately Serious	Serious
Fractured lower leg / ankle / foot	Less Serious	Serious
Fractured arm / collarbone / hand	Less Serious	Serious
Deep cuts / lacerations	Less Serious	Serious
Other head injury	Less Serious	Serious
Whiplash or neck pain	Slight	Slight
Shallow cuts / lacerations / abrasions	Slight	Slight
Sprains and strains	Slight	Slight
Bruising	Slight	Slight
Shock	Slight	Slight

Figure 15: Classification of injury severity using CRASH reporting system [3]

Appendix III – Data set

The dataset only includes the accidents which were reported to the police.

A. Accidents dataset - List of Variables

The variables not in bold were dismissed as were not relevant for the study.

- Accident Index
- Police Force: 51 categories, ranged from 1 to 98 referring to the area.
- Number of Vehicles
- Number of Casualties
- **Date (DD/MM/YYYY):** The day was transformed into month (factor variable with 12 levels for each of the months of the year).
- **Day of Week (int):** The day of the week contained 7 categories from "Monday" to "Saturday".
- **Time (HH:MM):** The hour was extracted to be able to split the variable into 4 categories: "Morning", "Afternoon", "Evening" and "Noon".
- Location Easting OSGR
- Location Northing OSGR
- Longitude
- Latitude
- Local Authority (District): 418 number of districts, from 1 to 941.
- Local Authority (Highway Authority - ONS code): 207 values representing the highway districts.
- **1st Road Class :** 6 categories: (1) Motorway, (2) A(M), (3) A, (4) B, (5) C, (6) Unclassified). A(M) means that the dual carriage road has a hard shoulder and has been upgraded to motorway [75]. The data was regrouped into 4 levels: (B) Big, (M) Medium, (S) Small, and (U) Unclassified [76].
- 2nd Road Class: The variable 2nd Road Class was removed as contains different values than the specified in [77].
- 1st Road Number
- **Road Type:** 8 categories: (1) Roundabout, (2) One-way street, (3) Dual carriage, (6) Single Carriage, (7) Slip road, (9) Unknow, (12) One-way street / Slip road, and (-1) Data missing or out of range. To reduce the number of categories 9 and -1 became unknown and 2 and 12 became one-way streets.
- **Speed limit:** The speed limit goes from 20 to 70 miles per hour (6 categories) and there is a last category for missing data or out of range variables with -1 code.
- **Junction Detail:** 10 categories: (0) Not at junction or within 20 metres, (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, and (-1) Data missing or out of range.
- **Junction Control:** 6 categories: (0) Not at junction or within 20 metres, (1) Authorised person, (2) Auto traffic signal, (3) Stop sign, (4) Give way or uncontrolled, and (-1) Data missing or out of range.
- 2nd Road Number
- **Pedestrian Crossing-Human Control:** 4 categories: (0) None within 50 metres, (1) Control by school crossing patrol, (2) Control by another authorised person, and (-1) Data missing or out of range.
- **Pedestrian Crossing-Physical Facilities:** 7 categories (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, and (-1) Data missing or out of range.
- **Light Conditions:** 6 categories: (1) Daylight, (4) Darkness - lights lit, (5) Darkness - lights unlit, (6) Darkness - no lighting, (7) Darkness - lighting unknown, and (-1) Data missing or out of range.
- **Weather Conditions:** 10 categories: (1) Fine no high winds, (2) Raining no high winds, (3) Snowing no high winds, (4) Fine + high winds, (5) Raining + high winds, (6) Snowing + high winds, (7) Fog or mist, (8) Other, (9) Unknown, and (-1) Data missing or out of range.
- **Road Surface Conditions:** 8 categories (1- Dry, 2- Wet or damp, 3- Snow, 4- Frost or ice, 5- Flood over 3cm. deep, 6- Oil or diesel, 7- Mud, -1- Data missing or out of range)
- **Special Conditions at Site:** 9 categories: (0) None, (1) Auto traffic signal – out, (2) Auto signal part defective, (3) Road sign or marking defective or obscured, (4) Roadworks, (5) Road surface defective, (6) Oil or diesel, (7) Mud, and (-1) Data missing or out of range. 2 and 3 have been grouped together.
- **Carriageway Hazards:** 9 categories: (0) None, (1) Vehicle load on road, (2) Other object on road, (3) Previous accident, (4) Dog on road, (5) Other animal on road, (6) Pedestrian in carriageway - not injured, (7) Any animal in carriageway (except ridden horse), and (-1) Data missing or out of range. 5 and 7 were grouped together.
- **Urban or Rural Area:** 3 categories: (1) Urban, (2) Rural, and (3) Unallocated.

- Did Police Officer Attend Scene of Accident: 3 categories: (1) Yes, (2) No, and (3) No - accident was reported using a self-completion form (self rep only).
- Lower Super Ouput Area of Accident_Location (England & Wales only)

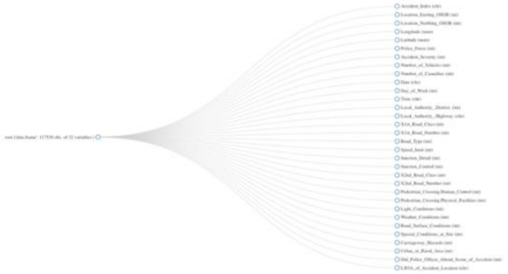


Figure 16: Variables Accidents Dataset

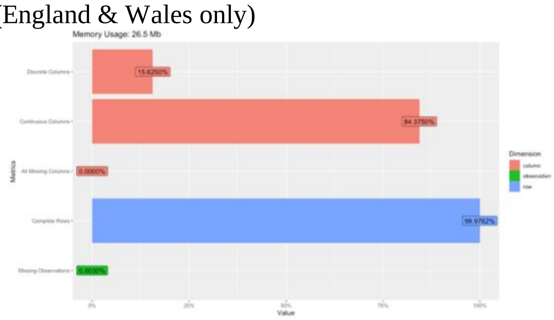


Figure 17: Initial Exploration - Variables Accidents Dataset

The summary of variables can be found at “STATS19 Variable lookup data guide” [77].

APPENDIX IV – ASSOCIATIONS – CRAMER’S TEST GOODMANKRUSKAL PACKAGE

Cramer’s V test values goes from 0 (no association) to 1 (complete association) [35] and is explained as “the square root of a normalized chi-square value” as per [36]. The Cramer’s V test was performed in the stratified sample. By looking at the results, the higher associations were the following:

- Accident severity-urban-rural
- Timeday-light
- Month-light -weather-road
- Road1class-roadtype-speedlimit-junctiondetail-junctioncontrol-crossing-urbanrural
- Weekday
- Roadtype-accidentseverity-road1class-speedlimit-junctiondetail-junctioncontrol-crossing-road-special-carriagehazards-urbanrural

Table 14: Higher Associations in Cramer's Test - GoodmanKruskal package

Only roadtype, speedlimit, crossing and urbanrural has some kind of association with accidentseverity.

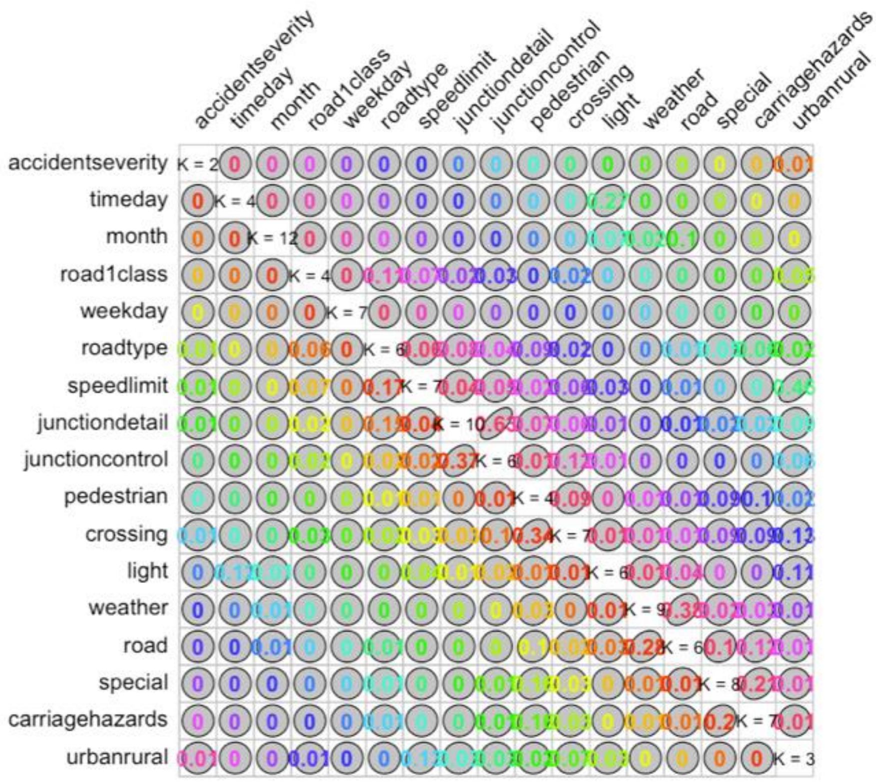


Figure 18: Cramer's V test – GoodmanKruskal package [36]

Appendix V – Correlations

The correlation between the various factors and accident severity was really low. The elements more correlated with the accident severity were the following:

- roadtype_singlecarriage: 0.074
- speedlimit_60: 0.087
- junctiondetail_notjunction: 0.077
- pedestrian_unknown: - 0.072

Table 15: Category Variables most highly correlated with Accident Severity

Some high correlations were seen between categories of the same variable. As none of them is 1 or -1, is ensured not multicollinearity [78].

- speedlimit_70 – road1class_B: 0.63
- junctiondetail_roundabout – roadtype_roundabout: 0.68
- junctioncontrol_unknown – junctiondetail_nojunction: 0.92

Table 16: Category Variables most highly correlated

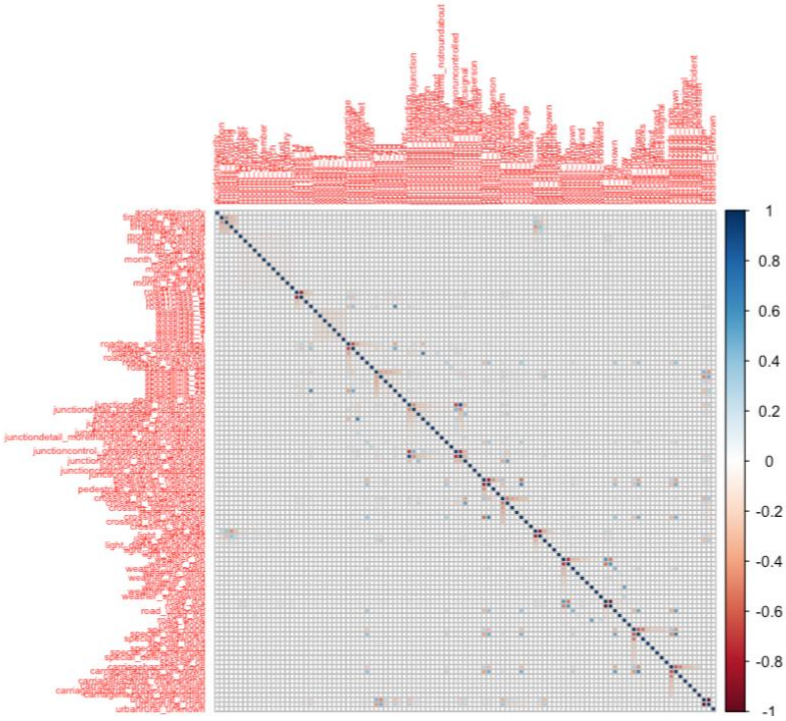


Figure 19: Correlations

Appendix VIII – Feature Selection

The Fselector package [37], revealed as most important variables junctiondetail, speedlimit, roadtype, crossing and urbanrural.

attr_importance		pedestrian	0.07322740	weather	0.04784122
junctiondetail	0.10470530	junctioncontrol	0.07175383	month	0.03971343
speedlimit	0.09571275	light	0.06678310	road	0.03729103
roadtype	0.08924270	special	0.05042110	road1class	0.03053590
crossing	0.08675019	carriagehazards	0.04974397	weekday	0.02604243
urbanrural	0.07572015			timeday	0.02539551

Table 17: Most important Variables as per Fselector package

The Boruta package [38] [39] revealed as most important variables the speedlimit, roadtype, road1class, urbanrural, road, weather, pedestrian and crossing.

	meanImp	medianImp	minImp	maxImp	normHits	decision
timeday	0.6469286	0.7339801	-2.57277837	3.3458614	0.0000000	Rejected
month	-1.2600320	-0.5102172	-3.63914479	0.3957947	0.0000000	Rejected
road1class	16.8769140	17.1473834	6.31330221	27.8683431	1.0000000	Confirmed
weekday	-0.5286061	-0.5908550	-2.60188184	1.5105081	0.0000000	Rejected
roadtype	18.2994752	18.5890821	8.60938892	29.6017878	1.0000000	Confirmed
speedlimit	20.2588924	21.4476570	5.79580740	29.6273096	1.0000000	Confirmed
junctiondetail	-1.6463148	-1.2799449	-3.80643971	-0.2995825	0.0000000	Rejected
junctioncontrol	-8.4773047	-8.5786664	-9.67296560	-6.3109561	0.0000000	Rejected
pedestrian	8.2332207	7.9982504	3.23538852	13.7046391	0.8888889	Confirmed
crossing	7.2687952	7.2388845	1.71209575	14.4807472	0.8888889	Confirmed
light	-1.3021980	-0.9045923	-3.56093903	-0.2391108	0.0000000	Rejected
weather	14.2878786	15.3121302	1.47839760	22.8582889	0.9595960	Confirmed
road	15.6542368	17.0089084	0.06711911	25.1486792	0.9393939	Confirmed
special	-2.1025496	-2.1779149	-4.16672873	0.8033388	0.0000000	Rejected
carriagehazards	2.8938643	2.9210923	-3.92975478	9.9000892	0.3434343	Confirmed
urbanrural	15.8187928	16.8661826	2.13524583	25.4104329	0.9696970	Confirmed

Figure 20: Boruta package results including the meanImp

Pedestrian was the most important variable, followed by roadtype, speedlimit, special, urban rural and crossing, when using the LASSO regression using the glmnet package [42].

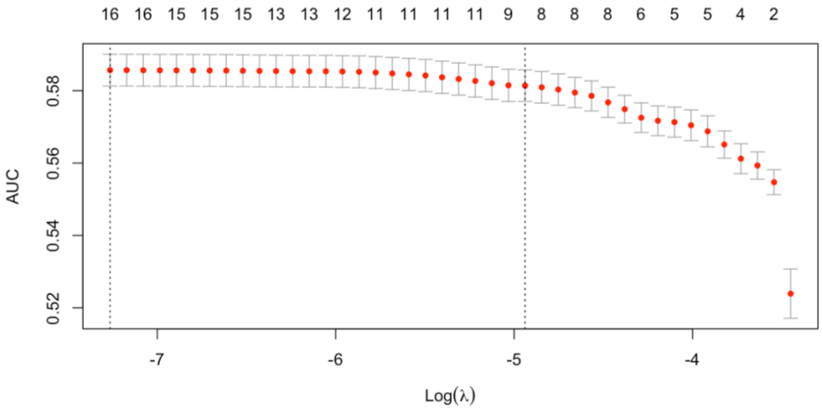


Figure 21: (Area under the ROC curve (AUC) - Number of Variables – Lambda) – LASSO Cross-validation results

(Intercept)	timeday	road1class	roadtype	speedlimit	junctioncontrol	pedestrian	crossing	light
-0.33	0.01	0.03	0.09	0.09	0.06	-0.41	-0.06	-0.03
weather	road	special	carriagehazards	urbanrural				
-0.04	-0.02	-0.08	-0.01	-0.06				

Figure 22: Variables coefficients and intercept LASSO