UHASSELT
KNOWLEDGE IN ACTION

School of Transportation Sciences

Master of Transportation Sciences

Master's thesis

Utilising Machine learning and statistical approach to predict near-miss accidents using i-Dreams data

Kizito Lule Ssentongo

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences, specialization Traffic Safety

SUPERVISOR :
Prof. dr. Muhammad ADNAN

CO-SUPERVISOR :
dr. Muhammad Wisal KHATTAK

2024
2025

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

School of Transportation Sciences

Master of Transportation Sciences

Master's thesis

Utilising Machine learning and statistical approach to predict near-miss accidents using i-Dreams data

Kizito Lule Ssentongo

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences, specialization Traffic Safety

SUPERVISOR :
Prof. dr. Muhammad ADNAN

CO-SUPERVISOR :
dr. Muhammad Wisal KHATTAK

Acknowledgments

I would like to express my gratitude to everyone who has supported me throughout the course of writing this thesis.

I am deeply grateful to my supervisor, Dr. Muhammad Adnan, for his guidance, valuable insight, and continuous encouragement throughout the entire research process. His expertise and direction have been instrumental in shaping this work.

I am also grateful to all the professors and staff of the Master of Transportation Sciences (Traffic Safety) at Hasselt University, whose teaching and feedback have laid the foundation for this research. Special thanks to all the contributors of the iDreams project, whose hard work and dedication have led to my gaining access to the naturalistic driving data used in this study.

I would also like to extend thanks to my fellow students and friends for the support, collaboration and input during the long days of writing and analysis.

Thank you all.

Kizito Lule Ssentongo
June, 05, 2025

Abstract

Near-miss events are high-risk driving situations that do not result in crashes but have a high potential to cause harm. In the field of road safety, these events are increasingly recognised as valuable indicators for proactive safety analysis. In light of this, this study investigates appropriate modelling methods to predict the frequency of the occurrence of these events using naturalistic driving data derived from the EU-funded project iDreams. This study specifically focuses on headway-related near-miss incidents with the aim of identifying the key trip-level and driver-level predictors of near-miss frequency. The study evaluates the effectiveness of both statistical and machine learning models in predicting risk.

A dataset comprising 4,481 trips linked to 47 drivers in Belgium was analysed. Each trip was labelled with the count of dangerous headway events. Two statistical models – Poisson and Negative Binomial regression – were developed to estimate associations between near-miss frequency and predictors, including trip duration, distance, average speed, time of day, road environment, driver age, driver experience, and income. Additionally, a Random Forest regressor was also implemented to assess non-linear relationships and feature importance.

Results showed that trip distance (across all the models) and average speed were consistently associated with higher frequencies of near-miss events across all the models. Time of day had minimal influence, even though night or dusk trips were generally linked to fewer near-miss events, possibly due to increased driver caution or low traffic density. Older drivers had lower event frequencies yet reversely driving experience showed a mild positive relationship with near-miss risk. Overall, the Negative Binomial model was a better statistical fit than Poisson, while the Random Forest model achieved strong predictive performance ($R^2 = 0.743$).

Keywords
Machine learning, near-miss accidents, traffic safety, prediction modelling

Word Count
11986

Table of Contents

List of figures

List of tables

1 Introduction

1.1 Background and rationale for the study

According to the World Health Organisation, road traffic crashes are a global health challenge responsible for an estimated 1.19 million deaths in 2023, translating to 15 deaths per 100,000 people. Traffic injuries are also the leading cause of death for persons aged 5-29, and they result in significant economic and social loss. With the global road network and motor vehicle fleet size expected to grow significantly in the future, it is increasingly crucial to address road safety issues.

The conventional method of addressing road safety primarily involves the analysis of crash data before an intervention can be deployed. This method, although valuable, has limitations. According to Sarkar, Rao, & Chatterjee (2024), the traditional method is inherently reactive – relying entirely on the occurrence of crashes, which often have severe outcomes. This, according to scholars, can lead to a delayed identification of risks and ultimately hinder timely intervention. In order to take on a more proactive approach to road safety interventions, it is therefore important to use Surrogate Safety Measures (SSMs), e.g Time To Collision (TTC), which are more frequent and serve as indicators of potential safety problems (Sarkar, Rao, & Chatterjee, 2024; Wang, et al. 2024).SSMs aim to predict and prevent future crashes without waiting for an accumulation of sufficient crash data. They are used to evaluate the safety level of traffic systems by focusing on near-miss incidents and other indicators of unsafe interactions among road users (Singh & Das, 2021).

Near-misses can be understood as the performance of an evasive manoeuvre by a driver to avoid a crash, and this manoeuvre may be in the form of sudden braking and/or rapid steering operations without resulting in an accident (Arai et al., 2001; Hanowski et al., 2007). Identifying these near-misses may be done either manually – through data collection techniques like video recordings and field analysis (Siregar, Agah, & Hidayatullah (2018) – or by use of automated tools which enable the automatic identification, classification, and evaluation of near-misses for instance the Federal Highway Administration's Surrogate Safety Assessment Model (SSAM).

Based on findings from a study by Dong, et al. (2024), integrating machine learning (ML) techniques into determining near-miss incidents significantly advances traffic safety analysis by leveraging large datasets and detecting complex safety risks that traditional methods may overlook. Additionally, as evidenced by Lu, Grembek, & Hansen (2022), ML methods process a large amount of data in real time and thus facilitate proactive safety interventions. Despite these benefits, the use of ML in near-miss detection is still understudied and thus requires further research to establish functional models scalable to different traffic systems.

This study, therefore, aims to address this knowledge gap by examining how both ML and statistical methods can be used to predict the frequency of near-miss events in naturalistic driving data. According to Lord & Mannering (2010), including statistical approaches is key since this study models event frequency rather than probability. The study utilises both trip-level and driver-level variables to identify factors that contribute to increased safety risk. The findings are intended to support proactive safety interventions in traffic systems.

1.2 Problem statement

Road crashes are caused by a combination of human, vehicle, and driving environment-related factors. Although advancements have been made to tackle these safety issues, the conventional method that involves the accumulation of crash data is limited in its ability to capture all crash-influencing factors in sufficient detail.

Besides the fact that crash data is often underreported and infrequent, scholars have also highlighted the ethical implications of waiting for crashes to happen before safety interventions are deployed. The use of SSMs presents road safety specialists with an opportunity to adopt more proactive measures. SSMs, especially when applied to naturalistic driving data, help to provide early indicators of potential crash risks by analysing near-miss accidents and any other unsafe interactions.

Naturalistic driving data–collected by smart in-vehicle technology during real-world driving conditions–enables the capture of a comprehensive and continuous dataset on driver behaviour, vehicle trajectory, driving environment (road type, weather), etc. This data is essential in determining SSMs, which in turn helps in understanding the contributing factors to the occurrence of near-miss events and how these events evolve across different driving conditions.

Studies have shown that integrating machine learning (ML) techniques and SSMs is key to enhancing a proactive approach to road safety. In determining SSMs, ML algorithms have been found to significantly outperform conventional statistical methods owing to their accuracy and ability to capture, process, and analyse an extensive range of naturalistic driving data (Dong et al., 2024; Driessen et al., 2024; Behboudi, Moosavi and Ramnath, 2024).

However, despite their significance, Behboudia, Moosavi and Ramnath (2024) state that the integration of ML in SSMs is still a relatively understudied domain. Das et al. (2023) also state that there are research gaps in analysing the suitability of SSMs in mixed traffic situations and that the domain still lacks viable frameworks for combining and using multiple SSMs. In the context of Belgium, near-miss prediction using ML techniques is an understudied field despite the country's unique driving environment and road safety record. Additionally, although some studies model near-misses as a probability outcome, this study adopts a trip-level frequency approach in order to identify contributing factors and gain greater insight into exposure risk patterns (Cai et al., 2021). This trip-level frequency modelling thus warrants the use of statistical modelling methods and, in the context of this study, offers the opportunity to compare the effectiveness of both ML and statistical methods.

This study, therefore, aims to apply ML and statistical approaches to naturalistic driving data from Belgium in order to develop a tailor-made and context-aware near-miss prediction methodology to enhance the precision of traffic safety assessments.

1.3 Research objectives and questions

This study utilises data from the iDreams (Intelligent Driver and Road Environment Assessment and Monitoring System) project. iDreams is a road safety initiative funded by the European Union with an aim of developing an integrated platform that helps prevent drivers from getting too close to the boundaries of unsafe vehicle operations, thus keeping them within a "Safety Tolerance Zone". The iDreams platform works by integrating real-time sensor data, physiological monitoring, and environmental conditions to dynamically assess driving risk and provide context-aware interventions to drivers both during and after trips. In the development of the platform, naturalistic driving data was collected from vehicles fitted with a range of sensors and equipment in five countries: Belgium, Germany, Greece, Portugal, and the United Kingdom. Vehicles were driven under typical real-world driving conditions, capturing a range of data related to the vehicle, driver, and environmental conditions, such as speed, distance, time, driver fatigue and distraction, weather conditions, etc.

This study utilises data captured in Belgium, and, therefore, the following are the objectives;

- To investigate how trip-level and driver-level characteristics influence the frequency of near-miss events in iDreams naturalistic driving data.
- To assess the statistical relationship between key predictors and the frequency of dangerous headway events.
- To develop and evaluate both a statistical and a machine learning model for predicting the frequency of near-miss events.
- To draw a comparison between the performance of the statistical approach and the machine learning model.
- To aid future safety assessment frameworks by identifying the most influential features that contribute to an elevation in driving risk levels.

In order to achieve the above-mentioned objectives, the following research questions have to be answered;

1. Which trip-level and driver-level characteristics are mainly associated with the frequency of near-miss events during naturalistic trips?
2. To what extent can statistical approaches explain the variation in the frequency of near-miss events using trip data and driver attributes?
3. How does the performance of a machine learning model compare to a statistical approach in predicting near-miss events?
4. What insights can be drawn from both approaches (machine learning and statistical) about the most relevant contributors to near-miss driving behaviour?

2 Literature Review

2.1 Near-Miss Accidents and SSMs

According to the USA Occupational Safety and Health Act (OSHA), a near-miss is an incident in which no property was damaged and no personal injury was sustained but where, given a slight shift in time or position, damage or injury easily could have occurred. A near-miss is an unplanned event that can precede events in which a loss or injury could occur (Siregar, Agah, & Hidayatullah, 2018). Near-misses may also be referred to as "near accidents," “near hits,” or "near collisions." They can also be defined as sudden braking and rapid steering operations by the driver without resulting in an accident (Arai , et al. 2001). A near-miss accident is understood as the performance of an evasive manoeuvre by the driver to avoid a vehicle accident (Hanowski et al., 2007).

Near accidents are fundamentally unclear and can potentially be interpreted in two ways: either as a wakeup call highlighting a potential source of danger or as a success indicating that margins were good enough (McMullen & Markman, 2000; Dillon & Tinsley, 2008; as cited by Terum & Svartdal, 2019).

The Swedish Traffic Conflict Technique (TCT) Observer Manual, which was published in 2018, also classifies near-miss accidents as traffic conflicts. The manual elaborates on the concept of the “Safety Pyramid” that was introduced by Hydén (1987). According to this manual, the traffic process can be seen as a number of elementary events which differ in their degree of severity.

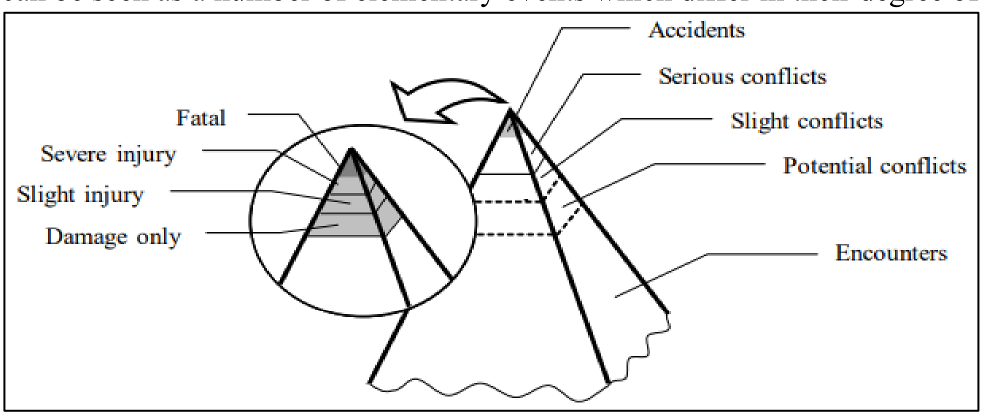

Figure 1: Safety Pyramid (Hydén, 1987)

The lower part of this pyramid represents the normal and safe interactions that happen most of the time between road users. At the extreme, the top of the pyramid represents the most severe events, such as fatal or injury-leading accidents. These severe events are also known to be very infrequent in comparison with the total number of events. According to this TCT, a conflict’s severity is defined at the moment when one of the road users starts taking an evasive action. This conflict severity is based on two indicators:

- Time-to-Accident (TA) - time remaining to a collision when the relevant road user takes the evasive action;
- Conflicting Speed (CS) - speed of the relevant road user when they take the evasive action.

Other Near-miss indicators

According to Jiang, et al. (2021), the common indicators of traffic conflicts can typically be in 3 forms. The first form of these indicators measures risk aversion behaviour. It determines whether there is a conflict by observing whether or not an aversion behaviour exists, as well as the severity based on the urgency. These assessments are mostly qualitative in nature, for example, steering manoeuvres and evident deceleration. The second form of indicators measures the proximity in space and time, with the two most common indicators being the time-to-collision (TTC) and the post-encroachment time (PET). According to Li, et al. (2023), TTC is defined as the time remaining before a collision if none of the vehicles involved changes their speeds and directions. PET is the time separation between two vehicles passing each other in a conflict point where the two paths cross each other.

According to Lu, et al. 2021, in a basic TTC calculation, acceleration is considered to be zero. TTC can therefore be determined by the following equation

$$\boxed{TTC = \begin{cases} \frac{gap}{v_f - v_l}, v_f > v_l \\ \infty, v_f \leq v_l \end{cases}}$$

Where;

- v_f - Velocity of the following vehicle
- V_l - Velocity of the leading vehicle
- *gap*- the distance headway minus vehicle length of the leading vehicle
- If $vf \leq vl$, it means that the vehicle is safe, and if $vf > vl$ and TTC ≤ threshold, it means that the vehicle is safe. According to studies cited by Kuang, Qu, & Wang (2015), the TTC threshold has varied between 1.5 s and 4s.

The third form, according to Jiang et al., measures characteristics of the vehicle's own movement, such as deceleration, with the most common indicator of a vehicle's own movement characteristics being the deceleration rate to avoid a crash (DRAC). Lu, et al. (2021) define DRAC as the squared differential speed between a following vehicle and its corresponding leading vehicle, divided by their closing gap. DRAC can also be defined as the minimum deceleration rate required by the following vehicle to come to a timely stop (or match the leading vehicle's speed) and hence avoid a crash (Cooper & Ferguson, 1976; as cited by Kuang, Qu, & Wang, 2015). DRAC can, therefore, be denoted as;

$$\mathrm{DRAC} = \begin{cases} \frac{(V_2-V_1)^2}{D_{1-2}}, \mathrm{if} V_2 > V_1 \\ 0, \mathrm{otherwise} \end{cases}$$

that is,

$$\mathrm{DRAC} = \frac{V_2-V_1}{\mathrm{TTC}}$$

Where V1 is the leading vehicle's speed, V2 is the following vehicle's speed, and D1–2 is the gap between the two vehicles. A higher DRAC value indicates a more dangerous car-following scenario. In general, Kuang, Qu, & Wang (2015) state that TTC is negatively related to DRAC. They also cite 2 thresholds for DRAC, citing The American Association of State Highway and Transportation Officials (AASHTO, 2004), which suggests that a given vehicle is in conflict if its DRAC exceeds a threshold of 3.4 m/s^2. The other, slightly lower threshold, is recommended by Archer (2005) at 3.35 m/s^2 for most drivers.

Besides PET and TTC, Gore et al. (2023) also highlight additional conflict indicators, also called surrogate safety measures (SSM). These SSMs report the space or time proximity between different road users to a projected collision point. They include Deceleration Rate (DR), Proportion of Stopping Distance (PSD), Time Integrated TTC (TIT), and Modified time to collision (MTTC). A study conducted by Vogel (2003) also found that although using TTC provided good sensitivity to traffic changes, using headway as a safety indicator was better in safety enforcement since it is consistent across different locations. A similar view is also held by Ramezani-Khansari, Nejad, & Moogeh (2020), who agree that, in car following scenarios, braking time headway is more stable and significant.

2.2 Factors contributing to near-miss accidents

Near-miss accidents are influenced by a combination of human, vehicle, and environmental-related factors (Bekelcho, et al., 2024; Jomnonkwao, et al., 2023).

2.2.1 Human factors

These factors are pivotal in contributing to near-miss accidents since they have a direct effect on the decision-making process, vehicle control, and reaction times. Key factors include driver distraction and driver fatigue.

Driver Distraction

Distraction while driving is a growing concern, particularly due to the increased use of mobile phones and in-vehicle infotainment systems. Studies indicate that cell phone distraction is responsible for 6% of crashes, while all forms of distraction cause 29% of crashes (Blincoe et al., 2023). According to information from the American National Highway Traffic Safety Administration (NHTSA), 3,308 people were killed in 2022 in vehicle crashes that involved distracted drivers, representing 7.8% of total fatalities. According to the American National Institutes of Health (NIH), doing something else while driving e.g. eating, talking on the phone, texting, etc, increases the risk of a crash. A series of NIH-funded studies found that novice drivers were up to 8 times more likely to crash or have a near-miss when dialling a phone or reaching for a phone or other object. While adults were more than twice as likely to crash or have a near miss when dialling (NIH, 2014).

Driver Fatigue

Driver fatigue is a critical contributing factor to near-miss incidents since it significantly impairs a driver's ability to safely operate a vehicle. Fatigue reduces alertness, affects decision-making, and slows reaction time, thereby increasing the risk of crashes or near-misses (NHTSA, 1998; RoSPA, 2024). Fatigue-impaired driving may lead to several errors, such as unintentional lane departure, missing traffic signals, and a delayed response to sudden traffic hazards. Considering the dangers that it poses, several studies have highlighted that driver fatigue is highly prevalent. A systematic review of research articles on road crashes related to driver sleepiness, published between 2000 and 2020, found that 28% of drivers experienced at least one episode of uncontrollable sleepiness while driving. Additionally, approximately 5% of these drivers had a crash or near-miss accident due to sleepiness (Saleem, 2022). According to data from the NHTSA, it is estimated that in 2017, there were 91,000 reported crashes in the USA involving sleepy drivers. These crashes led to over 50,000 injuries and nearly 800 fatalities (NHTSA, 2017). In Europe, a survey conducted by the European Transport Workers' Federation (ETF) found that 60% of truck drivers reported driving while feeling fatigued, and around 30% of them fell asleep at the wheel at least once within the previous year (Vitols & Voss, 2021). Similarly, a 2010 study of French drivers found that 28% of them had experienced at least one episode of severe sleepiness while driving in the previous 12 months, and 11% reported at least one near-miss accident for the same period, with close to half of these incidents being sleep-related (Sagaspe, et al., 2010).

2.2.2 Vehicle Dynamics

This encompasses factors such as speed, acceleration, braking performance, and steering. Driving at excessive speeds and rapidly accelerating can significantly contribute to near-miss accidents. High-speed driving reduces the available time for a driver to identify and react to unexpected events in traffic, and this increases the likelihood of a crash or near-miss incident. A study by Guillen, et al. (2020) analysed automobile insurance telematics and found that speeding is associated with an increased risk of acceleration events, which are an indicator of near-miss incidents. Similarly, the study by Bekelcho et al. (2024), which analysed near-miss incidents among truck drivers in southern Ethiopia, found that about 72% of truckers had experienced a near-miss accident and that the majority of these near-misses (26%) were attributed to speeding.

Another key element in vehicle dynamics is harsh/sudden braking. These events are typically characterised by rapid deceleration and are indicative of traffic emergency situations where a driver must swiftly react in order to avoid a crash (Bagdadi, 2013). These manoeuvres can destabilise the vehicle and may lead to a loss of control. Additionally, according to research, drivers in urban settings may be more prone to sudden stopping owing to the nature of traffic in city environments (Guillen, et al. 2020). Similarly, sudden steering inputs, which are often performed to avoid unexpected obstacles, can compromise vehicle stability and may represent an even higher risk scenario than sudden braking. A study by Smith, Najm, & Lam (2003) found that drivers engaged in last-second steering actions at shorter distances compared to their braking responses.

2.2.3 Environmental factors

These are aspects directly affecting the driving environment. They may be weather conditions and road-related factors. Adverse weather conditions like rain, fog, snow, or ice have a significant influence on the occurrence of traffic crashes and, inherently, near-miss incidents. For instance, rain and fog can impair visibility and thus make it challenging for drivers to judge distances and see obstacles, while snow and ice create slippery surfaces that make it difficult to control vehicles, especially during sudden stops. According to data from the NHTSA, between 2007 and 2016, weather-related factors were responsible for 21% of vehicle crashes, 19% of injuries, and 16% of crash fatalities in the USA. Findings from Bekelcho et al. (2024) also revealed that the combination of reduced visibility and slippery road surfaces during adverse weather conditions increased the risk of driver errors and near-miss events.

Road-related factors such as road layout and design, road geometry, and road surface conditions also play a role in the frequency of near-miss accidents. According to findings from the study by Siregar, Agah, and Hidayatullah (2018), complex road junctions with unconventional designs may pose a challenge to drivers and are thus prone to high traffic conflict rates. Concerning road geometry, Garnaik, Giri, and Panda (2023) identify a direct relationship between several road geometric factors and an increased risk of crashes or near-misses. For example, small radius horizontal curves, increased levels of superelevation, limited sight distance, narrow lanes, etc., were all found to increase risks related to speeding, erratic manoeuvring, improper steering, and loss of traction. Additionally, poor road conditions, such as potholes, uneven surfaces, pavement roughness and skid resistance, directly affect ride comfort, can distract drivers, lead to a loss of vehicle control and, ultimately, potential near-misses (Bekelcho et al., 2024; Mkwata & Chong, 2022).

2.3 Machine learning in near miss prediction

Identifying and predicting near-miss accidents is critical in order to achieve proactive traffic safety management. Traditionally, approaches that predict near-miss incidents often employ statistical techniques such as regression analysis, time series analysis, and correlation studies. However, these methods may be limited when handling large volumes of heterogeneous data for complex and non-linear interactions such as near misses (Obasi & Chizubem, 2023). A systematic review of statistical models by Slikboer et al. (2020) found that traditional methods had a reduced predictive performance due to several reasons, including poor variable selection and inadequate validation.

Machine learning techniques are, thus, more effective alternatives that are capable of finding patterns within large datasets, for instance, naturalistic driving data. Algorithms such as Random Forest, Decision Trees, and other deep-learning models have been applied to predict the severity of injuries and near-miss accidents. For instance, the study by Obasi & Chizubem (2023) utilised a stacked sparse autoencoder (SSAE) that demonstrated a high accuracy when predicting injury severity. A comparative study by Hossain et al. (2021) also found that ensemble machine learning methods, e.g. XGBoost and random forest, effectively handle complex data structures and outperform traditional statistical methods in predicting accident severity. These conclusions are consistent with the findings made by Wahab & Jiang (2019) and Sufian & Varadarajan (2023), all of whom found that machine learning models had a better and more accurate predictive performance compared to traditional models, thus highlighting the potential of using ML in traffic safety analysis.

This study used a two-pronged modelling approach employing both a machine learning and a statistical approach. For machine learning, the Random Forest model was used to predict the frequency of near-miss events across different trips. This prediction was based on both trip-level and driver-level characteristics such as trip duration, distance, trip average speed, time of day, driver age, and driver experience. This approach handles both numerical and categorical variables, offers feature importance ranking, and is flexible.

Poisson regression was selected as the statistical modelling approach because the dependent variable – frequency of near-miss events – is a count variable that follows a distribution more consistent with Poisson assumptions. According to Akram et al. (2023) and Coxe et al. (2009), applying standard linear models to count outcomes may lead to biased and inconsistent estimates, and as such, Poisson regression models provide more appropriate analyses for count data.

2.4 ML methodologies best suited for near-miss prediction in iDreams

The iDreams dataset contains a large volume of multimodal naturalistic driving data that includes GPS readings, driver fatigue levels, vehicle dynamics (braking, steering, acceleration), and other factors like speeding, headway distance, lane departures, readings for forward collision warning, and overtaking events. The safety and driving events' metrics are recorded over time and thus require a sequential analysis.

Additionally, the data is imbalanced, with the near-miss events expected to be generally fewer compared to normal driving instances/incidents. This imbalance may negatively affect the performance of the ML methodology as the machine learning model may become biased toward the majority class, thus leading to poor detection of rare near-miss incidents. Therefore, this will necessitate the use of class-balancing techniques, as illustrated by (Chawla, et al. 2002).

The table below shows the final recommended ML methods to detect and predict near-misses in iDreams data

ML Method	**Best Suited For**	**Advantages**	**Challenges**	**iDreams Data Type**	**Source(s)**
Random Forest	Predictive modelling of dangerous event frequency	Handles non-linear interactions and ranks feature importance	May overfit without tuning, it is also less interpretable than Generalised Linear Models	Naturalistic driving data (for instance, as per this study, e.g. distance, average speed, duration, etc.)	Xue et al. (2019) Hossain et al. (2021)
SMOTE + XGBoost	Handling class imbalance in near-miss prediction	Improves recall in rare event detection	Over-sampling may introduce synthetic noise	Data that shows rare events like tailgating, late braking, extreme cornering, etc.	Chawla, et al. (2002)

XGBoost (Tree-based model)	Feature-based near-miss prediction	High accuracy and interpretable insights	Less effective for sequential event modelling	Braking force, lane deviation, and steering angle variance	Iranitalab & Khattak (2017)
LSTM networks	Time-series driver behaviour analysis, e.g. acceleration patterns, delays in reaction time.	Captures temporal dependencies and is useful for real-time near-miss detection	Requires a large dataset, thus making it computationally expensive	Driver fatigue, acceleration, braking, steering, and reaction time	Zhang, Yang, & Yang (2023) Mili et al. (2023)
Isolation Forest (IF)	Detecting high-risk driving patterns	Unsupervised detection of outliers	May flag too many false positives	Driving anomalies, e.g. sudden speed drops	Liu, Ting, & Zhou (2009)
Autoencoders (Anomaly Detection)	Detecting unknown near-miss events	Works well for unsupervised detection of rare events	Needs fine-tuning for threshold selection	Lane departures, sudden braking, erratic acceleration patterns	Dong et al. (2018) Ip, Artur, & Mihaita (2024)

Table 1: ML methods for iDreams data

3 Methodology

This study involved both a literature review and modelling. The literature review was conducted with academic material that was identified and sourced from conventional academic databases like Google Scholar and Research Gate. Owing to the specific nature of the study, AI tools like ChatGPT were also used to search and identify relevant academic publications.

The primary focus of the search was on publications related to statistical or machine learning modelling, with a preference for studies conducted on crash or near-miss prediction, definitions and indicators of near-miss accidents, and the use of models in naturalistic driving data.

3.1 Overview of modelling approach

This study adopts a quantitative research design that is based on the analysis of secondary naturalistic driving data collected from a set of Belgium-based drivers as part of the iDreams project. This study's key aim is to use both statistical and machine-learning techniques to predict the frequency of near-miss incidents. It specifically focuses on driving events classified as *"Headway Level 2"*. Analysis in this study is conducted at the trip level, with each individual trip considered as a separate entry for analysis.

This study employs a two-part modelling approach. Initially, a Poisson regression model is used to explore the statistical relationships between the frequency of near-miss events and trip-level & driver-level characteristics. However, based on diagnostic indicators from the Poisson model, particularly high values for deviance and Pearson chi-square statistics, overdispersion was detected. Therefore, a Negative Binomial regression model was applied using the same predictors. Afterwards, a Random Forest model is applied to assess the predictive power of the same features through a machine learning framework.

The original dataset consisted of information on over 17,000 trips, from which **4,966** trips were selected based on their association with one of 53 drivers. It is from these drivers that key driver characteristics were extracted, e.g. driver age and length of their driving experience. After further filtration, trips without complete driver information were excluded, and this reduced the number of trips to **4,481**, which was the final modelling dataset.

Data processing, feature extraction, and model development were carried out using Python and supported by MS Excel. Only trips that included at least one dangerous event were selected and retained for modelling. This selection criteria is consistent with prior near-miss prediction studies, e.g. Sun et al. (2024). Since near-miss events are more frequent in naturalistic driving datasets, trips with at least one dangerous event offer necessary variation for identifying risk factors and also prevent interfering with the model's ability to detect patterns due to class imbalance.

The figure below provides a schematic overview of the entire methodology for this study

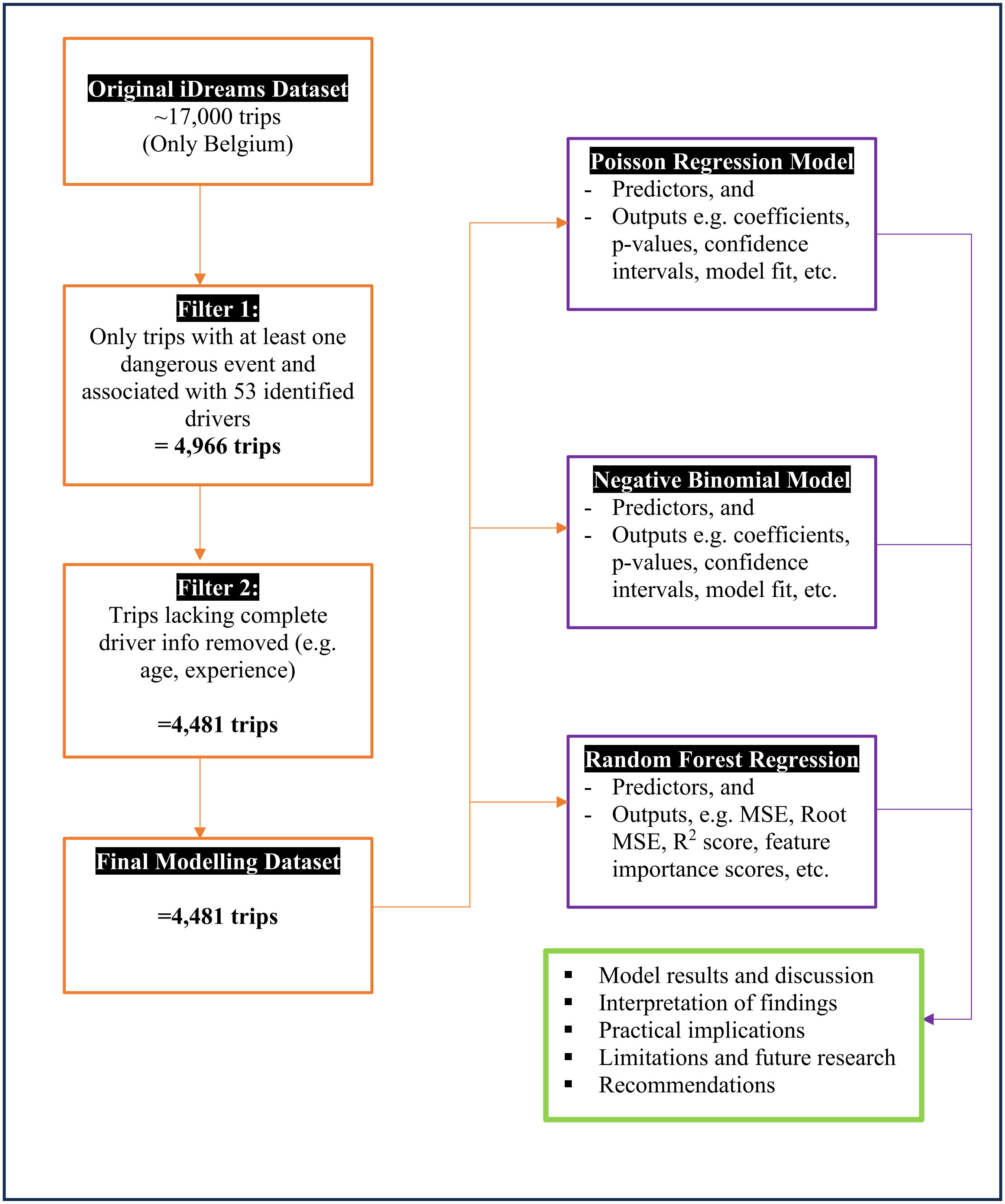

Figure 2: Methodology Schematic Representation

3.2 Data Source

This study uses trip-level naturalistic driving data from the iDreams project. The project involved, among other things, the use of instrumented vehicles equipped with advanced in-vehicle monitoring technologies across several European countries, including Belgium.

The vehicles were fitted with various sensors and telematics systems, including the Mobileye driver assistance system. Mobileye collected high-frequency data on forward vehicle dynamics, thus allowing for the measurement of continuous time headway readings. These time headway values were then discretised into 4 safety levels:

- -1: *Invalid or missing value*
- 0: *Safe Headway*
- 1: *Moderate caution*
- 2: *High risk (dangerous event)*

In this study, **level 2** is used as a stand-in for near-miss incidents and therefore serves as the dependent variable. The trip files for each trip analysed contained both contextual and telemetry information from which trip-level features were extracted e.g. duration, speed, distance. The trips that were retained were those in which the trip driver identifier (*short_id*) matched a valid entry in the driver dataset. After filtering, the final dataset used for modelling included 4,481 trips representing 47 unique drivers.

3.3 Dependent Variable

In this study, the dependent variable is the number of dangerous events per trip. This variable is defined as the count of *headway level 2* readings within each individual trip. These events are extracted from the driving data using Python scripts. These Python scripts scan the JSON trip files for occurrences of the value **2** under the field for headway level labelled **data_ME_AWS_hw_level.**

Within the iDreams dataset, this level 2 is used as an indicator of a high-risk following distance between a subject vehicle and the vehicle ahead. This categorisation is based on discretised time headway measurements provided by the Mobileye ADAS sensor. Level 2 represents driving situations where the time headway drops below a critical safety threshold, signalling a potential near-miss condition that could escalate into a collision if an evasive action is not taken.

In order to ensure model relevance, only trips that included at least one *headway level 2* event were retained in the dataset. The total number of these events for each trip was computed and recorded as a numerical value (**dangerous_events).** This value, per trip, serves as the dependent variable for all the three models.

3.4 Independent Variables

This study uses both trip-level and driver-level features as independent variables that may influence the frequency of the dangerous headway events (**dangerous_events**). The selection of these variables was primarily influenced by their availability, interpretability, and their overall relevance to prior research on driving risk research. All the variables were extracted from the JSON trip files and driver questionnaire data using Python and MS Excel. The complete list of trips and their associated variables were compiled into a single CSV file. The final variable selection is as follows;

3.4.1 Trip-level variables

At the start of the analysis, variables like *distraction events* and *weather* were considered. However, they were excluded from the final model due to inconsistent data recording and a low volume of valid entries across the dataset. As such, the following trip-level variables were considered;

- **Duration** (seconds): Total time of the trip, extracted from the trip JSON file and validated by calculating trip start and end times.
- **Day/Night Indicator** (binary): Derived from the field **data_ME_AWS_time_indicator.** Since the original format of this variable was text, it was recoded to 0 for "day" and 1 for "night" or "dusk".
- **Distance** (kilometres): Total trip distance as recorded in the trip data.
- **Average Speed** (km/h): Trip duration is converted to hours, and this variable is computed by dividing distance by duration.
- **Road Environment**: Categorisation that indicates the road type on which a trip occurred. Road environment categories were proportionally assigned to trips based on each driver's self-reported weekly driving exposure across 3 road types, i.e. urban, rural, and motorway (coded as 1, 2, 3, respectively, for modelling). Trips were first sorted in descending order of average speed per driver and road environment categories assigned based on the driver's reported exposure proportions e.g. for a driver with 100 trips and a reported driving exposure of 60% urban, 30% rural, and 10% motorway, after sorting the trips by descending average speed, 60 trips would be assigned code 1, 30 trips code 2, and 10 trips code 3.

3.4.2 Driver-level variables

- **Age** (years): Age of the driver at the time of data collection.
- **Gender** (binary). Binary coding of recorded driver genders (0 = male, 1 = female).
- **Driving Experience** (years): Number of years that the driver had been licensed at the time of data collection.
- **Income** (euro): A range of monthly earnings per driver. To enable modelling, the ranges were converted into midpoint values as follows:

Original Range	Midpoint Used
Less than €1.000	500
€1.000 to €2.000	1500
€2.000 to €3.000	2500
€3.000 to €4.000	3500
€4.000 to €5.000	4500
Over €5.000	5500

Table 2: Income Ranges and Midpoints

- **Driving range**: This is an estimated weekly driving distance per driver recorded as a categorical range. To enable modelling, these ranges were converted into midpoint values as follows:

Original Range	Midpoint Used (km)
Up to 50 km	25
50 – 100 km	75
100 – 500 km	300
500 – 1000 km	750

Table 3: Weekly Driving Distance and Midpoints

For the final dataset, only drivers for whom these variables were available were included. Trips with incomplete driver information were excluded, resulting in a total of **4,481 trips** associated with **47 drivers.**

3.5 Feature Extraction

Due to storage and computation constraints, the data was handled in batches of 100-200 files, allowing for efficient extraction. The following key variables were extracted from each trip file:

- short_id (unique driver identifier)
- duration
- distance
- data_ME_AWS_hw_level (used to count instances of dangerous events = headway level 2)
- data_ME_AWS_time_indicator (used to classify time of day; “day, night, or dusk”)

Driver-related features were extracted from the iDreams questionnaire and matched to trips using the **short_id** field. Only trips from drivers present in this field were retained. Python’s pandas library was used to merge and verify data consistency.

3.6 Modelling Techniques

This study employed Poisson regression, Negative Binomial regression and Random Forest regression modelling approaches to predict the frequency of near-miss events / dangerous events (*headway level 2*) per trip. All models were developed and executed using Python.

3.6.1 Count Regression Models

Poisson regression was selected as an initial approach because the dependent variable – **dangerous_events** – is a non-negative count variable. The model estimates how trip-level and driver-level features influence the expected number of near-miss events. A Generalised Linear Model (GLM) framework with a Poisson distribution and log link function was applied using the *statsmodels* package in Python.

The following were the included predictors and their mathematical relationships:

- [+] Duration (in seconds)
- [+] Distance (in kilometres)
- [+] Average speed (in km/h)
- [-] Day/night indicator (recoded: 0 = day, 1= night, and dusk)
- [+] Road environment (ordinal categories; 1= urban, 2 = rural, 3 = motorway)
- [-] Driver age (in years)
- [-] Driver gender (0=male, 1=female)
- [+] Driver income (midpoints of ranges in euro)
- [+] Driving experience (in years)
- Range of weekly driving distance (midpoints of ranges in km)

The outputs of the model included regression coefficients, statistical significance (p-values), and exponentiated coefficients (Exp (B)), confidence intervals, and diagnostics regarding model fit. However, during the analysis, the Poisson model showed signs of overdispersion and therefore, in order to account for this, a Negative Binomial regression model was estimated using the same predictors. The Negative Binomial model produced improved fit results and was found to be better suited for this dataset compared to the Poisson model.

3.6.2 Random Forest Regression

In order to account for potential nonlinearities, feature interactions, and to complement the statistical model, a Random Forest Regressor was developed using Python's **scikit-learn** library. This ML technique builds multiple decision trees and averages their outputs to improve accuracy and reduce overfitting.

Just like in the Poisson model, the same predictor variables were used for this Random Forest model. However, here the dataset was split into a training set (80%) and a testing set (20%) to validate model performance.

Evaluation of the model focused on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score, and feature importance scores.

4 Results

This chapter presents the findings derived from the descriptive analysis and modelling of near-miss events using the filtered iDreams dataset. The aim is to examine how trip-level and driver-level features relate to the frequency of near-miss events, i.e. *headway level 2* events. The results provide an overview of the dataset and a presentation of the model outputs (both statistical and machine learning).

4.1 Descriptive Statistics

4.1.1 Overview of the Dataset

The main iDreams dataset contained well over 17,000 trips. From this, 4966 trips were randomly extracted based on 2 criteria;

- The trip file had to contain one of the 53 unique driver identifiers (short_id) that were found the iDreams driver dataset
- The trip had to have at least one timestamped near-miss event, i.e. value 2 under data_ME_AWS_hw_level in the trip JSON file.

For modelling purposes, these 4,966 trips were further reduced to 4,481 after excluding trips from drivers whose information was incomplete (e.g. age, and experience). These 4,481 trips were taken by a total of 47 unique drivers.

4.1.2 Driver Characteristics

a) Age and experience;

- Drivers' ages ranged from 20 to 79 years (mean age = 47, median age = 43, SD = 18)

Age Range	No. of Drivers
20 – 30	12
31 – 40	7
41 – 50	9
51 – 60	4
61 – 70	9
71 – 80	6
	47

Table 4: Driver Age Range

- Drivers' years of experience ranged from 2 to 55 years (mean = 27, median = 24, SD = 17)

Experience Range (years)	No. of Drivers
2 – 10	12
11 – 20	4
21 – 30	12
31 – 40	4
41 – 50	10
50+	5
	47

Table 5: Driver Experience Range

b) Gender; 30 male, 17 female. 69% of the trips were taken by males while 31% were taken by females.

c) Income;

Income Range	No. of Drivers
Less than €1.000	1
€1.000 to €2.000	4
€2.000 to €3.000	11
€3.000 to €4.000	9
€4.000 to €5.000	12
Over €5.000	7
Unstated	3
	47

Table 6: Driver Income Range

d) Weekly road exposure

This is a range that shows the reported range of weekly driving distances per driver.

Average Weekly Driving Distance	No. of Drivers
Up to 50 km	1
50 – 100 km	9
100 – 500 km	26
500 – 1000 km	9
Unknown	2
	47

Table 7: Range of weekly driving distance

4.1.3 Trip Characteristics

- Trip duration: The original dataset featured trip durations captured in seconds. The shortest trip analysed was about 1.5 minutes (94 seconds), and the longest trip lasted approximately 5 and a half hours. [mean = 24 minutes, median duration = 17 minutes SD = 21 (minutes)].
- Trip distance: Distances ranged from 0.3 km to 615 km. (mean distance = 23 km, median distance = 12.5 km, SD = 28.7).
- Average speed: This category was not expressly available in the dataset; however, it was calculated and included per trip [$\frac{\text{Distance x 3600}}{\text{Duration}}$]. Average speeds ranged from 2.8 km/h to 113 km/h. The median value of average speed across the trips was approximately 46 km/h.
- Road environment: After assigning trip road types based on driver information, 29% of trips were on urban roadways, 44% on rural, and 27% on motorways.
- Time of day: Trips were categorised by driving conditions; day, dusk, and night. For modelling purposes, both dusk and night were similarly coded.

Time of day	Percentage share of trips
Day	89%
Night	9.5%
Dusk	1.5%

Table 8: Trip time of day distribution

4.1.4 Near-miss Event Distribution

There were a total of 62,903 recorded near-miss events across all the trips. Male drivers accounted for about 93% of these events, with females at 7%. After assigning road environment categories to the trips, over 90% of all near-miss incidents were in the urban roadways. The median number of dangerous events was 7. The distribution of these dangerous events is skewed, with event counts per trip ranging from 1 at the lowest to 303 at the highest. This is further illustrated by the fact that around only 13% of all the trips account for over 50% of recorded near-miss events.

4.2 Count Regression Model Output

The study initially used a Poisson regression model to examine the relationship between the frequency of near-miss events and both trip-level characteristics & driver demographics. This modelling approach was deemed appropriate owing to the count nature of the dependent variable, which followed a discrete, non-negative distribution. However, the results showed signs of overdispersion, and thus, to counter this effect, a Negative Binomial regression was also conducted.

Both models used 4,481 trips, with the occurrence of the near-miss event being the dependent variable while trip and driver characteristics were the independent variables. The models were run on a CSV file that contained all the necessary data for each of these trips. The table below shows how each of these variables was labelled in the CSV file.

Variable	Column Label in CSV
Near-miss event count	dangerous_events
Trip duration (in seconds)	duration
Time of day indicator	day_night
Trip distance (in kilometres)	distance
Trip average speed (in km/h)	average_speed
Road environment category	road_environment
Driver age (in years)	age
Driver experience (in years)	experience
Driver gender (binary)	gender
Driver weekly driving distance	stc_weekly
Driver income range	income

Table 9: Count Regression Model Variables

4.2.1 Poisson Regression Results

This model was initially selected due to the discrete, non-negative nature of the dependent variable. The modelling was done in Python using the statsmodels library. A log-link Poisson model was selected. The table below shows a summary of the model output

Metric	Value
Log-likelihood	-24,033
Deviance	31,797
Pearson chi-square	36,200
Degrees of freedom	4,285
Pseudo R-squared	0.9999

Table 10: Poisson Regression Results

Based on these outputs, it was hypothesised that the high values for deviance and the Pearson-chi square suggested the presence of overdispersion within the dataset. It is likely that the variance of the outcome exceeded the mean, and therefore, a more flexible model would be appropriate.

The table below shows coefficient estimates for the Poisson model

Variable	Coefficient	Std. Error	p-value	Exp(β)	95% CI for Exp(β)
Intercept	2.1324	0.071	<0.001	8.435	[7.35, 9.69]
Duration	6.66e–10	2.21e–09	0.763	≈1.00	[1.000, 1.000]
Distance	0.0120	0.000	<0.001	1.012	[1.012, 1.012]
Average Speed	0.0139	0.000	<0.001	1.014	[1.013, 1.015]
Time of Day	-0.3748	0.013	<0.001	0.688	[0.670, 0.706]
Driver Age	-0.0489	0.003	<0.001	0.952	[0.946, 0.959]
Experience	0.0450	0.003	<0.001	1.046	[1.039, 1.053]
Gender	-0.0529	0.010	<0.001	0.949	[0.930, 0.968]
Weekly distance	1.97e-05	1.88e-05	0.295	1.000	[1.000, 1.001]
Driver Income	1.22e-05	3.76e-06	0.001	1.000	[1.000, 1.000]
Road environment	0.1683	0.010	<0.001	1.183	[1.161, 1.206]

Table 11: Poisson Model Coefficients

- Distance and average speed were found to be significantly associated with an increased frequency of near-miss events. Trip duration was not a statistically significant predictor.

4.2.2 Negative Binomial Regression Results

Due to indications of overdispersion in the Poisson regression model, a Negative Binomial regression model was estimated. This model is suitable for count data that exhibits high levels of variability that may not be well-handled by the Poisson model.

The same predictor variables were used, and the model was fitted using the Generalised Linear Model (GLM) framework in Python.

The table below shows a summary of the model output

Metric	Value
Log-likelihood	-14,200
Deviance	2,407
Pearson chi-square	2,800
Degrees of freedom	4,285
Pseudo R-squared	0.5153

Table 12: Negative Binomial Regression Results

These results are a substantial improvement over the Poisson model. All key metrics significantly reduced and thus suggesting a better-fitting model.
The table below shows coefficient estimates for the Negative Binomial model

Variable	Coefficient	Std. Error	p-value	Exp(β)	95% CI for Exp(β)
Intercept	2.199	0.243	<0.001	9.019	[5.61, 14.50]
Duration	5.33e–10	4.89e–09	0.913	≈1.00	[1.000, 1.000]
Distance	0.0250	0.001	<0.001	1.025	[1.023, 1.027]
Average Speed	0.0105	0.002	<0.001	1.011	[1.007, 1.014]
Time of Day	-0.4575	0.053	<0.001	0.633	[0.570, 0.703]
Driver Age	-0.0437	0.011	<0.001	0.958	[0.936, 0.978]
Experience	0.0367	0.012	0.002	1.037	[1.014, 1.062]
Gender	-0.1456	0.040	<0.001	0.865	[0.798, 0.935]
Weekly distance	-0.0003	7.58e-05	<0.001	1.000	[1.000, 1.001]
Driver Income	-1.112e-05	1.41e-05	0.431	1.000	[1.000, 1.000]
Road environment	0.1308	0.036	<0.001	1.140	[1.062, 1.224]

Table 13: Negative Binomial Model Coefficients

The Negative Binomial model confirmed most of the trends that were previously observed in the Poisson regression. However, it offered more reliable estimates due to an improvement in the handling of overdispersion. Key findings from this model include:

- The baseline expected count of near-miss events corresponds to the intercept value
- Trip duration was not a statistically significant predictor of near-miss event frequency.
- A 1 km increase in trip distance is associated with a corresponding 2.5% increase in expected number of near-miss events.
- Each 1 km/h increase in average speed slightly increases the expected number of near-miss events by 1%.
- Night or dusk trips have up to 37% lower near-miss events compared to daytime trips
- Female drivers are associated with approximately 13.6% fewer near-miss events compared to males.
- Drivers moving from lower speed roads into higher speed roads, e.g. driving from urban to rural or rural to motorway, increase potential near-miss event risk by close to 14%
- The income of the driver had no statistically significant relationship with near-miss events.
- Each additional year in the age of the driver reduces the expected amount of near-miss incidents by approximately 4.3%
- Reversely, for each extra year of driving experience, the data showed a 3.2% increase in the expected number of near-miss events.

Note: The effects of driver age and experience on the frequency of near-miss events appear to be countering each other. Thus, there is a need for a more detailed study into the specific ‘cut-off points’ at which an increase in both age and/or experience leads to a unidirectional movement in expected near-miss event frequency.

4.3 Random Forest Model Results

In addition to the count regression models, a Random Forest regressor was developed to assess the predictive strength of the same set of trip-level and driver-level variables. This was chosen due to the model's capability in handling complex, non-linear relationships and interactions. The dataset, variables, and their arrangement were the same as for the count regression models.

The dataset was randomly split into 2 sets; 80% was used as a training set while the remaining 20% was a test set. A Random Forest regressor with 100 trees was trained, and the performance of the model was evaluated on the test set.

The table below shows the performance:

Metric	Value
Mean Squared Error (MSE)	96.93
Root Mean Squared Error (RMSE)	9.85
R^2 Score	0.743

Table 14: Random Forest Model Results

The model was able to explain approximately 74.3% of the variance in the number of dangerous events in the unseen data. The RMSE value is suggestive of moderate error in the predicted number of events. It can be hypothesised that this moderate error is due to variability and skewed dispersion in the driving data.

Furthermore, the model extracted feature importance scores to show the relative contribution of each predictor to the overall prediction accuracy.

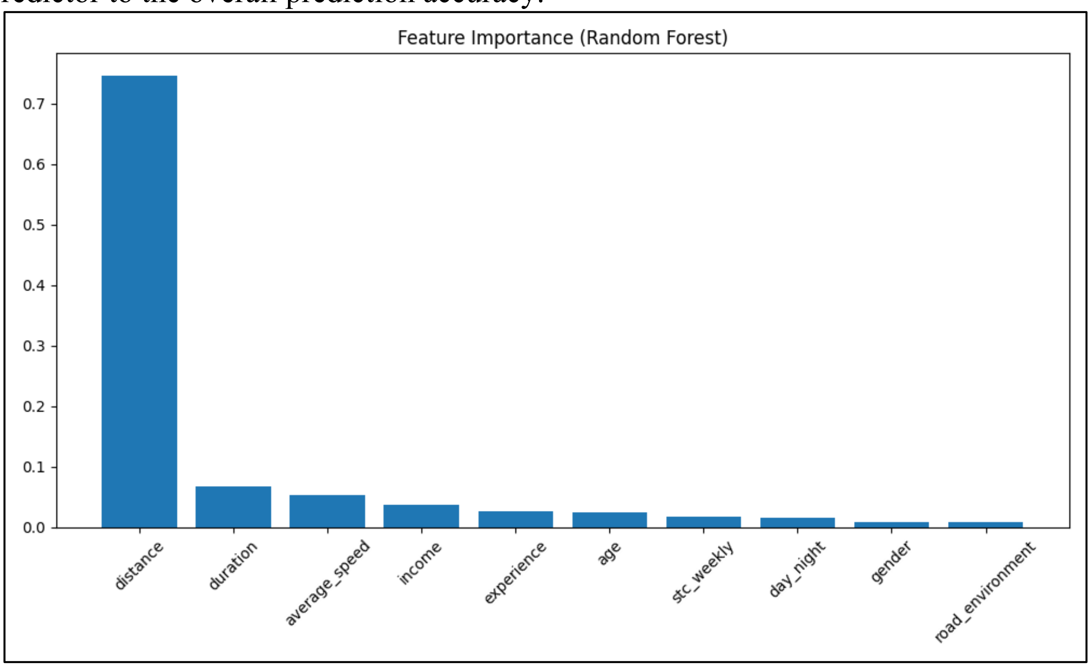

Figure 3: Random Forest Feature Importance Ranking

According to this ranking, trip distance was the most important predictor, greatly contributing to the model's predictions. Other trip-level features, such as duration, average speed, and income, showed moderate influence, while road environment, gender, and time of day contributed relatively little. The most important driver-level predictor was income, while driver gender was not highly ranked.

5 Discussion

This chapter provides an interpretation of the results from both the statistical and machine learning models. The main aim was to investigate which trip-level and driver-level factors had a significant influence on the frequency of near-miss incidents.

5.1 Interpretation of Key Findings

5.1.1 Trip-Level Influences

- Trip distance was found to be the most influential predictor across all models. In both of the statistical models, trip distance was positively associated with near-miss event frequency. This finding was further reinforced by the Random Forest model, which showed that distance accounted for over 70% of model importance. This result from the models supports the literature that longer trips increase exposure time and thus the likelihood of being involved in risky situations (Shen et al., 2020; Bagdadi, 2013).
- Average speed was also found to be a significant risk factor, with higher speeds leading to increased near-miss event frequency. This is also in line with the generally accepted safety viewpoint that links speed with crash risk and near-miss event frequency.
- In a seemingly counterintuitive finding, night/dusk time driving was associated with fewer near-miss events compared to daytime for both models. According to the literature, although the low light conditions in the night time may be considered riskier, this finding likely reflects cautious driving behaviour due to factors like an increase in alertness, presence of law enforcement, & reduced traffic density, etc. (Sun et al., 2024; Masello et al., 2023).

5.1.2 Driver-level Characteristics

- The Random Forest model highlighted that income was the most important driver-level predictor. This possibly points to the influence of socioeconomic factors on driving exposure and risk-taking behaviour.
- Driver age showed a negative relationship with the frequency of near-miss events, with older drivers generally being less involved in dangerous events than younger ones. According to the literature, this finding is consistent with recorded changes in driving behaviour due to increasing age, for example, decreased likelihood for speeding, decreased impaired driving, and a generally more risk-averse driving style (NHTSA, 2023).
- Reversely, driving experience had a mild positive effect on near-miss event frequency. Even though this possibly reflects a level of behavioural complacency or a false sense of confidence, there is still a need for a more thorough determination of actual driving experience. This is because, for modelling purposes, this study considered experience only in years, i.e. how long the driver had been with a driving license and did not consider the actual amount of time and/or different driving scenarios the driver had been exposed to.

5.2 Practical Implications

Findings from this study offer a number of practical implications for the advancement of road safety, policy design, fleet management, and targeted driver behaviour interventions. The study identifies the key factors influencing the frequency of near-miss events and provides evidence to support strategies that mitigate risky driving behaviour before it results in a crash.
The practical implications of this study can be in the following areas;

1. Integrating prediction models into driver assistance systems. The hybrid modelling approach of using both statistical and machine learning models can offer complementary value in the development of driver assistance systems.
2. Tailor-made driver interventions. The study found that older drivers were associated with fewer near-miss events. By focusing on driver characteristics such as age, road safety experts can devise age-sensitive behavioural interventions such as programs targeting novice drivers, refresher courses for drivers with less road exposure, programs that help older drivers adapt to cognitive changes over time, etc.

   Additionally, since more experienced drivers showed a slight increase in near-miss event frequency, these safety interventions should also include campaigns that help discourage overconfidence.
3. Targeted speed management strategies. Findings from the modelling showed a significant and positive association between average speed and frequency of dangerous events. Since speed is one of the most critical behavioural factors contributing to road crashes, findings from this study support the implementation of context-sensitive speed enforcement strategies such as:
   - Adaptive speed warnings within in-vehicle driver feedback systems
   - Use of speed limiters in fleet vehicles
   - Risk-based insurance pricing that factors in speed
   - Targeted public awareness campaigns
4. Policy and road safety research. Results from this study reinforce the need to prioritise trip context and the monitoring of behaviour in the development of road safety programs. Stakeholders such as road safety agencies can use findings from this study as evidence to support the design of behaviorally-informed regulations. There is also a need to invest in the continued collection and analysis of large-scale, real-world driving datasets to make road safety interventions much more proactive.

5.3 Limitations and Future Research

While this study contributes to understanding factors that influence the frequency of near-miss events, it faced some limitations that may point to valuable directions for future research.

- Conducting only a trip-level analysis. Due to resource constraints and data processing complexity, this study conducted analysis only at the trip level rather than at the event level. The dataset used for modelling contained over 60,000 dangerous events (headway level 2), and as such, an event-level analysis was not feasible within the study scope. This led to data aggregation, which may not capture the characteristics of individual near-miss events. For future research, a more "zoomed-in" event-level analysis would better isolate the moment-by-moment features that distinguish dangerous from non-dangerous conditions within trips.
- Missing variables. Even though the study incorporated key trip and river variables, some other equally important behavioural and contextual factors were not included, e.g. driver distraction, fatigue, drowsiness, and environmental conditions (weather, road type, congestion, etc.). These factors significantly influence driver performance, and their absence may reduce the explanatory power of the models.

  Future work on such a dataset should aim to integrate such variables into the model to give an all-round perspective.
- Limited generalisability. The dataset used was based on car drivers in Belgium. The modelling process also did not differentiate vehicle types or driving contexts. This makes the findings very area-specific and harder to generalise to:
  - Other geographical, cultural, or regulatory contexts
  - Different vehicle types, e.g. trucks or buses.
- Reliance on a single surrogate safety indicator. This study uses time headway as its sole measure of near-miss risk. While headway is a well-established metric, its use in isolation may not fully capture complex trip-level driving conditions. It is important, therefore, that similar research considers integrating multiple surrogate safety measures such as Time To Collision (TTC), Post-Encroachment Time (PET), or Deceleration Rate to Avoid Crash (DRAC) in order to build a more multidimensional understanding of risk.
- Cross-sectional nature of the study. The data and information used in the study are simply a "snapshot" of a particular moment in time for the drivers and their environment. It is worth investigating how driving styles may have evolved over time to get a better understanding of the impact that targeted behaviour-based interventions may have on near-miss frequency for the same group of drivers. This can help to further improve proactive road safety measures.
- To create a richer and more robust prediction system, future research can complement the Negative Binomial and Random Forest approaches used in this study with more flexible machine learning models, e.g. Gradient Boosted Trees, XGBoost.

6 Conclusion and Recommendations

An important insight from the study is the comparative difference between the modelling approaches (statistical and machine learning). While some predictors consistently influenced both models, others significantly varied between the two approaches.

For instance, in the Poisson model, a change in the *road environment* variable was associated with a 14% increase in expected near-miss events. Yet, the same variable was ranked as the least important by the Random Forest model. Additionally, *trip duration* was found not to be statistically significant in the count models, yet the Random Forest model ranked it as the second most important predictor.

This is evidence that while count models may be appropriate when explaining relationships or effect sizes, machine learning models are better equipped to identify complex patterns. This further illustrates the importance of adopting a hybrid approach (statistical + machine learning) in order to get a holistic understanding of influencing factors in near-miss prediction.

In conclusion, this study contributes to road and traffic behaviour research in several ways.

- It provides empirical evidence that supports the importance of both trip-level and driver-level features in predicting near-miss risk.
- It demonstrates the value of combining both machine learning approaches and statistical approaches when analysing surrogate safety data, i.e. using both Random Forest regression and Negative Binomial regression.
- It highlights the vital role played by naturalistic driving data for proactive risk modelling. This is key in ensuring that transport planners do not depend entirely on crash-based analyses.

Additionally, based on the findings, the following recommendations are proposed:

- Incorporating trip distance and speed thresholds in systems that monitor drivers and give them feedback during driving.
- Developing real-time risk identification platforms that provide personalised and well-tuned risk assessments and recommendations to drivers.
- Improving the framework upon which naturalistic driving data is collected, processed, shared, and analysed across different vehicle manufacturers, road safety agencies, and researchers. This would help in expanding the availability of this data for key proactive safety interventions.

7 References

Akram, M., Cerin, E., Lamb, K. E., & White, S. R. (2023). Modelling count, bounded and skewed continuous outcomes in physical activity research: beyond linear regression models. *International Journal of Behavioral Nutrition and Physical Activity*.

Al-Shehari, T., Al-Razgan, M. S., Alfaqih, T., Alsowail, R., & Pandiaraj, S. (2023). Insider Threat Detection Model Using Anomaly-Based Isolation Forest Algorithm. *IEEE Access*. doi:10.1109/ACCESS.2023.3326750

Arai , Y., Nishimoto, T., Ezaka , Y., & Yoshimoto, K. (2001). Accidents and Near-Misses Analysis by Using Video Drive-Recorders in a Fleet.

Archer, J. (2005). *Indicators for traffic safety assessment and prediction and their application in micro-simulation modelling: A study of urban and suburban intersections.* Stockholm: Royal Institute of Technology.

Bagdadi, O. (2013). Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F: Traffic Psychology and Behaviour*, 117-126.

Batutin, A. (2024, June 21). *In-depth Guide to Machine Learning (ML) Model Deployment*. Retrieved February 9, 2025, from https://shelf.io/blog/machine-learning-deployment/

Behboudia, N., Moosavib, S., & Ramnath, R. (2024). *Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Reviewof Machine Learning Techniques.* doi:10.48550/arXiv.2406.13968

Bekelcho, T., Birgoda, G. T., Leul, H., Maile, M., Alemayehu, M., & Olani, A. B. (2024). *Near miss road traffic accidents and associated factors among truck drivers in Gamo zone, southern Ethiopia by using a contributory factors interaction model.* Front Public Health. doi:10.3389/fpubh.2024.1386521

Blincoe, L., Miller, T., Wang, J.-S., Swedler, D., Coughlin, T., Lawrence, B., . . . Dingus, T. (2023). *The economic and societal impact of motor vehicle crashes, 2019 (Revised).* National Highway Traffic Safety Administration.

Cai, M., Yazdi, M. A., Mehdizadeh, A., Hu, Q., Vinel, A., Davis, K., . . . Rigdon, S. E. (2021). The association between crashes and safety-critical events: Synthesized evidence from crash reports and naturalistic driving data among commercial truck drivers. *Transportation Research Part C: Emerging Technologies*. doi:https://doi.org/10.1016/j.trc.2021.103016.

Carrascosa, I. P. (2024, December 3). *Integrating Machine Learning into Existing Software Systems*. Retrieved February 8, 2025, from https://www.kdnuggets.com/integrating-machine-learning-into-existing-software-systems

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 321-357.

Cooper, D. F., & Ferguson, N. (1976). Traffic studies at t-junctions – a conflict simulation. *Traffic Engineering & Control*.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment, 91*(2), Abstract. Retrieved from https://doi.org/10.1080/00223890802634175

Das, T., Samandar, M. S., Autry, M. K., & Rouphai, N. (2023). *Surrogate Safety Measures: Review and Assessment in Real-World Mixed Conventional and Autonomous Vehicle Platoons.* doi:10.1109/ACCESS.2023.3248628

Dillon, R., & Tinsley, C. (2008). *How near-misses influence decision making under risk: A Missed Opportunity for Learning.*

Dong, C., Shao, C., Li, J., & Xiong, Z. (2018). An Improved Deep Learning Model for Traffic Crash Prediction. *Journal of Advanced Transportation*.

Dong, Y., Zhang, L., Farah, H., Zgonnikov, A., & Arem, B. v. (2024). *Data-driven Semi-supervised Machine Learning with Surrogate Measures of Safety for Abnormal Driving Behavior Detection.* Delft University of Technology.

Driessen, T., Stefan, D., Heikoop, D., Dodou, D., & Winter, J. d. (2024). Using mobile devices for driving test assessment: a study of acceleration and GPS data. *The International Journal of Transportation Research*. doi:10.1080/19427867.2024.2352198

Filipsson, F. (2024, August 2). *Integrating Machine Learning Platforms with Existing Systems*. Retrieved February 8, 2025, from https://redresscompliance.com/integrating-machine-learning-platforms-with-existing-systems/

Garnaik, M. M., Giri, J. P., & Panda, A. (2023). Impact of highway design on traffic safety: How geometric elements affect accident risk. *Scientific journal of the European Ecocycles Society*, 83-92. doi:10.19040/ecocycles.v9i1.263

Gore, N., Chauhan, R., Easa, S., & Arkatkar, S. (2023). Traffic conflict assessment using macroscopic traffic flow variables: A novel framework for real-time applications. *Accident Analysis and Prevention*.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M., & Elpidorou, V. (2020). *Can automobile insurance telematics predict the risk of near-miss events?* Retrieved from diposit.ub.edu/dspace/bitstream/2445/154515/1/691701.pdf

Hanowski, R. J., Olson, R. L., Bocanegra, J., Hickman, J. S., Dingus, T. A., & D., S. J. (2007). *Critical Incidents that Occur in the 10th and 11th Hour of Driving in Commercial Vehicle Operations: Does Risk Increase in the 11th Hour?* Virginia Tech Transportation Institute .

Hossain, M. A., Ahmed, S., Ray, S. K., & Mbhuiyan, M. (2021). A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity. *ATA Science and Computational Intelligence.* London. doi:10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069

iDreams. (n.d.). *About iDreams*. Retrieved January 20, 2025, from What is iDreams: idreamsproject.eu/wp/

iDreams. (n.d.). *Objectives*. Retrieved January 20, 2025, from About iDreams: https://idreamsproject.eu/wp/project/

Ip, W., Artur, G., & Mihaita, A.-S. (2024). Traffic accident prediction via three-dimensional convolution autoencoder and victim-party demographic data. *30th Anniversary of Transportation Research.* Crete.

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

Jiang, R., Zhu, S., Chang, H., Wu, J., Ding, N., Liu, B., & Qiu, J. (2021). Determining an Improved Traffic Conflict Indicator for Highway Safety Estimation Based on Vehicle Trajectory Data. *Sustainability, 13*(16). doi:http://dx.doi.org/10.3390/su13169278

Jomnonkwao, S., Hantanong, N., Champahom, T., Se, C., & Ratanavaraha, V. (2023). Analyzing Near-Miss Incidents and Risky Riding Behavior in Thailand: A Comparative Study of Urban and Rural Areas. *Safety*(Traffic Safety Culture). Retrieved from https://doi.org/10.3390/safety9040090

Kuang, Y., Qu, X., & Wang, S. (2015). A tree-structured crash surrogate measure for freeways. *Accident Analysis and Prevention, 77*, 137-148.

Li, D., Fu, C., Sayed, T., & Wang, W. (2023). An integrated approach of machine learning and Bayesian spatial Poisson model for large-scale real-time traffic conflict prediction. *Accident Analysis and Prevention*.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2009). Isolation Forest. (pp. 413-422). IEEE.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 291-305.

Lu, C., He, X., van Lint, H., Tu, H., Happee, R., & Wang, M. (2021). Performance evaluation of surrogate measures of safety with naturalistic driving data. *Accident Analysis and Prevention*.

Lu, J., Grembek, O., & Hansen, M. (2022). *Connecting Surrogate Safety Measures to Crash Probablity via Causal Probabilistic Time Series Prediction.* University of California, Berkeley.

Malhotra, R. (2024). *Best Practices for Integrating Machine Learning Models into Software Applications*. Retrieved February 9, 2025, from https://www.valuecoders.com/blog/software-engineering/integrating-machine-learning-models-into-software-applications/

Marqas, R., & Özyurt, F. (2024). Comparative Analysis of LSTM, XGBoost, and Hybrid Approaches in Credit Card Fraud Detection. *Proceeding Book of 4th International Artificial Intelligence and Data Science Congress ICADA.* AS-Proceedings.

Masello, L., Castignani, G., Sheehan, B., Guillen, M., & Murphy, F. (2023). Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis & Prevention*. Retrieved from https://doi.org/10.1016/j.aap.2023.106997

McMullen, M., & Markman, K. (2000). Downward counterfactuals and motivation: The “wake-up call” and the “Pangloss” effect. . *Pers. Soc. Psychol. Bull.*(26), 575–584.

Mili, M., Kerkeni, A., Abdallah, A. B., & Bedoui, M. H. (2023). *ICU Mortality Prediction using Long Short-Term Memory Networks.* Retrieved from https://arxiv.org/abs/2308.12800

Mkwata, R., & Chong, E. E. (2022). Effect of pavement surface conditions on road traffic accident - A Review. *ICCEE.* Kuching. doi:https://doi.org/10.1051/e3sconf/202234701017

Ness, S., Eswarakrishnan, V., Sridharan, H., Shinde, V., Janapareddy, N. V., & Dhanawat, V. (2024). Anomaly Detection in Network Traffic Using Advanced Machine Learning Techniques. IEEE. doi:10.1109/ACCESS.2025.3526988

NHTSA. (1998). *Drowsy Driving and Automobile Crashes*. National Highway Traffic Safety Administration. Retrieved from www.nhtsa.gov/sites/nhtsa.gov/files/808707.pdf

NHTSA. (2017, October). Drowsy Driving 2015. *Traffic Safety Facts*. National Highway Traffic Safety Administration. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812446

NHTSA. (2023). *Countermeasures That Work - Older Drivers*. Retrieved from nhtsa.gov: https://www.nhtsa.gov/book/countermeasures-that-work/older-drivers/understanding-problem

NHTSA. (2024). *Driver Electronic Use in 2023*. National Highway Traffic Safety Administration. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813660

NHTSA. (n.d.). *Distracted Driving*. Retrieved February 1, 2025, from www.nhtsa.gov/risky-driving/distracted-driving

NHTSA. (n.d.). *How Do Weather Events Impact Roads*. Retrieved from Road Weather Management Program: ops.fhwa.dot.gov/weather/q1_roadimpact.htm

NIH. (2014, February). NIH News in Health. 4. National Institute of Health: Department of Health and Human Services. Retrieved from https://testdomain.nih.gov/sites/nihNIH/files/2014/February/NIHNiHFeb2014.pdf

Obasi, I. C., & Chizubem, B. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon*. doi:10.1016/j.heliyon.2023.e18812

OpenAI. (2024). ChatGPT (29 april 2024) [Large language model]. https://chatgpt.com/. (n.d.).

OpenAI. (2025). ChatGPT (18 February 2025) [Large language model]. https://chatgpt.com/. (n.d.).

Ramezani-Khansari, E., Nejad, F. M., & Moogeh, S. (2020). Comparing time to collision and time headway as safety criteria. *Pamukkale University Journal of Engineering Sciences*.

RoSPA. (2024, March). *Road safety factsheet: Driver fatigue and road collisions*. Royal Society for the Prevention of Accidents. Retrieved from www.rospa.com/getmedia/25999a67-6e67-4a8c-b32d-13b5c1090010/Driver-Fatigue-Factsheet-2022-updated.pdf

Sagaspe, P., Taillard, J., Bayon, V., Lagarde, E., Moore, N., Boussuge, J., . . . Philip, P. (2010). Sleepiness, near-misses and driving accidents among a representative population of French drivers. *Journal of Sleep Research, 19*(4). Retrieved from https://doi.org/10.1111/j.1365-2869.2009.00818.x

Saleem, S. (2022). *Risk assessment of road traffic accidents related to sleepiness during driving: a systematic review.* Lahore: Deparrtment of Community Medicine, King Edward Medical University. Retrieved from https://applications.emro.who.int/EMHJ/V28/09//1020-3397-2022-2809-695–700-eng.pdf

Sarkar, D. R., Rao, K. R., & Chatterjee, N. (2024). A review of surrogate safety measures on road safety at unsignalized intersections in developing countries. *Accident Analysis and Prevention*.

Shen, S., Benedetti, M., Zhao, S., Wei, L., & Zhu, M. (2020). Comparing distance and time as driving exposure measures to evaluate fatal crash risk ratios. *Accident Analysis and Prevention*. doi:10.1016/j.aap.2020.105576

Singh, D., & Das, P. (2021). A Review on Surrogate Safety Measures in Safety Evaluation and Analysis. *6th Conference of the Transportation Research Group of India (CTRG-2021).* doi:10.1007/978-981-19-4204-4_7

Siregar, M. L., Agah, H. R., & Hidayatullah, F. (2018). Near-Miss Accident Analysis for Traffic Safety Improvement at a 'Channelized' Junction With U-Turn. *International Journal of Safety and Security Engineering*.

Slikboer, R., Muir, S. D., Silva, S. S., & Meyer, D. (2020). *A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data.* doi:10.1186/s13643-020-01475-7

Smith, D. L., Najm, W. G., & Lam, A. H. (2003). Analysis of Braking and Steering Performance in Car-Following Scenarios. *Journal of Passenger Cars: Electronic and Electrical Systems*, 248-255.

Sufian, M. A., & Varadarajan, J. (2023). *Enhancing Prediction and Analysis of UK Road Traffic Accident Severity Using AI: Integration of Machine Learning, Econometric Techniques, and Time Series Forecasting in Public Health Research.* Leicester. Retrieved from https://arxiv.org/pdf/2309.13483

Sun, S., Guillen, M., Pérez-Marín, A. M., & Ni, L. (2024). Determining Driving Risk Factors from Near-Miss Events in Telematics Data Using Histogram-Based Gradient Boosting Regressors. *Journal of Theoretical and Applied Electronic Commerce Research*.

Terum, J. A., & Svartdal, F. (2019). Lessons learned from accident and near-accident experiences in traffic. *Safety Science*.

Vitols, K., & Voss, E. (2021). *Driver Fatigue in European Road Transport.* Brussels: European Transport Workers' Federation.

Vogel, K. (2003). A comparison of headway and time to collision as safety indicators. *Accident Analysis and Prevention*. doi:10.1016/s0001-4575(02)00022-2

Wahab, L., & Jiang, H. (2019). *A comparative study on machine learning based algorithms for prediction of motorcycle crash severity.* doi:10.1371/journal.pone.0214966

Wang, Y., Xu, C., Liu, P., Li, Z., & Chen, K. (2024). Assessing the predictability of surrogate safety measures as crash precursors based on vehicle trajectory data prior to crashes. *Accident Analysis and Prevention*.

WHO. (2023). *Global status report on road safety.* Geneva: World Health Organisation.

Xue, Q., Wang, K., Lu, J. J., & Liu, Y. (2019). Rapid Driving Style Recognition in Car-Following Using Machine Learning and Vehicle Trajectory Data. *Journal of Advanced Transportation*. Retrieved from https://doi.org/10.1155/2019/9085238

Zhang, Z., Yang, H., & Yang, X. (2023). A Transfer Learning-Based LSTM for Traffic Flow Prediction with Missing Data. *Journal of Transportation Engineering Part A Systems*. doi:10.1061/JTEPBS.TEENG-7638

8 Appendices

8.1 Appendix 1: Poisson Regression Model (Python Script)

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

file_path =
r"C:\Users\lule4\Downloads\thesis\trip_data\data_for_modeling.csv
"
df = pd.read_csv(file_path)

df = df.dropna(subset=[
    'dangerous_events', 'duration', 'distance', 'average_speed',
'day_night',
    'age', 'experience', 'gender', 'stc_weekly', 'income',
'road_environment'
])

formula = (
    "dangerous_events ~ duration + distance + average_speed +
day_night + "
    "age + experience + gender + stc_weekly + income +
road_environment"
)

poisson_model = smf.glm(
    formula=formula,
    data=df,
    family=sm.families.Poisson()
).fit()

print(poisson_model.summary())
print("\nLog-Likelihood:", poisson_model.llf)
print("AIC:", poisson_model.aic)
```

8.2 Appendix 2: Negative Binomial Regression Model (Python Script)

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

file_path =
r"C:\Users\lule4\Downloads\thesis\trip_data\data_for_modeling.csv
"
df = pd.read_csv(file_path)

df = df.dropna(subset=[
    'dangerous_events', 'duration', 'distance', 'average_speed',
'day_night',
    'age', 'experience', 'gender', 'stc_weekly', 'income',
'road_environment'
])

formula = (
    "dangerous_events ~ duration + distance + average_speed +
day_night + "
    "age + experience + gender + stc_weekly + income +
road_environment"
)

nb_model = smf.glm(
    formula=formula,
    data=df,
    family=sm.families.NegativeBinomial()
).fit()

print(nb_model.summary())
print("\nLog-Likelihood:", nb_model.llf)
print("AIC:", nb_model.aic)
```

8.3 Appendix 3: Random Forest Model (Python Script)

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score
import matplotlib.pyplot as plt
import numpy as np

file_path =
r"C:\Users\lule4\Downloads\thesis\trip_data\data_for_modeling.csv"
df = pd.read_csv(file_path)

df = df.dropna(subset=[
    'dangerous_events', 'duration', 'distance', 'average_speed',
'day_night',
    'age', 'experience', 'gender', 'stc_weekly', 'income',
'road_environment'
])

X = df[['duration', 'distance', 'average_speed', 'day_night', 'age',
        'experience', 'gender', 'stc_weekly', 'income',
'road_environment']]
y = df['dangerous_events']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Random Forest Regression Results")
print(f"MAE: {mae:.4f}")
print(f"MSE: {mse:.4f}")
print(f"R²: {r2:.4f}")

importances = rf.feature_importances_
features = X.columns
indices = np.argsort(importances)[::-1]

plt.figure(figsize=(10, 6))
plt.title("Feature Importance (Random Forest)")
plt.bar(range(X.shape[1]), importances[indices], align="center")
plt.xticks(range(X.shape[1]), [features[i] for i in indices],
rotation=45)
plt.tight_layout()
plt.show()
```

8.4 Appendix 4: AI Logbook

Date: 18 February 2025
AI Tool Used: ChatGPT (GPT-4o)

Prompt 1

User Prompt:

Find me publications on near miss accidents. Specifically in road crashes.

AI Response:
The AI produced a list of academic and technical sources focused on near-miss incidents within road safety:

- Near-Miss Road Traffic Accidents and Associated Factors Among Truckers (PMC – BMC Public Health)(Mekonen et al., 2024)
- Exploring Spatial Associations Between Near-Miss and Police-Reported Crashes (ScienceDirect)(Khan et al., 2023)
- Prevalence and Predictors of Near Misses and Road Traffic Crashes Among Long-Distance Bus Drivers (Springer)(Quarshie et al., 2024)
- Near-Miss Accident Analysis for Traffic Safety Improvement at a 'Channelized' Junction (ResearchGate)(Wardhana & Darlis, 2018)
- Predictable is Preventable: Tracking Pedestrian Near-Miss Incidents (NJ Crossing Guards)(Rutgers CAIT, 2023)
- Accidents and Near-Misses Analysis Using Video Drive-Recorders in a Fleet Test (NHTSA)(Yamamoto et al., 2013)
- Car-to-Pedestrian Contact Situations in Near-Miss Incidents and Real-World Accidents (NHTSA)(Zhang et al., 2011)
- Developing a Near-Miss Reporting System for Roadside Responders (AAA Foundation)(Michael et al., 2021)
- Drive Video Analysis for the Detection of Traffic Near-Miss Incidents (arXiv)(Kataoka et al., 2018)
- Using SlowFast Networks for Near-Miss Incident Analysis in Dashcam Videos (arXiv)(Khan et al., 2024)

Prompt 2

User Prompt:

Search for: publications on machine learning modelling in near-miss or accident prediction; preferably stick to academic sources.

AI Response:

- Using Near Misses, Artificial Intelligence, and Machine Learning to Predict Future Serious Casualties (Wiley)(Meaker et al., 2024)
- Application of Optimized Machine Learning Techniques for Prediction of Accident Outcomes (ScienceDirect)(Abellán et al., 2018)
- Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis (MDPI)(Khan et al., 2023)
- An Explainable Machine Learning Approach to Traffic Accident Fatality Prediction (arXiv)(Al Mamun et al., 2024)
- Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review (arXiv)(Seth et al., 2024)
- Anticipating Traffic Accidents with Adaptive Loss and Large-Scale Incident Database (arXiv)(Yao et al., 2018)

Prompt 3

User Prompt:

Filter out near miss publications to focus on road safety related.

AI Response:

The AI filtered entries from Prompt 1 to include only those with direct application to road traffic safety:

- Near-Miss Road Traffic Accidents and Associated Factors Among Truckers (PMC – BMC Public Health)(Mekonen et al., 2024)
- Exploring Spatial Associations Between Near-Miss and Police-Reported Crashes (ScienceDirect)(Khan et al., 2023)
- Prevalence and Predictors of Near Misses and Road Traffic Crashes Among Long-Distance Bus Drivers (Springer)(Quarshie et al., 2024)
- Near-Miss Accident Analysis for Traffic Safety Improvement at a 'Channelized' Junction (ResearchGate)(Wardhana & Darlis, 2018)
- Predictable is Preventable: Tracking Pedestrian Near-Miss Incidents (NJ Crossing Guards)(Rutgers CAIT, 2023)
- Accidents and Near-Misses Analysis Using Video Drive-Recorders in a Fleet Test (NHTSA)(Yamamoto et al., 2013)
- Car-to-Pedestrian Contact Situations in Near-Miss Incidents and Real-World Accidents (NHTSA)(Zhang et al., 2011)
- Drive Video Analysis for the Detection of Traffic Near-Miss Incidents (arXiv)(Kataoka et al., 2018)