A MACHINE LEARNING-DRIVEN APPROACH FOR ACCIDENT PREDICTION AND TRAFFIC SAFETY ANALYSIS IN NAMIBIA

By

Ngolo Maria

218094701

Submitted in partial fulfilment of the requirements for the degree of

Master of Data Science

in the

Faculty of Computing and Informatics

Department of Informatics

at

Namibia University of Science and Technology

Supervisor: Dr Richard Maliwatu

METADATA

Study Title	A machine learning –driven approach for accident prediction and traffic safety analysis in Namibia
Student **Name**	Maria Kaunapawa Ngolo
Student Number	218094701
Faculty	Faculty of Computing and Informatics
Department	Department of Informatics
Programme	Master of Data Science (09MADS)
Status	Dissertation
Supervisor	Dr Richard Maliwatu
First Co-Supervisor	N/A
Digital Signature of Student	
Digital signature of Supervisor	
Digital Signature of Programme Coordinator (or equivalent)	
Digital Signature of FCI HDC Representative	

DECLARATION

I, Maria Ngolo, hereby declare that this thesis titled "**A Machine Learning-Driven Approach for Accident Prediction and Traffic Safety Analysis in Namibia**" is my own work. It is being submitted in partial satisfaction of the criteria for the degree of Master of Science, at the Namibia University of Science and Technology (NUST), Namibia.

Signature: Maria Ngolo____MNgolo________Date: ___03/12/2024_____________

ABSTRACT

Globally, road traffic accidents contribute a large portion of injuries, fatalities, and significant economic losses and ongoing research has projected that by 2030, car crashes would be the 5th top reason for loss of life around the world. The key cause of traffic accidents is hard to determine nowadays because of a complex mix of factors, such as road conditions, weather conditions, and the mental condition of the drivers, to list a few. Without a thorough understanding of the characteristics and causes, intelligence-led countermeasures to decrease crashes cannot be created or implemented. Therefore, if traffic accident characteristics can be better understood, it might be easier to take some mitigative action. Nowadays, the utility of machine learning methods in the field of road traffic crashes is gaining traction. The objective of this dissertation is to analyse historical data for a five-year period (2018-2023) and to understand the patterns in accident occurrences by making use of machine learning methods. Machine learning models as such Random Forest, Support Vector Machine, K-Nearest Neighbours, Association Rule Algorithm (AARA), and k-Clustering were employed on the dataset. The Apriori Association Rule algorithm explored the rules with high lift and high support, respectively. The research shows that the Random Forest model is the reliable model in predicting crash severity, reaching an accuracy of approximately 81%. Factors such as junction type, poor road sign conditions, uncontrolled traffic, weather, lighting, road surface, vehicle type, and driver behaviours were identified as the significant variables influencing road accidents. Pedestrian, rollovers, and collision are the leading crash causes of the road accidents, and they are associated with uncontrolled traffic and daylight. Additionally, the research shows notable differences in accident rates by region, month, year, day of the week, and hour of the day, underscoring the impact of geographical features, seasonal trends, and commuting habits on accident rates. The findings indicate that high traffic volumes and urban congestion are the main causes of the greatest accident rates in metropolitan areas, especially in the Khomas Region of Namibia. Further, more accidents are happening more toward the weekend as compared to weekdays and during night hours.

ACKNOWLEDGEMENTS

First and foremost, I want to thank God for His love and mercy that endures forever. My sincere gratitude goes to my research supervisor, Dr. Richard Maliwatu, for his guidance and continuous support. Without his supervision, this work would not have seen the light of day. I shall be forever grateful. I also want to thank all my lecturers from the time I started my studies at the Namibia University of Science and Technology. This project is essentially a culmination of the various aspects I learned from each of them.

Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

Abbreviation/Acronym	Description
AADT	Annual Average Daily Traffic
ANN	Artificial Neural Network
BRF	Balanced Random Forest
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Tree
EB	Empirical Bayes
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GDP	Gross Domestic Product
GIS	Geographic Information System
HSID	Hotspot Identification
IoT	Internet of Things
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
OLS	Ordinary Least Squares
RWIS	Roadway Weather Information System
SEM	Spatial Error Model
SLM	Spatial Lag Model
SSAE	Stacked Sparse Autoencoder
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
WHO	World Health Organisation
XGB	Extreme Gradient Boosting

CHAPTER ONE: INTRODUCTION

1.1 Background of the study

Car accidents are posing significant threats to public safety and economic development, making them a critical matter worldwide (Smith & Johnson, 2018) and have become a prevalent major global public health issue, as highlighted by the WHO in its 2023 Global Status Report on Road Safety (WHO, 2023a).The increasing petition for road transportation in current periods has brought about frequent challenges, one of the most serious being traffic congestion. This congestion importantly rises the likelihood of accidents, often leading to severe injuries and fatalities (Harrou et al., 2024). Moreover, traffic crashes account for most casualties and injuries worldwide because of the rapid proliferation of vehicle ownership and traffic demand, especially for young people aged 15 to 29 (Jia et al., 2018). According to the World Health Organisation (WHO), traffic accidents result in between 20 and 50 million injuries and 1.35 million fatalities per year(*Road Traffic Injuries*, 2021). Apart from personal and family suffering (such as physical injuries, emotional distress, and trauma experienced by individuals and their families) because of road traffic accidents, these accidents can also lead to significant economic burdens at both individual and national levels as a result of limited spatial road safety analysis strategies and poor traffic control strategies (Berhanu et al., 2023).

In addition, the expenses of medical care and the impact on productivity result in financial losses for those who die or become disabled in accidents. Not to mention, another factor that contributes to these economic challenges is family members who must take time off from their jobs or education to provide care for the injured. Remarkably, these road accidents cost 3% of most countries' gross domestic product (GDP) (*Road Traffic Injuries*, 2021).

Namibia faces challenges in ensuring road safety and minimising crashes due to its diverse landscape and rapidly expanding towns and cities(Namibia Ministry of Works and Transport, 2020). Thus, the deployment of innovative machine learning techniques present an opportunity to optimise collision prediction and improve road safety operations within the country (L. H. Brown & Jones, 2021). Recent research, for instance, emphasises the use of high-resolution risk maps, which are created using satellite imagery and GPS data, to pinpoint accidents-prone locations even in the absence of crash data. This invention provides scalable methods for proactive accident prevention and has been used in cities all around the world (Garcia et al., 2020; Khan et al., 2024).

The rapid increase in the number of smart devices and the development of the Internet of Things (IoT) have resulted in a significant rise in the amount of data that is now accessible (Zanella et al., 2014). The arrival of this continuous stream of data can have a crucial impact on the development of precise accident forecasting algorithms. Algorithms for machine learning may use data from various sources such as road sensors, GPS-equipped automobiles, and weather stations, to give estimated insights with greater accuracy. The combination of data facilitates the collection of useful data (Johnson et al., 2022). This approach not only enables authorities to anticipate possible regions where collisions are most likely to occur, but it also gives them the ability to make timely decisions with the goal of effectively deploying resources and ensuring road safety.

Namibia's road network plays a crucial role in fostering economic growth and connecting remote areas, yet it still faces challenges in effectively implementing road safety regulations and lacks real-time collision forecasting systems. Nevertheless, there is a chance to transform road safety management by employing proactive accident model predictions (Garcia et al., 2020) that apply algorithmic learning approaches, driven by breakthroughs in data analytics.

The main benefit of machine learning is its capacity to adapt to different situations, given that each country's road safety challenges are unique. Multiple factors, such as road conditions, wildlife crossings, and cultural standards, impact road safety in Namibia. Machine learning-powered methods enable the adaptation of forecasting models to overcome specific obstacles. Using local information and statistics can significantly improve the precision of estimates, thereby providing an important opportunity to establish a safer road environment (Smith et al., 2021).

Machine learning provides the advantage of doing multi-factor analysis, in contrast with standard accident inspection, which focuses on a limited set of parameters such as road conditions and vehicle speeds. A complete analysis can consider several elements, such as road structure, driver behaviour, traffic patterns, and socio-economic considerations (Ardakani et al., 2023). This thorough strategy provides a greater awareness of the basic causes that lead to accidents, allowing regulators to utilise specific steps to address the core issues that cause road safety dangers.

Furthermore, machine learning-driven accident prediction systems provide a notable benefit in their capacity to adjust and develop using real-time data (Chen et al., 2018). These systems can constantly acquire knowledge from new data inputs, hence improving their ability to make accurate predictions as time goes on. This adaptability permits the introduction of dynamic road safety measures that may effectively address evolving situations and developing trends. For example, if the algorithm notices a significant spike in traffic due to an unforeseen occurrence, it can suggest diverting traffic or expanding the presence of authorities to prevent collisions.

Initiatives for better road safety frequently ignore the contribution of members of the public and neighbourhood groups. Machine learning models could foresee accidents and communicate with the public via user-friendly interfaces and applications (Cheng et al., 2019). These technologies provide drivers and pedestrians with up-to-date information about the probability of accidents, allowing them to make well-informed decisions and develop safer behaviours while travelling. To ensure timely public updates, they use various communication techniques, such as dedicated mobile applications, SMS alerts, GPS navigation services, public displays, social media platforms, websites, and push notifications. Furthermore, the incorporation of feedback mechanisms allows these systems to consistently learn from road users' experiences, promoting a collaborative approach to improving traffic safety.

The successful implementation of any traffic safety strategy relies on the cooperative efforts of several stakeholders, such as government institutions, infrastructure authorities, and technology vendors. Machine learning not only improves the ability to predict collisions and facilitates them, but it also supports the creation of data-driven policies. By examining the consequences of security measures or facility adjustments, the government can make sensible choices that successfully address the fundamental factors that cause crashes. This association corresponds to Namibia's goal of improving road safety as one of its greater objectives for prosperity (Ardakani et al., 2023b).

The road traffic accident study has made significant progress towards developing widely recognised frameworks such as the Human Factors Theory and the Haddon Matrix. The application of these approaches has played a major part in improving our comprehension of the complicated features of crashes (Afrin & Yodo, 2020; Jilani et al., 2023). These beliefs emphasise the importance of human behaviour, vehicle factors, and the environment in contributing to accidents. Afrin and Yodo (2020) conducted a study that explores various strategies to alleviate road traffic congestion. Addressing this element is crucial for improving road safety and preventing accidents. The report emphasises the importance of adopting sustainable transportation solutions and building a strong transportation infrastructure. It is crucial, especially for regions such as Namibia, that have diverse terrains and varying traffic flow patterns (Smith et al., 2021).

Researchers frequently use a variety of statistical regression methods to model the severity of crash damage. Past studies have focused on leveraging traditional methods for data analysis (Kumar et al., 2020). While mathematical interpretation of these approaches can enhance the understanding of predictor variables, they may face limitations like reduced predictive accuracy and biased model estimation (Keilmann & Weber, 2021). So, machine learning (ML) methods have grown as an alternate to the old ways of looking at factors related to the severity of the injury in order to get useful information from large amounts of complex data (Keilmann & Weber, 2021b). Therefore, this study seeks to utilise machine learning methodologies to forecast vulnerable regions and anticipate the timing of accidents.

1.2 Problem Statement

According to the World Economic Forum's Quality of Road Infrastructure, Namibia has retained its position at the top of the list for the best roads in Africa for the fifth consecutive year. Namibia was graded as one of the countries with the best quality roads in Africa, with a score of 5.2 out of seven (World Economic Forum, 2023). Beside this recognition, Namibia continues to face a concerning trend of traffic accidents, with a significant rise in injuries and fatalities. At the beginning of the year, the Namibian Police (NamPol) revealed concerning statistics- revealing a 5% increase from last year in car accidents during the 2023/2024 festive period (Namibian Police, 2024). Despite the concerning rise, there was a modestly positive development as the mortality rate declined by 14% and this shows that there was progress in emergency and healthcare facilities. However, the overall road safety situation remains worrisome, as there has been a 5% surge in crashes and a 2% uptick in injuries throughout the examined timeframe. According to NamPol's figures for the 2023–2024 year, there were more than 400 traffic incidents, leading to 724 injuries and 57 fatalities. Furthermore, there was a significant 19% surge in the number of fatalities on the road in 2023 compared to 2022, resulting in 184 deaths from 173 fatal crashes. This indicates a rise from the 155 fatalities in 149 crashes recorded in the preceding year. The substantial increase in fatal crashes on roads underscores the critical requirement for the introduction of innovative projects to solve road safety issues in Namibia.

Moreover, the increasing occurrence of road traffic accidents in Namibia requires urgent intervention, as statistics show a death rate of 466 per 100,000 inhabitants and 18 injuries per 100,000 individuals in 2022. The given figures were presented at the 7th UN Global Road Safety Week in Windhoek, emphasising the problem's seriousness not only in Namibia but worldwide (United Nations, 2023). Annually, car accidents cause more than 1.3 million deaths worldwide, with up to 50 million people suffering injuries. The United Nations cautions that if swift and decisive action is not taken, traffic accidents might result in 500 million injuries and 13 million more fatalities over the course of the next ten years, particularly in less developed and middle-income nations like Namibia. In addition, the World Health Organisation (WHO) has calculated that Namibia has a yearly death rate of 30.4 per 100,000 individuals due to car accidents. This statistic emphasises Namibia's status as one of the most perilous countries in Sub-Saharan Africa in terms of traffic safety. The average fatal accident rate for the region is 26.6 road deaths per 100,000 individuals, according to the WHO's 2018 report (World Health Organisation, 2019).

Therefore, Namibia is experiencing a critical and ongoing road safety crisis due to a disturbingly elevated incidence of traffic accidents and fatalities. Despite its relatively small population and a reasonable car ownership rate—401,476 registered vehicles in 2021 (15,867.7 per 100,000 individuals)—Namibia is ranked in the top 20 countries globally for the highest road traffic accident death rate per 100,000 people. Nearly every day, reports of road accidents emerge, and the annual increase in incidents has become deeply concerning. The disproportionate rate of road traffic deaths jeopardises lives and positions Namibia precariously regarding road safety. Yet, the fundamental causes of these accidents are inadequately understood, prompting urgent enquiries into the principal components contributing to the increasing rates of road accidents and their underlying causes.

1.3 Aim of the Study

The purpose of the study is to employ a machine method that incorporates historical accident data to uncover actionable insights from historical data to help towards accident prevention policies and programmes.

1.3.1 Research Objective

In order to increase road safety and lower the frequency of traffic accidents, this study intends to create an accident prediction model for Namibia's road network that is powered by machine learning.

Sub- Objectives:

The main objective was broken up into the following sub-objectives:

I. To identify important features that contribute to crash severity

II. To enhance Model Performance with Historical Context

III. To evaluate Model Effectiveness Using Historical Records

1.3.2 Research Question:

How can a reliable machine learning model be developed to predict road accidents in Namibia and enhance road safety using historical accident data?

Sub Questions:

I. What are the major causes of traffic accidents in Namibia?

According to the results, the major causes including human errors such as reckless and *negligent driving, poor road conditions, and inadequate traffic control and speeding.*

II. What are the predominant types of traffic accidents?

The most dominant accident involves pedestrians, followed by collisions, and rollovers.

III. Out of the commonly used machine learning techniques such as Random Forest, Support Vector machine and K-Nearest neighbours (KNN), which model can be effectively predicting road data?

Based on the model results, Random Forest outweigh the other models, with an accuracy of 81%, while the rest are 70% and 69%, respectively.

1.3.3 Significance of the study:

The research intends to boost public safety by targeting the rising incidence of traffic accidents and helping road authorities with improving management, efficient budgeting, and appropriate resource distribution. It employs a data-driven, proactive strategy for road safety that minimises the economic consequences of accidents for both individuals and the nation. Further, it complements the current studies on road safety in Namibia, delivering valuable insights for future initiatives and policy formulations.

1.3.4 Scope of the study

The study aims to analyse accident data in Namibia, covering a 5-year period (2028-2023) and covering all 14 regions and not specific regions, with the aim to uncover actionable insights from Namibia's accident historical data at a national level only to help towards accident prevention policies and programmes. The study followed a positivist research philosophy because the study focused on noticeable phenomena and quantitative data to be able to develop predictive models. The secondary data, collected from the MVA fund and the Namibian National Road Safety Council (NRSC), were applied to machine learning techniques such as random forest, support vector machine, and K-nearest neighbours (KNN). Variables that are considered in the models are crash priority, which is the dependent variable; regions; crash types; crash cause; vehicle type; road condition; weather; visibility; time; day of the week, etc. However, it has excluded suburbs since it has more missing values; age and citizenship data were not available.

1.4 Ethical Considerations

The study conducted an ethical review by obtaining a signed and approved ethical clearance from Namibia University of Science and Technology (NUST) to ensure data privacy and fairness while also implementing transparency and accountability measures for the model's deployment.

1.5 Organisation of the dissertation

The dissertation's remaining sections are structured as follows:

- Chapter two

  Covers the literature review, both theoretical and empirical literature, discussing what other researchers and book authors have done on similar topic to this research study, what other research employed and what was their outcomes.

- Chapter three

  Covers the methodology, on how the study was carried out and this includes Saunders' research onion (Research philosophy, Research approach, Research strategy, Choices, Time horizon and Techniques and procedures research).

- Chapter four

  Presents the results and discussion of the study from the descriptive analysis based on the model's results including interpretations.

- Chapter five

  Covers the conclusion highlighting the study's findings of the study, limitations and recommendations

- References
- Appendix

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The state of the art in the relevant disciplines for the described work—such as related works, data, and machine learning algorithms—is introduced in this chapter. This study reviewed the literature to uncover pertinent papers on machine learning in crashes, traffic accidents, algorithms, and techniques to bolster the study. This allowed the study to identify the models and variables used in previous studies. This chapter delves into both theoretical and empirical literature, offering valuable insights and context for the research conducted in this study. Looking at relevant concepts, methodologies, and research findings related to machine learning in road accidents can help this study to discover knowledge gaps that this study will solve. This extensive examination will improve Namibian accident forecasts and traffic safety analysis.

2.2 Theoretical literature

2.2.1 Established Frameworks in Accident Causation

The theoretical review explains current theories (concepts or entire theories), their connections, how they have been expanded upon, and how new hypotheses have been developed. A collision on the road is regarded as a failure of the driver's system to monitor one or more behaviours necessary for the trip to be completed safely. Insufficient roadways and a lack of adequate and rational execution are the main causes of vehicle accidents (Ahmed et al., 2024). The main causes of traffic accidents are poor roads, careless driving, fatigued driving, drowsiness, stupor, sickness, using cell phones, eating and drinking in the car, being careless when making a mistake on the road, and other drivers' inability to respond appropriately to the situation (Sun, Liu, Chen, & He, 2019). Traumatic injuries from traffic accidents can affect any part of the body, regardless of the vehicle. The head, abdomen, pelvis, chest, and spine are among the body's most vulnerable areas, with catastrophic effects (Ahmed et al., 2023).

Similarly, well-established frameworks in the field of road traffic accident theories, such as the Human Factors Theory and the Haddon Matrix, have provided a foundation for understanding the complex nature of accidents (Jilani et al., 2023). These ideas highlight the significance of human behaviour (alcohol, speeding, mobile phone, drivers' gender, and age), vehicle-related factors, such as defects in the motor vehicle and the environment factors such as road, and weather conditions, in contributing to accidents. However, modern literature offers valuable enhancements to these perspectives. In their study, Afrin and Yodo, (2020) conducted a survey to examine solutions aimed at mitigating road traffic congestion, which is a critical determinant of road safety and accident prevention. The survey emphasised the immediate need for sustainable transportation solutions and a strong transportation infrastructure, especially in countries like Namibia, which is characterised by varied terrains and unpredictable traffic patterns (J. Smith & al., 2023).

2.2.2 Application of Machine Learning in Traffic Flow Forecasting

The book titled: "*Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers"* by Awad and Khanna (2015) offers a comprehensive and in-depth theoretical framework for understanding learning machines. While the book does not explicitly address road safety, it does offer fundamental principles and concepts that can be immediately employed to machine-learning-based research on traffic safety. It is essential to first grasp certain basic concepts in succession to fully comprehend the complexities of machine learning in traffic management and safety.

Hu et al. (2015) also discuss the use of machine method in practice, namely the use of the least squares support vector machine for traffic flow forecasting. New developments and breakthroughs in machine learning methods, as shown by this application, are a big step forward in making models that can predict traffic flow, which is a key part of predicting accidents and analysing traffic safety (Johnson et al., 2022). The incorporation of machine techniques -based models for predictions represents a major advance in traffic safety examination, delivering instant and reliable data about traffic conditions and enabling the implementation of preventative measures to reduce congestion and improve road safety (Johnson et al., 2023).

Ensemble learning methods are gradually employed for predicting road traffic crash severity due to their capability to merge the leads of many models. Mansoor et al.,(2023) proved that a two-layer ensemble model combining RF and XGBoost accomplished better accuracy and F1 scored relatively well compared to single-model methods. Similarly, Zhang et al.,(2023) introduced a data-driven soft sensor that uses ensemble learning techniques to predict the velocities of swarm motion. This exploration determines the adaptability of machine method in managing traffic-related difficulties, with a particular focus on forecasting the speeds of swarm movements. Accurate predictions in such scenarios are of paramount importance for proactive traffic management and accident prevention (Z. Islam & Abdel-Aty, 2023).

In addition , Cong et al., (2016) conducted a research that explores the practical applications of machine learning, specifically traffic flow prediction focusing on the use of the least squares support vector machine. The advancements and outcomes in machine learning techniques greatly streamline the creation of predictive models for traffic patterns, as seen in the example. Conducting traffic safety analysis and predicting accidents is a crucial component (Johnson et al., 2022). Machine learning algorithms have significantly improved the analysis of traffic safety by providing rapid and precise data on traffic conditions. This enables the use of proactive strategies that reduce traffic congestion and improve road safety (Johnson et al., 2023).

2.3 Empirical literature

2.3.1 Data Mining Techniques in Accidents Analysis

Empirical literature explores past studies in view of attempting to answer specific research questions. Numerous empirical studies have investigated the numerous factors that cause traffic accidents. Research has proven that machine learning models successfully anticipate crash severity. Ahmed et al.,(2023) performed a thorough assessment comparing numerous machine learning methods, which includes Random Forest (RF), Decision Jungle, and CatBoost. Their outcomes indicated that RF attained the highest accuracy of 81.45% in predicting accident severity using an extensive dataset from New Zealand traffic accident dataset (2016–2020). Likewise, a case study from Italy conducted by Meocci et al., (2023) applied Gradient Boosting to evaluate pedestrian accident risks, determining visibility, lighting conditions, and road features as critical factors. M. K. Islam et al., (2022) have examined crash severity and hotspots in Saudi Arabia utilising Gradient Boosting and Random Forest models. Their research highlighted essential elements and sites linked to serious accidents, offering significant insights for focused interventions. The results demonstrated that the RF model surpassed other classifiers in predicting crash severity, attaining the highest accuracy and robustness in classification.

Researchers commonly use data mining techniques to analyse road accident data and identify the specific elements that influence accident severity. According to Kumar and Toshniwal (2015), using data mining techniques like categorisation, association rule mining, and algorithmic clustering can help identify areas that are prone to accidents and assess the conditions around accident occurrences. This approach is quite beneficial for evaluating the relevant factors in traffic accidents. Finding accident hotspots is another critical factor that is typically the first stage of road safety studies. Inaccuracies in hotspot identification could produce inferior outcomes. Montella (2010) conducted a comparison of common hotspot identification (HSID) techniques.

The Empirical Bayes approach (EB), outperforming other HSID techniques, has established to be the greatest reliable and reliable technique. In the study by Sandor et al., (2014), accident hotspots are searched using the GPS coordinates of the accidents using the clustering approach (DBSCAN) as an option to the traditional HSID methods. Finding hotspots (or clusters) with shorter lengths and greater accident densities is made feasible by the DBSCAN technique. Low-density areas will also be eliminated during the process. The author employed performance measures to assess the efficacy of the model. Widely employed measures comprise false positive rate (FPR), sensitivity, specificity, accuracy, and precision. However, Roshandel et al. (2015) discovered that only a limited percentage of research utilises all these measures to thoroughly assess their models. The authors assert that it is critical to authenticate any forecasting model using a wide range of indications.

2.3.2 Effects of Traffic Congestion on Driver Behaviour

Isaksen and Johansen's (2021) study examined how traffic congestion affects driver behaviour in situations where pollutants have dispersed. They applied behavioural analytic methods to evaluate driver responses. According to their study, traffic congestion had a significant impact on how drivers behaved, causing them to change their behaviour after experiencing congestion. These behaviours included increased levels of stress, diminished tolerance, and an elevated inclination to partake in hazardous driving practices.

Wåhlberg et al. (2011) conducted an extensive survey study in England and the UK, focusing on crucial driver behaviour factors that contribute to traffic incidents. Wåhlberg et al. (2011) identified medical issues, poor vision, speeding, inexperience, and impaired driving (due to drugs or alcohol) as the six primary factors that contributed to the crashes in the survey's hypothetical scenarios. In addition, driving at night was significantly associated with an increase in the rate of RTA and this might be due to poor visibility at night, and sleepiness of the drivers.

Other environmental and geographical factors, such as visibility, weather conditions, land use, and road layout, influence traffic accidents in addition to drivers' behaviour. Salifu (2002) defines a collision hot zone as a specific geographic location where accidents, especially those resulting in injuries or fatalities, happen more frequently compared to surrounding areas. Anderson (2009) used the clustering technique to identify areas with a high concentration of traffic collisions. This study additionally connected the hotspots to environmental and land use data to obtain insights into the intricacies of traffic accidents and their correlation with specific geographical areas. According to the data, both the spatial lag model (SLM) and the spatial error model (SEM) outperform the ordinary least squares (OLS) model. In terms of p-value and correlation coefficient, the SEM demonstrates superior performance.

2.3.3 Real-time Traffic Management Systems

Allström et al. (2017) conducted a study that focused on the use of intelligent traffic management systems in urban areas to analyse the complexities of traffic management in smart cities. Their research focused on the importance of continuous monitoring and using data-driven technology to control traffic. The results highlighted the effectiveness of intelligent traffic management systems, which frequently use machine learning algorithms, in mitigating traffic congestion while improving overall traffic movement, resulting in increased road safety in urban areas.

2.3.4 Impact of Road Surface Characteristics and Influence of Weather Conditions on Road Safety

Moreover, the state of the road pavement has a substantial influence on traffic collisions. Wet road surfaces, caused by variables such as rainfall and heavy precipitation, increase slipperiness and the risk of accidents (Lee et al., 2018). The chi-square tests have demonstrated a significant link between road surface characteristics and road traffic accidents, particularly on curved routes (Ardakani et al., (2023). The relationship is particularly noticeable during certain seasons, especially the summer, when unfavourable weather conditions increase the likelihood of accidents. (Hayat & associates,2013). Muktar and Fono (2024) conducted research to develop a web-based machine learning method for predicting the severity of traffic incidents in Montreal. They aimed to identify the primary factors influencing the severity of accidents and utilise that knowledge to enhance road safety and improve the efficacy of emergency responders. The research employed categorisation techniques such as eXtreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), Random Forest (RF), and Gradient Boosting (GB). The dataset comprised traffic accident records from Montreal spanning from 2012 to 2021, obtained from the Société de l'Assurance Automobile du Québec (SAAQ). The models were assessed based on precision, recall, F1-score, and accuracy, with XGBoost attaining the greatest accuracy at 96%, followed by CatBoost at 95%, Random Forest at 93%, and Gradient Boosting at 89%.

Lee et al., (2018) found a significant correlation between the severity of accidents and the time of day. In addition, Adanu et al., (2018) conducted an analysis on the significance of accidents happening on weekends compared to weekdays. They discovered that although more accidents take place on weekdays, a greater percentage of severe injuries occurred on weekends. Weather conditions have a major effect on the incidence of road traffic accidents. Li et al., (2017) underlines the effect of weather conditions, such as fog, rain, and snow, on road safety. Small-scale increments (SSI), which are minor alterations in weather conditions, can cause decreased visibility, hence increasing the probability of accidents (Li et al., 2017). Dastoorpoor et al., (2016) additionally validate the positive correlation between weather conditions, including precipitation, temperature, and windstorms, and the incidence of road traffic accidents. Elassad et al., (2020) employed support vector machines and multi-layer perceptron to forecast accidents using weather conditions, with notable levels of accuracy. In addition, Lawal (2021) conducted an analysis of road accidents in the UK and Scotland, examining several scenarios and emphasising the influence of lighting conditions and weather on the severity of incidents. Weather conditions influence driver visibility, but it is important to acknowledge that a significant number of incidents occur during daylight hours and on dry road surfaces (Abbas et al., 2023).

Huang et al., (2019) and Chen and Wang et al., (2020) have conducted empirical research that corroborates the idea that including meteorological data has a substantial influence on accident prediction. Researchers have discovered that incorporating meteorological factors, such as temperature and rainfall, can greatly enhance the accuracy of models, especially in predicting weather-related occurrences. Furthermore, Wang and Liu, (2022) have conducted empirical research. Asha and Narasimhadhan, (2018) investigated the use of YOLO (You Only Look Once) and a correlation filter for vehicle counts in traffic management systems. The researchers utilised YOLO, a deep learning technique for object recognition, to accurately ascertain the real-time count of automobiles. The findings highlighted the critical requirement for accurate vehicle counting in order to effectively monitor traffic in real-time and identify congestion points. Having quick access to up-to-date vehicle count data enables traffic management authorities to make well-informed decisions and optimise traffic signal timing, resulting in reduced congestion and a lower probability of accidents.

Komsiyah and Desvania (2021) did an empirical study that focused on analysing and simulating traffic lights at intersections with three signals using fuzzy inference methods. To improve signal timing, the researchers utilised traffic signal control models based on fuzzy logic. Research has shown that implementing traffic signal control using fuzzy logic can effectively optimise signal timing, ease traffic congestion, and improve traffic safety. Using fuzzy logic, traffic lights can adapt to changing traffic conditions, which improves traffic flow and reduces accidents. Corovic et al., (2018) utilised the YOLO (You Only Look Once) algorithm to accurately detect and track traffic participants in real-time. This technology greatly enhances the capacity to monitor traffic in real-time by precisely recognising and monitoring vehicles and pedestrians. The results demonstrated that the YOLO algorithm has the capability to precisely identify traffic participants, facilitating prompt decision-making for traffic control and mitigating accidents.

In addition, Makaba et al., (2020) introduced a paradigm based on Bayesian networks to evaluate the financial implications of road traffic collisions. A Bayesian network framework model was developed using real-life road traffic collision data and expert knowledge to assess the cost of road traffic collisions. The study demonstrated that the Bayesian network-based framework is an effective tool for assessing the economic consequences of accidents, directing the allocation of resources for accident prevention, and reducing financial burdens. Lykov and Asakura

(2020) applied the tensor and continuum models to identify unforeseen traffic flow patterns in prominent urban areas. Their methodology identifies and manages abnormal traffic patterns, assisting authorities in promptly addressing traffic issues and mitigating safety hazards. Their investigation revealed that the tensor-based approach has the capability to identify and address traffic irregularities, hence enhancing safety.

Jain (2011) also investigated the application of fuzzy logic in traffic signal control automation. The study presented a novel traffic signal controller that utilises fuzzy logic to adjust signal timing in response to current traffic circumstances. The findings indicate that the traffic signal control system, which is based on fuzzy logic, effectively improves traffic flow, and minimises accidents caused by congestion by adjusting signal timing in real-time. Nakanishi and Auza's (2023) primary goal was to use optimisation approaches for traffic light control and coordination. The study aimed to optimise traffic light coordination to minimise congestion and improve traffic movement effectiveness. The study has shown that implementing optimisation tactics can successfully reduce traffic-related collisions and improve road safety by adjusting traffic signal timing to match traffic demand.

2.4.5 Applications, Evaluation and Comparison of machine learning for Accident Prediction Models

Furthermore, Liang et al., (2021) have investigated the use of neural networks, specifically deep learning models, to predict accidents. Research has shown that deep neural networks have the capability to effectively comprehend intricate patterns in accident data. However, attaining optimal performance may require substantial quantities of data and processing resources. Yuan et al., (2020) and Wang and Liu, (2022) have undertaken studies on the utilisation of ensemble techniques, such as random forest and gradient boosting, to forecast accidents. Research has demonstrated that ensemble models often outperform individual algorithms by using multiple learners, resulting in improved accuracy in predictions.

The empirical research thoroughly examines multiple machine learning methods for accident forecasting. Wang et al., (2023) assessed the suitability of regression models, including linear regression and logistic regression, for predicting accidents. The findings suggest that regression models, although they yield comprehensible outcomes, may encounter difficulties in capturing the intricate connections present in accident data. Johnson and Brown (2018) conducted a study to assess the effectiveness of decision tree algorithms, specifically classification and regression trees (CART), in predicting accidents. Studies have shown that decision trees are effective in identifying important decision rules in accident data, making them especially useful for understanding the elements that contribute to accidents.

Machine learning algorithms, which uncover complex relationships between the various elements determining the severity of an injury, can make highly accurate predictions. Li et al., (2023) used a deep learning algorithm, the stacked sparse autoencoder (SSAE), to estimate the severity of injuries in road accidents using the contributing elements. In order to assess the severity of accident injuries, Hijazi et al., (2023) contrasted the Extreme Gradient Boosting (XGBoost) model with other traditional machine learning (ML) techniques. Furthermore, a multi-task deep neural network (DNN) framework was presented by Hijazi et al., (2023) that can simultaneously forecast property damage, injury, and mortality in traffic accidents. However, there are many obstacles to improving the application of machine learning to damage harshness estimate. A lack of readily available data could impair the model's effectiveness, challenges in choosing suitable features, and the potential for data overfitting, which could impair the model's performance and reduce its accuracy in predicting the severity of traffic injuries.

A team of Santos et al., (2021) from Concordia University examined the accidents that happened in Montreal by using a balanced random forest method. The team collected accident data from three publicly available datasets: the National Road Network database, which included information on road segments; Past Environmental Dataset, which included weather-related data; and Montreal Vehicle Collisions (2012–2018). Among the models examined were a baseline model, XGB (XG boost), RF (random forest), and BRF (balanced random forest). The scholars used only 0.1% of the two billion negative samples generated. The researchers made predictions for the next hour in each segment. With a false positive rate (FPR) of 13%, the algorithms' total prediction accuracy for events in Montreal was 85%. Similar to this, a team of researchers from Bangladesh's North South University used a variety of machine learning techniques to gain understanding of and predict the severity of accidents that take place there (Siam et al., 2020). The group made use of 2015 accident statistics, weather, traffic accident severity, and road information. The authors used the agglomerative hierarchical clustering technique to generate homogenous groupings, and then they used a random forest to choose the predictor variables for each cluster. The authors then created prediction rules for each cluster using the decision tree (C5.0) models.

This study then used the decision tree (C5.0) models for each cluster to generate prediction rules. The authors substantiate numerous regulations, one of which asserts a higher risk of fatal accidents on rural, regional, and national roads without dividers. Low visibility and unfavourable weather are other factors that researchers have identified as influencing the likelihood of accidents or their severity (C. Xu et al., 2018). Theofilatos and Ziakopoulos (2018) compared factors within and outside of urban areas. The study found that travel movement and rate changes affect powered two-wheeler (PTW) crashes. Various factors, both inside and outside of metropolitan areas, influence the harshness of crashes. These factors include bicycles, young drivers, intersections, and collisions with objects.

Iranitalab and Khattak (2017) looked at support vector machines (SVM), nearest neighbour classification (NNC), multinomial logit (MNL), and RF analysis techniques side by side to see which one could best predict how bad traffic events would be. Based on the data, NNC demonstrates superior prediction performance compared to RF, SVM, and MNL when it comes to more severe accidents. In order to predict traffic accidents, Kuang et al., (2019) looked into a variety of machine learning techniques, with Bayesian networks, random forests, and K-nearest neighbour. The best model predicted 61% of events correctly, but it had a 38% false alert rate. A CART model was created by C. Chen et al., (2016), who used regression and classification trees to build a classifier. During the training and testing phases, this classifier forecasts accidents with a 55% accuracy rate. To forecast the probability of crashes on multi-lane roads, Saeed et al., (2019) employed Poisson, negative binomial, and negative multinomial regression models.

Silva et al., (2020) state that the most common methods used for these objectives are artificial neural networks, decision trees, evolutionary algorithms, support vector machines, and nearest neighbour classification. Many different contexts use multivariate response models because they can handle both regression and classification problems. The subsequent studies employ deep learning methodologies to forecast traffic incidents. To examine the probability of traffic crashes, Theofilatos et al., (2019) used real-time traffic data from metropolitan arterial highways and Bayesian logistic regression models. Theofilatos et al., (2019) conducted a comparative study of various machine learning and deep learning techniques such as kNN, naive Bayes, classification trees, random forests, SVM, shallow neural networks, and deep neural networks. The study determined that the deep learning methodology produced the most favourable outcomes, whereas less intricate methods such as naive Bayes exhibited somewhat inferior performance. The technique for forecasting the likelihood of road traffic crashes using a long-short-term memory (LSTM) model was presented by Tian and Zhang (2022). According to this concept, the risk is the frequency of accidents in a given area during a certain period of time.

In their investigation of car accidents in Iowa from 2006 to 2013, Acker and Yuan (2018) utilised a ConvLSTM configuration. The Iowa Department of Transportation (DOT) provided the data, which included crash records, rainfall data, reports from the Roadway Weather Information System (RWIS), and supplementary information on speed regulations, yearly average daily traffic (AADT), and traffic camera amounts. From 2006 to 2012, the researcher underwent guidance using reports and then conducted testing in 2013. During the testing phase, the researcher used data from the previous seven days to predict locations for the following seven days. The ConvLSTM model demonstrated greater predictive accuracy in comparison to the other models. Furthermore, the algorithm accurately predicted the case study incidents on December 8, 2013, a day marked by a heavy blizzard. Pradhan and Sameen (2019) performed an extensive review of the prior research on forecasting traffic accidents. The researchers found that the synthesis of multiple data sources, neural networks, and deep learning algorithms results in extremely high levels of accuracy and precision.

Researchers have utilised diverse algorithms in recent years to predict traffic accidents. Noorunnahar et al., (2023) utilised the XGboost algorithm, a tree-based technique, to forecast traffic patterns in Istanbul, the most densely populated city globally, achieving an accuracy rate of 85%. Hebert et al., (2019) utilised the random forest algorithm (RF) to evaluate the frequency of road traffic collisions in Montreal. The analysis included car crash data from 2012 to 2018, as well as historical weather data and the National Road network. The algorithms exhibited a 13% false-positive rate (FPR) and achieved an 85% accurate prediction rate for traffic occurrences in Montreal. Iranitalab and Khattak (2017) did a study to see how well different algorithms, such as multinomial logit (MNL), nearest neighbour classification (NNC), support vector machine (SVM), and RF analysis techniques, could predict how bad traffic accidents would be. Determined that NNC had the highest average forecast accuracy compared to the other two methods: RF, VM, and MNL. Santos et al., (2022) investigated the prediction capacities of three machine learning techniques, specifically RF, K-nearest neighbour, and Bayesian networks, in predicting traffic accidents. The highest level of precision is 62 percent; however, the incidence of false positives is 38 percent.

In addition, Khaldi et al., (2021) developed a data-driven soft sensor that utilises ensemble learning techniques to accurately forecast the velocities of swarm motion. The investigation demonstrates the versatility of artificial intelligence in addressing traffic-related problems by focusing on forecasting swarm motion speeds. Similarly, Khan and Das (2024) conducted a study to implement proactive traffic control and prevent accidents. The geographical characteristics and the accident's precise location are critical factors to consider. Rural areas have a higher incidence of road traffic accidents and fatalities, according to research by Kumar et al. (2020).

Shafabakhsh et al., (2017) employed geographic information system (GIS) data to examine the impact of geographical factors on accidents. In their case study, Acker and Yuan (2018) employed temporal and geographical factors to develop prognostic models. These linked indicators provide crucial data for accident investigations. Scientists utilised a range of machine learning methodologies to forecast traffic accidents. Duan et al., (2019) utilised a decision tree methodology on Apache Spark, achieving a prediction accuracy rate of 91%. Cheng et al., (2019) employed a random forest prediction model to provide accurate forecasts using traffic data. Dogru and Subasi (2018) observed that the random forest model they employed for analysing accident data patterns outperformed earlier models in terms of detection rate, accuracy, and sensitivity.

In 2020, the Namibia Road Safety Research Group conducted a study to assess the practicality of integrating animal crossing data into accident prediction algorithms. Their empirical investigation emphasises the importance of modifying models for regions with a higher prevalence of wildlife-related accidents. Furthermore, the Namibia Ministry of Works and Transport (2019) and Brown and Jones (2021) conducted research on modifying models to suit the unique features of local road conditions, such as gravel roads and unpaved surfaces. These experiments highlighted the necessity of using region-specific road data to improve the precision of projections.

An empirical investigation was conducted on accident prediction systems using historical accident data. Garcia et al., (2020) and the Namibia Road Safety Council (2020) evaluated the accuracy, precision, recall, and F1-score of accident prediction models. Based on their research, these models exhibit more precision in comparison to earlier techniques. Furthermore, the Namibia Institute of Transport Studies (2018) and Cao et al., (2017) conducted empirical research to compare machine learning-based systems with well-established accident prediction methods.

Various investigations continuously show that machine learning models surpass conventional methods in terms of predicting accuracy.

Table 2.1: The short summary of empirical literature review

Past studies By:	**Describe the experiment that was performed**	**Describe the ML classifier that was used**	**Describe the dataset that was used**	**Describe the results that were used achieved**	**Describe the evaluation metrics that were used to determine the model performance**	**Describe any notable limitations of the study**
Ahmed & al.(2023)	The study compared traffic accident severity prediction machine learning methods.	The researchers tested numerous models, including Random Forest (RF), Decision Jungle (DJ), AdaBoost (AB), XGBoost, Light Gradient Boosting Machine (LGBM), and CatBoost.	The study utilised traffic accident data from New Zealand, covering the years 2016-2020.	Among the models examined, the Random Forest algorithm displayed the most performance, with an accuracy of 81.45%.	The research predominantly employed accuracy as the criterion for assessing model performance.	The study did not specify any notable limitations.
Muktar, B., & Fono, V. (2024).	The research aims to create a machine learning-driven web solution for forecasting the severity of traffic accidents in Montreal. Its mission is to identify critical factors affecting accident severity and offer insights for targeted actions to improve road safety and emergency response.	The research employed categorisation techniques such as eXtreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), Random Forest (RF), and Gradient Boosting (GB).	The dataset comprised traffic accident records from Montreal spanning 2012 to 2021, obtained from the Société de l'Assurance Automobile du Québec (SAAQ).	The models were assessed based on precision, recall, F1-score, and accuracy, with XGBoost attaining the greatest accuracy at 96%, followed by CatBoost at 95%, Random Forest at 93%, and Gradient Boosting at 89%.	The study primarily used confusion metric to evaluate model performance.	Not specified
Meocci, M., et al. (2023)	The researchers created a predictive model designed to evaluate the severity of traffic incidents in Italy. The main aim was to pinpoint high-risk zones and elements leading to major	The research utilised a Gradient Boosting technique to develop the predictive model. This ensemble learning method amalgamates the predictions	The model was trained and validated utilising the ISTAT dataset, which contains comprehensive records of traffic incidents throughout Italy.	The Gradient Boosting model shows outstanding efficacy in forecasting the severity of road accidents	The performance of the model was assessed using accuracy metrics, indicating the proportion of correct predictions made by the model.	The study did not specify

	accidents, thereby improving road safety protocols.	of several weak learners to enhance accuracy and resilience.				
M. et al.(2022),	The study evaluated collision severity and hotspots in Saudi Arabia to identify crucial characteristics and places connected with serious accidents, giving significant insights for targeted interventions.	The research utilised various machine learning classifiers, such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), to assess and categorise accident severity.	The study did not specify the dataset	The results demonstrated that the RF model outperformed other classifiers in predicting crash severity, achieving the highest accuracy and robustness in classification.	The evaluation measures used to determine model performance included accuracy, precision, recall, the F1-score, and Area Under the Curve (AUC).	The study did not specify any limitation
Makaba et al. (2020)	The study created a methodology leveraging Bayesian networks to estimate the financial effects of road traffic collisions.	Bayesian Network.	The study constructed a Bayesian network architecture using actual road traffic collision data.	The framework effectively modeled the complex relationships between various factors contributing to road traffic collisions and their associated costs	Specific evaluation metrics are not detailed	The study did not specify limitations in the available sources.
Lykov and Asakura (2020a)	The researchers used tensor-robust principal component analysis in conjunction with continuum modelling to identify anomalous traffic patterns in extensive urban regions.	Tensor Robust Principal Component Analysis.	Specific dataset details are not provided.	Their analysis shows that the tensor-based approach can detect and rectify traffic anomalies, thereby improving safety.	Specific evaluation metrics are not detailed in the available sources.	No specify limitations in the available sources.
Jain (2011)	The study investigated the application of fuzzy logic in traffic signal control automation	Fuzzy Logic Controller.	No data-set used was specified	The results demonstrate that the fuzzy logic-based traffic signal control system significantly enhances traffic flow and reduces congestion-related accidents by	No specific evaluation metrics was mentioned	The study does not specify limitations

				dynamically modifying signal timing.		
Liang et al. (2021)	The research examined the application of neural networks, particularly deep learning models, to forecast accidents. Studies indicate that deep neural networks can proficiently analyse complex patterns in accident data.	Deep Neural Networks.	No dataset was specified	Studies indicate that deep neural networks can proficiently understand complex patterns in accident data.	No specific evaluation metric was mentioned	Attaining optimal performance may require substantial quantities of data and processing resources.

2.4 Research Gap

While many studies worldwide employ machine learning techniques to analyse and predict crash severity, there is an outstanding gap in Namibia. Research focusing on uncovering patterns and predicting crash severity using modern approaches remains limited, as most studies in the Namibian context rely on traditional statistical methods or descriptive analyses. For instance, Matengu et al., (2022) examined the socio-demographic characteristics of drivers involved in accidents and identified key factors such as speeding and alcohol consumption. Similarly, data from the Namibia Motor Vehicle Accident Fund (2023) highlights efforts to reduce pedestrian-related crashes through interventions like education, enforcement, and engineering. Despite these efforts, the methodologies used predominantly focus on statistical summaries, leaving room for advanced techniques. Therefore, the current studies aim to employ machine learning techniques on Namibian's data to be able to reveal actionable trends, uncovered patterned and inform data-driven strategies tailored to Namibia's unique road safety challenges.

2.5 Chapter Summary

Despite extensive research on road traffic accidents and the application of various machine learning algorithms to forecast their outcomes, there is still a significant lack of research that combines historical accident data with environmental variables, such as weather conditions and road surface characteristics, to create prediction models. Namibia serves as a prime example of a situation marked by a dearth of research on integrating local elements into prediction models. The aim of this project is to develop a customised machine-learning system that precisely addresses the challenges associated with Namibia's road network. The system will utilise a combination of environmental variables and historical accident data to accurately identify factors with the highest probability of accidents occurring thereby enhancing the precision of forecasts on the timing of these events. The goal of this study is to address these deficiencies and improve the effectiveness of accident prevention programmes by providing critical information to road safety authorities.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1. Overview

This chapter provides an overview of the research technique used to create a machine-learning accident prediction model for the road network in Namibia. The research methodology consists of several crucial stages, such as analysing the target population, collecting data, evaluating both dependent and independent variables, pre-processing the data, employing statistical and machine learning forecasting techniques, building models, and evaluating and predicting the performance of these models. Below are the two figures demonstrating the overall view of the methodology.

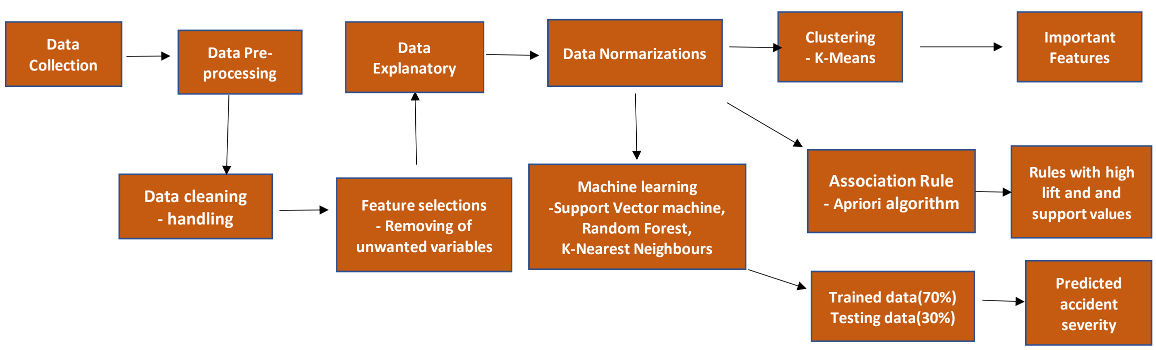


Figure 3.1: The general flow diagram for the Methodology

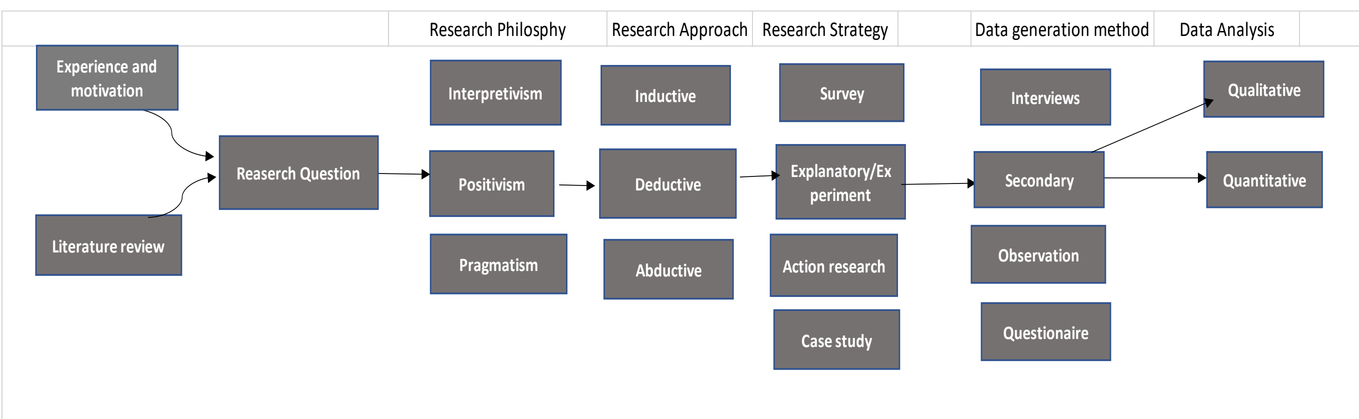


Figure 3.2: Model of the Research process (Saunders et al., 2019);(Oates, 2005)

3.2 Research Philosophy

The research philosophy employed in this dissertation is positivism with the aim of focusing on noticeable phenomena and quantitative data to be able to develop predictive models. The methodology aligns with the objective of identifying patterns or factors from historical road accident data that are associated with one another.

3.3 Research Approach

The study involves applying machine learning techniques to historical data to test established theories (crash prediction models and accident severity classification); therefore, the study used a deductive method in order to validate the hypothesis, and also explore theories on the factors that influence the severity of accidents and assess the effectiveness of models including support vector machines (SVM), random forests (RF), and KNN.

3.4 Research Strategy

The context of the research strategy is explanatory. The aim was to predict and describe the variables affecting traffic accidents in Namibia and to identify the underlying connections and patterns in the accident data to motivate this strategy.

3.5 Choice(s)

The study utilised a mixed methods approach, and this is because the study made use of both quantitative and qualitative data.

3.6 Time Horizon

The current study adopted a longitudinal time horizon, analysing data covering five-year period (January 2018 – December 2023) to allow identifying trends and change over time.

3.7 Techniques and Procedures

3.7.1 Study Population

Verschuren and Doorewaard (2010) define a research population as the actual portion of reality that a researcher wishes to investigate. This study utilised historical accident data from Namibia as its population.

3.7.2 Data source

This study employed both quantitative and qualitative secondary data collected from the Namibian National Road Safety Council (NRSC) and MVA Fund over a period of five years. The researcher encompassed all the regions of Namibia in the data set, starting on January 1, 2018, and ending in December 2023. While the time of 2018 - 2023 is considered recent, the conclusions and prediction models derived from it depended on existing trends and patterns to provide accurate analysis and projections. Improvements to infrastructure and the implementation of security measures may have impacted Namibia's road safety level, necessitating more up-to-date statistics.

3.7.3 The dependent variables

The crash priority is a critical factor in determining the magnitude of an event. The dependent variable in this study enabled the researcher to examine and comprehend the elements that influenced varying degrees of accident severity. Smith et al., (2017) showed that classifying incidents as "slight" or "severe" yields important insights.

3.7.4 Independent variables

The independent variables incorporate a broad range of elements that may influence crash severity. The main explanatory factors used in this study are listed in Table 1.

Table 3.1: List of independent variables considered in the study.

Variables	**Description**	**Types of Data**
Date Occurred	The date of the accident.	Date
Day of Week Occurred	The day of the week when the accident occurred.	Categorical
Number of Fatality	The count of fatalities in the accident.	Numerical
Number of non-injured	The count of individuals who were not injured in the accident.	Numerical
Number of Injured	The count of individuals who sustained injuries.	Numerical
Number of Persons Trapped	The count of individuals trapped in vehicles.	Numerical
Crash Time Occurred	The time of day when the accident took place.	Time
Crash Cause	The cause or reason behind the accident.	Categorical
Region	The geographical region where the accident occurred.	Categorical
Condition of Road Sign	The condition of road signs at the accident location.	Categorical
Junction Type	The type of junction or intersection.	Categorical

Light Condition	The lighting conditions at the time of the accident.	Categorical
Obstruction	Any obstructions present at the accident scene.	Categorical
Road Direction	The direction of the road at the accident location.	Categorical
Road Flat/Sloped	The terrain of the road (flat or sloped).	Categorical
Road Mark Condition	The condition of road markings.	Categorical
Road Mark Type	The type of road markings.	Categorical
Road Signs Clearly Visible	Whether road signs were clearly visible or not.	Categorical
Road Surface	The condition of the road surface.	Categorical
Road Surface Quality	The quality of the road surface.	Categorical
Road Surface Type	The type of road surface (e.g., asphalt, gravel).	Categorical
Road Type	The type of road (e.g., highway, rural road).	Categorical
Severe Wind	Whether severe wind conditions were present during the accident.	Categorical
Speed Limit on Road	The designated speed limit on the road.	Numerical
Traffic Control Type	The type of traffic control (e.g., traffic lights, stop signs).	Categorical
Visibility	The visibility conditions at the time of the accident.	Categorical
Weather	The prevailing weather conditions.	Categorical
Vehicle Make	The make or manufacturer of the vehicle involved.	Categorical
Vehicle Model	The model of the vehicle involved.	Categorical
Vehicle Type	The type of vehicle involved.	Categorical

Town	Where the accident occurred.	Categorical
Crash Type		Categorical

3.7.5 Data Pre-processing and Feature Engineering

The study undertook an exacting procedure for feature engineering and pre-processing to prepare the collected data for the training and testing of machine learning models. The additional goal of this approach was to extract essential information from the dataset, consequently improving the accuracy of our accident prediction models. Pedregosa et al., (2011) determined that data preparation responsibilities encompass managing missing values, transforming categorical variables into a suitable format, and validating the data's integrity. Feature engineering covers several types of approaches, such as feature scaling, one-hot encoding, and feature selection (Li et al., 2017).

The pre-processing involved addressing missing values by eliminating all null or unknown rows. Prior to the elimination of rows and columns, the percentage of missing data for each variable was calculated to assess the extent of missingness. Variables with over 50% missing values, such as suburbs and street names, were removed, while those with less than 50% underwent imputation. Feature selection was conducted to focus solely on the essential variables required for the research, as demonstrated in Table 3.1 above. Irrelevant columns, including crash number, crash name, and vehicle purpose, were eliminated. Non-numerical variables, including day of the week, vehicle types, and weather, etc , were initially encoded (0, 1, 2, … n), while certain numerical variables, such as crash time, were first categorised (e.g., 4-6 early morning, 6-12 morning, 12-17 afternoon, 17-20 evening, 20-00 midnight) and subsequently encoded from 0 to 5 for analytical purposes in machine learning and association rule.

3.7.6 Machine learning models

This section outlines the machine learning methodologies and association rule mining strategies that were utilised for predicting road accidents.

3.7.6.1 Support Vector Machines

SVMs are supervised learning models that utilise associated learning algorithms for classification and regression analysis. Vladimir Vapnik developed the support vector machine (SVM) in 1992, as detailed in his 1995 book. Vladimir Vapnik designed this technique to predict and classify both linear and non-linear data. The objective of support vector machines (SVM) is to identify the hyperplane with the largest possible margins. The maximum margin hyperplane is the one that provides the most distance between the decision classes for separation. This study refers to the training instances closest to the hyperplane with the widest margin as support vectors (Y. Xu et al., 2006). The support vector machine (SVM) is highly proficient in managing classification jobs involving multiple classes (Hsu et al., 2011). This was consistent with the accident severity categories, which included moderate, slight, and severe.

The SVM system can be absolutely defined mathematically as follows.

Consider a training set Q, denoted as $Q = \{xi, yi\}_{i=1}^{N}$ $with\ input\ vector\ xi = \{xi^1, \dots, xi^n\}^T \in R$ and target labels yi be in the range of -1 to 1.

According to the Vapnic Formula, must satisfy the following conditions:

$$\begin{cases} W^T \emptyset(xi) + b \geq 1, if\ yi = +1 \\ W^T \emptyset(xi) + b \leq 1, if\ yi = -1 \end{cases} \quad (1)$$

Which is equivalent to?

$$yi[W^T \emptyset(xi) + b] \geq 1, i = 1 \quad (2)$$

Where the weight vector (maximum margin) and b is the bias

3.7.6.2 Random forests (RF)

Breiman and Cutler (2001) established random forests, a supervised learning method and a non-linear model adept at handling both regression and classification tasks. This technique uses bootstrap sampling to build an ensemble of decision trees by altering the training dataset. During the node training phase of each tree, a predetermined ratio randomly selects features from the original dataset without replacement. People recognise random forests for their robustness and accuracy, particularly in managing large datasets. They can effectively handle absent data without significant loss of forecast accuracy. Breiman, (2001) states that the random forest algorithm learns from bagged datasets of randomly chosen features, explaining the calculation of feature importance. The Random Forest analysis integrates the results from various decision trees to create a predictor that is more robust against randomness and less susceptible to overfitting. The method is predicated on two fundamental ideas: random feature selection and bagging. In beginning, new training sets with sample sizes of n each are randomly selected from the original training set using replacement sampling for a specific training dataset of size n. The freshly created training set is then used to train K-base trees. To guarantee variation among the foundation trees, the researcher also chose at random a subset of attributes for each tree's growth in the Random Forest. The splitting process in the construction of a single decision tree continues unabated until the tree descends as far as it can.

The procedure for constructing a random forest entails the following steps:

- The bootstrapping sample technique produces K training sets from the original dataset.
- Training each tree on a distinct subset of the data and selecting a random collection of characteristics for each division in a tree.

- For k iterations, repeat step 2 to produce k decision trees.
- The random forest receives new inputs, and each tree produces a predictive outcome. A majority vote among all decision trees determines the ultimate classification outcome.

Cheng et al., (2020) found three principal parameters affecting the tuning efficacy of random forests: the total number of trees (n_estimators), the quantity of characteristics used for each node split (max_feature), and the maximum depth of the trees.

3.7.6.3 K-Nearest neighbours (KNN)

K-Nearest Neighbours (KNN) is a supervised machine learning method utilised for classification and regression applications (Peterson, 2009). KNN functions as a distance-based technique, preserving all instances associated with training data points in an n-dimensional space. Upon acquiring a new, unfamiliar data point, KNN use distance metrics to ascertain the k nearest instances and predicts the class of the data point by determining the most dominant class among the neighbours. The method computes the mean of the k nearest neighbours to get a prediction for continuous, real-valued input.

The algorithm pertaining to KNN is delineated as follows:

1) Enter the item and classify it as

   $(\boldsymbol{x_i y_i}) = \boldsymbol{items\ i\ recieved\ from\ decision\ tree}$

2) Proceed to the next level and acquire the subsequent item as

   $(\boldsymbol{x_{i+1}, y_{i+1}}) = \boldsymbol{item_{i+1}}$

3) Calculate the Euclidean distance between items.

   $\mathbf{euc_i} = \sqrt{(\mathbf{x_{i+1}} - \mathbf{x_i})^2 - (,\mathbf{y_{i+1}} -, \mathbf{y_i})^2}$

4) $\mathbf{ifeuc_i} \leq \mathbf{Cluster_i}$
5) $\mathbf{Cluster_i} = \mathbf{item_i}$
6) $\mathbf{endofif}$

3.7.6.4.K-means techniques

K-means is a simple unsupervised machine learning that addresses the clustering issue by separating a dataset into a specified number of clusters. The main goal is to create K centers, or centroids, which represent each cluster. The centroid represents the approximate center of a cluster (Azis & Atmajaya, 2016). It is imperative that one carefully selects these centres, as different placements may produce divergent outcomes. Align the centroids at the maximum separation to attain optimal results. Purwaningsih (2019) claims that the K-means approach groups data with analogous features into a single cluster while categorising data with different attributes into separate clusters. The formulas for Euclidean distance are as follows:

$$d = |x - y| \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

Were,

d = calculation of distance to the centre of the cluster

x = point coordinates of the object

y = centroid coordinate

$$\sum (i = 1)^n \ = n \text{ is the amount of data to be measured the distance, while } i$$
$$= \ 1 \text{ is the clustering process starting from the first iteration}$$

x_i, *coordinate point of the i object*

$y_i = i$ *centroid coordinate point*

3.7.6.5 Association rule

Association rule mining such as Apriori algorithm is a recognised data mining algorithm for identifying potential patterns in large datasets. Association rule mining is adaptable among other machine learning techniques due to its lack of a specified function and absence of dependent variables. ARA delineates the correlation between sets of objects according to a specified feature A and a feature B. In other terms, the presence of attribute A correlates with X% probability of simultaneously possessing attribute B, expressed as "A implies B" or A → B (Hsu & Chang 2020).

If incident A takes place, event B immediately occurs in C% of cases, and this pattern recurs in S% of all occurrences within the dataset.

Three parameters govern the generation of rules: support (S), confidence (C), and lift (L).

1. **The confidence value (C)** assesses the rule's truth, computed as the proportion of facts that include both point A and point B to those that include only attribute A. The confidence quantity represents the likelihood of B occurring given the presence of A.

$$C(A \to B) = P(A/B) = \frac{Supp(A \cup B)}{Supp(A)} \qquad (4)$$

2. **Support value (S)** represents the frequency of events that include both attribute A and attribute B. The support value signifies the likelihood of A and B occurring simultaneously.

$$S(A \rightarrow B) = P(A \cap B) \quad (5)$$

When support and confidence levels are sufficiently high, one can say that any future example containing A is going to also include feature B. It is essential to assess the degree of independence between feature A and feature B to prevent the emergence of "spurious" rules when both support and confidence are elevated (Makarova et al., 2020). The lift value is necessary as a metric to assess the impact of association rules.

3. ***Lift Value (L):*** Confidence and the likelihood of B showing up by itself are compared:

$$L(A \rightarrow B) = \frac{C(A \rightarrow B)}{S(B)} \quad (6)$$

The rule's items are likely to group together if the lift value is greater than or equal to 1, indicating a positive association. A lift value less than 1 indicates a low likelihood of grouping the rule's items together, implying a negative association. For a rule to be considered legitimate, the value must be at least 1

3.8. Model performance evaluation

Several parameters have been encompassed in the model evaluation. Based on the confusion matrix, the prediction was assessed using the following four metrics: F-measure, accuracy, precision, and recall.

3.8.1 Confusion Matrix

A confusion matrix is a technique that compares a classification model's estimates against the data's actual labels in order to assess how well it performs. This includes false negatives (FN), true negatives (TN), true positives (TP), and false positives (FP). According to Powers (2020a), TP and TN denote accurate predictions, but FP and FN denote inaccurate ones.

Table 3.2: Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

3.8.2. Accuracy

Accuracy is determined by estimating the percentage of correctly classified cases in the validation dataset. The maximum accuracy is 1, while the minimal is 0. In the formula, the numerator represents the correct predictions (true positives and true negatives), while the denominator contains all the algorithm's forecasts (right as well as wrong ones).

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

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (7)$$

3.8.3. Precision

A statistical metric called precision is used to assess how well a classification model performs, primarily in dual classification issues. It is represented as the proportion of accurate positive forecasts among all of the model's positive predictions. It is used as a metric to evaluate how good the accident prediction model performs. Think of it like a detective trying to crack a case. For the detective, gathering a significant amount of evidence is insufficient. Additionally, they must verify that the material they have gathered is pertinent and leads them in the correct direction. It will be more difficult to solve the case if they are dealing with a large number of false leads. It can be computed as:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

A high accuracy indicates that the model predicts few positive events incorrectly, as indicated by a low wrong positive rate. Conversely, low precision indicates that the model consistently forecasts positive results.

3.8.4. Recall

Recall metric, often called the measure of completeness, reflecting the percent of correct classified true positive tuples. It is determined by:

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Recall measures how well the model distinguishes the real positives from all other positive occurrences in the data. A model with a high recall accurately identifies the majority of positive cases and has a low rate of false negatives. Conversely, a low recall indicates that many successful cases are missing by the model. In situations where the cost of false negatives is large, like in fraud detection or medical diagnosis, recall is a helpful metric. A false negative can lead to missed diagnoses or fraudulent activity remaining undetected (Powers, 2020b).

3.8.5. F-Measure/F1-Score

Using both the classifier's recall and precision, F-score, is a statistical assessment metric for category, calculates a score between 0 and 1. This metric serves as a valuable compromise, acknowledging the importance of both recall and precision. It is particularly useful because it represents the harmonic mean of these two conflicting metrics, blending them into a single value as represented in the formula.

$$F - Measure = \frac{2(precision \times recall)}{precision + recall} \tag{10}$$

An F1-score of 1 indicates perfect precision and recall, whereas a score of 0 indicates that the model's predictions are completely incorrect. The distribution of the classes is unbalanced. At that point, there are many more instances of one class than the other. In particular, the F1 score is useful. Because a model that only predicts the majority class will be highly accurate in certain circumstances, accuracy may be a misleading metric. Because it takes into consideration both recall and precision, the F1-score offers a more insightful assessment of the model's performance (Powers, 2020).

3.9 Summary of the research methodology

Inclusion: by following and applying the methodology as outlined above, it is aligned with the research agenda to address traffic safety in Namibia broadly by providing a data-driven approach to uncover critical insights and propose actionable recommendations. The research was able to identify and compare the performance of the models in predicting severity and to identify which factors contributed more and what types of crashes were high in Namibia that increased accidents by using analytical tools such as clustering, association rule, RF, SVM, and KNN.

Table 3.3: The table below briefly outlines how each research question is addressed through tailored methods and tools

Research Questions	Tools	Outcome
How can a reliable machine learning model be developed to predict road accidents in Namibia and enhance road safety using historical accident data?	Data collection, cleaning, and pre-processing of historical accident data; ML model development and evaluation (Random Forest, SVM, KNN)	A machine learning model with the highest predictive accuracy is selected.
What are the major causes of traffic accidents in Namibia?	Data exploration and analysis using statistical methods clustering techniques, apriori rule	Can be either environment factors or human behaviours or both
What are the predominant types of traffic accidents?	clustering models, and visualisation of crash types	
Which machine learning model is most effective in predicting road accidents?	Model comparison using metrics like accuracy, precision, and recall (Random Forest, SVM, KNN)	A machine learning model with the highest predictive accuracy is selected.

Source (Author)

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Introduction

The study's methodology was the subject of Chapter 3. Chapter 4 establishes the findings of the study and further supports the research findings by looking at other findings of previous researchers to determine whether they are in agreement or contradiction and to highlight any unique insights or understandings derived from this investigation. The chapter covers the explanatory analysis and the use of machine learning models. Using data for Namibian road traffic accidents from 2018 to 2023, this chapter describes empirical findings and analysis about areas with high accident rates, patterns of when accidents happen, and important factors that contribute to those patterns. The primary focus is to evaluate road traffic accidents and to forecast on crash severity with the aid of Random Forest, Support Vector Machines, KNN, K-Means Clustering, Association Rules, and model evaluation and prediction.

4.2 Number of Accidents per Year

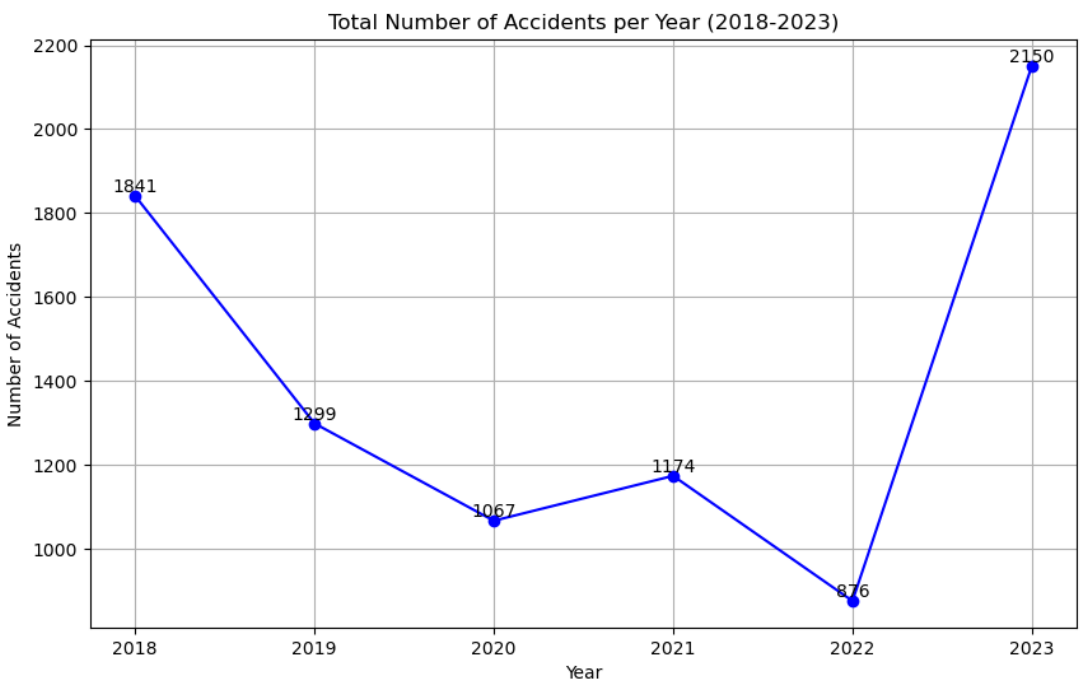


Figure 4.1: Total Number of Accidents per year (2018-2023)

Figure 4.1 illustrates the yearly rate of accidents over a six-year period, highlighting vehicle accidents trends in Namibia from 2018 to 2023. In 2018, there was an important spike in accidents, hitting 1,841 events. Namibia recorded the highest rate of road accidents in Africa in 2018, with around 23.9 accidents per 100,000 people, according to a 2020 Xinhu report (Adanu, E. K., Jones, S., & Odero, K., 2020).

The line graph illustrates a fluctuating trend of traffic accident rates in Namibia from 2018 to 2023, with significant drops seen between 2018 and 2020. The enforcement of limitations during the COVID-19 epidemic, which impacted transport, likely resulted in a general decline in traffic incidents in 2020. The COVID-19 pandemic led to a 42.05% decrease in road accidents mostly due to lockdown measures and travel restrictions that substantially decreased traffic levels in 2020. Ibarra-Rojas et al., (2021) also reported a significant reduction in vehicle accidents under COVID-19 limitations across many nations. Following the normalisation of traffic loads after the pandemic, the incidence of road accidents started to increase again from 2021, reaching its highest point in 2023. This increase highlights the need for ongoing and adaptive road safety initiatives. Garcia et al., (2020) emphasised the need for road safety initiatives to be both enduring and flexible with changing circumstances.

4.3 Number of Accidents per Month for 2018-2023

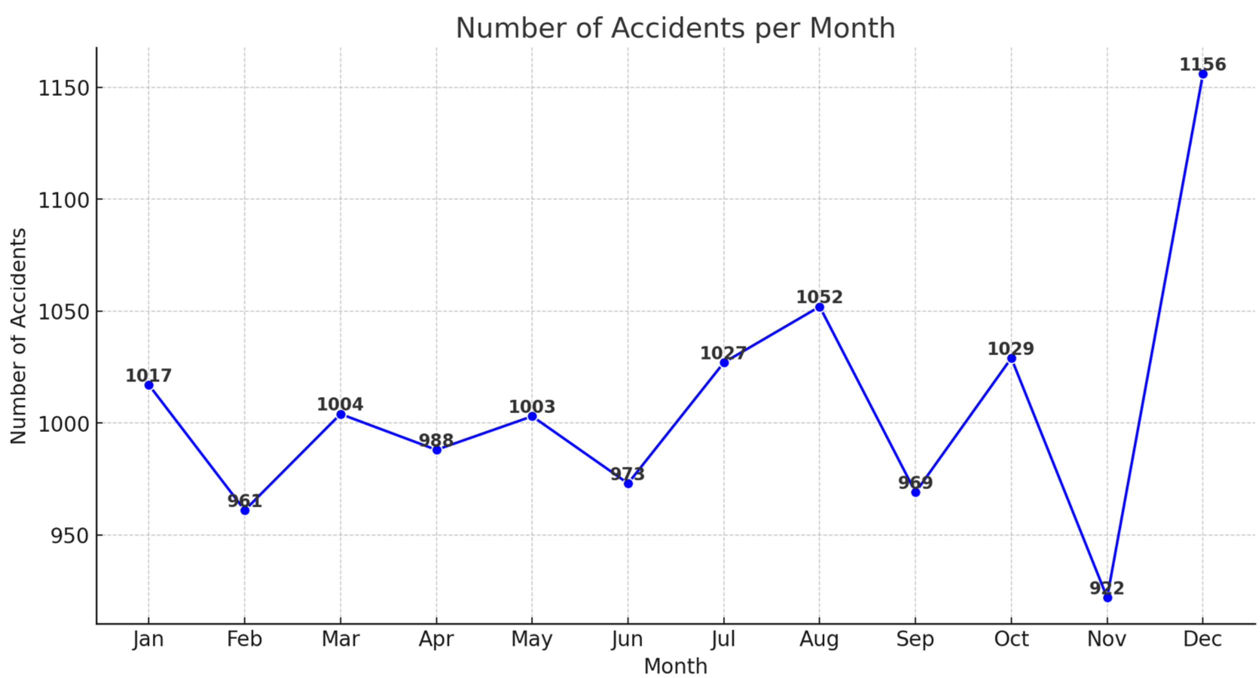


Figure 4.2: Number of Accidents per Month for 2018-2023

The seasonal review in Figure 4.2 indicates that December (1156), January (1071), March (1004) and August (1052) have the highest accident rates and this can be because of an increase in travel frequency during the New Year and festive season due to last-minute shopping, holiday gatherings, and family events. Adanu (2020) noticed a similar pattern, highlighting that holidays usually lead to a raised risk of accidents due to elevated levels of traffic and the potential for driving while intoxicated. Wang (2021) also found a significant increase in car crashes during holiday periods, attributed to risky behaviours like speeding and driving under the influence of alcohol or drugs. The Easter holiday, which typically falls in March or April, is likely to have contributed to the spike in accidents in March. Furthermore, adverse weather conditions, particularly in the rainy season, contributed to the rising accident rates. Lee et al., (2018) found that accidents are more likely to occur on wet roads because drivers experience reduced visibility and slippery conditions. Chen and Zhang (2019) emphasised that precipitation during these months can present a risk on roads, especially for novice drivers.

4.4. Number of Accidents by Day of the Week over the period of 5 years (2018-2023)

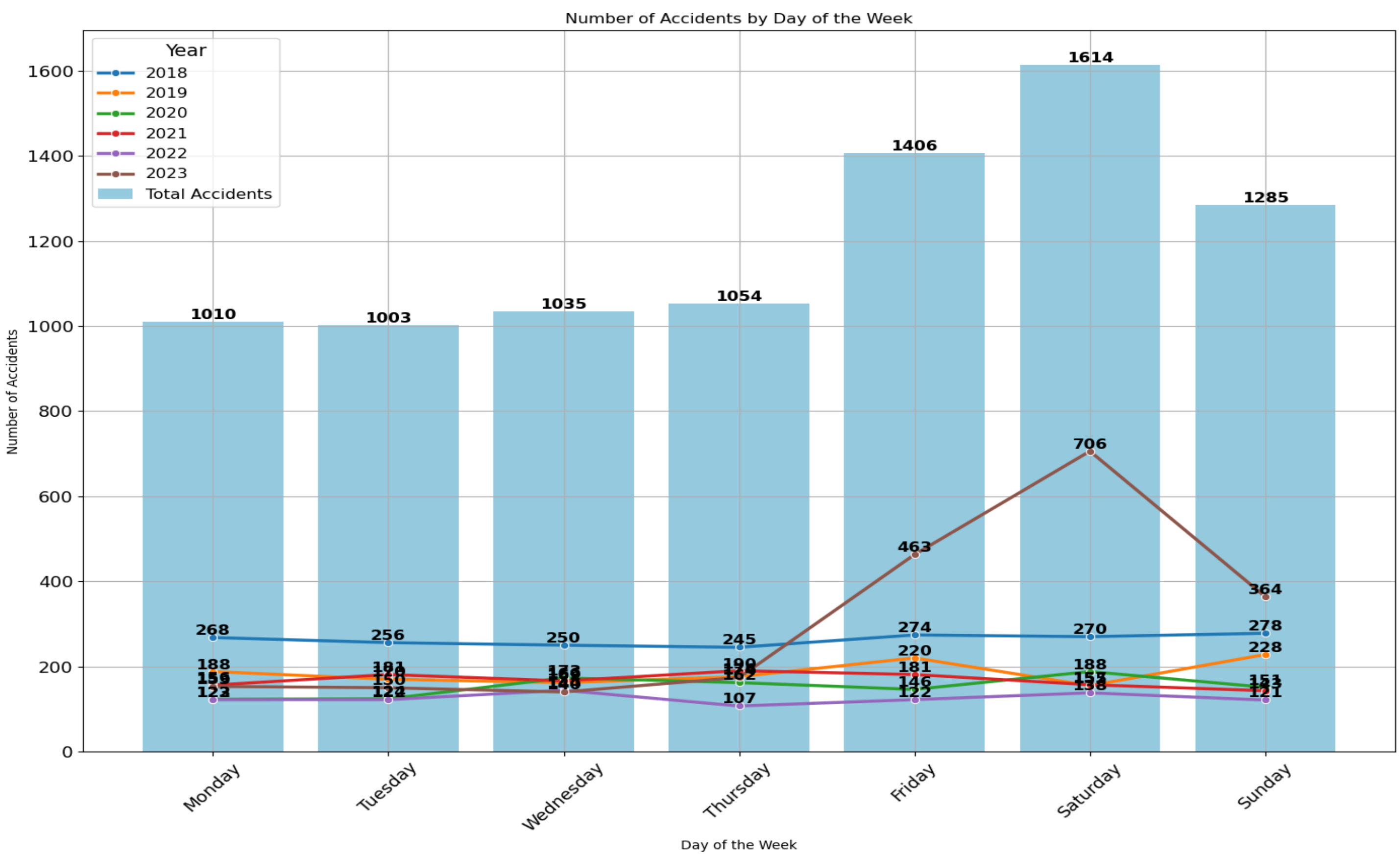


Figure 4.3: Number of Accidents by Day of the Week over the period of 6 years (2018-2023)

The chart 4.3 above displays the number of incidents on each day of the week from 2018 to 2023. The bar chart indicates the total daily accident count, whereas the line graphs reflect the annual accident trends over time. Statistics indicate that Saturdays and Fridays experience the highest number of accidents, with 1,614 accidents on Saturdays and 1,406 on Fridays. The pattern indicates that weekends, particularly late Fridays and Saturdays, pose significant risks. Increased travel, social gatherings, and the likelihood of driving under the influence during weekend recreational activities are likely leading to higher accident rates. Weekdays exhibit a lower incidence of accidents, with Monday through Thursday routinely recording similar figures (about 1,003 to 1,054). This signifies the more consistent commute patterns observed during the workweek. We expect consistent traffic volumes, with a minor influx of vehicles for work or school commutes. Despite a minor reduction in traffic on Sundays relative to Saturdays, the incidence of accidents remains high, especially in 2023, which had a total of 364 documented events. This indicates that the danger remains heightened owing to weekend activities and the potential for those coming home. The consistent occurrence of increased weekend accident rates throughout the years underscores the imperative for targeted road safety measures and enhanced law enforcement during these crucial periods. To successfully diminish the incidence of accidents, especially on Fridays and Saturdays, it is crucial to align public awareness initiatives with traffic enforcement strategies in accordance with these patterns. Despite the steady accident rates on weekdays indicating a decreased danger, it is imperative to sustain monitoring and safety protocols to ensure security.

Therefore, on average, more accidents happen during the weekends than weekdays. The findings indicate that Fridays and Saturdays consistently experience the highest accident rates, whereas weekdays show more consistent accident numbers. Saturday has the highest rate of accidents, making up 19.4% of all accidents. This finding aligns with Isaksen and Johansen's (2021) study that emphasised weekends as a period where road accidents are more common, attributed to factors such as increased leisure activities, social events, and alcohol consumption. It also supports Zhang et al.,'s (2020) findings that highlight how weekend accident risks rise due to higher traffic levels, more leisure activities, and reduced caution among drivers. In addition, findings are also in agreement with that of Nangombe and Neema (2012), in their study of the statistical analysis of road traffic fatalities in Namibia from 2007 to 2009. On the effect of the day of the week on fatality in a vehicle collision in Namibia, they found out that, Sundays and Saturdays were recording the highest number of fatalities with 147 and 146, respectively.

The number of accidents occurring on weekdays shows little variation from 2018 to 2023. For example, there are between 1,000 and 1,050 accidents reported annually on Mondays through Thursdays, indicating the regularity of weekday traffic patterns and commute habits. This corresponds to the study conducted by Abdel-Aty et al., (2018), who discovered that crashes on weekdays typically result from regular traffic patterns during peak hours, exhibiting less variability compared to weekends. The study by Chen and Li (2019) also suggests that weekend trips carry greater danger due to drivers being more likely to engage in risky actions such as speeding or driving under the influence of intoxicating substances.

4. 5. The average number of Accidents per hour over the period of 5 years

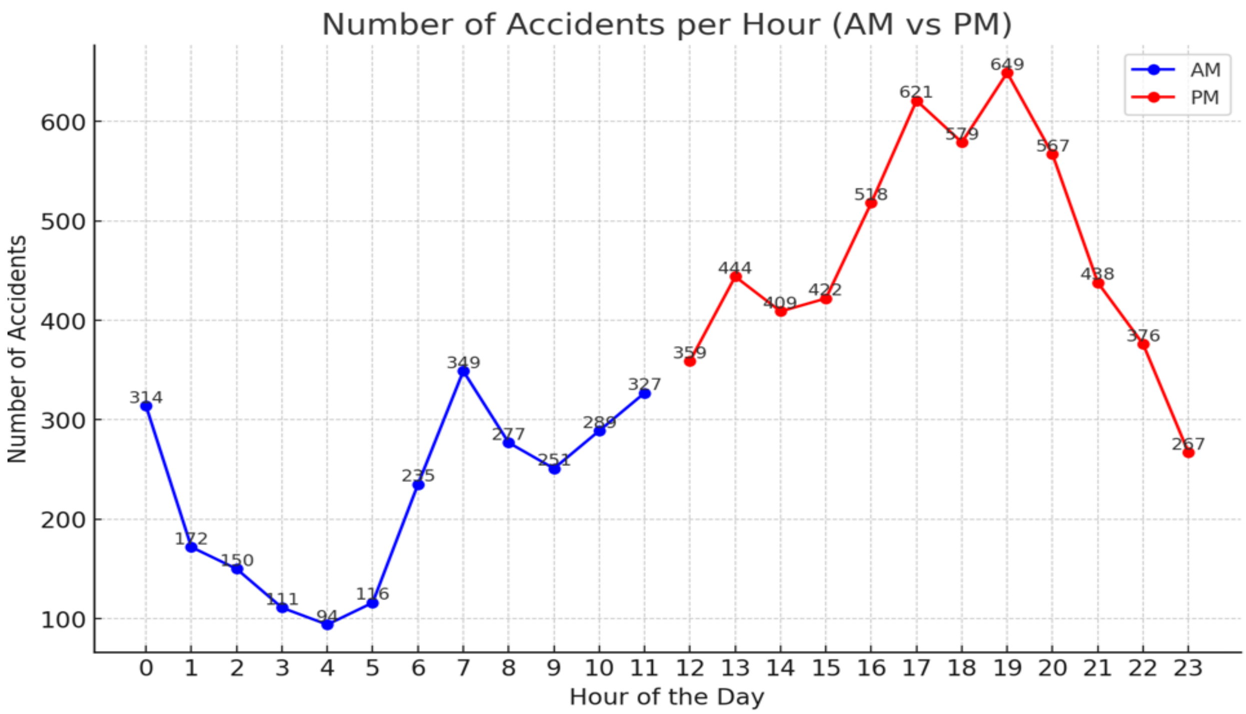


Figure 4. 4: The average number of Accidents per hour over the period of 6 years

Figure 4.4 displays the mean frequency of crashes happening each time during the day. The data showed a higher number of accidents at midnight, reaching a total of 314 accidents. This could be attributed to late-night socialising, restless driving, and the likely influence of alcohol or exhaustion. In the early morning hours, the occurrence of accidents drastically reduces, reaching its lowest point around 4:00 AM with 94 accidents. The researchers estimate this drop mostly due to the reduced quantity of automobiles on the highway during that time. As the day begins, the incidence of accidents climbs substantially, peaking at 7:00 AM with 349 occurrences. This rapid rise happens during the morning rush hour as more people are going to work and dropping off schoolchildren. When traffic grows, it raises the likelihood of accidents. After this peak, the accident rate remains reasonably consistent but slightly reduced, reflecting the typical patterns of early morning driving. The incidence of accidents increases in the afternoon, reaching a peak of 359 incidents at 12:00 PM and extending into the evening. At 5:00 PM, there is a peak of 649 accidents, showing the highest occurrence of events for the day. This rise coincides with the evening rush hour when folks return home from work or school, resulting in traffic congestion and an elevated risk of accidents. Subsequent to the peak, the accident rate falls progressively but stays notably high until late evening, with 488 accidents reported at 8:00 PM. The drop continues overnight, with the accident total down to 267 by 11:00 PM.

The study established that accidents increased significantly during the busy morning rush hours between 6:00 AM and 8:00 AM, with the highest occurrence at 7:00 AM, and this increase in traffic congestion and dangerous driving conditions is caused by more people driving to work or taking their children to school. Wang et al., (2023) discovered similar outcomes, suggesting that increased risks are associated with rush hour periods because of congested roads and tired drivers. Abdel-Aty (2018) also mentioned that the higher chance of accidents during morning commutes is caused by hurrying, pressure, and diversions. Between 4:00 PM and 7:00 PM, accidents significantly increase during the evening rush hour, reaching the peak at 6:00 PM. The rise is due to an increase in the number of vehicles on the road as individuals commute back from work or participate in post-school events. According to Chen et al., (2020), driver fatigue and impatience in the evening rush hour increase the level of danger, while Lee et al., (2018) noted that decreased visibility during this period also heightens risks. Similar findings of the prior years on road fatalities in Namibia by Nangombe and Neema (2012), also found that evening hours subsidises to the highest number of fatalities every day.

4.6. Number of Accidents per hour (2018-2023)

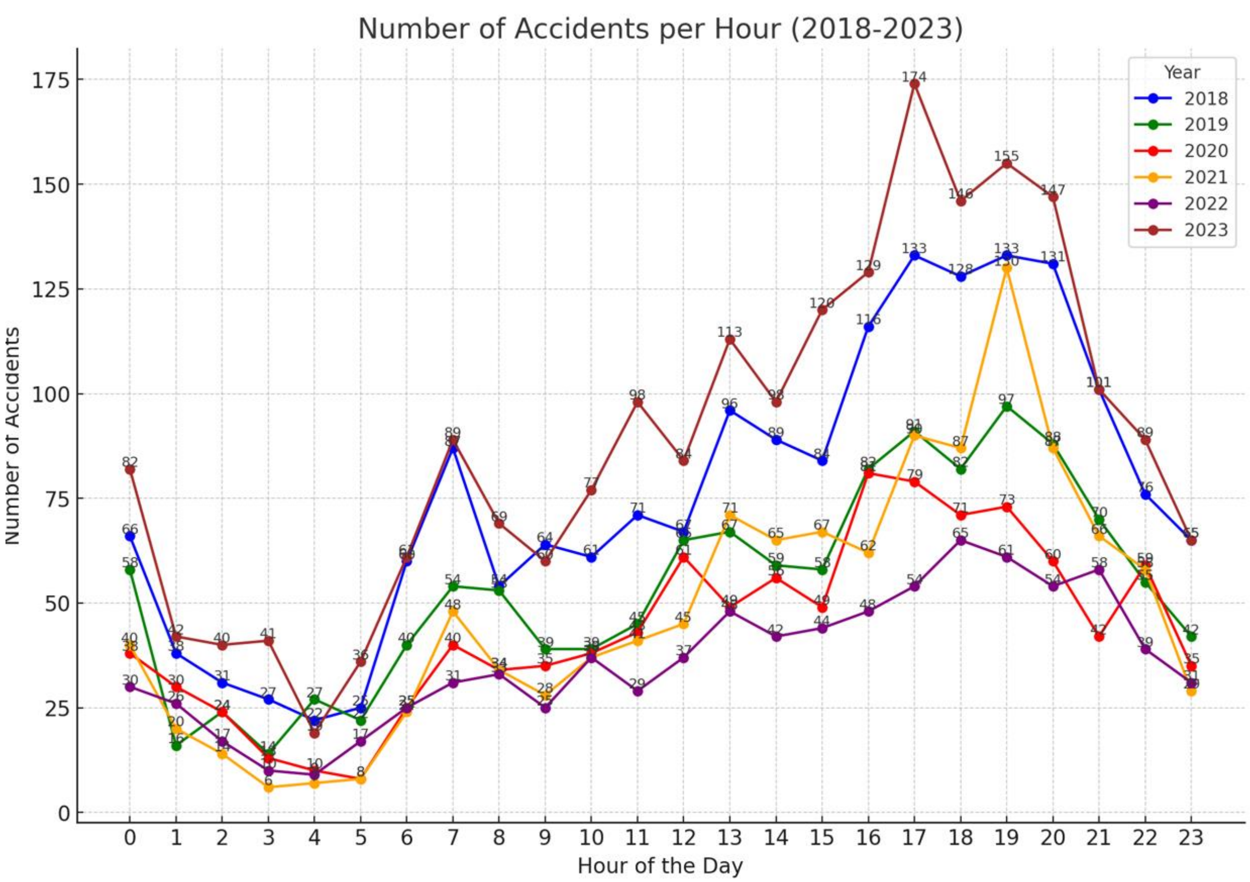


Figure 4. 5: Number of Accidents per hour (2018-2023)

The graph of accidents per hour from 2018 to 2023 shows recurring patterns in high accident rates. The graphs show that accidents reach their peak at midnight, during morning rush hours, and in the late afternoon to early evening each year. Multiple years show significant midnight accident counts, especially in 2023, with 82 accidents. More accidents during the morning typically occur between 7:00 and 8:00 AM, highlighting the risks associated with the early commute. Across all the years, the accident rates are high in the afternoon and evening, especially between 4:00 and 6:00 PM. The evening rush hour is a time when traffic and commuting are at their worst, thereby increasing the risk of accidents. The graph supports the assumption that commuting and late-night travel make particular times of day, regardless of the year, more dangerous on the roads.

4.7. Number of Accidents per Region for a period 2018-2023

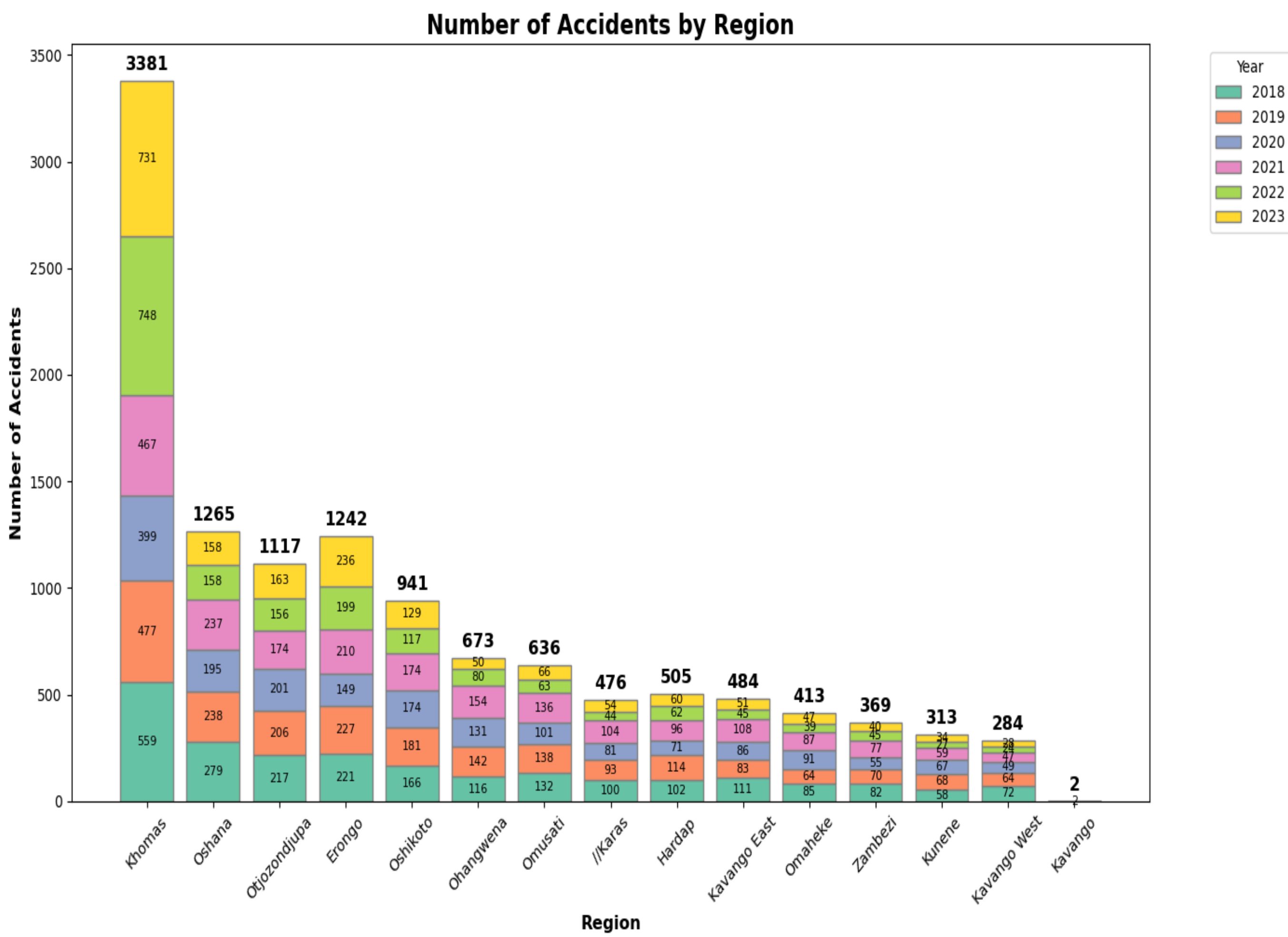


Figure 4. 6: Number of Accidents per Region for a period 2018-2023

The chart illustrates the distribution of traffic accidents in different areas of Namibia from 2018 to 2023. It is demonstrated that road accidents are more probable in areas characterised by high vehicle density, congested traffic, and substantial populations. Khomas region counts 3,381 accidents, representing the highest incidence of road traffic accidents (RTAs) in the nation. This is probably attributable to its designation as the most congested urban centre with the greatest traffic volumes. Prior research has designated Khomas as the location with the greatest incidence of road traffic accidents in Namibia (Chatukuta, 2020). Windhoek, the capital city of Namibia situated in the Khomas region, functions as the nation's economic and administrative centre, seeing significant daily vehicular traffic. The annual analysis in Khomas indicates a reasonably stable incidence of accidents, notably increasing to 731 in 2023. This growing trend could be attributed to deteriorating traffic conditions, increased vehicle utilisation, or possibly increased accident reporting. Considering that Khomas possesses the largest population (494,605), there exists a positive association between population size and accident rates. The results agree with the study by Peden et al., (2019) , who highlighted that areas with higher population densities tend to report more accidents due to higher road usage, coupled with the pressure of urban mobility and frequent interactions among vehicles and pedestrians. This finding is further supported by Chen et al., (2021), who identified a strong relationship between the number of vehicles on the road and accident rates, particularly in densely populated and economically active urban regions.

The Oshana and Erongo regions record elevated accident totals of 1,265 and 1,242 incidents, respectively. Oshana's strategic position in northern Namibia, is that of functioning as a pivotal junction for numerous main routes and this exacerbates its elevated accident rate. This strategic location enhances traffic volumes in the area, thereby elevating the probability of accidents. The annual data for Oshana indicates a decline in accidents in recent years, especially in 2022 and 2023, suggesting effective road safety measures or modifications to traffic legislation that have favourably influenced the area. Conversely, Erongo, being a seaside region with substantial tourist engagement and active mining operations, encounters heightened traffic volumes that lead to a higher accident rate. The annual statistics in Erongo indicate that the incidence of accidents has remained stable, with no significant fluctuations, which implies a persistent level of risk that may necessitate continuous safety interventions to avert any escalation. Donald Matthys (June 11, 2021) states that the Roads Authority indicated that regions with elevated accident rates correspond to the highest number of registered vehicles, predominantly owned by residents of Windhoek (Khomas region; 167,088 vehicles or 41%), Oshakati (Oshana region; 35,638 vehicles or 9%), Walvis Bay (Erongo region; 24,542 vehicles or 6%), and Swakopmund (Erongo region; 22,453 vehicles or 6%).

On the other hand, regions like Zambezi (369 accidents), Kunene (313 accidents), and Kavango West (284 accidents) exhibit notably fewer cases of road traffic accidents. These areas are typically more rural and less populated, resulting in reduced traffic volumes and, thus, fewer crashes. The annual analysis in these regions reveals consistent accident totals without notable fluctuations, signifying a stable yet persistent level of risk that has stayed constant over time. The regions with a lower population and fewer vehicle activities may contribute to a lower likelihood of accidents occurring.

4.8 Average number of accidents per top 6 regions by top towns

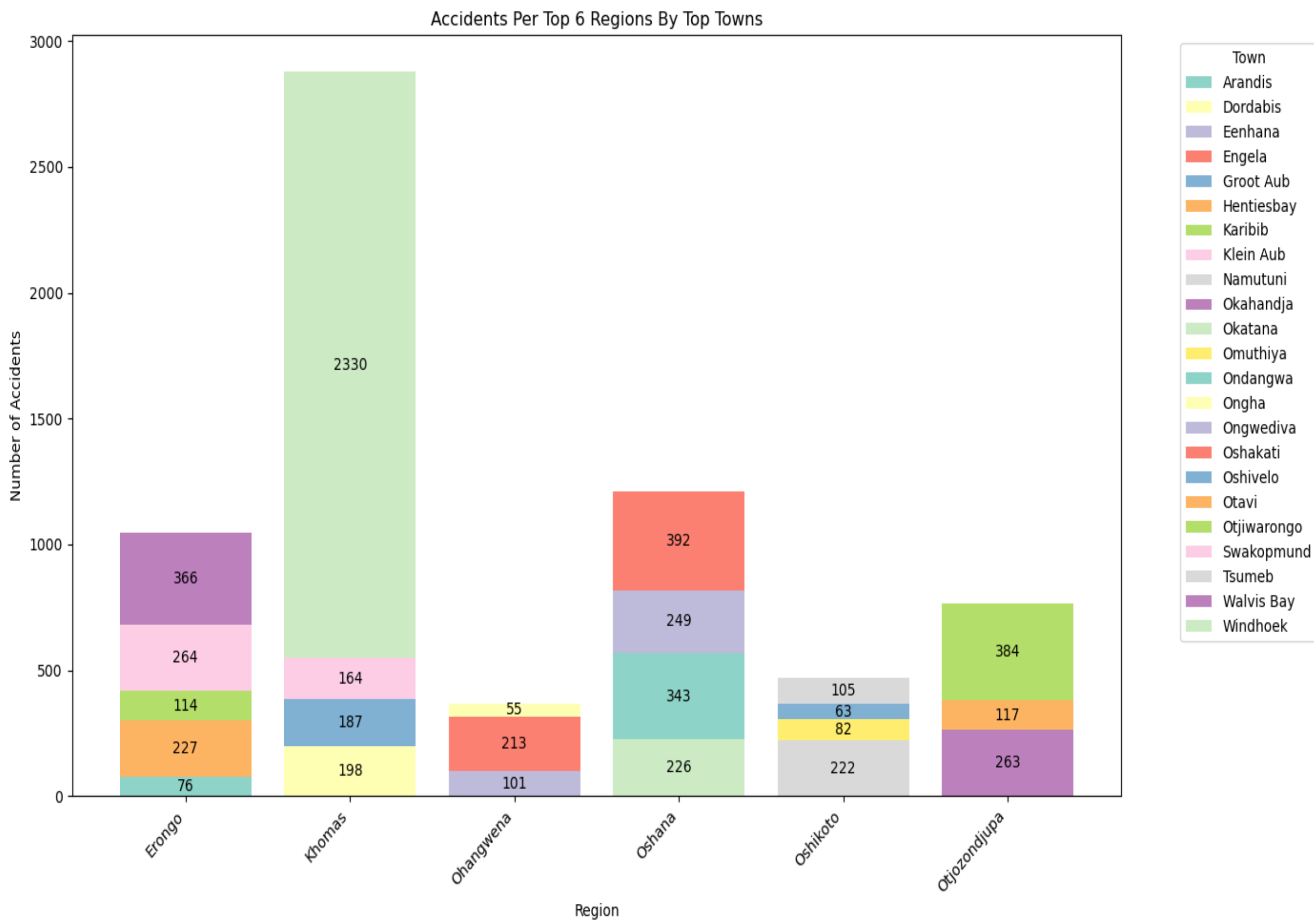


Figure 4. 7: Average number of accidents per top 6 regions by top towns

The bar graph shows the distribution of traffic accidents across six regions in Namibia with higher accident rates, emphasising the prominent towns within each region. Windhoek has a significant number of accidents in the Khomas region, with a total of 2,330 incidents, thereby confirming its status as the most accident-prone area by a substantial margin. This demonstrates the effect of urbanization on traffic incidents, as greater population density and traffic volume result in an increased frequency of accidents. On the other hand, regions such as Erongo and Oshana show a more balanced distribution of accidents across various towns. Erongo, Walvis Bay, and Swakopmund are key contributors, demonstrating their importance as major urban centres. The concentration of accidents in Oshana, particularly in Oshakati, indicates that regional traffic safety initiatives are more effective than those targeting individual towns. The graph demonstrates a notable correlation between urbanisation and accident frequency, with densely populated regions like Khomas and Erongo showing increased accident rates. Future traffic safety initiatives should focus on these areas, especially Windhoek, to lower the rising number of fates. The distribution of accident counts in regions like Otjozondjupa indicates that regional strategies should account for the specific needs of different towns to ensure effectiveness.

4.9 Distribution of Different Types of Accidents

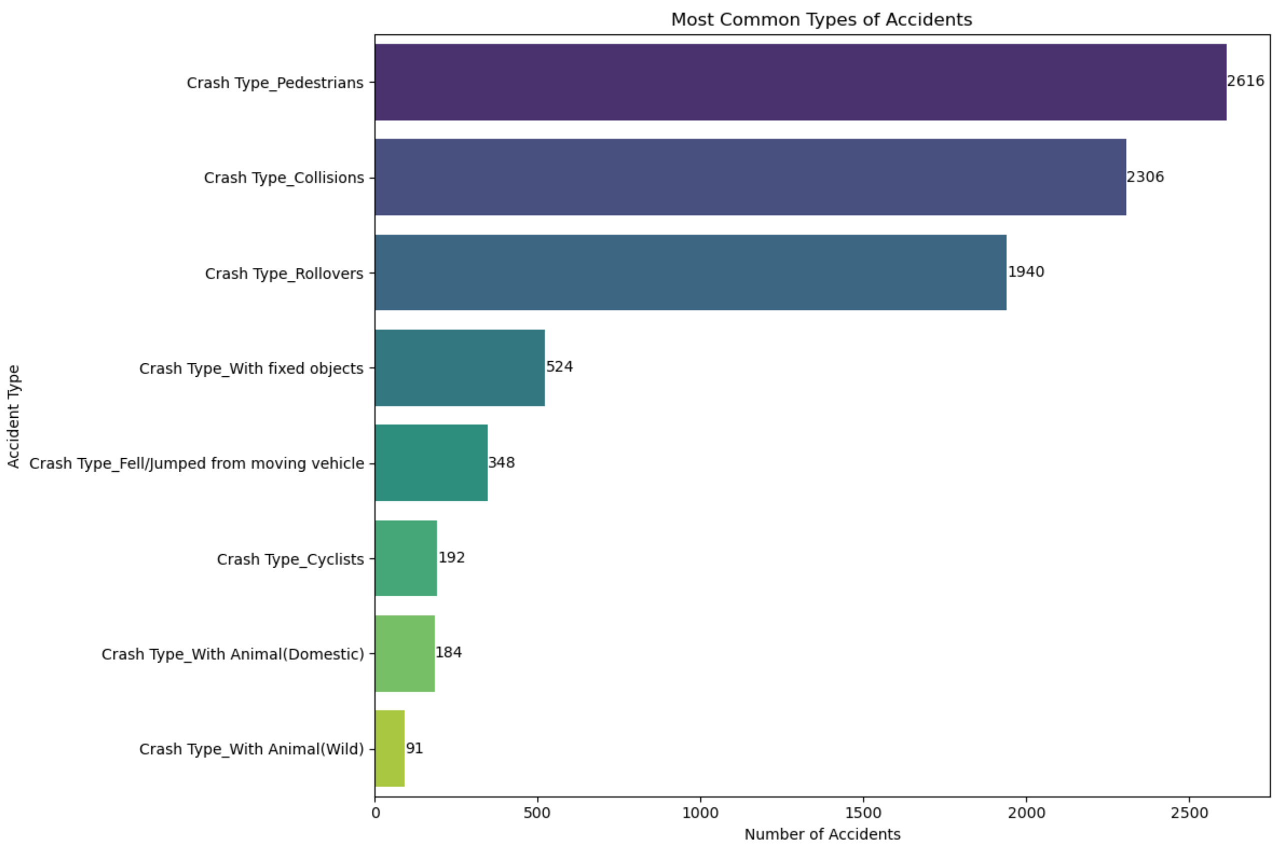


Figure 4. 8: Distribution of Different Types of Accidents

Figure 4.8 indicates the distribution of various types of accidents, with pedestrian-related incidents occurring most frequently with 2,616 cases. The World Health Organisation's Global Status Report on Road in low- and middle-income countries indicates that collisions represent the second most prevalent category of accidents, indicating a necessity for improved vehicle interaction and traffic management strategies. Wegman et al.,'s (2020) research suggests that enhancing traffic signals, enforcing speed limits, and improving road design can significantly reduce vehicle collisions. Moreover, rollovers constitute a substantial fraction of accidents, indicating possible concerns regarding vehicle stability and control. According to the National Highway Traffic Safety Administration (NHTSA, 2020), sharp manoeuvres, high speeds, and poor road conditions frequently lead to rollovers, with SUVs and trucks presenting a heightened risk due to their raised centres of gravity. The results aligned with the findings by Bester and Geldenhuys (2018) who reported that rollover incidents are mostly due to high speeds and poor road conditions. However, some studies' findings are in disagreement for instance the study conducted by Mthembu and Moyo (2020) found more car and fixed object crashes than pedestrian or rollover accidents. This shows that improving road safety with advanced vehicle stability technologies and better infrastructure is crucial for reducing the dangers of pedestrian accidents, collisions, and rollovers.

4.10 Vehicles involved in Accidents

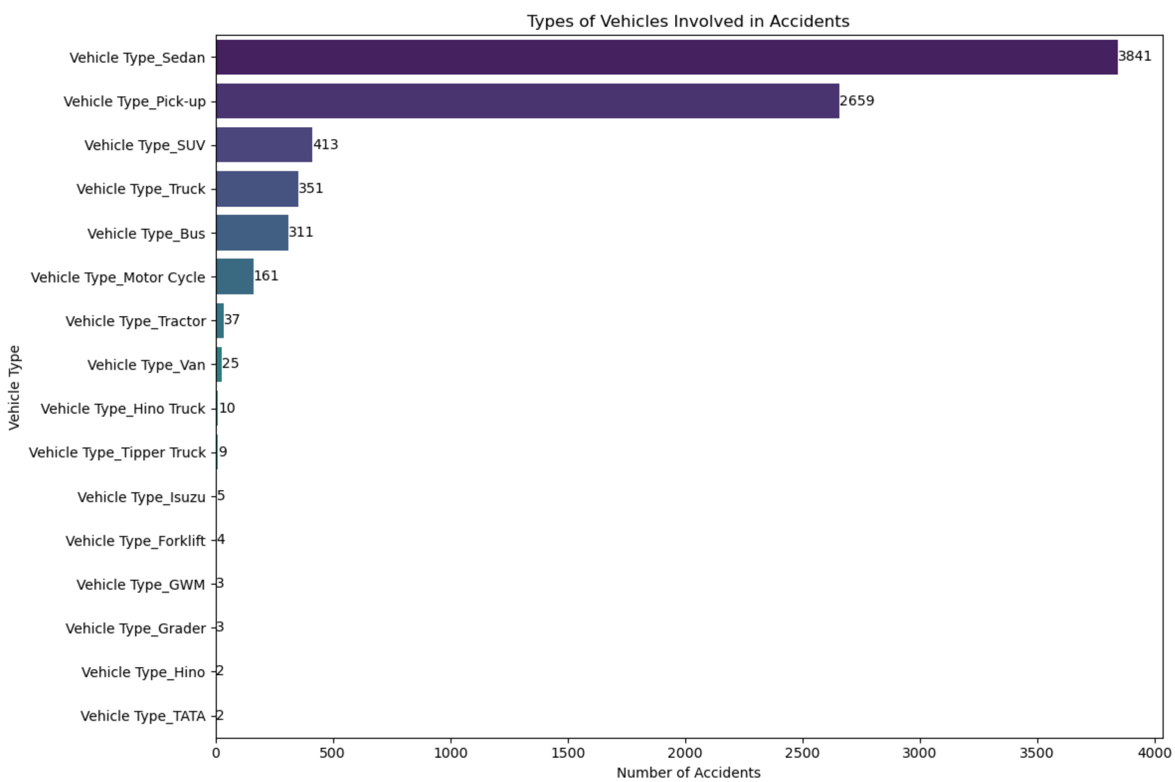


Figure 4. 9: Vehicles involved in accidents

The figure 4.9 categorises cars involved in accidents, indicating that sedans contribute 49% of accidents, and trucks (34%) are the most frequently implicated vehicle types. SUVs (413), trucks (351), and buses (311) follow in ranking; however, motorcycles (161) and other less common vehicles, such as tractors, vans, and Hino trucks, demonstrate significantly lower accident involvement. Mushonga (2020) found that cars and trucks were frequently involved in road accidents because of their widespread use for personal and corporate travel and they are always on the road and similarly, the study conducted by Maponga et al., (2021) observed that cars and trucks dominate the road and cause more accidents.

This suggests that commonly used vehicles for personal or public transport, particularly sedans and trucks, exhibit a higher likelihood of accidents due to their increased frequency and regular use. Zambon and Fedeli (2013) found that smaller passenger vehicles, including sedans and pickups, are involved in road accidents more often than larger commercial vehicles, due to their higher presence on the road, which aligns with this observation.

4.11 common causes of Crashes

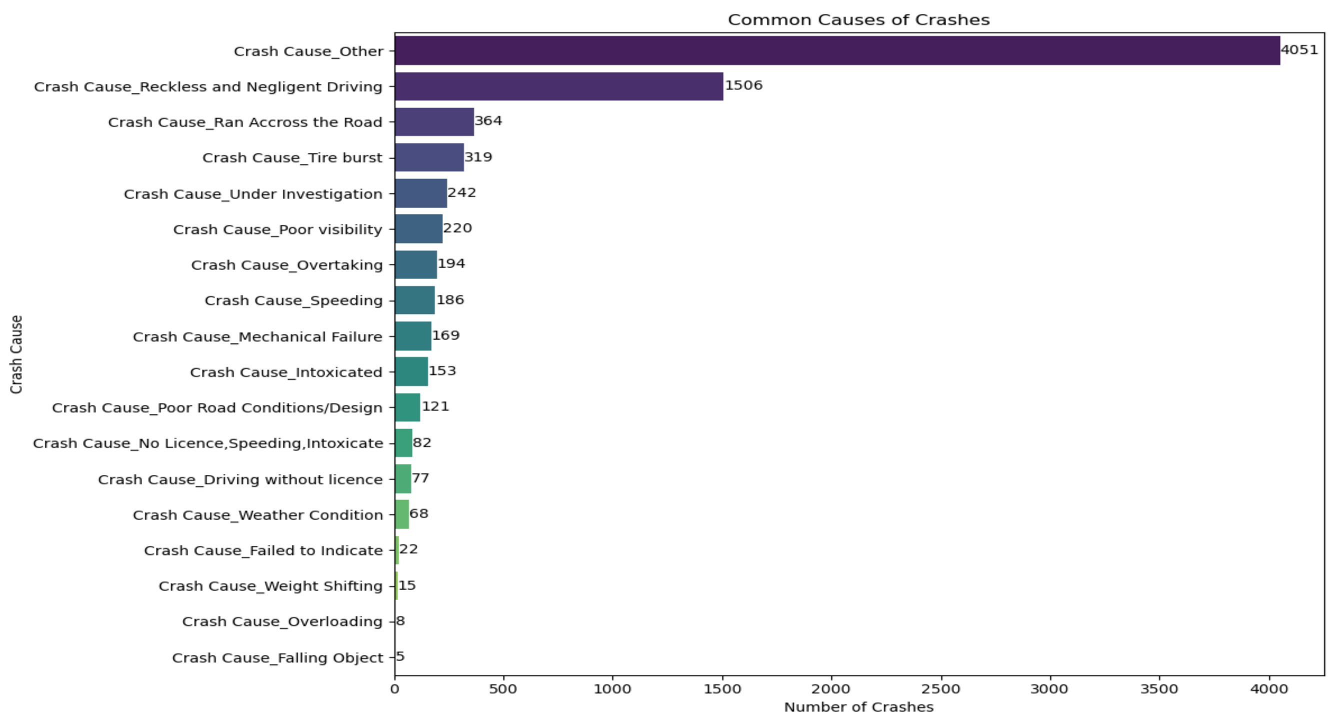


Figure 4. 10: common causes of crashes

Figure 4.10 illustrates the primary causes of accidents, with "reckless and negligent driving" accounting for 1,506 accidents overall. Factors that add to the problem are street crossings, tire blowouts, limited visibility, overtaking, and driving too fast. Labelling 4,051 crashes as "other" suggests a range of rare or unknown reasons for the incidents. Adanu and colleagues (2020) along with NHTSA (2015) point out that reckless driving plays a major role in causing road accidents, especially in sub-Saharan Africa, where lack of driver training and high speeds are common problems. Around 90% of global road traffic accidents are caused by human error, such as irresponsible and careless driving, as stated by the World Health Organisation. It indicates that human mistakes and driving habits are crucial factors in road crashes, emphasising the importance of enhancing traffic regulations, enforcement, and driver education to reduce these incidents.

4.12 Results of the models fitted

4.12.1 Models: Target Variable: Crash Priority: Moderate = 0, Severe =1, Slight = 2

The study utilised machine learning algorithms such as support vector machines (SVM), random forests, and K-nearest neighbours (KNN) to predict the crash priority. The study further employed K-means clustering and association criteria to discover patterns within the data. The dataset was separated into two parts: 70% for training and 30% for testing.

4.12.2 Random Forest:

Confusion Matrix:

$$\begin{bmatrix} 50 & 32 & 231 \\ 5 & 290 & 106 \\ 40 & 123 & 1907 \end{bmatrix}$$

Table 4.1 Model Performance for the Random Forest model

Class	Precision	Recall	F1-score
0	0.53	0.16	0.25
1	0.65	0.72	0.69
2	0.85	0.92	0.88
Accuracy			0.81

The Random Forest model was implemented with n_estimators=200 and max_depth=20.The Random Forest mode l achieved a high accuracy rate of 81%. Therefore, Random Forest demonstrated outstanding results compared to o ther models in predicting 'Slight' accidents. Robust performance has been supported by similar studies. Sharma et a l., (2021) indicated that Random Forest models effectively capture complex interactions among input features such as speed, vehicle type, and road conditions, enhancing their predictive capability for accident severity.

The confusion matrix indicates that the model was particularly good at identifying cases classified as 'Slight' (‘CRASH PRIORITY = 2’), which showed the highest precision and recall scores. Table 4.1 above indicates that, Class 0 (Moderate), has a precision of 0.53, recall of 0.16, and F1-score of 0.25 indicate a struggle to accurately predict 'Moderate' accidents, as the low recall shows that the model often misses moderate accidents, which could be critical for safety planning and interventions. For Class 1 (Severe), the precision of 0.65, recall of 0.72, and F1-score of 0.69 show that the model performs moderately well in identifying 'Severe' crashes, but with a notable false negative rate, suggesting that while the model can predict severe accidents reasonably well, it still fails to capture some severe cases, which could impact targeted interventions. For Class 2 (Slight), a high recall (0.92) and F1-score (0.88) suggest that this class was the easiest to predict, meaning the model is very effective in identifying slight accidents, which could help in understanding less critical crashes and implementing preventative measures for them.

The model exhibited difficulties in predicting "moderate" collisions; however, its robustness in the slight and severe categories indicates potential practical applications. Moderate incidents are challenging to differentiate from other types of accidents because of their comparable characteristics (Kim et al., 2019).

Table 4.2: The importance scores of the top 10 input features:

Feature	Importance
Number of vehicles involved	0.1895
Crash Type	0.1278
Speed limit	0.1251
Day of week occurred	0.1194
Crash time occurred	0.1194
Junction Type	0.1128
Crash cause	0.0846
Number of persons involved	0.0788
Condition of road sign	0.0290
Road direction	0.0100

Table 4.2 above shows the top important features that are mostly contributing to the crash priority using the Random Forest model. It is indicated that the number of vehicles involved was the most influential factor with 18%, followed by crash type (12%) and the speed limit on the road. These findings are in agreement with the research conducted by Afghari et al. (2021), which found that human factors are one of the most significant predictors of crash severity. This indicates that interventions aimed at reducing the number of people involved in accidents, enforcing stricter speed limits and improving road safety in adverse weather conditions could significantly reduce accident severity.

4.12.3 Support Vector Machine (SVM)

Confusion Matrix:

$$\begin{bmatrix} 71 & 47 & 195 \\ 28 & 306 & 67 \\ 306 & 180 & 1584 \end{bmatrix}$$

Table 4.3. Model Performance for the SVM model

Class	Precision	Recall	F1-score
0	0.18	0.23	0.20
1	0.57	0.76	0.66
2	0.86	0.77	0.81
Accuracy			0.70

The SVM model, which used an RBF kernel, achieved an overall accuracy of 70%. The model performed well at distinguishing 'severe' and 'slight' accidents but had difficulty predicting' moderate' accidents, similar to the Random Forest model. The model's precision, recall, and F1-score for 'moderate' accidents were 0.18, 0.23, and 0.20, respectively. The poor performance shows that the SVM model has trouble telling the difference between moderate accidents and other types. This is probably because "moderate" and other severity levels share some features.

In addition, Zhao et al., (2020) also found that SVM often struggles with intermediate categories in multi-class task classifications due to its susceptibility to pattern overlap. The model showed acceptable accuracy for "severe" accidents with a precision of 0.57, a recall of 0.76, and an F1-score of 0.66. This suggests that, although SVM effectively identifies numerous severe crashes, it often misclassifies certain less severe accidents as severe, potentially resulting in unnecessary resource allocation or misdirected interventions. This result is consistent with previous research indicating that SVM models effectively identify high-risk crash scenarios, although they often overestimate severity (Han et al., 2020). The model demonstrated strong performance for "slight" accidents, achieving a precision of 0.86 and an F1-score of 0.81, indicating that SVM can effectively identify less critical crashes. This insight may aid in comprehending patterns of minor accidents and developing more efficient preventative measures.

4.12.4.K-nearest Neighbours (KNN)
Confusion Matrix:

$$\begin{bmatrix} 101 & 42 & 170 \\ 55 & 298 & 48 \\ 377 & 162 & 1531 \end{bmatrix}$$

Table 4.4. Model Performance for the KNN model

Class	Precision	Recall	F1-score
0	0.19	0.32	0.24
1	0.59	0.74	0.66
2	0.88	0.74	0.80
Accuracy			0.69

Table 4.4 shows the KNN model performance results, which consider the 10 nearest neighbours. KNN achieved an accuracy of 69%. Its performance closely mirrored that of the SVM model, particularly in its struggles with predicting 'Moderate' accidents. The precision for 'Moderate' accidents was 0.19, with a recall of 0.32. Although slightly better than SVM, this still indicates that the model fails to accurately classify many moderate accidents. For 'Severe' accidents, the model had a precision of 0.59 and a recall of 0.74, meaning that it could correctly identify most severe crashes but still left room for improvement. The recall for 'Slight' accidents was lower than that of the Random Forest and SVM models, at 0.74, but the precision remained high at 0.88. This suggests that while KNN is effective at identifying slight accidents, it does miss some cases, which could impact accident prevention strategies focused on less critical incidents. Kim et al., (2019) noted similar findings, demonstrating that KNN models often struggle with edge cases due to the nature of distance-based classification, particularly when classes are not well-separated.

Figure 4.11 ROC distribution for all the trained models

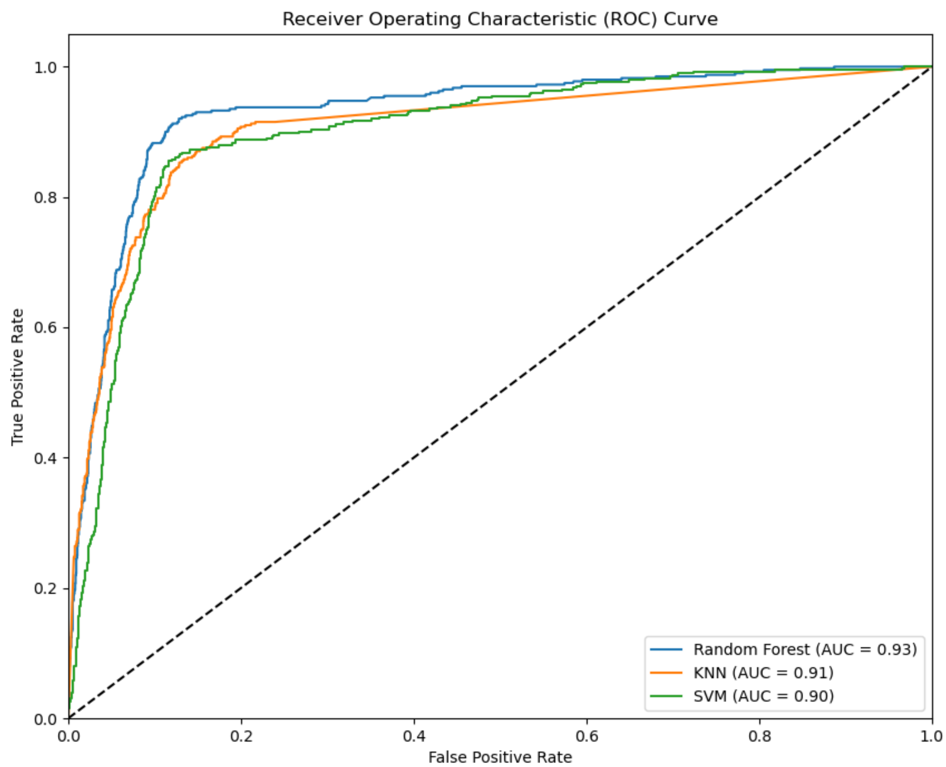


Figure 4.11: ROC distribution for all the trained models

The figure 4.11 shows that the ROC (Receiver Operating Characteristic) curve analysis provides an important measure of a model's ability to distinguish between different classes—in this case, the severity of accidents ('Moderate', 'Severe', and 'Slight'). The Area under the Curve (AUC) summarises the overall performance of the model across all classification thresholds.

The ROC curve analysis confirmed that, although almost all models showed good performance, the Random Forest model outperformed the others, achieving the highest AUC of 93%, indicating its superior ability to distinguish between accident severities. A higher AUC reflects better performance, and Random Forest's result suggests it is the most reliable model for predicting crash severity across all categories, in agreement with the study by Sharma et al. (2021). In contrast, both SVM and KNN had lower AUC values, highlighting their difficulty in accurately predicting 'moderate'' accidents.

4.12 .5 K-Means Clustering

The study determined the best n for K-mean clustering , by running K-means when n=3,4,5 and came to conclusion by choosing the best n for k-means ,and it was done by comparing silhouette score of each k-means , the higher the silhouette score the better the clustering. When n=3 the silhouette score is0.58, when n= 4 the silhouette score is 0.47and when n=5 the silhouette score of 0.42.This suggest that as the number of n increases , the silhouette score becomes less (see the appendix). Therefore, the best n for k-means is 3. K-Means clustering analysis identified unique patterns in crash data among three clusters: Cluster 0 had 2712 cases, Cluster 1 had 3717 cases, and Cluster 2 had 2,008 cases, as shown in Table 4.12.5. The silhouette score of 0.58 suggests that the clusters are somewhat well-defined, indicating distinct patterns in the dataset regarding crash severity. These groupings differ depending on important characteristics like accident severity, number of individuals involved, weather conditions, and road conditions, helping pinpoint root causes of accidents and direct specific interventions.

Table 4.5 Cluster results:

Cluster	summary of the cluster cases (data per each Cluster)
0	2712
1	3717
2	2008

Silhouette Score for Means Clustering: 0.5801570439569687

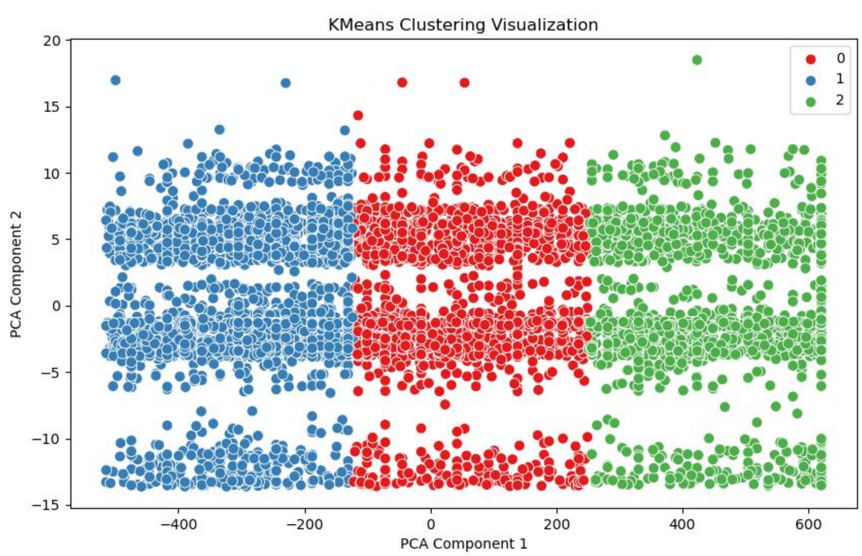


Figure 4.12: clustering

Table 4.6: Most significant features			
Feature	**Cluster 0 (Mean)**	**Cluster 1 (Mean)**	**Cluster 2 (Mean)**
Crash Priority	1.67	1.67	1.63
Number of Persons Involved	3.6	3.68	3.6
Number of Vehicles Involved	1.44	1.49	1.5
Crash Time (Minutes)	570.27	915.07	170.96
Road Surface Quality	2.79	2.75	2.81
Road Surface Type	4.17	4.16	4.28
Speed Limit	10.67	10.65	10.44
Weather Conditions	0.15	0.13	0.16

Table 4.6 shows the most important features of the clustering analysis. Cluster 0 contains 2 712accidents and has a mean crash priority of 1.67 (moderate). The number of persons involved averages 3.60, and the number of vehicles involved is 1.44. The mean crash time is around 570 minutes, indicating a tendency for accidents to occur in the late morning or early afternoon. This cluster is characterised by well-maintained roads (mean road surface type: 4.17) and higher speed limits (mean of 10.67), suggesting accidents on main roads or highways. These findings align with Afghari et al., (2021), who found that accidents on well-maintained roads often occur at higher speeds, contributing to accident severity. Thus, this cluster implies the need for stricter speed enforcement and improved signage to mitigate crash severity on well-maintained roads.

Cluster 1 consists of 3,717 accidents with a mean crash priority of 1.67. It has a slightly higher number of persons involved (3.68) compared to Cluster 0, and the average number of vehicles involved is 1.49. The mean crash time is later in the day, around 915 minutes, suggesting a peak in accidents during the afternoon or early evening. This cluster stands out due to poorer road surface quality (mean of 2.75) and the slightly higher involvement of vehicles, indicating that accidents here may occur in urban settings or areas with poorer infrastructure. According to Chang et al., (2022), lower road quality can significantly increase accident risk, especially in densely populated areas. This cluster highlights the importance of investing in road infrastructure improvements and urban traffic management.

Cluster 2 contains 2,008 accidents with a slightly lower crash priority of 1.63, representing relatively less severe accidents. The number of persons involved averages 3.60, with 1.50 vehicles involved on average. The mean crash time for this cluster is earlier, around 171 minutes, suggesting that accidents are more frequent during the early morning hours. The road surface quality is relatively high (mean of 4.28), and the speed limit is similar to the other clusters (mean of 10.44). This cluster likely represents accidents on well-maintained roads during periods of low traffic, such as early morning hours. According to Sharma et al., (2021), accidents in such conditions are often the result of human error rather than road conditions, highlighting the importance of driver awareness campaigns.

4.12 .6 Association Rule

Table 4.6 represent the outcomes of the associated rules. The table was set with a lift>1 and support>0.6 to get high lift and high support rules, and the rest of the rules, are included in appendix. The rule has items included under the antecedents (left-hand side of the association rule), which represent the items that were observed as a condition or premise. The antecedents specify what is occurring and are, thus, considered the starting point of the rule. The rule has more than one item as the consequent (right-hand side of the association rule), which represents the items that are found to occur in conjunction with the antecedents or predicted to occur in the future. The consequent indicates the outcome or the result that is associated with the group of antecedents. This study presented a total of 20 rules with lift values greater than 1 and support values greater than 0.6. A high support indicates a higher proportion of the items in the set.

The researcher noticed that 'Crash type= Head sides⇒Light condition daylight; vehicles involved few; crash priority _moderate (conf. 0.9; lift 2.9). Confidence (0.90 or 90%): The probability that a recorded crash type as head sides, will contain Light condition daylight; vehicles involved few; crash priority _moderate. Lift (2.9): The probability of finding head sides, Light condition daylight; vehicles involved few; crash priority _moderate occurring together is 2.9 times higher than finding any one of them independently in the dataset. The Lift value is greater or equal to 1, indicating that this rule can most likely be used for prediction rather than only a random guess. This implies to all the rules with high lift. On the other poor, with high support, Condition of road sigh poor; Junction type Slipway⇒ $Crash\ priority$_slight, traffic control type uncontrolled and weather clear (Conf. 0.765, support 0.64) are high frequent rules which are highly related to accident. More association rules see the appendix.

Table 4.7 High Lift and high Support

Rule	Antecedents	Consequents	Support	Confidence	Lift
1	{'CRASH TYPE_Head Side'})	{'LIGHT CONDITION_Daylight', 'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Moderate'}	0.467	0.914	2.912
2	{'CRASH TYPE_Head rear collision'})	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Slight'}	0.568	0.815	2.904
3	{'ROAD SURFACE_Dry', 'CRASH TYPE_Head Side'})	{'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Slight'})	0.514	0.812	2.787
4	{'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'LIGHT CONDITION_Daylight', 'ROAD SURFACE_Dry'}	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People', 'WEATHER_Clear'}	0.415	0.725	2.126

5	{'ROAD TYPE_Freeway', 'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'LIGHT CONDITION_Daylight'	0.412	0.733	1.635
		High Support			
6	{'JUNCTION TYPE_Slipway'	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'}	0.849	0.849	1.281
7	{'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.834	0.707	1.288
8	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'LIGHT CONDITION_Daylight'})	0.752	0.700	1.275
9	{'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People', 'WEATHER_Clear'})	0.736	0.732	1.528
10	{'CONDITION OF ROAD SIGN_Poor', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'WEATHER_Clear’	0.640	0.765	1.288

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

The study's objective was to uncover actionable insights from historical data to help develop accident prevention policies and programmes. The analysis included human, environmental, and road factors. Human factors are referred to as those factors that involve human behaviours such as speeding, overtaking, and recklessness, while environmental factors are factors such as weather[rain, snow and ice, severe winds, excessive temperatures] and visibility[poor lighting, obstructed views, sun glare]. Coupled with human errors and environmental factors are road factors such as potholes, uneven surfaces, and inadequate signage which cause accidents.

The study investigated the primary factors that contribute to the high rate of road accidents in Namibia. Firstly, the analysis revealed that the random forest model excels in predicting crash severity, particularly slight crash priority, and this can be because more accidents are slight over-weighing the moderate and severe accidents with an accuracy of approximately 81%. The analysis revealed that most accidents have a slight crash priority, but it also revealed that pedestrians are the most common crash type in Namibia. Moreover, the higher accident rate in Namibia, particularly in 2023, indicates a tendency for most accidents to occur during rush hours and nighttime. This trend persisted across all the six years studied, indicating a trend rather than a seasonal pattern. Meanwhile, many accidents occur on Fridays, with a higher frequency on Saturdays. This suggests a need for more resources to be on duty during the morning, evening, and weekend hours. Additionally, research shows that sedans and pickups are the vehicle types most frequently involved in crashes.

Moreover, the study revealed a high rate of accidents in regions with high populations (Khomas, Erongo, and Oshana). According to the clustering analysis, Cluster 1 was characterised by poor road conditions and a larger number of vehicles involved. This implies that poor road conditions are associated with a larger number of vehicles. For instance, in the Khomas region, particularly on the Katutura side, substandard road conditions correlate with a higher number of vehicles. Researchers discovered a significant correlation between collision categories and variables such as lighting conditions, vehicle participation, and crash precedence through association mining. Unregulated traffic, minimal vehicle involvement, and moderate crash severity are often associated with head-on and rear-end incidents. Pedestrian accidents regularly happen at uncontrolled intersections in daylight and on dry road surfaces, resulting in minor injuries. Road conditions, especially on motorways, in clear weather, and at slipway intersections, can lead to minor crashes and unexpected traffic. A particular type of vehicle, such as sedans also came out through association rule, affects the incidence of slipway incidents, which often lead to minor collisions on dry road surfaces. Poor road sign conditions, uncontrolled traffic, and slipway junctions frequently trigger minor accidents. Therefore, factors like junction type, weather, lighting, road surface, vehicle type, and driver behaviours play a crucial role in determining the frequency and severity of accidents.

5.1 Contribution

The main contribution of the study is that it is added to the literature review in the Namibian context. This added literature will act as a source of knowledge to provide new ideas for utilising machine learning in predictive traffic accidents rather than traditional statistical techniques.

5.2 Assumptions and Limitations

The study experienced some difficulties. The first hurdle was that the research proposal failed to kick start as planned since the researcher waited to get an approved and signed ethical clearance that took time. Tis hiccup delayed the study to inquire data from MVA Fund and National Road Safety Council on time. The accuracy of the model relies on the quality of historical accident data. This became the second hurdle as the study experienced incomplete data as some factors were missing from the dataset (such as the age, sex, citizenship and some had more missing values) and led to less reliable predictions. Lastly, the historical accident data used to train the model contained biases, like favouring certain groups or areas, which could result in unfair or inaccurate predictions.

5.3 Recommendations

Based on the findings and conclusions of the study the following recommendations are provided to National Road Safety Council and other stakeholders responsible for road safety:

1. **Increased Traffic Enforcement:**
    - **Seasonal Periods:** Deploy additional traffic officers during the festive seasons like the New Year and Christmas to monitor traffic flow and enforce traffic rules.
    - **Peak Hours:** Allocate more resources to traffic management during morning and evening rush hours.
2. **Road Infrastructure Improvement:**
    - **Prioritise Populated Areas**: Invest in improving road conditions, especially in densely populated regions and towns.
3. **Public Awareness Campaigns:**
    - **Pedestrian Awareness:** Conduct awareness campaigns to educate pedestrians about road safety rules and how to protect themselves from accidents.
    - **Driver Education:** Implement comprehensive driver education programmes to improve driver behaviours and reduce accidents.

5.4 Future Work

- **Machine Learning:** Explore the use of machine learning techniques, such as neural networks, to predict accident hotspots and identify potential risk factors.
- **Demographic Factors:** Since it was proven that the crash severity in Namibia is high and we know that Windhoek is the highest, future researcher should consider factors like suburb, citizenship, to identify where exactly in Windhoek are accidents concentrated and also to identify if more accidents are caused by residents or non-residents to gain deeper insights into accident patterns.

References

Abbas, M. S., Yliopisto, O., Päivärinta, T., Peltonen, E., & Peltonen, E. (2023). Accident prediction using machine learning: analyzing weather conditions and model performance. *OuluREPO*.

Acker, B., & Yuan, M. (2018). Network-based likelihood modeling of event occurrences in space and time: A case study of traffic accidents in Dallas, Texas, USA. *Cartography and Geographic Information Science*, *46*(1), 21–38.

Adanu, E. K., Hainen, A., & Jones, S. (2018). Latent class analysis of factors that influence weekday and weekend single-vehicle crash severities. *Accident Analysis & Prevention*, *113*, 187–192.

Afrin, T., & Yodo, N. (2020). *A Systematic Review on Urban Road traffic Congestion*.

Ahmed, S., & al., et. (2024). Comprehensive analysis of machine learning models for traffic accident prediction: A case study from New Zealand. *Electronics*, *13*(8), 3036.

Allström, A., Barceló, J., Ekström, J., Grumert, E., Gundlegård, D., & Rydergren, C. (2017). *Traffic management for smart cities BT - Designing, Developing, and Facilitating Smart Cities* (V. Angelakis, E. Tragos, H. C. Pöhls, A. Kapovits, & A. Bassi (Eds.); pp. 211–240). Springer.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, *41*(3), 359–364. https://doi.org/10.1016/j.aap.2008.12.014

Ardakani, S. P., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a Machine-Learning-Enabled data analysis. *Sustainability*, *15*(7), 5939. https://doi.org/10.3390/su15075939

Asha, C. S., & Narasimhadhan, A. V. (2018). *Vehicle counting for traffic management system using YOLO and correlation filter BT - 2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. 1–6.

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, concepts, and applications for engineers and system designers*.

Berhanu, Y., Dietrich, S., Teklu, B., & Alemayehu, E. (2023). Spatial analysis of road traffic accidents: identifying hotspots for improved road safety in Addis Ababa, Ethiopia. *Cogent Engineering*, *10*(2). https://doi.org/https://doi.org/10.1080/23311916.2023.2269655

Brown, L. H., & Jones, P. L. (2021). Leveraging Machine Learning for Traffic Safety Enhancement in Urban Areas. *IEEE Intelligent Transportation Systems Magazine*, *13*(2), 8–17.

Brown, L. M., & Jones, M. N. (2021). Regional Road Safety Challenges: Customizing Models for Gravel Roads and Unpaved Surfaces. *Road Safety Management*, *9*(4), 200–216.

Cao, T. D., Manolescu, I., & Tannier, X. (2017). Extracting Linked Data from statistic spreadsheets. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Chen, S. J., Hsieh, M. J., Chen, K. C., & Hsu, Y. L. (2018). A Deep Learning Approach for Real-Time Traffic Accident Prediction in Urban Environments. *Sensors*, *18*(11), 3946.

Cheng, R., Zhang, M.-M., & Yu, X.-M. (2019). *Prediction Model for Road Traffic Accident Based on Random Forest BT - DEStech Transactions on Social Science, Education and Human Science*.

Cong, Y., Wang, J., & Li, X. (2016). Traffic Flow Forecasting by a Least Squares Support Vector Machine with a Fruit Fly Optimization Algorithm. *Procedia Engineering*, *137*, 59–68. https://doi.org/10.1016/j.proeng.2016.01.234

Corovic, A., Ilic, V., Duric, S., Marijan, M., & Pavkovic, B. (2018). *The real-time detection of traffic participants using YOLO algorithm BT - 2018 26th Telecommunications Forum (TELFOR)*. 1–4.

Council, N. R. S. (2020). Performance Evaluation of Accident Prediction Models: A Comparative Study. *Journal of Road Safety Analysis*, *8*(2), 75–92.

Dastoorpoor, M., Idani, E., Khanjani, N., Goudarzi, G., & Bahrampour, A. (2016). Relationship between air pollution, weather, traffic, and Traffic-Related mortality. *Trauma Monthly*, *21*(4). https://doi.org/10.5812/traumamon.37585

Dogru, N., & Subasi, A. (2018). Traffic accident detection using random forest classifier. *IEEE Xplore*. https://doi.org/10.1109/lt.2018.8368509

Duan, G., Wang, X., Huang, T., & Kurths, J. (2019). An improved Group Similarity-Based Association rule mining algorithm in complex scenes. *International Journal of Pattern Recognition and Artificial Intelligence*, *34*(02), 2059005. https://doi.org/10.1142/s0218001420590053

Elassad, Z. E. A., Mousannif, H., & Moatassime, H. A. (2020). Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study. *Traffic Injury Prevention*, *21*(3), 201–208. https://doi.org/10.1080/15389588.2020.1723794

Garcia, M. A., & al., et. (2020). Predictive Modeling for Traffic Accidents: A Comprehensive Review of Models and Evaluation Metrics. *Accident Analysis & Prevention*, *144*, 105676.

Harrou, F., Zeroual, A., Kadri, F., & Sun, Y. (2024). Enhancing road traffic flow prediction with improved deep learning using wavelet transforms. *Results in Engineering*, *23*. https://doi.org/https://doi.org/10.1016/j.rineng.2024.102342

Hebert, A., Guedon, T., Glatard, T., & Jaumard, B. (2019). High-Resolution road vehicle collision prediction for the city of Montreal. *IEEE Xplore*. https://doi.org/10.1109/bigdata47090.2019.9006009

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2011). A practical guide to support vector classification. In *Data Science Association*.

Iranitalab, A., & Khattak, A. (2017). Iranitalab and Khattak (2017). *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

Isaksen, E. T., & Johansen, B. G. (2021). Congestion pricing, air pollution, and individual-level behavioral responses. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3832230

Islam, M. K., Reza, I., Gazder, U., Akter, R., Arifuzzaman, M., & Rahman, M. M. (2022). Predicting road crash severity using classifier models and crash hotspots. *Appl. Sci.*, *12*, 11354. [Google Scholar] [CrossRef]

Islam, Z., & Abdel-Aty, M. (2023). Traffic conflict prediction using connected vehicle data. *Analytic Methods in Accident Research*, *39*, 100275. https://doi.org/10.1016/j.amar.2023.100275

Jain, P. (2011). *Automatic traffic signal controller for roads by exploiting fuzzy logic BT - Computer Networks and Information Technologies* (V. V Das, J. Stephen, & Y. Chaba (Eds.); Vol. 142, pp. 273–277). Springer.

Jia, R., Khadka, A., & Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis and Prevention*, *121*(September), 223–230. https://doi.org/10.1016/j.aap.2018.09.018

Jilani, U., Asif, M., Zia, M. Y. I., Rashid, M., Shams, S., & Otero, P. (2023). A Systematic Review on Urban Road traffic Congestion. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-023-10700-0

Johnson, K. L., & Brown, L. M. (2018). Decision Trees for Accident Prediction: A Case Study on Classification and Regression Trees (CART). *Accident Analysis & Prevention*, *120*, 180–195.

Johnson, M., Smith, A., & Brown, C. (2022). Machine Learning Applications in Traffic Flow Prediction. *Journal of Traffic Safety Analysis*, *30*(3), 245–260.

Johnson, M., Smith, A., & Brown, C. (2023). Machine Learning-Driven Predictive Models for Enhanced Traffic Safety Analysis. *Transportation Research*, *45*(4), 567–582.

Keilmann, E., & Weber, T. (2021). Perspektive Olympiaseelsorge. *Academia – Ein Verlag in Der Nomos Verlagsgesellschaft EBooks*, 373–382. https://doi.org/10.5771/9783896659552-373

Khaldi, B., Harrou, F., Benslimane, S. M., & Sun, Y. (2021). A Data-Driven soft sensor for swarm motion speed prediction using ensemble learning methods. *IEEE Sensors Journal*, *21*(17), 19025–19037. https://doi.org/10.1109/jsen.2021.3087342

Khan, M. N., & Das, S. (2024). Advancing traffic safety through the safe system approach: A systematic review. *Accident Analysis & Prevention*, *199*, 107518. https://doi.org/10.1016/j.aap.2024.107518

Komsiyah, S., & Desvania, E. (2021). Traffic lights analysis and simulation using Fuzzy inference system of MamDANI on Three-Signaled Intersections. *Procedia Computer Science*, *179*, 268–280. https://doi.org/10.1016/j.procs.2021.01.006

Kumar, S., Mahima, S. K., Srivastava, D. K., Kharya, P., Sachan, N., & Kiran, K. (2020). Analysis of risk factors contributing to road traffic accidents in a tertiary care hospital. A hospital based cross-sectional study. *Chinese Journal of Traumatology*, *23*(3), 159–162.

Lawal, A. M. (2021). Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study. In *Coventry University*.

Lee, J., Chae, J., Yoon, T., & Yang, H. (2018). Traffic accident severity analysis with rain-related factors using structural equation modeling – A case study of Seoul City. *Accident Analysis & Prevention*, *112*, 1–10.

https://doi.org/10.1016/j.aap.2017.12.013

Li, H., Gao, Q., Zhang, Z., Zhang, Y., & Ren, G. (2023). Spatial and temporal prediction of secondary crashes combining stacked sparse auto-encoder and long short-term memory. *Accident Analysis & Prevention*, *191*, 107205. https://doi.org/10.1016/j.aap.2023.107205

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Liang, T. U., Wang, V. W., & Chen, X. Y. (2021). Evaluating Deep Neural Networks for Accident Prediction: Insights and Challenges. *Transportation Research Part B: Methodological*, *146*, 49–68.

Lykov, S., & Asakura, Y. (2020). Anomalous traffic pattern detection in large urban areas: Tensor-based approach with continuum modeling of traffic flow. *International Journal of ITS Research*, *18*(1), 13–21.

Makaba, T., Doorsamy, W., & Paul, B. S. (2020). Bayesian Network-Based Framework for Cost-Implication Assessment of Road Traffic Collisions. *International Journal of Intelligent Transportation Systems Research*, *19*(1), 240–253. https://doi.org/10.1007/s13177-020-00242-1

Mansoor, U., & al., et. (2023). Ensemble machine learning models for predicting traffic accident severity: A comparative analysis. *Computational Intelligence*, *39*(5), 1082–1095.

Matengu, M., & al., et. (2022). *Factors Associated with Road Traffic Accidents in Windhoek*. Repository of University of Namibia. Retrieved from UNAM Repository

Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis & Prevention*, *42*(2), 571–579. https://doi.org/10.1016/j.aap.2009.09.025

Muktar, B., & Fono, V. (2024). Toward safer roads: Predicting the severity of traffic accidents in Montreal using machine learning. *Electronics*, *13*(15), 3036. https://doi.org/10.3390/electronics13153036

Nakanishi, Y. J., & Auza, P. M. (2023). Connected vehicles and driving automation systems. In *Springer handbooks* (pp. 1079–1113). https://doi.org/10.1007/978-3-030-96729-1_50

Noorunnahar, M., Chowdhury, A. H., & Mila, F. A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PLoS ONE*, *18*(3), e0283452. https://doi.org/10.1371/journal.pone.0283452

Oates, B. J. (2005). *Researching Information Systems and Computing*. SAGE Publications.

Organization, W. H. (2019). *Global Status Report on Road Safety 2018*. World Health Organization.

Organization, W. H. (2023). *Global Status Report on Road Safety 2023: Country Profiles*. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pradhan, B., & Sameen, M. I. (2019). Review of Traffic Accident Predictions with Neural Networks. In *Advances in*

Science, Technology & Innovation/Advances in science, technology & innovation (pp. 97–109). https://doi.org/10.1007/978-3-030-10374-3_8

Road Traffic Injuries. (2021). https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Roshandel, S., Zheng, Z., & Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention*, *79*, 198–211. https://doi.org/10.1016/j.aap.2015.03.013

Salifu, M. (2002). *Analysis of accident potential at unsignalised urban junctions in Ghana*.

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. *Computers*, *10*(12), 157. https://doi.org/10.3390/computers10120157

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. *Journal of Safety Research*, *80*, 254–279. https://doi.org/10.1016/j.jsr.2021.12.007

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students*. Pearson Education.

Shafabakhsh, G. A., Famili, A., & Bahadori, M. S. (2017). GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *Journal of Traffic and Transportation Engineering (English Edition)*, *4*(5), 290–299.

Silva, P. B., Andrade, M., & Ferreira, S. (2020). Machine learning applied to road safety modeling: A systematic literature review. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(6), 775–800. https://doi.org/10.1016/j.jtte.2020.07.004

Smith, A., & Johnson, B. (2018). Enhancing Road Safety through Predictive Analytics. *Journal of Transportation Engineering*, *144*(6), 4018034.

Smith, A., Johnson, M., & Brown, C. (2021). Advancements in Machine Learning for Traffic Flow Prediction. *Journal of Traffic Safety Analysis*, *35*(2), 123–138.

Smith, A., Jones, B., & Brown, C. (2021). Assessing the Transport Needs in Namibia: A Comprehensive Survey. *Journal of Sustainable Transportation*, *25*(4), 567–582.

Smith, J., & al., et. (2023). Advancing Road Safety Through Technological Innovations: A Case Study of Namibia. *Journal of Transportation Research*, *45*(3), 321–335.

Smith, T., Victoroff, J., & Hobbs, G. R. (2017). Amputation, atrocity, and the anatomy of revenge: A relational psychoanalytic perspective on mass violence. *Psychoanalysis, Culture & Society*, *22*(1), 68–81.

Studies, N. I. of T. (2018). Machine Learning-Based Models for Accident Prediction: A Comprehensive Evaluation. *Journal of Transport Safety and Security*, *10*(4), 245–262.

Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing machine learning and deep learning methods for Real-Time Crash Prediction. *Transportation Research Record Journal of the Transportation Research Board*, *2673*(8), 169–181. https://doi.org/10.1177/0361198119841571

Theofilatos, A., & Ziakopoulos, A. (2018). Examining Injury Severity of Moped and Motorcycle Occupants with

Real-Time Traffic and Weather Data. *Journal of Transportation Engineering Part A Systems*, *144*(11). https://doi.org/10.1061/jtepbs.0000193

Transport, N. M. of W. and. (2019). Customizing Accident Prediction Models for Local Road CNamibia Ministry of Works and Transport (2019) onditions: A Road Safety Initiative. *Namibia Transport Research Journal*, *7*(3), 121–138.

Transport, N. M. of W. and. (2020). *National Road Safety Strategy 2020-2030*. Government of Namibia.

Verschuren, P., & Doorewaard, H. (2010). *Designing a Research Project (2nd ed.)*. Eleven International Publishing.

Wåhlberg, A. A., Dorn, L., & Kline, T. (2011). The Manchester Driver Behaviour Questionnaire as a predictor of road traffic accidents. *Theoretical Issues in Ergonomics Science*, *12*(1), 66–86. https://doi.org/10.1080/14639220903023376

Wang, J., Ji, L., Ma, S., Sun, X., & Wang, M. (2023). Analysis of Factors Influencing the Severity of Vehicle-to-Vehicle Accidents Considering the Built Environment: An Interpretable Machine Learning model. *Sustainability*, *15*(17), 12904. https://doi.org/10.3390/su151712904

Wang, X. Y., & Liu, C. D. (2022). Comparative Analysis of Ensemble Models for Accident Prediction: Random Forest vs. Gradient Boosting. *Transportation Research Part C: Emerging Technologies*, *126*, 103120.

Xu, C., Bao, J., Wang, C., & Liu, P. (2018). Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *Journal of Safety Research*, *67*, 65–75. https://doi.org/10.1016/j.jsr.2018.09.013

Xu, Y., Zomer, S., & Brereton, R. G. (2006). Support vector Machines: a recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry*, *36*(3–4), 177–196. https://doi.org/10.1080/10408340600969486

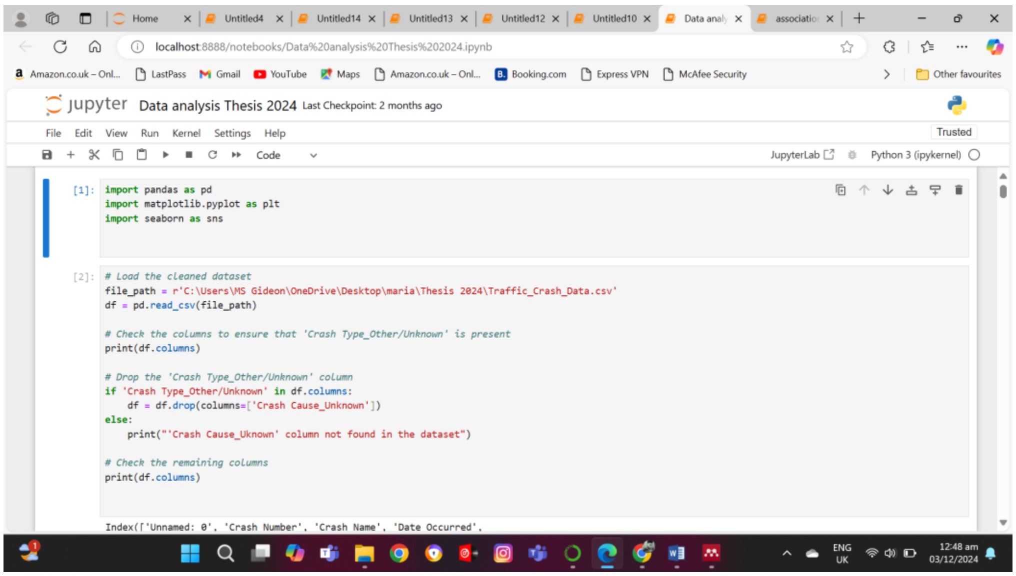
Yuan, Q., Zhang, T., & Cui, H. (2020). Deep learning based crash prediction models: Comparisons and discussions. *Accident Analysis & Prevention*, *139*, 105494.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Journals & Magazine | IEEE Xplore*. https://doi.org/10.1109/jlt.2018.8368509

Appendix A: Implementation Details

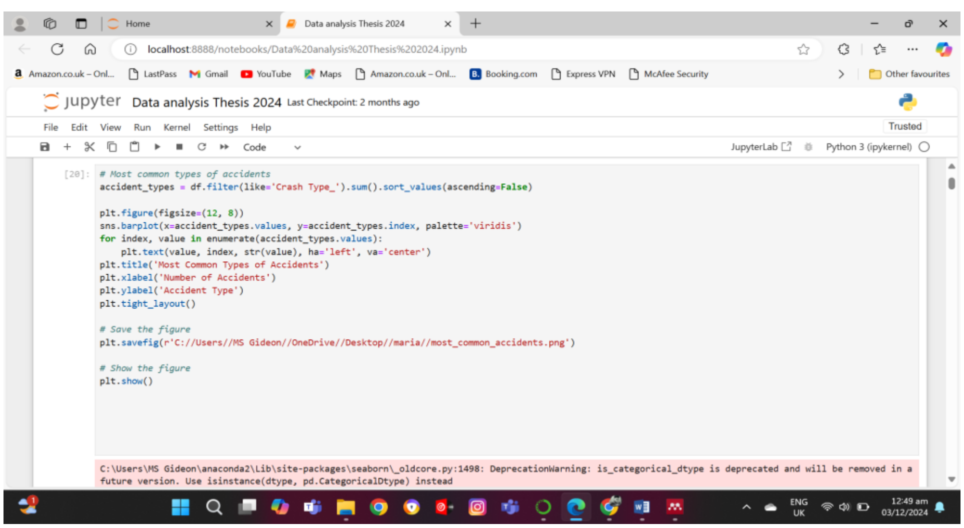
For the analysis, the study makes use of python, and the below are screenshots of some of the codes and codes themselves that were used from the packages, data loading, data cleaning, pre-processing, feature selections etc.

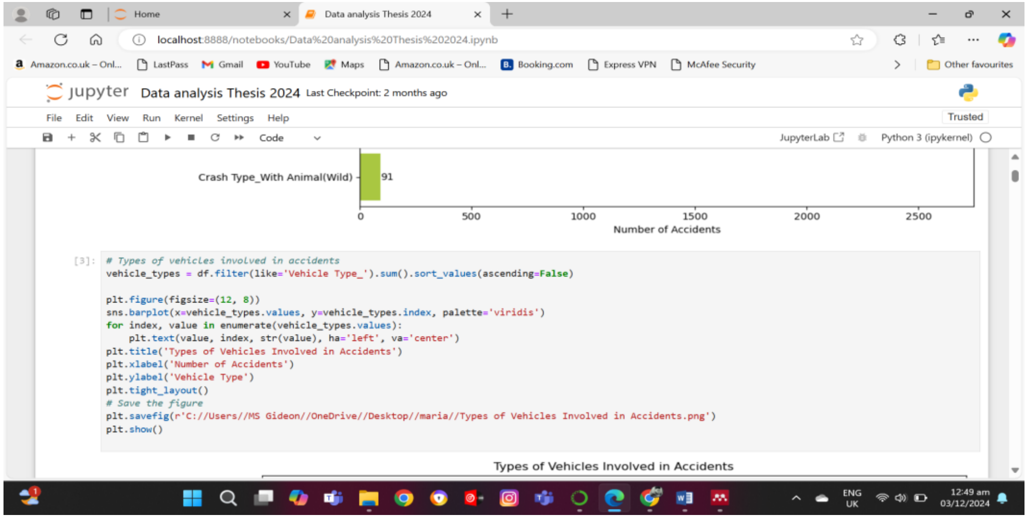
1. Packages and load of data

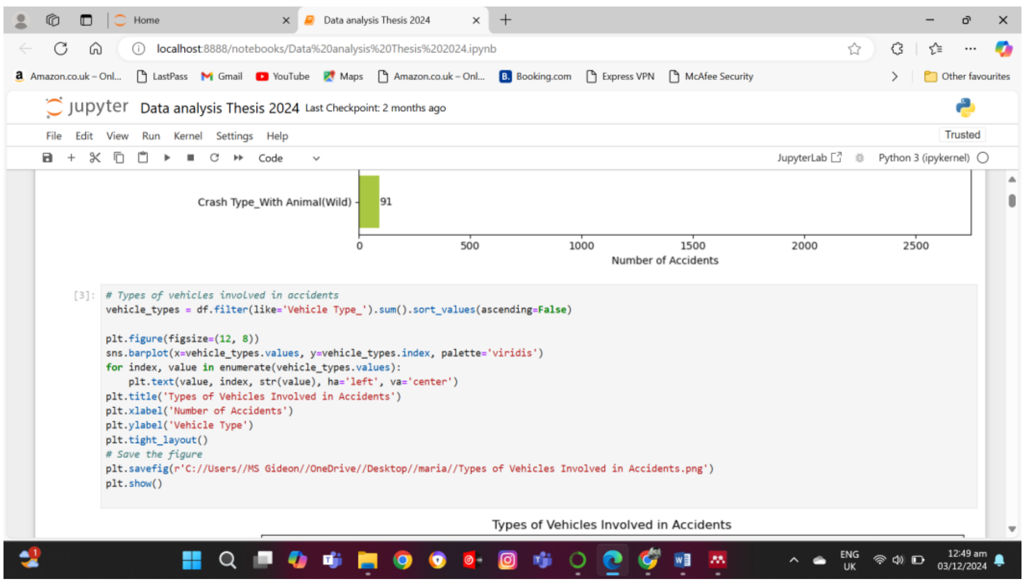


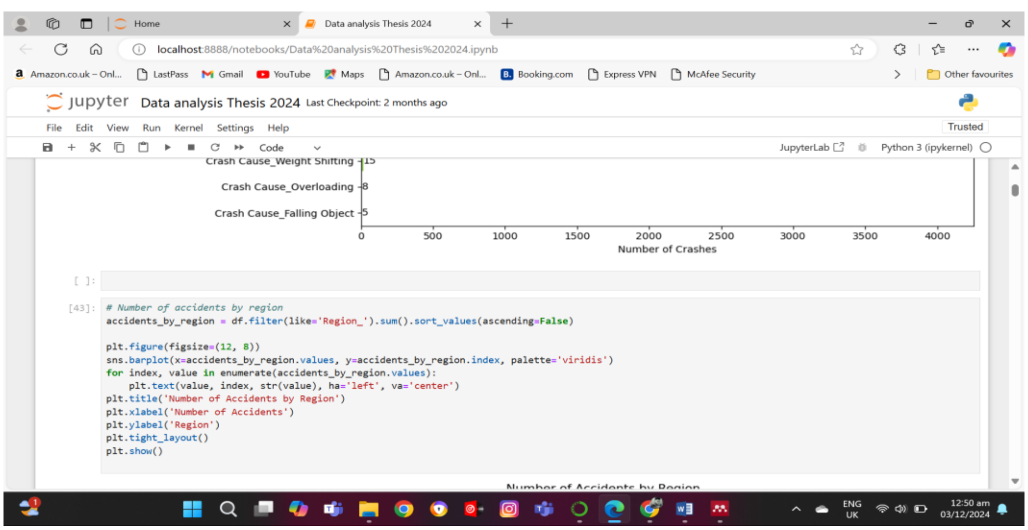
2. Explanatory charts

The descriptive statistics figures are generated using the below codes:

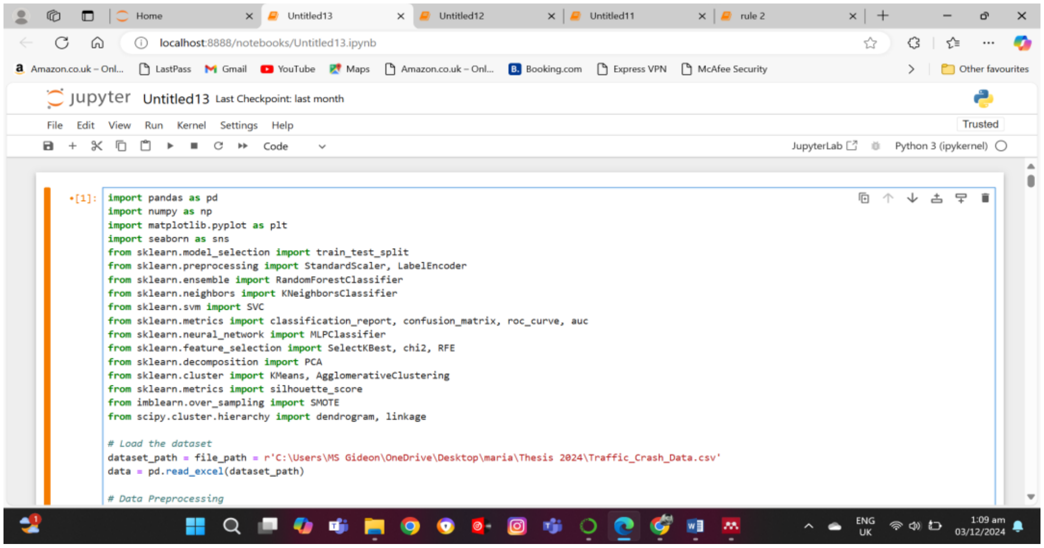



Data analysis Thesis 2024
localhost:8888/notebooks/Data%20analysis%20Thesis%202024.ipynb
Jupyter Data analysis Thesis 2024 Last Checkpoint: 2 months ago
File Edit View Run Kernel Settings Help
Trusted
Crash Type_With Animal(Wild)
91
0
500
1000
1500
2000
2500
Number of Accidents
[3]:
# Types of vehicles involved in accidents
vehicle_types = df.filter(like='Vehicle Type_').sum().sort_values(ascending=False)
plt.figure(figsize=(12, 8))
sns.barplot(x=vehicle_types.values, y=vehicle_types.index, palette='viridis')
for index, value in enumerate(vehicle_types.values):
plt.text(value, index, str(value), ha='left', va='center')
plt.title('Types of Vehicles Involved in Accidents')
plt.xlabel('Number of Accidents')
plt.ylabel('Vehicle Type')
plt.tight_layout()
# Save the figure
plt.savefig(r'C://Users//MS Gideon//OneDrive//Desktop//maria//Types of Vehicles Involved in Accidents.png')
plt.show()
Types of Vehicles Involved in Accidents
12:49 am
03/12/2024


Data analysis Thesis 2024
localhost:8888/notebooks/Data%20analysis%20Thesis%202024.ipynb
Jupyter Data analysis Thesis 2024 Last Checkpoint: 2 months ago
File Edit View Run Kernel Settings Help
Trusted
Crash Type_With Animal(Wild)
91
0
500
1000
1500
2000
2500
Number of Accidents
[3]:
# Types of vehicles involved in accidents
vehicle_types = df.filter(like='Vehicle Type_').sum().sort_values(ascending=False)
plt.figure(figsize=(12, 8))
sns.barplot(x=vehicle_types.values, y=vehicle_types.index, palette='viridis')
for index, value in enumerate(vehicle_types.values):
plt.text(value, index, str(value), ha='left', va='center')
plt.title('Types of Vehicles Involved in Accidents')
plt.xlabel('Number of Accidents')
plt.ylabel('Vehicle Type')
plt.tight_layout()
# Save the figure
plt.savefig(r'C://Users//MS Gideon//OneDrive//Desktop//maria//Types of Vehicles Involved in Accidents.png')
plt.show()
Types of Vehicles Involved in Accidents
12:49 am
03/12/2024

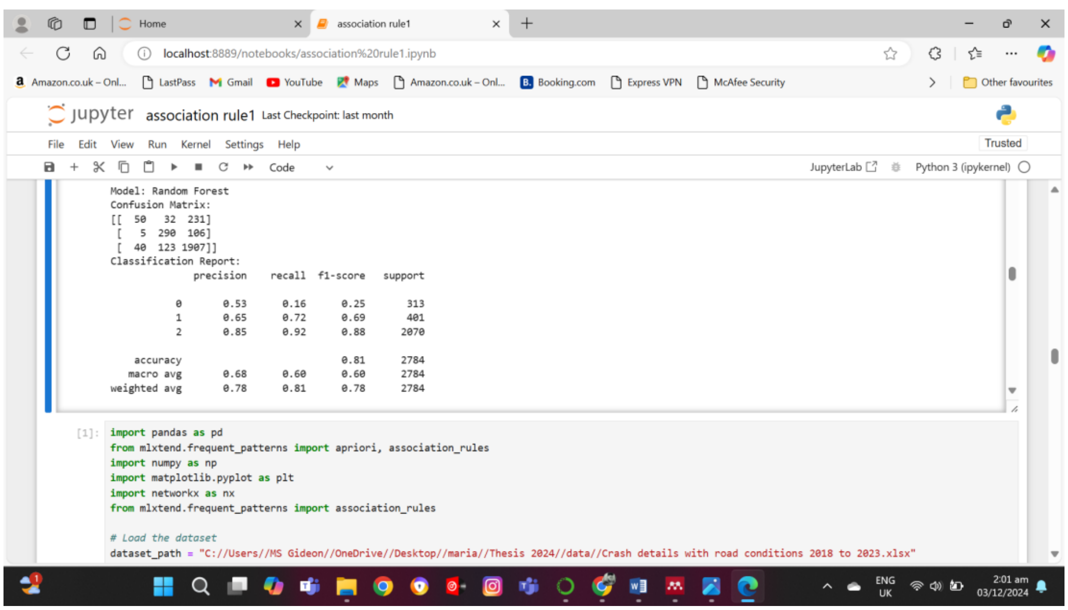

Data analysis Thesis 2024
localhost:8888/notebooks/Data%20analysis%20Thesis%202024.ipynb
Jupyter Data analysis Thesis 2024 Last Checkpoint: 2 months ago
File Edit View Run Kernel Settings Help
Trusted
Crash Cause_Weight Shifting
15
Crash Cause_Overloading
8
Crash Cause_Falling Object
5
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Crashes
[ ]:
[43]:
# Number of accidents by region
accidents_by_region = df.filter(like='Region_').sum().sort_values(ascending=False)
plt.figure(figsize=(12, 8))
sns.barplot(x=accidents_by_region.values, y=accidents_by_region.index, palette='viridis')
for index, value in enumerate(accidents_by_region.values):
plt.text(value, index, str(value), ha='left', va='center')
plt.title('Number of Accidents by Region')
plt.xlabel('Number of Accidents')
plt.ylabel('Region')
plt.tight_layout()
plt.show()
Number of Accidents by Region
12:50 am
03/12/2024

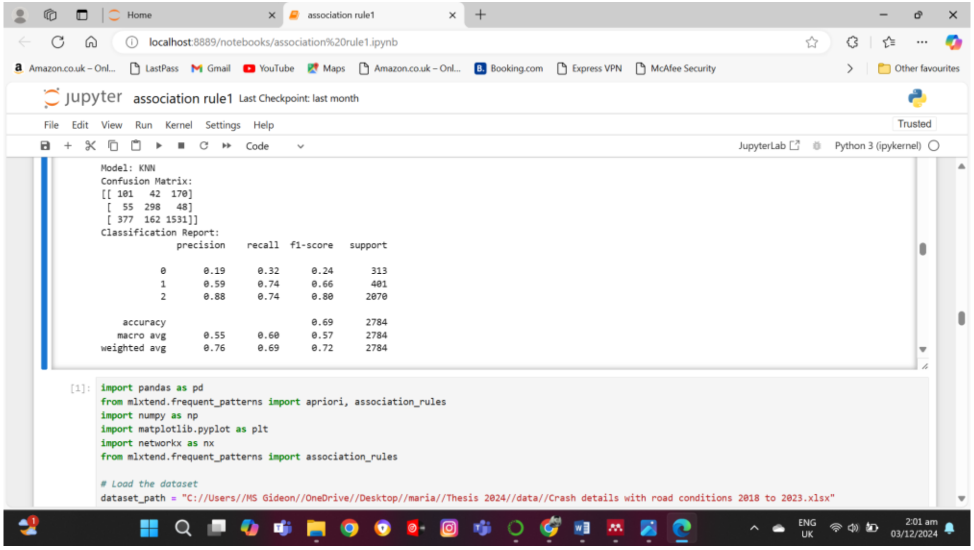
3. Packages needed for the models

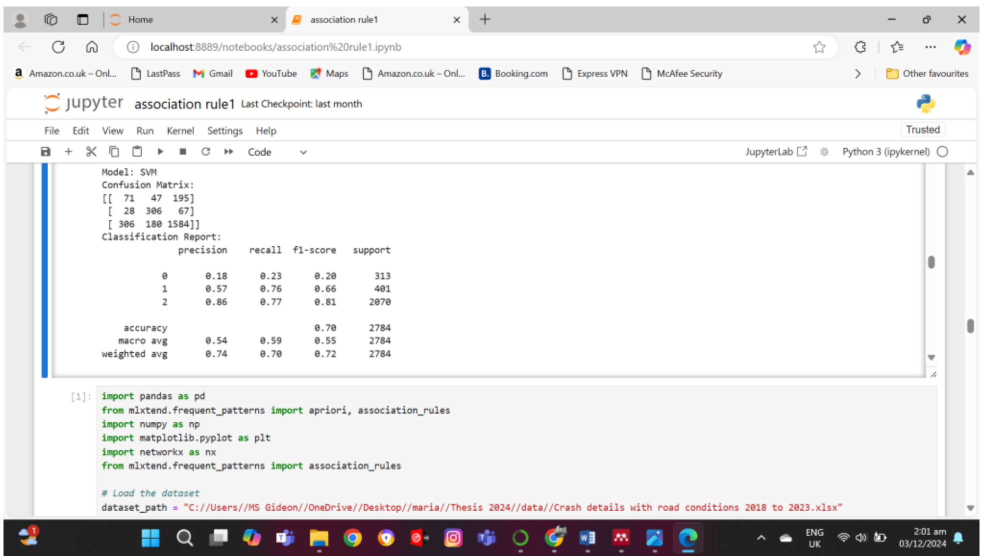


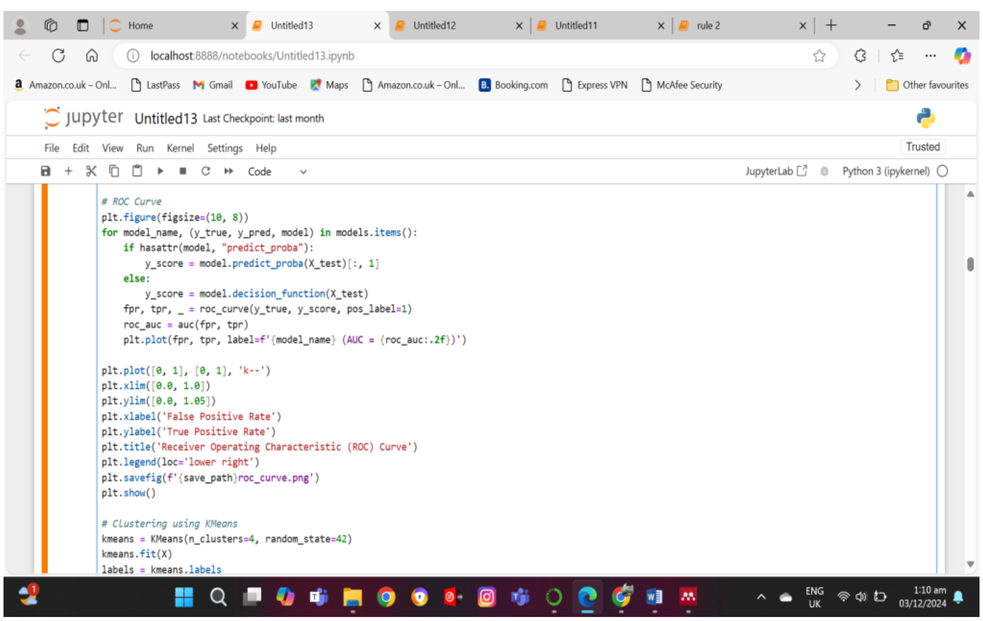
4. Model's results

The outcomes of each of the model are showed here



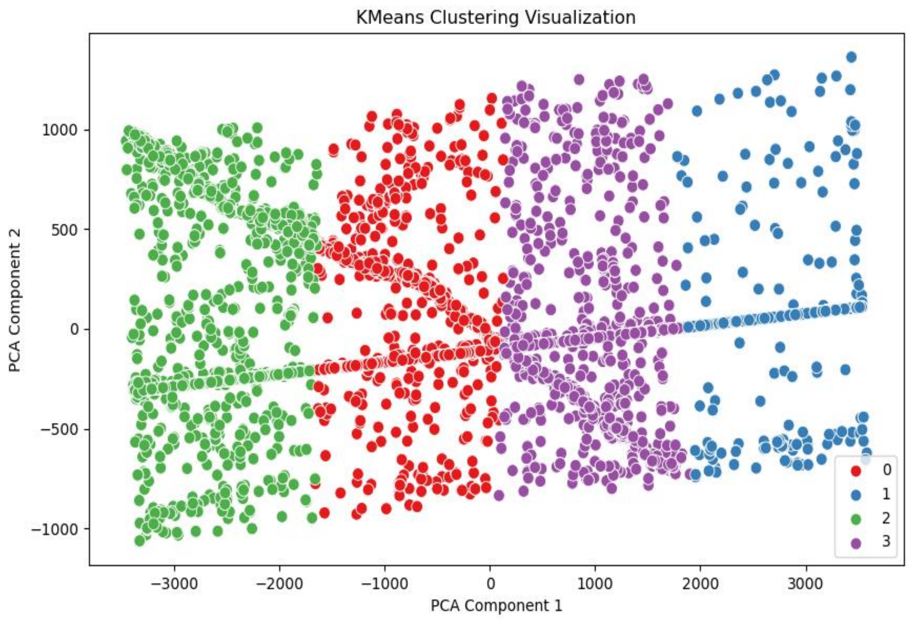

jupyter association rule1 Last Checkpoint: last month
Model: KNN
Confusion Matrix:
[[101   42 170]
 [55 298   48]
 [377 162 1531]]
Classification Report:
precision recall f1-score support
0 0.19 0.32 0.24 313
1 0.59 0.74 0.66 401
2 0.88 0.74 0.80 2070
accuracy 0.69 2784
macro avg 0.55 0.60 0.57 2784
weighted avg 0.76 0.69 0.72 2784
[1]:
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
import numpy as np
import matplotlib.pyplot as plt
import networkx as nx
from mlxtend.frequent_patterns import association_rules
# Load the dataset
dataset_path = "C://Users//MS Gideon//OneDrive//Desktop//maria//Thesis 2024//data//Crash details with road conditions 2018 to 2023.xlsx"

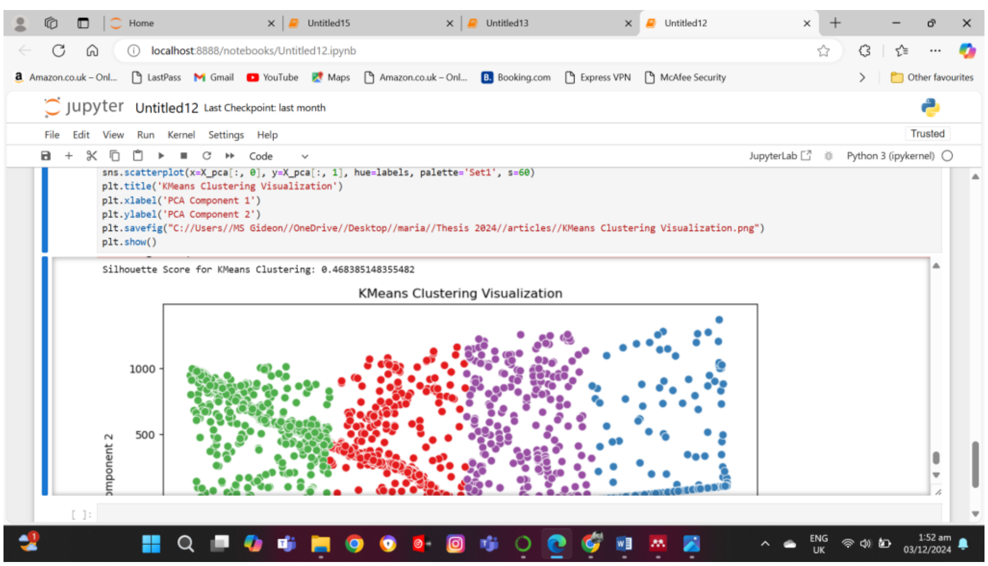

jupyter association rule1 Last Checkpoint: last month
Model: SVM
Confusion Matrix:
[[71   47 195]
 [28 306   67]
 [306 180 1584]]
Classification Report:
precision recall f1-score support
0 0.18 0.23 0.20 313
1 0.57 0.76 0.66 401
2 0.86 0.77 0.81 2070
accuracy 0.70 2784
macro avg 0.54 0.59 0.55 2784
weighted avg 0.74 0.70 0.72 2784
[1]:
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules
import numpy as np
import matplotlib.pyplot as plt
import networkx as nx
from mlxtend.frequent_patterns import association_rules
# Load the dataset
dataset_path = "C://Users//MS Gideon//OneDrive//Desktop//maria//Thesis 2024//data//Crash details with road conditions 2018 to 2023.xlsx"

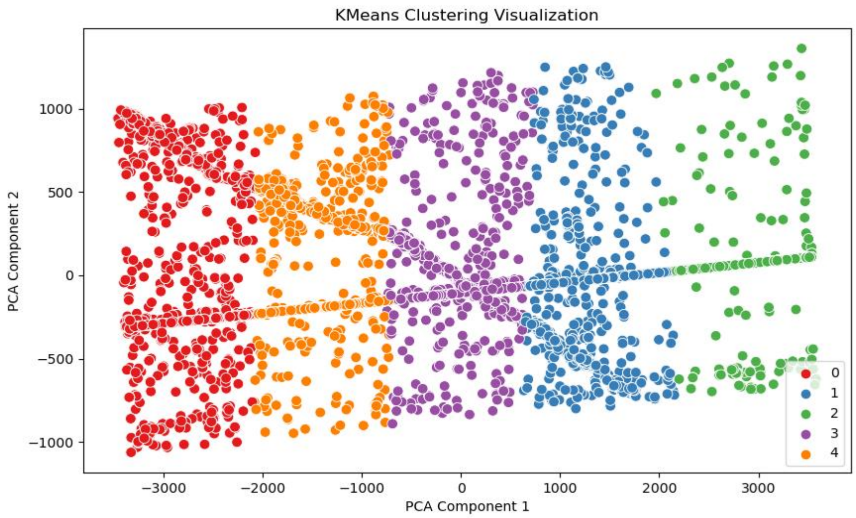

jupyter Untitled13 Last Checkpoint: last month
# ROC Curve
plt.figure(figsize=(10, 8))
for model_name, (y_true, y_pred, model) in models.items():
if hasattr(model, "predict_proba"):
y_score = model.predict_proba(X_test)[:, 1]
else:
y_score = model.decision_function(X_test)
fpr, tpr, _ = roc_curve(y_true, y_score, pos_label=1)
roc_auc = auc(fpr, tpr)
plt.plot(fpr, tpr, label=f'{model_name} (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc='lower right')
plt.savefig(f'{save_path}roc_curve.png')
plt.show()
# Clustering using KMeans
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(X)
labels = kmeans.labels

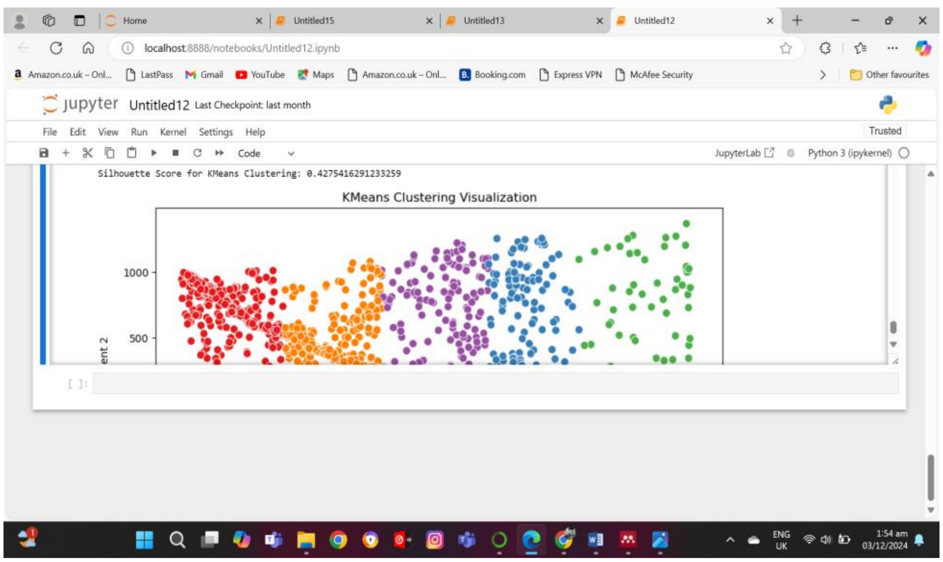
5. Clustering charts

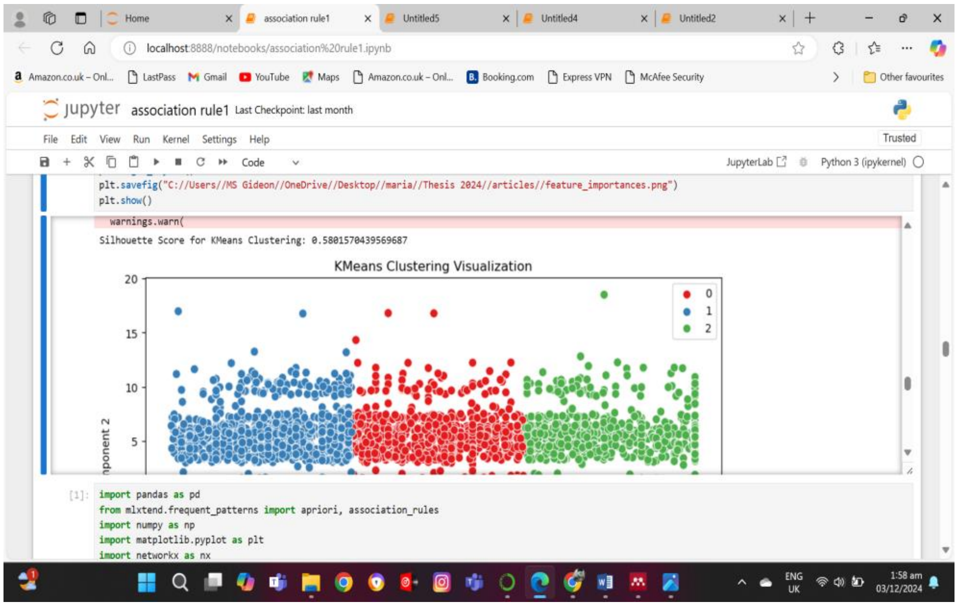
The study wanted to choose to best n for k-means clustering, therefore, below is the K-Means Clustering when n=3,4,5and the score of each clustering



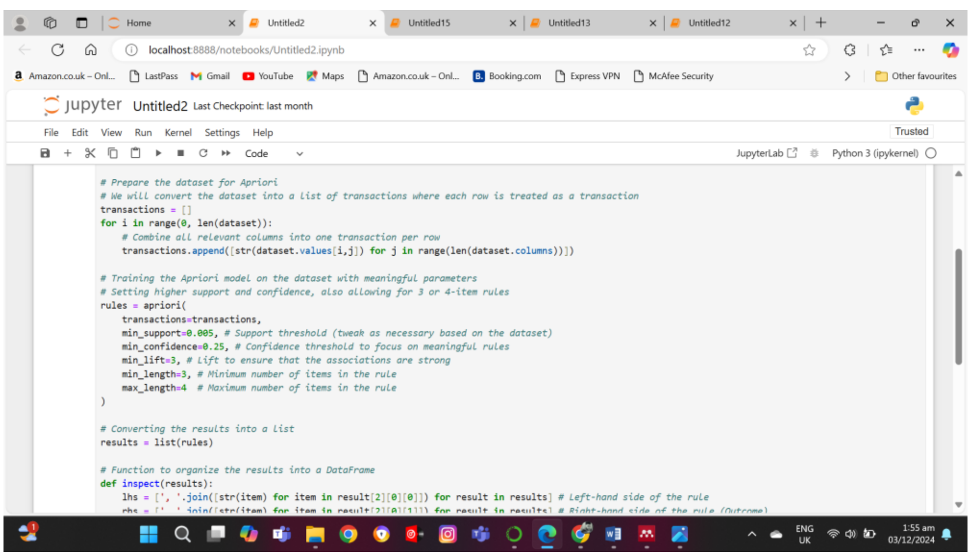



KMeans Clustering Visualization
PCA Component 2
PCA Component 1
1000
500
0
−500
−1000
−3000
−2000
−1000
0
1000
2000
3000
0
1
2
3
4


Home
Untitled15
Untitled13
Untitled12
localhost:8888/notebooks/Untitled12.ipynb
jupyter Untitled12 Last Checkpoint: last month
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Silhouette Score for KMeans Clustering: 0.4275416291233259
KMeans Clustering Visualization
1000
500
1:54 am
03/12/2024

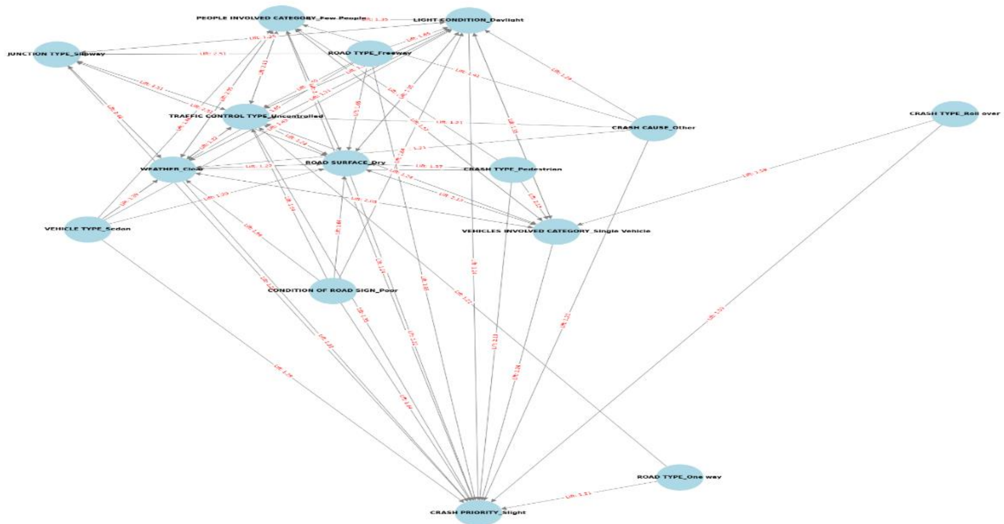


6. Codes: Apriori rules



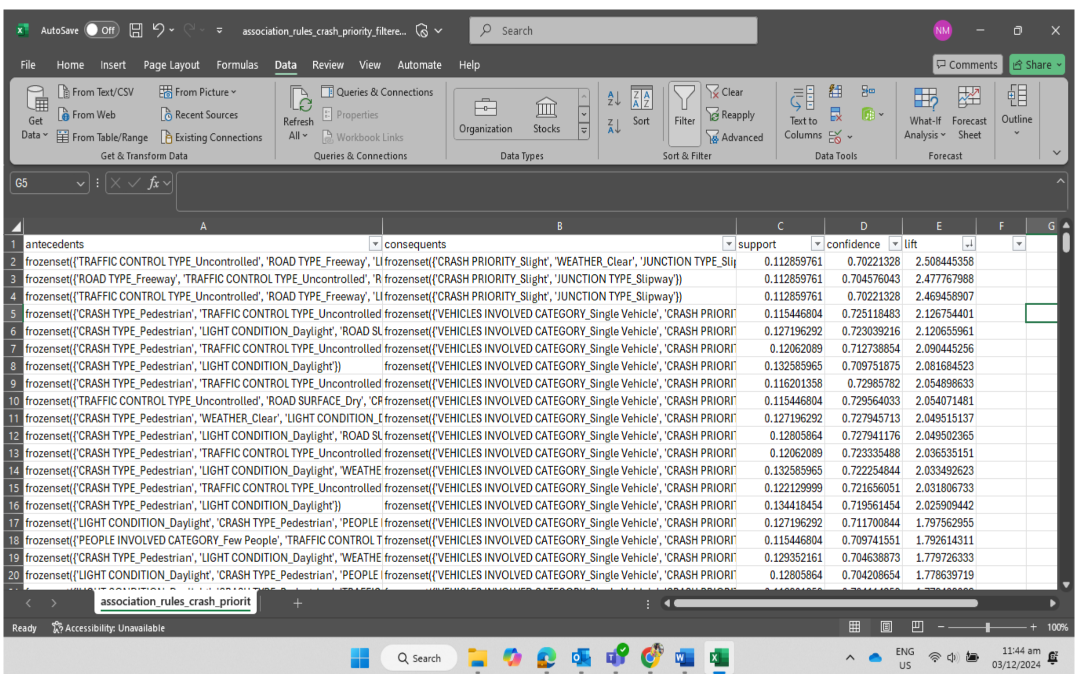
7. Network chart for the apriori rule

The chart below is the network chart that display how rules are connected to one another and indicated that most of the factors combined result in crash priority (severe) every time

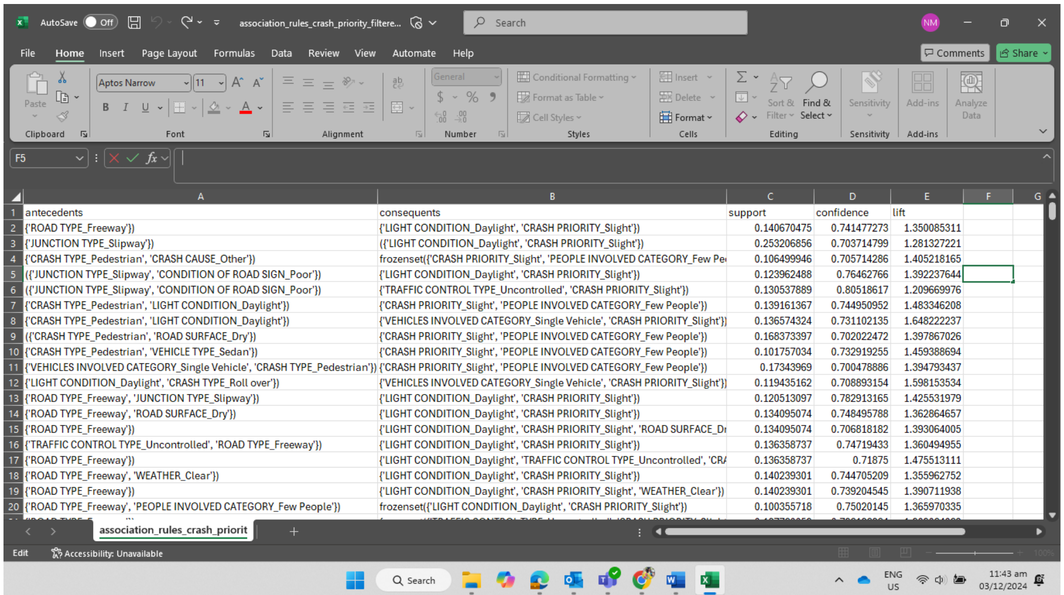


8. Results: Apriori rule

Since python was used for the analysis, to get the results for the rule, it was saved as an excel fine, and below is the screenshot of some of the rules generated



	antecedents	consequents	support	confidence	lift
2	frozenset({'TRAFFIC CONTROL TYPE_Uncontrolled', 'ROAD TYPE_Freeway', 'LI	frozenset({'CRASH PRIORITY_Slight', 'WEATHER_Clear', 'JUNCTION TYPE_Sli	0.112859761	0.70221328	2.508445358
3	frozenset({'ROAD TYPE_Freeway', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'R	frozenset({'CRASH PRIORITY_Slight', 'JUNCTION TYPE_Slipway'})	0.112859761	0.704576043	2.477767988
4	frozenset({'TRAFFIC CONTROL TYPE_Uncontrolled', 'ROAD TYPE_Freeway', 'LI	frozenset({'CRASH PRIORITY_Slight', 'JUNCTION TYPE_Slipway'})	0.112859761	0.70221328	2.469458907
5	frozenset({'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.115446804	0.725118483	2.126754401
6	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight', 'ROAD SU	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.127196292	0.723039216	2.120655961
7	frozenset({'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.12062089	0.712738854	2.090445256
8	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.132585965	0.709751875	2.081684523
9	frozenset({'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.116201358	0.72985782	2.054898633
10	frozenset({'TRAFFIC CONTROL TYPE_Uncontrolled', 'ROAD SURFACE_Dry', 'CR	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.115446804	0.729564033	2.054071481
11	frozenset({'CRASH TYPE_Pedestrian', 'WEATHER_Clear', 'LIGHT CONDITION_[	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.127196292	0.727945713	2.049515137
12	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight', 'ROAD SU	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.12805864	0.727941176	2.049502365
13	frozenset({'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.12062089	0.723335488	2.036535151
14	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight', 'WEATHE	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.132585965	0.722254844	2.033492623
15	frozenset({'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.122129999	0.721656051	2.031806733
16	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.134418454	0.719561454	2.025909442
17	frozenset({'LIGHT CONDITION_Daylight', 'CRASH TYPE_Pedestrian', 'PEOPLE I	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.127196292	0.711700844	1.797562955
18	frozenset({'PEOPLE INVOLVED CATEGORY_Few People', 'TRAFFIC CONTROL T	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.115446804	0.709741551	1.792614311
19	frozenset({'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight', 'WEATHE	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.129352161	0.704638873	1.779726333
20	frozenset({'LIGHT CONDITION_Daylight', 'CRASH TYPE_Pedestrian', 'PEOPLE I	frozenset({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORIT	0.12805864	0.704208654	1.778639719



	antecedents	consequents	support	confidence	lift
2	{'ROAD TYPE_Freeway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.140670475	0.741477273	1.350085311
3	{'JUNCTION TYPE_Slipway'})	({'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.253206856	0.703714799	1.281327221
4	{'CRASH TYPE_Pedestrian', 'CRASH CAUSE_Other'})	frozenset({'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few Pe	0.106499946	0.705714286	1.405218165
5	({'JUNCTION TYPE_Slipway', 'CONDITION OF ROAD SIGN_Poor'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.123962488	0.76462766	1.392237644
6	({'JUNCTION TYPE_Slipway', 'CONDITION OF ROAD SIGN_Poor'})	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRASH PRIORITY_Slight'})	0.130537889	0.80518617	1.209669976
7	{'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People'})	0.139161367	0.744950952	1.483346208
8	{'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORITY_Slight'})	0.136574324	0.731102135	1.648222237
9	({'CRASH TYPE_Pedestrian', 'ROAD SURFACE_Dry'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People'})	0.168373397	0.702022472	1.397867026
10	{'CRASH TYPE_Pedestrian', 'VEHICLE TYPE_Sedan'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People'})	0.101757034	0.732919255	1.459388694
11	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH TYPE_Pedestrian'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People'})	0.17343969	0.700478886	1.394793437
12	{'LIGHT CONDITION_Daylight', 'CRASH TYPE_Roll over'})	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORITY_Slight'})	0.119435162	0.708893154	1.598153534
13	{'ROAD TYPE_Freeway', 'JUNCTION TYPE_Slipway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.120513097	0.782913165	1.425531979
14	{'ROAD TYPE_Freeway', 'ROAD SURFACE_Dry'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.134095074	0.748495788	1.362864657
15	{'ROAD TYPE_Freeway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight', 'ROAD SURFACE_Dr	0.134095074	0.706818182	1.393064005
16	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'ROAD TYPE_Freeway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.136358737	0.74719433	1.360494955
17	{'ROAD TYPE_Freeway'})	{'LIGHT CONDITION_Daylight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRA	0.136358737	0.71875	1.475513111
18	{'ROAD TYPE_Freeway', 'WEATHER_Clear'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.140239301	0.744705209	1.355962752
19	{'ROAD TYPE_Freeway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight', 'WEATHER_Clear'})	0.140239301	0.739204545	1.390711938
20	{'ROAD TYPE_Freeway', 'PEOPLE INVOLVED CATEGORY_Few People'})	frozenset({'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.100355718	0.75020145	1.365970335

9. Apriori rule: Tables: High lift

Rule	Antecedents	Consequents	Support	Confidence	Lift
1	{'CRASH TYPE_Head Side'})	{'LIGHT CONDITION_Daylight', 'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Moderate'}	0.467	0.914	2.912
2	{'CRASH TYPE_Head rear collision'})	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Slight'}	0.568	0.815	2.904
3	{'ROAD SURFACE_Dry', 'CRASH TYPE_Head Side'})	{'VEHICLES INVOLVED CATEGORY_Few Vehicles', 'CRASH PRIORITY_Slight'})	0.514	0.812	2.787
4	{'CRASH TYPE_Pedestrian', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'LIGHT CONDITION_Daylight', 'ROAD SURFACE_Dry'}	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People', 'WEATHER_Clear'}	0.415	0.725	2.126
5	{'ROAD TYPE_Freeway', 'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'LIGHT CONDITION_Daylight'	0.412	0.733	1.635
6	{'CONDITION OF ROAD SIGN_Poor', 'LIGHT CONDITION_Daylight',	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRASH PRIORITY_Slight'})	0.422	0.814	1.223

	'JUNCTION TYPE_Slipway'}				
7	{'JUNCTION TYPE_Slipway', 'LIGHT CONDITION_Daylight', 'CRASH CAUSE_Overtaking'})	{'CRASH PRIORITY_Slight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'WEATHER_Clear'})	0.500	0.771	1.214
8	{'VEHICLE TYPE_Sedan', 'JUNCTION TYPE_Slipway'})	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRASH PRIORITY_Moderate', 'ROAD SURFACE_Dry'}	0.520	0.732	1.213
9	{'ROAD TYPE Freeway'}	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRASH PRIORITY_Slight', 'ROAD SURFACE_Dry'}	0.437	0.726	1.202
10	{'VEHICLE TYPE Sedan', 'LIGHT CONDITION Daylight', 'WEATHER Clear', 'JUNCTION TYPE Slipway'}	{'TRAFFIC CONTROL TYPE Uncontrolled', 'CRASH PRIORITY Slight'}	0.515	0.799	1.200

10. Apriori: High Support

Rule	Antecedents	Consequents	Support	Confidence	Lift
1	{'JUNCTION TYPE_Slipway'}	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'}	0.849	0.803	1.281
2	{'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'LIGHT CONDITION_Daylight', 'CRASH PRIORITY_Slight'})	0.834	0.707	1.288
3	{'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'WEATHER_Clear', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'LIGHT CONDITION_Daylight'})	0.752	0.700	1.275
4	{'CRASH TYPE_Pedestrian', 'LIGHT CONDITION_Daylight'})	{'CRASH PRIORITY_Slight', 'PEOPLE INVOLVED CATEGORY_Few People', 'WEATHER_Clear'})	0.736	0.732	1.528
5	{'CONDITION OF ROAD SIGN_Poor', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'TRAFFIC CONTROL TYPE_Uncontrolled', 'WEATHER_Clear'	0.640	0.765	1.288
6	{'VEHICLE TYPE_Sedan', 'JUNCTION TYPE_Slipway', CRASH CAUSE_Overtaking})	{'TRAFFIC CONTROL TYPE_Uncontrolled', 'CRASH PRIORITY_Slight',	0.6202	0.732	1.213
7	{'CONDITION OF ROAD SIGN_Poor',	{'CRASH PRIORITY_Slight',	0.6182	0.766	1.572

	'ROAD SURFACE_Dry', 'JUNCTION TYPE_Slipway'})	'TRAFFIC CONTROL TYPE_Uncontrolled', 'LIGHT CONDITION_Daylight'})			
8	{'CRASH TYPE_Pedestrian', 'CRASH CAUSE_Speeding', 'LIGHT CONDITION_Daylight', 'WEATHER_Clear'})	({'VEHICLES INVOLVED CATEGORY_Single Vehicle', 'CRASH PRIORITY_Moderate', 'ROAD SURFACE_Dry'})	0.617	0.703	1.776
9	{'CRASH CAUSE_Recklessand Negligent', 'ROAD TYPE_Freeway', 'JUNCTION TYPE_Slipway'})	{'CRASH PRIORITY_Slight', 'WEATHER_Clear', 'LIGHT CONDITION_Daylight', 'ROAD SURFACE_Dry'})	0.612	0.755	1.507
10	{'JUNCTION TYPE_Slipway', 'WEATHER_Clear', 'CRASH CAUSE_Reckless and Negligent})	{'CRASH PRIORITY_Slight', 'LIGHT CONDITION_Daylight'})	0.601	0.701	1.277

11. Code: Descriptive Statistics

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the cleaned dataset
file_path = r'C: \Users\MS Gideon\OneDrive\Desktop\maria\Thesis 2024\Traffic_Crash_Data.csv'
df = pd.read_csv(file_path)

# Display the first few rows of the dataset to ensure it loaded correctly
print(df.head())

# Convert 'DATE OCCURED' to datetime format and extract year and month
df['DATE OCCURED'] = pd.to_datetime(df['DATE OCCURED'], errors='coerce')
df['Year'] = df['DATE OCCURED'].dt.year
df['Month'] = df['DATE OCCURED'].dt.month

# Number of accidents per year
accidents_per_year = df['Year'].value_counts().sort_index()

plt.figure(figsize=(10, 6))
sns.lineplot(x=accidents_per_year.index, y=accidents_per_year.values, marker='o', color='purple')
for index, value in enumerate(accidents_per_year.values):
   plt.text(accidents_per_year.index[index], value, str(value), ha='center', va='bottom')
plt.title('Number of Accidents per Year')
plt.xlabel('Year')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Number of accidents per month
accidents_per_month = df['Month'].value_counts().sort_index()
```

```
plt.figure(figsize=(10, 6))
sns.lineplot(x=accidents_per_month.index, y=accidents_per_month.values, marker='o', color='purple')
for index, value in enumerate(accidents_per_month.values):
    plt.text(accidents_per_month.index[index], value, str(value), ha='center', va='bottom')
plt.title('Number of Accidents per Month')
plt.xlabel('Month')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# Most common types of accidents
accident_types = df['CRASH TYPE'].value_counts().sort_values(ascending=False)

plt.figure(figsize=(12, 8))
sns.barplot(x=accident_types.values, y=accident_types.index, palette='viridis')
for index, value in enumerate(accident_types.values):
    plt.text(value, index, str(value), ha='left', va='center')
plt.title('Most Common Types of Accidents')
plt.xlabel('Number of Accidents')
plt.ylabel('Accident Type')
plt.tight_layout()
plt.show()

# Types of vehicles involved in accidents
vehicle_types = df['VEHICLE TYPE'].value_counts().sort_values(ascending=False)

plt.figure(figsize=(12, 8))
sns.barplot(x=vehicle_types.values, y=vehicle_types.index, palette='viridis')
for index, value in enumerate(vehicle_types.values):
    plt.text(value, index, str(value), ha='left', va='center')
plt.title('Types of Vehicles Involved in Accidents')
plt.xlabel('Number of Accidents')
plt.ylabel('Vehicle Type')
plt.tight_layout()
plt.show()

# Number of accidents by region
accidents_by_region = df['REGION'].value_counts().sort_values(ascending=False)

plt.figure(figsize=(12, 8))
sns.barplot(x=accidents_by_region.values, y=accidents_by_region.index, palette='viridis')
for index, value in enumerate(accidents_by_region.values):
    plt.text(value, index, str(value), ha='left', va='center')
plt.title('Number of Accidents by Region')
plt.xlabel('Number of Accidents')
plt.ylabel('Region')
plt.tight_layout()
plt.show()


# Number of accidents by day of the week
accidents_by_day_of_week = df['DAY OF WEEK OCCURED'].value_counts().sort_index()

plt.figure(figsize=(10, 6))
sns.barplot(x=accidents_by_day_of_week.index, y=accidents_by_day_of_week.values, palette='viridis')
for index, value in enumerate(accidents_by_day_of_week.values):
    plt.text(index, value, str(value), ha='center', va='bottom')
plt.title('Number of Accidents by Day of the Week')
```

```
plt.xlabel('Day of the Week')
plt.ylabel('Number of Accidents')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
# Convert 'Date Occurred' to datetime format if not already done
df['Date Occurred'] = pd.to_datetime(df['Date Occurred'], errors='coerce')

# Extract the year from 'Date Occurred'
df['Year'] = df['Date Occurred'].dt.year

# One-hot encoding categorical variables if not already done
day_of_week_columns = [col for col in df.columns if col.startswith('Day Of Crash_')]
accidents_by_day_of_week = df[day_of_week_columns].sum().sort_index()

# Renaming the columns for better readability
accidents_by_day_of_week.index = accidents_by_day_of_week.index.str.replace('Day Of Crash_', '')

# Set the order of the days of the week
ordered_days = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
accidents_by_day_of_week = accidents_by_day_of_week.reindex(ordered_days)

plt.figure(figsize=(14, 10))
sns.barplot(x=accidents_by_day_of_week.index, y=accidents_by_day_of_week.values, palette='viridis', label='Total
Accidents')
for index, value in enumerate(accidents_by_day_of_week.values):
plt.text(index, value, str(value), ha='center', va='bottom', fontsize=12, fontweight='bold', color='black')

# Prepare data for line plot
df['Day of Week'] = df['Date Occurred'].dt.day_name()
day_order = ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
accidents_per_day_year = df.groupby(['Year', 'Day of Week']).size().unstack().reindex(columns=day_order)

# Plot line graphs for each year
colors = sns.color_palette('tab10', n_colors=len(accidents_per_day_year.index))
for idx, year in enumerate(accidents_per_day_year.index):
  sns.lineplot(x=accidents_per_day_year.columns, y=accidents_per_day_year.loc[year], marker='o', label=year,
color=colors[idx], linewidth=2.5)
  for day, value in accidents_per_day_year.loc[year].items():
    plt.text(day, value, f'{int(value)}', ha='center', va='bottom', fontsize=12, fontweight='bold', color=colors[idx])

plt.title('Number of Accidents by Day of the Week')
plt.xlabel('Day of the Week')
plt.ylabel('Number of Accidents')
plt.legend(title='Year', fontsize=12, title_fontsize=14)
plt.xticks(rotation=45, fontsize=13)
plt.yticks(fontsize=13)
plt.tight_layout()
plt.show()
```

Appendix B: Ethics clearance certificate

1. Ethics Clearance

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

FACULTY RESEARCH ETHICS COMMITTEE (F-REC)

DECISION/FEEDBACK ON THE RESEARCH PROPOSAL

Dear **Maria Ngolo (218094701)**

RESEARCH TOPIC: A MACHINE LEARNING-DRIVEN APPROACH FOR ACCIDENT PREDICTION AND TRAFFIC SAFETY ANALYSIS IN NAMIBIA.

Supervisor (if applicable): **Dr Richard Maliwatu**

Qualification registered for (if applicable): Master of Data Science

(Reference number of applications: **FACULTY RESEARCH ETHICS COMMITTEE REGISTRATION NUMBER: FREC - 10/24**)

Re: Ethical screening application No: FREC - 10/24

The Faculty of **Computing and Informatics** Ethics Screening Committee of the Namibia University of Science and Technology reviewed your application for the above-mentioned research. The research as set out in the application has been:

Approved	X

(Indicate with an X, and N/A if not applicable and proceed)

We would like to point out that you, as a researcher, are obliged to maintain the ethical integrity of your research, adhere to the ethical guidelines of NUST, and remain within the scope of your research proposal and supporting evidence as submitted to the F-REC. Should any aspect of your research change from the information as presented to the F-REC, which could affect the possibility of harm to any research subject, you are under the obligation to report it immediately to your supervisor or F-REC as applicable in writing. Should there be any uncertainty in this regard, you must consult with the F-REC.

We wish you success with your research and trust that it will make a positive contribution to the questfor knowledge at NUST.

Any ethical issues that need to be highlighted?	Why are these issues important?	What must/could be done to minimize the ethical risk?
No	N/A	N/A

Recommendation: The application is approved.

Sincerely,

Prof. Suama L Hamunyela
Chairperson: Faculty Ethics Screening
CommitteeTel: +264-61-207-2922
CC: Co-supervisor: None

NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY
P/Bag 13388 Windhoek NAMIBIA
2024 -04- 11
OFFICE OF THE DEAN
FACULTY OF COMPUTING & INFORMATICS

2. Information Usage Disclaimer

The disclaimer below was signed to be provided with data from NRSC

Information Usage Disclaimer

The National Road Safety Council (NRSC) is a statutory body established in terms of section 2 of the National Road Safety Act, Act 9 of 1972. The functions of the NRSC include amongst others, the collection of information in connection with road safety and the making available thereof to authorities and persons concerned and the public generally;

Copy right

1. Unless otherwise specified, the NRSC is the exclusive and exclusive owner of all rights, licenses, titles, and interests in and to all lists, libraries, databases, images, compilations, files, and other material developed and posted for inclusion in our data system, including ownership of any trade secrets or copyright pertaining thereto.
2. Without the prior consent of the NRSC, no license will be granted to edit, publish, transmit, create any derivative work from, sell, distribute, commercially use, or otherwise exploit, in whole or in part, data, information, geographical attributes, and reports.
3. This information is provided free of charge to relevant stakeholders for the intended purposes and third parties subject to the following disclaimer:

Disclaimer:

1. The NRSC, its consultants, and its officers specifically state that this product/ document/information is for the intended purposes only.
2. All right reserved. No part of the document or products or information may be reproduced, translated, or transmitted, in any form means of electronic, mechanical, photocopying, recording, or otherwise, without prior written permission of the NRSC.
3. This product is for the intended purpose and not for legal, engineering or surveying purposes.
4. This information is provided with the understanding that, while every effort is made to ensure that the data, information, graphics, etc. are accurate and up to date, the NRSC, its consultants, and officers accept no responsibility for accuracy or correctness, or for any consequences of relying on it by way of the conclusions drawn from such information.
5. NRSC accepts no liability, however arising, for any use of said information or reliance placed on it.

Signature

Chief Road Safety Research Officer
Sub Division Road Safety Research

Receiver Details

MariaNgolo