Application of Machine learning and Big Data in Enhancing Road Safety by Predicting Accidents

Kirimi, Dennis Mwenda

148988

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Analytics Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

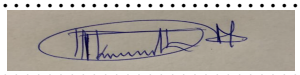
Nairobi, Kenya

October 2024

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Kirimi Dennis Mwenda**

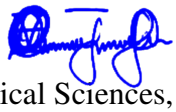
Signature: 

Date: September 27, 2024

Approval

The thesis of Kirimi Dennis Mwenda was reviewed and approved by the following:

Dr. Evans Otieno Omondi

Supervisor,  September 27, 2024

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Background: Road traffic accidents have been a recurring global problem that cause over 1.3 million lives annually, with a significant impact on the mortality of children and young people. Effective accident prevention strategies have been required due to the ensuing expenses to society, the legal system, and the hospital system. The goal of the current work is to create a safe system that integrates engineering, traffic control, and vehicle standards for accident predictions to improve road safety.

Method: The algorithms Gradient Boosting Classifier, XGB Classifier, and Random Forest Classifier were utilized to forecast accident severity and produce an interactive map that could locate accident hotspots and estimate accident frequencies.

Results: In total 660,679 cases the dataset provided a solid basis for examining trends and patterns in traffic safety with the Gradient Boosting Classifier accurately predicting the model with over 85%. A total of 16.74% were classified as serious accidents and 83.26% as slight. In the validation phase, our final accident prediction model, utilizing Gradient Boosting Classifier algorithm, showcased outstanding performance metrics. With an accuracy score of 0.85, it demonstrated exceptional proficiency and was highly likely to be accurate when predicting a positive instance. However, both its precision stood at 0.71, indicating that while it was accurate, it was still effective at detecting all positive instances or achieving total correctness.

Conclusion: The accident severity prediction model demonstrated accuracy and reliability, as evidenced by its successful external validation using high-quality registry data within the UK. This model holds significant potential for deployment in improving road safety analytics, enabling stakeholders to effectively assess and compare accident severity outcomes across different regions and transportation systems.

KEY WORDS: *Road Safety, Machine learning, algorithms, Big Data, Accident Prediction, Road Safety, Classification Techniques.*

Table of contents

List of figures

Acknowledgement

With deep appreciation, I would like to thank everyone who helped shape this Research dissertation on the "Application of Machine Learning in Predicting Road Accident Severity."

First and foremost, I would want to express my sincere gratitude to my supervisor (Dr. Evans Omondi) for all of his help and support during the Research dissertation process. This study's emphasis and direction have been greatly influenced by the knowledge and insights that have been given.

I express my gratitude to Strathmore University for furnishing the requisite resources and fostering a favorable atmosphere for conducting this study. Without their assistance, the chance to investigate the relationship between machine learning and traffic safety would not have been feasible. In addition, I would want to thank my family, friends, and colleagues for their support and understanding throughout this project. This Research dissertation is the result of a team effort, and for that, I am very appreciative to everyone who has contributed to its development.

Dedication

My beloved family, whose unending love, support, and understanding have been the cornerstone of my academic path, is the recipient of my research dissertation dedication. Your consistent support has given me the strength to pursue this scientific endeavor and has inspired my goals.

To my wonderful supervisor, Dr. Evans Omondi, I am truly appreciative. This study has evolved as a result of your advice, knowledge, and mentoring, and I sincerely appreciate the priceless lessons and insights you have imparted.

I also dedicate this idea to Strathmore University, an organization that has offered a platform for invention and discovery in addition to the academic atmosphere required for progress. My love of research and my insatiable curiosity have been greatly influenced by my experiences at Strathmore.

This dedication serves as a token of my gratitude for my family, supervisor, and Strathmore University's combined impact and support as I pursue knowledge and academic greatness.

Chapter 1

Introduction

1.1 Background of the study

Road traffic crashes are a continual human problem and a development issue. The World Health Organization (WHO) estimates that over 1.3 million deaths occur each year from roads globally, with road traffic accidents being the top cause of child and youth mortality (age group of 5-29 years). Besides the human losses, traffic accidents are a secondary burden to health services and cause extensive expenditures from medical bills, court actions, mobilized emergency rescue units, insurance payments, and workplace absence (Hua et al., 2023). According to the WHO, sub-optimal interaction among elements involving road traffic systems like vehicles, drivers, and roads is the root of all cases involving road collisions. It describes a safe system for road safety that includes interventions of engineering measures (infrastructure designs), traffic operations control (legislation and enforcement), and motor vehicle standards (Fiorentini and Losa, 2020). Nevertheless, this development is progressing quickly in most countries with sound road safety initiatives. The only way to realize substantive leaps in international road safety is by developing creative means involving modern technologies such as artificial intelligence and extensive data analysis. As observed, there are various causes of this increase, such as poor traffic system management, ineffective infrastructure, and driver's acts . In addition to these things, people believe that accidents happen every day while one is going home. However, solving such problems necessitates knowledge of the dimensions of the issue and novel strategies for limiting the frequency and intensity of future accidents.

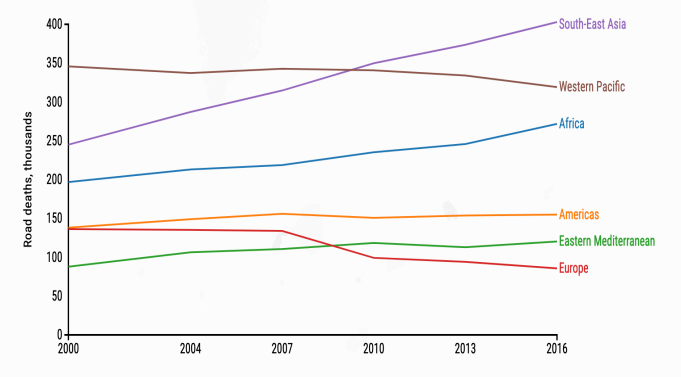


Figure 1.1: Deaths on Roads

Road accidents cause severe injuries and property damage, resulting in losses on the parts of the victims, their families, and entire communities. Accidents lead to heavy human suffering, such as deaths, cases of serious harm, and expenses on health that take more extended periods. Road crashes also cause extensive property damage, translating into tremendous financial problems for individuals, insurance companies, and the government. Road incident economics go further than initial medical costs (Fiorentini and Losa, 2020). Accidents lead to short-term disability/injuries that result in loss of productivity over the long term, higher insurance rates, and stress on the health care system. Such conditions are an economic strain placed upon people and the national economy. Socially, road accidents also go beyond the individuals that are directly affected. These affect family members and communities in that they cause emotional stress and, in turn, have prolonged physical health implications. These rippling effects affect the community, leading to poor welfare and low living standards. Developing appropriate intervention programs and prevention strategies relies largely on understanding these social effects.

Road accidents globally remain a big problem that needs more innovative responses. This study examined how machine learning and big data can be utilized to foretell national transport system road accidents using the UK as the case study. Analyzing and appreciating these contributing factors to accidents lead to developing specific preventative measures and enhancing road safety strategies to reduce the number of occurrences and the degree of injuries and attendant financial costs to humans and society. Legislation was an example

of a policy implemented in the UK, such as an amended Road Traffic Act requiring seat belts, motorcycle helmets, and child booster seats. With its subsequent Amendment, the Road Traffic Act 1988 establishes compliance rules (Gillespie, 2012). Offenders are also penalized according to these provisions. Further regulations specifically address vulnerable road users, including cyclists and pedestrians. Other measures to ensure the safe use of roads include speed limits, traffic signals, pedestrian crossings, and divided highways. This involved constant educational campaigns that encouraged safe driving habits. The police enforced violations of dangerous traffic offenses such as speeding, driving under the influence (DUI), and mobile phone use as a driver. In addition, technological developments introduced anti-lock braking systems (ABS), Electronic stability control (ESC), and collision avoidance systems into vehicles that helped prevent or reduce injury when an accident occurred. This resulted in a global decline in deaths caused by car accidents. There were still distractions, impaired behavior, aggression, or ordinary mistakes in human behaviors. Crashes were influenced by environmental factors such as bad weather, snowy and icy roads, and poor visibility (Gutierrez-Osorio and Pedraza, 2020). However, emerging revolutionary technologies such as vehicle-to-vehicle (V2V)communications, autonomous emergency braking, and advanced driver assistance systems are yet to enter the consumer market.

Big data has led to massive advancements in machine learning for road safety predictions that could transform road safety into an evidence-based practice informed by data-driven insights and predictive analytics. Artificial intelligence also includes one field called machine learning, in which algorithms are developed, through which computers can learn about patterns from vast amounts of data by themselves instead of being specially programmed. With increasing numbers of weather pattern sensors being included in roads and a large variety of information being recorded, such as accident rates on each road, congestion levels, driver's behavior, and vehicle maintenance records, the traffic data volume will be huge like never before. This research dissertation focuses on the transportation sector whereby big data is analyzed using various machine learning techniques such as Gradient Boosting Classifier, random forest, and XGB Classifier to find patterns, interactions, and predictive correlations. It is possible to detect the most significant risk-related factors in this situation leading to

such accidents. Therefore, relevant measures like modifying streets and highways, setting up speed limitation signs, warning drivers, etc., and standardizing vehicles may be taken. Analyzing such transportation systems, big data using machine learning algorithms made it possible to develop predictive models and thus estimate chances for accidents in forthcoming conditions (Sangare et al., 2021). These algorithms allowed picking the subtle mix of factors resulting in severe and slight collisions. One would, for instance, relate the previous crashes on a particular road by implementing algorithms such as Gradient Boosting and XG Boost while considering parameters such as traffic density, road geometry, number of casualties, longitudes, latitudes, time of day, weather events, and driver profiles (Ghandour et al., 2020). Machine learning combined with big data for improving road safety was demonstrated using exploratory analysis, identifying risk factors, and modeling. Through a series of supervised and unsupervised analyses, several different methods that utilized data-driven approaches for augmenting existing interventions to lower road casualties was tested using datasets obtained from various sources dealing with transportation.

The findings of these models shed light on areas such as the higher chances of accidents experienced by motorcyclists riding through rural highways during rainy morning rush hours, as well as other riders cruising on urban streets under precise conditions. For instance, there was an increased possibility of death in a crash if the fog is present along the winding mountainside highway over a clear, straight interstate (Ghandour et al., 2020). Predictions were also made in real time using new data incorporated within the calibrated models. For instance, an alert was sent in case of increased risk of multiple pileup accidents on a crowded motor highway, such as a warning if a driver traveled at excessive speed into a corner or when the sensor detects possible danger ahead (Formosa et al., 2020). Road accidents, especially fatalities, can be curtailed using big data and machine learning intelligence for deeper analysis, insights, and pre-emptive warnings. Preventive action on accident avoidance is much better than reacting after a crash occurs. Utilizing machine learning in transport data analytics leads to life-saving crash prevention systems.

The intended objective of this dissertation was to use machine learning methods such as Gradient Boosting, XG Boost, and random forest that can discover intricate relationships between movement data statistics and the possibility of traffic accidents (Fiorentini and Losa, 2020). The strengths of each algorithm suggested its suitability to be applied in this context. The preferred method for classifying accident versus no accident scenarios was based on Gradient Boosting. Random Forest algorithm handled missing values, identified the most likely causes, and accommodated nonlinear interactions among variables. Unlike many other approaches, the XGB Classifier was good at spotting small patterns in big datasets that contained many variables and lots of noise. The random forest approached an ensemble of many decision trees stabilizes forecasts while increasing accuracy (Gan et al., 2020).

Aggregated and anonymized transport information sourced from the UK STATS19 accident database, Road Safety Data portal, Highways England, and the Department for Transport was used to train and test these models. Some important input variables included weather conditions, the nature of roads used, vehicle types, demographic statistics, and prior accidents' severities available in historical records. Such data was utilized to instill the candidate models and choose the best predictive accuracy parameters (Gutierrez-Osorio and Pedraza, 2020). Evaluating a model's performance well against large volumes of transport data in diverse circumstances was essential. This ensured the predictions were valid for operational use, predictive alerts, and road safety improvement. Several algorithms were tried and optimized to obtain the best accuracy and precision, taking advantage of historical accident patterns in the data. A model would not go live until it demonstrated a reliable performance only once. The dissertation demonstrated the potentiality of employing Machine Learning Techniques for predicting and preventing road accidents. The algorithms found connections between traffic accidents and conditions connected with them, such as weather, traffic density, road geometry, vehicle type, demographics, etc, by training the models using historical transportation data. Tuning the models helped integrate them into real-time systems where the alerts could be predicted for improved road safety. Nevertheless, it was subjected to intensive vetting of considerable samples before its implementation for a strong output (Gan et al., 2020). Machine learning has enormous predictive capacities, but its practical

was carefully prepared and tested. This dissertation aimed to show a proof-of-concept for a machine learning-based way of avoiding accidents in cities, not creating a commercial product. Such models needed further work to be implemented, but the methods showed possible changes that intelligent machines could offer regarding rescue applications for transport.

1.2 Statement of the problem

Road safety has been a major concern because more and more accidents are having a big influence on society and the economy. Efforts to proactively improve road safety have been hampered by the absence of efficient predictive measures to foresee and prevent accidents. Creating a reliable machine learning model that both predicts and lessens the likelihood of accidents is the difficult part. The public, law enforcement, and transportation authorities are among the stakeholders who urgently want a focused solution that tackles the unique dynamics of accident-prone locations and significantly lowers the number of traffic accidents. The absence of efficient predictive methods to foresee and avoid traffic accidents is the issue being addressed. In order to significantly lower the number of traffic accidents, the emphasis is on creating a machine learning model that can forecast accidents with accuracy.

1.3 Research objectives

1.3.1 General objective

The main objective of this study is to develop a machine learning algorithm to predict accident severity.

1.3.2 Specific objectives

1. To develop and apply machine learning models within the project time-frame that predict accident severity using historical data and quantifiable metrics.

2. To apply an interactive geospatial tool that combines spatial analysis and machine learning for effective investigation of accident trends by interested parties.

3. To analyze temporal trends in accidents, considering day, time, and seasons, to enhance understanding of road safety dynamics within the project duration.

1.4 Justification of the study

This study was very important since it provided useful insights and workable answers for the urgent problem of road safety. Multiple stakeholders were served by the creation of an intelligent geographic visualization tool and the accurate prediction of accident severity. Above all, preemptive insights helped law enforcement organizations make smart resource allocation decisions and carry out focused actions. Urban planners were able to optimize road layouts for increased safety by using this technology to inform infrastructure improvements. Insurance companies also improved risk assessment models, which will result in more equitable premiums.

The dissertation intended to lower accident rates and promote safer roads; therefore, it had a practical impact on the broader public. Policymakers created evidence-based road safety plans to save lives and mitigate societal and economic costs by recognizing temporal patterns and influential expenses related to mishaps. In conclusion, the dissertation tackled a significant societal issue head-on by offering information and practical solutions to a wide range of stakeholders, including the public and law enforcement, to build more resilient and safer communities.

1.5 Significance of the study

The urgent concern of road safety, has been a crucial issue affecting society and the economy, highlighted the importance of the study. The growing incidence of collisions emphasized the necessity of effective preventative actions which enhanced traffic safety in a proactive manner. The goal of the project was to close this gap by creating a dependable machine learning model that could both anticipate and lessen the chance of accidents. By achieving its goals, the research made a significant contribution to a number of aspects related to road safety.

First off, the development of machine learning algorithms to forecast the severity of accidents based on past data and quantitative metrics was set to provide a proactive element to traffic safety initiatives. Through timely insights provided by this technique, stakeholders are better equipped to carry out interventions and preventive measures. Furthermore, the research endeavors to augment stakeholder involvement by the utilization of interactive mapping, a concise geospatial instrument that combined spatial analysis and machine learning. It envisaged that this technology would provide a user-friendly platform for stakeholders to effectively investigate and understand accident trends, including the general public, law enforcement, and transportation authorities.

Moreover, the examination of temporal patterns in accidents, integrated variables like day, time, and season, aims to enhance comprehension of the dynamics of road safety. With the help of this temporal analysis, decision-makers were able to make well-informed choices that resulted in focused interventions and policies that lowered the total number of traffic accidents. The study's findings ultimately had the potential to maximize resource allocation by focusing interventions on accident-prone areas and reducing the financial toll that accidents take on property damage, medical expenses, and lost productivity. The study's significance essentially stemmed from its ability to transform the way we think about road safety and create a more sustainable and safe transportation environment.

1.6 Research assumptions

Data availability and accuracy: It was assumed that the historical data, which was used to train machine learning models, was reliable and indicative of the various elements that contributed to traffic accidents. It is also assumed that the necessary information was accessible for the designated period of time, including accident reports and pertinent metrics.

Generalizability of the model: The machine learning models created in this work was able to generalize outside of the training set. It is anticipated that the models would accurately forecast the degree of accidents in a variety of settings and circumstances, even ones that aren't specifically mentioned in the historical data.

Folium tool efficiency: It was anticipated that the geospatial tool created with Folium, which combines machine learning and geographical analysis, was successful in giving stakeholders an easy-to-use platform to investigate and examine trends in accidents. It was anticipated that the tool would operate effectively and add to a thorough comprehension of the dynamics of road safety.

Representativeness of temporal pattern: It was anticipated that the temporal trends examined-taking into account the day, time, and seasons-would be indicative of more general trends in the dynamics of road safety. It was anticipated that the results would provide trustworthy insights into how accidents change throughout various time periods.

Engaging stakeholders: It was anticipated that users of the proposed geospatial tool, such as the general public, law enforcement, and transportation authorities, would interact with it actively. It was expected that the tool would help these stakeholders identify accident patterns and meet their information needs.

Effects of interventions: It was anticipated that the severity and frequency of traffic accidents would decrease when preventive actions based on the insights and predictive models from the GIS tool were put into practice. The research predicated on the hypothesis that interventions led to improvements in road safety.

Chapter 2

Literature Review

2.1 Introduction

Road safety is an urgent issue that has significant effects on the health of society and the stability of the economy. In addition to endangering lives, the rising number of accidents places a significant financial strain on local governments. The World Health Organization (WHO) reported that millions of people have been dying in traffic accidents each year, making it one of the world's top causes of death. The effects are not limited to the death toll; they also have widespread social repercussions that influence public health systems, worker productivity, and general quality of life.

Beyond the death toll, traffic accidents put a strain on economies through direct costs like lost production and higher insurance premiums, as well as indirect costs like medical bills and property damage. Proactive efforts to improve road safety have been hampered by insufficiently effective forecasting tools. This research dissertation explored the creation and evaluation of a trustworthy machine learning model in an effort to close this important gap. This model intended to meet the pressing needs of stakeholders, including the general public, law enforcement, and transportation authorities, by not only accurately predicting accidents but also considerably lowering their likelihood. By means of this investigation, the study aimed to offer a targeted remedy customized to the distinct characteristics of accident-prone areas, thereby reducing the effects of traffic accidents on society and economy.

2.1.1 Challenges of absent predictive measures for accidents

One of the biggest obstacles to road safety was the lack of effective accident prediction tools, which made it difficult to take preventative action to lessen the effects of traffic accidents. Road safety was a dynamic and complicated field, and traditional approaches that depended solely on historical data and rule-based systems had not been able to adequately capture this. Recent research, like those by Amiri et al. (2023) and Zhao et al. (2023), indicated that the models in use frequently failed to adjust to the dynamic nature of traffic patterns, meteorological conditions, and driver behavior.

Although data collection technology have advanced, one major obstacle still exists: road conditions' inherent unpredictability. According to Amiri et al. (2023), real-time fluctuations that lead to accident occurrences are not adequately taken into account by predictive models that rely solely on static data. Furthermore, Zhao et al. (2023) drew attention to the difficulty of combining various datasets into a coherent forecasting framework, such as those pertaining to weather, vehicle kinds, and road conditions.

These difficulties highlighted the urgent need for novel strategies, like machine learning, to get beyond the drawbacks of traditional prediction metrics. A sophisticated machine learning model has the potential to both predict accidents accurately and dynamically adapt to the many factors influencing road safety, as this research dissertation addressed. This offered a promising solution to the current problems in accident prediction.

2.1.2 The need for reliable accident forecasting models to enable targeted prevention

The ongoing high rate of traffic accidents highlighted the need for creative approaches to improve road safety. The European Commission for Mobility and Transport (ECMT) and

the National Highway Traffic Safety Administration (NHTSA) claimed that the present approaches to accident prediction and prevention are insufficient in that they did not offer precise and prompt solutions. Because they mostly depended on past data and rule-based systems, conventional methods found it difficult to adjust to the complex and dynamic character of road settings.

Recent research, such as that by Yamout (2022) and Guo et al. (2021), demonstrated the urgent need for a paradigm change toward more advanced prediction models, especially those based on machine learning. According to Yamout (2022), conventional models were unable to adapt quickly enough to include real-time data streams, which causes delays in the detection of new accident-prone situations. Guo et al. (2021) expounded upon the constraints of rule-based systems, highlighting their incapacity to identify intricate patterns amidst the extensive datasets currently at hand.

In order to solve this pressing issue, a dependable machine learning model specifically designed for accident prediction and likelihood reduction was developed and evaluated as part of this research project. Through the utilization of sophisticated algorithms and real-time data analytics, the suggested model aimed to surpass the constraints of current methodologies. To usher in a new era of proactive road safety interventions, the objective was to give stakeholders, including law enforcement and transportation authorities, a tool that not only accurately forecasts probable accidents but also significantly contributes to preventive efforts.

2.2 Theoretical literature review on road safety challenges and predictive measures

The issues surrounding road safety are complex and include a range of elements that have been affecting how frequently and how seriously accidents occur. Research by Andrey and

Mills (2003) indicated that the effects of accidents go beyond simple harm to people and property and affect many facets of society welfare. This review emphasized the seriousness of road safety concerns by examining the literature in order to provide a thorough knowledge.

Starting with foundational studies, McMillen et al. (2002) emphasized how accidents have repercussions that extended beyond the immediate physical damage, impacting social cohesiveness and mental health. As we looked more closely at the research, Bhavan (2019) highlighted the burden that accidents place on healthcare systems and the overall economy. The often-ignored hidden costs became important factors to take into account when assessing the complete impact of road safety issues.

Furthermore, Sangare et al. (2021) research from 2021 highlighted the unequal distribution of accidents among regions, emphasizing the geographical aspects of road safety concerns. Their research offered a geographical viewpoint and shows that certain regions have more difficulties than others because of things like poor infrastructure and high population density. The research on road safety that used spatial analysis enhanced our knowledge by emphasizing the necessity of targeted actions.

Theoretical literature unveiled a complex picture of road safety issues that combined spatial, economic, and human factors. Since the goal of this study project was to improve accident prediction metrics, a comprehensive grasp of these issues was necessary to create focused and efficient treatments.

2.2.1 Predictive measures and their limitations in road safety

A thorough review of the road safety literature illustrated the complex environment in which existing predictive methods operated as well as the difficulties associated with their implementation. Zhao et al. (2023) claimed that conventional predictive measures mostly depended

on past data and preset criteria. Even while these techniques were fundamental, they were not very flexible when it came to changing road conditions. Rule-based systems' static nature made it difficult for them to react quickly to changing driving habits and traffic patterns, which could cause delays in spotting new accident-prone situations.

This investigation began when Zhang et al. (2014) clarified established prediction metrics by highlighting their dependence on past data and predetermined rules. The study claimed that these models frequently run into problems when trying to adjust to the dynamic and ever-changing nature of road surroundings. Rule-based systems' static nature made it difficult for them to record fluctuations in traffic patterns and driving behaviors in real time, which caused delays in the identification of potentially dangerous situations as they emerged (Zhang et al., 2014).

Building on this groundwork, Abellán et al. (2013) stated that their study explored the difficulties posed by incorporating a variety of datasets into prediction models. The authors pointed out that although improvements in data collection technology have resulted in a flood of useful data, efficiently combining these datasets has been a challenging endeavor. The study addressed the challenges that came with integrating data from several sources, including weather, vehicle kinds, and road infrastructure. Together, these results highlighted the necessity of a paradigm change in predictive metrics. Our goal in this research dissertation was to use advanced machine learning algorithms—which illustrated the ability to dynamically adapt to the complexities of road safety dynamics—to solve these highlighted constraints as we move forward.

2.2.2 Gaps in road safety predictive measures

While progress had been made in understanding and creating predictive metrics, it is clear from exploring the topography of road safety literature that significant gaps still existed, demanding attention and creative solutions. A seminal study by Cai et al. (2022) found that

real-time data stream integration was frequently overlooked in favor of applying standard prediction models in existing literature. This left a big hole in the models' ability to predict accidents in a timely and accurate manner since they were unable to account for the dynamic nature of road settings.

Moreover, the current corpus of literature often homogenizes issues related to road safety without paying enough attention to the spatial details that influenced the likelihood of accidents. Ziakopoulos and Yannis (2020) highlighted the importance of spatial analysis in recognizing regional patterns and creating focused interventions. A significant gap in the literature is caused by the paucity of research that included geographical viewpoints, which made it difficult to gain a thorough understanding of accident dynamics at both the macro and micro levels.

Furthermore, even if machine learning has become more and more prominent in recent studies, a thorough analysis of how it integrated with predictive models in the context of road safety was still in its infancy. According to Xu et al. (2018), machine learning algorithms may be able to function in concert with conventional models in an as-yet-undiscovered area to provide a more resilient predictive framework. This gap highlighted the unrealized potential for cutting-edge approaches that could adjust to the dynamically changing complexity of road safety. By combining geospatial analysis, real-time data streams, and the advantages of machine learning over conventional models, this research dissertation tended to close these gaps. The goal was to close these gaps and make a substantial contribution to the development of predictive metrics for road safety that are both adaptable and effective.

2.3 Supervised machine learning models in predicting accident severity

Because the goal of this dissertation was to use supervised machine learning to transform road safety, it was critical to conduct a thorough investigation of the effectiveness of particular algorithms in forecasting the severity of accidents. Gan et al. (2020) claim that the Random Forest algorithm demonstrated extraordinary promise in this field. Because of its ensemble design, Random Forest provided robustness against over fitting and improved overall model accuracy by combining predictions from several decision trees. The adaptability of Random Forest to many input features, such as weather patterns and road configurations, provided a thorough comprehension of the factors that influenced accident severity. Visual aids like decision tree diagrams helped readers better understand Random Forest's complex decision-making process by capturing their attention.

To take things a bit further, there was interest in applying the XGBoost classifier to forecast accident severity. As stated by Delen et al. (2017), XGBoost gave a potent framework that demonstrated the significance of several features in assessing accident severity in addition to offering excellent predictive accuracy. Through the use of strategies like feature significance plots, which helped stakeholders pinpoint the main causes of catastrophic occurrences, the interpretability of this ensemble method was improved. These graphics, which frequently took the shape of bar charts or SHAP values, increased reader interest and gave readers a better comprehension of the underlying patterns that XGBoost discovered during the predictive modeling phase.

The literature demonstrated the effectiveness of gradient boosting techniques and decision trees by broadening the scope. Abellán et al. (2013) investigated Random forest and decision trees, which provide decision-making transparency by allowing stakeholders to understand the reasoning behind accident severity estimates. Gradient boosting, on the other hand, worked exceptionally well at enhancing predictive accuracy by iteratively strengthening weak

learners (Jamal et al., 2021). By comparing these supervised algorithms to one another, comparative performance graphs was incorporated to better understand the relative advantages and disadvantages of each method and help anticipate accident severity.

2.3.1 Exploring crucial parameters in accident severity prediction

Comprehending the dynamics of traffic accidents required a thorough investigation of multiple factors, each of which contributed differently to the total accident scene. Andrey and Mills (2003) indicated that weather conditions were critical components that impacted the severity of accidents. According to their research, inclement weather – such as rain and fog-strongly corresponded to a higher risk of serious accidents. Bad weather created complications that influenced driving behavior and accident outcomes by altering vision and road surface conditions. Such discoveries improved the accuracy of prediction models and enabled proactive measures to be taken in advance of adverse weather.

According to Lobo et al. (2019), a road's type had a significant impact on the severity of an accident. Their research showed that there are differences in the patterns of accidents on roads as opposed to residential or urban regions. Highways tend to have greater accident rates due to their larger traffic volumes and speeds, which highlighted the importance of road type as a crucial element in predictive models. Pie charts that illustrated accidents were distributed among various road types as an example of a graphic representation that could visually highlight the importance of this parameter, drawing the reader in which promoted deeper comprehension of its consequences on accident severity.

Li et al. (2021) stated that vehicle attributes are another component that warranted consideration. Their study highlighted how different vehicle types had differing effects on the severity of accidents. For example, because of their bulk and structural differences, accidents involving heavier vehicles, such trucks, had more severe repercussions. This parameter en-

hanced the predictive power of models by exploring the intricacies of different vehicle kinds, enabling customized actions depending on the unique cars in question. In order to provide a comprehensive understanding of the complex relationship between vehicle attributes and accident severity, visual aids such as bar graphs showed the frequency of accidents involving various vehicle kinds were used in addition to the textual presentation. The complex interplay between several factors that affect accident severity, such as weather, road kinds, and vehicle attributes was highlighted. To contribute to the creation of a predictive model that took these important factors into account in a comprehensive manner for improved road safety, this dissertation seek to combine and expand on these insights.

2.3.2 Strengths and weaknesses of existing models in the context of road safety

Strengths: The predictive models that are now in use for road safety have demonstrated excellent strengths, especially with regard to their ability to utilize previous data to provide intelligent forecasts. An analysis by Wu and Levinson (2021) found that ensemble models like Gradient Boosting and Random Forests have demonstrated a strong ability to capture intricate patterns in large data-sets. With its ability to handle a wide range of input features, including as weather, vehicle characteristics, and types of roads, these models provide a comprehensive understanding of the factors impacting the severity of accidents. Additionally, decision trees-a crucial part of ensemble models-allow stakeholders to understand the reasoning behind forecasts, which promotes well-informed decision-making. This is due to their interpret-ability and transparency. These models' strengths are that they can be applied to a variety of data-sets, which gives them a strong basis for predictive accuracy.

Weaknesses: Nevertheless, there are certain shortcomings to the current road safety models. One significant flaw, as noted by Fiorentini et al. (2023), is the possibility of over-fitting, especially with ensemble models. These models' complexities are useful for catching

subtleties, but they can also cause the training data to overemphasize noise, which limits the models' applicability to novel and untested settings. Furthermore, complicated ensemble models can be challenging to interpret, making it harder for non-experts to understand the underlying decision-making processes. Furthermore, current models frequently ignored the temporal component of accident data, which included changing traffic patterns and driver behaviors, which restricted the models' capacity to dynamically adjust to changing road settings. In order to improve predictive models, this study addressed these shortcomings.

2.4 Intelligent machine learning-integrated geo-spatial visualization tools

Within the ever-evolving realm of modern research, the nascent subject of machine learning and geographic visualization tools holds great promise for a variety of fields. The integration of machine learning algorithms with geospatial visualization tools has gone beyond traditional bounds, allowing a paradigm shift in the study and transmission of geographical data, according to recent studies by Yuan (2021). This synthesis made it possible to produce complex visualizations that feature predicted insights from machine learning models in addition to showcasing spatial patterns. Stakeholders may have discovered hidden trends in geographical data that traditional methods might miss by combining the power of predictive analytics with geospatial visualization in a seamless way. This study aimed to investigate this combination further, investigated new approaches to leverage the complementary strengths of geospatial visualization and machine learning for improved understanding of intricate spatial dynamics.

2.4.1 Combining advanced tools with folium library for interactive geographic visualization

Researchers have been using the Folium library and other state-of-the-art tools more and more as they embarked on the exciting path of interactive geographic visualization in order to overcome the constraints of traditional mapping methodologies. Recent research, like that of Rajamani and Iyer (2023), indicated that Folium is a particularly strong Python module that easily interfaces with data sources and enables researchers to produce dynamic, interactive maps. Its intuitive interface made it simple to integrate a variety of data-sets, which improved the visualization experience. Folium, an essential tool in the geospatial toolbox, has made it easier to overlay machine learning findings on maps and provided stakeholders with a user-friendly platform to explore and comprehend intricate spatial patterns. In addition to enhancing Folium's interactive features, the integration of tools such as D3.js and Mapbox allowed for the production of aesthetically appealing and highly informative geographic visualizations. The goal is to investigate the unrealized potential of these technologies and find new approaches to utilize features in order to create complex and captivating interactive geographic visualization applications.

2.4.2 Dissecting the road safety quilt

Within the dynamic field of road safety analysis, FOLIUM is a useful instrument that provided a distinct perspective for deciphering the complexities of collisions. The visualization landscape has undergone a revolution owing to the integration of FOLIUM into road safety studies, as stated by Rajamani and Iyer (2023). Because of its interactive mapping features, researchers have analyzed accident data on an easy-to-use platform and gain a thorough grasp of the contributing causes and spatial patterns. This study investigated the current uses of FOLIUM in accident analysis, determined how its dynamic mapping capabilities in conjunction with geospatial visualization enabled stakeholders to identify regions that are prone to accidents and provide practical insights. This study project aimed to explore the diverse contributions of FOLIUM, from its little-known uses to its unexplored potential.

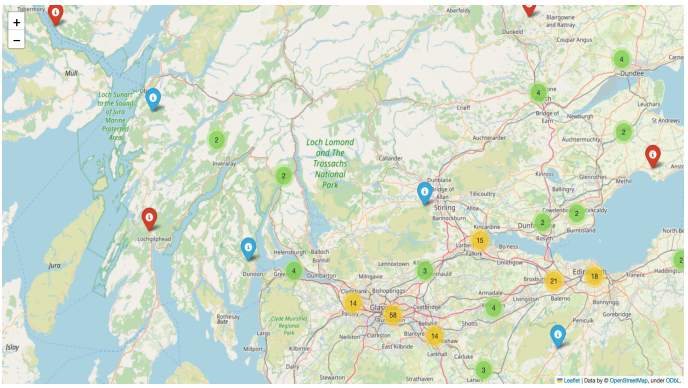


Figure 2.1: Illustration of folium mapping

2.4.3 Evaluating folium's potential for analysis of road safety

Benefits uncovered: Folium, a powerful Python interactive mapping package, revealed a wealth of advantages for road safety study. One of its greatest benefits, according to Jamal et al. (2021), is the easy integration of various data sources, which made it possible to create dynamic and eye-catching maps. Through unparalleled depth, stakeholders could examine spatial trends and accident data thanks to the real-time interactivity. Moreover, Folium made it easier to overlay machine learning forecasts on maps, which provided a comprehensive view of regions that are prone to accidents. Its ability to handle a wide range of data types made it an essential tool for academics trying to glean insights from intricate data-sets on road safety.

Overcoming obstacles: Nevertheless, there were certain obstacles to overcome when integrating Folium into analyses of traffic safety. One significant issue was that it could not handle massive data-sets, which could impair its performance on large accident databases, as Delen et al. (2017) suggest. The complexity of real-time interaction resulted in computing overhead, which would affect how responsive the tool was. Furthermore, thorough validation processes were necessary to guarantee the correctness of prediction models linked with Folium. Taking care of these issues was essential to maximizing Folium's potential in road safety analysis. The objective was to balance these advantages and difficulties to develop a

comprehensive understanding of Folium's contribution to the advancement of road safety research.

2.5 Temporal trends in accident incidents and road safety dynamics

Examining day-to-day dynamics: The temporal nuances of traffic safety accidents provided an engrossing story that exposed patterns that go beyond randomness. Abellán et al. (2013) claimed that the day of the week has a significant impact on the frequency of accidents. Their research showed that weekends frequently see an increase in accidents, which may be related to more recreational activities and maybe lax driving practices. On the other hand, weekdays showed different patterns with more accidents around cities, particularly during rush hour. Comprehending these daily fluctuations was essential to customizing therapies to certain temporal settings. This study explored the temporal aspects of road safety with the goal of synthesizing current knowledge and illuminating the subtleties influencing accident occurrences on various days of the week.

Handling the hours and seasons: Time is not only woven together by the weekly cycle, but also by the ebb and flow of the seasons and the passage of time. Theofilatos and Yannis (2014) provided insight into the effects of the seasons by highlighting how changes in the weather affect traffic conditions and, in turn, the frequency of accidents. Their study also explored the temporal aspects of accidents according to the time of day. They discovered that the risk environment changed during the day, following different patterns. The purpose of the study was to deepen and broaden this investigation by examining the ways in which the interaction of hours and seasons affects the temporal fluctuations in traffic accidents. Deciphering these patterns was essential to creating focused tactics that complement the dynamic temporal rhythms present in traffic safety.

2.5.1 Methodologies for analyzing road safety trends

Cracking the Temporal Mysteries: Examining the temporal patterns in traffic safety was a complex process requiring advanced techniques. Yamout (2022) stated that a popular method is time-series analysis, in which accident data from the past is examined to find recurrent patterns and trends. With the use of this statistically based approach, researchers identified trends in accident rates over time and gained understanding of how traffic safety changed over time. Furthermore, spatial-temporal analysis based on GIS has become more popular. According to Guido et al. (2022), this methodology combined temporal data with geographic information systems (GIS) providing a thorough investigation of how accidents occur in both place and time. Through the application of temporal patterns over geographic maps, researchers revealed complex correlations between characteristics specific to a certain period of time and the incidence of accidents. In order to improve our comprehension of temporal trends in road safety, this study aimed to expand upon these approaches by investigating novel directions and honing already-established methods.

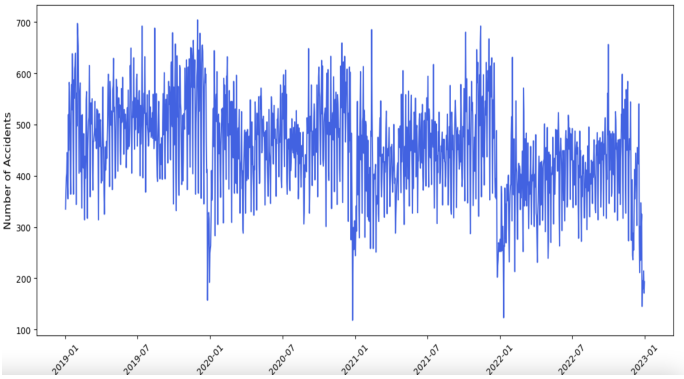


Figure 2.2: Time-Series Analysis of Accident over time

2.5.2 Evaluating past contributions and limitations

Deciphering past contributions: A wealth of important contributions have embellished the canvas of road safety research, offering priceless insights into the intricate dynamics at work. Guido et al. (2022) stated that previous studies made a substantial contribution to our comprehension of the complex variables affecting road safety. Research utilizing geographic

analysis had clarified the regional differences in accident frequencies, identifying hot spots and facilitating the application of focused treatments. Furthermore, as Zheng et al. (2014) pointed out, predictive modeling had provided a proactive lens that enabled stakeholders to foresee possible accident-prone locations based on past data. The foundation that the proposed research aimed to establish has been shaped by these contributions taken together. The research endeavors to enhance current approaches by incorporating and consolidating these insights, thereby exploring the nuances of road safety dynamics in terms of both space and time.

Overcoming obstacles to future development: Nevertheless, there are still certain limits that road safety research must overcome. Predictive model comprehension is a major difficulty, as Abellán et al. (2013) pointed out. The complicated interplay between socioeconomic characteristics, driving patterns, and contextual changes in road infrastructure necessitated careful analysis. Furthermore, despite much exploration, the temporal dynamics of road safety frequently fell short of providing a thorough picture. The goal of the suggested study was to overcome these restrictions by tackling the problems of temporal nuances and generalizability. Through the integration of advances in machine learning and spatial-temporal analysis, the project delved to transcend previous constraints by expanding our understanding of the dynamics of road safety and guiding the development of more efficacious preventive measures.

2.6 Identification of gaps in the current literature

An overview of the literature on road safety to date showed a situation characterized by both remarkable advancements and enduring difficulties. While research had significantly advanced our understanding of the variables affecting road safety, Saifuzzaman and Zheng (2014) pointed out that there is still a clear void in the thorough integration of spatial and temporal dynamics. Numerous publications that are now under publication primarily concentrated on spatial or temporal factors, which limits the comprehensive understanding

needed for proactive accident avoidance. Furthermore, limits are apparent in the extrapolation of results to various scenarios. Research frequently lacked the depth required to take socioeconomic circumstances, driver behavior, and variances in road infrastructure into consideration. The goal was to fill in these gaps and limitations by critically analyzing the literature.

The examined literature laid the groundwork for the planned research, but it was deficient in a few crucial areas. Overall, it aligned with the problem statement and research objectives. Hossain et al. (2019) stated that the constraints noted in the body of current literature were consistent with the gap in effective predictive measures that had been established. One enduring difficulty was the lack of comprehensive models capable of reliably predicting and preventing accidents in real-time. While there was research on predictive modeling, there hasn't been much done to incorporate machine learning into a dynamic, real-time framework. By creatively fusing machine learning with geographic visualization, this project delved to close these gaps and give stakeholders a powerful tool for identifying, comprehending, and reducing the hazards associated with road safety. This idea presented itself as a progressive step toward filling in the identified gaps and moving road safety research into new areas of accuracy and efficacy by critically analyzing the body of existing literature.

2.6.1 Bridging the literature gap

It is critical to close the gaps in the literature on road safety in order to propel the field into previously unexplored areas of accuracy and efficacy. Ang et al. (2022) claimed that the current drawback of static predictive models was addressed by the combination of machine learning and geographic visualization. This novel method shifted the paradigm in accident prediction and prevention by enabling real-time response to changing road conditions. The research quested to close this gap and advance road safety studies from retrospective analysis to proactive, predictive actions. A tool that could dynamically modify forecasts based on changing spatial and temporal elements, in addition to identifying accident-prone regions, would be beneficial to stakeholders such as law enforcement and transportation authorities. In addition, the suggested study was in accordance with the perspective given by Mekonnen

et al. (2022), who stressed the significance of context-specific methods in research on road safety. To bridge the gaps in the existing research, a technique that took into account the subtle differences in various locations, types of roads, and demographic characteristics must be developed. By giving stakeholders individualized information, this precision-oriented strategy hoped to enable more focused and successful responses. This project is important because it directly tackles the ongoing issues that societies dealing with the social and economic effects of traffic accidents face. Its importance goes beyond scholarly curiosity. Essentially, the research delved to provide useful recommendations that could change the road safety landscape in addition to making a contribution to the scholarly conversation.

2.7 Literature summary

Important discoveries from the careful analysis of the literature on road safety have highlighted the need for a paradigm change in predictive modeling. Theofilatos and Yannis (2014) have observed that although previous research provided useful understanding of the temporal and spatial aspects affecting road safety, there was still a discernible lack of integration of dynamic, real-time predictive models. Traditional methods were not as applicable to the changing road conditions because of their static character. By presenting a novel combination of machine learning and geographic visualization, the study quested to close this gap and promised to move beyond reactive accident analysis to proactive accident avoidance. The potential for this project to completely change how stakeholders view and handle road safety issues emphasizes how important it is. This study attempts to skillfully weave together a narrative that not only captures the essence of our academic pursuits but also heralds in a new era in the quest for safer roads as we traverse from the familiar realms of current research findings to uncharted territories of methodology and results.

Chapter 3

Methodology

3.1 Introduction

The primary goal of the research project was to analyze accident severity by using machine learning techniques applied to a large dataset downloaded from Kaggle. The dataset offered a wealth of information for comprehending the variables determining accident severity because it included a wide range of parameters linked to traffic accidents.

3.2 Overview of the dataset

The dataset was made up of records from different accidents, each identified by an index or unique identifier. Important details about the accidents included their severity, date, location (latitude and longitude), illumination, district area, number of victims, number of cars involved, kind of road, road surface, urban/rural classification, weather, and kind of vehicle. The 660,679 cases and 14 unique features in the dataset provided a solid basis for examining trends and patterns in traffic safety. Detailed data preprocessing was the first step, where redundant features were eliminated and missing values were addressed. This created a strong basis for the supervised machine learning algorithms that followed. Using information from a dataset with more than 660,000 entries, these algorithms-which included ensemble approaches and decision trees-went through a rigorous training and assessment process.

By integrating the Folium library, the approach expanded its use beyond the realm of prediction to the geographical domain (Ajagbe et al., 2020). The development of an intelligent geographic visualization tool that dynamically interfaced with machine learning insights was

made easier by this connection. This platform, which was accessible by design, provided stakeholders with an instantaneous, interactive platform for easily exploring patterns in traffic safety.

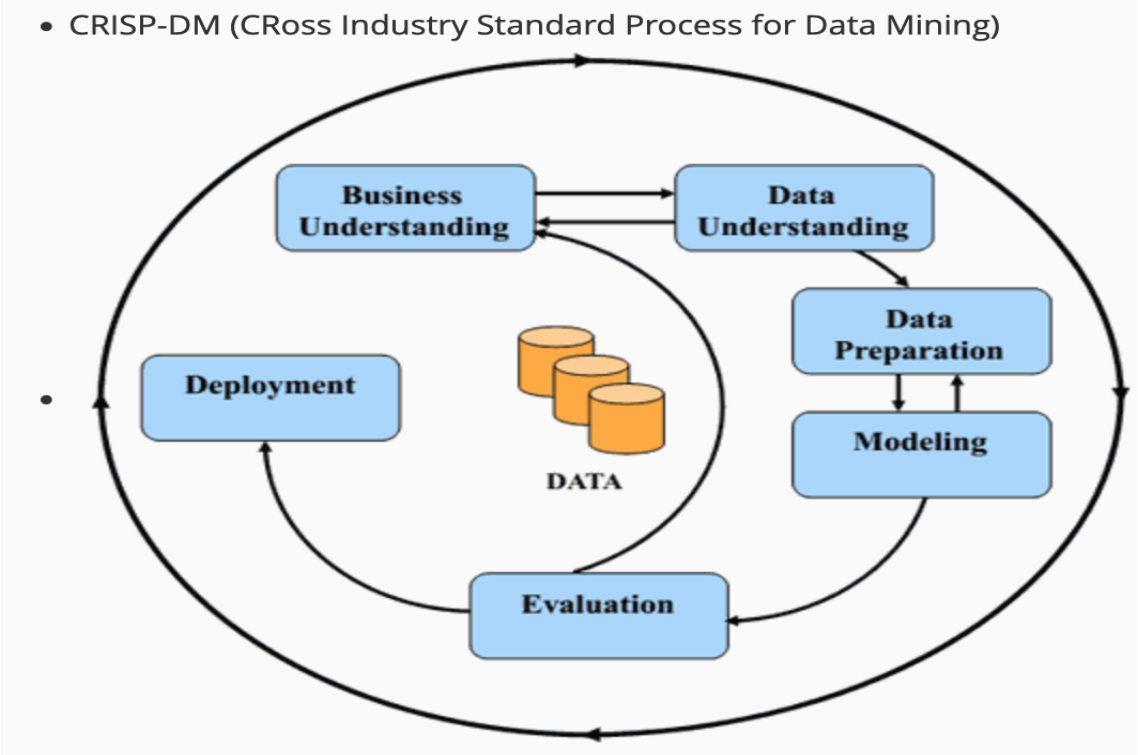


Figure 3.1: Steps to follow in Machine learning

3.3 Data collection

The data were acquired from Kaggle, a reputable platform known for hosting datasets pertinent to machine learning and data science challenges. The data was downloaded and stored in csv format. The data was then exported to Jupyter notebook for exploratory analysis and machine learning modelling. From the data, the outcome variable is accident severity which represent how accidents were classified into classes such as serious, slight and fatal. The predictor variables (covariates) are given in Table 3.1

Accident severity: representing how accidents were classified into classes like serious, slight, and fatal. The predictor variables (covariates) are given in Table 3.1

Table 3.1: Description and definition of variables

Variable/Covariate	**Description**
Accident Date	Date of the accident, used to identify temporal patterns and correlations influencing accident severity.
Number of Casualties	Total number of people injured or killed in the accident.
Longitude and Latitude	Geographical coordinates of the accident location.
Weather Conditions	Weather at the time of the accident (e.g., clear, raining, foggy).
Road Surface Conditions	Condition of the road surface during the accident (e.g., ice, wet, dry).
Road Type	Type of road where the accident occurred (e.g., highway, residential, main road).
Urban or Rural Area	Classification of the accident location as urban or rural.
Number of Vehicles	Total number of vehicles involved in the accident.

3.4 Data-preprocessing

3.4.1 Feature inspection

To grasp the breadth of information covered, specific factors were explored, including vehicle type. Various vehicle categories were delineated using vehicle type and its unique characteristics. This approach facilitated an understanding of the diverse range of vehicles implicated in accidents, spanning from standard cars to buses, bicycles, and agricultural vehicles.

3.4.2 Managing missing values

Mechanism of missing values: MICE successfully imputed missing values in the dataset. The mechanism employed by MICE involved performing multiple imputations using predictive mean matching (PMM) method with 5 imputations, maximum 5 iterations, and a random seed of 123. This method imputed missing values by drawing from the conditional distribution of each variable given the observed data. The imputed values for GPS coordinates contributed to a more comprehensive map of the study area, enhancing the spatial coverage and increasing the number of locations available for analysis. The following results were observed after this mechanism was employed

3.4.3 Adjustment of class imbalance in accident severity column

In the methodology employed for this dissertation, the 'Fatal' class within the 'Accident_Severity' column was combined with the 'Serious' category due to the former's insignificance in the dataset. The 'Fatal' class accounted for only 1.1% of the total cases, a proportion that is widely regarded as insufficient to contribute meaningfully to the model's predictive power. Combining classes that represent less than 5% of the dataset is a recognized practice in machine learning, particularly in the context of imbalanced datasets (Rubaidi et al., 2022).

This adjustment was necessary to address the issue of sparsity, which could lead to model instability, overfitting, and poor generalization. By merging the 'Fatal' and 'Serious' categories, the project enhanced the model's robustness and reliability, allowing for more meaningful distinctions without the noise or irregularities introduced by an underrepresented class. Additionally, this approach mitigated the potential bias towards the majority class, ensuring a more balanced and accurate prediction of severe accidents.

According to Chamseddine et al. (2022), the practice of combining underrepresented classes is supported by studies that have demonstrated improvements in model performance when classes representing less than 5% of the data are merged. This technique proved advantageous

in enhancing predictive accuracy and generalizability, both critical for applications in road safety where the accurate prediction of severe accidents is paramount. Consequently, this methodological adjustment was integral to improving the efficacy of the model in predicting significant accident events, thereby contributing to the overarching goal of enhancing road safety through data-driven insights.

3.5 Feature engineering

Feature engineering involved comprehensive examination and manipulation of key elements such as weather conditions, road type, and vehicle type to uncover hidden patterns and correlations. This process included generating new features and altering existing ones, such as extracting temporal aspects from accident dates and encoding categorical variables for compatibility with machine learning methods. Scaling and normalization techniques were employed to mitigate biases during model training, ensuring numerical features were brought to a common scale. Setting feature engineering as a top priority made it easier to create a carefully customized dataset that was suitable for predictive modeling, which in turn made it possible to create accurate and perceptive machine learning models. A crucial step in the machine learning process, feature selection, was handled carefully. Categorical variables with multiple unique values were transformed into a format appropriate for model training using label encoding. The scikit-learn module's LabelEncoder function was used to systematically identify and modify the categorical columns that were intended for label encoding. The methodical methodology adopted in this study enhanced the dataset's prediction capacities and supported data-driven decision-making in road safety initiatives by optimizing the numerical representation of categorical variables for machine learning algorithms.

3.6 Model training

The accident severity prediction dataset was split into training and testing sets to provide a strong model evaluation. The split ratio of 80% training and 20% testing was used to ensure

that there was enough data for model training and still a significant amount for evaluation. The Gradient Boosting Classifier, XGB Classifier, and Random Forest algorithms were selected for the model training process because they each have specific benefits in managing intricate datasets and forecasting the severity of accidents.

3.6.1 Random forest algorithm

To forecast accident severity (Y), the Random Forest model made use of an ensemble of decision trees in the research project. Every tree in the forest was built using a bootstrap sample, or sample taken with replacement, from the training set. When a node was split during the tree-building process, the best split was selected at random from a subset of features rather than taking into account every feature. By reducing the connection between trees, this method lowered the variance of the forest and improved overall forecast accuracy. The model is given in equation (3.1).

$$Y = RF(X) + e \tag{3.1}$$

- Y: The Accident Severity, which was the target variable aimed to predict in the research project.
- X: Represented the matrix of attributes/features in the dataset in the research project.
- RF: Denoted the Random Forest algorithm, which was a collection of decision trees, in the research project.
- e: Was the error term, capturing the randomness and the model's inability to explain the variability fully, in the research project.

3.6.2 Gradient boosting classifier

The Gradient Boosting Classifier was employed to model the probability that an accident had a particular level of severity, specifically focusing on the two-class outcome after merging the 'Fatal' and 'Serious' categories. Gradient Boosting is an ensemble learning method that builds a predictive model through an iterative process, where weak learners, typically decision trees, are sequentially added to correct the errors of the preceding models (Bentéjac et al., 2021). This approach was particularly effective in handling the imbalanced classes present in the research project. The model is given in equation (3.2)

$$P(Y = k|X) = \frac{e^{F_k(X)}}{\sum_{j=1}^{K} e^{F_j(X)}} \tag{3.2}$$

where:

- Y: The outcome representing Accident Severity in the research project, with $k = 1$ representing 'Slight' and $k = 2$ representing the merged 'serious' and 'fatal' class as 'Serious'.
- X: Represents the attributes/features in the dataset.
- $F_k(X)$: The model's output score for class k, which was computed as the sum of the predictions from the sequence of decision trees.
- K: The total number of classes, which is 2 in this case.
- e: The base of the natural logarithm.

The model iteratively minimized the loss function by adding trees that correct the residual errors of the previous models. This method allowed improvement of the model's performance, particularly in distinguishing between the 'Slight' and 'Serious' accidents, thereby enhancing the predictive accuracy for road safety analysis.

3.6.3 XGBoost Classifier

Using the two-class outcome, the XGBoost Classifier was used to model the likelihood that an accident had a specific degree of severity. Extreme Gradient Boosting, according to Amjad et al. (2022), is a sophisticated and scalable implementation of the Gradient Boosting framework. It minimizes the loss function by using a gradient descent technique with decision trees as base learners (Ramraj et al., 2016). The equation (3.3) provides the model.

$$P(Y = k|X) = \frac{e^{F_k(X)}}{\sum_{j=1}^{K} e^{F_j(X)}} \tag{3.3}$$

where:

- Y: The outcome representing Accident Severity in the research project, with $k = 1$ representing 'Slight' and $k = 2$ representing the merged 'Serious' and 'Fatal' class.
- X: Represents the attributes/features in the dataset in the research project.
- $F_k(X)$: The score for class k, computed as the sum of the outputs from the ensemble of decision trees.
- K: The total number of classes, which is 2 in this case.
- e: The base of the natural logarithm.

By employing regularization approaches to prevent overfitting, XGBoost improved the model and made it capable of managing the dataset's imbalanced classes. The project improved the performance of the road safety analysis by utilizing XGBoost's efficiency and accuracy to distinguish between the 'Slight' and 'Serious' categories.

3.6.4 Model evaluation

Accuracy assessment: Accuracy involves quantifying the overall correctness of the model's perfomance and is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

.

Thorough performance analysis: A confusion matrix was used in the research endeavor to give a thorough performance analysis across different accident severity categories.

Table 3.2: Confusion Matrix

		Predicted Class	
	Class 1	Class 2	Class 3
Actual Class			
Class 1	**True Positive**	a_{12}	a_{13}
Class 2	a_{21}	**True Positive**	a_{23}
Class 3	a_{31}	a_{32}	**True Positive**

The confusion matrix is given in Table 3.2

Handling Imbalanced Dataset: Precision (P), recall (R), and F1-score ($F1$) was employed to rectify imbalances in the dataset. These metrics are defined in equation (3.4):

$$P = \frac{TP}{TP+FP}, \; R = \frac{TP}{TP+FN}, \; F1 = \frac{2 \times P \times R}{P+R}, \tag{3.4}$$

These measures guaranteed a fair evaluation of the study endeavor by evaluating the model's capacity to recognize situations with varying degrees of severity.

Discriminatory power assessment: The area under the ROC curve (AUC-ROC) was used to assess the discriminatory capacity of the model. This provided a thorough understanding of the model's ability to distinguish between different severity classifications. By utilizing these many metrics, the study aimed to provide comprehensive understanding of the Random Forest model's functioning.

3.7 Model selection and optimization

Several machine learning models were carefully compared during the selection process, with a focus on giving priority to those that closely matched the distinctive features of accident severity prediction. The model's capacity to identify different degrees of severity was taken into account in addition to performance criteria like accuracy, to make sure that it was in line with the overall objective of improving road safety. The study fine-tuned the hyperparameters to increase the predictive power of the models. Hyperparameters were adjusted using random method and grid search.

3.8 Deployment and spatial visualization

This involves spatial visualization to effectively transform the models' prediction insights into practical steps to enhance road safety. The visualization improved accessibility to model insights and enabled law enforcement and decision-makers to fully recognize and understand road safety patterns.

Chapter 4

Results and interpretation

4.1 Introduction

This chapter presents the data analysis process, key findings, and interpretations of the results obtained from the study. The chapter begins by presenting descriptive statistics to offer an overview of the data's main characteristics, followed by a detailed examination of key variables and their relationships. Subsequent sections focus on the application of machine learning algorithms used to derive insights from the data.

4.2 Accident severity distribution in UK

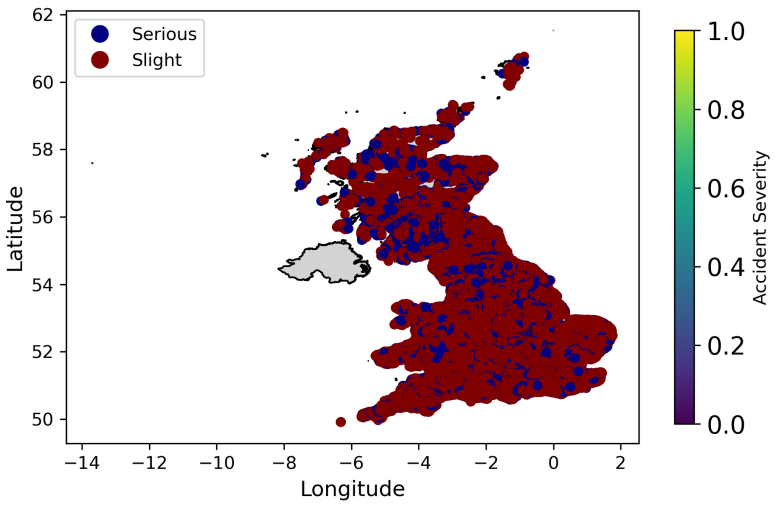


Figure 4.1: Accident severity distribution in the UK

Figure 4.1 illustrated how accident severity occurred within the geographical region of UK. It is clear that there were geographically concentrated areas with higher accident severity in the United Kingdom. The map illustrated the distribution of accidents, showing clear clustering patterns for minor accidents (red) and serious accidents (blue). In urbanized areas, especially in the south of the UK, around London, Birmingham, and Manchester, where traffic volume and population density are usually higher, accidents were concentrated. The proportion of blue dots indicated the relative frequency of serious incidents in the central and northern regions. This could be the result of a number of variables, including the state of the environment, traffic patterns, and road infrastructure.

On the other hand, there were less accidents reported in rural areas in Scotland, Wales, and the northern half of the UK, despite the fact that serious accidents seemed to occur more frequently in these sparserly populated areas. Existing research on rural accident severity patterns suggested that this could be caused by slower emergency responses in remote areas, increased speeds on rural roads, and fewer safety precautions (Theofilatos and Yannis, 2014). The distribution was consistent with the theory that less crowded places with fewer but more hazardous incidents tend to have more serious accidents, whereas metropolitan areas recorded a larger frequency of accidents, most of which were mild in nature. The underlying reasons for these regional variations should be further studied in order to generate customized policy measures meant to increase road safety across various

Strong significance was shown by a p-value of less than ($p > 0.001$), which implied that the observed relationship or difference is statistically significant. From Table 4.1 indicated that with a p-value less than 0.001, the light condition variable proved to be statistically significant to accident severity. The road surface conditions demonstrated a statistical significance of ($p > 0.001$). In terms of road type, they did portray statistical significance with the predictor variable of ($p < 0.001$), The weather showed some effect, however weather conditions such as wet, stormy, and snowy weather were statistically significant to the accident severity predictor column with ($p < 0.001$). P-values above 0.05 showed that the vehicle type did not significantly affect the severity of the collision. When comparing the severity of accidents,

Table 4.1: Factors influencing accident severity

Dependent: Severity	**Accident**		**Serious**	**Slight**	**Total**	**p**
Total N (%)		88217 (13.4)	563801 (85.3)	660679		
light conditions	Darkness_lights_off	7968 (9.0)	34273 (6.1)	43921 (6.6)		<0.001
	Darkness_lights_on	19490 (22.1)	110483 (19.6)	131878 (20.0)		
	Daylight	60759 (68.9)	419045 (74.3)	484880 (73.4)		
road surface	Dry	61708 (70.0)	381049 (67.6)	448547 (67.9)		<0.001
	Snow	2572 (2.9)	21607 (3.8)	24407 (3.7)		
	Wet/flood	23937 (27.1)	161145 (28.6)	187725 (28.4)		
road type	Dual carriageway	11746 (13.3)	85863 (15.2)	99424 (15.0)		<0.001
	Roundabout	3665 (4.2)	40185 (7.1)	43992 (6.7)		
	Single carriageway	70540 (80.0)	419563 (74.4)	496663 (75.2)		
	Slip road	2266 (2.6)	18190 (3.2)	20600 (3.1)		
weather	Rainy	10729 (12.2)	77589 (13.8)	89311 (13.5)		<0.001
	Snowy	674 (0.8)	6410 (1.1)	7123 (1.1)		
	Stormy	2284 (2.6)	18147 (3.2)	20678 (3.1)		
	Sunny	74530 (84.5)	461655 (81.9)	543567 (82.3)		
vehicle type	Heavy_vehicles	6252 (7.1)	40265 (7.1)	47108 (7.1)		0.526
	Light_vehicles	72786 (82.5)	465461 (82.6)	545446 (82.6)		
	Motorbikes/bicycle	9179 (10.4)	58075 (10.3)	68125 (10.3)		
urban or rural	Rural	37313 (42.3)	196087 (34.8)	239001 (36.2)		<0.001
	Urban	50904 (57.7)	367714 (65.2)	421678 (63.8)		

urban areas and rural areas showed a statistically significant correlation ($p < 0.001$). These results revealed that the severity of accidents was greatly influenced by characteristics such light conditions, road types, surfaces, weather, and urban/rural locations.

4.3 Model development and evaluation

4.3.1 Random Forest Analysis

The Random Forest model's accuracy was 0.7467, indicating that approximately 74.67% of the predictions made by the model were correct. The confusion matrix in Table 4.2 presents the distribution of predictions.

Table 4.2: Confusion Matrix: Random Forest Model

		Predicted	
		Severe	**Slight**
Actual	**Severe**	5280	14096
	Slight	19374	93386

Table 4.2 presents the distribution of predictions. The confusion matrix revealed that the model correctly classified 52,800 severe accidents and 93,386 slight accidents. However, it misclassified 14,096 severe accidents as slight and 19,374 slight accidents as severe. The Random Forest model demonstrated reasonable accuracy, particularly in predicting slight accidents (93,386 correctly classified) compared to severe ones (52,800 correctly classified). The slightly lower accuracy in severe accident prediction might be attributed to class imbalance, where the number of slight accidents substantially exceeded that of severe accidents. The model's robustness, indicated by its performance across multiple decision trees, provided a reliable prediction mechanism, albeit with some trade-offs in precision for less frequent classes. The Random Forest model effectively captured the complexity of the data, balancing bias and variance while achieving commendable performance in accident severity prediction. Further tuning and class rebalancing might improve its predictive accuracy, particularly for severe accidents.

4.3.2 XGBoost Analysis

Table 4.3: Classification Report: XGBoost Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0 (Slight)	0.26	0.19	0.22	19376
1 (Severe)	0.87	0.91	0.89	112760
Accuracy		0.80		
Macro Avg	0.56	0.55	0.55	132136
Weighted Avg	0.78	0.80	0.79	132136

The XGBoost model achieved an accuracy of 0.8023, indicating that 80.23% of the predictions were correct. The detailed classification metrics and the confusion matrix are presented in Table 4.3 and Table 4.4 respectively. The confusion matrix indicated that the model

Table 4.4: Confusion Matrix: XGBoost Model

		Predicted	
		Severe	**Slight**
Actual	**Severe**	3651	15725
	Slight	10397	102363

correctly identified 3,651 slight accidents and 102,363 severe accidents. However, 15,725 severe accidents were misclassified as slight, while 10,397 slight accidents were misclassified as severe. The XGBoost model demonstrated a strong predictive capability for severe accidents, with a precision of 0.87 and recall of 0.91. The F1-score of 0.89 for severe accidents highlighted the model's effectiveness in balancing precision and recall. The slightly lower performance for slight accidents, with a precision of 0.26 and recall of 0.19, was likely due to the class imbalance, where severe accidents were more prevalent in the dataset. The weighted average F1-score of 0.79 reflected the overall performance, suggesting that the model was better suited for predicting severe accidents, which was critical for the task. The XGBoost model provided robust predictions, particularly for severe accidents, making it a valuable tool for accident severity prediction. The application of regularization techniques and gradient boosting contributed to its high accuracy and strong performance metrics.

4.3.3 Gradient Boosting Classifier

The Gradient Boosting Classifier was utilized to predict accident severity due to its strength in combining weak learners to form a robust predictive model. This approach is well-suited for capturing complex, non-linear relationships in the data, making it an ideal choice for the task at hand.

Model Tuning and Validation

- **Hyper-parameter tuning:** The hyper-parameters were optimized by assigning class weights to account for the class imbalance in the data, with weights of 0: 2.5, 1: 1.0.

This adjustment was crucial for improving the model's performance, particularly in predicting the minority class (less severe accidents).

- **Cross-validation:** The model's generalization capability was assessed using K-fold cross-validation, ensuring that the evaluation metrics were reliable and not overly optimistic.

Model Evaluation

- **Performance metrics:** The model's predictive accuracy was evaluated using metrics such as accuracy, precision, recall, and the F1-score. The confusion matrix was also analyzed to understand the distribution of correctly and incorrectly classified instances.
- **Baseline Comparison:** The performance of the Gradient Boosting model was compared against other models. The tuning process led to an increase in overall accuracy and significant improvements in precision and recall for the minority class, as shown in Table 4.5.

Table 4.5: Optimized Gradient Boosting Model Performance

Metric	Value
Accuracy	0.8449
Precision (Class 0)	0.38
Recall (Class 0)	0.09
F1-Score (Class 0)	0.15
Precision (Class 1)	0.86
Recall (Class 1)	0.97
F1-Score (Class 1)	0.91

The Gradient Boosting model demonstrated improved predictive accuracy, particularly after the hyper-parameter tuning process. The model achieved an accuracy of 0.8449, with a significant enhancement in the recall of the minority class (less severe accidents), highlighting the effectiveness of the class weighting strategy. These results underscore the model's ability to generalize well across different classes, making it a reliable choice for predicting accident severity.

4.4 Model Comparison

Table 4.6: Model Comparison

Metric	XGBoost Classifier	Gradient Boosting Classifier
Accuracy	0.8023	0.8449
Precision (Class 0)	0.26	0.38
Recall (Class 0)	0.19	0.09
F1-Score (Class 0)	0.22	0.15
Precision (Class 1)	0.87	0.86
Recall (Class 1)	0.91	0.97
F1-Score (Class 1)	0.89	0.91

The comparison in Table 4.6 demonstrates that the Gradient Boosting model outperformed the XGBoost model in terms of overall accuracy and recall for the majority class (Class 1). The optimized Gradient Boosting model achieved a higher accuracy of 0.8449 compared to 0.8023 for XGBoost. Additionally, the recall for the minority class (Class 0) improved significantly after tuning, highlighting the effectiveness of class weighting in the Gradient Boosting model. The decision tree visualized in the provided diagram represents one of the ensemble trees from the Gradient Boosting model. This model was trained to predict the target variable based on the features in the dataset. The structure of the tree offers insights into the significance of different features and their interactions in making predictions. It reveals that the Number of Vehicles involved in an incident is the most crucial predictor in this model, forming the root node of the tree. This indicates that the number of vehicles plays a primary role in distinguishing between different outcomes in the dataset. Specifically, the model suggests that incidents involving fewer than 1.5 vehicles tend to result in a different set of outcomes compared to incidents with more vehicles.

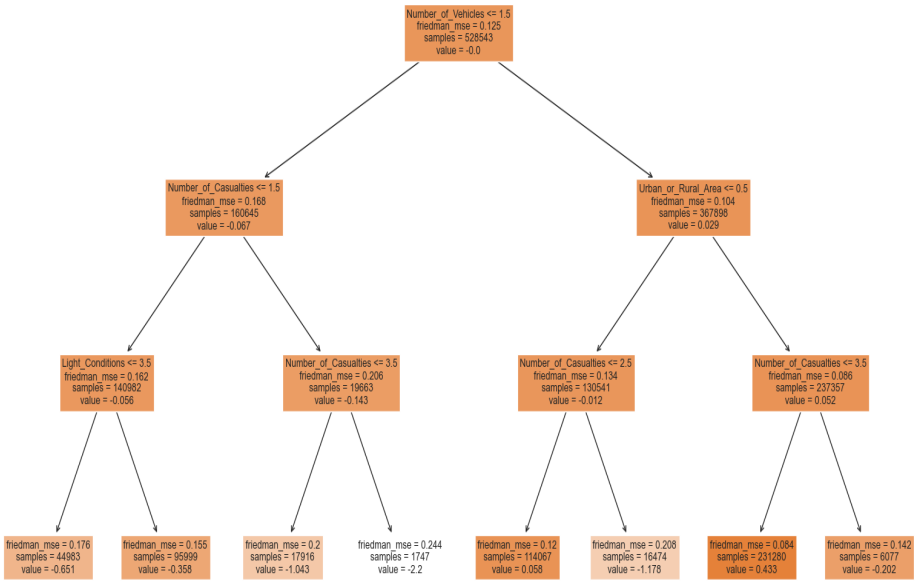


Figure 4.2: Feature Importance of GBoosting

Figure 4.2 illustrated the importance of the features used by the model. The decision tree revealed that the Number of Vehicles involved in an incident was the most crucial predictor in this model, forming the root node of the tree. This indicated that the number of vehicles played a primary role in distinguishing between different outcomes in the dataset. Specifically, the model suggested that incidents involving fewer than 1.5 vehicles tended to result in a different set of outcomes compared to incidents with more vehicles.

- **Number of Casualties:** The subsequent splits in the tree further emphasized the importance of the number of casualties. For instances where the number of vehicles was fewer than 1.5, the model split based on whether the number of casualties was less than or equal to 1.5. This split significantly contributed to the model's ability to

differentiate between outcomes, highlighting the impact of the number of casualties on the target variable.

- **Urban or Rural Area:** The tree indicated that whether the incident occurred in an urban or rural area was another important factor, particularly when the number of vehicles was greater than 1.5. The model showed that incidents in rural areas (denoted by values less than or equal to 0.5) led to a distinct set of outcomes compared to urban incidents. This suggested that the location of the incident affected the likelihood of certain outcomes.

- **Light Conditions:** The influence of light conditions was also significant, particularly when the number of casualties was fewer than or equal to 3.5. The model used this feature to further refine its predictions, suggesting that poor light conditions (denoted by values less than or equal to 3.5) contributed to different outcomes.

The decision tree's splits revealed the hierarchical importance of features in predicting the target variable. The model's emphasis on the Number of Vehicles and Number of Casualties as primary predictors aligned with logical expectations, as these factors were directly related to the severity and likelihood of certain outcomes.

Compared to other models that might have considered multiple features simultaneously, the decision tree approach provided a clear, interpretable structure that ranked the importance of features in a sequential manner. This characteristic was particularly useful in understanding the complex interactions between variables in predicting outcomes.

However, the decision tree also had limitations in that it might have oversimplified relationships by focusing on sequential splits rather than capturing potential interactions between multiple variables simultaneously, as methods like SHAP did.

We benchmarked the performance of our models, including Random Forest, XGBoost, and a tuned Gradient Boosting model with class weights, to evaluate their effectiveness in predicting accident severity.

Table 4.7: Comparison of key performance metrics

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)
Random Forest	0.7467	0.27	0.87	0.27	0.83	0.27	0.85
XGBoost	0.8023	0.26	0.87	0.19	0.91	0.22	0.89
Gradient Boosting (Unadjusted)	0.8137	0.62	0.85	0.01	1.00	0.01	0.92
Gradient Boosting (Class Weights)	0.8449	0.38	0.86	0.09	0.97	0.15	0.91

Table 4.7 presents a comparison of key performance metrics, including accuracy, precision, recall, and F1-score, for each model. The analysis revealed several insights. The Gradient Boosting model with class weights achieved the highest accuracy of 0.8449, indicating superior performance compared to the other models. This model demonstrated improved precision for class 0 (0.38) and class 1 (0.86) compared to both XGBoost and Random Forest. Additionally, it showed a notably high recall for class 1 (0.97), making it effective in identifying severe accidents. However, recall for class 0 was relatively low (0.09).

XGBoost achieved an accuracy of 0.8023 and exhibited high recall for class 1 (0.91), suggesting good performance in identifying severe accidents. Despite this, the precision for class 0 was low (0.26), indicating challenges in correctly classifying less severe accidents.

Random Forest attained an accuracy of 0.7467. Although it demonstrated high precision for class 1 (0.87), the recall for class 0 (0.27) and precision for class 0 (0.27) were relatively low.

Overall, the Gradient Boosting model with class weights adjustment displayed the best balance between accuracy, precision, and recall, especially for the severe class. This balance proved crucial for effective accident severity prediction.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

An extensive examination of the results obtained by comparing the effectiveness of different machine learning models used to forecast the severity of accidents is presented in this chapter. The best performing model was the Gradient Boosting model with class weights, which showed excellent accuracy and recall for serious accidents and indicated its potential for use in road safety applications. The chapter goes on to address how these findings could affect accident prevention tactics and offers suggestions for improving prediction models. The model's actual use and its effects on raising traffic safety and lowering accident-related fatalities are the main topics of the concluding remarks.

5.2 Discussion

This study's main goal was to create a machine learning model that could forecast the severity of accidents and pinpoint the critical variables affecting the results of traffic accidents. From the study comparison analysis revealed that, out of all the classifiers that were assessed, including Random Forest and XGBoost, the Gradient Boosting model with class weights performed better. The model's efficacy in predicting accident severity was highlighted by its 0.8449 accuracy, which was accompanied by enhanced precision and recall for both severe and less severe accident classes. The results are consistent with other studies, highlighting

the effectiveness of ensemble approaches in managing class imbalance and improving classification task predicting performance (Yamout, 2022).

In important performance criteria, the Gradient Boosting Classifier—optimized with class weights—performed better than the XGBoost and Random Forest models. It performed better overall than Random Forest (0.7467) and XGBoost (0.8023), achieving an accuracy of 0.8449. Gradient Boosting produced an especially noteworthy increased recollection for severe accidents (Class 1) with a score of 0.97, demonstrating its efficacy in identifying serious accidents—a critical component for focused interventions. On the other hand, the XGBoost model performed poorly in classifying less severe accidents (Class 0), despite having a high recall for severe accidents (0.91). This suggests that the model had trouble accurately classifying these cases. While Random Forest also demonstrated excellent precision (0.87) for severe accidents, its overall efficacy was limited by poor memory and precision for less severe incidents.

The study demonstrated how well the Gradient Boosting model predicted the severity of accidents, highlighting the potential benefits of adopting advanced machine learning models in accident prediction. Legislators, insurance providers, and traffic management organizations should implement sophisticated machine learning algorithms such as Gradient Boosting to enhance prediction accuracy. This is in line with Poufinas et al. (2023) when he states that these models can improve real-time decision-making by enabling quicker and more precise responses to serious incidents, ultimately reducing casualties and increasing overall traffic safety. Significant determinants of accident severity identified in the study included the number of cars involved, the number of casualties, weather conditions, and whether the environment was urban or rural. Policies and procedures related to traffic safety should incorporate these key predictors. For instance, specific interventions like increased police presence, better signage, or improved road infrastructure could be targeted at high-risk areas based on these criteria. Additionally, data-driven public awareness programs could be developed to target high-risk behaviors, such as speeding and driving in hazardous conditions (Hughes et al., 2016).

Several important factors that predict accident severity were identified by the Gradient Boosting model's decision tree analysis from our study. The most important predictor was found to be the quantity of cars involved, which was followed by the number of casualties and the location of the accident—rural versus urban. These results are consistent with the body of research that highlights the influence of environmental factors and traffic density on accident outcomes (Sangare et al., 2021). The results of statistical significance testing showed that there were significant relationships between accident severity and light conditions, road types, surfaces, and urban/rural locations ($p < 0.001$). Severity was also significantly impacted by weather, especially extreme weather like storms and snow ($p < 0.001$). These findings support earlier research by Li et al. (2021)that emphasizes the significance of environmental and contextual elements in road safety by highlighting their significance in accident severity.

The durability of the Gradient Boosting model in handling imbalanced datasets is demonstrated by its ability to balance accuracy, precision, and recall—especially for severe incidents. This is in line with Zhao et al. (2023) findings, which showed how class weighting might enhance model performance for uncommon events. All things considered, the study's findings offer insightful information on how well different machine learning algorithms can anticipate the severity of accidents. Because of its greater performance, the Gradient Boosting model may be used in actual road safety systems to provide more precise forecasts and more precisely targeted safety measures. In the future, research could investigate how to improve forecast accuracy and reliability even further by incorporating new features or using different models.

5.3 Strengths and limitations of the study

The study offers insightful information about the prediction of accident severity through machine learning algorithms; its strengths and limitations shed light on the opportunities and drawbacks of employing these models for road safety. Advances in this subject can be made possible by improving the resilience of predictive models through a knowledge of these factors. This study excels in demonstrating the higher predictive performance of the

Gradient Boosting model, which captures complicated, non-linear correlations within the data, outperforming traditional models (Ghandour et al., 2020). The ability of sophisticated machine learning approaches to uncover crucial elements impacting accident severity is demonstrated by the model's durability in handling huge datasets with a variety of predictive features. Additionally, a deeper understanding was made possible by the interpretability that the application of SHAP analysis offered.

Nonetheless, a number of restrictions must be recognized, chiefly those about the availability and quality of the data. Critical information, including exact weather data at the moment of the accident, driver conduct, and real-time traffic circumstances, was frequently missing from accident databases that were already in existence (Fiorentini and Losa, 2020). This restriction might have limited the model's capacity to accurately represent the intricate relationships affecting the severity of accidents (Hu et al., 2020). In order to improve the data landscape, future research should strive to gather thorough, high-resolution data using cutting-edge technology like car sensors, traffic cameras, and Internet of Things devices.

The models' limited applicability to other contexts with different road conditions, traffic patterns, and regulatory environments stems from their development using data from particular geographic regions, raising concerns about the generalizability of the findings (Theofilatos and Yannis, 2014). To understand these models' broader effectiveness, validation across a variety of situations that take into account variations in infrastructure, weather, and socioeconomic factors is essential. Furthermore, one drawback in reflecting the dynamic nature of accident occurrence and severity is the study's use of historical accident data. Creating real-time prediction frameworks that incorporate live data feeds could enhance the temporal dynamics and adaptability of the models and allow for proactive interventions like targeted emergency response based on predicted accident severity or dynamic speed adjustments (Xu et al., 2018).

When using these models, ethical issues and bias mitigation are especially important because biased data might result in unfair risk assessments (Mekonnen et al., 2022). Model outputs may unintentionally be influenced by geography, vehicle type, or socioeconomic status biases. Future research has to concentrate on locating and reducing these kinds of biases in order

to guarantee just and equitable forecasts for a variety of demographics. The final major challenge is incorporating predictive insights into the current policy and decision-making processes. This is because incorporating predictive models into traffic management systems, emergency response protocols, and insurance assessments is necessary to turn technical discoveries into useful applications. In the end, a more data-driven approach to road safety will be facilitated by addressing these constraints and investigating the recommended areas of additional research to improve the prediction power, interpretability, and practical value of machine learning models.

One of the significant limitations of this study was the inability to successfully implement spatial models for predicting accident severity. Initially, spatial random forest models were explored to account for the geographic correlation between accident locations and severity outcomes. However, the results indicated a low accuracy of only 15%, which significantly underperformed compared to non-spatial models. This poor performance could be attributed to the lack of meaningful spatial autocorrelation in the data. Upon further investigation, there appeared to be no statistically significant spatial dependence between the geographic coordinates and the severity of accidents. According to Wu (2020) and Fotheringham et al. (2009), spatial models like Geographically Weighted Regression (GWR) are most effective when there is a clear spatial relationship between variables, which was not observed in this dataset.

This result disproved the hypothesis that common environmental elements, traffic patterns, or road conditions may affect accident severity based on geographic proximity. The lack of geographical correlation reduced the applicability of spatial models in this situation. On the other hand, without requiring spatial information, conventional machine learning models like Random Forest and Gradient Boosting showed improved prediction power and accuracy. Because adding geographical models did not improve the model's performance, they were excluded from the final analysis. This restriction emphasized the necessity for more thorough spatial data or sophisticated methods to identify minute geographic trends that could affect the severity of an accident.

5.4 Recommendations

The results of this study on the use of machine learning models to predict accident severity lead to the following recommendations to improve road safety and direct future research:

5.4.1 Policy recommendations

The results of this study highlight the urgent need for data-driven policy interventions that use cutting-edge machine learning algorithms, such as gradient boosting, to improve road safety. The Gradient Boosting model exhibits greater performance in predicting the severity of accidents, indicating the possibility of incorporating this kind of predictive analytics into traffic management and public safety plans. Machine learning algorithms should be used by transportation authorities and policymakers to identify high-risk areas and driver habits. Peden (2004) supports that this would enable targeted interventions that have the potential to drastically lower accident rates.

First, implementing predictive models in real-time traffic monitoring systems can provide actionable insights to prevent severe accidents. Authorities can utilize these models to forecast accident hotspots and deploy preventive measures, such as enhanced traffic enforcement, improved road signage, and adaptive speed limits, tailored to specific locations and times of day. (Kennedy et al., 2018) states that such interventions align with evidence suggesting that predictive analytics can improve public safety outcomes by preemptively addressing risk factors .

Secondly, integrating machine learning insights into driver education and licensing programs could enhance awareness of behaviors associated with severe accidents, such as speeding and distracted driving. By educating drivers on the predictors of high-risk scenarios, tailored training programs can be developed to mitigate these behaviors, ultimately leading to safer driving practices (Cutello et al., 2020). Additionally, Cutello et al. (2020) suggests that insurance companies could use these insights to adjust premium rates based on predictive risk assessments, promoting safer driving through financial incentives.

Furthermore, investment in data infrastructure is crucial to support the continuous collection and analysis of traffic data. A robust data framework would enable the seamless integration of machine learning models into daily traffic management operations, enhancing the accuracy and timeliness of accident severity predictions. Policymakers should prioritize funding for technologies that facilitate real-time data analytics and model deployment to maximize the benefits of predictive modeling in public safety (Mbuh et al., 2019).

In conclusion, embracing machine learning technologies for accident severity prediction offers a transformative approach to road safety management. By integrating predictive analytics into policy frameworks, transportation authorities can develop more proactive, data-driven strategies that not only forecast but also prevent severe accidents, thereby safeguarding public health and reducing economic losses associated with traffic incidents.

5.4.2 Recommendations for further studies

Future research should focus on exploring the integration of more advanced deep learning techniques, such as neural networks, to enhance the accuracy of accident severity predictions. According to Matos et al. (2024), deep learning models, particularly recurrent and convolutional neural networks, have shown superior performance in handling complex and non-linear relationships in traffic data, which traditional machine learning models might overlook. Additionally, investigating the role of real-time data integration, including weather conditions, road surface status, and vehicle telemetry data, could provide a more holistic approach to predicting accident severity Matos et al. (2024).

Furthermore, it is recommended to examine the ethical implications of predictive analytics in road safety, particularly concerning data privacy and the potential biases in model predictions. As highlighted by Matos et al. (2024), addressing these biases is critical to ensuring fair and equitable application of predictive models across diverse populations. Expanding research in these areas will further advance the field and contribute to safer and more efficient transportation systems.

5.5 Conclusion

The study effectively showed how machine learning algorithms can anticipate the severity of accidents, and the Gradient Boosting model turned out to be the most successful. When it came to accident severity prediction, the Gradient Boosting model outperformed the Random Forest and XGBoost models by a considerable margin. It outperformed XGBoost (0.8023) and Random Forest (0.7467) with an accuracy of 0.8449, demonstrating its superior ability to handle complicated datasets and produce correct predictions. Important variables that were found to be significant predictors of accident severity included the number of cars involved, the number of casualties, and the location of the event—rural versus urban. These variables play a crucial role in predicting the severity of accidents, and the Gradient Boosting model effectively used them to improve forecast accuracy. Clear insights into feature relevance were obtained from the Gradient Boosting model's decision tree analysis. This improved interpretability made it easier to understand how different factors affect the outcome of accidents, which is important for putting targeted safety measures in place. This was in line with Chamseddine et al. (2022) concept of improving interpretability in machine learning models. The results highlight how machine learning methods, in particular Gradient Boosting, might be used to increase traffic safety. Through the use of such cutting-edge methods, interested parties can create accident response and prevention plans that are more precise and successful. The model's capacity to strike a balance between memory and accuracy, particularly in the case of serious incidents, points to important advantages for both policy development and traffic safety management. The better performance of the Gradient Boosting model overall indicates its potential to improve road safety predictions and direct strategic responses. On the basis of these results, future studies could improve predictive models and investigate other characteristics that might affect the severity of accidents.

References

Abellán, J., López, G., and De OñA, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications*, 40(15):6047–6054.

Ajagbe, S. A., Oladipupo, M. A., and Balogun, E. O. (2020). Crime belt monitoring via data visualization: a case study of folium. *International Journal of Information Security, Privacy and Digital Forensic*, 4(2):35–44.

Amiri, Z., Heidari, A., Navimipour, N. J., Unal, M., and Mousavi, A. (2023). Adventures in data analysis: A systematic review of deep learning techniques for pattern recognition in cyber-physical-social systems. *Multimedia Tools and Applications*, pages 1–65.

Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., and Amjad, U. (2022). Prediction of pile bearing capacity using xgboost algorithm: modeling and performance evaluation. *Applied Sciences*, 12(4):2126.

Andrey, J. C. and Mills, B. E. (2003). *Collisions, casualties, and costs: Weathering the elements on canadian roads*. Institute for Catastrophic Loss Reduction London, ON, Canada.

Ang, K. L.-M., Seng, J. K. P., Ngharamike, E., and Ijemaru, G. K. (2022). Emerging technologies for smart cities' transportation: geo-information, data analytics and machine learning approaches. *ISPRS International Journal of Geo-Information*, 11(2):85.

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967.

Bhavan, T. (2019). The economic impact of road accidents: the case of sri lanka. *South Asia Economic Journal*, 20(1):124–137.

Cai, Q., Abdel-Aty, M., Zheng, O., and Wu, Y. (2022). Applying machine learning and google street view to explore effects of drivers' visual environment on traffic safety. *Transportation research part C: emerging technologies*, 135:103541.

Chamseddine, E., Mansouri, N., Soui, M., and Abed, M. (2022). Handling class imbalance in covid-19 chest x-ray images classification: Using smote and weighted loss. *Applied Soft Computing*, 129:109588.

Cutello, C. A., Hellier, E., Stander, J., and Hanoch, Y. (2020). Evaluating the effectiveness of a young driver-education intervention: Learn2live. *Transportation research part F: traffic psychology and behaviour*, 69:375–384.

Delen, D., Tomak, L., Topuz, K., and Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*, 4:118–131.

Fiorentini, N. and Losa, M. (2020). Handling imbalanced data in road crash severity prediction by machine learning algorithms. *Infrastructures*, 5(7):61.

Fiorentini, N., Pellegrini, D., and Losa, M. (2023). Overfitting prevention in accident prediction models: Bayesian regularization of artificial neural networks. *Transportation research record*, 2677(2):1455–1470.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., and Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. *Accident Analysis & Prevention*, 136:105429.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2009). Geographically weighted regression. *The Sage handbook of spatial analysis*, 1:243–254.

Gan, J., Li, L., Zhang, D., Yi, Z., and Xiang, Q. (2020). An alternative method for traffic accident severity prediction: using deep forests algorithm. *Journal of advanced transportation*, 2020:1–13.

Ghandour, A. J., Hammoud, H., and Al-Hajj, S. (2020). Analyzing factors associated with fatal road crashes: a machine learning approach. *International journal of environmental research and public health*, 17(11):4111.

Gillespie, A. A. (2012). A duty to note the details of drivers under the road traffic act 1988, s. 172? atkinson v dpp [2011] ewhc 3363 (admin). *The Journal of Criminal Law*, 76(2):110–112.

Guido, G., Shaffiee Haghshenas, S., Shaffiee Haghshenas, S., Vitale, A., Astarita, V., Park, Y., and Geem, Z. W. (2022). Evaluation of contributing factors affecting number of vehicles involved in crashes using machine learning techniques in rural roads of cosenza, italy. *Safety*, 8(2):28.

Guo, Q., Angah, O., Liu, Z., and Ban, X. J. (2021). Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors. *Transportation Research Part C: Emerging Technologies*, 124:102980.

Gutierrez-Osorio, C. and Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of traffic and transportation engineering (English edition)*, 7(4):432–446.

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., and Sadeek, S. N. (2019). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, 124:66–84.

Hu, Q., Cai, M., Mohabbati-Kalejahi, N., Mehdizadeh, A., Alamdar Yazdi, M. A., Vinel, A., and Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. part 2: Prescriptive modeling. *Sensors*, 20(4):1096.

Hua, J., Li, L., Ning, P., Schwebel, D. C., He, J., Rao, Z., Cheng, P., Li, R., Fu, Y., Li, J., et al. (2023). Road traffic death coding quality in the who mortality database. *Bulletin of the World Health Organization*, 101(10):637.

Hughes, B. P., Anund, A., and Falkmer, T. (2016). A comprehensive conceptual framework for road safety strategies. *Accident Analysis & Prevention*, 90:13–28.

Jamal, A., Zahid, M., Tauhidur Rahman, M., Al-Ahmadi, H. M., Almoshaogeh, M., Farooq, D., and Ahmad, M. (2021). Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study. *International journal of injury control and safety promotion*, 28(4):408–427.

Kennedy, L. W., Caplan, J. M., and Piza, E. L. (2018). *Risk-based policing: Evidence-based crime prevention with big data and spatial analytics*. University of California Press.

Li, G., Liao, Y., Guo, Q., Shen, C., and Lai, W. (2021). Traffic crash characteristics in shenzhen, china from 2014 to 2016. *International journal of environmental research and public health*, 18(3):1176.

Lobo, A., Ferreira, S., Iglesias, I., and Couto, A. (2019). Urban road crashes and weather conditions: Untangling the effects. *Sustainability*, 11(11):3176.

Matos, J., Gallifant, J., Chowdhury, A., Economou-Zavlanos, N., Charpignon, M.-L., Gichoya, J., Celi, L. A., Nazer, L., King, H., and Wong, A.-K. I. (2024). A clinician's guide to understanding bias in critical clinical prediction models. *Critical Care Clinics*, 40(4):827–857.

Mbuh, M., Metzger, P., Brandt, P., Fika, K., and Slinkey, M. (2019). Application of real-time gis analytics to support spatial intelligent decision-making in the era of big data for smart cities. *EAI Endorsed Transactions on Smart Cities*, 4(9).

McMillen, C., North, C., Mosley, M., and Smith, E. (2002). Untangling the psychiatric comorbidity of posttraumatic stress disorder in a sample of flood survivors. *Comprehensive Psychiatry*, 43(6):478–485.

Mekonnen, A. A., Beza, A. D., Sipos, T., et al. (2022). Estimating the value of statistical life in a road safety context based on the contingent valuation method. *Journal of Advanced Transportation*, 2022.

Peden, M. M. (2004). *World report on road traffic injury prevention*. World Health Organization.

Poufinas, T., Gogas, P., Papadimitriou, T., and Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. *Risks*, 11(9):164.

Rajamani, S. K. and Iyer, R. S. (2023). Use of python modules in ecological research. In *Perspectives on the Transition Toward Green and Climate Neutral Economies in Asia*, pages 182–206. IGI Global.

Ramraj, S., Uzir, N., Sunil, R., and Banerjee, S. (2016). Experimenting xgboost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40):651–662.

Rubaidi, Z. S., Ammar, B. B., and Aouicha, M. B. (2022). Fraud detection using large-scale imbalance dataset. *International Journal on Artificial Intelligence Tools*, 31(08):2250037.

Saifuzzaman, M. and Zheng, Z. (2014). Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transportation research part C: emerging technologies*, 48:379–403.

Sangare, M., Gupta, S., Bouzefrane, S., Banerjee, S., and Muhlethaler, P. (2021). Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning. *Expert Systems with Applications*, 167:113855.

Theofilatos, A. and Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72:244–256.

Wu, D. (2020). Spatially and temporally varying relationships between ecological footprint and influencing factors in china's provinces using geographically weighted regression (gwr). *Journal of Cleaner Production*, 261:121089.

Wu, H. and Levinson, D. (2021). The ensemble approach to forecasting: a review and synthesis. *Transportation Research Part C: Emerging Technologies*, 132:103357.

Xu, C., Ji, J., and Liu, P. (2018). The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation research part C: emerging technologies*, 95:47–60.

Yamout, O. (2022). *An Application of Neural Networks in Predictive Construction Equipment Maintenance*. PhD thesis.

Yuan, M. (2021). Gis research to address tensions in geography. *Singapore Journal of Tropical Geography*, 42(1):13–30.

Zhang, X., Onieva, E., Perallos, A., Osaba, E., and Lee, V. C. (2014). Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, 43:127–142.

Zhao, J., Ma, R., Sun, J., Zhang, R., and Zhang, C. (2023). Modeling and analysis of vehicle path dispersion at signalized intersections using explainable backpropagation neural networks. *Fundamental Research*.

Zheng, L., Ismail, K., and Meng, X. (2014). Traffic conflict techniques for road safety analysis: open questions and some insights. *Canadian journal of civil engineering*, 41(7):633–641.

Ziakopoulos, A. and Yannis, G. (2020). A review of spatial approaches in road safety. *Accident Analysis & Prevention*, 135:105323.

Appendix A

Ethical Review Report

27th March 2024

Mr Kirimi Mwenda Dennis,
kirimi.dennis@strathmore.edu

Dear Mr Kirimi,

RE: Application of Machine Learning and Big Data in Enhancing Road Safety by Predicting Accidents

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2052/24.** The approval period is from **27th March 2024 to 26th March 2025.**
This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
27-Mar-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix B

Similarity Index

148988_Proposal .pdf

ORIGINALITY REPORT

9%	8%	7%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Strathmore University Student Paper	1%
2	Submitted to Bogazici University Student Paper	<1%
3	"Innovations in Electrical and Electronic Engineering", Springer Science and Business Media LLC, 2024 Publication	<1%
4	www.mdpi.com Internet Source	<1%
5	Submitted to Machakos University Student Paper	<1%
6	www.epstem.net Internet Source	<1%
7	ouci.dntb.gov.ua Internet Source	<1%
8	Gerald K. Ijemaru, Kenneth L.-M Ang, Jasmine K.P. Seng. "Transformation from IoT to IoV for	<1%

Appendix C

Codes used in the study

For a detailed view of the code used in the development and testing of the Accident Severity model, please refer to the following Jupyter Notebook available on my GitHub repository:

- **Accident Severity Model Code**: https://github.com/kirimi2022/Accident-Severity-model/blob/main/Accident%20Severity%201.ipynb

Appendix D

Gradio App Screenshots for Accident Severity Prediction

The following screenshots illustrate the Gradio app designed for predicting accident severity. This app offers both a single input interface and a bulk upload option for CSV files.

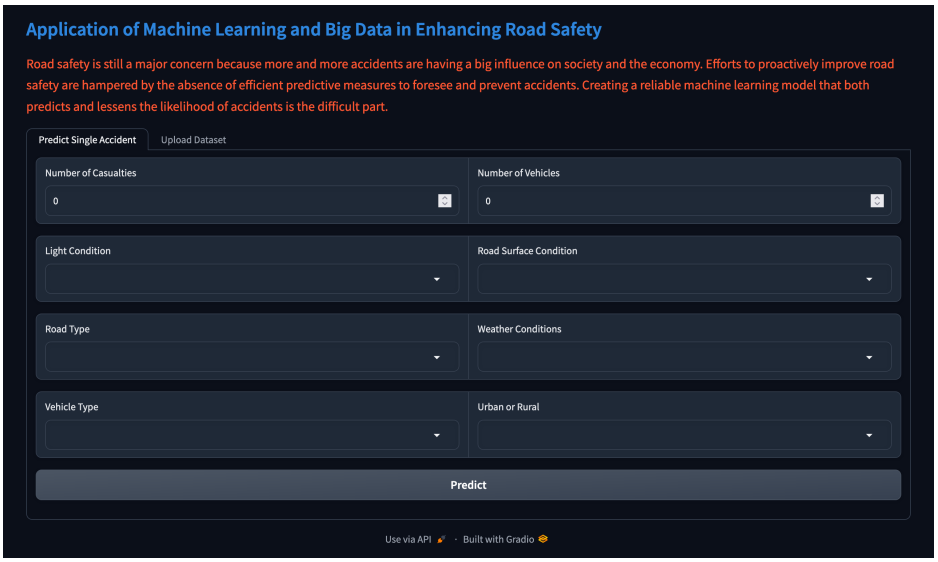


Figure 1: Single Input Interface for Accident Severity Prediction

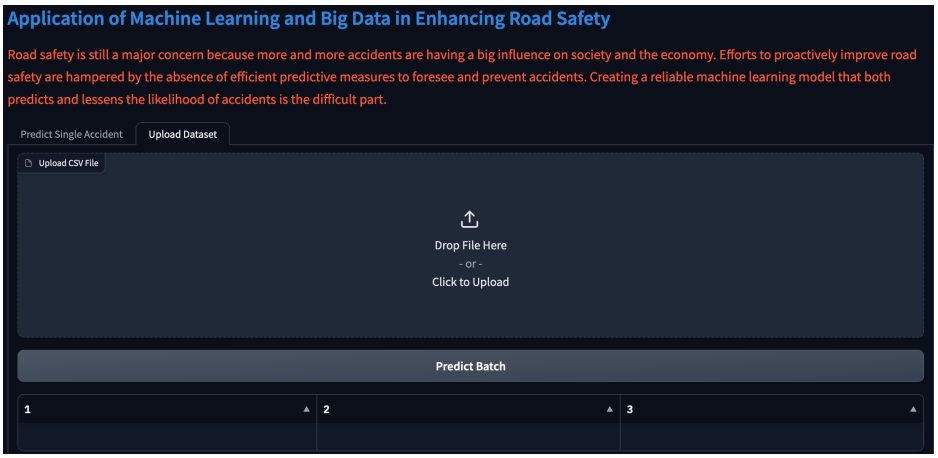


Figure 2: Bulk Upload Interface for CSV File Predictions