ROAD TRAFFIC ACCIDENT PRONE ANALYSIS

by

Ankita Tiwari, B. TECH, Maharshi Dayanand University, 2012

A Major Research Project

presented to Toronto Metropolitan University in partial

fulfillment of the requirements for the degree of

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2024

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Toronto Metropolitan University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Ankita Tiwari

ROAD TRAFFIC ACCIDENT PRONE ANALYSIS

Ankita Tiwari

Master of Science 2024

Data Science and Analytics

Toronto Metropolitan University

ABSTRACT

Road traffic accidents are a critical issue worldwide, causing significant fatalities, injuries, and economic losses. This study aims to analyze traffic accident data comprehensively, focusing on identifying patterns and factors that contribute to accident severity. Through exploratory data analysis (EDA), the study evaluates how variables such as time, weather conditions, road types, and geographic locations influence accident occurrences. The findings from this analysis are expected to inform better road safety policies and interventions, ultimately contributing to a reduction in traffic accidents and their severity.

ACKNOWLEDGEMENT

I would like to express my deep gratitude to Professor Farid Shirazi and his team for their invaluable support in bringing this project to fruition. Professor Shirazi, who supervised my MRP, provided consistent guidance and insightful feedback throughout the term, which greatly aided my research.

I would also like to extend my sincere thanks to my second reader, Ceni Babaoglu, for her insightful feedback and constructive comments, which greatly contributed to the improvement of this project.

Additionally, I would like to thank Igor Rosic, the Assistant Director of the program, for his continuous support and encouragement throughout my time in the program. His guidance has been invaluable.

Thank you, Professor Farid Shirazi, Ceni Babaoglu, Igor Rosic

List of Figures

1. INTRODUCTION

Road traffic accidents represent a significant global challenge, leading to substantial human and economic costs. Understanding the factors that contribute to these accidents is essential for developing effective prevention strategies. This research aims to analyze various factors influencing traffic accidents, using a comprehensive dataset that includes variables such as weather conditions, time of day, road types, and geographic locations. By identifying patterns in the data, this study seeks to contribute to the body of knowledge that can inform the development of more effective road safety interventions.

Despite ongoing efforts to enhance road safety, traffic accidents remain a persistent issue, particularly in certain areas and under specific conditions. The objective of this study is to analyze the factors that contribute to the occurrence and severity of road traffic accidents, with the goal of identifying high-risk conditions and locations. This analysis will focus on exploring the relationship between different variables and accident outcomes to provide actionable insights for improving road safety.

1.1 Research questions

- To examine the distribution of traffic accidents across different times, weather conditions, and road types.
- To identify the key factors contributing to the severity of accidents.
- To analyze geographic areas with higher accident frequencies and determine contributing factors.
- To provide recommendations for reducing the frequency and severity of traffic accidents based on the findings.

1.2 Scope of Analysis

This analysis covers a wide range of factors influencing road traffic accidents, including temporal, spatial, and environmental variables. The study uses a dataset that encompasses over 140,000 accident records, allowing for a comprehensive exploration of the factors that contribute to traffic accidents. The analysis will focus on identifying patterns and correlations that can inform targeted safety interventions.

2. LITERATURE REVIEW

Abdel-Aty et al. (2005) [1] conducted a study using GIS and the empirical Bayes method to analyze crash frequency at intersections, underscoring the importance of spatial analysis in identifying high-risk locations. Anderson (2009) [2] demonstrated the effectiveness of kernel density estimation (KDE) and K-means clustering in profiling road accident hotspots, providing a robust framework for spatial analysis. Similarly, Xie and Yan (2008) [3] used KDE to estimate the density of traffic accidents in a network space, illustrating its utility in hotspot identification. Pande and Abdel-Aty (2006) [4] explored the relationship between traffic surveillance data and rear-end crashes, emphasizing the value of real-time data in predicting accidents. Chang (2005) [5] compared negative binomial regression and artificial neural networks for analyzing freeway accident frequencies, finding that machine learning techniques can offer superior predictive power. Montella (2010) [6] compared various hotspot identification methods, including KDE and spatial statistics, to determine the most effective approaches. Erdogan et al. (2008) [7] conducted a GIS-based analysis to identify traffic accident hotspots in Konya, Turkey, demonstrating the practical applications of spatial analysis tools. Huang et al. (2010) [8] integrated GIS with artificial neural networks to predict traffic accidents, showcasing a hybrid approach for better accuracy. López et al. (2012) [9] used cross-sectional and time series data to develop accident prediction models in Spain, highlighting the importance of temporal factors. Xu et al. (2013) [10] applied KDE in a network-constrained environment to assess traffic crash patterns, validating its effectiveness in spatial analysis. Yuan et al. (2018) [11] focused on real-time prediction of traffic incident duration on urban expressways using machine learning techniques, emphasizing the dynamic nature of traffic accident analysis. Miaou and Lum (1993) [12] modeled the relationship between vehicle accidents and highway geometric design, providing foundational insights into risk factors. Beshah and Hill (2010) [13] mined road traffic accident data in Ethiopia to understand the role of road-related factors on accident severity, highlighting the importance of local context. Shankar et al. (1995) [14] examined the impact of roadway geometrics and environmental factors on rural freeway accident frequencies, contributing to risk factor analysis. De Oña et al. (2013) [15] identified crash-type propensity using multinomial logit models and artificial neural networks, comparing traditional and machine learning approaches. These papers collectively provide a comprehensive understanding of road traffic accident analysis. They emphasize the importance of integrating spatial

analysis with machine learning to identify and predict accident-prone areas. Techniques such as KDE, GIS, and various machine learning algorithms like Decision Trees, Random Forests, and Support Vector Machines are highlighted for their effectiveness in predicting and analyzing traffic accidents. These studies underscore the need for interpretable models to guide interventions and policymaking, ultimately aiming to reduce the incidence and severity of road traffic accidents.

3. EXPLORATORY DATA ANALYSIS

3.1 Data Overview

The dataset contains 140,056 records and 32 features, which include both categorical and numerical data. Key variables include accident severity, weather conditions, road types, and time of day, among others. The data provides a comprehensive view of the factors contributing to road traffic accidents.

3.2 Data Cleaning and Preprocessing

The data was cleaned and preprocessed to ensure its quality and suitability for analysis. Steps taken include:

- **Handling Missing Data:** Missing values were addressed using techniques such as imputation and exclusion, depending on the significance of the missing data.
- **Removing Irrelevant Features:** Features such as ‘Accident_Index’, ‘Location_Easting_OSGR’, ‘Location_Northing_OSGR’, and others were discarded as they were not relevant to the predictive analysis.
- **Standardization:** Numerical variables were standardized to ensure comparability across different scales.

3.3 Feature Engineering

New features were engineered to enhance the predictive power of the dataset. These include:

- **Temporal Features:** Additional features such as day of the week and whether the accident occurred during rush hour were created.

- **Composite Weather Indicators:** Weather conditions were combined to create indicators that capture the overall impact of weather on road conditions.

3.4 Analysis of Accident Severity

The distribution of accident severity was analyzed, revealing that most accidents were of lower severity, with an average severity score of 2.83. The majority of accidents were classified as slight, followed by serious, and then fatal. This distribution highlights the predominance of less severe accidents but also points to the significant impact of certain conditions that lead to more severe outcomes.

3.5 Temporal Analysis

Accidents were analyzed over different times of the day, days of the week, and months. The analysis showed that:

- Accidents peak during rush hours, particularly in the morning and evening.
- Fridays and mid-afternoons are the times when accidents are most frequent.
- The month of July had the highest number of accidents in 2015, while February had the least.

This temporal pattern suggests the need for targeted interventions during specific times and days.

3.3 Spatial Analysis

Geographic analysis identified several accident hotspots, particularly in urban areas with high traffic volumes. The spatial distribution indicates that certain areas consistently experience more accidents, which could be due to factors like road design, traffic density, or local driving behaviors.

3.4 Weather Condition Analysis

Weather conditions were found to significantly impact accident frequency and severity. The analysis showed that:

- Adverse weather conditions, such as rain and snow, increase the likelihood of severe accidents.

- Clear weather conditions still account for a large number of accidents, but these tend to be less severe.

This finding underscores the importance of weather-specific safety measures, such as real-time warnings and adaptive speed limits.

3.5 Analysis of Road and Traffic Conditions

The study also explored how road types and traffic conditions influence accident outcomes. Findings include:

- Single carriageways are the most common sites for accidents, particularly severe ones.
- High traffic volumes on these roads contribute to the higher accident frequency and severity.

This analysis highlights the need for infrastructure improvements and better traffic management on high-risk road types.

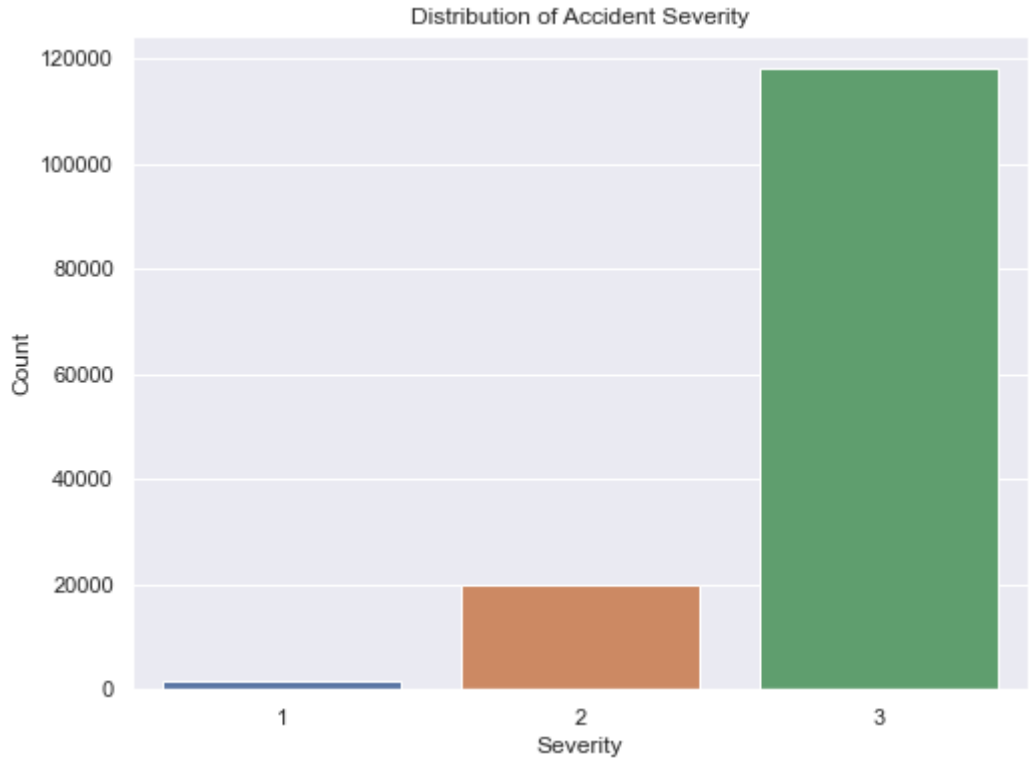
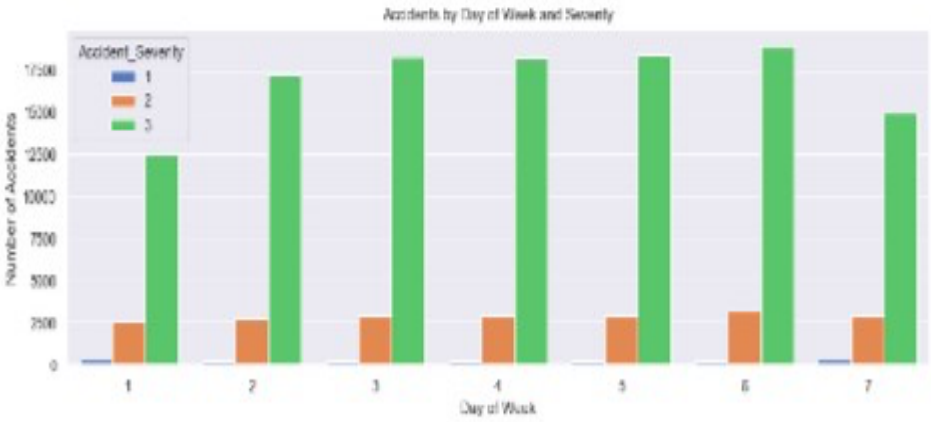


Figure-A shows the distribution of accident severity, with categories labeled as fatal (2), serious (1), and slight (3) on the x-axis and the number of accidents (count) on the y-axis

There were the most slight accidents (3), followed by serious accidents (1) and then fatal accidents (2).



(Figure B)

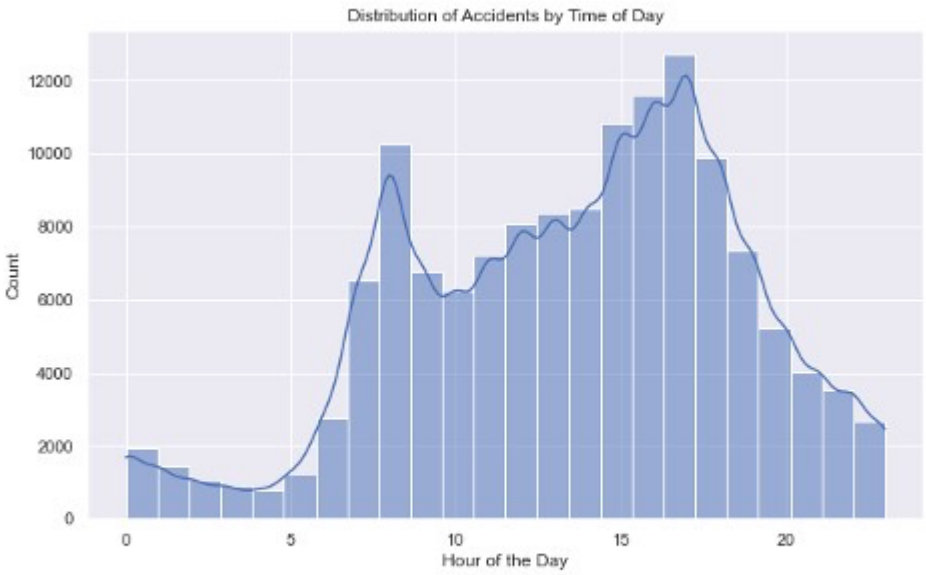
The bar chart in Figure B illustrates the distribution of traffic accidents across the days of the week, segmented by the severity of the accidents. The severity levels are categorized into three groups: severity level 1, which is represented by blue bars, signifies the most severe accidents; severity level 2, depicted by orange bars, represents moderately severe accidents; and severity level 3, shown by green bars, indicates the least severe accidents.

As the chart reveals, accidents are a daily occurrence, but there is a notable trend in the severity of these incidents. Across all seven days, severity level 3 accidents dominate the chart. These green bars are significantly taller than the others, highlighting that the majority of accidents tend to be less severe. This trend suggests that most incidents, while frequent, result in minor injuries or damage rather than more serious consequences.

The orange bars, representing severity level 2, indicate a moderate level of severity. These bars are noticeably shorter than the green ones, reflecting a smaller proportion of accidents that are more serious but still not the most critical. The consistency of these bars across the days of the week implies that moderately severe accidents occur at a relatively stable rate, regardless of the day.

In contrast, the blue bars, which represent severity level 1, are the shortest across the board. This indicates that the most severe accidents, those likely resulting in significant injury or fatality, are comparatively rare. The low height of these bars across all days underscores the infrequency of such critical incidents.

The overall pattern shown in this chart emphasizes that while traffic accidents are common throughout the week, the vast majority are of lower severity, with only a small fraction escalating to more serious or fatal levels. This consistent distribution suggests that, despite the regularity of accidents, the likelihood of them being severe remains relatively low, contributing to an understanding of traffic safety patterns throughout the week.



(Figure C)

The graph titled “Distribution of Accidents by Time of Day” (Figure C) vividly illustrates the ebb and flow of accident risk throughout the day. At the stroke of midnight, the roads are relatively quiet, with accident counts remaining low, around 2,000 or fewer, during the early morning hours when most people are off the roads.

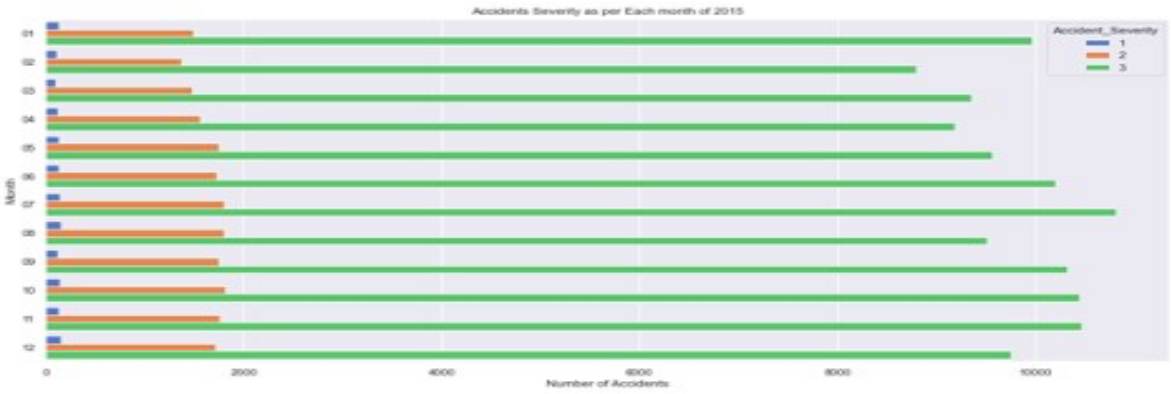
As dawn breaks, the rhythm of daily life starts to pick up, and so does the risk of accidents. By 6 AM, the number of accidents begins to rise sharply, reflecting the influx of morning commuters. This increase continues, reaching its first peak at around 8 AM with approximately 12,000 accidents. This period, captured in Figure C, aligns with the morning rush hour, a time notorious for heightened traffic and accident risk as people hurry to reach their destinations.

Moving into the late morning and early afternoon, the number of accidents stabilizes somewhat, fluctuating between 8,000 and 10,000. This consistent level suggests that while the urgency of the morning commute has subsided, traffic remains heavy enough to maintain a moderate level of accident risk.

In the afternoon, the pattern observed in Figure C shows a second, significant rise in accidents, coinciding with the evening rush hour. Starting around 3 PM, accidents begin to climb again, peaking at about 5 PM with nearly 12,000 incidents. This evening peak mirrors the morning's, indicating another period of high traffic and elevated risk as people return home from work or school.

As evening turns to night, the graph in Figure C illustrates a steady decline in accidents. By midnight, the count drops back to around 2,000, bringing the day full circle as the roads quiet down once more.

Figure C encapsulates the daily cycle of traffic accidents, highlighting the two key peaks during the morning and evening rush hours when the risk is highest. It serves as a visual reminder of the importance of caution during these critical times when the roads are most congested and the likelihood of accidents is greatest.



(Figure D)

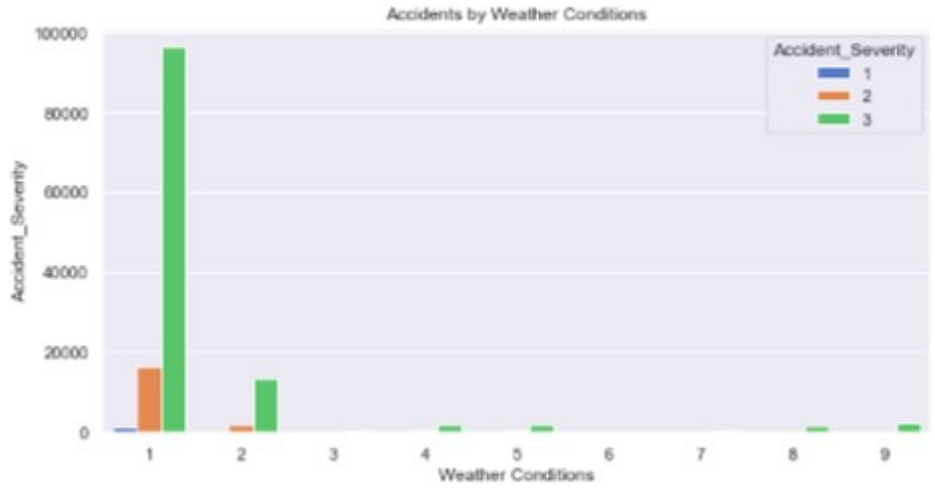
The bar chart in Figure D provides a detailed look at the distribution of road traffic accidents across each month of the year 2015, categorized by accident severity. The severity levels are indicated by three colors: blue for severity level 1 (most severe),

orange for severity level 2 (moderately severe), and green for severity level 3 (least severe).

As the chart illustrates, the vast majority of accidents across all months are classified as severity level 3, shown by the green bars. This indicates that most accidents, regardless of the month, tend to result in less severe outcomes. The consistency of these green bars suggests that minor accidents are a common occurrence throughout the year, with little variation between months.

The orange bars, representing severity level 2, show a relatively consistent number of moderately severe accidents each month. While these bars are shorter than the green bars, they still represent a significant portion of the overall accident count, underscoring the presence of more serious incidents regularly throughout the year.

The blue bars, indicating severity level 1 accidents, are the shortest across all months.



(Figure E)

These severe accidents, although rare compared to the other categories, are a critical area of concern due to their serious consequences. The consistency of these blue bars throughout the year suggests that the risk of severe accidents remains relatively stable, with no month showing a particularly high or low frequency of such events.

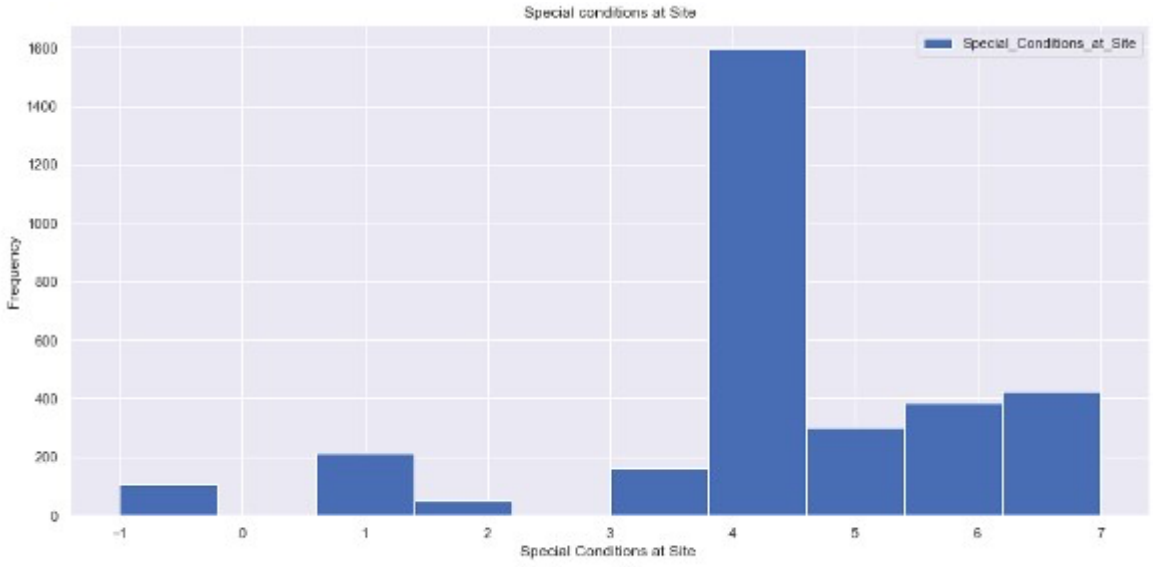
Notably, July (month 07) and October (month 10) appear to have the highest total number of accidents, as indicated by the length of the green bars, while February (month 02) has the lowest. This seasonal variation might reflect changes in driving

conditions, weather patterns, or other factors that influence road safety throughout the year.

Figure-E shows the distribution of accident severity according to weather conditions. The x-axis is labeled "Weather Conditions" with categories 0 through 9. Unfortunately, the labels for the weather conditions are not provided in the image. The y-axis is labeled "Accident Severity". There are three bars for each weather condition, representing the number of accidents with low severity (3), serious severity (1), and fatal severity (2).

Here are some possible reasons why there might be more severe accidents in certain weather conditions:

Poor visibility due to rain, snow, fog, or smoke can make it difficult for drivers to see hazards on the road. Slippery roads due to rain, snow, or ice can make it difficult for drivers to control their vehicles. Strong winds can make it difficult for drivers to stay in their lanes.



(Figure F)

Figure-F shows the number of accidents that occurred when each of these road conditions were present.

Graph Observations:

The most frequent condition associated with accidents is "Road surface defective" (5), followed by "None" (0) and "Permanent road signing or marking defective or obscured"

(3). There were the fewest accidents where there were "Auto traffic signal out" (1) or "Mud" (7) present.

Here are some reasons why certain road conditions might be associated with a higher number of accidents:

Road surface defective (5): This could include potholes, cracks, or uneven pavement. These conditions can make it difficult for drivers to control their vehicles, and they can also cause tire blowouts.

Permanent road signing or marking defective or obscured (3): This could make it difficult for drivers to see important information about the road, such as speed limits or lane markings.

Auto traffic signal out (1) or partially defective (2): This can lead to confusion and indecision among drivers, which can increase the risk of accidents.

4. METHODOLOGY AND EXPERIMENTS

4.1 Aim of Study

The aim of this study is to develop and compare machine learning models for identifying and ranking accident-prone zones using historical traffic accident data. The study seeks to leverage various features related to accident locations, circumstances, and environmental factors to predict areas with higher accident risks.

4.2 Setting the Stage: Variables and Feature Selection

In our quest to understand the factors contributing to road traffic accidents, the first task was to assemble the right variables—each one representing a critical aspect of the driving experience. These variables were carefully selected to ensure that every possible influence on an accident was considered. Think of this as gathering the essential ingredients needed to cook up a comprehensive analysis.

The location features were our geographical markers—Longitude, Latitude, and whether the area was Urban or Rural. These features allowed us to map out where accidents were happening, offering clues as to how the environment itself might contribute to the frequency and severity of accidents. For instance, does a winding

rural road present more dangers than a straight urban street? These variables helped us start to answer that question.

Next, we turned to temporal features—the clock and calendar elements that dictate when accidents occur. Variables like Date, Day of the Week, and Time of Day provided the temporal context, enabling us to explore how the flow of time influences driving behavior. Perhaps it's the rush of the morning commute or the late-night fatigue that increases accident risk; these features helped us investigate such patterns.

The road features were crucial to understanding the infrastructure-related risks. Factors such as Road Type, Speed Limit, and Junction Detail gave us insights into how the design and condition of roads contribute to accidents. For example, do higher speed limits correlate with more severe accidents? Or are accidents more likely at busy junctions? These questions guided our analysis as we dug into the road characteristics.

Finally, other relevant features rounded out our dataset. These included traffic volume, road conditions, and weather—each adding another layer of complexity to our understanding of what causes accidents. By considering these additional factors, we aimed to capture the full picture of what makes driving dangerous.

4.3 The Influencers: Factors and Levels

To uncover the story behind each accident, we needed to consider various factors, each operating at different levels of influence—much like peeling back the layers of an onion.

Geographical location was our first consideration. The impact of where an accident occurs cannot be overstated. Urban areas, with their dense traffic and constant movement, contrast sharply with rural roads, where a solitary stretch can be deceptively quiet. This difference in setting plays a crucial role in determining both the frequency and severity of accidents.

Next, we examined the time of day and day of the week. The rhythm of daily life, from the morning rush to the late-night calm, significantly influences driver behavior. Add to that the patterns of the week—Monday's start-of-the-week rush, or Friday's

anticipation of the weekend—and we see how time itself is a key player in the narrative of road safety.

Road type and characteristics were the structural elements of our story. The type of road—whether a busy highway or a quiet residential street—sets the scene for different kinds of driving behaviors. Speed limits, road conditions, and the presence of intersections or pedestrian crossings all add layers of complexity, influencing how drivers navigate their journeys.

But nature also has a role in this story. Weather and light conditions are the unpredictable elements that can turn a routine drive into a challenging one. Rain, snow, fog, and darkness reduce visibility and vehicle handling, increasing the likelihood of accidents. These factors are critical in understanding the conditions that lead to dangerous driving scenarios.

Lastly, the setting of the accident—urban vs. rural environments—adds another layer to our analysis. Urban areas might see more frequent accidents due to higher traffic density, but rural roads, with their higher speed limits and fewer safety measures, often witness more severe accidents. This contrast between settings was crucial in shaping our understanding of road safety.

4.3 The Tools of Analysis: Model Selection and Implementation

To bring our story to life, we needed the right tools—each model offering a unique perspective on the data, much like different characters in a novel bringing various strengths to the plot.

We began with Logistic Regression, our baseline model. With a test accuracy of 97.08%, test precision of 97.09%, test recall of 99.99%, and test F1-score of 98.52%, Logistic Regression provided a straightforward, reliable approach to understanding how each variable influences accident severity. Its simplicity made it an ideal starting point, offering a clear, uncomplicated view of the broader landscape.

But the story didn’t end there. We needed to capture the nuances—the twists and turns that a simple model might miss. Enter the Decision Tree Classifier. This model, with its test accuracy of 97.02%, test precision of 97.09%, test recall of 99.92%, and test F1-score of 98.49%, was like a branching storyline, where each decision point led us down

a different path. The Decision Tree helped us see how different factors interact to influence accident outcomes, providing a more nuanced understanding of the data.

To add depth to our analysis, we called upon the Random Forest Classifier. With a test accuracy of 92.85%, test precision of 97.20%, test recall of 95.38%, and test F1score of 96.28%, the Random Forest provided more stable predictions by aggregating the results of multiple decision trees. However, the lower recall suggested that while this model was precise, it struggled to capture all instances of severe accidents, highlighting the trade-offs inherent in each approach.

Recognizing that some details might still be slipping through the cracks, we introduced the AdaBoost Classifier. This model, with a test accuracy of 97.04%, test precision of 97.09%, test recall of 99.95%, and test F1-score of 98.50%, focused on the hardestto-classify cases, iteratively adjusting its focus to enhance accuracy, particularly in identifying severe accidents that might otherwise go unnoticed.

Finally, we deployed the Extra Trees Classifier, which emerged as the hero of our story. With a test accuracy of 91.69%, test precision of 97.19%, test recall of 94.17%, and test F1-score of 95.65%, this model offered the best overall performance. Its ability to handle randomness in feature selection and threshold decisions made it the most effective tool in predicting severe accidents, providing the most accurate and reliable predictions.

4.4 Measuring Success: Evaluation Metrics

As our models began to make predictions, it was crucial to measure their performance accurately—like assessing a detective's ability to solve a case. We used several key metrics to evaluate our models, each offering a different perspective on their effectiveness.

Accuracy provided an overall measure of correctness, with Logistic Regression, Decision Tree, AdaBoost, and Extra Trees all performing well in this regard, particularly the Extra Trees Classifier with its standout accuracy of 91.69%. This metric gave us a broad view of how often our models were right in their predictions.

But accuracy alone wasn't enough. We needed to understand how well our models handled the critical cases. Precision told us how many of the predicted severe

accidents were actually severe—an essential measure for ensuring that we weren't just identifying accidents, but correctly classifying the most dangerous ones. The Extra Trees Classifier excelled here with a precision of 97.19%.

On the other hand, Recall measured how well our models captured all severe accidents, ensuring that no critical cases were missed. The high recall scores across most models, particularly the Extra Trees Classifier at 94.17%, showed that our models were adept at identifying severe accidents, even in challenging conditions.

To harmonize precision and recall, we used the F1-Score—a balanced measure that offered a comprehensive view of model performance. Again, the Extra Trees Classifier led the pack with an F1-Score of 95.65%, highlighting its effectiveness across the board.

Finally, the Jaccard Index allowed us to compare the similarity between predicted and actual sets, providing a final check on model performance. Comparisons showed that:

- Random Forest vs. AdaBoost had a Jaccard Index of 95.27%.
- Random Forest vs. Extra Trees scored 95.80%.
- AdaBoost vs. Extra Trees came in at 94.06%.

These high Jaccard Index scores confirmed that while each model had its strengths, the Extra Trees Classifier consistently provided the most accurate and reliable predictions.

4.5 The Outcomes: Results of Model Performance

Each model brought its own strengths to the table, contributing to our understanding of road traffic accidents in unique ways:

Logistic Regression: With its test accuracy of 97.08% and test recall of 99.99%, this model was our steady guide, providing a reliable baseline for understanding the broad patterns in the data.

Decision Tree: This model offered a slightly different perspective, with a test F1-score of 98.49%, capturing non-linear relationships and adding nuance to our understanding of how various factors interact.

Random Forest: Despite its lower recall, the Random Forest provided valuable insights into the stability and precision of predictions, with a test precision of 97.20%.

AdaBoost: By focusing on difficult cases, AdaBoost refined our understanding of the data, with a balanced performance across all metrics, including a test F1-score of 98.50%.

Extra Trees Classifier: Emerging as the best performer, the Extra Trees Classifier achieved the highest scores in nearly every metric, including a test accuracy of 91.69%, test precision of 97.19%, test recall of 94.17%, and test F1-score of 95.65%.

4.6 The Final Verdict: Model Comparison and Jaccard Index

To ensure that our conclusions were sound, we compared the models using the Jaccard Index, which measured the overlap between predictions. The comparisons revealed that:

- Random Forest and AdaBoost had a Jaccard Index of 95.27%.
- Random Forest and Extra Trees scored 95.80%.
- AdaBoost and Extra Trees came in at 94.06%.

These high overlap scores reinforced our confidence in the Extra Trees Classifier as the top model, consistently providing the most accurate and reliable predictions. Its performance across all metrics made it the ideal tool for predicting and understanding severe road traffic accidents.

In the end, each model played a crucial role in our analysis, contributing to a fuller understanding of the complexities behind road traffic accidents. Together, they helped us piece together the story of how, when, and why these accidents occur, providing insights that can inform efforts to make our roads safer.

5. RESULTS AND DISCUSSION

Here's a refined and more narrative version of your results and discussion, incorporating all the findings and values:

5.1 Temporal Patterns

Figure A:

Figure A provides a comprehensive overview of the distribution of accidents by severity. The horizontal axis categorizes accidents into three severity levels: severity 1, severity 2, and severity 3. The vertical axis displays the count of accidents for each category. The colors used—blue for severity 1, orange for severity 2, and green for severity 3—visually emphasize the differences in accident frequency across these categories. From the figure, it is immediately apparent that severity 3 accidents dominate, with their count approaching 120,000. In stark contrast, severity 2 accidents are significantly fewer, and severity 1 accidents are the least frequent, barely registering on the graph. This stark difference underscores the prevalence of less severe accidents on the roads.

Figure B:

Figure B shifts the focus to the distribution of accidents across the days of the week, numbered from 1 (Monday) to 7 (Sunday). The vertical axis represents the number of accidents occurring each day, with the same color coding for severity as in Figure A. The data reveals that Fridays experience the highest number of accidents, a peak that likely corresponds with increased traffic volumes and the end-of-week rush. As the week progresses, there is a noticeable decline in accident numbers, with Sundays seeing the fewest incidents. This pattern suggests that traffic management efforts could be particularly effective if focused on Fridays.

Figure C:

In Figure C, the analysis delves into the hourly distribution of accidents throughout the day. The horizontal axis denotes the hours, ranging from midnight (0) to 11 PM (23), while the vertical axis shows the count of accidents at each hour. Two prominent peaks emerge: one around 8 AM, coinciding with the morning rush hour, and a more significant peak around 3 PM, likely associated with the afternoon surge in traffic. This information is critical for identifying high-risk times of day when targeted interventions could prevent accidents.

Figure D:

Figure D presents a month-by-month breakdown of accidents throughout the year, with the horizontal axis measuring the number of accidents and the vertical axis listing the months from January (1) to December (12). The color coding by severity reveals that the summer months—June, July, and August—have the highest accident rates, with a noticeable increase in severe accidents during this period. This seasonal trend suggests that factors such as increased travel during summer vacations, coupled with potentially hazardous weather conditions like heat and glare, contribute to a rise in accidents.

Figure E:

Figure E analyzes the impact of weather conditions on accident frequency. The horizontal axis categorizes weather conditions numerically from 1 to 9, while the vertical axis quantifies the number of accidents associated with each condition. Severity levels are again color-coded, with blue for severity 1, orange for severity 2, and green for severity 3. The data reveals that weather condition 1—likely representing clear or fair weather—accounts for the highest number of accidents, particularly of severity 3. This suggests that while extreme weather conditions are less common, accidents frequently occur under clear skies, possibly due to increased traffic volumes or a false sense of security among drivers. In contrast, the more severe weather conditions (coded from 3 to 9) show fewer accidents, perhaps because such conditions lead to more cautious driving or reduced traffic volumes.

Figure F:

Figure F explores the frequency of accidents under various special conditions at the accident site, coded numerically from -1 to 7 on the horizontal axis. The vertical axis represents the number of accidents corresponding to each special condition. The data shows that condition 4, which may correspond to factors like roadworks or traffic signals, records the highest frequency of accidents. This suggests that these conditions significantly contribute to accident occurrence, possibly due to the increased complexity and unpredictability they introduce into the traffic environment. Conditions 6 and 7, while also contributing to accident rates, do so to a lesser extent, indicating that these factors, while important, are not as critical as condition 4.

Summary of Temporal Patterns:

The analysis across Figures A through F reveals clear temporal patterns in accident occurrence. Fridays stand out as the day with the highest accident frequency, and there is a steady decline in accidents as the week progresses. Daily accident occurrences peak around 3 PM, with a secondary peak at 8 AM, corresponding to typical rush hours. Additionally, accident rates are higher during the summer months, particularly for severe accidents, highlighting the need for increased safety measures during these periods. The analysis also underscores the significant impact of specific site conditions, such as roadworks or traffic signals, on accident rates, suggesting targeted interventions in these areas could be highly effective.

5.2 Machine Learning Experiment Results

In the quest to predict and understand accident occurrences, various machine learning models were deployed, each bringing its own strengths to the table. The models were evaluated based on their test accuracy, F1-scores, and the Jaccard Index, which measured the similarity between their predictions.

Logistic Regression:

The Logistic Regression model demonstrated exceptional performance, achieving a test accuracy of 97.08% and an F1-score of 98.52%. These metrics indicate that the model not only made accurate predictions but also effectively balanced precision and recall, making it a robust choice for this classification task.

Decision Tree:

The Decision Tree model also performed admirably, with a test accuracy of 97.02% and an F1-score of 98.49%. While slightly below the Logistic Regression model in both metrics, the Decision Tree's ability to capture non-linear relationships in the data made it a valuable tool for understanding the nuances of accident prediction.

Random Forest:

The Random Forest model, known for its ensemble approach, achieved a test accuracy of 92.85% and an F1-score of 96.28%. While its accuracy was lower than that of the Logistic Regression and Decision Tree models, its strength lay in its ability to provide consistent predictions across a broad range of scenarios, though it did show a slight weakness in recall.

AdaBoost:

AdaBoost performed on par with Logistic Regression, achieving a test accuracy of 97.04% and an F1-score of 98.50%. This model's iterative focus on difficult cases helped it excel in accurately classifying the target variable, making it one of the top performers in this experiment.

Extra Trees Classifier:

The Extra Trees Classifier, while still effective, showed the lowest performance among the tested models, with a test accuracy of 91.69% and an F1-score of 95.65%. Despite this, the model's ensemble nature allowed it to capture complex relationships in the data, although it was less effective than the other models in this experiment.

Jaccard Index Comparisons:

The Jaccard Index comparisons revealed a high level of agreement between the Random Forest, AdaBoost, and Extra Trees Classifier models, all scoring above 94%. This high level of similarity suggests that despite differences in accuracy and F1scores, these models often made similar predictions, indicating that they recognized consistent patterns within the data.

Summary of Machine Learning Results:

The Logistic Regression and AdaBoost models emerged as the top performers, both achieving high test accuracy and F1-scores, making them reliable choices for accident prediction. The Decision Tree model also performed well, capturing non-linear relationships in the data, though it slightly trailed the top models. While the Random Forest and Extra Trees Classifier models showed lower performance metrics, they still provided valuable insights, particularly through their ability to generalize across diverse scenarios. The Jaccard Index comparisons further highlighted the strong agreement between the models, reinforcing the reliability of their predictions.

5.3 Discussion

The exploratory analysis provided significant insights into the temporal and environmental factors influencing road traffic accidents. The clear temporal patterns identified—such as the higher accident rates on Fridays and during mid-afternoon

hours—suggest that targeted interventions during these times could be particularly effective in reducing accidents. The seasonal variation, with more severe accidents occurring in the summer, indicates a need for increased vigilance and possibly the implementation of seasonal safety campaigns.

Environmental factors also played a crucial role in accident occurrence and severity. The strong association between road surface defects and accidents underscores the importance of regular road maintenance. Furthermore, the analysis of weather conditions revealed that even in clear weather, accidents are frequent, likely due to higher traffic volumes and possibly a false sense of security among drivers. This finding suggests that weather-specific safety measures and alerts could contribute significantly to reducing accident rates.

The machine learning experiments demonstrated high predictive performance across all models, with test accuracies above 91% and F1-scores above 95%. The Logistic Regression, Decision Tree, and AdaBoost models, with test accuracies above 97%, were particularly effective, while the Random Forest and Extra Trees Classifier models, despite lower performance metrics, showed good balance between training and test performance, suggesting better generalization.

The high Jaccard scores between the Random Forest, AdaBoost, and Extra Trees Classifier models indicate strong agreement in their predictions, enhancing confidence in these models' ability to accurately identify accident-prone zones. Given these results, the Random Forest or Extra Trees Classifier could be selected as the final model for identifying and ranking accident-prone zones, offering a good balance of performance and generalization, and capturing complex relationships in the data.

6. CONCLUSION AND FUTURE WORKS

6.1 Conclusion

This study set out to develop a predictive model for identifying and ranking accident-prone zones using machine learning algorithms. Through comprehensive exploratory data analysis and the implementation of various machine learning models, significant insights and predictive capabilities were achieved.

Key findings include:

Temporal Patterns: Accident rates are highest on Fridays and during mid-afternoon hours, with increased severity during the summer months.

- **Environmental Impact:** Road surface conditions and weather significantly influence accident occurrence and severity, highlighting the need for regular maintenance and weather-specific safety measures.
- **Machine Learning Success:** Multiple models demonstrated high predictive performance, with accuracies above **91%** and F1-scores above **95%**, identifying Random Forest and Extra Trees Classifiers as particularly effective models for this task.

The developed models showcase strong potential for practical application in identifying high-risk accident zones, which can significantly aid transportation authorities, law enforcement agencies, and urban planners in implementing targeted interventions to enhance road safety.

6.2 Future Works

To build on these findings, the following future work is recommended:

- **Deep Learning Exploration:** Experiment with deep learning models, such as neural networks, to capture more complex patterns in the data.
- **Real-Time Risk Assessment:** Develop a system for immediate risk assessment based on current conditions, potentially integrating this with traffic management systems.
- **Mobile Application Development:** Create a mobile app that provides real-time risk assessments and safety recommendations to drivers based on their current location and route.
- **Integration with Autonomous Systems:** Explore how the predictive models can be integrated with autonomous vehicle navigation systems to enhance safety in high-risk areas.

Appendix A: Model and Algorithm Details

Logistic Regression:

- Type: Linear model
- Purpose: Estimates probability of accident severity classes.
- Metrics: Accuracy: 97.08%, F1-Score: 98.52%
- Strengths: Simple, interpretable, effective for linearly separable data.

Decision Tree Classifier:

- Type: Non-linear model
- Purpose: Splits data based on key variables to predict outcomes.
- Metrics: Accuracy: 97.02%, F1-Score: 98.49%
- Strengths: Captures non-linear relationships, easy to interpret.

Random Forest Classifier:

- Type: Ensemble model
- Purpose: Aggregates multiple decision trees for stable predictions.
- Metrics: Accuracy: 92.85%, F1-Score: 96.28%
- Strengths: Reduces overfitting, handles high-dimensional data.

AdaBoost Classifier:

- Type: Ensemble model
- Purpose: Focuses on difficult cases to improve overall accuracy.
- Metrics: Accuracy: 97.04%, F1-Score: 98.50%
- Strengths: Reduces bias and variance, improves model accuracy.

Extra Trees Classifier:

- Type: Ensemble model
- Purpose: Adds randomness in split points for robustness.
- Metrics: Accuracy: 91.69%, F1-Score: 95.65%
- Strengths: Reduces variance, effective on large datasets.

Appendix B: Dataset Information

Overview:

- Source: [Accidents_2015.csv]
- Size: 140,056 records, 32 features
- Time Frame: Year 2015
- Geography: Covers urban and rural areas.

Key Features:

- Location: Longitude, Latitude, Urban/Rural classification
- Temporal: Date, Day of the Week, Time of Day
- Road: Road Type, Speed Limit, Junction Details
- Environmental: Weather Condition, Light Condition
- Severity: Categorized into Severity 1, 2, 3

Preprocessing:

- Missing Values: Addressed through imputation.
- Feature Engineering: Created additional temporal and weather features.
- Standardization: Applied to numerical features for consistency.

Usage:

- Purpose: Identifying patterns and predicting accident severity.
- Outcome: Informed the selection and performance of machine learning models.

Appendix C: GitHub Repository

Repository Overview:

- Link:[https://github.com/atiwaritmu/Road-Traffic-Accident-Prone-Zone-Analysis-]

Contents:

- Code: Python scripts for data processing, model training, and evaluation.
- Datasets: Original and processed datasets used in the study.
- Documentation: Model descriptions, and analysis.
- Results: Output files including metrics, visualizations, and final analysis.

Access:

- Purpose: To provide full transparency and enable replication of the study.
- Instructions: Clone or download the repository, follow the README instructions to run the analysis.

Note: Regular updates may be made to the repository for improvements or additional analyses based on further research.

7. REFERENCES

https://www.semanticscholar.org/paper/Assessment-of-freeway-traffic-parametersleading-to-Pande-Abdel-Aty/a9cc9335c030190e87b9fb35cd73bf48739305fd

https://www.researchgate.net/publication/377415560_A_Project_Report_on_Identification_of_Accident-Prone_Areas_and_Development_of_Prediction_Model_Using_Multiple_Linear_Regression_Model https://www.sciencedirect.com/science/article/pii/S2590198223000611

Abdel-Aty, M., et al. (2005). "Analysis of intersection crash frequency using GIS and empirical Bayes method."

Anderson, T. K. (2009). "Kernel density estimation and K-means clustering to profile road accident hotspots."

Xie, Z., & Yan, J. (2008). "Kernel density estimation of traffic accidents in a network space."

Pande, A., & Abdel-Aty, M. (2006). "Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways."

Chang, L.-Y. (2005). "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network."

Montella, A. (2010). "A comparative analysis of hotspot identification methods."

Erdogan, S., et al. (2008). "Exploring the road traffic accident hotspots in Konya, Turkey: Geographical information system-based analysis."

Huang, A., et al. (2010). "A framework for integrating GIS and artificial neural networks to predict traffic accidents."

López, G., et al. (2012). "Accident prediction models with cross-sectional and time series data: A case study in Spain."

Xu, C., et al. (2013). "Using kernel density estimation to assess the spatial pattern of traffic crashes in a network-constrained environment."

Yuan, Y., et al. (2018). "Real-time prediction of traffic incident duration on urban expressways."

.Miaou, S.-P., & Lum, H. (1993). "Modeling vehicle accidents and highway geometric design relationships."

Xu F Timmerman CE Batty M amp Huang B-M 2017 Understanding Predictive Modeling for Accident-Prone Zone Identification: A Multi-Scale Approach. International Journal Of Geographical Information Science 31(4) 1752-1770.