Report on Road Accidents Dataset

Abstract:

Road accidents are a major public health and safety issue that affects millions of people every year. Understanding the factors and patterns that contribute to road accidents is essential for designing and implementing effective interventions and policies to prevent and reduce them. In this report, I present a data analysis project that uses a comprehensive dataset of road accidents reported over multiple years in the UK. The dataset contains various attributes related to accident status, vehicle and casualty preferences, demographics, and severity of casualties. It also includes factors such as pedestrian details, casualty types, road maintenance worker involvement, and the Index of Multiple Deprivation (IMD) deciles for casualties' home areas. I used various methods and techniques, such as descriptive statistics, visualizations, machine learning, and predictive modeling, to explore and analyze the dataset, and to identify the key features that influence accidents. Moreover, I interpreted the performance and results of our models and discussed the implications and limitations of my findings. My project aims to provide valuable insights for policymakers, and analysts who are interested in studying and improving road safety and transportation systems.

Leatreture Review:

The World Health Organization (WHO), as outlined in the 2018 Global Status Report on Road Safety (GSRRS), confirms that the global toll of deaths due to road traffic-related incidents reached a staggering 1.35 million in 2016 [1]. Simultaneously, the Pan American Health Organization (PAHO) reports that traffic accidents emerged as the second leading cause of death among young adults aged 15–29 in 2016. Of particular concern is the revelation that 47% of all fatalities resulting from traffic accidents involve vulnerable road users, including motorcyclists, cyclists, and pedestrians.

Despite substantial efforts involving the implementation of technological infrastructure and the enforcement of stringent traffic policies, the incidence of accidents remains unacceptably high, surpassing expectations. This predicament is exacerbated by the intricate challenge of accurately identifying the root causes of traffic accidents. Mechanical issues, adverse weather conditions, mental and physical fatigue, negligence, and road imperfections, among other factors, contribute to the complexity of the problem.

In the contemporary landscape, the deployment of prediction models as tools to mitigate mortality in traffic accidents has become a tangible reality. The insights derived from these models empower policymakers, transportation safety designers, and researchers to discern contributing factors and formulate recommendations, fostering significant strides in reducing

accident rates [3,4]. Noteworthy studies, sponsored by transportation-related institutions or companies [4,5,6,7,8,9], underscore the transformative potential of prediction models. However, these models grapple with challenges such as high data dimensionality resulting from information imbalances and the intricate management of large-scale datasets. Therefore, enhancing prediction models demands a strategic focus on exploring diverse data sources and devising solutions to address associated issues.
Given that these models rely on real-world data, authors often turn to government platforms and internet services for data collection. Integrating information from internet services into prediction models facilitates the establishment of real-time information channels, thereby augmenting accuracy. However, this approach encounters challenges, including incongruities in values and metrics across different sources due to variations in experimental design, acquisition protocols, equipment utilization, and data volume. Consequently, it is imperative to underscore the current state of learning-based traffic accident predictions and delineate the primary research challenges inherent in this domain.

Description of analysis methods:

Generally, the objective of this project is to identify patterns and influential factors in accidents in England during the year 2022. As customary in any data analysis undertaking, the project unfolds through several pivotal stages. Initially, the data is subjected to a meticulous cleaning process, addressing null values and outliers. Subsequently, a thorough analysis of the dataset is conducted, employing descriptive and distributional statistics to gain insights into the underlying patterns. The third stage involves the application of advanced engineering techniques to address specific challenges, culminating in the utilization of machine learning methodologies to intelligently process the data and derive effective solutions for the identified issues.

This project employs a dual-pronged approach to data analysis. The initial phase leverages statistical methods, providing a foundational understanding of the dataset. Subsequently, the project transitions to a more sophisticated stage where intelligent methods such as Random Forests, Decision Trees, and Deep Neural Networks are deployed. These advanced methodologies enable a nuanced exploration of intricate patterns and contribute to a more comprehensive understanding of the factors influencing accidents in the specified context.

At first, I will explain the dataset briefly. It is comprehensive data on road accidents in the UK which columns are:

1. **Status**: The status of the accident (e.g., reported, under investigation).
2. **Accident_Index**: A unique identifier for each reported accident.
3. **Accident_Year**: The year in which the accident occurred.
4. **Accident_Reference**: A reference number associated with the accident.
5. **Vehicle_Reference**: A reference number for the involved vehicle in the accident.

6. **Casualty_Reference**: A reference number for the casualty involved in the accident.
7. **Casualty_Class**: Indicates the class of the casualty (e.g., driver, passenger, pedestrian).
8. **Sex_of_Casualty**: The gender of the casualty (male or female).
9. **Age_of_Casualty**: The age of the casualty.
10. **Age_Band_of_Casualty**: Age group to which the casualty belongs (e.g., 0-5, 6-10, 11-15).
11. **Casualty_Severity**: The severity of the casualty's injuries (e.g., fatal, serious, slight).
12. **Pedestrian_Location**: The location of the pedestrian at the time of the accident.
13. **Pedestrian_Movement**: The movement of the pedestrian during the accident.
14. **Car_Passenger**: Indicates whether the casualty was a car passenger at the time of the accident (yes or no).
15. **Bus_or_Coach_Passenger**: Indicates whether the casualty was a bus or coach passenger (yes or no).
16. **Pedestrian_Road_Maintenance_Worker**: Indicates whether the casualty was a road maintenance worker (yes or no).
17. **Casualty_Type**: The type of casualty (e.g., driver/rider, passenger, pedestrian).
18. **Casualty_Home_Area_Type**: The type of area in which the casualty resides (e.g., urban, rural).
19. **Casualty_IMD_Decile**: The IMD decile of the area where the casualty resides (a measure of deprivation).
20. **LSOA_of_Casualty**: The Lower Layer Super Output Area (LSOA) associated with the casualty's location.

Analyze:

- **Step 1:** Clean Data

In the initial phase of data refinement, it became apparent that certain columns contained redundant information across all rows. Recognizing the impediment these columns posed to our analytical pursuits and investigative endeavors, a decisive step was taken to expunge them from the dataset. Notable among these excluded columns were 'Status' and 'Accident_Index.'

Proceeding to the subsequent phase of data purification, meticulous attention was devoted to addressing the presence of null values and rectifying duplicate rows. A methodical approach was employed to tackle outlier data and instances of missing information. This involved judiciously eliminating data wherever feasible and substituting it with the remaining values within the

respective columns. Noteworthy columns subjected to this treatment included ['casualty_home_area_type', 'casualty_imd_decile', 'lsoa_of_casualty'].

It is imperative to underscore the meticulous handling of outlier data, executed through the discerning removal of anomalous entries and the strategic replacement of missing values. The cleansing process culminated in a refined dataset, devoid of redundancies and poised for subsequent stages of analysis and interpretation. The commitment to data integrity and the meticulous resolution of anomalies lay the foundation for robust and reliable insights to be derived from the dataset.

- **Step 2:** statistical analysis

  In the subsequent phase, a meticulous statistical analysis of the refined dataset was conducted, yielding a nuanced understanding of the data. Leveraging statistical methodologies, redundant features were discerned and subsequently eliminated. Presented below are insightful descriptive charts encapsulating the outcomes of the statistical analyses.

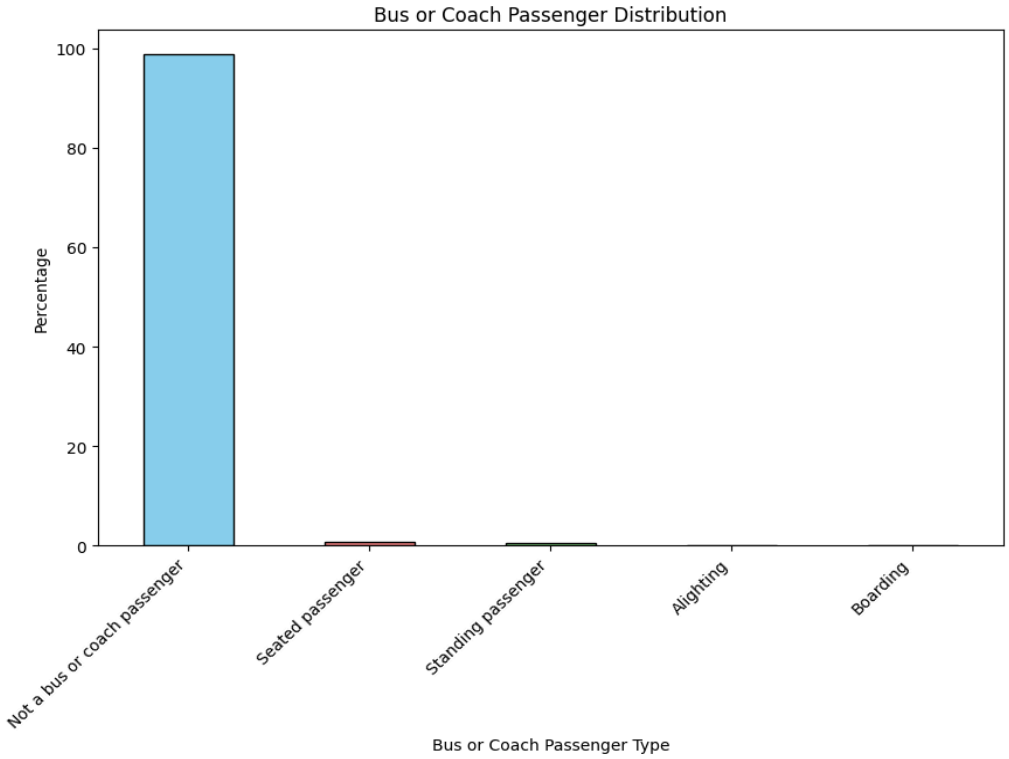

figure 1. Bus or Coach Passenger Distribution

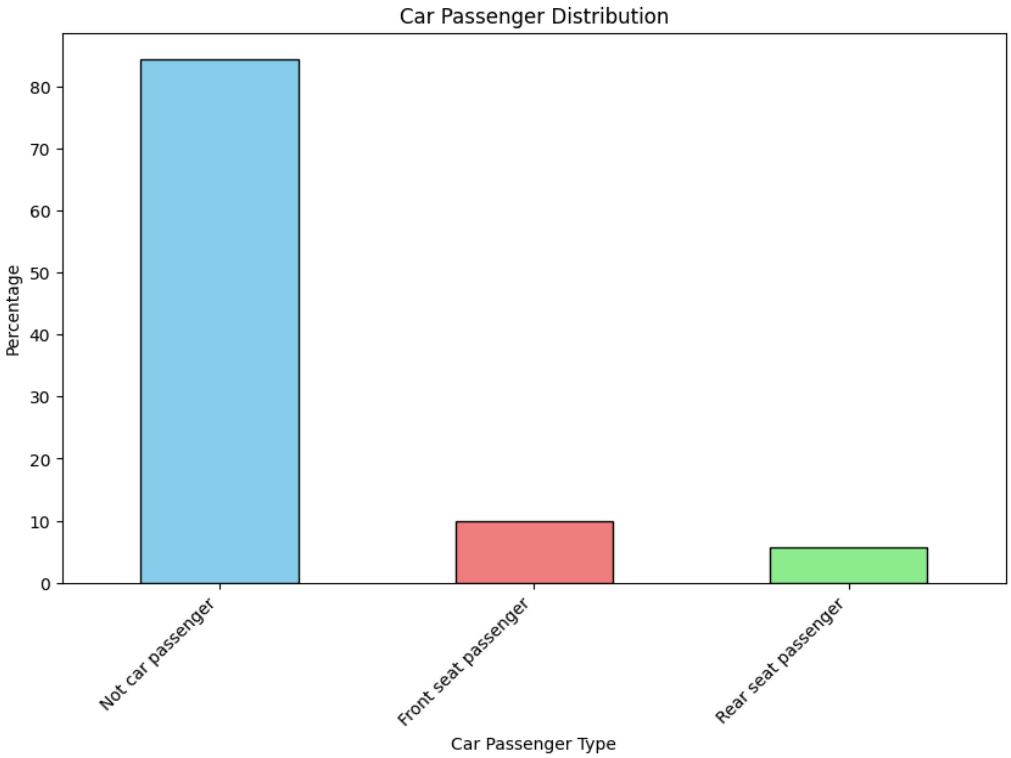

Figure 2. Car Passenger Distribution

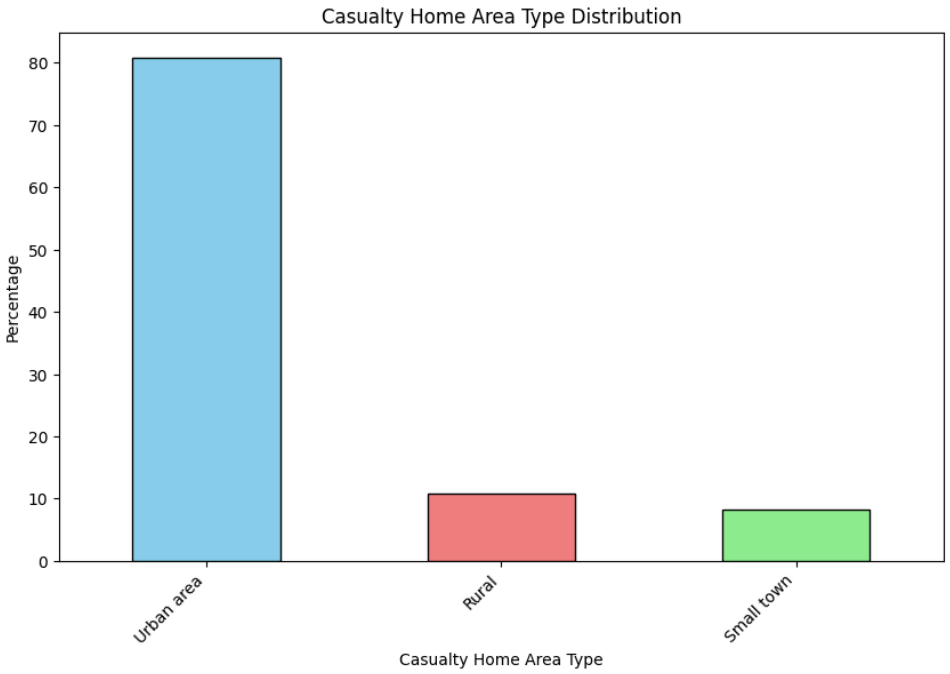

Figure 3. Casualty Home Area Type Distribution

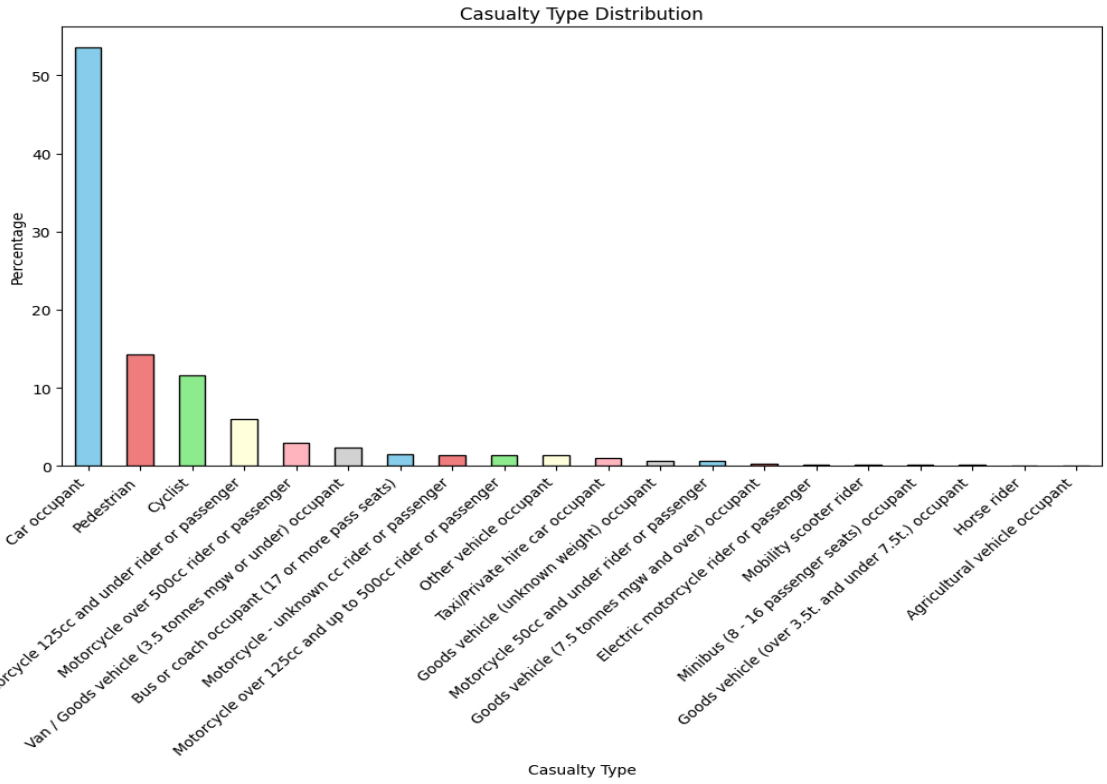

Figure 4. Casualty Type Distribution

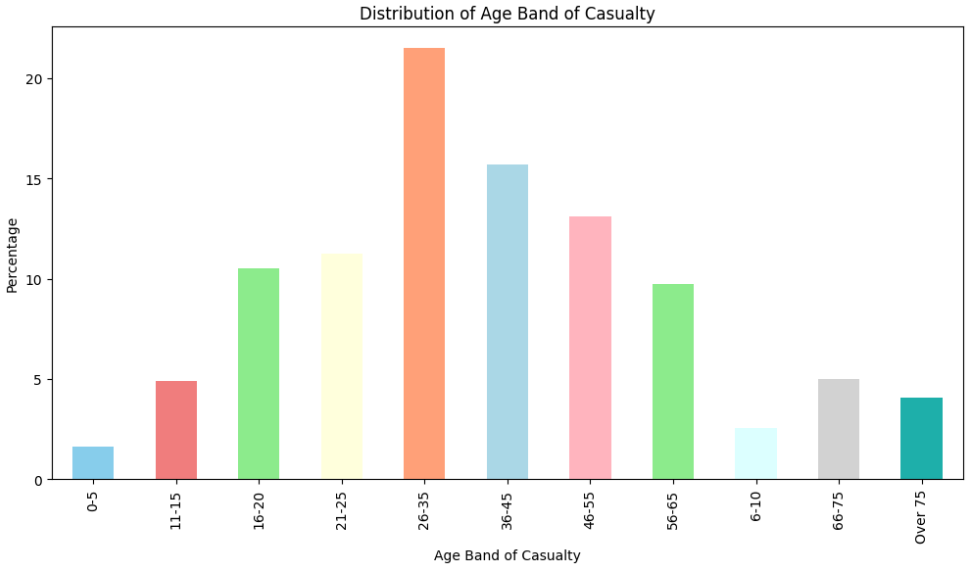

Figure 5. Age Band of Casualty

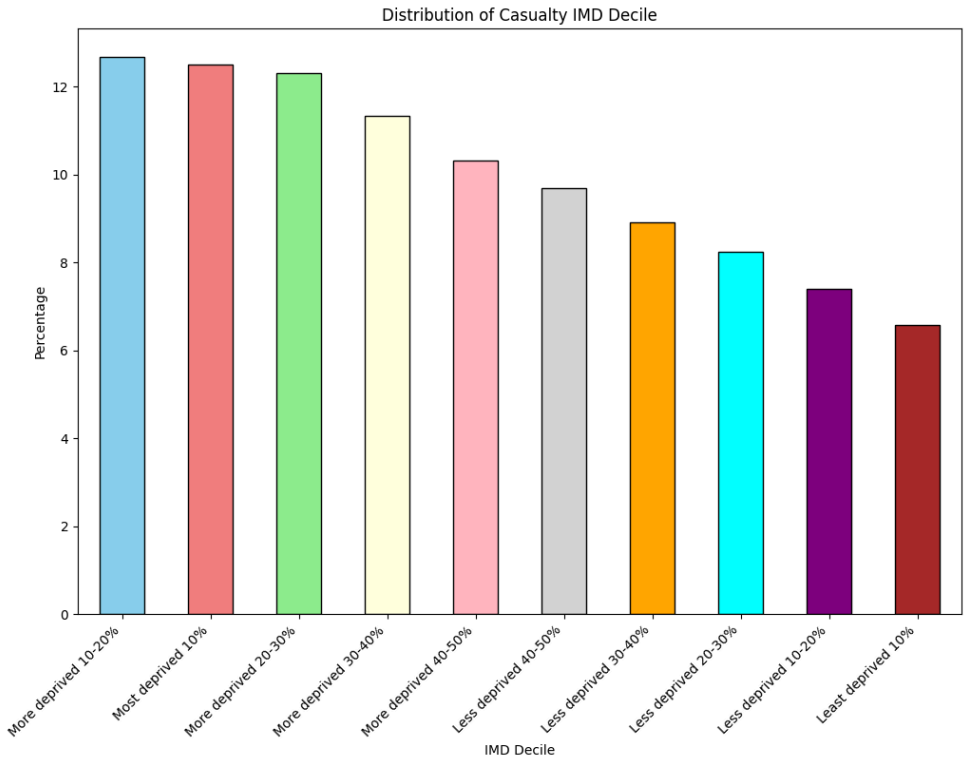

Figure 6. IMD Decile Casualty

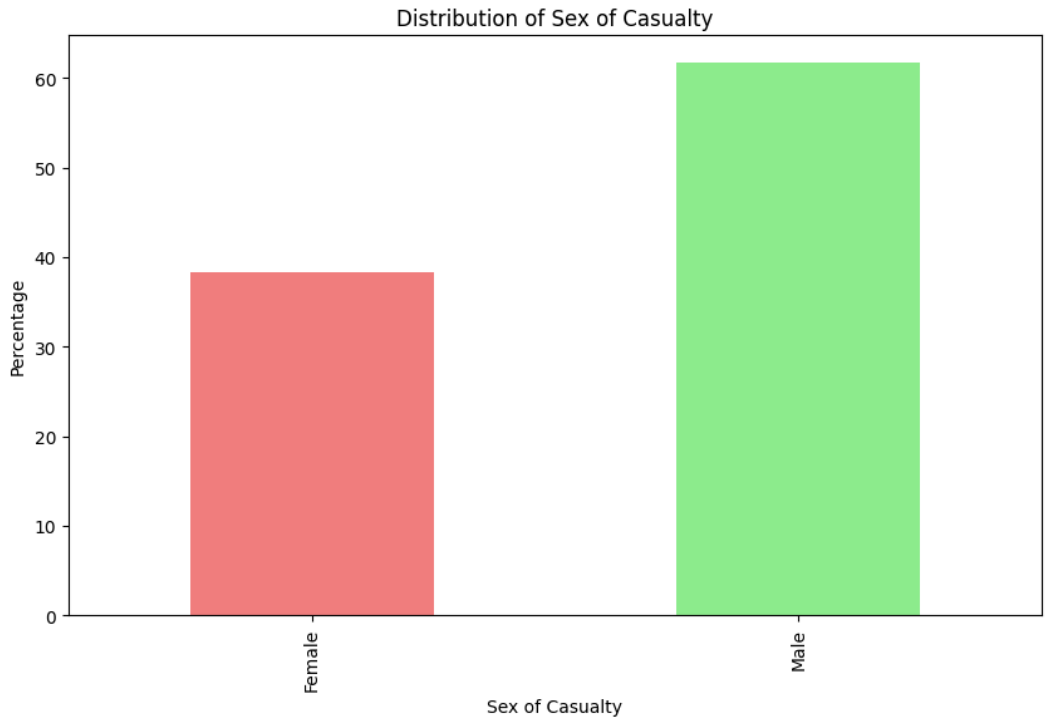

Figure 7. Sex of Casualty

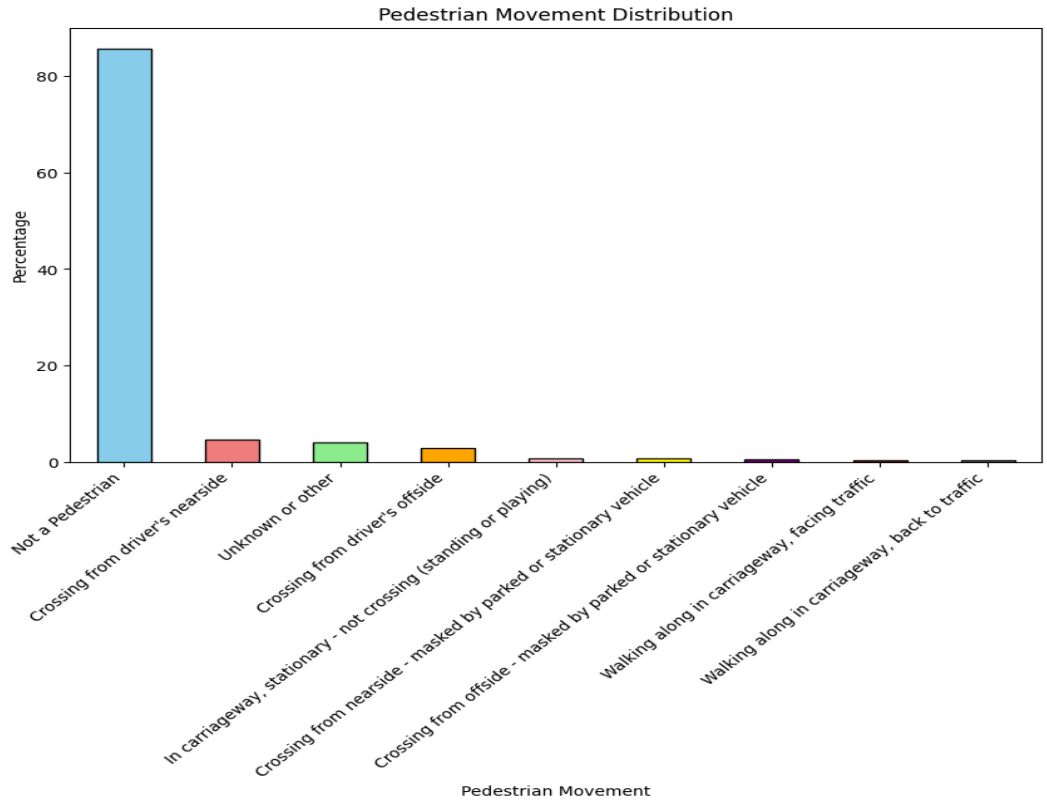

Figure 8. Pedestrian Movement Casualty

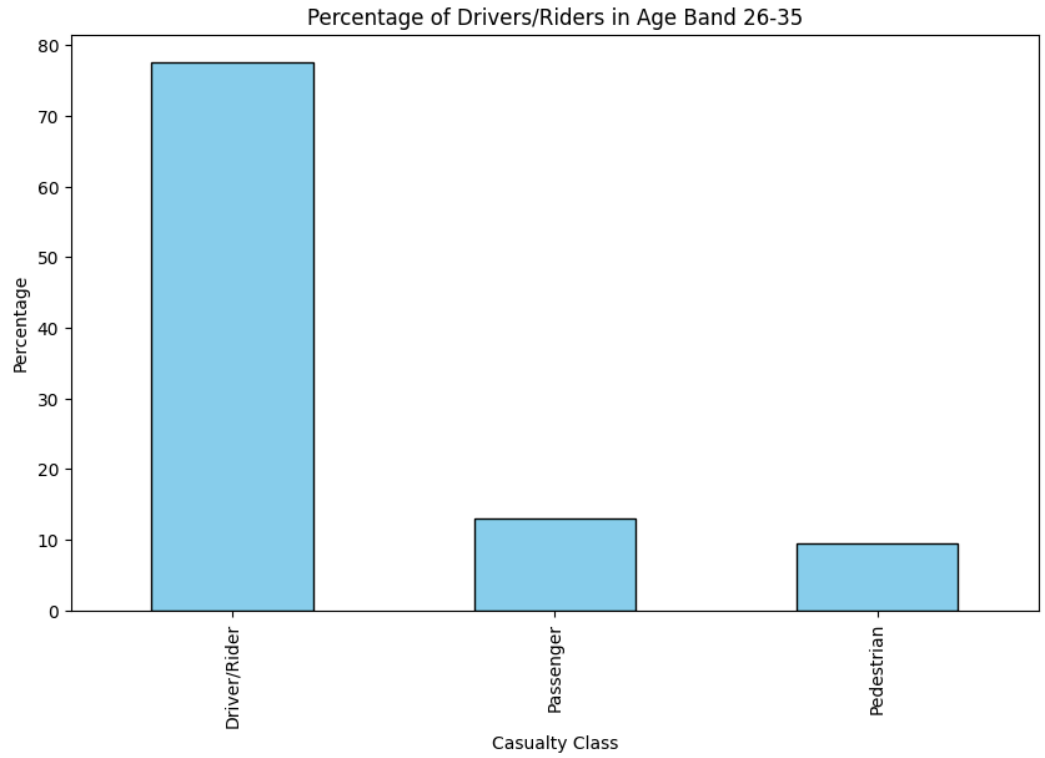

Figure 9. Percentage of Driver/Riders in Age Band 26-35 Casualty

Discussion of Statistical Studies:

The visual representations stemming from the statistical exploration distinctly reveal a predominant demographic among accident victims—predominantly male, with a notable majority being car drivers. Incidences involving passengers and bus drivers are comparatively infrequent. In contrast, pedestrians and construction workers emerge as less prevalent figures in accident scenarios. The age distribution of victims highlights a concentration within the 26 to 35-year age bracket, constituting a significant proportion of the total. Urban locales serve as the focal points for approximately 80% of reported accidents. Furthermore, a discernible pattern emerges, indicating a higher frequency of accidents transpiring in socioeconomically deprived areas.

Building upon the data cleaning procedures previously executed, the statistical analyses not only provided a comprehensive overview of accident dynamics but also facilitated the identification and elimination of superfluous features. This meticulous approach enhances the precision of subsequent analyses and fortifies the foundation for deriving meaningful insights from the dataset.

Step 3: Features Engineering and Training and Testing

Subsequently, following the meticulous data cleaning process, an in-depth data analysis was conducted utilizing sophisticated machine learning methodologies. Employing a diverse array of algorithms, such as random_forest, gradient_boosting, and decision_tree, the dataset underwent rigorous testing to discern the most impactful patterns. The results of these comprehensive tests unveiled a set of distinctive features that emerged as the paramount contributors to the overall analysis. The top features, indicative of their significance in influencing the outcomes, were identified as follows:

- Age Band of Casualty: This feature exhibited a substantial importance level, accounting for approximately 52.27% of the overall impact.
- Sex of Casualty: With a notable importance value of 34.75%, the gender of the casualty played a crucial role in determining patterns within the dataset.
- Casualty IMD Decile: This feature, assessing the Index of Multiple Deprivation (IMD) decile for casualties, contributed significantly, with an importance rating of 7.21%.
- Casualty Home Area Type: The nature of the area in which the casualty resides was a pivotal factor, representing 5.76% of the overall importance.

These findings underscore the nuanced relationships and correlations within the dataset, shedding light on the pivotal role these features play in influencing the outcomes of road accidents. It is noteworthy that this analysis not only enhances our understanding of the dataset but also lays the groundwork for informed decision-making and the implementation of targeted interventions to improve road safety.

Suggestions to reduce traffic accidents:

Based on the comprehensive analysis conducted, which involved the identification of influential factors contributing to accidents, several noteworthy suggestions emerge. These recommendations are poised to enhance road safety and mitigate the occurrence of accidents.

1. Firstly, considering the disproportionately higher frequency of accidents in areas characterized by greater deprivation, a strategic initiative is proposed to improve the overall infrastructure of roads within these regions. Concurrently, the implementation of speed control cameras and strategically placed traffic lights is strongly advocated to regulate vehicular movement and minimize accident risks.
2. Furthermore, the analytical findings underscore a concerning trend wherein passengers of private cars emerge as the demographic most susceptible to injuries. In contrast, buses and their passengers exhibit a lesser involvement in accidents. Consequently, an

imperative suggestion surfaces to bolster and refine public transportation systems, fostering their expansion and encouraging greater public utilization.

3. Urban areas, as delineated by the analysis, emerge as hotspots for accidents. This phenomenon is intricately tied to the concentrated nature of urban facilities, compelling individuals to traverse cities extensively for various aspects of their lives. To address this, a pivotal recommendation is to foster the equitable distribution of facilities across all regions, including smaller towns and rural areas, thereby alleviating the concentration of activities in urban centers.
4. Moreover, the analysis sheds light on the prominent role played by individuals aged 26-35 in accidents, marking them as a significant demographic. In response, a proposal is set forth to introduce more stringent driving regulations tailored to this age group, recognizing the need for heightened vigilance and compliance.
5. Considering the substantial contribution of pedestrians and cyclists to accidents, a proactive suggestion involves the augmentation of facilities dedicated to their safety. This encompasses the establishment of pedestrian lanes and dedicated cycling pathways, aimed at providing safer commuting options for these vulnerable road users.
6. Lastly, an observation regarding the predominant role of men in accidents prompts the recommendation for targeted educational initiatives. A call for enhanced educational programs specifically tailored for men is proposed, accompanied by the contemplation of more stringent regulatory measures to address this demographic's involvement in accidents comprehensively.

References:

1. World Health Organization. Global Status Report on Road Safety 2018; WHO: Geneva, Switzerland, 2018.

2. Pan American Health Organization. Status of Road Safety in the Region of the Americas 2018; PAHO: Washington, DC, USA, 2019.

3. Yasin Çodur, M.; Tortum, A. An artificial neural network model for highway accident prediction: A case study of Erzurum, Turkey. PROMET-Traffic Transp. 2015, 27, 217–225.

4. Hossain, M.; Muromachi, Y. A Bayesian network-based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accid. Anal. Prev. 2012, 45, 373–381.

5. Paikari, E.; Moshirpour, M.; Alhajj, R.; Far, B.H. Data integration and clustering for real-time crash prediction. In Proceedings of the 2014 IEEE 15th International Conference on Information

Reuse and Integration (IEEE IRI 2014), Redwood City, CA, USA, 13–15 August 2014; pp. 537–544.

6. Basso, F.; Basso, L.J.; Bravo, F.; Pezoa, R. Real-time crash prediction in an urban expressway using disaggregated data. Transp. Res. Part C Emerg. Technol. 2018, 86, 202–219.

7. Xu, C.; Wang, W.; Liu, P. A genetic programming model for real-time crash prediction on freeways. IEEE Trans. Intell. Transp. Syst. 2012, 14, 574–586.

8. Wang, L.; Abdel-Aty, M.; Shi, Q.; Park, J. Real-time crash prediction for expressway weaving segments. Transp. Res. Part C Emerg. Technol. 2015, 61, 1–10.

9. Lin, L.; Wang, Q.; Sadek, A.W. A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transp. Res. Part C Emerg. Technol. 2015, 55, 444–459