Classification Performance Evaluation of Traffic Accident Data Using Machine Learning

Efkan Efekan

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Civil Engineering

Eastern Mediterranean University
September 2020
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Civil Engineering.

Prof. Dr. Umut Türker
Chair, Department of Civil Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Civil Engineering.

Assoc. Prof. Dr. Mehmet Metin Kunt
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Tolga Çelik

2. Assoc. Prof. Dr. Mehmet Metin Kunt

3. Asst. Prof. Dr. Şevket Can Bostancı

ABSTRACT

Road traffic accidents, which are a global problem, cause huge losses both in economic and social areas, while traffic accidents lead to casualties, injuries and death. According to the World Health Organization report, more than 1.25 million deaths occur each year as a result of traffic accidents, on the other hand, non-fatal accidents affect more than 20 million people. Although Great Britain has the world's safest road records, research shows that 5 people are killed every day in road traffic accidents. In order to identify the most effective factors related to accidents, researchers have developed and effectively used large data sets containing various information about previous accidents. In this academic study, using the recorded traffic accidents data of Great Britain, statistical models will be used to identify and classify the parameters causing traffic accidents. a detailed procedure of injury severity prediction using the Support Vector Machine, k-Nearest Neighbour and Gaussian Naïve Bayes classification techniques will be discussed. Furthermore, feature selection methods including Chi-square, Random forest, Support vector machine recursive feature elimination and Light gradient boosting machine, will be debated to identify the most important attribute of the traffic accidents. According to the latest available data set in 2018, traffic accidents data, accuracy rate of 77.40% was calculated with the k-Nearest Neighbour method, 78.98% with SVM-RBF and 77.71% with Gaussian Naïve Bayes. As a result of the classification for the severity of casualty, SVM-RBF and GNB often performed the best, giving the same result, at a rate of 87.80%. Classification for vehicle type, the best accuracy value in both test data and training data was obtained with SVM-RBF method with 84.53% and 84.36, respectively. While the percentage of accuracy in the KNN and GNB classification

methods for the test phase was 82.20% and 83.33%, respectively, it was calculated as 82.24% and 82.95%, respectively, as a result of the analysis made for the training phase. Although there are close answers with three classification methods, SVM-RBF classification shows a better performance than other classification tools.

Keywords: Traffic Accident, Machine Learning, Feature Selection, Classification

ÖZ

Küresel bir sorun olan trafik kazaları hem ekonomik hem de sosyal alanlarda büyük kayıplara yol açarken trafik kazaları zayiat, yaralanma ve ölüme neden olmaktadır. Dünya Sağlık Örgütü raporuna göre, trafik kazaları nedeniyle her yıl 1,25 milyondan fazla ölüm meydana gelirken, ölümcül olmayan kazalar 20 milyondan fazla kişiyi etkiliyor. İngiltere dünyanın en güvenli yol kayıtlarına sahip olmasına rağmen, araştırmalar trafik kazalarında her gün 5 kişinin öldüğünü gösteriyor. Kazalarla ilgili en etkili faktörleri tanımlamak için araştırmacılar, önceki kazalar hakkında çeşitli bilgiler içeren büyük veri setleri geliştirmiş ve etkin bir şekilde kullanmışlardır. Bu akademik çalışmada, Büyük Britanya'nın kayıtlı trafik kazaları verileri kullanılarak, trafik kazalarına neden olan parametreleri tanımlamak ve sınıflandırmak için istatistiksel modeller kullanılacaktır. Destek Vektör Makinesi, k-En Yakın Komşu ve Gauss Naïve Bayes sınıflandırma tekniklerini kullanarak yaralanma şiddeti tahmininin ayrıntılı bir prosedürü tartışılacaktır. Ayrıca, trafik kazalarının en önemli niteliğini belirlemek için Ki-kare, Rastgele orman, Destek vektör makinesi özyinelemeli özellik eleme ve Hafif gradyan güçlendirme makinesi gibi özellik seçim yöntemleri tartışılacaktır. 2018 yılında mevcut olan en son verilere göre, trafik kazaları verileri %77,40 doğruluk oranı k-En Yakın Komşu yöntemiyle %78,98 SVM-RBF ile ve %77,41 ile Gauss Naïve Bayes ile hesaplanmıştır. Yaralanmanın ciddiyeti için yapılan sınıflandırma sonucunda SVM-RBF ve GNB genellikle en iyi performansı göstererek aynı sonucu %87,80 oranında verdi. Araç tipi sınıflandırması hem test verilerinde hem de eğitim verilerindeki en iyi doğruluk değeri sırasıyla %84,53 ve %84,36 ile SVM-RBF yöntemi ile elde edilmiştir. Test aşaması için KNN ve GNB sınıflandırma yöntemlerindeki doğruluk yüzdesi sırasıyla %82,20 ile

%83,34 iken, eğitim aşaması için yapılan analiz sonucunda sırasıyla %82,24 ve %82,95 olarak hesaplanmıştır. Üç sınıflandırma yöntemi ile yakın cevaplar olmasına rağmen, SVM-RBF sınıflaması diğer sınıflandırma araçlarından daha iyi bir performans göstermektedir.

Anahtar Kelimeler: Trafik Kazaları, Makine Öğrenimi, Özellik Seçimi, Sınıflandırma

ACKNOWLEDGMENT

I would like to express my gratitude to Assoc. Prof. Dr. Mehmet Metin Kunt for his valuable contribution and suggestion throughout the study. He encouraged me to explore my interests in transportation engineering and find innovations and helped develop ideas about my research. It is a great honour for me to complete my work and research under his guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	Artificial Neural Network
CART	Classification and Regression Tree
DfT	Department of Transport
DTC	Decision Tree Classifier
GB	Great Britain
GNB	Gaussian Naïve Bayes Classifier
KNN	K-Nearest Neighbour Classifier
LightGBM	Light Gradient Boosting Machine
MLP	Multilayer Perceptron
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
SDG	Sustainable Development Goals
SVM	Support Vector Machine
WHO	World Health Organization

Chapter 1

INTRTODUCTION

1.1 Introduction

Nowadays, traffic road accidents are a major problem that continues to cause deaths, injuries and deaths worldwide and cause great economic and social losses. Road accidents have become one of the eighth leading causes of death worldwide in the world health organization's road safety situation report, causing approximately 1.35 million deaths per year, and while more than 20 million suffering severe trauma, many require long-term and costly treatment. Road accidents cause loss of 1-3% of most countries' gross domestic product (World Health Organization, 2018). In addition, this report, published in 2018, underlines that road traffic injuries are the leading cause of death for children and young adults aged 5-29. Moreover, the 10-year Sustainable development Goal11 Target 3.6 (SDG), run by the (UN General Assembly, 2019) aims to halve global deaths and injuries from road traffic accidents by 2020.

Although the number of fatalities in traffic accidents has decreased in recent years in the United Kingdom, which has the best road safety in the world, it still poses a problem. According to statistical information published in the report called "Reported road casualties in Great Britain"(Department of Transportation, 2019). A total of 1,784 people died in road traffic accidents reported in Great Britain in 2018. The number of deaths in 2018 (1,784) was 1% lower compared to 1,793 deaths in

2017. A total of 160,597 casualties were found in traffic accident reported in 2018. This was announced as 6% lower and record low compared to 2017. 25,511 of them were seriously injured casualties while 133,302 were slightly injured casualties. Furthermore, while 279 young people between the ages of 17 and 24 died on motorways, it remained stable compared to 2017, and the number of road deaths among the elderly population over 60 increased from 559 in 2017 to 588 in 2018.

Studies of the extent of the problems of road crashes show that although traffic incidents often have a diverse and often complex history, traffic accident factors are often of similar causes. Human, vehicle, and environmental factors are associated with accident severity. Human factors play a critical role in traffic accidents. Human factors include gender, age, ability to drive, driving style and dangerous driving (drug and alcohol use), over speeding, violating traffic lights, risky driving behaviour (Abu-Zidan & Eid, 2015; Zhang et al., 2014). The vehicle constituent includes the vehicle type, poor technical state of a motor vehicle, and vehicle's skidding on pavement. The most common environmental factors affecting road traffic accidents are weather condition, heavy wind, light conditions. The infrastructural factors under environment are road type, road surface condition, junction location, and road lane type (Beshah & Hill, 2010; Lankarani et al., 2014).

This global problem requires more attention to reduce the severity and frequency of accidents (Alkheder et al., 2017). Historical data on previous accidents represents a tremendous opportunity for researchers to identify the most influential factors in such accidents, which plays a key role in finding appropriate solutions to mitigate this problem in the future. However, extracting information from this data is a very difficult task, as they are typically very large and high in size.

The main objectives of this road accident data classification are to identify the main and main factors causing road traffic crashes and to establish policies and preventive actions to reduce the severity of the accident. Machine learning algorithms are used to analyse data, extract hidden patterns, predict the severity of accidents, and summarize information in a useful format.

1.2 Background of Study

Traffic accident has a great economic effect due to traffic accidents, injury and death reasons. Many researchers attach great importance to identifying common factors that significantly affect traffic accidents and analysis. Scholars are trying to reduce the major effects of possible traffic accidents with limited budget resources, and to base accident precautions and scientific and objective investigation of their causes (Chong et al., 2005). For this purpose, various approaches applied by researchers such as data mining, machine learning, artificial intelligence, data fusion, social networks and so on to investigate historical accident datasets (Bello-Orgaz et al., 2016). In this way, it contributes to the reduction of the number of traffic accidents and the accident severity.

1.3 Objective and Scope of Study

In order to identify the most effective factors related to accidents, researchers have effectively used large data sets containing various information about traffic accidents. The aim of this study is to formulate an algorithm to deduce the major factors involved in an accident as a result of the severity of the injury. The quantitative analysis between injury severity and selected features was heavily researched. The purpose of this study is to examine the causes of serious injuries in the United Kingdom. The main purpose of this thesis is to understand the relationship between the classification method and the determining factors by

reducing the size of the database with feature selection, determining the accident severity and class, vehicle type and the most important factors in the number of casualties. A separate investigation will be carried out because of the high casualties of young and old drivers worldwide. In this way, it is aimed to find the factors that cause accidents for both old and young drivers. In this study, different machine learning classification algorithms are applied such as on Support Vector Machine with radial basis function kernel, Gaussian Naïve Bayes and k-Nearest Neighbour road traffic accident data set obtained from UK road traffic accident in 2014 and 2018.

1.4 Thesis Layout

Chapter 2 presents the previous studies on the using the classification tool of machine learning to predict the severity of injury. Chapter 3 shows the methodology section and describes the data description, general working structure of algorithms used in this study. Chapter 4, analysis section, describes how I apply algorithms in Python language. Chapter 5 presents results and discusses all the analysis results obtained from the Python. Chapter 6 concludes the all of chapters also presents the recommendations for further studies.

Chapter 2

LITERATURE REVIEW

This chapter includes previous studies that use the classification tool of machine learning to predict the severity of injury as a result of traffic accidents and to identify relevant factors causing the accident.

(Ospina-Mateus et al., 2019) analysed traffic accidents and crash and severity related factors in Cartagena, Colombia. They used the classification algorithms of Decision Tree (DT-J48), Rule Induction (PART), Support Vector Machines (SVMs), Naïve Bayes (NB), and Multilayer Perceptron (MLP) to predict the severity of the accident. As a result of the analysis, male and female motorcyclists between the ages of 20-39 predicted that they were more inclined to high-severe accidents.

(Bahiru et al., 2018) used data mining classification techniques J48, ID3, CART and Naïve Bayes to find the factors that caused traffic accidents. According to their results, the accuracy of the J48 classifier is higher than other classifiers, but according to the AUC and ROC results, Naïve Bayes classification accuracy was found better than others, even though its accuracy was lower than the J48 and CART classifiers. They concluded that speed limit, weather and lightning conditions, lane numbers, and accident time are the most important traffic accident factors, on the other hand, gender, age, area where accident occurred, and vehicle type are fewer effective factors of traffic accident severity.

Between 2005 and 2015, (Cigdem & Ozden, 2018) classified two main sub-clusters in Adana on the basis of ten-year accident data that consisted of fatal and non-fatal traffic accidents as a result of vehicle accidents. In this study, they investigated the effect of weather conditions on accident severity by using k-Nearest Neighbor, Naïve Bayes, Multilayer Perceptron, Decision Tree, Support Vector Machine methods. As a result of the analysis, Decision Tree, k-Nearest Neighbor and Multilayer Perceptron based models provided higher accuracy in classification of accidents than other models. While DTC and KNN algorithms performed slightly better in classifying fatal accidents in both datasets, MLP gave the highest accuracy and highest AUC rate in both non-fatal and fatal cases.

(Alkheder et al., 2017) used an artificial neural network (ANN) to estimate the severity of injury (death, severe, moderate and minor severe accidents) of traffic accidents by handling the 5973 traffic accident records that took place in Abu Dhabi from 2008 to 2013. Using 90% of the data set for training and 10% for testing purposes, ANN estimation performances were 81.6% and 74.6%, respectively. Based on the training data determined for death, severe, moderate and mild accidents, ANN prediction accuracy was 4.5%, 10.2%, 80.1% and 94.5% respectively. According to the test dataset, ANN prediction accuracy for death, severe, moderate and minor severe accidents was 0%, 0%, 78.4% and 82% respectively.

(Kumar & Toshniwal, 2017) used three popular classification algorithms, Classification and Regression Tree (CART), Naïve Bayes and Support Vector Machine, to analyse the Powered Two-Wheeled road accident data from Uttarakhand state in India and in various regions of Uttarakhand to identify the factors affecting the severity of these accidents. After the analysis, the classification accuracy of

CART (87.10%) was found to be better than the other two techniques according to the data in the entire Uttarakhand state.

(Iranitalab & Khattak, 2017) used Multinomial Logit (MNL), Nearest Neighbour Classification (NNC), Support Vector Machines (SVM) and Random Forests (RF) classification methods to investigate the crash severity estimation of traffic accidents in Nebraska, USA between 2012-2015. They stated that NNC has the best prediction performance in more severe crashes. In addition, RF and SVM had two other adequate performances, but MNL indicated that it was the weakest accuracy method.

(Castro & Kim, 2016), conducted a study using Bayesian network, decision trees and artificial neural networks to investigate the role of different factors in injury risk and identify the most common factors in an accident. According to the result, they showed that the three most common factors were light conditions, vehicle manoeuvre and road type. In addition, researchers found that the vehicle's age and weather conditions had no significant effect on the severity of injury.

(Al-Turaiki et al., 2016) used CHAID, J48 and Naïve Bayes classification techniques to determine the factors causing the severity of traffic accidents in Riyadh, Saudi Arabia. As a result of the study, he stated that distraction during vehicle use is an important factor leading to injuries and deaths, and the age of the car is also an important factor.

Based on two-year accident data in New Mexico, (Chen et al., 2016) used support vector machine (SVM) models to further understand and investigate the effects on

driver injury severity in tipping accidents. With the classification and regression tree (CART) model, they identified important factors for predicting driver injury severity.

Above, many studies have been mentioned to understand the severity of the traffic accidents by using various classification tools of machine learning. Researchers used classification methods such as Support Vector Machines (SVMs), Naïve Bayes (NB), and Multilayer Perceptron (MLP), Random Forests (RF), k-Nearest Neighbour (KNN). In this thesis study, Support Vector Machines with radial basis function kernel (SVM-RBF), Gaussian Naïve Bayes (GNB) and k-Nearest Neighbour (KNN) are discussed. KNN is extremely easy to implement in its most basic form and still performs highly complex classification tasks. It is a lazy learning algorithm since it has no special training stage and therefore does not require training before making real-time predictions. This makes the KNN algorithm much faster than other algorithms that require training, such as SVM. However, the KNN algorithm does not work well with high-dimensional data because with multiple dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension. Since KNN cannot work well with high dimensional data, SVM method has been used to close this gap. SVM is more effective in high-dimensional areas. SVM works relatively well when there is a clear division between classes. It is effective in cases where the size number is more than the number of samples. Naïve Bayes classification method is very simple, easy to apply and fast. High performance can be achieved with less training data and is not sensitive to irrelevant features. It can also be used for binary and multi-class classification problems.

Chapter 3

METHODOLOGY

3.1 Data and Methods

In this section, a detailed procedure of injury severity prediction using the Support Vector Machine, k-Nearest Neighbour and Gaussian Naïve Bayes classification techniques will be discussed. Furthermore, feature selection methods including Chi-square, Random forest, Support vector machine recursive feature elimination and Light gradient boosting machine, will be debated to identify the most important attribute of the traffic accidents.

3.2 Data Description

In this study, the data set used between 2014-2018, the traffic accident data published by the UK Department of Transport (DfT) and open to the public through the STATS19 database were used. The STATS19 data system contains 3 main files: accident, casualty and, vehicle. In addition, for this database reporting system, victims are classified as slight injury, serious injury or death.

All traffic accidents occurring on the highway and reported to the police within 30 days and causing death or personal injury to one or more vehicles are formed. This report includes the severity and location of traffic accidents that occur on roads where the public has motor vehicle access, except for private roads. Also included in this report are accidents when pedestrians are getting on or off the bus, and cyclists and horse riders injuring themselves or a pedestrian.

In the file to be reported for vehicles, it contains the details of the vehicle completely regardless of whether the vehicle is damaged for each vehicle involved or caused by an injury accident. The incidents in which the driver, rider and passenger were injured in the vehicles damaged in the accident are the examples in the vehicle reports.

In the report, casualty is reported for people who died or were injured in a traffic accident. Full details of the reports on which topics were included are set out in appendix.

3.3 Data Pre-Processing

Too much irrelevant and unnecessary information or noisy and incomplete data in the database may greatly mislead the analysis result. Therefore, various data pre-processing methods have been developed in order to eliminate this problem to obtain higher accuracy. There are many application methods such as data cleaning, data conversion, and data reduction, which are steps related to data pre-processing. In this thesis, feature selection, which is one of the steps of data conversion, will be used to determine the most important features in the data set. Since some features are unrelated to the scope of study, irrelevant data such as identity, latitude, longitude, time, year of accidents in the data set, and missing or unknown data were extracted before using the feature selection method. However, due to the unbalanced distribution of the available data, while analysing the factors that cause accidents in the UK an optimum yield results ranging between 80-90% can be achieved. Vehicles, casualties, accidents are shown in Table 1, Table 2, and Table 3 respectively. For example, in the vehicles file, 41.70% of 80.19% of traffic accidents that occur at intersection locations are “Not at or within 20m of junction”, 22.04%

are “Entering roundabout” and 16.45% are “Mid junction- on roundabout or on main road”. These parameters are represented as 0,1 and 8, respectively. The factors of the other factors discussed are given in the tables below.

Table 1: Factors of Vehicles

Factors	**Labels**	**Code**	**Percentage of Parameters (%)**
Vehicle Reference Number	Number of accident vehicles	1	54.15
	Number of accident vehicles	2	38.33
	Number of accident vehicles	3	5.60
Vehicle Type	Car	9	70.42
	Pedal cycle	1	8.01
	Van/Goods vehicle 3.5 tonnes maximum gross weight (mgw) and under	19	5.33
Vehicle Manoeuvre	Going ahead other	18	47.30
	Turning right	9	8.99
	Slowing or stopping	4	6.61
	Waiting to go ahead but held up	3	5.29
	Moving off	5	4.65
	Parked	2	4.25
	Going ahead right-hand bend	17	3.15
Skidding and Overturning	No skidding, jack-knifing or overturning	0	85.03
	Skidded	1	6.26
Junction Location	Not at or within 20 metres of junction	0	41.70
	Approaching junction or waiting/parked at junction approach	1	22.04
	Mid junction – on roundabout or on main road	8	16.45
1st Point of Impact	Front	1	49.26
	Back	2	16.91
	Offside	3	13.53
Sex of Driver	Male	1	63.32
	Female	2	27.05
Age Band of Driver	26 – 35	6	21.23
	36 – 45	7	16.66
	46 – 55	8	15.30
	21 – 25	5	9.75
	56 – 65	9	9.25
	16 – 20	4	6.47
	66 – 75	10	4.68
Vehicle Location Restricted Lane	On main c’way – not in restricted lane	0	92.58

Table 2: Factors of Casualties

Factors	Labels	Code	Percentage of Parameters (%)
Vehicle Reference	Number of accident vehicles	1	56.81
	Number of accident vehicles	2	39.78
Casualty Reference	Number of casualties	1	75.67
	Number of casualties	2	16.22
Casualty Class	Driver or rider	1	64.37
	Passenger	2	21.67
	Pedestrian	3	13.97
Sex of Casualty	Male	1	59.31
	Female	2	40.66
Age Band of Casualty	26 – 35	6	20.70
	36 – 45	7	15.08
	46 – 55	8	13.98
	21 – 25	5	11.33
	16 – 20	4	9.98
	56 – 65	9	8.74
	66 – 75	10	5.29
Casualty Severity	Fatal	1	64.37
	Serious	2	21.67
	Slight	3	13.97
Car Passenger	Not car passenger	0	81.58
	Front seat passenger	1	11.24
	Rear seat passenger	2	6.89
Casualty Home Area Type	Urban	1	72.19
	Rural	3	10.33
	Small town	2	7.81

Table 3: Factors of Accidents

Factors	Labels	Code	Percentage of Parameters (%)
Accident Severity	Slight	3	79.75
	Serious	2	18.89
	Fatal	1	1.36
Number of Vehicles	Number of accident vehicles	2	60.44
	Number of accident vehicles	1	29.23
Number of Casualties	Number of casualties	1	79.21
	Number of casualties	2	14.43
Road Type	Single carriageway	6	72.02
	Dual carriageway	3	15.87
Junction Detail	Not at junction or within 20 metres	0	42.46
	T or staggered junction	3	29.32
	Crossroads	6	9.31
Pedestrian Crossing-Physical Facilities	No physical crossing facilities within 50 metres	0	77.37
	Pedestrian phase at traffic signal junction	5	7.95
Light Conditions	Daylight	1	72.11
	Darkness – lights lit	4	20.18
Weather Conditions	Fine no high winds	1	80.91
	Raining no high winds	2	10.43
Road Surface Conditions	Dry	1	73.83
	Wet or damp	2	23.01
Speed limit	30 MPH	30	59.92
	60 MPH	60	12.55
	20 MPH	20	8.69
Urban or Rural Area	Urban	1	67.34
	Rural	2	32.61

3.4 Feature Selection for Classification

The feature selection that affects the performance of the Machine Learning model very much is the process of selecting and finding the most useful features in the data set. Attribute selection can basically identify attributes that do not work or work less well for existing attributes in the data set. Feature Selection provides an effective way to increase the forecast rate by removing irrelevant and unnecessary data, which can reduce and improve computation time.

Three main methods are used for feature selection in classification.

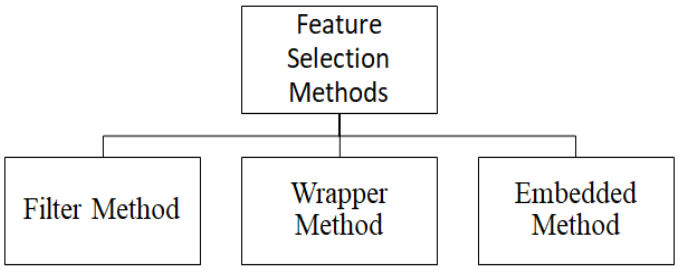


Figure 1: Feature Selection Methods

1) Filter Methods: There is a sharp mathematical criterion for evaluating the quality of a feature or a subset of features. This criterion is then used to filter irrelevant features.
2) Wrapper Methods: A classification algorithm is assumed to be available to evaluate how well the algorithm performs with a subset of features. A feature search algorithm is then wrapped around that algorithm to determine the respective feature set.
3) Embedded Methods: The solution of a classification model usually contains useful tips on the most relevant features. These features are isolated, and the classifier is retrained on the pruned features.

In this study, 4 different methods were applied to specify the related features from a dataset and to remove the irrelevant or partially related features from the dataset, thereby removing the negative performance on the model and obtaining better accuracy from the model. These methods are Chi-square, Random Forest, Linear Support Vector Machine and Light Gradient Boosting Machine.

3.4.1 Chi-squared

Chi-square is a numerical test that measures the deviation from the expected distribution given that the property event is independent of the class value. The chi square value is calculated from the following metrics such as true positives (tp), false positives (fp), true negatives (tn), false negatives (fn), probability of positive case count (Ppos), and probability of negative case number (Pneg)(Ikram & Cherukuri, 2017).

$$\begin{aligned} chi - square\_metric = {} & t(t_p, (t_p + f_p)P_{pos}) + t(f_n, (f_n \\ & + t_n)P_{pos}) + t(f_p, (t_p + f_p)P_{neg}) \\ & + t(t_n, (f_n + t_n)P_{neg}) \end{aligned} \quad (1)$$

where t (count, expect) = (count – expect)2/expect.

The chi-square approach consists of the following steps:

1) Specify the hypothesis
2) Devise an analysis plan
3) Examine sample data
4) Deduce results.

3.4.2 Support Vector Machine Recursive Feature Elimination (SVM-RFE)

Support vector machine (SVM) is a popular and efficient classification technique and is widely applied in many transportation systems such as traffic flow prediction (Li & Xu, 2020), vehicular traffic density estimation (Tyagi et al., 2012), railway electrification system (Jung et al., 2016) and public transportation planning system (Ul Haq et al., 2020). SVM recursive feature elimination (SVM-RFE) is a feature selection algorithm based on SVM. When creating the SVM learning model, the weights of the features are also calculated. SVM-RFE removes features with the lowest weights repeatedly. The order of feature removal represents the feature importance order (Guyon et al., 2002).

3.4.3 Random Forest

Random forest (RF) provides feature significance measurements as one of its useful derivatives. RF consists of 4-12 hundred decision trees, each built on random extraction of observations from the dataset and random extraction of features. The overall property significance is calculated as a reduction in node impurity weighted by the probability of reaching that node. When going deep into tree levels, the node impurity should be reduced, and therefore the effect of the node can be objectively measured by decreasing the impurity across the node. The Gini impurity is calculated for each node where it is possible to calculate the probability of the node, based on the number of samples reaching the node divided by the total number. In this case, higher values correspond to more important features (AlSagri & Ykhlef, 2020).

The general feature significance begins as follows:

1) Calculating nodes importance n_j of node j for every decision tree.

$$n_j = W_j C_j - W_{left\,(j)} C_{left\,(j)} - W_{right\,(j)} C_{right\,(j)} \quad (2)$$

- W_j: Node j reachability probability
- C_j : Gini impurity of node

2) Calculating the importance of each feature (F) in the tree.

$$F_j = \frac{n_j}{\sum_{i=1}^{m} n_i} \quad (3)$$

3) Calculating the importance of each feature in Random Forest.

$$Feature\ Importance\ (i) = \frac{\sum_{j=1}^{m} F_j}{k} \quad (4)$$

3.4.4 Light Gradient Boosting Machine (LightGBM)

LightGBM has a significant performance improvement, faster training rate, lower memory requirements, higher accuracy. In the traditional GBDT algorithm, the most time-consuming step is to find the most suitable partition point. According to the traditional solution, Pre-Sequence processing is used to enumerate all potential feature points according to pre-ordered feature values, while LightGBM replaces the traditional Pre-Sequence processing with the histogram algorithm(Zhang et al., 2020). LightGBM sorts the most suitable solution categories into 2 sub-clusters and divides them into a categorical feature and sorts the categories according to the training target in each department. The LightGBM sorts the histogram by its accumulated values (sum_gradient / sum_hessian) and then finds the best split in the histogram listed (Microsoft, 2020). In addition, LightGBM adds a limit to the depth of the tree based on the traditional Leaf-wise strategy to find the best split gain node. In this way, the algorithm provides high efficiency and prevents the problem of over-fitting due to the very deep tree structure.

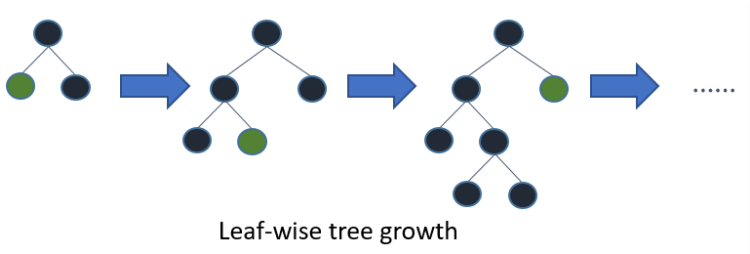


Figure 2: LightGBM Leaf-Wise Tree Growth (Microsoft, 2020)

3.5 Machine Learning Classification Techniques

After determining the factors related to the target variable with feature selection analysis, several classification tools of machine learning were used to calculate the estimated accuracy of these factors. Information about the methods used below is given.

3.5.1 Support Vector Machine Radial Basis Function (SVM-RBF)

SVM is a supervised machine learning algorithm that can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is often used for a smaller data set because the processing of the given training data takes more time than other classification methods (Liu et al., 2017). SVM is based on the idea of finding a hyper plane that best separates the features into different areas. The basic intuition developed here is the possibility of accurately classifying points in their respective regions or classes, no matter how far the support vector points are from the hyper plane.

Below are the equations of SVM that have been developing over the years.

$$(x_i, y_i), x_i \in R^n, y_i \in \{1, -1\}, i = 1,2, \dots \dots, \dots l \quad (5)$$

To find the most suitable solutions for w and b parameters, the following optimization problem should be solved with Lagrange (Acı & Ozden, 2018).

$$\min_{\mathrm{w,b}} \frac{1}{2} = w^T \mathrm{w} + \mathrm{c} \sum_{i=1}^{l} \delta(w, b; x_i, y_i) \quad (6)$$

$\delta(w, b; x_i, y_i)$ represents slack (misspecification) variable and $C \geq 0$ are the specified penalty parameter of the error term.

Two commonly used slack (misspecification)are given in Equation (7) and Equation (8);

$$max(1 - y_i)(w^T \sigma(x_{i)} + b), 0) \quad (7)$$

$$max(1 - y_i)(w^T \sigma(x_{i)} + b), 0)2 \quad (8)$$

Here, σ represents the function used to move the training data into a higher dimensional space. The decision function for each x test data is given in Equation (9);

$$f(x) = sgn(w^T \sigma(x) + b) \quad (9)$$

Finally, RBF kernel was used in this study.

$$K\left((x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (10)$$

$$\gamma = {}^{1}/_{2\sigma^2} \quad (11)$$

In the equation, $\|x_i - x_j\|^2$ shows the vector distance. γ parameter is free variable and determines the width of the function.

3.5.2 K-Nearest Neighbor Classifier

The Nearest Neighbour Classifier ensures consistently high performance without prior assumptions about data distributions in training samples. The closest neighbour classifier can be used with almost any type of data, if a suitable distance function is

available. The basic approach is the same as in multi-dimensional data. For any test sample, the k-closest neighbours in the training data are determined. The raid tag of these nearest neighbours is reported as the corresponding class tag. Large k values help reduce the effects of noisy points in the exercise data set, and the most appropriate k selection is usually done through cross-validation) (Liao & Vemuri, 2002). Having a versatile, simple and easy-to-implement algorithm, KNN does not need to adjust a few parameters or make additional assumptions while creating a model. However, as the number of samples, predictors and independent variables increases in the KNN algorithm, the algorithm slows down significantly.

3.5.3 Gaussian Naïve Bayes

Naive Bayes Classifiers are based on Bayes Theorem. These classifiers assume that the value of a particular property is independent of the value of any other property. In a controlled learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayes classifiers have a simple application in solving problems in many real-life situations, as they need a little training data to estimate the parameters required for classification. A frequent assumption when working with continuous data is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The Gaussian Naive Bayes divide the training data into classes by class and calculate the average and variance of each class. The following formula can be used to estimate the possibilities of a continuous dataset.

$$P\left(\frac{x_i}{y}\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}) \quad (12)$$

Where x_i = dependent variable, y = class variable. The parameters σ_y and μ_y are estimated using maximum likelihood.

3.6 Performance Metrics

The performance of the classifier model is defined from a matrix known as the confusion matrix, which shows true and misclassified samples for each class. Confusion matrix is an important data structure that helps calculate different performance measurements on specific data such as accuracy, f1-measure, recall and precision of the classification technique.

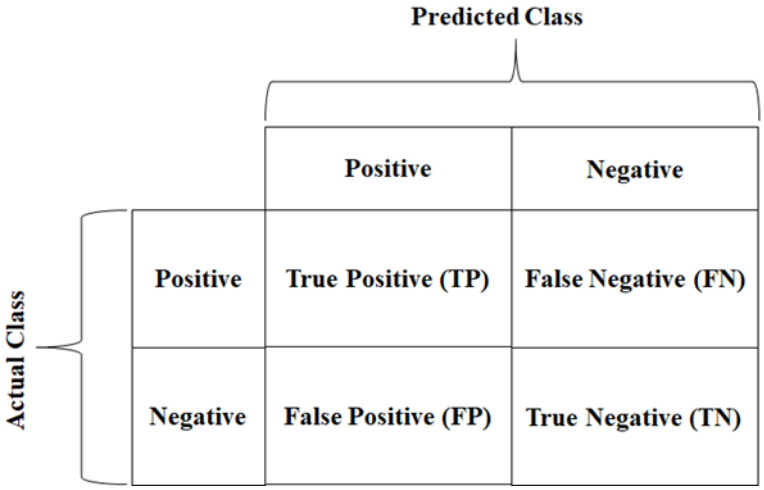


Figure 3: Performance Metrics

$$Accuracy = (\frac{TP + TN}{TP + FP + TN + FN}) \quad (13)$$

$$Precision = (\frac{TP}{TP+FP}) \quad (14)$$

$$Recall = (\frac{TP}{TP+FN}) \quad (15)$$

$$F1 - measure = (\frac{2 * Recall * Precision}{Recall + Precision}) \quad (16)$$

Chapter 4

ANALYSIS

The fourth part of the study consists of two main sub-sections. In the first part, feature selection analysis was done by using the data of traffic accidents in England in 2018. Also, in line with the result obtained from the feature selection in this section, if the sex and age of the drivers appear as a relevant parameter, a separate subset was created for them and a review was made within them. In the second part, after the feature selection, various classification tools of machine learning are used by selecting only the relevant features according to the target variable. The data set used here is traffic accident data between 2014-2018. The accuracy performance result of the matrix is specified with f1- measure, recall, precision.

The process used to predict and evaluate the findings was implemented in Python language with the machine learning module called Sci-kit learn which is where the algorithms were produced to find the results (Pedregosa et al., 2011). Sci-kit is an open source machine learning library that supports both supervised and unsupervised learning. It also provides various tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities.

In supervised learning, models use input data and target outputs (tags) to learn the function or map between them, and the invisible outputs can be estimated by combining the learned model with the input data. On the other hand, in unsupervised learning, unlabelled data focuses on natural learning from multidimensional (Mitchell et al., 2013).

It was used with Scikit-learn to create a predictive model to determine which factors are most important in traffic accidents.

```
In [1]: import matplotlib.pyplot as plt
        import numpy as np
        import pandas as pd
        import seaborn as sns
        import warnings
        from sklearn.feature_selection import SelectKBest
        from sklearn.feature_selection import SelectFromModel
        from sklearn.feature_selection import chi2
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.ensemble import ExtraTreesClassifier
        from sklearn.svm import LinearSVC
        from sklearn.preprocessing import MinMaxScaler
        from sklearn.feature_selection import RFE
        from sklearn.feature_selection import RFECV
        from sklearn.metrics import accuracy_score, f1_score
        from sklearn.model_selection import train_test_split
```

Figure 4: Importing Algorithms from Python Library

First, the read_csv () method provided by the pandas' module was used to read the csv file containing comma separated values and convert it to the pandas' DataFrame.

Pandas is used to perform activities such as loading and saving data, adding and deleting columns, deleting rows, selecting data, renaming columns and rows, and sorting data.

```
In [2]: df = pd.read_csv('C:/Users/efkan/OneDrive/Masaüstü/DATA/dftRoadSafetyData_Accidents_2018.csv')
        df.head()
```

```
C:\Users\efkan\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3058: DtypeWarning: Columns (0) have mixed types. S
pecify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[2]:

	Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of_(
0	2018010080971	529150.0	182270.0	-0.139737	51.524587	1	3	2	
1	2018010080973	542020.0	184290.0	0.046471	51.539651	1	3	1	
2	2018010080974	531720.0	182910.0	-0.102474	51.529746	1	3	2	
3	2018010080981	541450.0	183220.0	0.037828	51.530179	1	2	2	
4	2018010080982	543580.0	176500.0	0.065781	51.469258	1	2	2	

5 rows × 32 columns

Figure 5: Reading Data Set with Pandas

After that, irrelevant data such as accident index, latitude, longitude, local authority, date, time, day of week was dropped from the dataset.

```
In [3]: df=df.drop(['Accident_Index','Location_Easting_OSGR','Location_Northing_OSGR','Longitude','Latitude','L
                   'LSOA_of_Accident_Location','Police_Force','Did_Police_Officer_Attend_Scene_of_Accident','L
        df.head()
```

Out[3]:

	Accident_Severity	Number_of_Vehicles	Number_of_Casualties	Road_Type	Speed_limit	Junction_Detail	Pedestrian_Crossing Human_Contr
0	3	2	2	3	30	0	
1	3	1	1	6	30	2	
2	3	2	1	6	20	6	
3	2	2	1	3	30	7	
4	2	2	2	6	30	0	

Figure 6: Removing Unrelevant Factors

Then, filtering was performed by selecting factors that are generally between 80-90% of the total of factors in each parameter with DataFrame.isin () method, which provides filtering task in Pandas.

```
In [18]: keep1 = ['1','2']
         keep2 = ['3','6']
         keep3 = ['30','60','20']
         keep4 = ['0','3','6']
         keep5 = ['0','5']
         keep6 = ['1','4']
         keep7 = ['1','2']
         keep8 = ['1','2']
         keep9 = ['1','2']

In [19]: filter1 = df["Number_of_Vehicles"].isin(keep1)
         filter2 = df["Road_Type"].isin(keep2)
         filter3 = df["Speed_limit"].isin(keep3)
         filter4 = df["Junction_Detail"].isin(keep4)
         filter5 = df["Pedestrian_Crossing_Physical_Facilities"].isin(keep5)
         filter6 = df["Light_Conditions"].isin(keep6)
         filter7 = df["Weather_Conditions"].isin(keep7)
         filter8 = df["Road_Surface_Conditions"].isin(keep8)
         filter9 = df["Number_of_Casualties"].isin(keep9)

In [20]: df2= df[filter1 & filter2 & filter3 & filter4 & filter5 & filter6 & filter7 & filter8 & filter9]
```

Figure 7: DataFrame.isin Method

In the new DataFrame formed, the target variable and dependent variable were selected and the whole data set was divided into two groups. The first is the training set that trains the algorithm to generate a model, and the other is the test set of the model being tested to understand how accurate its predictions are.

Using the train_test_split () method in Scikit-learn to divide the data set into two groups, 30% of the data was selected as test data and the remaining 70% as training data.

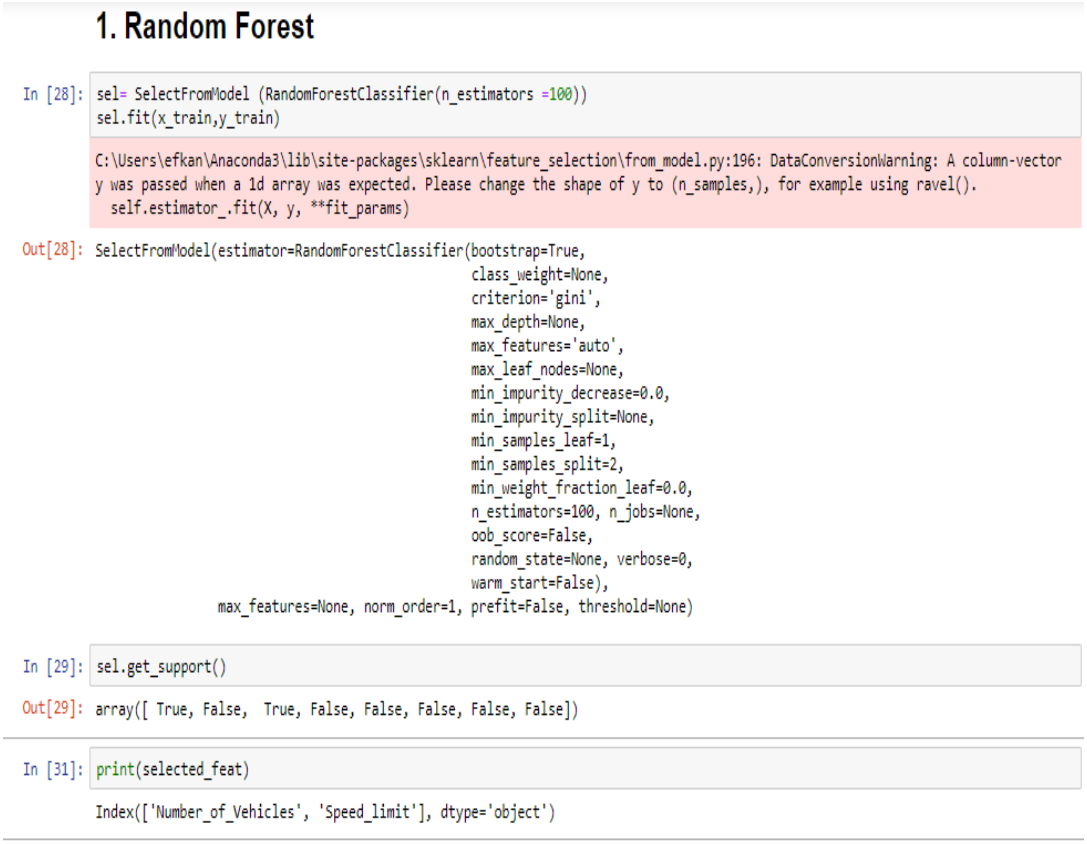
```
target = 'Accident_Severity'
```

```
X = df2.loc[:, df2.columns != target]
Y = df2.loc[:, df2.columns == target]
```

```
x_train, x_test, y_train, y_test = train_test_split(X, Y,
test_size=0.10,
random_state=0)
```

Figure 8: Selecting and Target Splitting Dataset

Feature selection algorithms are trained using the fit () method on test data (x_train) and test target (y_train) and optimum results are obtained in relation to the target output. These results are shown as true or false and the 4 different features used are ranked according to the total number of correct in the selection. In this way, it is understood which factors are how important.



1. Random Forest

```
In [28]: sel= SelectFromModel (RandomForestClassifier(n_estimators =100))
         sel.fit(x_train,y_train)

         C:\Users\efkan\Anaconda3\lib\site-packages\sklearn\feature_selection\from_model.py:196: DataConversionWarning: A column-vector
         y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
           self.estimator_.fit(X, y, **fit_params)

Out[28]: SelectFromModel(estimator=RandomForestClassifier(bootstrap=True,
                                                          class_weight=None,
                                                          criterion='gini',
                                                          max_depth=None,
                                                          max_features='auto',
                                                          max_leaf_nodes=None,
                                                          min_impurity_decrease=0.0,
                                                          min_impurity_split=None,
                                                          min_samples_leaf=1,
                                                          min_samples_split=2,
                                                          min_weight_fraction_leaf=0.0,
                                                          n_estimators=100, n_jobs=None,
                                                          oob_score=False,
                                                          random_state=None, verbose=0,
                                                          warm_start=False),
                                  max_features=None, norm_order=1, prefit=False, threshold=None)

In [29]: sel.get_support()

Out[29]: array([ True, False,  True, False, False, False, False, False])

In [31]: print(selected_feat)

         Index(['Number_of_Vehicles', 'Speed_limit'], dtype='object')
```

Figure 9: Fitting Method

All Algorithms in a Tablo

In [40]:
```
# put all selection together
feature_selection_df = pd.DataFrame({'Feature':feature_name, 'Chi-2':chi_support, 'Random Forest': sel.get_support() ,'SVM_Linea
# count the selected times for each feature
feature_selection_df['Total'] = np.sum(feature_selection_df, axis=1)
# display the top 100
feature_selection_df = feature_selection_df.sort_values(['Total','Feature'] , ascending=False)
feature_selection_df.index = range(1, len(feature_selection_df)+1)
feature_selection_df.head(10)
```

Out[40]:

	Feature	Chi-2	Random Forest	SVM_Linear	LightGBM	Total
1	Number_of_Vehicles	True	True	True	True	4
2	Speed_limit	True	True	False	True	3
3	Urban_or_Rural_Area	True	False	True	False	2
4	Weather_Conditions	False	False	True	False	1
5	Light_Conditions	False	False	False	True	1
6	Junction_Detail	False	False	False	True	1
7	Road_Type	False	False	False	False	0
8	Road_Surface_Conditions	False	False	False	False	0

Figure 10: Ranking of Factors According to Their Importance

After the feature selection, unrelated factors were determined and removed from the dataset, and the targets in the test data were estimated using the predict () method. Finally, the score was obtained using the accuracy_score () method.

1. KNN classifier

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(x_train, y_train)
knn_predictions = knn.predict(x_test)
print(accuracy_score(y_test, knn_predictions))
print(confusion_matrix(y_test, knn_predictions))
print(classification_report(y_test, knn_predictions))
```

```
C:\Users\efkan\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: DataConversionWarning: A column-vector y was passed when a
1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  This is separate from the ipykernel package so we can avoid doing imports until
```

```
0.7619249523001907
[[   0    2   51]
 [   0   36  883]
 [   0  187 3558]]
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        53
           2       0.16      0.04      0.06       919
           3       0.79      0.95      0.86      3745

    accuracy                           0.76      4717
   macro avg       0.32      0.33      0.31      4717
weighted avg       0.66      0.76      0.70      4717
```

Figure 11: Classification Accuracy Score Algorithms

All the steps in the supervised learning algorithm were mentioned above. The complete representation of these steps is shown on the flowchart below.

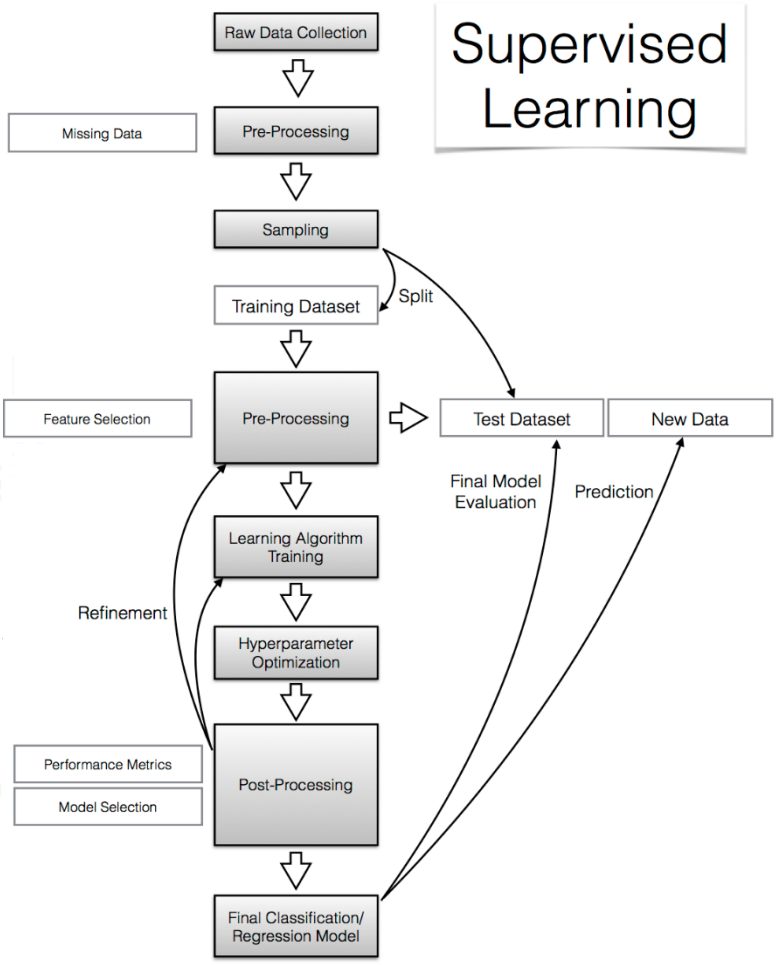


Figure 12: A Flowchart of Supervised Learning Algorithm (Raschka, 2014).

The algorithms produced throughout this study can be seen in the appendices of this study.

Chapter 5

RESULTS AND DISCUSSION

5.1 Results of Feature Selection for Classification

All parameters of vehicle accident and injured, which are the 3 main files of STATS19 database, are examined in this section.

5.1.1 Feature Selection for Casualty

In the csv file of casualty, two different independent variables, casualty severity and casualty class, were investigated by examining them separately. In addition, as a result of the analysis made for both independent variables in this study, especially when the severity of the casualties are examined, gender and age factor are important features because young women-men, old women-men, old men-women, young men-women, old-young by creating five different sub-sets consisting of drivers, the similar and different relationships between them are examined. Result tables of sub-sets are in appendix.

5.1.1.1 Feature Selection for Casualty Severity

As a result of the analysis on the severity of the casualties, the most important factor stands out as the gender of the victims. The age of the victims and the fact that the accident occurred in regions such as the city or the countryside are two other important factors in casualty severity. According to the result obtained, it is seen that it is not important the number of casualty and car passenger.

Table 4: Feature Selection of Casualty Severity

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Sex of Casualty	True	True	True	False	3
2	Vehicle Number	True	False	True	False	2
3	Casualty Home Area Type	True	False	True	False	2
4	Age Band of Casualty	False	True	False	True	2
5	Casualty Number	False	False	False	False	0
6	Car Passenger	False	False	False	False	0

5.1.1.1.1 Young Woman-Male Factors in Feature Selection of the Severity of Casualty

If the effect of less experienced girls' and boys' drivers between 16 and 20 years on this model is examined, the most important factor valid for both is seen as the number of vehicles involved. Again, for young drivers, the place where the casualties occur, and the car passenger appear as two features that affect this model.

5.1.1.1.2 Elderly Male-Female Factors in Feature Selection for Severity of Casualty

If the effects of women and men between the ages of 66 and 75 on this model are considered, the most important factor of the two is the number of vehicles. It seems that the number of elderly men who had a traffic accident is much higher than that of older women. Car passenger and casualty home area type features have almost the same importance in the elderly.

5.1.1.1.3 Elderly-Young Male and Female Driver Factors in Feature Selection for Severity of Casualty

While the common aspects of the causes of accidents between old – young male and female drivers were investigated, the vehicles involved in the accident and the place where the accident occurred were determined as the most important factors in this

model. Car travel factor is shown as important features for old-young female driver model.

5.1.1.1.4 Elderly-Young Drivers Factors in Feature Selection for Severity of Casualty

When the old and young drivers are examined, the common feature of the old and young drivers is the casualty area and the high number of accident vehicles plays an important role in this model.

5.1.1.2 Feature Selection for the Casualty Class

As a result of this study, car passenger is the most important feature in this model. The sex of the casualties and number of casualties have significance in the casualty class. In this model, young women-men, the elderly. It was examined under 5 different sub-sets for the vehicle type, including women-men, old men-women, young men-women, old-young drivers. Result tables of sub-sets are in appendix.

Table 5: Feature Selection for The Casualty Class

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	False	3
2	Sex of Casualty	False	False	True	True	2
3	Casualty Number	True	False	True	False	2
4	Age Band of Casualty	False	False	False	True	1
5	Vehicle Number	False	False	False	False	0
6	Casualty Home Area Type	False	False	False	False	0

5.1.1.2.1 Young Woman-Male Factor in Feature Selection for the Class of Casualty

Looking at the impact of girls' and boys' drivers between the ages of 16 and 20 on this model, the most important factor for young drivers is car passenger. Due to the high degree of casualty and accident number of vehicles in young male and female

drivers, young drivers are notable for the number of casualty and vehicle. Another factor is the location where the accident took place is not important for young drivers.

5.1.1.2.2 Elderly Female-Male Factor in Feature Selection for Class of Casualty

Looking at the impact of older women and men aged 66 to 75 on this model, they have almost the same structure as young men and women. As with young drivers, the most important factor in older drivers is seen as car travel. Similarly, to young drivers, elderly drivers are remarkable for the number of injured and number of vehicles, as the number of vehicles injured and injured in older male and female drivers is high. The last factor, the settlement where the accident was made, is not important for elderly drivers.

5.1.1.2.3 Elderly-Young Male and Female Driver Factors in Feature Selection for the Class of Casualty

According to the result obtained from this model, car travel seems to be the most important feature in both analyses. In addition, it is understood that the number of casualties in this model is high. At the same time, it is seen that the casualty home area type is not related to this model.

5.1.1.2.4 Elderly-Young Driver Factor in Feature Selection for Casualty Class

When the causes of accidents of older and young drivers are examined at the same time, while the car travel feature comes to the forefront as the common feature of old and young drivers the age band of casualty is not an insignificant feature.

5.1.2 Feature Selection for Vehicle Type

The vehicle type, which is the only independent variable in the vehicle's csv file, appears to be one step ahead of the other factors by the vehicle manoeuvre and the gender of the driver who uses the vehicle after the feature selection. Other equally important factors are the age of driver and the point of the first blow to the vehicle during the accident. In this model, the location of the vehicle and the skidding and overturning of the vehicle appear as irrelevant features in the classification made for the vehicle type. Again, in this study, young women - young men, old women - old men, old men - women, young men - women and old - young women - men were examined under 5 different sub-sets to understand its effect on vehicle type feature selection. Result tables of sub-sets are in appendix.

Table 6: Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Manoeuvre	True	True	True	True	4
2	Vehicle Reference	True	True	True	False	3
3	Sex of Driver	True	True	True	False	3
4	First Point of Impact	False	False	True	True	2
5	Age Band of Driver	False	True	False	True	2
6	Junction Location	True	False	False	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

5.1.2.1 Young Female-Male Factors in Feature Selection for Vehicle Type

It is stated that the importance and order of accident factors selected for young drivers as a result of the examination of this model of girls and boys between the ages of 16 and 20 is the same. The most important factor is seen as the number of vehicles and the part where the vehicle received the first blow. Vehicle manoeuvre and junction location also attract attention as other essentials.

5.1.2.2 Old Female-Male Factors in Feature Selection for Vehicle Type

Considering the impact of women and men aged 66 to 75 on this model, the most important factor for elderly drivers are the number of vehicles and first point of impact. Vehicle manoeuvre and junction location also attract attention as other fundamentals.

5.1.2.3 Elderly-Young Male and Female Driver Factor in Feature Selection for Vehicle Type

In the analysis made for this model, the reasons for the accident were found to be the same regardless of the elderly-younger female or male. As it can be understood in this model, as a result of the examination for vehicle type selection, men and women show a similar behaviour regardless of their age.

5.1.2.4 Elderly-Young Driver Factor in Feature Selection for Vehicle Type

This model gathers older and young drivers in one place and examines for vehicle type feature selection. As can be seen, it shows features of similar importance, such as results in other models.

5.1.3 Feature Selection for Accidents

Since there are two independent variables, these variables are examined separately, and their similarities and differences are specified. When we examine the factors in accidents, it is understood that the causes of accidents caused by environmental

reasons are an insignificant feature. Weather is seen as an unrelated feature at the end of the analysis for both independent variables. However, while its light condition does not affect the number of casualties in the accident, it is highly more effective in the accident severity. While the road type is insignificant in both examinations, the condition of the road's surface is only a little important for the number of casualties. As a result of both analyses, the number of vehicles and the speed limit are considered as the most important features, while the area where the accident occurred in the number of casualties is one step higher than the severity of the accident.

Table 7: Feature Selection for Accident Severity

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Speed Limit	True	True	True	True	4
2	Number of Vehicles	True	True	True	True	4
3	Urban or Rural Area	True	False	True	False	2
4	Light Conditions	False	False	True	True	2
5	Junction Detail	False	False	False	True	1
6	Weather Conditions	False	False	False	False	0
7	Road Type	False	False	False	False	0
8	Road Surface Conditions	False	False	False	False	0

Table 8: Feature Selection for Number of Casualties

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Urban or Rural Area	True	True	True	False	3
2	Speed limit	True	True	False	True	3
3	Number of Vehicles	True	True	True	False	3
4	Road Type	False	False	True	False	1
5	Road Surface Conditions	False	False	True	False	1
6	Junction Detail	False	False	False	True	1
7	Weather Conditions	False	False	False	False	0
8	Light Conditions	False	False	False	False	0

5.2 Evaluation of the Classification Methods

In this thesis, using the classification methods K-Nearest Neighbour, Support Vector Machine and Gaussian Naïve Bayes, the traffic accident data in England between 2014-2018 are processed and the results of the target variables in each main file are given in the tables below. The results between 2014-2017 are shown in tables and are included in the appendix.

5.2.1 Classification Results of the Number of Casualties in Accident

The traffic accidents data examined over the last 5 years have obtained a very high degree of accuracy using the classification methods of machine learning. Meanwhile, the classification method with the highest accuracy value was found as SVM RBF. While classifying the traffic accident data set in 2018 by these two methods, the highest result was found with an accuracy of 86.51%. However, while obtaining the highest accuracy rate with this method, the instability of the data distribution between the classes in the target variable appears. Precision, Recall, and f1-score values for one casualty were 87%, 100% and 93%, respectively, while these values were 0% for two casualties. In general, it is understood that SVM RBF classification method gives slightly better results than KNN. On the other hand, the GNB classifier has the worst accuracy value in the study in this dataset.

Table 9: Classification Results of Number of Casualties in Accident File (2018)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	87%	100%	91%	84.54%
		2	18%	5%	8%	
	SVM-RBF	1	87%	100%	93%	86.51%
		2	0%	0%	0%	
	GNB	1	89%	92%	90%	82.46%
		2	29%	21%	25%	
Training	KNN	1	87%	100%	91%	84.28%
		2	20%	5%	9%	
	SVM-RBF	1	86%	100%	93%	86.45%
		2	0%	0%	0%	
	GNB	1	88%	92%	90%	82.21%
		2	29%	20%	24%	

5.2.2 Classification Results of the Accident Severity

If we compare the classification accuracy percentages in the last 5 years, the worst accuracy rate was obtained according to 2018 data. Accuracy values for KNN, SVM-RBF and GNB in the trained data set were found as 78.98%, 77.40% and 77.41%, respectively. Considering the test data in 2014, SVM-RBF is the best performing classification algorithm with an accuracy rate of 84.88%. In the severity of accident severity, the least degree of accuracy is stated as KNN with 77.40%. According to weighted average of f1-measure in all 3 algorithms, the probability of avoiding traffic accidents in 2018 with a slight injury was found to be 78%. Although the mortality rate is 0% in 5 years, this value is expected to be slightly higher in the real world. If Recall equals zero, this means the pattern is broken and there are no positive cases in the input data. The 2018 accident severity test data set consists of 53 data for death, 919 for severe injury and 3745 for minor injury. The fact that the precision, recall, f1-score values in death are 0%, the death data which is approximately 1% of the total test data is due to the absence of a positive case

regarding death in this data set. In brief, this value was found due to the very low rate of death data in the test data set. This means that the data set is unbalanced distribution.

Table 10: Classification Results of The Accident Severity (2018)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	77.40%
		Serious	21%	3%	5%	
		Slight	79%	97%	88%	
		Weighted Average	62%	79%	70%	
	SVM-RBF	Fatal	0%	0%	0%	78.98%
		Serious	79%	100%	88%	
		Slight	62%	79%	70%	
		Weighted Average	63%	79%	70%	
	GNB	Fatal	0%	0%	0%	77.41%
		Serious	30%	6%	11%	
		Slight	80%	96%	87%	
		Weighted Average	68%	77%	71%	
Training	KNN	Fatal	0%	0%	0%	77.98%
		Serious	31%	5%	8%	
		Slight	79%	97%	88%	
		Weighted Average	69%	78%	71%	
	SVM-RBF	Fatal	0%	0%	0%	79.07%
		Serious	0%	0%	0%	
		Slight	79%	100%	88%	
		Weighted Average	63%	79%	70%	
	GNB	Fatal	0%	0%	0%	77.37%
		Serious	30%	7%	11%	
		Slight	80%	96%	87%	
		Weighted Average	69%	77%	71%	

5.2.3 Classification Results of Casualty Severity

As a result of the classification for the severity of injury, SVM-RBF and GNB often performed the best, giving the same result, at a rate of 87.80%. In general, almost 88% of accidents in the UK have recently been recovered from minor injuries. When Precision, recall and f1-measure values are examined, the probability of serious injury in traffic accident seems to be very low.

Table 11: Classification Results of The Casualty Severity (2018)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	87.49%
		Serious	14%	1%	1%	
		Slight	88%	100%	93%	
		Weighted Average	79%	87%	82%	
	SVM-RBF	Fatal	0%	0%	0%	87.80%
		Serious	0%	0%	0%	
		Slight	88%	100%	94%	
		Weighted Average	82%	91%	86%	
	GNB	Fatal	0%	0%	0%	87.80%
		Serious	0%	0%	0%	
		Slight	88%	100%	94%	
		Weighted Average	82%	91%	86%	
Training	KNN	Fatal	0%	0%	0%	87.28%
		Serious	12%	0%	1%	
		Slight	88%	100%	93%	
		Weighted Average	78%	87%	82%	
	SVM-RBF	Fatal	0%	0%	0%	87.59%
		Serious	0%	0%	0%	
		Slight	88%	100%	93%	
		Weighted Average	77%	88%	82%	
	GNB	Fatal	0%	0%	0%	87.59%
		Serious	0%	0%	0%	
		Slight	88%	100%	93%	
		Weighted Average	77%	88%	82%	

5.2.4 Classification Results of Casualty Class

In this thesis, pedestrians and passengers are included in the missing class without including pedestrians. Drivers are represented by number 1 and passengers by number 2. It is seen in the table below that almost 100% accuracy is achieved as a result of the analysis.

Table 12: Classification Results of The Casualty Class (2018)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	100%	100%	100%	99.95%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.95%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.95%
		2	100%	100%	100%	
Training	KNN	1	100%	100%	100%	99.95%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.95%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.95%
		2	100%	100%	100%	

5.2.5 Classification Results of Vehicle Type

In the classification made for the vehicle type, the best accuracy value in both test data and training data was obtained with the SVM-RBF method with 84.53% and 84.36%, respectively. While the accuracy percentage in the KNN and GNB classification methods for the test phase was found to be 82.20% to 83.33%, respectively, as a result of the analysis made for the training phase, it was calculated as 82.24% and 82.94%, respectively. The car parameter appears to be the most important factor affecting accuracy relative to the f1-score percentages.

Table 13: Classification Results of Vehicle Type (2018)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Pedal Cycle	35%	22%	22%	82.20%
		Car	86%	95%	90%	
		Van	14%	3%	6%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.53%
		Car	85%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	37%	24%	29%	83.33%
		Car	86%	97%	91%	
		Van	0%	0%	0%	
Training	KNN	Pedal Cycle	35%	21%	27%	82.24%
		Car	86%	94%	90%	
		Van	16%	4%	6%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.36%
		Car	84%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	36%	23%	28%	82.95%
		Car	86%	96%	91%	
		Van	0%	0%	0%	

Chapter 6

CONCLUSIONS AND RECOMMENDATIONS

In the light of the data shared by the world health organization about traffic accidents, more than 1 million people lose their lives in traffic every year. Traffic accidents are described as the eighth largest cause of death worldwide. It aims to cut this number in half by 2020 as a result of 10 years of work carried out by the united nations to reduce traffic losses.

In Britain, both road engineers and academics took a big step in improving road safety by publicly sharing traffic accident data at the end of each year in order to increase road safety. During this period from 2008 to 2018, there is a noticeable decrease in traffic accidents.

The main study aim of this thesis is to determine the factors causing the current traffic accidents in the UK and to increase road safety. In this context, there is a data set that is open to the public and consists of 3 main files. These are vehicle, accident and casualty. These three files are prepared to be linked to each other and it is expected that the main purpose of the accident severity will be calculated with high accuracy. For this reason, in this thesis, first, the relevant parameters for the target vector, accident severity, were obtained by using Chi-Square, Linear Support Vector Machine, Random Forest and Light Gradient Boosting Machine, which are feature selection algorithms. Then, supervised classification techniques of machine learning

were used to understand the accuracy of the relationship between these parameters. The classification techniques used are Support Vector Machine Radial Basis Function, Gaussian Naïve Bayes and k-Nearest Neighbour.

By applying these three classification methods to the processed data, the accuracy of the connection between the parameters has been determined. According to the latest available data set in 2018, traffic accidents data, classification accuracy rate of 77.40% was calculated with the K-Nearest Neighbour method, 78.98 % with SVM-RBF and 77.41% with Gaussian Naïve Bayes. As a result of the analysis made with test data for the number of casualties, and SVM appears to be the highest accuracy score with a percentage of 86.51%, while the GNB and KNN appear to be 82.46% and 84.54%. Although precision, recall and f1-score values are 0% in the second-class label of SVM, these values were found as 29%, 21% and 25%, respectively, in the GNB method. Although the accuracy value of the GNB classification method is lower than the other classification method, it is understood that this method is more ideal because the values in the second-class label are higher than both methods. As a result of analysis with test data for casualty severity, GNB and SVM appears to have the highest accuracy score with a percentage of 87.80%, while KNN appears to be 87.49%. Since precision, recall and f1-score values are not zero in the KNN second class label, it is understood that the KNN classification is better in this model, although the accuracy value is lower than the other classification method. As a result of the analysis with the test data for the vehicle, SVM appears to be the highest accuracy score with a percentage of 84.53%. Again, in this model, it is understood that KNN classification is better in this model, although the accuracy value is lower than the other classification method since the precision, recall and f1-score values are

not zero in the KNN second class label. Although there are close results with three classification methods, SVM-RBF classification shows a better performance than other classification tools. On the other hand, the KNN classification tool seems to be the most suitable model for this study, since the precision, recall and f1-score percentages are not zero in models other than accident severity.

The most important factor for accident severity was found to be the speed limit, and there were differences in the causes of accident severity according to driver's gender and age. The most important traffic accident factors for young drivers are car travel, vehicle manoeuvre and intersection location. It has been determined that the number of casualties in older male drivers is considerably higher than that of young men and women. As a result of decreased reflex in elderly drivers, vehicle manoeuvre factor stands out in accident factors.

In future studies, the reasons for the high number of fatal traffic accidents on roads in rural areas will be investigated. I would like to recommend for the authorities in charge of the accident database to reorganize accident reports to give more detailed information about the occurrence of the accidents.

REFERENCES

Abu-Zidan, F. M., & Eid, H. O. (2015). Factors affecting injury severity of vehicle occupants following road traffic collisions. *Injury-International Journal of the Care of the Injured, 46(1),* 136-141.

Acı, C., & Ozden, C. (2018). Predicting the severity of motor vehicle accident injuries in Adana-turkey using machine learning methods and detailed meteorological data. *International Journal of Intelligent Systems and Applications in Engineering, 6(1),* 72-79.

Al-Turaiki, I., Aloumi, M., Aloumi, N., Alghamdi, K., & Ieee. (2016). Modeling Traffic Accidents in Saudi Arabia using Classification Techniques. *2016 4th Saudi International Conference on Information Technology (Big Data Analysis) (Kacstit),* 15-19.

Alkheder, S., Taamneh, M., & Taamneh, S. (2017). Severity Prediction of Traffic Accident Using an Artificial Neural Network. *Journal of Forecasting, 36(1),* 100-108.

AlSagri, H., & Ykhlef, M. (2020). Quantifying Feature Importance for Detecting Depression using Random Forest. *International Journal of Advanced Computer Science and Applications, 11(5),* 628-635.

Bahiru, T. K., Singh, D. K., Tessfaw, E. A., & Ieee. (2018). Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity. *Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (Icicct)*, 1655-1660.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59.

Beshah, T., & Hill, S. (2010). Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development*,

Castro, Y., & Kim, Y. J. (2016). Data mining on road safety: factor assessment on vehicle accidents using classification models. *International Journal of Crashworthiness, 21(2)*, 104-111.

Chen, C., Zhang, G. H., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis and Prevention, 90*, 128-139.

Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica, 29(1)*.

Department of Transportation. (2019). *REPORTED ROAD CASUALTIES GREAT BRITAIN: 2018, Annual Report (1787321541)*.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46(1-3),* 389-422.

Ikram, S. T., & Cherukuri, A. K. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University-Computer and Information Sciences, 29(4),* 462-472.

Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis and Prevention, 108,* 27-36.

Jung, J., Chen, L., Sohn, G., Luo, C., & Won, J. U. (2016). Multi-Range Conditional Random Field for Classifying Railway Electrification System Objects Using Mobile Laser Scanning Data. Remote Sensing, 8(12), Article 1008.

Kumar, S., & Toshniwal, D. (2017). Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India. *European Transport Research Review, 9(2),* Article 24.

Lankarani, K. B., Heydari, S. T., Aghabeigi, M. R., Moafian, G., Hoseinzadeh, A., & Vossoughi, M. (2014). The impact of environmental factors on traffic accidents in Iran. *Journal of injury and violence research, 6(2),* 64.

Li, C., & Xu, P. (2020). Application on traffic flow prediction of machine learning in intelligent transportation. *Neural Computing & Applications.*

Liao, Y. H., & Vemuri, V. R. (2002). Use of K-Nearest Neighbor classifier for intrusion detection. *Computers & Security, 21(5),* 439-448.

Liu, P., Choo, K. K. R., Wang, L. Z., & Huang, F. (2017). SVM or deep learning? A comparative study on remote sensing image classification. *Soft Computing, 21(23),* 7053-7065.

Microsoft.(2020). LightGBM Release 2.3.2. Retrieved from https://lightgbm.readthedocs.io/_/downloads/en/latest/pdf/

Mitchell, R., Michalski, J., & Carbonell, T. (2013). *An artificial intelligence approach. Springer.*

Ospina-Mateus, H., Jimenez, L. A. Q., Lopez-Valdes, F. J., Morales-Londono, N., & Salas-Navarro, K. (2019). Using Data-Mining Techniques for the Prediction of the Severity of Road Crashes in Cartagena, Colombia. *Applied Computer Sciences in Engineering (Wea 2019), 1052,* 309-320.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). *Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12,* 2825-2830.

Raschka, S. (2014). Predictive Modeling, supervised machine learning, and pattern classification-the big picture. Sebastian Raschka.

Tyagi, V., Kalyanaraman, S., & Krishnapuram, R. (2012). Vehicular Traffic Density State Estimation Based on Cumulative Road Acoustics. *Ieee Transactions on Intelligent Transportation Systems, 13(3),* 1156-1166.

Ul Haq, E., Xu, H. R. O., Chen, X. H., Zhao, W. Q., Fan, J. P., & Abid, F. (2020). A fast hybrid computer vision technique for real-time embedded bus passenger flow calculation through camera. *Multimedia Tools and Applications, 79*(1-2), 1007-1036.

UN General Assembly. (2019). Sustainable Development Goals Road Safety All. Retrieved from https://www.unece.org/fileadmin/DAM/trans/roadsafe/publications/Road_Safety_for_All.pdf

World Health Organization. (2018). *Global status report on road safety 2018: Summary.*

Zhang, G. N., Yau, K. K. W., & Gong, X. P. (2014). Traffic violations in Guangdong Province of China: Speeding and drunk driving. *Accident Analysis and Prevention, 64,* 30-40.

Zhang, Y., Zhang, R. R., Ma, Q. F., Wang, Y. H., Wang, Q. Q., Huang, Z. H., & Huang, L. Y. (2020). A feature selection and multi-model fusion-based approach of predicting air quality. *Isa Transactions, 100,* 210-220.

APPENDICIES

Appendix A: Data Description

Accident Report

The following issues are included in the report while preparing the accident report:

- Injuries on the highway.
- Accidents related to boarding and alighting on passenger buses;
- Accidents at bus stops and intersections;
- Accidents where cyclists or riders injured themselves or a pedestrian;
- Accidents in Royal Parks (on public roads with motor vehicle access).

The following issues were not included in the report while creating the accident report.

- Accidents without personal injury;
- Accidents on private roads (excluding Royal Parks) or in cars parks;
- Accidents were reported to the police 30 or more days after the incident.

Vehicle Report

In the file to be reported for vehicles, it contains the details of the vehicle completely regardless of whether the vehicle is damaged for each vehicle involved or caused by an injury accident. The following are the topics included in this report.

- Vehicles where the driver / rider / passenger is injured;
- Vehicles damaged in an accident;
- Vehicles that damage a pedestrian (including vehicles parked inside or outside the pedestrian crossing);

- Vehicles colliding with another vehicle in the accident;
- Vehicles that do not cause damage, cause injury or cause an accident, but contribute to the accident (including parked, stationary, temporarily lifted or moving vehicles and non-towed vehicles).

Casualty Report

The following items are included in the report in the file of casualty to be reported for people who died or were injured in a traffic accident.

- Injury to vehicle passengers who suddenly manoeuvre or brake to avoid impact;
- A pedestrian who hurt herself in a parked vehicle;
- The person injured after falling from a vehicle;
- A person injured while getting on or off a bus;
- A person injured in a bus or other vehicle by a brake, sudden manoeuvre or collision;
- A person injured by the main road as a result of an accident that started on the public highway.

In STATS 20, which contains Instructions for the Completion of Traffic Accident Reports, all factors in 3 different data sets and the coding made to describe them in computer language can be accessed on the https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmentdata/file/230596/stats20-2011.pdf .

Appendix B: Factors in Feature Selection of Casualty Severity

Table B1: Young Male Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Number	True	True	True	True	4
2	Casualty Home Area Type	True	True	False	True	3
3	Car Passenger	True	True	False	True	3
4	Sex of Casualty	False	False	True	False	1
5	Age Band of Casualty	False	False	True	False	1
6	Casualty Number	False	False	False	False	0

Table B2: Young Female Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Number	True	True	True	True	4
2	Casualty Home Area Type	True	True	False	True	3
3	Car Passenger	True	True	False	True	3
4	Sex of Casualty	False	False	True	False	1
5	Age Band of Casualty	False	False	True	False	1
6	Casualty Number	False	False	False	False	0

Table B3: Old Male Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Casualty Number	True	True	True	True	4
2	Vehicle Number	True	True	False	True	3
3	Casualty Home Area Type	True	True	False	False	2
4	Car Passenger	False	True	True	False	2
5	Age Band of Casualty	False	False	True	False	1
6	Sex of Casualty	False	False	False	False	0

Table B4: Old Female Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Casualty Home Area Type	True	True	True	True	4
2	Vehicle Number	True	False	True	True	3
3	Car Passenger	True	True	False	True	3
4	Age Band of Casualty	False	False	True	False	1
5	Sex of Casualty	False	False	False	False	0
6	Casualty Number	False	False	False	False	0

Table B5: Elderly-Young Male Driver Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Casualty Home Area Type	True	True	False	True	3
2	Vehicle Number	True	False	True	False	2
3	Age Band of Casualty	True	False	False	True	2
4	Sex of Casualty	False	False	True	False	1
5	Casualty Number	False	False	True	False	1
6	Car Passenger	False	True	False	False	1

Table B6: Elderly-Young Female Driver Factors in Feature Selection of The Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Number	True	True	True	True	4
2	Casualty Home Area Type	True	True	True	True	4
3	Car Passenger	True	True	False	True	3
4	Age Band of Casualty	True	True	False	True	3
5	Sex of Casualty	False	False	True	False	1
6	Casualty Number	False	False	False	False	0

Table B7: Elderly-Young Drivers' Factors in Feature Selection for Severity of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Casualty Home Area Type	True	True	True	True	4
2	Vehicle Number	True	False	True	True	3
3	Sex of Casualty	True	False	True	True	3
4	Car Passenger	False	True	False	True	2
5	Age Band of Casualty	True	False	False	True	2
6	Casualty Number	False	False	False	False	0

Appendix C: Factors in Feature Selection of Casualty Class

Table C1: Young Male Factor in Feature Selection for The Class of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	False	True	2
3	Vehicle Number	False	False	False	True	1
4	Age Band of Casualty	False	False	True	False	1
5	Sex of Casualty	False	False	True	False	1
6	Casualty Home Area Type	False	False	False	False	0

Table C2: Young Female Factor in Feature Selection for The Class of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	False	True	2
3	Vehicle Number	False	False	False	True	1
4	Age Band of Casualty	False	False	True	False	1
5	Sex of Casualty	False	False	True	False	1
6	Casualty Home Area Type	False	False	False	False	0

Table C3: Elderly Male Factor in Feature Selection for The Class of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Sex of Casualty	False	False	True	True	2
3	Casualty Number	True	False	False	True	2
4	Age Band of Casualty	False	False	True	True	2
5	Vehicle Number	False	False	False	True	1
6	Casualty Home Area Type	False	False	False	True	1

Table C4: Elderly Female Factor in Feature Selection for The Class of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	False	True	2
3	Sex of Casualty	False	False	True	False	1
4	Age Band of Casualty	False	False	True	False	1
5	Vehicle Number	False	False	False	False	0
6	Casualty Home Area Type	False	False	False	False	0

Table C5: Elderly-Young Male Driver Factors in Feature Selection for The Class of Casualty

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	True	True	3
3	Sex of Casualty	False	False	True	False	1
4	Age Band of Casualty	False	False	False	True	1
5	Vehicle Number	False	False	False	False	0
6	Casualty Home Area Type	False	False	False	False	0

Table C6: Elderly-Young Female Driver Factors in Feature Selection for The Casualty Class

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	False	True	2
3	Vehicle Number	False	False	True	True	2
4	Sex of Casualty	False	False	True	False	1
5	Casualty Home Area Type	False	False	False	False	0
6	Age Band of Casualty	False	False	False	False	0

Table C7: Elderly-Young Driver Factors in Feature Selection for The Casualty Class

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Car Passenger	True	True	True	True	4
2	Casualty Number	True	False	True	True	3
3	Sex of Casualty	False	False	False	True	1
4	Vehicle Number	False	False	False	False	0
5	Casualty Home Area Type	False	False	False	False	0
6	Age Band of Casualty	False	False	False	False	0

Appendix D: Factors in Feature Selection of Vehicle Type

Table D1: Young Male Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	First Point of Impact	True	True	True	True	4
3	Vehicle Manoeuvre	True	True	False	True	3
4	Junction Location	True	True	False	True	3
5	Sex of Driver	False	False	True	False	1
6	Age Band of Driver	False	False	True	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Table D2: Young Female Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	First Point of Impact	True	True	True	True	4
2	Vehicle Reference	True	True	True	False	3
3	Junction Location	True	True	False	True	3
4	Vehicle Manoeuvre	True	True	False	False	2
5	Sex of Driver	False	False	True	False	1
6	Age Band of Driver	False	False	True	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Table D3: Old Male Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	Junction Location	True	True	True	True	4
3	First Point of Impact	True	True	True	True	4
4	Vehicle Manoeuvre	True	True	False	True	3
5	Age Band of Driver	False	False	True	False	1
6	Vehicle Location Restricted Lane	False	False	False	False	0
7	Skidding and Overturning	False	False	False	False	0
8	Sex of Driver	False	False	False	False	0

Table D4: Old Female Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	First Point of Impact	True	True	True	True	4
3	Vehicle Manoeuvre	True	True	False	True	3
4	Junction Location	True	True	False	True	3
5	Sex of Driver	False	False	True	False	1
6	Age Band of Driver	False	False	True	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Table D5: Old and Young Male Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	Age Band of Driver	True	True	True	True	4
3	Vehicle Manoeuvre	True	True	False	True	3
4	Junction Location	True	True	False	True	3
5	First Point of Impact	False	True	True	True	3
6	Sex of Driver	False	False	True	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Table D6: Old and Young Female Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	First Point of Impact	True	True	True	True	4
3	Age Band of Driver	True	False	True	True	3
4	Vehicle Manoeuvre	True	True	False	False	2
5	Sex of Driver	False	False	True	False	1
6	Junction Location	False	True	False	False	1
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Table D7: Elderly-Young Drivers' Factors in Feature Selection for Vehicle Type

#	Features	Chi Square	Random Forest	SVM Linear	LightGBM	Total
1	Vehicle Reference	True	True	True	True	4
2	Vehicle Manoeuvre	True	True	False	True	3
3	Sex of Driver	True	True	True	False	3
4	Junction Location	True	True	False	True	3
5	First Point of Impact	False	True	True	True	3
6	Age Band of Driver	True	False	True	True	3
7	Vehicle Location Restricted Lane	False	False	False	False	0
8	Skidding and Overturning	False	False	False	False	0

Appendix E: Classification Results

Table E1: Classification Results of Number of Casualties in Accident File (2014)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	88%	92%	90%	82.10%
		2	25%	16%	20%	
	SVM-RBF	1	87%	100%	93%	86.51%
		2	0%	0%	0%	
	GNB	1	88%	92%	90%	82.23%
		2	25%	17%	20%	
Training	KNN	1	87%	92%	90%	81.99%
		2	26%	17%	20%	
	SVM-RBF	1	86%	100%	93%	86.16%
		2	0%	0%	0%	
	GNB	1	88%	93%	90%	82.48%
		2	29%	20%	24%	

Table E2: Classification Results of Number of Casualties in Accident File (2015)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	86%	99%	90%	81.59%
		2	16%	9%	11%	
	SVM-RBF	1	87%	100%	93%	86.61%
		2	0%	0%	0%	
	GNB	1	88%	93%	90%	82.82%
		2	29%	20%	24%	
Training	KNN	1	87%	93%	90%	81.79%
		2	19%	11%	14%	
	SVM-RBF	1	86%	100%	93%	86.33%
		2	0%	0%	0%	
	GNB	1	88%	93%	90%	82.62%
		2	29%	18%	22%	

Table E3: Classification Results of Number of Casualties in Accident File (2016)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	87%	100%	92%	86.16%
		2	30%	8%	12%	
	SVM-RBF	1	86%	100%	93%	86.19%
		2	0%	0%	0%	
	GNB	1	87%	92%	90%	82.23%
		2	28%	19%	22%	
Training	KNN	1	86%	100%	92%	86.00%
		2	31%	0%	0%	
	SVM-RBF	1	86%	100%	93%	86.03%
		2	0%	0%	0%	
	GNB	1	87%	92%	90%	8.83%
		2	28%	18%	22%	

Table E4: Classification Results of Number of Casualties in Accident File (2017)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	87%	97%	93%	84.88%
		2	30%	7%	44%	
	SVM-RBF	1	87%	100%	93%	86.16%
		2	0%	0%	0%	
	GNB	1	88%	92%	90%	82.06%
		2	28%	19%	23%	
Training	KNN	1	86%	99%	92%	85.26%
		2	31%	7%	11%	
	SVM-RBF	1	86%	100%	93%	86.36%
		2	0%	0%	0%	
	GNB	1	88%	92%	90%	82.58%
		2	29%	20%	24%	

Table E5: Classification Results of The Accident Severity (2014)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	83.21%
		Serious	20%	4%	7%	
		Slight	85%	97%	91%	
		Weighted Average	75%	83%	78%	
	SVM-RBF	Fatal	0%	0%	0%	84.88%
		Serious	0%	0%	0%	
		Slight	85%	100%	92%	
		Weighted Average	72%	85%	78%	
	GNB	Fatal	0%	0%	0%	82.27%
		Serious	24%	9%	13%	
		Slight	86%	95%	90%	
		Weighted Average	76%	82%	78%	
Training	KNN	Fatal	0%	0%	0%	83.26%
		Serious	20%	4%	7%	
		Slight	85%	97%	91%	
		Weighted Average	75%	83%	78%	
	SVM-RBF	Fatal	0%	0%	0%	84.84%
		Serious	0%	0%	0%	
		Slight	85%	100%	92%	
		Weighted Average	72%	85%	78%	
	GNB	Fatal	0%	0%	0%	82.37%
		Serious	24%	9%	13%	
		Slight	85%	96%	90%	
		Weighted Average	76%	82%	78%	

Table E6: Classification Results of The Accident Severity (2015)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	84.99%
		Serious	24%	0%	10%	
		Slight	85%	100%	92%	
		Weighted Average	75%	83%	78%	
	SVM-RBF	Fatal	0%	0%	0%	85.06%
		Serious	0%	0%	0%	
		Slight	85%	100%	92%	
		Weighted Average	71%	84%	77%	
	GNB	Fatal	0%	0%	0%	82.63%
		Serious	24%	8%	12%	
		Slight	86%	96%	90%	
		Weighted Average	75%	82%	78%	
Training	KNN	Fatal	0%	0%	0%	84.61%
		Serious	26%	8%	1%	
		Slight	85%	100%	92%	
		Weighted Average	76%	83%	79%	
	SVM-RBF	Fatal	0%	0%	0%	84.72%
		Serious	0%	0%	0%	
		Slight	85%	100%	92%	
		Weighted Average	72%	85%	78%	
	GNB	Fatal	0%	0%	0%	82.10%
		Serious	24%	9%	13%	
		Slight	85%	95%	90%	
		Weighted Average	76%	82%	78%	

Table E7: Classification Results of The Accident Severity (2016)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	82.06%
		Serious	24%	3%	6%	
		Slight	83%	98%	90%	
		Weighted Average	73%	82%	76%	
	SVM-RBF	Fatal	0%	0%	0%	83.12%
		Serious	0%	0%	0%	
		Slight	83%	100%	91%	
		Weighted Average	69%	83%	75%	
	GNB	Fatal	0%	0%	0%	78.93%
		Serious	25%	15%	19%	
		Slight	84%	92%	88%	
		Weighted Average	74%	78%	76%	
Training	KNN	Fatal	0%	0%	0%	81.14%
		Serious	24%	3%	5%	
		Slight	83%	98%	90%	
		Weighted Average	72%	81%	75%	
	SVM-RBF	Fatal	0%	0%	0%	82.76%
		Serious	0%	0%	0%	
		Slight	83%	100%	91%	
		Weighted Average	67-8%	83%	75%	
	GNB	Fatal	0%	0%	0%	78.27%
		Serious	26%	15%	19%	
		Slight	84%	92%	88%	
		Weighted Average	74%	78%	76%	

Table E8: Classification Results of The Accident Severity (2017)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	78.69%
		Serious	27%	6%	10%	
		Slight	81%	96%	88%	
		Weighted Average	70%	79%	73%	
	SVM-RBF	Fatal	0%	0%	0%	80.63%
		Serious	0%	0%	0%	
		Slight	81%	100%	89%	
		Weighted Average	65%	81%	72%	
	GNB	Fatal	0%	0%	0%	79.21%
		Serious	31%	7%	11%	
		Slight	81%	97%	88%	
		Weighted Average	72%	80%	74%	
Training	KNN	Fatal	0%	0%	0%	78.67%
		Serious	27%	7%	11%	
		Slight	81%	96%	88%	
		Weighted Average	71%	79%	73%	
	SVM-RBF	Fatal	0%	0%	0%	80.52%
		Serious	0%	0%	0%	
		Slight	81%	100%	89%	
		Weighted Average	65%	81%	72%	
	GNB	Fatal	0%	0%	0%	79.32%
		Serious	33%	7%	12%	
		Slight	81%	97%	88%	
		Weighted Average	72%	79%	73%	

Table E9: Classification Results of The Casualty Severity (2014)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	88.85%
		Serious	13%	4%	6%	
		Slight	91%	97%	94%	
		Weighted Average	84%	89%	86%	
	SVM-RBF	Fatal	0%	0%	0%	90.82%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	82%	91%	86%	
	GNB	Fatal	0%	0%	0%	90.82%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	82%	91%	86%	
Training	KNN	Fatal	0%	0%	0%	89.07%
		Serious	11%	4%	5%	
		Slight	91%	97%	94%	
		Weighted Average	84%	89%	86%	
	SVM-RBF	Fatal	0%	0%	0%	91.17%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	83%	91%	87%	
	GNB	Fatal	0%	0%	0%	91.17%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	83%	91%	87%	

Table E10: Classification Results of The Casualty Severity (2015)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	90.99%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	89%	91%	87%	
	SVM-RBF	Fatal	0%	0%	0%	90.99%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	89%	91%	87%	
	GNB	Fatal	0%	0%	0%	90.99%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	89%	91%	87%	
Training	KNN	Fatal	0%	0%	0%	90.96%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	83%	91%	87%	
	SVM-RBF	Fatal	0%	0%	0%	90.96%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	83%	91%	87%	
	GNB	Fatal	0%	0%	0%	90.96%
		Serious	0%	0%	0%	
		Slight	91%	100%	95%	
		Weighted Average	83%	91%	87%	

Table E11: Classification Results of The Casualty Severity (2016)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	89.31%
		Serious	20%	3%	5%	
		Slight	90%	99%	94%	
		Weighted Average	80%	89%	84%	
	SVM-RBF	Fatal	0%	0%	0%	89.99%
		Serious	0%	0%	0%	
		Slight	90%	100%	95%	
		Weighted Average	81%	90%	85%	
	GNB	Fatal	0%	0%	0%	89.99%
		Serious	0%	0%	0%	
		Slight	90%	100%	95%	
		Weighted Average	80%	89%	84%	
Training	KNN	Fatal	0%	0%	0%	88.56%
		Serious	16%	2%	4%	
		Slight	90%	99%	94%	
		Weighted Average	82%	89%	84%	
	SVM-RBF	Fatal	0%	0%	0%	89.52%
		Serious	0%	0%	0%	
		Slight	90%	100%	94%	
		Weighted Average	81%	90%	85%	
	GNB	Fatal	0%	0%	0%	89.52%
		Serious	0%	0%	0%	
		Slight	90%	100%	94%	
		Weighted Average	80%	90%	85%	

Table E12: Classification Results of The Casualty Severity (2017)

Data	Classification Techniques	Class	Precision	Recall	f1-score	Accuracy
Testing	KNN	Fatal	0%	0%	0%	88.46%
		Serious	20%	2%	3%	
		Slight	89%	99%	94%	
		Weighted Average	81%	88%	84%	
	SVM-RBF	Fatal	0%	0%	0%	88.92%
		Serious	0%	0%	0%	
		Slight	89%	100%	94%	
		Weighted Average	79%	89%	84%	
	GNB	Fatal	0%	0%	0%	88.92%
		Serious	0%	0%	0%	
		Slight	89%	100%	94%	
		Weighted Average	79%	89%	84%	
Training	KNN	Fatal	0%	0%	0%	88.14%
		Serious	21%	2%	3%	
		Slight	89%	99%	94%	
		Weighted Average	81%	88%	83%	
	SVM-RBF	Fatal	0%	0%	0%	88.61%
		Serious	0%	0%	0%	
		Slight	89%	100%	94%	
		Weighted Average	79%	89%	84%	
	GNB	Fatal	0%	0%	0%	88.61%
		Serious	0%	0%	0%	
		Slight	89%	100%	94%	
		Weighted Average	79%	89%	84%	

Table E13: Classification Results of The Casualty Class (2014)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
Training	KNN	1	100%	100%	100%	99.98%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.98%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.98%
		2	100%	100%	100%	

Table E14: Classification Results of The Casualty Class (2015)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	100%	100%	100%	99.96%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.96%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.96%
		2	100%	100%	100%	
Training	KNN	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.97%
		2	100%	100%	100%	

Table E15: Classification Results of The Casualty Class (2016)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	100%	100%	100%	99.92%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.92%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.92%
		2	100%	100%	100%	
Training	KNN	1	100%	100%	100%	99.93%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.93%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.93%
		2	100%	100%	100%	

Table E16: Classification Results of The Casualty Class (2017)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	1	100%	100%	100%	99.99%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.99%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.99%
		2	100%	100%	100%	
Training	KNN	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	SVM-RBF	1	100%	100%	100%	99.97%
		2	100%	100%	100%	
	GNB	1	100%	100%	100%	99.97%
		2	100%	100%	100%	

Table E17: Classification Results of The Vehicle Type (2014)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Pedal Cycle	34%	21%	26%	79.18%
		Car	85%	95%	89%	
		Van	16%	4%	7%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	83.29%
		Car	83%	100%	91%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	39%	28%	33%	81.94%
		Car	86%	95%	90%	
		Van	0%	0%	0%	
Training	KNN	Pedal Cycle	31%	25%	28%	79.41%
		Car	85%	95%	88%	
		Van	14%	4%	6%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	83.17%
		Car	83%	100%	91%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	40%	30%	34%	82.10%
		Car	86%	95%	90%	
		Van	0%	0%	0%	

Table E18: Classification Results of The Vehicle Type (2015)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Pedal Cycle	33%	11%	16%	82.98%
		Car	86%	97%	91%	
		Van	10%	0%	1%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.15%
		Car	84%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	39%	28%	32%	83.04%
		Car	86%	96%	91%	
		Van	0%	0%	0%	
Training	KNN	Pedal Cycle	36%	12%	18%	82.75%
		Car	86%	92%	89%	
		Van	18%	1%	1%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	83.62%
		Car	84%	100%	91%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	39%	28%	32%	82.35%
		Car	86%	95%	90%	
		Van	0%	0%	0%	

Table E19: Classification Results of The Vehicle Type (2016)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Pedal Cycle	35%	22%	27%	82.20%
		Car	86%	94%	90%	
		Van	14%	3%	6%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.53%
		Car	85%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	37%	24%	29%	83.33%
		Car	86%	97%	91%	
		Van	0%	0%	0%	
Training	KNN	Pedal Cycle	35%	21%	27%	82.24%
		Car	86%	94%	90%	
		Van	16%	4%	6%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.36%
		Car	84%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	36%	23%	28%	82.95%
		Car	86%	96%	91%	
		Van	0%	0%	0%	

Table E20: Classification Results of The Vehicle Type (2017)

Data	Classification Techniques	Class Labels	Precision	Recall	f1-score	Accuracy
Testing	KNN	Pedal Cycle	35%	18%	24%	83.14%
		Car	86%	96%	91%	
		Van	19%	2%	3%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.62%
		Car	85%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	38%	22%	27%	83.46%
		Car	86%	96%	91%	
		Van	0%	0%	0%	
Training	KNN	Pedal Cycle	37%	16%	22%	83.50%
		Car	86%	96%	91%	
		Van	18%	1%	3%	
	SVM-RBF	Pedal Cycle	0%	0%	0%	84.78%
		Car	85%	100%	92%	
		Van	0%	0%	0%	
	GNB	Pedal Cycle	37%	23%	28%	83.64%
		Car	86%	96%	91%	
		Van	0%	0%	0%	

Appendix F: Algorithms of Feature Selection

Random Forest

```
sel= SelectFromModel (RandomForestClassifier(n_estimators =100))
sel.fit(x_train,y_train)
sel.get_support()
selected_feat = x_train.columns[(sel.get_support())]
print(selected_feat)
```

Chi-Square

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.preprocessing import MinMaxScaler
X_norm = MinMaxScaler().fit_transform(X)
chi_selector = SelectKBest(chi2, k=3)
chi_selector.fit(X_norm, y_train)
chi_support = chi_selector.get_support()
```

Linear Support Vector Machine

```
svm = LinearSVC()
rfe = RFE(svm)
rfe.fit(x_train,y_train)
rfe.support_
```

LightGBM

```
lgbc=LGBMClassifier(n_estimators=500,          learning_rate=0.05,
num_leaves=32, colsample_bytree=0.2,    reg_alpha=3, reg_lambda=1,
min_split_gain=0.01, min_child_weight=40)
embeded_lgb_selector = SelectFromModel(lgbc, max_features=8)
embeded_lgb_selector.fit(x_train, y_train)
embeded_lgb_support = embeded_lgb_selector.get_support()
embeded_lgb_feature                                                =
X.loc[:,embeded_lgb_support].columns.tolist()
print(str(len(embeded_lgb_feature)), 'selected features')
```

Appendix G: Algorithms of Classification Techniques

K-Nearest Neighbour

```
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier()
KNN.fit(x_train, y_train)
KNN_predictions = KNN.predict(x_test)
print(accuracy_score(y_test, KNN_predictions))
print(confusion_matrix(y_test, KNN_predictions))
print(classification_report(y_test, KNN_predictions))
```

Support Vector Machine with Radial Basis Function

```
SVM =SVC (kernel='rbf', random_state=0, gamma=.01, C=1)
SVM.fit(x_train,y_train)
SVM_predictions = SVM.predict(x_test)
print(accuracy_score(y_test,SVM_predictions))
print(confusion_matrix(y_test, SVM_predictions))
print(classification_report(y_test, SVM_predictions))
```

Gaussian Naïve Bayes

```
GNB = GaussianNB()
GNB.fit(x_train,y_train)
GNB_predictions = GNB.predict(x_test)
print(accuracy_score(y_test, GNB_predictions))
print(confusion_matrix(y_test, GNB_predictions))
print(classification_report(y_test, GNB_predictions))
```