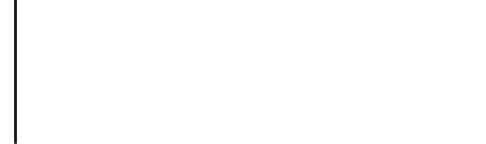
College of Business,

Technology and



Engineering

MSc Dissertation Report

**Road Traffic Accident Risk Prediction using a Classification Model**

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in [Big Data Analytics]

|  |  |
| --- | --- |
| Student Name | Rita Chinwenwa Uzoka |
| Student ID | 30034087 |
| Supervisor | Olamilekan Shobayo |
| Date of Submission | January 9th,2023 |

This dissertation does NOT contain confidential material and thus can be made available to staff and students via the library.

# Acknowledgements

To the Most High, the Most Powerful, I give thanks, who made sure that this project was finished in good health even though I felt ill during the process, my late father for raising me to be the person I am, the tutors at Sheffield Hallam University for the knowledge and skills to implement this project because the skills used in this project encompass everything I learned as a student, placement, and more, and my Supervisor for his guidance and professionalism to ensure an excellent result. Additionally, the assistance, prayers, and support of friends and family contributed to the accomplishment of this effort.

**ABSTRACT**

Due to the rapid urbanisation process and the surge in vehicle numbers, catastrophic traffic accidents have occurred, resulting in fatalities and significant financial losses. Dealing with accident-prone areas, including their classification, identification, and modification of priority, has gained popularity in recent years as a means of raising and improving the level of safety on the road network. It is effective to reduce traffic-related injuries and fatalities as well as the damage caused by accidents by proactively reducing high-risk regions. One method for analysing traffic accidents is to use a geographic information system (GIS) to look at geographical and temporal trends in places where accidents frequently occur. In this work, accidents were investigated as temporal occurrences using spatial statistical techniques based on GIS. We specifically examined the usage of localization patterns and hot spot distribution with the use of temporal information, and we constructed a highly accurate machine learning model for traffic accident risk prediction based on the spatiotemporal correlation pattern. The full city's accident data is mapped using geographic information software (Arc GIS 10.2). To help transportation authorities and policymakers, the study set out to forecast risks for RTAs. It considered classification models such as support vector machine techniques, logistic regression, decision trees, k-nearest neighbour, Rainforest, gradient boost, and neural networks. Accuracy, precision, recall, and the F1 Score were the four assessment measures that were used to assess these classification models. Hence, this is a hybrid model of geospatial clustering (Hotspot Analysis) and baseline models. Empirical findings and analysis demonstrate that, when compared to the other methods, decision tree produced the best outcomes with an F1 score of 50%.

Contents

[Acknowledgements ii](#_Toc124154418)

Abstract………………………………………………………………………………………….………iii

1.0 Introduction……………………………………………………………………………………….1

[1.1 Statement of Problem 4](#_Toc124154421)

[1.2 Aim of the Research 5](#_Toc124154422)

[1.3 Research Objective 5](#_Toc124154423)

1.4 Research Question……………………………………..………………………..…………...5

[1.5 Research Benefits 6](#_Toc124154424)

[1.6 Achieving the Research Objective 7](#_Toc124154425)

[1.7 Structure of the report 7](#_Toc124154426)

[2.0 Introduction 9](#_Toc124154427)

[2.1 Focus of Study 10](#_Toc124154428)

[2.2 Major Factors 11](#_Toc124154429)

[2.2.1 Collection of Data 11](#_Toc124154430)

[2.2.2 Major Factor Summary 13](#_Toc124154431)

[2.3 Selecting a method 13](#_Toc124154432)

[2.3.1 Comparison and Regression 13](#_Toc124154433)

2.3.2 Categorization Method……..…………………………………………………………….14

[2.3.3 Neural networks and hybrid systems 15](#_Toc124154434)

2.3.4 Clustering Method……..……………………………………………………………….…16

[2.3.5 Selecting a Method Summary 17](#_Toc124154435)

[2.4 Model Deployment 17](#_Toc124154436)

[2.4.1 Development of back-end systems 18](#_Toc124154437)

[2.4.2 Implementation Summary 19](#_Toc124154438)

[2.5 system evaluation 19](#_Toc124154439)

[2.5.1 Model Evaluation 19](#_Toc124154440)

[2.5.2 System Evaluation Summary 20](#_Toc124154441)

[2.6 Literature Review Summary 21](#_Toc124154442)

[3.0 Introduction 22](#_Toc124154445)

[3.1 Research Design 23](#_Toc124154446)

[3.1.1 Research Philosophy 23](#_Toc124154447)

[3.1.2 Justification for Research Philosophy 23](#_Toc124154448)

[3.1.3 Research Ethics 24](#_Toc124154449)

[3.1.4 Research Method (Methodological Choice) 25](#_Toc124154450)

3.1.5 Research Strategy…..…...……………………………………………………………….25

3.1.6 Research Time…………………………………………………………………………….26

[3.1.7 Techniques and Procedures 27](#_Toc124154451)

[4.0 Introduction 48](#_Toc124154457)

[4.1. Map Visualization 48](#_Toc124154458)

[4.2 Exploring Data 50](#_Toc124154459)

[4.2.1 Distribution of Traffic Accidents Spatial 50](#_Toc124154460)

[4.2.2 Pattern of Traffic Accidents Over Time 51](#_Toc124154461)

[4.2.3 Hot Spot Analysis 54](#_Toc124154462)

[4.3. Final Dataset 55](#_Toc124154463)

[4.4 Modelling Result 57](#_Toc124154464)

[4.4.1 Pre-processing of the final dataset 57](#_Toc124154465)

[4.4.2 Data cleansing and variable removal 57](#_Toc124154466)

[4.5 Testing and Training of Datasets 60](#_Toc124154467)

[4.5.1 Assembling a dataset 61](#_Toc124154468)

[4.5.2 Testing and Algorithm Computing 62](#_Toc124154469)

5.0 [Discussion of Findings 65](#_Toc124154471)

6.0 [Conclusion 67](#_Toc124154472)

[6.1 Limitations of Research 67](#_Toc124154473)

[6.3 Scope for further research 69](#_Toc124154474)

[6.4 Recommendation 70](#_Toc124154475)

7.0 References [84](#_Toc124154476)

Appendix A [85](#_Toc124154477)

Appendix B [91](#_Toc124154486)

Appendix n………………………………………………………………………………………...108

##### 

**CHAPTER ONE**

# INTRODUCTION

## Background of Study

Around the world, road traffic collisions claim the lives of about 1.3 million people annually while injuring between twenty and fifty million people. Road users who are more susceptible to accidents, such as cyclists and bikers account for more than half of all traffic-related fatalities and injuries. Due to medical costs for the injured, as well as missed wages for the dead or disabled, road traffic accidents not only cause human pain but also inflict a heavy financial burden on victims and their families. In general, traffic accidents hurt the economy, costing countries about 3% of their GDP each year (WHO, 2022).

A total of 1,752 reported road deaths were recorded in Great Britain in 2019, according to final figures on reported road casualties, which is about the same as it has been since 2012: 153,158 deaths from documented traffic incidents of all severity levels, a 5% drop from 2018 (Department for Transport, 2020). A fundamental problem for society is how to effectively address traffic accidents and, hence, improve road safety because of the fatalities and associated financial and social consequences. Road safety has been the focus of intense research and practise in the field of transportation. The UK Highways Agency is constantly paving new roads and resurfacing existing ones. The government declared in December 2014 that it would invest £15 billion in enhancing and expanding UK highways. The application of this funding entails the construction of technologically advanced smart highways for congested roads linking London, Birmingham, Manchester, and Yorkshire. (Department for Transport & Highways Agency, 2015). The UK has been branded the "whiplash capital of the world" despite having the third-safest roads in the world, with 2.9% vehicle deaths for every 100,000 people per year, only behind Sweden (2nd) but ahead of nations like Switzerland (4th) and the Netherlands (5th). According to the Association of British Insurers (ABI), more than 1,500 whiplash claims are filed in the UK every day, incurring a huge cost to the insurance sector of more than £2 billion annually (Lee, 2022).

One of the most crucial issues facing traffic safety engineers worldwide is where to apply safety protective methods so that they can have the greatest influence on traffic safety. The study of environmental (like distance from a U-turn, markets, bridges, important buildings, and residential areas), individual (like age and gender of drivers, time of day), and technological (like power, maximum speed, and size of cars) variables (Yakar, 2015) (Li, Ma, Zhu, Zeng, and Wang, 2018) helps find high-risk road areas and figure out accident patterns. The prediction of the risk of road accidents is one of the key research topics in traffic safety. The geometry of the road, the flow of traffic, the traits of the drivers, and the surroundings of the road all have a major impact on the likelihood of traffic accidents. Numerous studies have been conducted, including those on identifying dangerous locations and "hot spots" (i.e., sites on a section of road or highway that have a higher accident risk than expected at some threshold levels of significance, which are referred to as "hotspots," "black spots," or "high accident locations").Accurate hot spot detection is the most logical way to lower accident frequency (Al-Omari, Shatnawi, Khedaywi, & Miqdady, 2020). Studies have been carried out to estimate accident rates and examine the factors that contribute to them. (Wenqi, Dongyu, & Menghua, 2017). Using spatial-temporal analysis in the GIS (Geographical Information System) environment, solutions are successful in finding trends in the accident area. Accidents that are not distributed randomly throughout time or geography typically raise concerns about their origins and causes. Contrary to conventional approaches, spatial thinking aids in recognising patterns and providing explanations for their qualities. Academics have been interested in the numerical study of spatial point distributions for a long time. The well-known Moran's I statistic is one of the earliest works that was produced (Moran, 1948). The Berry and Marble edited collection is a key early work on universal geographical analysis (Berry & Marble, 1968). Road accident hotspots can now be measured using a variety of techniques other than points on a map. (Maher & Summersgill, 1996), for instance, concentrated on traffic accidents and network parts, particularly the ideal length. For the quantification of accident hotspots utilising the interpolation technique, Sabel et al. (2005) adopted a surface-based modelling approach. Later publications by Cliff and Ord (Cliff, 1973), Cliff and Ord (Cliff & Ord, 1981), Ord and Getis (Ord & Getis, 1995a), Anselin (Anselin, 1995), and Gatrell et al. (1996) provide examples of how spatial point analysis theory has evolved. A significant number of recent studies, including Cho's (Tam Cho, 2003), have used the methods that were created over the preceding 50 years because of the overview by Getis and Ord (Ord & Getis, 1995a; Tam Cho, 2003) (Gundogdu, 2010). It is necessary to analyse spatiotemporal trends in the areas of collisions in the road. This can be done by using geospatial technology (Choudhary, Ohri, & Kumar, 2015). Many transportation organisations rely on GIS because it is an effective tool for displaying accident data and identifying hot spots. As previously stated, one method of accident analysis employs GIS technology to analyse temporal patterns in accident-prone areas.

Building a reliable system for predicting road accidents is a crucial task in traffic accident prevention. If the likelihood of a traffic accident in a specific area can be anticipated, we can communicate this information to the adjacent cars to warn them or influence them to take a less risky route. However, due to the multiplicity of factors that might influence road accidents, numerous studies have been conducted in recent years to determine the variables impacting the frequency of traffic and urban accidents as well as their impact on risk modeling. For instance, Le (Le, Liu, & Lin, 2020) conducted a study to pinpoint high-risk locations and the concentration of accidents for various seasons and times. Ha and Thill (2011) studied how location and location-dependent factors affected the likelihood of accidents occurring in cities. Deublein et al. (2013) examined the estimation of the frequency of traffic accidents and the management of infrastructure to address this risk. Delmelle et al. (2012) studied the factors influencing bicycle and pedestrian accidents in various databases while considering the high effectiveness of problems with a small number of samples. Variable and fixed factor data that people obtained with the aid of mobile phones were completed by VGI data-based risk assessment, which is particularly challenging. For instance, the rate of road accidents varies greatly between different geographic areas. Additionally, increment weather like snow or fog can limit road visibility and traffic capacity, increasing the likelihood of traffic accidents. The frequency of traffic accidents fluctuates during the day, maybe reflecting the health of the drivers (Ren, Song, Wang, Hu, & Lei, 2018). Several outputs, including analysis of categorization, forecast, anomalies, and clusters can be achieved using data mining techniques, which are available in data science. Machine learning (ML) operations are data mining methods. ML is characterised as a technique that enables computers to learn on their own without the need for intricate coding and that may be used to make provisions for data analysis, decision making, and data preparation for real-life situations. As a method, ML is said to be primarily concerned with enhancing the capacity of computer systems to obtain data and utilise that data to learn for themselves. Data analysis is the first step in learning so that future societal decision-making can consider any patterns found in the information (Bokaba, Doorsamy, & Paul, 2022).

### 1.1 Statement of Problem

Accidents are one of the biggest issues with road transportation around the globe since they result in numerous injuries and monetary losses. According to World Health Organization reports, there are more than 1 million traffic-related deaths and 20 to 50 million injuries worldwide every year (WHO, 2022). Around 50 million people are injured, and 1.25 million people die because of road traffic accidents (RTAs) each year (Abdulhafedh, 2017). Worldwide, transportation authorities have been working to put mechanisms in place to reduce RTAs (road traffic accidents). However, despite the implementation of several regulations and safety measures, RTAs have not considerably diminished, making this a challenging task. The inability to accurately forecast risk-level RTAs is a contributing factor in this failure. Therefore, having a system that could predict accident risks with ease would be very helpful in lowering accident rates.

### 1.2 Aim of the Research

We seek to accurately forecast risk levels (RTAs) by examining public datasets and external data sources related to this domain.

### 1.3 Research Objective

This work has the following objectives:

1. To comprehend the current work, complete a literature review.
2. Use exploratory data analysis to pinpoint the elements that may affect machine learning.
3. Classify unsafe locations using spatial statistics on the secondary dataset.
4. Applying support vector machine algorithms, logistic regression, k-nearest neighbour, Rainforest, gradient boosting, decision trees and neural networks on a secondary dataset with RTA features. These specific ML regression models are used mostly because of their distinctive qualities and because they are well-liked in the literature.
5. Evaluate and compare how well the models work by using other well-known evaluation methods, such as accuracy, precision, recall, and the F1 score.
6. The best ML regression model is employed to forecast new regions.

**1.4 Research Questions**

How can the risk of road traffic accident be predicted?

### 1.5 Research Benefits

Both in the short and long terms, this initiative offers the following benefits:

1. The UK government and traffic authorities will be able to plan road measures to help solve the accidents in such locations with the help of identifying the places that have high and low risks.
2. promising way to predict or lower the risk of traffic accidents, which would make the world a safer place.
3. Additionally, UK insurers might use the prediction system as a benchmark for fees, lowering their financial losses. People who drive within the normal or expected limits of acceleration, speed, stopping, and turning (Noked, 2010) often pay less for their insurance.
4. increase the attention and clarity of UK traffic incidents.
5. Determine the potential for more investigation into a more complete system.

The benefits centre on supporting the UK government, UK traffic regulators and authorities, large populations living in London and its environs, and UK road accident insurers in terms of analysis and risk prediction of road traffic accidents, as well as learning about the arbitrary variables that influence accidents on the road. The process below was used to determine whether these benefits had been realised.

### 1.6 Achieving the Research Objective

The project strategy for a study (Appendix A), where appropriate ethical approval was acquired prior to the project's beginning (Appendix B), contains the project's initial proposal. The literature review on data gathering and the factors employed, various types of algorithms, system development, and finally system testing were the first steps in the project flow. Following the identification of the secondary research (a quantitative and qualitative dataset) and the use of Arc GIS Pro 10.2 to carry out a spatial analysis on the secondary research to pinpoint hot- and cold-spot locations in London, the model creation stage, when the dataset was cleaned to retain important factors before the algorithm was trained and tested using a classification method, and an analysis of the findings to identify relevant factors for the model, followed by the research methodology, we were able to conveniently predict road traffic risk areas in the UK.

### 1.7 Structure of the report

The study starts with Chapter 1, which includes the study's history, its problem description, its goals, its target and query, and its limits. Before moving on to the research methodology in chapter three, where we explain the research design, method, strategy, data collection methods, data analysis, and data modelling, the literature review will be discussed in chapter two, which includes the key variables, algorithm selection, model implementation, and system testing of the study. In chapter four, the results and discoveries are presented, along with a list of all statistical techniques used during data analysis. Discussions of the findings follow in Chapter 5, where the results are discussed. The conclusion will be presented in the sixth and last chapter of this study.

.

#### 1.8 Summary of Chapter

The study's background examines the evolution of traffic accidents over time, examines the death toll from accidents globally, examines how technology has helped reduce accidents, examines the financial toll these incidents have on individuals, families, and the nation at large, and examines the statistics of deaths from traffic accidents in the UK over time. Even though the UK is regarded as one of the safest nations, it is still regarded as one of the "whiplash" nations due to road traffic accidents. The chapter concludes with a discussion of the research limits and the overall structure of the paper. It also includes a statement of the research problem, research aims, objective, and questions the study seeks to address.

**CHAPTER TWO**

**LITERATURE REVIEW**

## 2.0 Introduction

This chapter is broken up into four main sections: key variables, algorithm selection, model implementation, and system testing. Each section is devoted to a different part of the relevant literature that needs to be analysed critically.

|  |  |
| --- | --- |
| Literature Area | Rationale for Research |
| Major Factors | Describe the techniques for gathering data, including how to get a dataset or create one from primary sources.  Describe the process used to decide which influential factors to include in the model. |
| Selecting a method | Choose the algorithm that best fits the project's goals in terms of its functionality.  Depending on the data used, each sort of algorithm has a different set of implications.  Verify a decision using other developed systems. |
| Model Deployment | Choose the most effective method for implementing the algorithm.  If an improper algorithm type is chosen, the model may be overly ineffective and have low accuracy. |
| system evaluation | Find out what needs to be done to make algorithms more accurate so that the best algorithm can be chosen. |

*Table 2.1: Justification of the Literature*

## 2.1 Focus of Study

The goal of the literature is to demonstrate the variety of motives behind a study's choice of methodology. A factor in this is the data type, which may be divided into primary and secondary datasets. To address the issue of poor accuracy caused by a lack of data, studies that created prediction models employing secondary road accident data were discussed. Other studies investigated using both main and secondary RTA datasets. The modelling of the traffic risk of secondary events, however, was where MLP excelled.

Exploratory data analysis on the dataset of traffic accidents on the roads was recommended by another study. The investigation was used by the authors to identify the factors that negatively affect traffic accidents. The characterization and modelling of traffic flow will be demonstrated in a study. Regression and clustering techniques were used in the investigation, and they produced some extremely encouraging findings. The definition of a traffic accident hotspot was recently studied using GIS (Geographical Information Systems) and spatial Clustering (Analysis) extensively. Road experts have historically located hotspots by comparing count data from various places and classifying the areas according to severity. However, the rising use of GIS has prompted researchers to adopt cutting-edge methods like ArcGIS Pro10.2 hot spot analysis (Spatial Clustering) to examine the connections that are accidentally significant. A hotspot for traffic accidents, however, has no universally accepted definition, which leaves significant space for discussion. The Hot Spot Analysis tool in ArcGIS 10.2 was employed to determine the hotspot areas in London based on the results of these investigations. With this approach, the model's accuracy may be increased while the results can be evaluated quickly and simply by taking a few important factors into account. This study was successful in determining which characteristics should be included in the framework to evaluate various detrimental effects on traffic accidents. Promising outcomes were shown by the framework. A study employing ML classifiers was presented last. According to the study, the suggested model can make accurate traffic predictions. Most machine learning (ML) classifiers are affected by the size of the dataset and their capacity to deal with overfitting issues. They are used in a variety of situations, including urban and rural areas, motorways, and roads. Promising outcomes when comparing ML Classifiers were shown by a study by (Bokaba et al., 2022). Using precision, recall, f1 score, ROC curve, true positive rate (TPR), and false positive rate, the study examined various models.

## 2.2 Major Factors

The variables must be taken into analysis before choosing an algorithm. This relates to the method used to gather the data as well as the precise purpose of the study, including the intended application of the data.

### 2.2.1 Collection of Data

(Fu & Zhou, 2011) chose an improved LM-BP network to predict RTAs. Humans, vehicles, weather, traffic path, and other factors of randomness in traffic accidents are taken as inputs. The study acknowledges that the back propagation model is deserving of being used as a prediction model, but the raw data wasn't sufficient for the study, so the result was less than ideal. Additionally, (Wenqi et al., 2017) developed a new road traffic accident prediction model (TAP-CNN) by using traffic accident influencing factors, such as traffic flow, weather, and light, to build a state matrix to describe the traffic state. TAP-CNN outperforms the traditional CNN in predicting road accidents, but the training data needs more data, and factors related to road structure and alignment can be used to achieve higher accuracy. Furthermore, Viswanath (Viswanath, Preethi, Nandini, & Bhuvaneshwari, 2021) also used data mining methods to create an accident prediction model using the Apriori algorithm and support vector machines. The dataset used was for Bangalore roads and included information on place, time, type of accidents, road condition, number of fatalities, etc. The researcher was able to predict the risk of accidents using these models, and as a result, a web application with a user interface was created to display the result to the user after they provided such inputs; however, because the dataset used here only covered a small area of land, using the same model elsewhere would not yield the same precise results.

To simplify the gathering of participation data in the research area, Kaffash Charandabi, Gholami, and Abdollahzadeh Bina (2022) created a crowdsourcing platform. The system allowed volunteers to instantly record the time and position of traffic control cameras, accidents, and road damage. A robot was created in the Telegram environment to gather voluntary data because Telegram is widely used in Iran, and some local groups of graduate and undergraduate students took part in the initiative. Following a quick introduction and dispatching the Telegram bot, the volunteers gathered the needed data. Because of the limited time available for this project, using or including crowdsourcing as a method of data collection may render the project impractical. However, secondary data already contains all the information needed to create a recommendation model that will help this case study achieve its goal.

### 2.2.2 Major Factor Summary

The shortcomings of the papers include a lack of sufficient raw data, factors, and geographic areas. All these issues were considered and addressed in my work, which led to the data, factors, and geographic locations used for my work. More information on this is covered in Chapter three.

## 2.3 Selecting a method

The kind of algorithm chosen will depend on how much computation is necessary to produce a given output. Considering this, the algorithms that were discussed were regression, classification, clustering, and neural networks, which according to the article appear to be the most popular algorithmic categories. The functionality of various single algorithm types is merged to generate hybrid or mixed algorithms, which conduct numerous computations simultaneously to serve a particular purpose.

### 2.3.1 Comparison and Regression

Regression analysis and correlation both concentrate on the relationships between various variables. The correlation coefficient is a metric for determining whether two variables are linearly related. By employing real-time traffic data instead of more conventional logistic regression techniques, this paper (Theofilatos, Yannis, Kopelias, & Papadimitriou, 2016) suggests several rare event logit models to be applied to model the occurrence of traffic accidents. The dependence of findings on stratified sampling and classification is the fundamental drawback of the rare-events logistic regression, though it wasn't possible to include all explanatory factors in the models due to serious problems with multicollinearity among traffic variables.

Maryam et al. (2016) or Guodong et al. (2018) did not find a significant link between RTAs and air pollution (as measured by NO, CO, NO2, NOx, PM10, SO2, and O3 rates) in Iran or China, respectively. According to this study, air quality may not have a significant impact on our model's ability to forecast RTAs in Iran or China, respectively. This knowledge helped us decide whether to use air quality data in our investigation. The effectiveness of Generalized Estimating Equation (GEE) and Support Vector Machine (SVM) models of real-time traffic flow (loop detector) and roadway geometry features was investigated in this research (Abdel-Aty & Abdalla, 2004). The study also noted that the model's accuracy decreased after being used on different roadways, suggesting that overfitting might have occurred. To train a more complete model, our effort intends to employ a wider variety of predictor variables (such as weather and road information). Furthermore, a binary classification problem, such as "Risk = 0 or 1," describes our RTA prediction project. Generalized Estimating Equation (GEE) is therefore not appropriate because it is better suited for linear regression. Despite being appropriate for binary classification, support-vector machines (SVM) (Sun, Sun, & Chen, 2014) require a very lengthy time to train the model when there is a large amount of data (in our example, several hundred thousand rows with more than 30 characteristics). This model was tested, but because the training time was excessively long (more than 2 hours), other models were chosen. These articles prioritized algorithm correctness over variable accuracy.

**2.3.2 Categorization Method**

(Bokaba et al., 2022) assigned labels to each class as part of the classification process divides data into several unique categories. It considered five ways to deal with missing data and classifiers, including naive Bayes, logistic regression, k-nearest neighbor, AdaBoost, support vector machine, and random forest. Five evaluation criteria, including accuracy, rootmean-square error, precision, recall, and receiver operating characteristic curves, were used to assess these classifiers. The evaluation also included dimensionality reduction techniques and parameter tweaking. The empirical findings and analysis demonstrate that, when compared to the other combinations, the RF classifier and multiple imputations via chained equations produced the best performance. K-nearest neighbours (KNN) and C-means clustering (CM) methods for real-time RTA prediction were compared by Lv et al. (2009). According to this study, KNN is a more straightforward, quick, and efficient model than CM. KNN was used in our research as one of the models for comparison. Although these papers provide helpful experimentation advice, they do not address the implications of the chosen variables that have already been found.

### 2.3.3 Neural networks and hybrid systems

The ideas behind ANN were outlined by Azad Abdulhafedh (Abdulhafedh, 2017): As a type of computational intelligence tool that can be applied to prediction and classification issues using a learning mechanism akin to the cognitive system's learning process in the human brain, ANNs can accurately simulate exceedingly complicated non-linear functions. Input layers, hidden layers, and output layers make up the network body. To evaluate the risk of traffic accidents, Charandab et al. (2022) developed a generalized regression neural network configured using a self-organizing map. 22 different predictor factors are considered in this hybrid predictive model when estimating the risk of a traffic collision (features). According to a quality evaluation of the suggested approach for various scenarios, the average accuracy of the accident risk prediction was roughly 90.74 percent. In a supervised learning process, using a learning approach that would yield the appropriate results, such as back propagation, these models can be trained to estimate any nonlinear function with the required level of accuracy. Comparing ANNs to statistical models, several advantages can be found; regression models, for instance, require a pre-established connection or functional form in addition to this. However, one of this paper's most significant weaknesses is the paucity of data. The more training data there is, the more certain it looks that this model can be applied to many scenarios. Additionally, data from numerous roads with a variety of conditions can be analysed and compared, which was not possible for this study because it was difficult to get data from different roads. One problem with using first-hand research instead of second-hand research is this.

#### 2.3.4 Clustering Method

 It is a classification of items based on how similar and dissimilar they are to one another. It is crucial since it establishes the fundamental grouping among the available, unlabeled data. There are no standards for effective clustering. Which criteria the user chooses to meet their needs depends on them (Clustering in machine learning.2022). Fancello et al. (2018) employed clustering to separate accident data into homogenous clusters, and then they applied Poisson or negative binomial ("NB") modelling to each cluster. Likewise, Park (Park, Kim, & Ha, 2016) gathered extensive information about traffic accidents on Seoul's highways and developed a workflow for making predictions based on k-means cluster analysis and logistic regression. One drawback of these research is that they did not incorporate the spatiotemporal correlation pattern of actual traffic accidents into their model which can be achieved through spatial statistics using ArcGIS Pro 10.2 (Gupta & Singh, 2014). Without these details, the model's ability to forecast outcomes may suffer.

However, there is no agreed-upon definition of a hotspot for traffic accidents, leaving room for much debate. Based on the findings of these studies, the Hot Spot Analysis tool (Spatial clustering) in ArcGIS Pro 10.2 was used to identify the hotspot areas in London. Using this method has the advantages of being quick and easy to evaluate the results by taking into consideration a few key parameters thereby increasing the accuracy of the model.

### 2.3.5 Selecting a Method Summary

Upon the analysis of the different types of classification, we can say that the selection of the right algorithm is based on three steps: 1) Feature Selection: As part of the feature engineering, we will be using the correlation algorithm because it enables us to see the degree of correlation, if any, between various variables. This is a crucial stage in the preprocessing of pipelines for machine learning. We can decrease the number of features in a dataset by using the correlation matrix to find variables that have a good degree of correlation fit for the model. The variables chosen need to be justified or discussed. 2. Spatial Clustering: The important hot spots are found using a hot spot analysis tool (Arc GIS Pro 10.2) based on spatial autocorrelation. This study demonstrates how important spatial statistics are; as a result, it is important to analyse them in depth and make improvements. The analytical procedure validates the hot spot analysis results by contrasting them with places with high accident rates, or "black spots," which demand immediate adjustment in accordance with pertinent factors like road conditions, design features, surface conditions, etc. 3) Algorithm Selection: We resolved to compare the classification algorithms of six traditional machine learning languages to determine which is the best model based on metrics. This will ensure we consider variable accuracy over algorithm accuracy.

## 2.4 Model Deployment

Once an algorithm has been chosen, a choice must be made as to how it will be implemented. This selection affects both the programming language used to code the algorithms and train the model as well as the software used to support the development environment. In terms of the algorithm computation, the user presentation, and the software environment needed to build this, this constitutes the back end of the system development.

### 2.4.1 Development of back-end systems

Renggli et al. (2021) demonstrated how various characteristics of data quality spread over different phases of machine learning development and that creating machine learning models may be thought of as following a similar procedure to that used for creating traditional software. The close correlation between a machine learning model's quality and the quality of the training or assessment data is a key distinction between the two. So, improving the dataset with methods like data cleaning, data integration, and label acquisition is often one of the best ways to make a machine learning model more accurate, fair, and resilient.

After learning the steps necessary to create an accurate machine learning model, we must decide which programming language to employ. (Ozgur, Colliau, Rogers, & Hughes, 2017) compared Python and R: The programmer reads their data into a data frame when first using R, uses a built-in model by utilising R's formula language, and can afterwards see the model summary output. R focuses on a considerably smaller subset of statistical activities and data, making it much simpler for a programmer to get started. The programmer has a lot more options when starting off with Python. These options can include deciding how they want to read their data, what kind of structure to use to keep it in, which machine learning tool to employ, and what kinds of objects the tool even permits to be entered. R, however, focuses on a considerably smaller subset of statistical activities and data, making it much simpler for a programmer to get started. Other researchers, including Yang Wenzhuo (Yang, 2020), used Matlab (a well-known computing software) to read the Excel data and execute data selection (Mahmoudabadi, 2010) for creating machine learning models. Although Matlab and R have some benefits, Python will perform better due to its capability to carry out these processes of data cleaning, exploratory data analysis, and machine learning models (Kaliraja, Chitradevi, & Rajan, 2022) using the Python Jupyter Notebook environment on Windows 10. Pandas, NumPy, Seaborn, Matplotlib pyplot, and other Python libraries are used as important libraries. Imbalanced-Learn's pipeline resampling and Scikit-Learn's Random Forests Numerous frameworks, like TensorFlow, Keras, PyTorch, and Scikit-Learn, which are frequently used for future technologies, are available. When compared to Matlab, these frameworks are simpler to use (Vezeteu, Morariu, & Năstac, 2021).

### 2.4.2 Implementation Summary

According to an analysis of the literature, basic research has an impact on the development phase, especially on preprocessing the data and producing the model in accordance with a set of requirements. Hence the decision was made to use Python because it can perform the processes of data cleaning, exploratory data analysis, integration, label acquisition, and machine learning modelling in one environment known as Jupyter. For data selection and cleansing, Python and R seem to be the most popular programming languages. Jupyter can be used to model machine learning with Python tools such as Sckit-learn. To make sure the results are relevant to the end users, the system must be tested and validated after it has been constructed.

.

## 2.5 system evaluation

The final step was system evaluation, which involved gathering measurable data on the accuracy and precision of the algorithms and then validating them with new data after the recommendation system produced the desired results.

### 2.5.1 Model Evaluation

For evaluating the prediction's precision and accuracy, several baseline models, including Lasso, Ridge, Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Multilayer Perceptron (MLP), and Autoregressive Moving Average Model (ARMA), all of which were implemented by ScikitLearn, were shown to have values for the MNSE (mean square error), RSE (root square error), and MAE (mean absolute error) (Ren, Song, Wang, Hu, & Lei, 2018). R squared, however, was not regarded in this work as a crucial criterion for evaluation, as described by Chukwutoo C. Ihueze (Ihueze & Onwurah, 2018) and Mohammad Hesam Rashidi's (Rashidi, Keshavarz, Pazari, Safahieh, & Samimi, 2022) research. The classification model's prediction confusion is displayed in a confusion matrix. The confusion matrix is used to calculate precision, recall, and finally the F1 score, which was used as another metric to evaluate the model’s prediction and accuracy (Iveta, Radovan, & Mihaljevi, 2021). When assessing the model's performance, additional metrics should be considered in addition to the accuracy metric. After gaining an understanding of the relevant metrics and values for appropriate model fit, let's examine the split ratio and cross-validation folds. Using a split ratio of 75:25, 80:20, and 90:10, some machine learning algorithms performed better at 75:25, while others performed better at 90:10. Therefore, 80:20 was viewed as the middle value between the two divides, i.e., 75:25 and 90:10, In machine learning, k = 10 is a popular cross-validation value (Sowdagur, Rozbully-Sowdagur, & Suddul, 2022).

### 2.5.2 System Evaluation Summary

Since our model aims to provide a probability of the next period's traffic accident, which varies from 0 to 1, and could be treated as a regression task, metrics like RMSE, MAPE, RSE, and R squared could be calculated directly. In recent years, researchers have not only relied on accuracy as the only metric for evaluating the model's performance but also felt it was appropriate to add other metrics as well. Additionally, if we select a threshold to assign a positive or negative category to the route with estimated danger, we may then utilise several metrics, such as the F1 score, accuracy, recall, and precision for classification jobs, to verify the model's effectiveness. To create more thorough comparisons that will be achieved via test-and-learn, we employ classification metrics, and the highest F1 score is chosen.

### 2.6 Literature Review Summary

This study project was informed by the primary learning from the literature, which included identifying the most efficient way to develop and evaluate the system's operation and determining the steps the system should take. To enable the implementation of each stage based on previously published studies that were reviewed, specialised research approaches were needed depending on the system architecture. The best technique to train the model was determined to be a dataset, which must be acquired before the core research can be conducted. The variables were then chosen in accordance with secondary research, A hybrid model of geospatial clustering (Hotspot Analysis) and baseline models of gradient boosting, neural network regression, decision trees, rainforests, support vector machines, KNN, and logistic regression were identified, and we used classification metrics (F1 score, accuracy, recall, and precision) to pick the model with the highest F1 score. For developing and producing algorithm accuracy data, the Python Jupyter Notebook was found to be the best choice. Overall, the literature research demonstrated that there isn't a perfect solution out there. As a result, when using RTAs, it is necessary to constantly combine and evaluate various strategies to identify the one that performs the best.

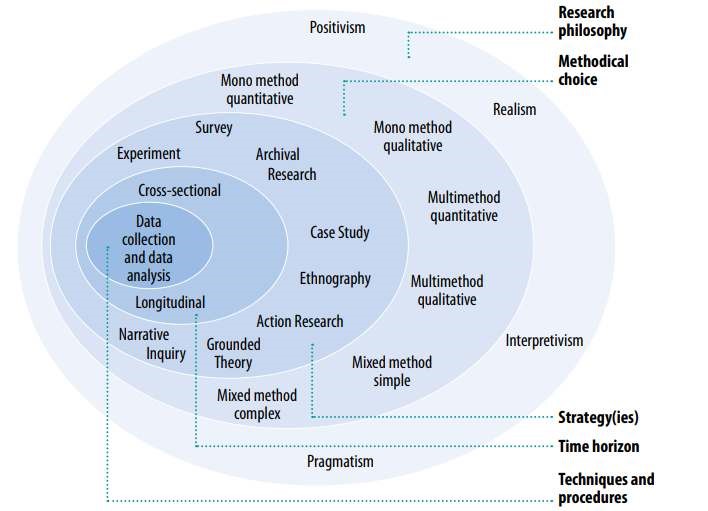
# CHAPTER THREE

# METHODOLOGY OF RESEARCH

## 3.0 Introduction

This chapter will attempt to provide a thorough explanation of the research design, including the methodological guiding principles of the study, the research setting and sample, data collection and analysis techniques, and a rationale for each of their respective adoptions.

## 3.1 Research Design



***Fig 3.1 Mark Saunders’s Research Onion (***Saunders & Tosey, )

The Saunders' onion will be used to illustrate each section of this chapter, from its outer layer to its center, under the corresponding headings in the paragraphs that follow.

### 3.1.1 Research Philosophy

A set of presumptions and views regarding the growth of knowledge are referred to as "research philosophy" (Saunders, Lewis, & Thornhill, 2007). However, this study uses a pragmatic perspective, which holds that the value of research lies in its application to real-world problems. It also recognizes that no single point of view can ever provide a complete picture and that there may be different realities (Saunders et al., 2007).

### 3.1.2 Justification for Research Philosophy

The main issue with the research is the inability to accurately predict risk level RTAs.Our intention to give practical information that is grounded in respondent experience and, therefore, of practical significance to the users heavily motivated my decision to adopt pragmatism as my overall philosophical perspective. Choosing a pragmatic strategy and setting clear research goals went hand in hand. By learning about road traffic accidents’ historical data and the research's perceived advantages early on, these goals were achieved. We were able to clarify and improve the research objectives by looking through theoretical and grey literature for knowledge gaps that would be useful to the research. The fact that different approaches or types of knowledge were used to address the main problem supports the pragmatist's contention that it is entirely feasible to work with various approaches and types of knowledge; multiple approaches are frequently feasible and may even be extremely suitable within one case study (Saunders et al., 2007). As a result, rather than researching a theory and reporting conclusions, a practical approach was used to create a workable deliverable. Instead of examining a range of possibilities, the investigation focused its findings on one possible deduction for the problem, necessitating the use of many forms of data.

### 3.1.3 Research Ethics

Ethics were upheld during data collection, processing, and presentation according to SHU norms (Sheffield Hallam University, 2017). As a result of basing every conversation on literature and primary research, the research was not slanted towards personal experiences to prevent a problem with professionalism. A consistent record of every study was kept, and according to Byne (2016), it was kept in a private folder that was only accessible by the researcher. Before beginning the research, official approval was obtained.

### 3.1.4 Research Method (Methodological Choice)

In keeping with the pragmatic and deductive nature of the research, a mixed method of qualitative and quantitative research was used for this study (Park et al., 2016). To obtain a historical dataset of accidents, causalities, LSOA, and their attributes, two different datasets that contain both quantitative and qualitative data were combined to create one dataset. Quantifiable data, which refers to the fact that the data consist of numbers that can be analysed using statistical techniques, and qualitative data, which are typically non-numerical and cannot be analysed using statistical techniques, were collected from the UK Traffic website (Chiang, Jhangiani, & Price, 2015).

By concentrating on the principle of actionable knowledge during the data examination, pragmatism assisted in the formulation of an analysis plan. This principle facilitated practical decision-making about the selection of analytical methodologies for various components of the obtained data. Consequently, rather than being constrained by pragmatism's innate emphasis on practice, we have exploited the emphasis on practise as a catalyst for novel ways of knowing and comprehending (Kelly & Cordeiro, 2020).

#### 3.1.5 Research Strategy

This project's goal was to demonstrate how a recommendation model might be used in the future, according to the system design and potential long-term uses and evaluations. It is therefore both a case study and a lesson (which explores a contemporary phenomenon inside its real-life environment, where the boundaries between phenomenon and context are not readily visible and numerous sources of information are utilized)and a demonstration of the system's functionality (Schell, 1992). Road traffic accidents in London are the subject of this case study. The primary objective of this study is to be able to identify risk level (high and low) areas in the UK in terms of road traffic accidents, which will ultimately result in a safer environment for London and the entire UK, supporting the Mayor of London's goals and objectives, which were established in 2018.

#### 3.1.6 Research Time

A mixed approach was used, including both qualitative and quantitative data from secondary research (i.e., data that had previously been gathered and examined by another party). When secondary data is required for the study, the researcher must investigate a variety of sources to find it, including published or unpublished data or records, company brochures and other books that offer pertinent information for the study, company web sites, and published or unpublished data or records. The research was done using a cross-sectional time horizon to balance out the effects of these factors and get it done faster (Sahay, 2016).

It could be predicted, as Yakar (Yakar,2015) predicted, that using different time periods as the dependent variable, the correlations will rise as the time periods become longer, reach a maximum, and then decline, possibly looking somewhat like a normal distribution. However, this should instead be done with due consideration given to what we know about how the independent variable behaves. Accident frequency is calculated over time based on the predictor's reliability and stability (given that it is a variable that changes over time, such as behaviour; childhood does not). Simply put, if the predictor is unstable over time, do not use a long period; otherwise, use as long a period as possible (af Wåhlberg, 2003). Moreso (Yakar,2015) explained that the length of the period used to identify black spots can range from 1 to 5 years, with a term of 3 years being the most common. A 5-year timeframe was chosen by the researcher to strike a balance between a lengthy term for accumulating numerous incidents and a short period to prevent the location from shifting too much, which is irrelevant to this study because there is sufficient information on accidents that were really documented to identify the black spots. Since our data predictor, which is the accident frequency, fluctuates with time, we chose the 3-year (2018–2020) period considering the predictor's stability and accuracy as suggested by Whlberg (2003).

### 3.1.7 Techniques and Procedures

Start

Collection of Data

EDA and Data

Preprocessing

Hot Spot Analysis

using Geospatial

Clustering (ArcGIS

Pro)

Final Dataset (After

merging output of

the Hotspot Analysis

with the initial

dataset)

Training the ML

Model

Select the model that

performs the best

Test the model on

new dataset or

parameters

*Fig 3.2 Flow Chart of the System Design*

#### 3.1.7.1 Sampling Method

Thus, considering the hierarchies underlying the data is also necessary for research comparison and synthesis. For studies intending to gather behavioral and attitudinal data based on "hierarchical" sampling strategies, ML analysis appears to be both straightforward and warranted (e.g., multi-stage sampling). Using a step-by-step procedure, multi-stage sampling is the process of going from a broad sample to a narrow sample (Dupont, Papadimitriou, Martensen, & Yannis, 2013). Multi-stage sampling is mostly used to choose samples that are concentrated in a small number of geographic areas, which saves time and money (Taherdoost, 2016). A multi-stage sampling method was used to narrow down the records of road traffic accident data for the whole United Kingdom to England and then to the 33 boroughs in London. This reduced costs and increased efficiency.

##### 3.1.7.2 Research Sample and Data Sources

Two different datasets were used, the first being the Road Traffic Accident Report. In the STATS19 data collection, personal injury collisions reported to the police are the primary source for statistics on road safety in Great Britain, and most statistics that are released are national statistics. Statistics on traffic accidents were evaluated by the UK Statistics Authority and verified as national statistics in July 2009, again in 2013, and again in 2019, for which a further compliance check will be done. One can download and customize data on road accidents in Great Britain that resulted in personal injuries (Stats 19). This information includes information on crashes that were reported to the police and involved at least one injury. In one accident, there may be several casualties (Department for Transport, 2022).

A cross-sectional time horizon of 3 years from 2018 to 2020 was used and narrowed the geographical area from all of Britain to London, which is in accordance with the research technique known as multi-sampling as previously described. I downloaded a snapshot of reported road collisions, by severity, numbers, and rates of reported road casualties by road user type, for Great Britain, from 1979 (Department for Transport, 2020). London was used because of the Mayor’s Office and Transport for London's dedication to enhancing road safety. The London Assembly evaluated the Transport Strategy in 2018 and the modification in 2022 following a thorough consultation by Transport for London (TfL). People's use of transportation, including the Tube, train, and bus systems, can shape London's streets, where they live, work, and spend their leisure time (Copyright, Greater London Authority, 2018).

The second dataset contains LSOAs as well as the names of London boroughs. For each Lower Super Output Area (LSOA) in Greater London, the LSOA Atlas offers an overview of demographic and associated statistics(LSOA) ,Greater London Authority (GLA), 2014 .

##### 3.1.7.3 Data Collection Methods

For the road accident dataset and the UK London datastore's LSOAs, respectively, secondary data that are already readily accessible at the UK Traffic website were used in this study. For the implementation of any machine learning method, having the right dataset is crucial. The information was derived from a UK road accident analysis dataset with 36 factors, which includes 331,370 observations of fatal and injury crash records (January 2018–December 2020). Each observation includes the following information: longitude, latitude, date, and time; LSOA of accident; number of accidents; light conditions; road conditions; pedestrians; number of casualties; route type and fatalities; cause; speed; police attendance at the scene of accident; LSOA districts; and the general weather condition at the time of the crash. The dependent variable, "Collision Severity," consists of numbers from 1 to 3 that describe how serious an accident was. (1) Collisions resulting in fatalities; (2) collisions resulting in serious injuries; and (3) minor injuries. The dataset was reduced to 7,051 observations and 38 factors following integration with the LSOA data; thus, observations with blank LSOAs were filtered away because it would be challenging to link the accidents to their respective London boroughs. The variables that make up the dataset column are specified in the dataset description (Appendix C), which may be opened, accessed, and downloaded for free from the URL within the reference. The variables were extracted and downloaded to a folder on my PC.

##### 3.1.8 Data Analysis Method

The processes of cleaning data, editing incomplete records or automatically imputed data, transforming non-normal data, and calculating aggregated scores frequently precede the examination of quantitative data. We conduct some exploratory data analysis (EDA) and pre-process this dataset because the dataset in its original form is not ready for data analysis. Because they are diverse in terms of their quantitative and qualitative makeup and range, the input data should be standardized. Quantification and a zero–one range of normalization were applied to the qualitative data.

We first explore and visualize the data from the UK Accidents Dataset using our experimental methodology to determine the number of predictor variables we can use for modelling and which predictor factors may have a greater impact on accident prediction.

#### 3.1.8.1 Distribution of Traffic Accidents Spatial

We use the 2014 London Datastore Land Mass to make the choropleth map of the number of traffic accidents in the borough of London from 2018 to 2020, which you can see in FIG. 4. 5.We did this to see if the number of accidents is related to a region's location.

##### 3.1.8.2 Pattern of Traffic Accidents Over Time

Before we looked at the temporal patterns of the number of traffic accidents (heatmap), we looked at how to figure out if the number of accidents that happen every day changed over time.

There is little doubt that the patterns of traffic accidents vary significantly during the day. More frequently than during off-peak hours, road accidents happen specifically during rush hours, as seen in Fig 4.6

##### 3.1.8.3 Spatio-Temporal Accident Hot Spot Analysis

Road location can have an impact on accident rates; some places can be classified as low- and high-risk areas. Accident trends might vary according to seasonal factors and the time of year. The last ten years have seen an increase in the investigation of their spatial features. According to Charandabi et al. (2022) and Ren et al. (2018), studying accident hot spots is an important and well-liked strategy in this regard.

In recent years, enough research has been done to come up with the right steps that could help prevent accidents and, as a result, lower the number of accidents (Butt et al., 2021).

A hot spot analysis' main objective is to gather data that will aid decision-makers in taking the proper actions to prevent and lessen traffic accidents. Accident data are typically employed as a measure of indices to calculate the occurrence of new accidents that are likely to occur. With already-available data and a GIS, it is possible to predict the temporal patterns of accidents and get accurate findings that are in line with actual circumstances. To research the connection between accident occurrence and surrounding features, data collection is both significant and required. Accidents are temporal events; hence, accident hotspots have been identified and modelled using spatial statistical techniques based on geographic information systems (GIS) (Aghajani, Dezfoulian, Arjroody, & Rezaei, 2017) (Gupta & Singh, 2014).

Moran’s method of spatial autocorrelation and the Getis-OrdGi\* statistic was applied to determine the temporal patterns and accident distribution, as well as to examine the hot spots. This tool is designed to figure out how evenly distributed or concentrated the features or spatial data are. It's important to note that by simultaneously observing the feature's location and qualities, this technique assesses the spatial feature distribution pattern. What kind of spatial feature distribution is it—random, distributed, or clustered—as determined by the analysis's findings? Using Z-Score and P-Value, this tool computes the Moran Index (or statistic) and assesses the significance of the result. We can determine whether to reject the null hypothesis based on the z-scores and p-values, which are statistical significance metrics (Aghajani et al., 2017).

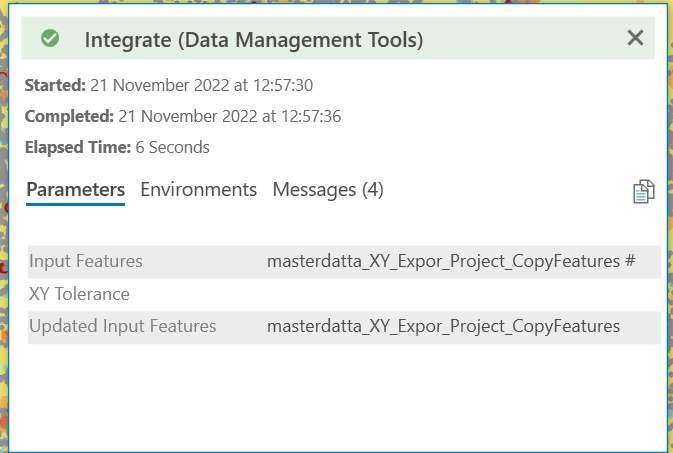
In essence, they show whether the observed geographical clustering of high or low values is stronger than what would be predicted from a random distribution of those same values. No FDR (false discovery rate) correction is reflected in the p-value or z-score fields. For a feature, a significant hot zone is indicated by a high Z score and low P value. A substantial cold region is indicated by a modest P value and a low negative Z score. The clustering becomes stronger as the Z score increases (or decreases) (Ord & Getis, 1995). The Integrate tool is used to group similar features before launching the Collect Events tool. The Collect Events tool combines coincident points and generates a new Output Feature Class that contains each distinct location found in the Input Feature Class. The total number of incidents at each distinct site is then stored in a new field called ICOUNT. This tool will only combine features if their X and Y centroid coordinates match exactly (Getis, 1996).

The Getis-OrdGi statistic (Getis-OrdGi function) is calculated for all the features in the data as part of the study of a "hot spot," which depicts accident concentration in a certain place or region. The estimated Z-score shows the locations of tiny and large clusters. Getis-OrdGi-based cluster mapping has been utilised to produce the desired hot areas for all the events (Ord & Getis, 1995).

## Process

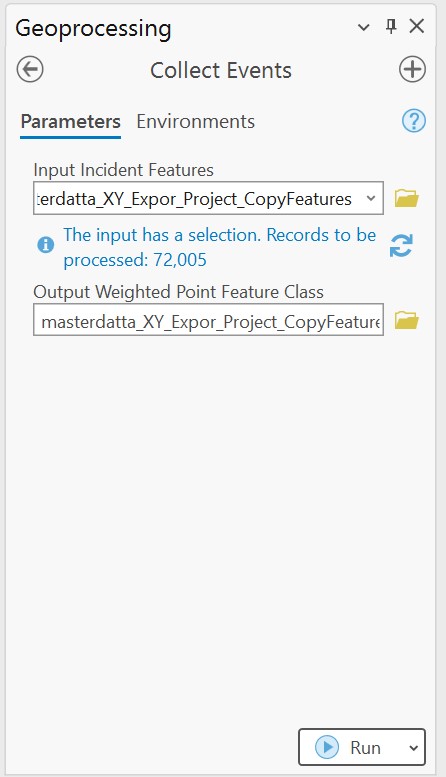
Using ArcGIS Pro 10.2, we performed a spatial analysis on 72,005 records. Performing a spatial analysis as described in the literature will improve the accuracy of the model. Individual events are not sufficient for the Hot Spot Analysis and Spatial Autocorrelation (Moran’s I) tools. When the input feature class contains coincident features, Collect Events can be used to build weighted data points.

1. Before using the collect event tool, which will only join features if their X and Y centroid coordinates are the same (Getis, 1996), the integrate tool is used to join nearby features that are close to each other.



*Fig 3.3 The integrate tool setup on ArcGIS Pro 10.2*

2 Secondary: After using the integrate tool to get snapshots of features, the output from the integrate tool will be used as the input while using the collect events tool setup, as seen in Fig. 4.9, which still has a total of 72005 records. Coinciding points are combined by Collect Events, which produces a new Output Feature Class including all the distinctive locations included in the Input Feature Class. The total number of incidents at each distinct site is then stored in a new field called ICOUNT. The ICOUNT, which is the result of the Collect Event tool setup, has a reduced record of 70,150 as the total number of weighted data points from the 72005. To carry out spatial autocorrelation and hot spot analysis, the data points need to be weighted, so out of 72005 data points, only 70150 could be weighted. This means that the 70,150 will be used for the auto-spatial correlation and the hot spot analysis; this 70,150 has the spatial weights needed for the spatial analysis (Getis, 1996).



*Fig 3.4 The Collect Event tool setup on Arc GIS Pro 10.2*

### The spatial autocorrelation technique (Moran's I approach)

Using this method, you can find out if a set of spatial features is spread out evenly or if they are all in one place.

The fact that this approach examines the spatial feature distribution pattern while concurrently observing the feature's position and properties is noteworthy. What is the state (random, dispersed, or clustered) of the spatial feature distribution, according to the analysis's findings? In fact, this tool computes the Moran index (or statistic) and assesses the significance of the result using the Z-score and P-value. We may determine whether to reject the null hypothesis based on each attribute using z-scores and p-values, which are statistical significance measures. In essence, they show whether the observed geographical clustering of high or low values is stronger than what would be predicted from a random distribution of those same values. False Discovery Rate) in the z-score and p-value fields.

A high Z score and low P value for a feature denote a significant hot area, according to the correction (discovery rate). There is a substantial cold area where the P value is low, and the negative Z score is low. The intensity of the clustering increases with a larger (or lower) Z score (Ord & Getis, 1995; Getis, 1996).

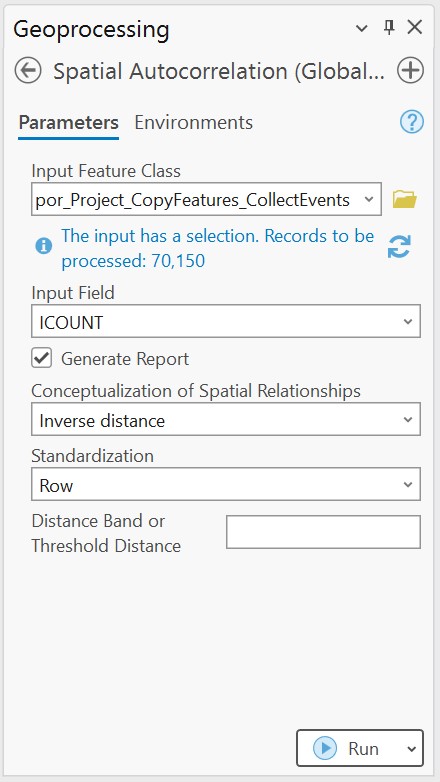


Fig 3.5 Spatial Autocorrelation Setup on ArcGIS Pro 10.2

### Hot Spot Analysis

The Getis-Ord Gi\* statistic, which is pronounced "G-i-star," is calculated for each feature in a dataset by the Hot Spot Analysis tool. The resulting p-values and z-scores indicate where spatial clusters of characteristics with high or low values are found. Each feature is examined in relation to its neighbours for this tool to function. Despite not always being a statistically significant hot spot, a feature with a high value is intriguing. For a feature to be a statistically significant hot spot, it must both have a high value and be surrounded by additional features that have a high value. The local sum for a feature and its neighbours is compared proportionally to the total number of features when the local sum differs significantly from the expected local sum. A statistically significant z-score is produced when the difference is too great to be the result of random chance. When the FDR adjustment is used, multiple testing and spatial dependency are taken into consideration when adjusting statistical significance. See Fig. 4.13 The setup feature of Hot Spot Analysis on ArcGIS Pro

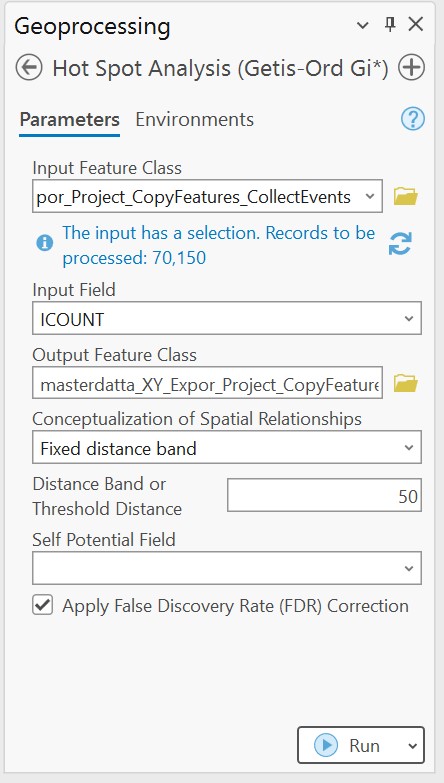


Fig. 3.6 The setup feature of Hot Spot Analysis on ArcGIS Pro

Whether or not the FDR correction is used, the Gi Bin field detects statistically significant hot and cold regions. Features in the +/-3 bins reflect statistical significance with a confidence level of 99 percent; those in the +/-2 bins reflect a confidence level of 95 percent; those in the +/-1 bins reflect a confidence level of 90 percent; and features in bin 0 do not exhibit statistically significant clustering. Statistical significance is determined using the p-value and z-score fields without FDR correction. The essential p-values determining confidence levels are decreased to account for geographical dependence and multiple testing when the optional Apply False Discovery Rate (FDR) Correction parameter is checked (Ord & Getis, 1995).

The Getis-Ord local statistic is given as:

### (1)

### 

where is the attribute value for feature , is the spatial weight between feature and *,* is equal to the total number of features and:

‾X = (2)

S = (3)

Where static is a z-score, so no further calculations are required

*Fig 3.7 The Gi\* Statistic equation* (Ord & Getis, 1995)

### Interpretation

#### The Gi\* statistic returned for each feature in the dataset is a z-score. As the z-score rises, the clustering of high values gets thicker for statistically significant positive z-scores (hot spot). For negative statistically significant z-scores, the clustering of low values becomes stronger as the z-score decreases (cold spot).

#### Use of representations in data analysis

#### When using trend analysis methods (such as Hot Spot Analysis' clustering), the null hypothesis (H0) is that either the features themselves or the values associated with those features are entirely random (CSR).

#### Another possibility (H1) The alternative hypothesis is that there is not total spatial randomization (CSR), either in the features themselves or in the values associated with those features, for the pattern analysis tools (Clustering done in Hot Spot Analysis).

#### 3.1.8.4 Modelling

To further determine the correlation between the variables and the relationship between the independent variables and the dependent variable, the output of ArcGIS' Hot Spot Analysis is used. A correlation matrix is one of the tools that was used to help with this choice. A table of correlating coefficients between different variables is called a correlation matrix. The correlation matrix can only be used for numerical datasets. The Pearson test is a method for choosing features as well. It is used with numerical attributes in datasets.

Following dataset modelling, we choose the top-performing algorithm. We started with five machine learning and deep learning algorithms for this selection. These models were chosen with an eye on their related uses. With the development of machine learning, many academics are beginning to concentrate on the prediction of traffic accidents in real-time. Lv used the k-nearest neighbour approach and feature variables based on Euclidean metrics to forecast traffic accidents (Lv, Tang, & Zhao, 2009). Park gathered extensive information on Seoul's highway traffic accidents and developed a prediction methodology based on kmeans cluster analysis and logistic regression (Park et al., 2016). To predict the risk, which is mostly used for classification and regression, many machine learning models are used, such as support vector machines, decision trees, artificial neural networks, gradient boosting, logistic regression, K-nearest neighbor, and random forests. The sample data point is accepted as input, and the algorithm performs classification and multiple regression. The risk of accidents is predicted by this module. A model with high accuracy and performance is chosen after comparison between various models, and it is then further tuned to produce an effective output.

Support Vector Machine Algorithm

Support vector machines (SVMs) are a class of supervised learning techniques used for regression, outliers’ detection, and classification. They may be used to solve both linear and non-linear binary classification problems. When the number of dimensions is greater than the number of samples, it is still useful in high-dimensional settings. It is also memory efficient because it only uses a portion of the training points (known as support vectors) in the decision function (Lessmann, Stahlbock, & Crone, 2006). There is no inherent reason why SVMs cannot be extended to highly multivariable datasets; however, doing so frequently necessitates performing a previous variable reduction step, such as PCA (Principal Component Analysis), first. Most applications of SVMs are applied to datasets with relatively small numbers of variables, like those typically obtained in analytical chemistry (Brereton & Lloyd, 2010).

Decision Trees

are a non-parametric supervised learning technique for regression and classification. By learning straightforward decision rules derived from the data attributes, the objective is to develop a model that predicts the value of a target variable. An approximate piecewise constant can be thought of as a tree. straightforward to comprehend and interpret. One can imagine trees. Minimum data preprocessing is required. Other methods frequently call for data normalization, the creation of dummy variables, and the elimination of blank values. However, keep in mind that this module does not handle missing values. The cost of employing the tree (i.e., making predictions about the data) increases logarithmically with the quantity of data points needed to train the tree, which can manage data that is both numerical and categorical. The implementation of Scikit-Learn, however, does not handle categorical variables.

artificial neural network

is a supervised learning technique that trains on a dataset to learn a function, where and are the input and output dimensions, respectively. It can learn an approximator for a non-linear function for either classification or regression, given a set of features and a target. In contrast to logistic regression, hidden layers allow for the presence of one or more non-linear layers between the input and output layers. The number of hidden neurons, layers, and iterations are a few of the hyperparameters that must be tuned in MLP. MLP is sensitive to feature scaling; the multi-layer perceptron (MLP) technique, which trains using backpropagation, is implemented by MLPClassifier. Without making any assumptions about how the event is mathematically represented, relationships between independent and dependent variables can be created. Mehta, Jain, Agarwal, and Bomnale (2012) say that ANN models are better than regression-based models because they can handle noisy data.

Gradient Boosting

For classification and regression issues, gradient boosting can be utilized. Recently, other intriguing ideas (such as XGBoost, LightGBM, and CatBoost) that place equal emphasis on speed and accuracy have been added to the family of gradient boosting algorithms (Mehta et al., 2022).

K Nearest Neighbour

KNN is a classification algorithm based on feature similarity. It examines the data, calculates their distance and similarity, and groups them according to K values. There are numerous techniques to compute distance (Labib, Rifat, Hossain, Das, & Nawrine, 2019). One of the simplest classifiers that may effectively handle classification problems is the k-NN classifier, sometimes referred to as the "lazy learner." It is a type of instance-based approach. Regression and classification are both used in the approach, which is supervised. The k indicates how many closest neighbours a model can consider. The method classifies the new data into groups that are connected to the available classes based on similarity measurements between the new data. The algorithm's simplicity in implementation is one of its merits (Bokaba et al., 2022).

RainForest

The RF model uses ensemble learning and is based on trees, which are used to build predictive models. In keeping with its name, the classifier constructs a forest from trees; more trees result in a more robust forest. The data samples are used by RF to build decision trees, which are then calculated and voted on to produce the best results. Bokaba et al. (2022) say that the algorithm is the best way to figure out how important features are in a group of datasets.

Logistic Regression

To translate the outcomes of linear functions into sigmoid functions, LR is a well-known classification approach. The method is simple to implement and can be easily extended to problems involving multiple classes. Also known as one of the basic ML techniques (Bokaba et al., 2022).

#### 3.2 Justifications for the Data Analysis Method Used

The ordinal character of the data was a crucial factor in the decision to use the data analytics strategy. To understand the locations and times of accidents, geographic statistics maps are essential. To collect such statistics, methodologies to characterise and model the spatial and temporal data are used. With the help of spatial statistics, one can understand the patterns, trends, and behaviours of geographic events, as well as determine why these phenomena occur and summarize the distribution of various phenomena. Thus, it supports our ability to decide more precisely. The studies of spatial statistics, including techniques for distribution analysis, pattern recognition, and spatial linkages, have been used in this research. Using spatial features alongside non-spatial statistical methods (neighborhood, connections, and spatial relationships) (Aghajani et al., 2017).

Additionally, all spatial processing was completed using Arc GIS 10.2 and its add-ons, and it was made on statistical metrics because most statistical analyses are made on the null hypothesis. Complete spatial randomness (CSR), which can apply to both the characteristics and the values associated with them, is the null hypothesis for the pattern analysis tools (clustering done in hot spot analysis). Look at the z-scores and p-values that the pattern analysis tools (Gupta & Singh, 2014) give you to see if you can reject the null hypothesis.

#### 3.3 Summary of Chapter

By searching through theoretical (H0, H1) and grey literature for knowledge gaps that would be helpful to the research, I was able to clarify and improve the research objectives using a pragmatic strategy. The pragmatist's claim that it is fully possible to deal with diverse types of knowledge is supported by the fact that different methods or types of knowledge were employed to address the primary problem. Considering this, a mixed method of quantitative and qualitative research was adopted to obtain numerical and categorical data, and the records of road traffic accident data for the entire United Kingdom were reduced to England and then to the 33 boroughs of London using a multi-stage sampling technique from the UK Traffic Website. Using the GIS to test the theory that the spatial feature distribution pattern is random, the respective p values and z scores as seen in FIG. 4.11 were used to reject the null hypothesis, resulting in the spatial feature distribution pattern of the weighted 70150 records being clustered and not random, and hence the 70150 weighted data points having less than 5% likelihood of noise, which makes the data very fit for modelling as this will improve the accuracy of the model. By addressing the spatiotemporal pattern with ArcGIS, we were able to improve the model's accuracy.

Our model attempts to provide a probability of future period traffic collisions ranging from 0 to 1. Because it can be viewed as a regression job, metrics like RMSE and PCC may be calculated directly. Additionally, if we select a threshold to assign the road with an estimated risk to a favorable or unfavorable category, we might use some metrics for classification tasks to verify the model's effectiveness. This study follows the spatial statistics of the theory on FIG that the larger the z score, the more significant it is. A location is classified as a "hot spot" or "high-risk area" if its z score is greater than or equal to - 0.207, otherwise as a "cold spot" or "low-risk area.” To do more thorough comparisons, we therefore employ classification metrics.

**CHAPTER FOUR**

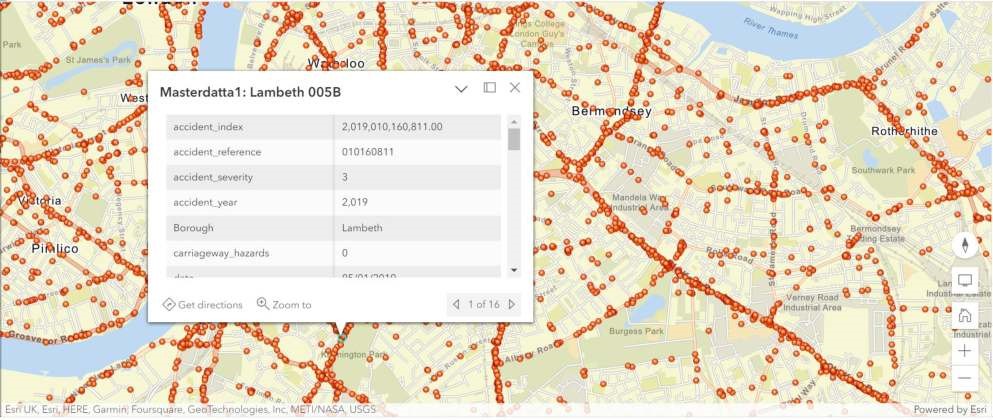
**RESULTS**

## 4.0 Introduction

The purpose of this chapter is to summarize the findings from the research as well as their methodology. The chapter will begin by outlining the outcomes of the data analysis techniques that were used, including exploratory data analysis, spatial analysis using ArcGIS 10.2 software for the hot spot analysis and statistics, correlation analysis, and feature selection prior to modeling. Finally, the outcomes of the modelling will be presented to forecast the risk levels. A summary of everything covered will be provided at the end of the chapter.

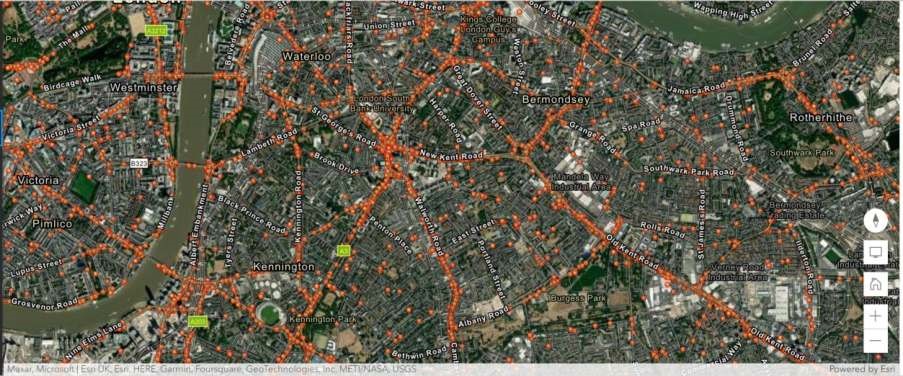
## 4.1. Map Visualization

When the accident locations are visualized on a map, it is seen that most accidents tend to cluster at road crossings.



*Fig 4.1: ArcGIS Map visualization of RTAs January 2018 – December 2020 in London.*

*Each red data point has all the features with respect to their longitudes and latitudes, hence the red data points indicate accidents that have occurred between January 2018 – December 2020 in London.*



*Fig 4.2: ArcGIS Map visualization of RTAs January 2018 – December 2020 in London. (Hybrid View)*

*A Hybrid view showing initial ArcGIS Map depicts all accident data points for the London area for the period of January 2018 to December 2020. Almost the whole road network is clogged with incidents. 38 features and 72005 observations in total are available.*

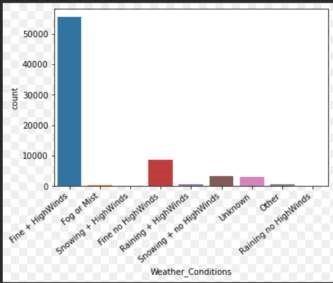
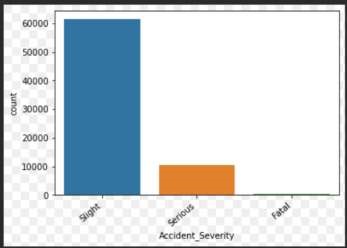


*Fig 4.3: ArcGIS Map visualization of RTAs January 2018 – December 2020 in London. (Street View)*

*A Street view of initial ArcGIS Map depiction of all accident data points for the London area for the period of January 2018 to December 2020. Almost the whole road network is clogged with incidents. 38 features and 72005 observations in total are available*.

## 4.2 Exploring Data

The datasets underwent a fundamental Exploratory Data Analysis (EDA). The datasets underwent a fundamental Exploratory Data Analysis (EDA).

*Fig 4.4: EDA of*

*key features*

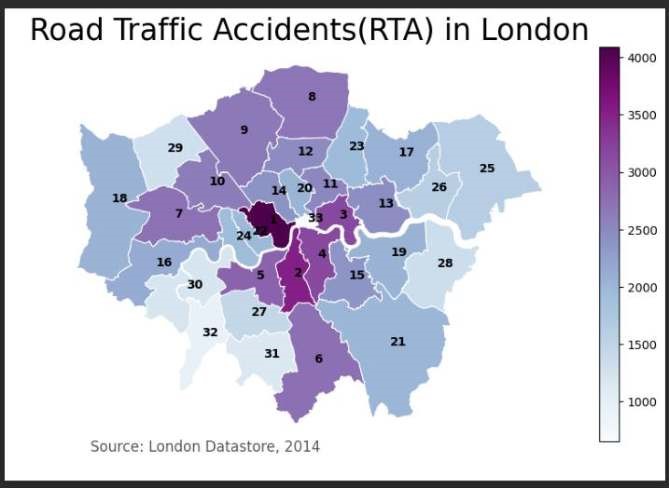
*The visualizations demonstrate that:*

Most accidents had a Severity of "3" (Slight injury). Accidents with Severity "2" (Serious) and Severity "1" are rare (Fatal).

Surprisingly, most accidents occur when the weather is good, and the roads are dry.

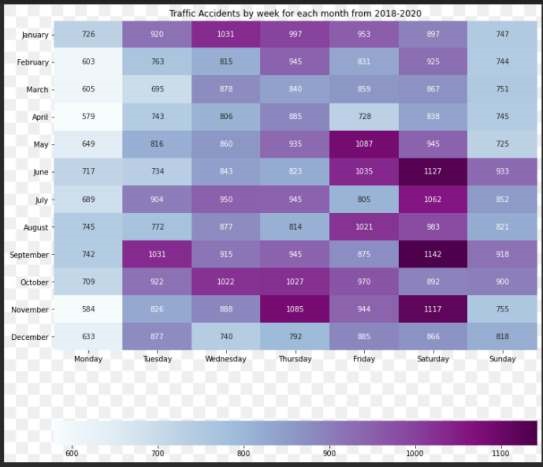
### 4.2.1 Distribution of Traffic Accidents Spatial

Figure 4.5 illustrates how the incidence of traffic accidents is not uniformly distributed and is strongly correlated with a region's geographic location. The London Borough of Croydon, the second largest of London's boroughs, offers affordable housing with good transportation links into central London and out to nearby Surrey and the South Coast. It has recorded the highest road accident rate. Typically, the highest traffic accident regions are in the major commercial and business areas.



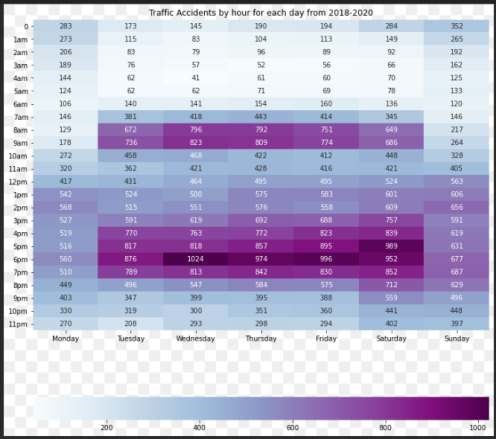
*Figure 4.5 Choropleth Map showing total RTAs in London boroughs from 2018-2020*

### 4.2.2 Pattern of Traffic Accidents Over Time



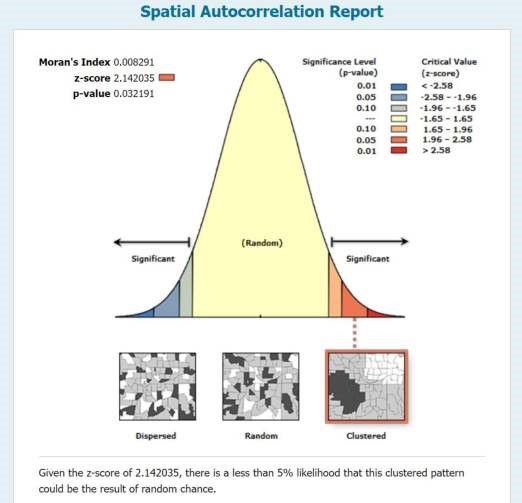
*Figure 4.6 Heatmap of RTAs January 2018 – December 2020 in London.*

*Heatmap displaying the number of RTAs by day for each month from 2018 to 2020 in all London boroughs. The image reveals that Saturdays often have the most RTAs, with greater RTA counts in the autumnal months, particularly in September.*



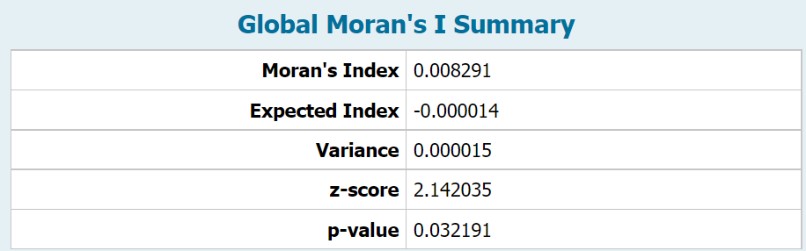
*Fig 4.7 Heatmap of RTAs January 2018 – December 2020 in London by day and hour*

*Heatmap displaying the number of RTAs in each London borough from 2018 to 2020, broken down by day and hour. The highest RTA instances occur during "rush hour" commuting times, which are 8–9 am and 4–6 pm, as expected.*



*Fig 4.8 An analysis of spatial autocorrelation*

Fig 4.8 shows that there is less than 5% likelihood that the 70150 weighted data points could be a result of random chance, random chance is noise (non-high density). Hence there is less noise in these 70150 records which will be good for modelling as it will improve the accuracy of the model.

  *Fig 4.9 Global Moran's I Synthesis*

### 4.2.3 Hot Spot Analysis

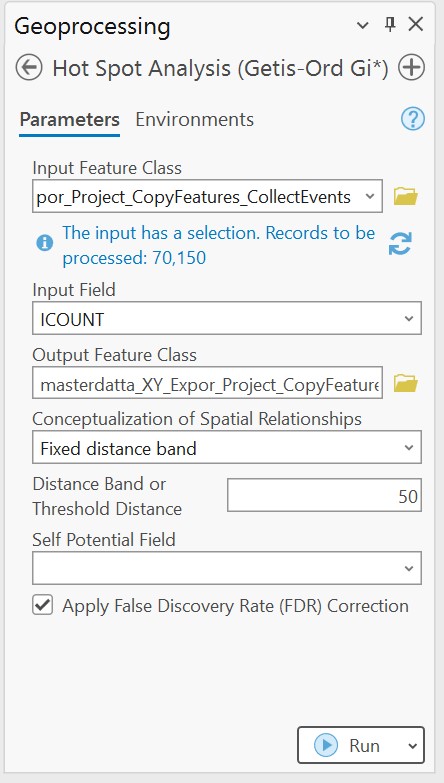
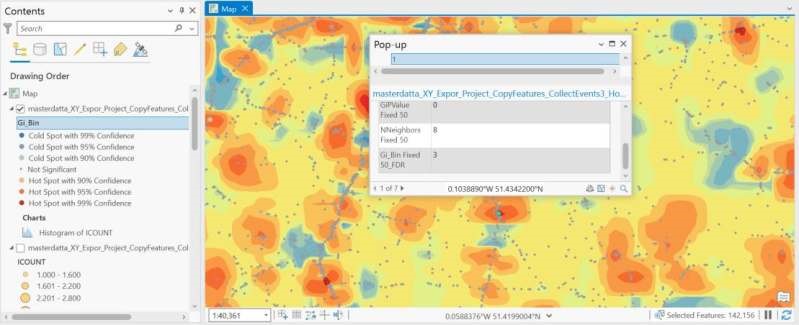
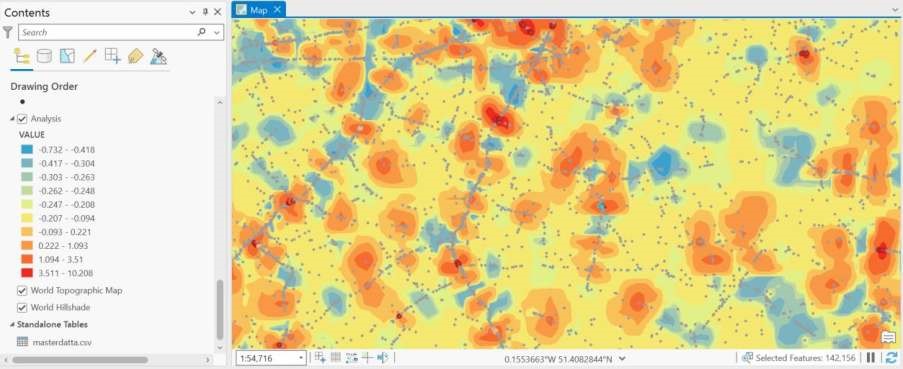


Fig 4.10 The setup feature of Hot Spot Analysis on ArcGIS Pro



*Fig 4.11: Hot Spot Analysis*

*Each red data point represents the hotspot, and each blue data point represents cold spots has all the features with respect to their longitudes and latitudes for the 70,150 records. The above pop has a Gi bin of +3 which means 99% confidence level of hot spot area.*

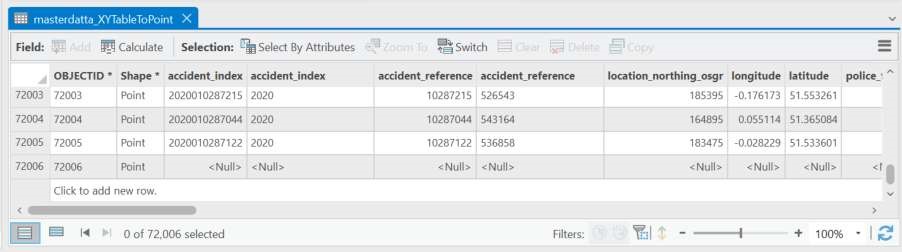


*Fig 4.12 Hot Spot Analysis and zcores*

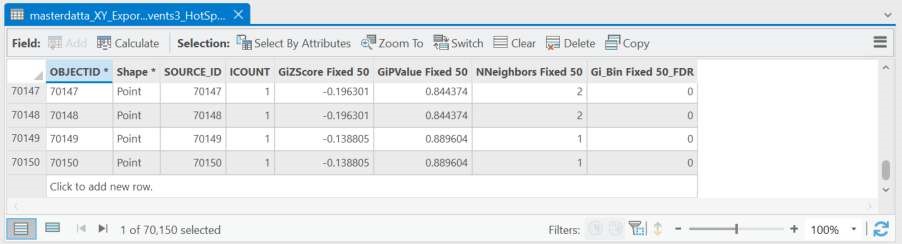
*Shows the Hot Spot and Cold Spot Areas, the legend shows that the Cold Spots areas start from the top with blue legends and low z scores and Hot spot areas are at the bottom with reddish colors legends and high z scores of 3.511 etc*.

## 4.3. Final Dataset

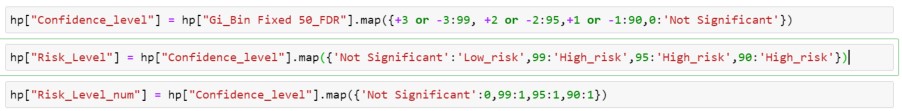
The final dataset is a merged dataset of the initial dataset of 72005 records that was uploaded to ArcGIS Pro and the hot spot analysis table of 70150 records that is the result of the analysis; the remaining 1,855 out of the 72005 that was uploaded to the software will be omitted after the merge because it was not analysed by the software because the Arc GIS Pro 10.2 software's collect events tool couldn't get their 1,855 coinciding points as previously explained.



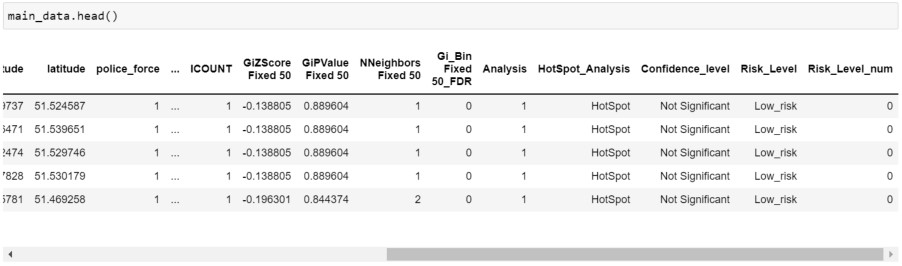
*Fig 4.13 The initial dataset used for the Hot Spot Analysis*



*Fig 4.14 Hot Spot Analysis table*



*Fig 4.15: Mapping based on [Getisetal.,1996], [Ordetal.,1995].*



*Fig 4.16 the final dataset after the merger and mapping.*

## 4.4 Modelling Result

After the models had been found, the model could be prepped and trained using the dataset.

### 4.4.1 Pre-processing of the final dataset

Before the dataset can be utilised for analysis, it must be cleaned and standardised. Based on the literature analysis, the decision was made to perform this task in Python. The Appendix contains the code for pre-processing the entire dataset. The procedure followed to prepare and alter the dataset for use is shown in Figure 4.20. The selection of the influential variables in these steps was informed by the secondary research analysis.

Final Dataset

Into Python, read the dataset (Pandas)

Eliminate unnecessary variables

Convert character class types to

integer and date types.

Export the dataset as a CSV file

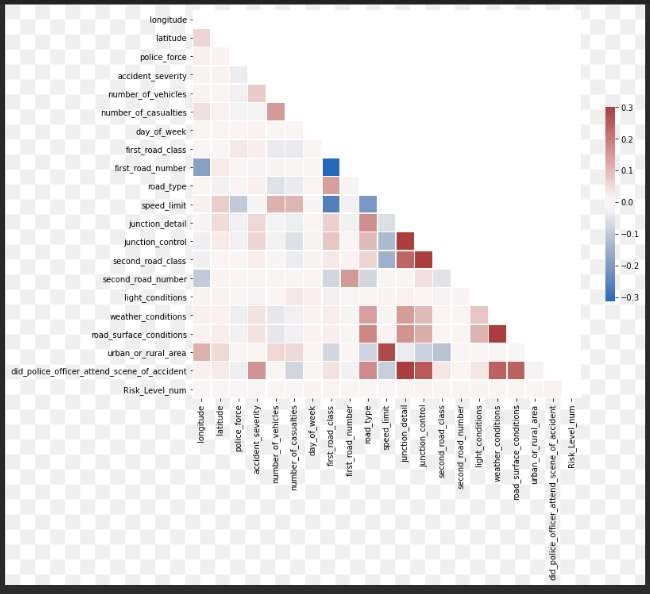
*Fig 4.17 Dataset Pre-process Outline*

### 4.4.2 Data cleansing and variable removal

To improve the effectiveness of the model's prediction, Slim et al. used data cleaning techniques in which redundant variables were first removed from the dataset. Considering this, the dataset was first read into Python before the excluded variables were eliminated, leaving the chosen influential variables (Slim, Hush, Ojah, & Babbitt, 2018). Variables offered limited room for study because the primary research respondents didn't identify them as influential, which meant they couldn't be used to make predictions. So, the target variable and 20 explanatory variables were kept in the dataset, while the other 17 were taken out.

|  |  |
| --- | --- |
| Name of the Variable | The basis for the selection |
| Risk\_Level | Target variable (for classification algorithm), enable the risk levels to be predicted using other variables |
| Risk\_Level\_num | Target variable (for regression algorithm), enable the risk levels to be predicted using other variables |
| longitude | Determine the accident's coordinates' location |
| latitude | Determine the accident's coordinates' location |
| Police\_force | Determine the police force at the scene |
| Accident\_severity | Determine how sever the accident is based on the severity levels. |
| number\_of\_vehicles | Determine the vehicles involved in the accidents |
| number\_of\_casualties | Determine the number of casualties involved in the accidents |
| date | Determine the date the accident occurred. |
| day\_of\_week | Determine the day of the week the accident occurred. |
| First\_road\_class | Determine the classification of the first road class based on the category of the road the accident occurred, roadway conditions play a major role in occurrence of accidents. |
| First\_road\_number | Determine the numbering of the classification of the first road class based on the category of the road the accident occurred, roadway conditions play a major role in occurrence of accidents. |
| Road\_type | Determine the classification of the type of road the accident occurred, roadway conditions play a major role in occurrence of accidents |
| Speed\_limit | Determine the speed limit of the road the accident occurred. |
| Junction\_detail | Determine the attributes of the junction to which the accident occurred |
| Junction\_control | Determine the attributes to which the junction is being controlled. |
| Second\_road\_class | Determine the classification of the second road class based on the category of the road the accident occurred, roadway conditions play a major role in occurrence of accidents. |
| Second\_road\_number | Determine the numbering of the classification of the second road class based on the category of the road the accident occurred. |
| Light\_conditions | Determine the light conditions of the road when the accident occurred. |
| weather\_conditions | Determine the weather conditions of the area when the accident occurred. |
| Road\_surface\_conditions | Determine the road surface conditions of the road. |
| Urban\_or\_rural\_area | Determine if the area is defined as urban or rural |
| did\_police\_officer\_attend\_scene\_of\_accident | Determine if the police attend the scene of the accident, this gives an insight if police is nearby. |

*Table 4.1 Justification for the features*



*Fig 4.18* correlation matrix

Figure 4.18 displays a correlation matrix for the dataset's numerical characteristics. To lessen the effects of multicollinearity, it is common practice to eliminate all but one characteristic from a set of strongly correlated features. Multicollinearity does not seem to have much of an impact on other models, such as Random Forest. Therefore, the model chosen will determine if it is necessary to eliminate linked features. Hence the features selected above show low multicollinearity and can thus be used for modelling.

## 4.5 Testing and Training of Datasets

The dataset was imported into a Jupiter notebook using Python, and sckit-learn was then used to train the model using the dataset.

### 4.5.1 Assembling a dataset

The preprocessed dataset needed to be imported into a Jupyter notebook and prepared using the steps indicated in the Sckit-Learn module function for the algorithms to be computed.

load the Pandas Library

Load the CSV File

Choose factors that can help you forecast target factors

Place Target Factor

Load Sckit-Learn Modules

Split the data into test and train at

90/10 ratio.

*Fig 4.19 Sckit-Learn Flow*

To prepare the dataset and run the algorithm inside the notebook, Pal's code from 2018 was used. To set the goal and explanatory variables, the Pandas Library and the pre-processed CSV dataset were imported. To divide the data into test and train datasets at 90/10 ratio for classification and regression algorithm, the Sckit-Learn Modules were then imported. Thus, 90% of the data can be used to train the model, and 10% to test it.

### 4.5.2 Testing and Algorithm Computing

Following the indicated flow, after the model had been trained in the sckit Learn environment, each individual algorithm was computed to obtain a range of scores.

Load necessary modules

Place Algorithm for Model

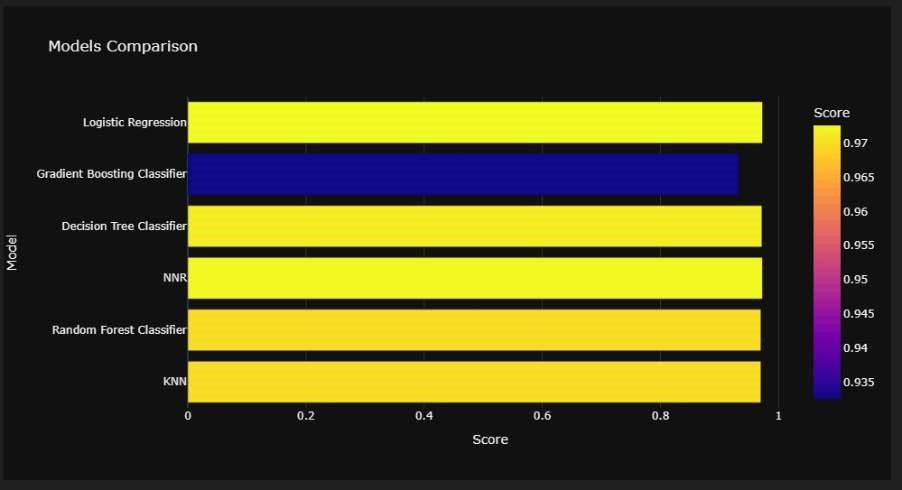
Train algorithm on train data and forecast

using the test data

show the algorithm scores.

*Fig 4.20 Algorithm Computation*

A variety of classification algorithms, including KNN, RF, NNR, Gradient Boost, and Decision Trees, were generated from the sckit learn resource (Ray 2017) under the influence of the literature. Each individual algorithm was specified as the model and then executed in the Jupyter notebook once the environment was ready.



*Fig 4.21 The result of the metrics used to evaluate the classification algorithms used for the project*.

**Metric for evaluation**.

Fig 4.25 shows that the classification model will yield a better result than the regression model, however since this model has imbalanced class, we will measure the evaluation based on the F 1 score and not the accuracy because accuracy is known not to be the best metric for imbalanced class.

Accuracy=TP+TN/TP+TN+FN+FP

Precision is defined as the ratio of accurately predicted positive observations to all observations in the expected positive class.

Precision = TP/ TP + FP

Recall is defined as the ratio of accurately anticipated positive observations to all actual positive observations, and is calculated as follows:

Recall = TP/TP + FN

Precision and Recall are combined to create the F1 Score.

Precision and Recall are typically less useful than F1 Score because it considers both false positives and false negatives. F1 Score = 2 \* TP/ 2 \* TP + FN + FP

where the letters TP, TN, FP, and FN, respectively, stand for true positive, true negative, false positive, and false negative (Yu, L., Du, Hu, Sun, Han, & Lv, 2021).

A screenshot of a computer

Description automatically generated with medium confidence

*Fig 4.22 The result of the F1 score(macro) used to evaluate the classification algorithms used for the project*

Table

Description automatically generated

*Fig 4.23 Metrics of Algorithms*

# CHAPTER FIVE

## Discussion of Findings

Delivering a road traffic accident risk prediction model that might help the traffic authority create strategies to reduce RTAs was the main goal of this project. As a result, the primary research question was answered. However, to achieve the project's goals, the initial project objective had to be met. The literature review was thorough in terms of determining what was known about the subject, namely the method for variable identification, the choice of algorithm, and the implementation of the algorithm.

The development and application of a road traffic accident risk prediction model considered a variety of potential contributing factors. The study's selection of variables is restricted to three primary categories: the state of the road, weather effects, and accident cause types. Previous literature did not consider the driver's emotional state of mind or experiential influences. Each red data point in Figures 4.1, 4.2, and 4.3 has all the characteristics in terms of longitudes and latitudes; therefore, the red data points represent accidents that took place in London between January 2018 and December 2020, with practically the whole road network being jam-packed with incidents. There are 38 features and 72005 observations available. Fig 4.4 displays the exploratory data analysis of important attributes and shows the following: The severity for most accidents was "3" (slight injury). Severity "2" (serious) and severity "1" accidents are uncommon (fatal).

Surprisingly, the best weather and dry roads are when accidents happen most. While in Fig. 4.11, each red data point denotes a hotspot, and each blue data point denotes a cold spot, both of which contain all the features for the 70,150 records with respect to their longitudes and latitudes. Gi bin +3 indicates 99% confidence in the hot spot location for the pop. The numerical features of the dataset are represented by a correlation matrix in Figure 4.18. It is usual to remove all but one feature from a collection of strongly correlated features to mitigate the impacts of multicollinearity. Other models, like Random Forest, appear to be relatively unaffected by multicollinearity. As a result, the model selected will determine whether connected features need to be removed. Because of this, the features chosen above exhibit less multicollinearity, making them suitable for modelling. All these data are part of the dataset analysed in this study.

A comparison of the criteria used to assess the classification algorithms employed for the project is shown in Figure 4.21,4.22 and 4.23. Fig 4.21 shows the a very good accuracy of all models employed, however, Logistic regression is observed to have the highest accuracy of 97.3%,according to (Iveta, Radovan, & Mihaljevi, 2021) the F1 score is a better metric used to evaluate the performance of an imbalanced dataset and an F1 score of 50% and above is termed as a good fit; all classification models used showed good fit models of F1 score of about 49% with Decision Tree having the highest F1 score of 50%,hence Decision Tree is chosen as the best model for this study.

**CHAPTER SIX**

## Conclusion

In this study, we gathered extensive data on road accidents and developed a machine learning model based on spatial clustering for estimating the likelihood of traffic accident occurrence. According to the results of the pattern analysis, the risk of traffic accidents is not evenly distributed throughout both location and time. Both regional spatial correlations and pronounced periodic temporal patterns are evident. Road traffic accident risk is difficult to forecast directly. As a result, we developed a machine learning model based on a classification model to characterise the risk of traffic accidents and capture their spatial and temporal patterns using ArcGIS 10.2 Pro. The performance comparison between the anticipated risks (Figure 4.21 and 4.22) and the F1 score (Fig 4.23) demonstrates the precision and efficacy of our approach. Following are the main conclusions of the investigation on this specific RTA dataset:

* Exploratory data analysis of important variables
* Spatial clustering was employed to define the hot spot analysis using ArcGIS 10.2 Pro, and a variety of ML techniques were used because of their popularity and unique properties. The analysis revealed that Decision Tree performed best when compared to the others, having the best F1 score of 50%.
* In addition, the evaluation performance was measured using accuracy, precision, and recall.

This technique is simple to implement in the traffic accident warning system, which will assist people in avoiding road accidents by selecting safer locations.

### Limitations of Research

1. This model forecast the risks of where accidents will occur but not when
2. Studies emphasize correlations between traits and road accidents rather than accident foretelling (Sager et al.).
3. one of many cases investigated in this study is an instrument that will communicate the likelihood of a traffic collision occurring in a region of interest.
4. Inaccessible to the public: the bulk of studies publish their findings in subscription journals in technical language. Additionally, it will assist the traffic authority in developing plans to lower RTAs.
5. Spatiotemporal dependency is a crucial aspect of traffic, and it can be used to increase the precision of traffic forecast. By using cross-correlation analysis, Yue looks at the spatiotemporal dependence of traffic flow and demonstrates its significance in assessing traffic predictability (Yue & Yeh, 2008). For large-scale traffic speed prediction, Asif presented an unsupervised learning approach Based on real-time traffic and event data (Asif et al., 2013), Pan proposed a model to forecast the spatiotemporal impact of incidents occurring on their surrounding traffic (Pan, Demiryurek, Shahabi, & Gupta, 2013). To capture the temporal dynamics and spatial dependencies of network-wide traffic, Yu created a spatiotemporal recurrent convolutional network (Yu, H., Wu, Wang, Wang, & Ma, 2017) (Yu Haiwang,2017).
6. According to Abou Ellasad (Abou Ellasad, 2020), crash-related observations typically result in unbalanced data sets since the target classes are not evenly represented (Elamrani Abou Elassad, Mousannif, & Al Moatassime, 2020). In imbalanced datasets, machine learning approaches are not very good at predicting the less represented class, especially if they are not adequately optimized.

**6.2 Delimitations of the Research**

Through the following methods, the researcher has tried to lessen the consequences of the research's limitations.

* Using a GIS (Geographical Information System) to apply spatial statistics to identify the spatial and temporal patterns of road accidents. As temporal phenomena, accidents have been studied in this study using GIS-based spatial statistical approaches to locate and model accident hot spots. Specifically, the researcher (Aghajani et al., 2017), looked at the usage of localization patterns and hot spot distribution with the aid of temporal information. Decision-makers can take appropriate action to reduce road accidents by using hot spot analysis with data creation. The use of GIS (Geographical Information Systems) like ArcGIS Software etc. for the spatial and temporal patterns analysis of hot spots of road traffic accidents has also been identified by researchers like (Gupta & Singh, 2014), (Le et al., 2020)etc.
* Data Imbalance was addressed by using a split ratio of 90:10 since we have majority of the low-risk values and minority of the high-risk values, the split ratio will include a portion of the minority class to be trained.

### 6.3 Scope for further research

The creation of a smartphone application that will assist drivers in deciding on a route for a ride is another aspect of this endeavour. It is also possible to implement a call-out to the driver using the mapping service, which would also declare the likelihood of risk along a selected route in addition to the instructions. In the future, service provider businesses like Uber, City Taxi, Bolt, and others may implement this. Additionally, this will help in improving the surveillance of regions that are prone to accidents and in providing emergency assistance in the event of one. The dangers identified by this algorithm can be used to improve the road safety signs that are posted along roadways.

### 6.4 Recommendation

* The accuracy of the model could be further improved by completely addressing the imbalance of the target variables.
* Tuning hyperparameters for modelling can be introduced.
* Test out different models, such as the artificial neural network and the XGBoost
* The most recent traffic rules, road factors, road speed changes, etc. are better reflected in the current model, which uses accident data from 2018 to 2020. More predictors, such as population density and traffic volume, could enrich the dataset.

**7.0 References**

Abdel-Aty, M., & Abdalla, M. F. (2004). Linking roadway geometrics and real-time traffic characteristics to model daytime freeway crashes: generalized estimating equations for correlated data. *Transportation Research Record, 1897*(1), 106-115.

Abdulhafedh, A. (2017). Road crash prediction models: different statistical modeling approaches. *Journal of Transportation Technologies, 7*(02), 190. af Wåhlberg, A. E. (2003). Some methodological deficiencies in studies on traffic accident predictors. *Accident Analysis & Prevention, 35*(4), 473-486.

Aghajani, M. A., Dezfoulian, R. S., Arjroody, A. R., & Rezaei, M. (2017). Applying GIS to identify the spatial and temporal patterns of road accidents using spatial statistics (case study: Ilam Province, Iran). *Transportation Research Procedia, 25*, 2126-2138.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019). Comparison of machine learning algorithms for predicting traffic accident severity. Paper presented at the *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT),* 272-276.

Al-Omari, A., Shatnawi, N., Khedaywi, T., & Miqdady, T. (2020). Prediction of traffic accidents hot spots using fuzzy logic and GIS. *Applied Geomatics, 12*(2), 149-161.

Anderson, T. (2007). Comparison of spatial methods for measuring road accident ‘hotspots’: a case study of London. *Journal of Maps, 3*(1), 55-63.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical*

*Analysis, 27*(2), 93-115.

Asif, M. T., Dauwels, J., Goh, C. Y., Oran, A., Fathi, E., Xu, M., Dhanya, M. M., Mitrovic, N., & Jaillet, P. (2013). Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE*

*Transactions on Intelligent Transportation Systems, 15*(2), 794-804.

Berry, B. J., & Marble, D. F. (1968). *Spatial analysis: a reader in statistical geography*. Prentice-Hall.

Block, R. L., & Block, C. R. (1995). Space, place and crime: Hot spot areas and hot places of liquor-related crime. *Crime and Place, 4*(2), 145-184.

Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Sciences, 12*(2), 828.

Boonserm, E., & Wiwatwattana, N. (2021). Using Machine Learning to Predict Injury Severity of Road Traffic Accidents During New Year Festivals From Thailand’s Open

Government Data. Paper presented at the *2021 9th International Electrical Engineering Congress (iEECON),* 464-467.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst, 135*(2), 230-267.

Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., Koh, T. W., & Sherazi, H. H. R. (2021). Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities. *IEEE Access, 9*, 47516-47529.

Cheng, W., & Jia, X. (2015). Exploring an alternative method of hazardous location identification: using accident count and accident reduction potential jointly. *Journal of Transportation Safety & Security, 7*(1), 40-55.

Chiang, I. A., Jhangiani, R. S., & Price, P. C. (2015). Conducting Experiments. *Research Methods in Psychology-2nd Canadian Edition,*

Choudhary, J., Ohri, A., & Kumar, B. (2015a). Identification of road accidents hot spots in varanasi using qgis. *Organized by Department of Civil Engineering, Indian Institute of*

*Technology (Banaras Hindu University), Varanasi-221005 Uttar Pradesh, India,* , 7.

Choudhary, J., Ohri, A., & Kumar, B. (2015b). Spatial and statistical analysis of road accidents hot spots using GIS. Paper presented at the *3rd Conference of Transportation*

*Research Group of India (3rd CTRG),*

Cliff, A. D. (1973). No title. *Spatial Autocorrelation,*

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: models & applications*. Taylor & Francis.

*Clustering in Machine Learning.* (2022). [https://www.geeksforgeeks.org/clustering-inmachine-learning/. https://www.geeksforgeeks.org/clustering-in-machine-learning/](https://www.geeksforgeeks.org/clustering-in-machine-learning/)

Copyright Greater London Authority, 2. (2018). *Mayor's Transport*

*Strategy.* [https://www.london.gov.uk/programmes-strategies/transport/our-visiontransport/mayors-transport-strategy-2018.](https://www.london.gov.uk/programmes-strategies/transport/our-vision-transport/mayors-transport-strategy-2018.) [https://www.london.gov.uk/programmesstrategies/transport/our-vision-transport/mayors-transport-strategy-2018](https://www.london.gov.uk/programmes-strategies/transport/our-vision-transport/mayors-transport-strategy-2018)

Dastoorpoor, M., Idani, E., Khanjani, N., Goudarzi, G., & Bahrampour, A. (2016). Relationship between air pollution, weather, traffic, and traffic-related mortality. *Trauma Monthly, 21*(4)

Delmelle, E. C., Thill, J., & Ha, H. (2012). Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians. *Transportation, 39*(2), 433-448.

Department for Transport. (2020). *Reported road casualties Great Britain, annual report: 2019.* [https://www.gov.uk/government/statistics/reported-road-casualties-great-britainannual-report-2019.](https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2019.) [https://www.gov.uk/government/statistics/reported-road-casualtiesgreat-britain-annual-report-2019](https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2019)

Department for Transport. (2022a). *Reported road collisions, vehicles and casualties tables for Great Britain.* Reported road casualties, Great Britain: annual report 2021. Reported road casualties, Great Britain: annual report 2021

Department for Transport. (2022b). *Road accidents and safety*

*statistics.* [https://www.gov.uk/government/collections/road-accidents-and-safety-](https://www.gov.uk/government/collections/road-accidents-and-safety-statistics.)

[statistics.](https://www.gov.uk/government/collections/road-accidents-and-safety-statistics.) [https://www.gov.uk/government/collections/road-accidents-and-safetystatistics](https://www.gov.uk/government/collections/road-accidents-and-safety-statistics)

Department for Transport, & Highways Agency. (2015). *Road investment strategy: 2015 to*

*2020.* [https://www.gov.uk/government/collections/road-investmentstrategy.](https://www.gov.uk/government/collections/road-investment-strategy.)<https://www.gov.uk/government/collections/road-investment-strategy>

Deublein, M., Schubert, M., Adey, B. T., Köhler, J., & Faber, M. H. (2013). Prediction of road accidents: A Bayesian hierarchical approach. *Accident Analysis & Prevention, 51*, 274-

291. 10.1016/j.aap.2012.11.019

Dupont, E., Papadimitriou, E., Martensen, H., & Yannis, G. (2013). Multilevel analysis in road safety research. *Accident Analysis & Prevention, 60*, 402-411.

Elamrani Abou Elassad, Z., Mousannif, H., & Al Moatassime, H. (2020). Class-imbalanced crash prediction based on real-time traffic and weather data: a driving simulator study. *Traffic Injury Prevention, 21*(3), 201-208.

Fu, H., & Zhou, Y. (2011). The traffic accident prediction based on neural network. Paper presented at the *2011 Second International Conference on Digital Manufacturing & Automation,* 1349-1350.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers,* , 256-274.

Getis, A. (1996). Local spatial statistics: an overview. *Spatial Analysis: Modelling in a GIS*

*Environment.,* , 261-277.

Gianfranco, F., Soddu, S., & Fadda, P. (2018). An accident prediction model for urban road networks. *Journal of Transportation Safety & Security, 10*(4), 387-405.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis &*

*Prevention, 35*(2), 273-285. 10.1016/S0001-4575(02)00005-2

Gundogdu, I. B. (2010). Applying linear analysis methods to GIS-supported procedures for preventing traffic accidents: Case study of Konya. *Safety Science, 48*(6), 763-769. 10.1016/j.ssci.2010.02.016

Gupta, R., & Singh, M. (2014). Accident black spot validation using GIS. Paper presented at the *15th Esri India User Conference,*

Ha, H., & Thill, J. (2011). Analysis of traffic hazard intensity: A spatial epidemiology case study of urban pedestrians. *Computers, Environment and Urban Systems, 35*(3), 230240.

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary versus secondary.

Ihueze, C. C., & Onwurah, U. O. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. *Accident Analysis & Prevention, 112*, 21-29.

Iveta, M., Radovan, A., & Mihaljević, B. (2021). Prediction of Traffic Accidents Severity

Based on Machine Learning and Multiclass Classification Model. Paper presented at the *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO),* 1701-1705.

Kaffash Charandabi, N., Gholami, A., & Abdollahzadeh Bina, A. (2022). Road accident risk prediction using generalized regression neural network optimized with self-organizing map. *Neural Computing and Applications, 34*(11), 8511-8524.

Kahneman, D., Ben-Ishai, R., & Lotan, M. (1973). Relation of a test of attention to road accidents. *Journal of Applied Psychology, 58*(1), 113.

Kaliraja, C., Chitradevi, D., & Rajan, A. (2022). Predictive Analytics of Road Accidents Using Machine Learning. Paper presented at the *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE),* 1782-

1786.

Kelly, L. M., & Cordeiro, M. (2020). Three principles of pragmatism for research on organizational processes. *Methodological Innovations, 13*(2), 2059799120937242.

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. Paper presented at the *2019 7th International Conference on Smart Computing & Communications (ICSCC),* 1-5.

Le, K. G., Liu, P., & Lin, L. (2020). Determining the road traffic accident hotspots using GISbased temporal-spatial statistical analytic techniques in Hanoi, Vietnam. *Geo-Spatial Information Science, 23*(2), 153-164.

Lee, E. (2022). *What is Whiplash?* [https://cpdonline.co.uk/knowledge-](https://cpdonline.co.uk/knowledge-base/care/whiplash/)

[base/care/whiplash/.](https://cpdonline.co.uk/knowledge-base/care/whiplash/) [https://cpdonline.co.uk/knowledgebase/care/whiplash/#:~:text=Despite%20having%20the%20third%20safest%20roads% 20in%20the,been%20dubbed%20the%20whiplash%20capital%20of%20the%20world.](https://cpdonline.co.uk/knowledge-base/care/whiplash/#:~:text=Despite%20having%20the%20third%20safest%20roads%20in%20the,been%20dubbed%20the%20whiplash%20capital%20of%20the%20world.)

Lessmann, S., Stahlbock, R., & Crone, S. F. (2006). Genetic algorithms for support vector machine model selection. Paper presented at the *The 2006 IEEE International Joint Conference on Neural Network Proceedings,* 3063-3069.

Li, Y., Ma, D., Zhu, M., Zeng, Z., & Wang, Y. (2018). Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network. *Accident Analysis & Prevention, 111*, 354-363.

Liu, G., Chen, S., Zeng, Z., Cui, H., Fang, Y., Gu, D., Yin, Z., & Wang, Z. (2018). Risk factors for extremely serious road accidents: Results from national Road Accident Statistical Annual Report of China. *PLoS One, 13*(8), e0201587.

*LSOA Atlas*

*Greater London Authority (GLA).* (2014). [https://data.london.gov.uk/dataset/lsoaatlas.](https://data.london.gov.uk/dataset/lsoa-atlas.)<https://data.london.gov.uk/dataset/lsoa-atlas>

Lv, Y., Tang, S., & Zhao, H. (2009). Real-time highway traffic accident prediction based on the k-nearest neighbor method. Paper presented at the *2009 International Conference*

*on Measuring Technology and Mechatronics Automation, , 3* 547-550.

Maher, M. J., & Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention, 28*(3), 281-296.

Mahmoudabadi, A. (2010). Comparison of weighted and simple linear regression and artificial neural network models in Freeway Accidents Prediction. Paper presented at the *2010 Second International Conference on Computer and Network Technology,* 392396.

Mehta, K., Jain, S., Agarwal, A., & Bomnale, A. (2022). Road Accident Prediction Using Xgboost. Paper presented at the *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI),* 50-56.

Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention, 26*(4), 471-482.

Miaou, S., & Lum, H. (1993). Modeling vehicle accidents and highway geometric design relationships. *Accident Analysis & Prevention, 25*(6), 689-709.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical*

*Society.Series B (Methodological), 10*(2), 243-251.

Noked, N. (2010). Providing a corrective subsidy to insurers for success in reducing traffic accidents.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis, 27*(4), 286-306.

Ozgur, C., Colliau, T., Rogers, G., & Hughes, Z. (2017). MatLab vs. Python vs. R. *Journal of*

*Data Science, 15*(3), 355-371.

Pan, B., Demiryurek, U., Shahabi, C., & Gupta, C. (2013). Forecasting spatiotemporal impact of traffic incidents on road networks. Paper presented at the *2013 IEEE 13th International Conference on Data Mining,* 587-596.

Pape-Köhler, C. I., Simanski, C., Nienaber, U., & Lefering, R. (2014). External factors and the incidence of severe trauma: time, date, season and moon. *Injury, 45*, S93-S99.

Parizi, F. S., Fromm, J., Deshpande, S., & Patel, S. (2018). RoyalFlush: Non-invasive water level monitor to prevent toilet overflows. Paper presented at the *Proceedings of the 8th International Conference on the Internet of Things,* 1-8.

Park, S., Kim, S., & Ha, Y. (2016). Highway traffic accident prediction using VDS big data analysis. *The Journal of Supercomputing, 72*(7), 2815-2831.

Patton, C. L. (2011). Induction, deduction and cyclical movement: A review of qualitative research methods. *The Qualitative Report, 16*(5), 1421-1425.

Peden, M., Scurfield, R., Sleet, D., Mathers, C., Jarawan, E., Hyder, A. A., Mohan, D., Hyder, A. A., & Jarawan, E. (2004). *World report on road traffic injury prevention*. World Health Organization.

Qin, X., Chen, Z., & Shaon, R. R. (2019). Developing jurisdiction-specific SPFs and crash severity portion functions for rural two-lane, two-way intersections. *Journal of Transportation Safety & Security, 11*(6), 629-641.

Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.

Rashidi, M. H., Keshavarz, S., Pazari, P., Safahieh, N., & Samimi, A. (2022). Modeling the accuracy of traffic crash prediction models. *IATSS Research,*

Renggli, C., Rimanic, L., Gürel, N. M., Karlaš, B., Wu, W., & Zhang, C. (2021). A data quality-driven view of mlops. *arXiv Preprint arXiv:2102.07750,*

Sahay, A. (2016). Peeling Saunder's research onion. *Research Gate, Art,* , 1-5.

Saunders, M., Lewis, P., & Thornhill, A. (2007). Research methods. *Business Students 4th*

*Edition Pearson Education Limited, England,*

Saunders, M., & Tosey, P.Research Design.

Schell, C. (1992). The value of the case study as a research strategy. *Manchester Business*

*School, 2*(1), 1-15.

*Self Organizing Maps – Kohonen Maps.* (2022). [https://www.geeksforgeeks.org/selforganising-maps-kohonen-maps/. https://www.geeksforgeeks.org/self-organising-mapskohonen-maps/](https://www.geeksforgeeks.org/self-organising-maps-kohonen-maps/)

Sherman, L. W. (1995). Hot spots of crime and criminal careers of places. *Crime and*

*Place, 4*, 35-52.

Sikdar, P., Rabbani, A., & Dhapekar, N. K. (2017). Hypothesis of data of road accidents in

India-review. *International Journal of Civil Engineering and Technology (IJCIET), 8*(6), 141-146.

Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting Student Enrollment Based on

Student and College Characteristics. *International Educational Data Mining Society,*

Sowdagur, J. A., Rozbully-Sowdagur, B. T. B., & Suddul, G. (2022). An Artificial Neural Network Approach for Road Accident Severity Prediction. Paper presented at the *2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC),* 267-270.

Sun, J., Sun, J., & Chen, P. (2014). Use of support vector machine models for real-time prediction of crash risk on urban expressways. *Transportation Research Record, 2432*(1), 91-98.

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *How to Choose a Sampling Technique for Research (April 10, 2016),*

Tam Cho, W. K. (2003). Contagion effects and ethnic contribution networks. *American*

*Journal of Political Science, 47*(2), 368-387.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2016). Predicting road accidents: a rare-events modeling approach. *Transportation Research Procedia, 14*, 3399-3405.

Vezeteu, P. V., Morariu, A. R., & Năstac, D. I. (2021). Adaptive data predictions for the energy sector at national level. Paper presented at the *2021 IEEE 27th International Symposium for Design and Technology in Electronic Packaging (SIITME),* 292-297.

Viswanath, D., Preethi, K., Nandini, R., & Bhuvaneshwari, R. (2021). A Road Accident Prediction Model Using Data Mining Techniques. Paper presented at the *2021 5th International Conference on Computing Methodologies and Communication (ICCMC),* 1618-1623.

Weisburd, D., & Mazerolle, L. G. (2000). Crime and disorder in drug hot spots: Implications for theory and practice in policing. *Police Quarterly, 3*(3), 331-349.

Wenqi, L., Dongyu, L., & Menghua, Y. (2017a). A model of traffic accident prediction based on convolutional neural network. Paper presented at the *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE),* 198-202.

Wenqi, L., Dongyu, L., & Menghua, Y. (2017b). A model of traffic accident prediction based on convolutional neural network. Paper presented at the *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE),* 198-202.

WHO. (2022). *Road traffic injuries.* [https://www.who.int/health-topics/road-](https://www.who.int/health-topics/road-safety#tab=tab_1.)

[safety#tab=tab\_1.](https://www.who.int/health-topics/road-safety#tab=tab_1.)<https://www.who.int/health-topics/road-safety#tab=tab_1>

Yakar, F. (2015). Identification of accident-prone road sections by using relative frequency method. *Promet-Traffic&Transportation, 27*(6), 539-547.

Yang, W. (2020). Short-term traffic prediction based on historical trend method. Paper

presented at the *2020 International Conference on Urban Engineering and Management Science (ICUEMS),* 438-441.

Yu, H., Wu, Z., Wang, S., Wang, Y., & Ma, X. (2017). Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors, 17*(7), 1501.

Yu, L., Du, B., Hu, X., Sun, L., Han, L., & Lv, W. (2021). Deep spatio-temporal graph convolutional network for traffic accident prediction. *Neurocomputing, 423*, 135-147.

Yue, Y., & Yeh, A. G. (2008). Spatiotemporal traffic-flow dependency and short-term traffic forecasting. *Environment and Planning B: Planning and Design, 35*(5), 762-771.

**Appendix A**

**Task 2: Research Project Plan:**

**Road Traffic Accident Risk Prediction using a Classification Model**

**Rita Chinwenwa Uzoka**

*30034087*

# INTRODUCTION AND JUSTIFICATION

The World Health Organization estimates that 50 million individuals are impacted by and injured in traffic accidents each year, killing 1.2 million people. A multitude of issues, including traffic congestion, air pollution, and traffic accidents, have been brought on by the explosion of vehicles in modern society because of the rapid rise of urbanization. Both significant economic losses and human casualties have been brought on by these issues. Around 1.25 million people die in traffic accidents each year, according to the World Health

Organization's Global Status Report on Road Safety, which was published in 2015. Real-time traffic flow prediction made possible by massive traffic data and machine learning has helped people avoid gridlock by selecting less congested routes. To forecast or lower the danger of traffic accidents, big traffic data and machine learning may also offer a promising option (WHO, 2022).

# RESEARCH QUESTION, AIMS & OBJECTIVES

We look at open datasets and other data sources related to this field to try to accurately predict risk levels (RTAs).

## Research question

How can the risk of road traffic accident be predicted?

### Objectives

The following goals pertain to this work:

* Conduct a literature review to understand the existing work.
* To identify the factors that could have an impact on machine learning, use exploratory data analysis.
* Classify dangerous areas using spatial statistics on the secondary dataset.
* Using support vector machine methods, logistic regression, k-nearest neighbour, Rainforest, gradient boosting, and neural networks on a secondary dataset containing RTA characteristics These ML classifiers are employed mostly due to their distinguishing features and popularity in the literature.
* Evaluate and compare the performance of the classification models using various widely used evaluation techniques, such as accuracy, precision, recall, and the F1 score.
* To predict new regions, the best ML classifier is used.

### Deliverable

To investigate how localization patterns and hot spot distribution are used with temporal data, and to construct a highly precise machine learning model for predicting the likelihood of traffic accidents based on the spatiotemporal correlation pattern.

Literature review

This section analyses the literature within the domain that will assist to identify and justify the proposed project. The literature review focuses on the present research on road accident prediction and different models used to achieve it, it addresses the gaps to implement new solutions to the research problem.

One can get a general idea of the danger areas based on historical statistics of traffic accidents that have been gathered throughout the years. The risk distribution, however, varied significantly depending on the hour of the day, day of the week, and month of the year. Complex elements, such as crowd density, traffic flow, weather, events, etc., can also have an impact on accident risk. A fine-grained and dynamic prediction of accident risk cannot be made using historical statistics. Therefore, it is crucial to incorporate machine learning technology into traffic accident risk prediction to forecast the dynamic accident risk change in a precise and quantitative manner. Chen uses the Stack Denoising Autoencoder's human mobility properties to extrapolate the risk of traffic accidents in Japan (Liu et al., 2018) They did not, however, consider the cyclical and spatial distribution patterns of traffic accidents. The deficiencies in road engineering, crash investigation and post-crash management, legislative and driver education makes roads to be unsafe. Globally, road accidents are the main cause of death per year, reducing millions of lives on a yearly basis. Hence, a system that can predict the occurrence of traffic accidents or accident-prone areas can potentially save lives. The people don’t have to depend on the government to implement traffic engineering etc, they take ownership of the likeliness of accidents to occur and will be able to mitigate it which will save lives and money a company always must incur due to loss of life and goods,

(Ogwueleka, Misra, Ogwueleka, & Fernandez-Sanz, 2014)used Artificial Neural Networks able to relate input with output, enable large number of variables and can tolerate error, ANN provide solution to non-algorithm problems, able to use historical data to deliver fresh solutions, while this model provides effective prediction of road accidents its inability to identify accidents hot spots for better management of drivers along the road was not considered by the authors.

(Oyetunji, Oladeji, Falana, & Idowu, 2017)used naïve bayes a machine learning classifier as an effective prediction model for road accidents in Nigeria, the author demonstrated how the model classified each accident, weather was not included in the factors for prediction, not all states of the country were used for experiment also there is need to use a mechanism that can identify accident hot spots.

The authors (Al Najada & Mahgoub, 2016)implemented combined methodologies of a big data analytics architecture, machine learning classifiers, this system will be more feasibility on locations that utilize road engineering because of the use of road segments in the experiments, hence this experiment mayn’t work well in developing countries, the authors used the different machine learning for different purposes to which identify patterns and accidents hot spots was not included which is more feasible in a developing country. The precision and recall metric were used to confirm accuracy of the prediction model due to data imbalance as against most authors who used the accuracy metric in machine learning classifier algorithm for prediction, (Hébert, Guédon, Glatard, & Jaumard, 2019) the author recognises the effect of data imbalance, but weather was not used as a factor for prediction, also the author didn’t recognise patterns that could identify accident hot spots.

# RESEARCH DESIGN

## Research Philosophy

The term philosophy of research refers to a system of belief and assumptions on knowledge development, this research project tends to provide a practical approach which informs future practise to a problem or gap identified through review of various literature, this type of research philosophy is called Pragmatism, hence this research philosophy will be applied in the methodologies and techniques (Saunders, Lewis, & Thornhill, 2007). **Research Approach**

We have identified our research problem which is the research question and have adopted practical solutions or algorithms for improving the efficiency of the research problem and collect data to explore and solve it. This type of research approach is a mixed method of qualitative and quantitative research (Park et al., 2016). Hence this project is centred towards a mixed method of qualitative and quantitative research**.**

## Methodology

The three main areas of this project (EDA, Spatio-temporal Analysis and Machine Learning) will require a combined and independent evaluation of the research goals. Another methodology to collect and analyse data is required for each of those evaluations: - EDA: Both qualitative and quantitative data can be used, also qualitative data can be transformed to quantitative data for data analysis which makes it a mixed-method methodology. Machine Learning The nature of machine learning means that models are almost fully evaluated on quantitative data, in particular the precision and processing time percentages. A simple comparison of results from various ML models will allow for a fair evaluation while controlling all possible external variables, such as the use of the same data set.

Machine Learning Model: machine learning algorithm is considered for predictive modelling. In other words, by applying a function f to the independent variable x we can predict the value of a dependent variable y.

Generally, to forecast future outcomes, we use predictive modelling or analytics. We collect past data known as historical data on an event for this purpose systematically. We train a statistical model afterwards. To forecast future results, we use a trained statistical model. There are a variety of classifiers used for Machine Learning, the classifiers to be used are support vector machine methods, logistic regression, k-nearest neighbour, Rainforest, gradient boosting, and neural networks. Support vector machine produces considerable precision and less calculating capacity, and it can be used for both regression and classification tasks, as an abbreviation of SVM. However, it is widely used for classification purposes. Rainforest is mainly used for large dataset as an algorithm for constructing decision trees. Naïve Bayes is a machine learning algorithm used to resolve classification issues, it is mainly used in text classification, which includes a large data set for training. As the requirements of the Machine Learning model are developed, further research will be carried to determine if more suitable technologies are available.

**Time Horizon**

The time horizon will be cross-sectional because the project is conducted within 3months.

## Data Collection and Data Analysis Methodology Data Source

I obtained data for Great Britain from 1979 on reported road crashes broken down by severity, quantity, and rates of recorded road casualties by category of road users (Department for Transport, 2020). London was chosen because of its commitment to increasing traffic safety, as did Transport for London. An extensive consultation by Transport for London led to the London Assembly's evaluation of the Transport Strategy in 2018 and the revision in 2022. (TfL). The streets of London, where people live, work, and spend their free time, can be shaped by how they get around, including through using the train, bus, and tube systems (Copyright, Greater London Authority, 2018).

The names of London boroughs and LSOAs are both included in the second dataset. The

LSOA Atlas provides an overview of demographic and related statistics for each Lower Super Output Area (LSOA) in Greater London. The following is a list of the boroughs that make up London: The City of London, Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Camden, Croydon, Ealing, Enfield, Greenwich, Hackney, Hammersmith and

Fulham, Haringey, Harrow, Havering, Hillingdon, Hounslow, Islington, Kensington and

Chelsea, Kingston upon Thames, Lambeth, Lewisham, Merton, Newham, Redbridge,

Richmond upon Thames, Southwark, Sutton, Tower Hamlets, Waltham Forest, Wandsworth, Westminster (LSOA)Greater London Authority (GLA), 2014

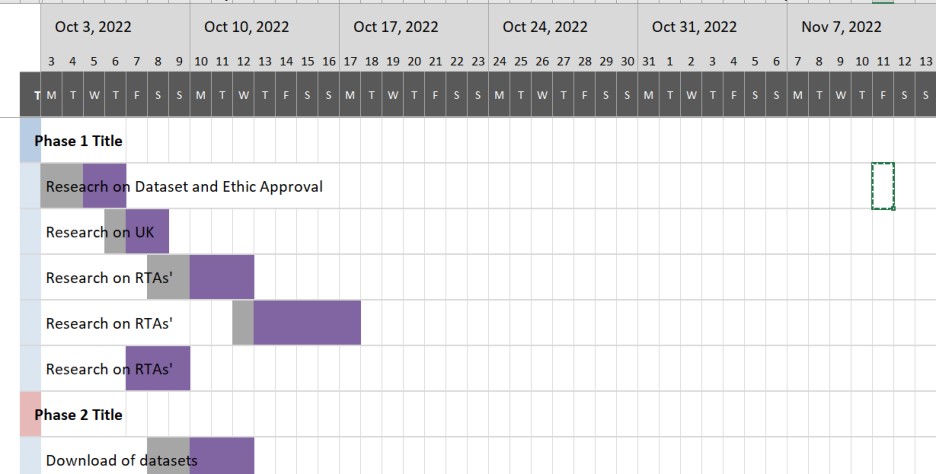
## Spatial Analysis

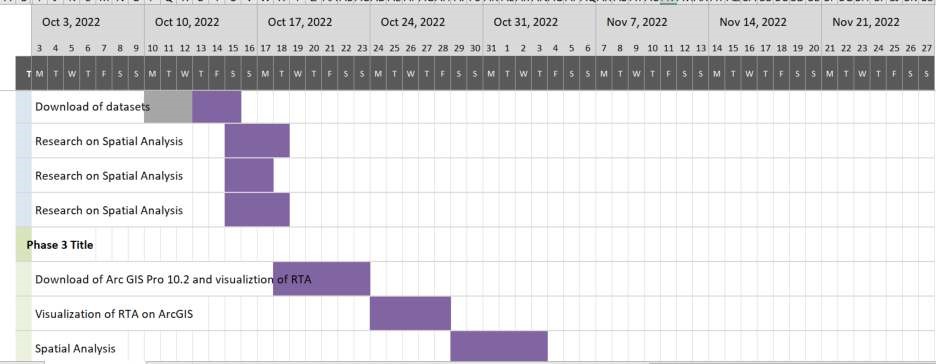
We can summarise the distribution of a variety of phenomena using spatial statistics. As a result, it aids in our decision-making. The use of spatial statistical analyses, including methods for pattern recognition, distribution analysis, and geographical correlations, allows the direct integration of spatial variables with conventional non-spatial statistical methods (neighbourhood, connections, and spatial relationships). spatial autocorrelation using the Global Moran's I tool, as well as Getis-OrdGi-based analysis of hot regions for all incidents. Arc GIS 10.2 and its add-ons will be used for all spatial processing.

# ETHICS, RISKS, AND ISSUES

|  |  |  |
| --- | --- | --- |
| **Potential Risks** | Issues | Mitigations |
| **Not securing Access to a Specialist Data Analysis and ML**  **Computer system** | Existing Computer will not be able to perform Data Analysis and ML due to Large Dataset, query etc | Access to any cloud infrastructure like AWS etc before the start of the project. |
| **Not securing access to the Arc GIS Pro**  **10.2** | Approval of securing access to Arc GIS Pro 10.2 throughout the duration of the project | Initiation to be made at an early stage, subscription to be secured. |
| **Saved data loss** | Redo of the project | Cloud backup of files |

Gantt Chart





References

Abdulkabir, M., Tunde, R. S., & Edem, U. A. (2015). Trend analysis on road traffic accident in nigeria. *Science Innovation, 3*(5), 52-57.

Al Najada, H., & Mahgoub, I. (2016). Anticipation and alert system of congestion and accidents in VANET using big data analysis for intelligent transportation systems. Paper presented at the *2016 IEEE Symposium Series on Computational Intelligence (SSCI),* 1-

8.

Hébert, A., Guédon, T., Glatard, T., & Jaumard, B. (2019). High-resolution road vehicle collision prediction for the city of montreal. Paper presented at the *2019 IEEE International Conference on Big Data (Big Data),* 1804-1813.

Ihueze, C. C., & Onwurah, U. O. (2018). Road traffic accidents prediction modelling: An analysis of anambra state, nigeria. *Accident Analysis & Prevention, 112*, 21-29.

Ogwueleka, F. N., Misra, S., Ogwueleka, T. C., & Fernandez-Sanz, L. (2014). An artificial neural network model for road accident prediction: A case study of a developing country. *Acta Polytechnica Hungarica, 11*(5), 177-197.

Oyetunji, M. O., Oladeji, F. A., Falana, O. J., & Idowu, P. A. (2017). Prediction of road traffic accident in nigeria using naive baye’s approach. *Advances in Multidisciplinary & Scientific Research Journal, 3*(1), 23-30.

Ren, H., Song, Y., Wang, J., Hu, Y., & Lei, J. (2018). A deep learning approach to the citywide traffic accident risk prediction. Paper presented at the *2018 21st International*

*Conference on Intelligent Transportation Systems (ITSC),* 3346-3351.

Ukoji, V. N. (2014). Trends and patterns of fatal road accidents in nigeria (2006– 2014). *Internet: Http://Www.Ifra-Nigeria.Org/IMG/Pdf/Fatalroad-Accidents-*

*Nigeria.Pdf Nov, 28*

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

*2018.* (). Retrieved from <https://www.who.int/publications/i/item/9789241565684>

World Health Organization. (2021). Road traffic injuries. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

**Appendix B**

**UREC2 RESEARCH ETHICS PROFORMA FOR STUDENTS UNDERTAKING LOW RISK PROJECTS WITH**

**HUMAN PARTICIPANTS**

This form is designed to help students and their supervisors to complete an ethical scrutiny of proposed research. The University [Research Ethics Policy s](https://www.shu.ac.uk/research/ethics-integrity-and-practice)hould be consulted before completing the form. The initial questions are there to check that completion of the UREC 2 is appropriate for this study. The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements to ensure compliance with the General Data Protection Act (GDPR). This involves informing participants about the legal basis for the research, including a link to the University research data privacy statement and providing details of who to complain to if participants have issues about how their data was handled or how they were treated (full details in module handbooks). In addition the act requires data to be kept securely and the identity of participants to be anonymized. They are also responsible for following SHU guidelines about data encryption and research data management. Information on the [Ethics Website](https://www.shu.ac.uk/research/quality/ethics-and-integrity/guidance-and-legislation)

The form also enables the University and College to keep a record confirming that research conducted has been subjected to ethical scrutiny.

The form may be completed by the student and the supervisor and/or module leader (as applicable). In all cases, it should be counter-signed by the supervisor and/or module leader, and kept as a record showing that ethical scrutiny has occurred. Some courses may require additional scrutiny. Students should retain a copy for inclusion in their research projects, and a copy should be uploaded to the relevant module Blackboard site.

Please note that it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the College Health and Safety Service.

**Checklist Questions to ensure that this is the correct form**

**1. Health Related Research with the NHS or Her Majesty’s Prison and Probation Service (HMPPS)or with participants unable to provide informed consent**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Does the research involve?    • Patients recruited because of their past or present use of the NHS | No |
| • Relatives/carers of patients recruited because of their past or present use of the NHS | No |
| • Access to data, organs or other bodily material of past or present NHS patients | No |
| • Foetal material and IVF involving NHS patients | No |
| • The recently dead in NHS premises | No |
| • Prisoners or others within the criminal justice system recruited for health-related research**\*** | No |
| • Police, court officials, prisoners or others within the criminal justice system**\*** | No |
| • Participants who are unable to provide informed consent due to their incapacity even if the project is not health related | No |
| 2. Is this a research project as opposed to service evaluation or audit? | No |
| *For NHS definitions of research etc. please see the following website* <http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf> |  |

If you have answered **YES** to questions **1 & 2** then you **MUST** seek the appropriate external approvals from the NHS, Her Majesty’s Prison and Probation Service (HMPPS) under their independent Research Governance schemes. Further information is provided below.

[https://www.myresearchproject.org.uk](https://www.myresearchproject.org.uk/Signin.aspx)

**NB** College Teaching Programme Research Ethics Committees (CTPRECS) provide Independent Scientific Review for NHS or HMPPS research and initial scrutiny for ethics applications as required for university sponsorship of the research. Applicants can use the IRAS proforma and submit this initially to their CTPREC.

**1. Checks for Research with Human Participants**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will any of the participants be vulnerable?  *Note: Vulnerable’ people include children and young people, people with learning disabilities, people who may be limited by age or sickness, people researched because of a condition they have, etc. See full definition on ethics website* | No |
| 2. Are drugs, placebos or other substances (e.g., food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind? | No |
| 3. Will tissue samples (including blood) be obtained from participants? | No |
| 4. Is pain or more than mild discomfort likely to result from the study? | No |
| 5. Will the study involve prolonged or repetitive testing? | No |
| 6. Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants?  *Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, or topics that are anxiety provoking, etc.* | No |
| 7. Will anyone be taking part without giving their informed consent? | No |
| 8. Is it covert research?  *Note: ‘Covert research’ refers to research that is conducted without the knowledge of participants.* | No |
| 9. Will the research output allow identification of any individual who has not given their express consent to be identified? | No |

If you have answered **YES** to any of these questions you are **REQUIRED** to complete and submit a UREC 3 or UREC4). Your supervisor will advise. If you have answered **NO** to all these questions then proceed with this form (UREC 2).

**General Details**

|  |  |
| --- | --- |
| Name of student | Rita Chinwenwa Uzoka |
| SHU email address | C0034087@my.shu.ac.uk |

|  |  |  |
| --- | --- | --- |
| Course or qualification (student) | MSc Big Data Analyst | |
| Name of supervisor | Olamilekan Shobayo | |
| email address | os6440@exchange.shu.ac.uk |  |
| Title of proposed research | **Road Traffic Accident Risk Prediction using a Classification Model** | |
| Proposed start date | October 3rd,2022 | |
| Proposed end date | January 8th ,2023 | |
| Background to the study and scientific rationale for undertaking it. | Building a reliable method for predicting traffic accidents is a crucial step in the avoidance of traffic accidents. We can advise surrounding cars to warn them or prompt them to take a less risky route if the likelihood of a traffic collision in a specific area can be forecast. However, due to the multiplicity of elements that might influence traffic accidents, it is highly challenging to predict with accuracy the risk of an accident. For instance, the number of road accidents varies  greatly between regions. Furthermore, unfavorable weather conditions like snow or fog can lower road visibility and traffic capacity, which raises the risk of traffic accidents. The number of traffic accidents varies throughout the day depending on the time of day, presumably reflecting the health of the drivers. Although a lot of study has been done on the primary components that contribute to traffic accidents. it is still difficult to estimate the likelihood of accidents occurring dynamically | |
| Aims & research question(s) | How to predict road traffic accident risk areas in UK? | |
| Methods to be used for:  1. recruitment of participants,    2.data collection,    3. data analysis. | Data Collection: The data collection method to use is secondary data that is already readily accessible at the UK Traffic website.  Data Analysis: exploratory data analysis and spatial analysis will be used. | |
| Outline the nature of the data held, details of anonymisation, storage and disposal procedures as required. | The data came from a UK road accident study dataset that had 36 components and 331,370 observations of crash events with fatalities and injuries (January 2018– December 2020). Each observation includes longitude, latitude, date, time, Lsoa of accident, number of accidents, light conditions, road conditions, pedestrians, number of casualties, type of route, fatalities, cause, speed, police presence at the scene of accident, LSOA districts, and | |
|  | overall weather at the time of the crash. | |

1. **Research in Organisations**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc.)? | No |
| 2. If you answered YES to question 1, do you have granted access to conduct the research?  *If YES, students please show evidence to your supervisor. PI should retain safely.* |  |
| 3. If you answered NO to question 2, is it because:   1. you have not yet asked 2. you have asked and not yet received an answer C. you have asked and been refused access.     *Note: You will only be able to start the research when you have been granted access.* |  |

1. **Research with Products and Artefacts**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programmes, databases, networks, processes, existing datasets, or secure data? | Yes |
| 2. If you answered YES to question 1, are the materials you intend to use in the public  domain?    *Notes: ‘In the public domain’ does not mean the same thing as ‘publicly accessible’.*   * *Information which is 'in the public domain' is no longer protected by copyright (i.e., copyright has either expired or been waived) and can be used without permission.* * *Information which is 'publicly accessible' (e.g., TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.*     *If you answered YES to question 1, be aware that you may need to consider other ethics codes.*  *For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.* | Publicly accessible |
| 3. If you answered NO to question 2, do you have explicit permission to use these materials as data?  *If YES, please show evidence to your supervisor.* |  |

|  |  |
| --- | --- |
| 4. If you answered NO to question 3, is it because:   1. you have not yet asked permission 2. you have asked and not yet received and answer 3. you have asked and been refused access.     *Note You will only be able to start the research when you have been granted* | **A/B/C** |

*permission to use the specified material.*

**Adherence to SHU policy and procedures**

|  |  |
| --- | --- |
| **Personal statement** | |
| I can confirm that:   * I have read the Sheffield Hallam University Research Ethics Policy and Procedures * I agree to abide by its principles. | |
| **Student** | |
| Name: Rita Chinwenwa Uzoka | Date: October 3rd 2022 |
| Signature: Rita | |
| **Supervisor or other person giving ethical sign-off** | |
| I can confirm that completion of this form has not identified the need for ethical approval by the FREC or an NHS, Social Care or other external REC. The research will not commence until any approvals required under Sections 3 & 4 have been received and any necessary health and safety measures are in place. | |
| Name: Olamilekan Shobayo Date: October 4th 2022 | |
| Signature: | |
| Additional Signature if required by course: | |
| Name: Date: | |
| Signature: | |

**Please ensure the following are included with this form if applicable, tick box to indicate:**

**Yes No N/A** Research proposal if prepared previously

Any recruitment materials (e.g. posters, letters, etc.)

Participant information sheet

Participant consent form

Details of measures to be used (e.g. questionnaires, etc.) Outline interview schedule / focus group schedule

Debriefing materials

Health and Safety Project Safety Plan for Procedures



College of Business,

Technology and   
Engineering



College of Business,

Technology and   
Engineering

Research Skills and   
Dissertation Module  
(55-706556)*.*

**PUBLICATION PROCEDURE FORM**

In this module, while you create your own research question or topic area, your supervisor makes a significant intellectual contribution to this work as the research progresses. Your supervisor will make the decision on whether your work merits publication based on the quality of the work you have produced. Your supervisor will co-author the paper for publication with you and your supervisor will both be listed as authors. You are required to sign the declaration below to confirm that you understand and will follow this procedure.

Declaration:

|  |  |  |
| --- | --- | --- |
| I Rita Chinwenwa Uzoka. Confirm that I understand will comply with the Publication Procedure outlined in the Module Handbook and the Blackboard Site. | | |
| **Student**: | Signature  Rita | Date  26/12/2022 |
| **Supervisor**: | Signature | Date  28/12/2022 |

**Appendix n**

**Fig 4.6,Fig 4.7 Heatmap of RTAs January 2018 – December 2020**

#!/usr/bin/env python

# coding: utf-8

# In[1]: import pandas as pd import numpy as np

heat\_df =pd.read\_csv('mast

erdatta\_1.csv') heat\_df.head()

# In[2]:

from datetime import datetime, timedelta

# In[4]:

heat\_df['date'] = pd.to\_datetime(heat\_df['date'])

# In[5]:

#Verify data type after data type

heat\_df.info()

# In[6]:

def to\_date\_format(n):

date\_str =(datetime(1899/12/30) + timedelta(n-1)).strftime("%d-%m-%Y") date\_date = datetime.strptime(date\_str, "%d-%m-%Y") return date\_date

# In[8]:

heat\_df['year'] =heat\_df['date'].dt.year heat\_df['month'] =heat\_df['date'].dt.month heat\_df['day'] = heat\_df['date'].dt.day

# In[9]:

heat\_df['date'].head()

# In[10]:

heat\_df.head()

# In[12]:

heat\_df["month"] = heat\_df["month"].map({1:'January',

2:'February',3:'March',4:'April',5:'May',6:'June',7:'July',8:'August',9:'September',10:'Oc tober',11:'November',12:'December'})

# In[14]:

heat\_df["day"] = heat\_df["day"].map({1:'Monday',

2:'Tuesday',3:'Wednesday',4:'Thursday',5:'Friday',6:'Saturday',7:'Sunday'})

# In[15]: heat\_df.head()

# In[17]:

# Find the monthly total for accidents for each day of the week (combine years 2018-

2020) def accidents\_month(df, month): RTAs\_month\_list = [] mon, tues, wed, thurs, fri, sat, sun = 0,0,0,0,0,0,0

for idx, row in heat\_df.iterrows():

if row["month"] == month and row["day"] == "Monday": mon +=1 elif row["month"] == month and row["day"] == "Tuesday": tues +=1 elif row["month"] == month and row["day"] == "Wednesday": wed +=1 elif row["month"] == month and row["day"] == "Thursday": thurs +=1 elif row["month"] == month and row["day"] == "Friday": fri +=1 elif row["month"] == month and row["day"] == "Saturday": sat +=1 elif row["month"] == month and row["day"] == "Sunday": sun +=1 else: a=0

RTAs\_month\_list.append(mon)

RTAs\_month\_list.append(tues)

RTAs\_month\_list.append(wed)

RTAs\_month\_list.append(thurs)

RTAs\_month\_list.append(fri)

RTAs\_month\_list.append(sat)

RTAs\_month\_list.append(sun)

total = mon+tues+wed+thurs+fri+sat+sun

return(RTAs\_month\_list)

# In[18]:

#Extract Road Traffic Accidents data by month january = accidents\_month(heat\_df, "January") february = accidents\_month(heat\_df, "February") march = accidents\_month(heat\_df, "March")

april = accidents\_month(heat\_df, "April")

may = accidents\_month(heat\_df, "May")

june = accidents\_month(heat\_df, "June")

july = accidents\_month(heat\_df, "July")

august = accidents\_month(heat\_df, "August") september = accidents\_month(heat\_df, "September") october = accidents\_month(heat\_df, "October") november = accidents\_month(heat\_df, "November") december = accidents\_month (heat\_df, "December")

# In[19]:

# Make matrix for the heatmap

columns = ["Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday",

"Sunday"]

rows = ["January", "February", "March", "April", "May", "June", "July", "August",

"September", "October", "November", "December"]

data = np.array([january, february, march, april, may, june, july, august, september, october, november, december]) heatMap = pd.DataFrame(data=data, index=rows, columns=columns)

# In[20]:

heatMap

# In[21]:

# **Make heatmap by months** import seaborn as sns import matplotlib.pyplot as plt fig =

plt.figure(figsize=(12,12)) map1 = sns.heatmap(heatMap, annot=True, fmt="d", cmap='BuPu', cbar\_kws={"orientation": "horizontal"} ) map1.set\_title("Traffic Accidents by week for each month from 2018-2020")

# In[23]:

#Function to determine number of accidents for each day of the week by hour

(combine years 2018-2020)

def accidents\_hour(df, hour): RTAs\_hour\_list = [] mon, tues, wed, thurs, fri, sat, sun = 0,0,0,0,0,0,0

for idx, row in df.iterrows():

if row["Hour"] == hour and row["day"] == "Monday": mon +=1 elif row["Hour"] == hour and row["day"] == "Tuesday": tues +=1 elif row["Hour"] == hour and row["day"] == "Wednesday": wed +=1 elif row["Hour"] == hour and row["day"] == "Thursday": thurs +=1 elif row["Hour"] == hour and row["day"] == "Friday": fri +=1 elif row["Hour"] == hour and row["day"] == "Saturday": sat +=1 elif row["Hour"] == hour and row["day"] == "Sunday": sun +=1 else: a=0

RTAs\_hour\_list.append(mon)

RTAs\_hour\_list.append(tues)

RTAs\_hour\_list.append(wed)

RTAs\_hour\_list.append(thurs)

RTAs\_hour\_list.append(fri)

RTAs\_hour\_list.append(sat)

RTAs\_hour\_list.append(sun)

total = mon+tues+wed+thurs+fri+sat+sun

return(RTAs\_hour\_list)

# In[24]:

hour\_0 = accidents\_hour(heat\_df, 0) hour\_1 = accidents\_hour(heat\_df, 1) hour\_2 = accidents\_hour(heat\_df, 2) hour\_3 = accidents\_hour(heat\_df, 3) hour\_4 = accidents\_hour(heat\_df, 4) hour\_5 = accidents\_hour(heat\_df, 5) hour\_6 = accidents\_hour(heat\_df, 6) hour\_7 = accidents\_hour(heat\_df, 7) hour\_8 = accidents\_hour(heat\_df, 8) hour\_9 = accidents\_hour(heat\_df, 9) hour\_10 = accidents\_hour(heat\_df, 10) hour\_11 = accidents\_hour(heat\_df, 11) hour\_12 = accidents\_hour(heat\_df, 12) hour\_13 = accidents\_hour(heat\_df, 13) hour\_14 = accidents\_hour(heat\_df, 14) hour\_15 = accidents\_hour(heat\_df, 15) hour\_16 = accidents\_hour(heat\_df, 16) hour\_17 = accidents\_hour(heat\_df, 17) hour\_18 = accidents\_hour(heat\_df, 18) hour\_19 = accidents\_hour(heat\_df, 19) hour\_20 = accidents\_hour(heat\_df, 20) hour\_21 = accidents\_hour(heat\_df, 21) hour\_22 = accidents\_hour(heat\_df, 22) hour\_23 = accidents\_hour(heat\_df, 23) hour\_24 = accidents\_hour(heat\_df, 24)

# In[25]:

# Make matrix for the heatmap

columns = ["Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"]

rows = ["0", "1am", "2am", "3am", "4am", "5am", "6am", "7am", "8am", "9am",

"10am", "11am", "12pm", "1pm", "2pm", "3pm","4pm","5pm","6pm","7pm","8pm","9pm","10pm","11pm"]

data = np.array([hour\_0, hour\_1, hour\_2, hour\_3, hour\_4, hour\_5, hour\_6, hour\_7, hour\_8, hour\_9, hour\_10, hour\_11, hour\_12, hour\_13, hour\_14, hour\_15, hour\_16, hour\_17, hour\_18, hour\_19, hour\_20, hour\_21, hour\_22, hour\_23])

heatMap\_by\_hour = pd.DataFrame(data=data, index=rows, columns=columns)

# In[26]:

# Make heatmap import seaborn as sns import matplotlib.pyplot as plt fig = plt.figure(figsize=(12,12))

ax = sns.heatmap(heatMap\_by\_hour, annot=True, fmt="d", cmap='BuPu', cbar\_kws={"orientation": "horizontal"} ) ax.set\_title("Traffic Accidents by hour for each day from 2018-2020")

# In[ ]:

**Figure 4.5 Choropleth Map showing total RTAs in London boroughs from 2018-2020**

#!/usr/bin/env python

# coding: utf-8

# In[1]:

pip install descartes

# In[1]:

import geopandas as gpd from geopandas.tools import sjoin import pandas as pd

import matplotlib.pyplot as plt from fiona.crs import from\_epsg from descartes.patch import PolygonPatch

# In[3]: #Load file

df =pd.read\_csv('masterdatta.csv')

#df.head()

df.head()

# In[4]:

#list of boroughs in London

borough\_df = df['Borough'].unique() print("No. of boroughs:", len(borough\_df)) for i in range(len(borough\_df)):

print(i+1, borough\_df[i])

# In[5]:

#Get total counts of RTAs for each borough (from 2009-2014) borough = []

rta = []

for i in borough\_df:

count = 0

for idx, row in df.iterrows(): if row["Borough"] == i:

count +=1 rta.append(count) borough.append(i)

# In[6]: type(borough)

# In[7]:

#Create a dataframe with the total number of RTAs by borough (2018–2020).

rta\_all = pd.DataFrame() rta\_all["borough"] = borough

rta\_all["rta\_all"] = rta

rta\_all.sort\_values(by='rta\_all', ascending=False, inplace = True)

#Add a column ranking the boroughs by the number of RTA incidents, from most to least. rank = [] for i in range(33):

b = str(i+1)

rank.append(b)

rta\_all["rank"] = rank

# Load and prepare geo data

# In[15]:

# Load a shapefile and specify the file directory.

geo\_df=gpd.read\_file('London\_geodata\_info/ESRI/London\_Borough\_Excluding\_MHW.shp')

# GEOdataframe is the type of data (not a normal dataframe). type(geo\_df)

# In[16]:

# View GEOdataframe geo\_df.head()

# In[17]:

#View London map without any data

geo\_df.plot()

# In[18]:

# Set the "geometry" coordinates to the borough geometry's centre (for labels).

geo\_df["center"] = geo\_df["geometry"].centroid geo\_center = geo\_df.set\_geometry("center") geo\_center

# In[23]:

merged = geo\_df.set\_index("NAME").join(rta\_all.set\_index("borough"))

merged

# Create choropleth map # In[27]:

# set the choropleth's range

chmin = merged["rta\_all"].min() chmax = merged["rta\_all"].max()

# make a figure and axes for Matplotlib fig, ax = plt.subplots(1, figsize=(10, 6))

#make a map

merged.plot(column="rta\_all", cmap="BuPu", linewidth=0.8, ax=ax, edgecolor= "1.0")

# drop the axis ax.axis("off") # include a title

ax.set\_title('London RTAs (Road Traffic Accidents)', fontdict={'fontsize': '25',

'fontweight' : '3'})

# make a data source annotation

ax.annotate('Source: London Datastore, 2014',xy=(0.1, .08), xycoords='figure fraction', horizontalalignment='left', verticalalignment='top', fontsize=12, color='#555555')

# make a colorbar legend

sm = plt.cm.ScalarMappable(cmap="BuPu", norm=plt.Normalize(vmin=chmin, vmax=chmax))

# empty array for the data range sm.\_A = []

# include the colorbar to the figure cbar = fig.colorbar(sm) #include rank labels

texts = [] for i in range(len(merged)): x = merged["center"][i].x y = merged["center"][i].y label = merged["rank"][i]

texts.append(plt.text(x, y, label, fontsize = 10, fontweight='bold' ))

# In[15]:

fig.savefig("map\_export.png", dpi=300)

# In[ ]:

**Data Exploration**

#!/usr/bin/env python

# coding: utf-8

#!/usr/bin/env python

# coding: utf-8

# In[2]:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import random

import numpy as np

path="dft-road-casualty-statistics-accident-1979-2020.csv"

# In[3]:

a=pd.read\_csv(path)

# In[4]:

a.tail(5000)

# In[5]:

import numpy as np

a.shape

# In[6]:

# Filter data between two dates

filtered\_df = a.loc[(a['accident\_year'] <= 2020)

& (a['accident\_year'] >= 2018)]

# In[7]:

filtered\_df.shape

# In[8]:

filtered\_df.head(50)

# In[9]:

filtered\_df.tail(50)

# In[10]:

filtered\_df.info()

# In[11]:

b=pd.read\_csv("Lsoa.csv")

# In[70]:

#write dataFrame to CSV file

filtered\_df.to\_csv("C:\\Users\\ritau\\Downloads\\filtered\_df1.csv")

# In[12]:

masterdatta = filtered\_df.merge(b, on= 'lsoa\_of\_accident\_location')

# In[13]:

masterdatta.head()

# In[67]:

masterdatta.to\_csv("C:\\Users\\ritau\\Downloads\\masterdatta.csv")

# In[69]:

masterdatta.to\_csv("C:\\Users\\ritau\\Downloads\\masterdatta1.csv")

# In[14]:

masterdatta.tail()

# In[15]:

masterdatta.shape

# In[16]:

masterdatta["Accident\_Severity"] = masterdatta["accident\_severity"].map({1:'Fatal', 2:'Serious',3:'Slight'})

# In[17]:

masterdatta['Weather\_Conditions'] = masterdatta['weather\_conditions'].map({1:'Fine + HighWinds', 2:'Fine no HighWinds',3:'Fog or Mist',4:'Other',5:'Raining + HighWinds',6:'Raining no HighWinds',7:'Snowing + HighWinds',8:'Snowing + no HighWinds',9:'Unknown'})

# In[18]:

masterdatta['Road\_Surface\_Conditions'] = masterdatta['road\_surface\_conditions'].map({1:'Dry', 2:'Flood over 3cm. deep',3:'Frost or ice',4:'Snow',5:'Unknown',6:'Wet or Damp'})

# In[19]:

masterdatta.head()

# In[20]:

a, axs = plt.subplots(1, 2, figsize=(8, 4), gridspec\_kw=dict(width\_ratios=[4, 3]))

sns.scatterplot(data=masterdatta, x="day\_of\_week", y="number\_of\_casualties", ax=axs[0])

sns.histplot(data=masterdatta, x="day\_of\_week", shrink=.8, alpha=.8, legend=True, ax=axs[1])

a.tight\_layout()

# In[21]:

# plot of accident severity

wplot = sns.countplot(data=masterdatta,x="Accident\_Severity")

wplot.set\_xticklabels(wplot.get\_xticklabels(), rotation=40, ha="right")

plt.show()

# In[22]:

# plot weather conditions

wplot = sns.countplot(data=masterdatta,x="Weather\_Conditions")

wplot.set\_xticklabels(wplot.get\_xticklabels(), rotation=40, ha="right")

plt.show()

# In[23]:

# plot of road surface conditions

wplot = sns.countplot(data=masterdatta,x="Road\_Surface\_Conditions")

wplot.set\_xticklabels(wplot.get\_xticklabels(), rotation=40, ha="right")

plt.show()

# Data Cleaning

# In[24]:

hp = pd.read\_excel("Hot\_Spot\_Analysis.xlsx", sheet\_name = ('HotSpot'))

# In[25]:

hp.head()

# In[26]:

hp["HotSpot\_Analysis"] = hp["Analysis"].map({0:'ColdSpot',1:'HotSpot'})

# In[27]:

hp.head()

# In[28]:

hp.info()

**ArcGIS Pro 10.2 mapping of Spatial Clustering**

# In[29]:

hp["Confidence\_level"] = hp["Gi\_Bin Fixed 50\_FDR"].map({+3 or -3:99, +2 or -2:95,+1 or -1:90,0:'Not Significant'})

# In[30]:

hp["Risk\_Level"] = hp["Confidence\_level"].map({'Not Significant':'Low\_risk',99:'High\_risk',95:'High\_risk',90:'High\_risk'})

# In[31]:

hp["Risk\_Level\_num"] = hp["Confidence\_level"].map({'Not Significant':0,99:1,95:1,90:1})

# In[32]:

hp.head()

# In[33]:

main = pd.read\_excel("Hot\_Spot\_Analysis.xlsx", sheet\_name = ('mainfile'))

# In[34]:

main.head()

# In[35]:

main\_data=main.merge(hp,on='id')

# In[36]:

main\_data.head()

# In[37]:

main\_data.shape

# In[38]:

main\_data.info()

**Correlation Matrix**

# In[39]:

main\_data.corr(method = 'pearson')

# In[40]:

corr =main\_data.select\_dtypes(include=[np.number]).corr()

mask = np.zeros\_like(corr, dtype=np.bool)

mask[np.triu\_indices\_from(mask)] = True

f, ax = plt.subplots(figsize=(11, 9))

# Generate a custom diverging colormap

cmap = sns.palette="vlag"

# Draw the heatmap with the mask and correct aspect ratio

sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,

square=True, linewidths=.5, cbar\_kws={"shrink": .5});

# Feature Engineering

# In[57]:

#write dataFrame to CSV file

main\_data.to\_csv("C:\\Users\\ritau\\Downloads\\main\_data.csv")

# In[3]:

maindata = pd.read\_csv("main\_data1.csv")

# In[4]:

maindata.head()

# In[5]:

maindata.info()

# In[6]:

from datetime import datetime, timedelta

# In[7]:

maindata['date'] = pd.to\_datetime(maindata['date'])

# In[8]:

maindata.info()

# In[9]:

import matplotlib.pyplot as plt

get\_ipython().run\_line\_magic('matplotlib', 'inline')

maindata.hist(figsize=(20,16), color = 'r');

plt.show();

# In[10]:

maindata.corr(method = 'pearson')

# In[11]:

corr =maindata.select\_dtypes(include=[np.number]).corr()

mask = np.zeros\_like(corr, dtype=np.bool)

mask[np.triu\_indices\_from(mask)] = True

f, ax = plt.subplots(figsize=(11, 9))

# Generate a custom diverging colormap

cmap = sns.palette="vlag"

# Draw the heatmap with the mask and correct aspect ratio

sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,

square=True, linewidths=.5, cbar\_kws={"shrink": .5});

# In[12]:

sns.heatmap(corr,cmap="crest")

# In[13]:

fig, ax = plt.subplots(1, 2, figsize=(10,3))

sns.distplot(maindata['Risk\_Level\_num'], ax=ax[0])

sns.distplot(np.cbrt(maindata['Risk\_Level\_num']), ax=ax[1])

plt.show()

print(maindata['Risk\_Level\_num'].skew().round(2))

print(np.cbrt(maindata['Risk\_Level\_num']).skew().round(2))

# In[14]:

fig, ax = plt.subplots(1, 2, figsize=(10,3))

sns.distplot(maindata['number\_of\_casualties'], ax=ax[0])

sns.distplot(np.log(maindata['number\_of\_casualties']), ax=ax[1])

plt.show()

print(maindata['number\_of\_casualties'].skew().round(2))

print(np.log(maindata['number\_of\_casualties']).skew().round(2))

**Modelling**

# In[15]:

maindata.shape

# In[16]:

from sklearn.preprocessing import MinMaxScaler

y = maindata['Risk\_Level']

X= maindata.drop(['date','Risk\_Level','Risk\_Level\_num'],axis =1)

scaler = MinMaxScaler()

data = pd.DataFrame(scaler.fit\_transform(X), index=y.index)

data.columns = X.columns

data

# In[17]:

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.10, random\_state=42)

# In[18]:

X\_train.shape

# In[19]:

X\_test.shape

# In[20]:

from sklearn.neighbors import KNeighborsClassifier

# In[21]:

knn = KNeighborsClassifier(n\_neighbors=3)

# In[22]:

knn\_model = knn.fit(X\_train, y\_train)#trains the classification model

# In[23]:

predict\_knn = knn\_model.predict(X\_test)#predicts the class of the data using classification model

# In[24]:

predict\_knn[:10]#Let's see the first 10 predictions

# In[25]:

from sklearn.metrics import confusion\_matrix, classification\_report

# In[26]:

print(classification\_report(y\_test,predict\_knn))

# In[27]:

print(confusion\_matrix(y\_test,predict\_knn))

# In[28]:

from sklearn.metrics import accuracy\_score, recall\_score, roc\_auc\_score, precision\_score, f1\_score

f1\_score(y\_test,predict\_knn, average='micro')

# In[29]:

knn=f1\_score(y\_test,predict\_knn, average='macro')

print(knn)

# In[30]:

f1\_score(y\_test,predict\_knn, average='weighted')

# In[31]:

recall\_score(y\_test,predict\_knn,average='weighted')

# In[32]:

recall\_score(y\_test,predict\_knn,average='macro')

# In[33]:

precision\_score(y\_test,predict\_knn,average='weighted')

# In[34]:

from sklearn. model\_selection import cross\_val\_score

scores = cross\_val\_score(knn\_model, X\_train, y\_train, cv=10)

acc\_knn = round(scores.mean(),2)

print("Accuracy:",acc\_knn)

print("Accuracy:",round(scores.mean()\*100,2),"%")

# In[35]:

from sklearn.ensemble import RandomForestClassifier

# In[36]:

rd = RandomForestClassifier(n\_estimators=150)

rd\_model = rd.fit(X\_train,y\_train)

# In[37]:

predict\_rd = rd\_model.predict(X\_test)

# In[38]:

predict\_rd

# In[39]:

from sklearn.metrics import f1\_score

print(classification\_report(y\_test,predict\_rd))

#print(f1\_score(y\_test,predict\_rd, average=None))

# In[40]:

rf = f1\_score(y\_test,predict\_rd, average='macro')

print(rf)

# In[41]:

from sklearn.model\_selection import cross\_val\_score

scores1 = cross\_val\_score(rd\_model, X\_train, y\_train, cv=10)

acc\_rf = round(scores1.mean(),2)

print("Accuracy:",acc\_rf)

print("Accuracy:",round(scores1.mean()\*100,2),"%")

# In[103]:

from sklearn.svm import SVC

svc = SVC(kernel='rbf', C=1, gamma='auto')

# In[104]:

svc\_model = svc.fit(X\_train,y\_train)

# In[105]:

predict\_svc = svc\_model.predict(X\_test)

# In[106]:

predict\_svc[:10]

# In[107]:

print(classification\_report(y\_test,predict\_svc))

# In[108]:

print(confusion\_matrix(y\_test,predict\_svc))

# In[37]:

scores2 = cross\_val\_score(svc\_model, X\_train, y\_train, cv=10)

print("Accuracy:",round(scores2.mean()\*100,2),"%")

# In[42]:

from sklearn.neural\_network import MLPClassifier

from sklearn.datasets import make\_classification

clf = MLPClassifier(random\_state=1,activation='relu', max\_iter=300).fit(X\_train, y\_train)

predict\_nn = clf.predict(X\_test)

predict\_nn

# In[43]:

print(classification\_report(y\_test,predict\_nn))

# In[44]:

nn = f1\_score(y\_test,predict\_nn, average='macro')

print(nn)

# In[45]:

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

acc\_nnr = accuracy\_score(y\_test, predict\_nn)

print(f"Accuracy Score of Neural Network Classifier is :{acc\_nnr}")

# In[48]:

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

from sklearn.ensemble import GradientBoostingClassifier

gb = GradientBoostingClassifier()

gb\_predict = gb.fit(X\_train, y\_train)

y\_pred\_gb = gb\_predict.predict(X\_test)

acc\_gb = accuracy\_score(y\_test, y\_pred\_gb)

conf = confusion\_matrix(y\_test, y\_pred\_gb)

clf\_report = classification\_report(y\_test, y\_pred\_gb)

print(f"Accuracy Score of Gradient Boost Classifier is :{acc\_gb}")

print(f"Confusion Matrix : \n{conf}")

print(f"Classification Report : \n{clf\_report}")

# In[49]:

y\_pred\_gb

# In[50]:

gb = f1\_score(y\_test,y\_pred\_gb, average='macro')

print(gb)

# In[51]:

from sklearn.tree import DecisionTreeClassifier

dt = DecisionTreeClassifier()

dt\_predict=dt.fit(X\_train, y\_train)

y\_pred\_dt = dt\_predict.predict(X\_test)

acc\_dt = accuracy\_score(y\_test, y\_pred\_dt)

conf = confusion\_matrix(y\_test, y\_pred\_dt)

clf\_report = classification\_report(y\_test, y\_pred\_dt)

print(f"Accuracy Score of Decision Tree is : {acc\_dt}")

print(f"Confusion Matrix : \n{conf}")

print(f"Classification Report : \n{clf\_report}")

# In[53]:

dt = f1\_score(y\_test,y\_pred\_dt, average='macro')

print(dt)

# In[57]:

from sklearn.linear\_model import LogisticRegression

lr = LogisticRegression()

lr\_predict =lr.fit(X\_train, y\_train)

y\_pred\_lr = lr\_predict.predict(X\_test)

acc\_lr = accuracy\_score(y\_test, y\_pred\_lr)

conf = confusion\_matrix(y\_test, y\_pred\_lr)

clf\_report = classification\_report(y\_test, y\_pred\_lr)

print(f"Accuracy Score of Logistic Regression is : {acc\_lr}")

print(f"Confusion Matrix : \n{conf}")

print(f"Classification Report : \n{clf\_report}")

# In[58]:

lr = f1\_score(y\_test,y\_pred\_lr, average='macro')

print(lr)

# In[59]:

models = pd.DataFrame({

'Model' : ['KNN','Random Forest Classifier','NNR','Decision Tree Classifier',

'Gradient Boosting Classifier','Logistic Regression'],

'Score' : [acc\_knn, acc\_rf, acc\_nnr, acc\_gb, acc\_dt, acc\_lr]

})

models.sort\_values(by = 'Score', ascending = False)

# In[131]:

pip install plotly

# In[60]:

import folium

import matplotlib.pyplot as plt

from folium.plugins import HeatMap

import plotly.express as px

plt.style.use('fivethirtyeight')

get\_ipython().run\_line\_magic('matplotlib', 'inline')

pd.set\_option('display.max\_columns', 32)

px.bar(data\_frame = models, x = 'Score', y = 'Model', color = 'Score', template = 'plotly\_dark', title = 'Models Comparison')

# In[61]:

models = pd.DataFrame({

'Model' : ['KNN','Random Forest Classifier','NNR','Decision Tree Classifier',

'Gradient Boosting Classifier','Logistic Regression'],

'F1\_Score' : [knn, rf, nn, dt, gb, lr]

})

models.sort\_values(by = 'F1\_Score', ascending = False)

# In[62]:

import folium

import matplotlib.pyplot as plt

from folium.plugins import HeatMap

import plotly.express as px

plt.style.use('fivethirtyeight')

get\_ipython().run\_line\_magic('matplotlib', 'inline')

pd.set\_option('display.max\_columns', 32)

px.bar(data\_frame = models, x = 'F1\_Score', y = 'Model', color = 'F1\_Score', template = 'plotly\_dark', title = 'Models Comparison')

# In[63]:

rfscore\_df = pd.DataFrame(data=[rf,acc\_rf],

columns=['Random Forest Score'],

index=["F1","Accuracy"])

# In[64]:

Knnscore\_df = pd.DataFrame(data=[knn,acc\_knn],

columns=['KNN'],

index=["F1","Accuracy"])

# In[65]:

gbscore\_df = pd.DataFrame(data=[gb,acc\_gb],

columns=['KNN'],

index=["F1","Accuracy"])

# In[66]:

nnscore\_df = pd.DataFrame(data=[nn,acc\_nnr],

columns=['Neural Network'],

index=["F1","Accuracy"])

# In[67]:

dtscore\_df = pd.DataFrame(data=[dt,acc\_dt],

columns=['Decision Tree Classifier'],

index=["F1","Accuracy"])

# In[68]:

lrscore\_df = pd.DataFrame(data=[lr,acc\_lr],

columns=['Logistic Regression'],

index=["F1","Accuracy"])

# In[69]:

models\_df = round(pd.concat([Knnscore\_df,rfscore\_df, nnscore\_df,dtscore\_df,gbscore\_df,lrscore\_df], axis=1),3)

# In[70]:

model\_metric = models\_df

import matplotlib

colours = ["lightgreen","lightpink","#00FF00"]

colourmap = matplotlib.colors.LinearSegmentedColormap.from\_list("", colors)

back\_colour = "#fbfbfb"

fig = plt.figure(figsize=(9,7)) # make a figure

gs = fig.add\_gridspec(2, 1)

gs.update(wspace=0.1, hspace=0.5)

ax0 = fig.add\_subplot(gs[0, :])

sns.heatmap(model\_metric.T, cmap=colourmap,annot=True,fmt=".1%",vmin=0,vmax=0.95, linewidths=1.5,cbar=False,ax=ax0,annot\_kws={"fontsize":12})

fig.patch.set\_facecolor(back\_colour) # background color of the figure

ax0.set\_facecolor(back\_colour)

ax0.text(1,-4.15,'Model Comparison',fontsize=18,fontweight='bold',fontfamily='serif')

ax0.text(0,-1.0,'Although the best accuracy is achieved by logistic regression,\n is this sufficient? In this situation, does the F1 score (macro) matter more?',fontsize=14,fontfamily='serif')

ax0.tick\_params(axis=u'both', which=u'both',length=0)

plt.show()

# In[71]:

new\_data = pd.read\_csv("new\_data.csv")

# In[72]:

new\_data.head()

# In[73]:

new\_data.info()

# In[74]:

predict\_nd = dt\_predict.predict(new\_data)

# In[327]:

predict\_nd

# In[ ]: