Analyzing and Modeling Colorado Crash Data with Comparisons of ID3, Random Forest, Naïve Bayes, and Logistic Regression

Sai Srikar Emani, Saurav Avinash Changde, Kurt Cushman, Harshitha Sai Addepalli
University of Colorado Denver – Department of Computer Science
Denver, United States of America
saisrikar.emani@ucdenver.edu, sauravavinash.changde@ucdenver.edu, kurt.cushman@ucdenver.edu, harshithasai.addepalli@ucdenver.edu

Abstract – Extensive research has been conducted both analyzing and modeling traffic data in an attempt to predict future accidents. This study looks to build on previous works, targeting a mountainous region, by examining crash data from Colorado, in the United States. Colorado provides a unique test subject, because of a mix of sizable metropolitan areas, rural areas, and varying weather patterns including rainstorms and snow. Previous work has shown varying success across several test methodologies, including Tobit regression and Bayesian spatial modeling for location. This study compiles the crash data, analyzes it, then leverages four machine learning models: ID3, Random Forest (RF), Naïve Bayes, and Logistic Regression (LR) to examine any correlations and identify trends. After examination, it was found that drivers age 16-25 are, overall, in less incidents than individuals 25-35. Correlations between age, injury severity, and number of vehicles was identified.

Keywords – crash, traffic patterns, vehicle collisions, inclement weather

I. INTRODUCTION

Over the last several years, the National Highway Traffic Safety Administration (NHTSA) has identified several disturbing trends. Despite the overall number of police-reported vehicle traffic crashes dropping from approximately 6.76 million in 2019 to approximately 5.25 million in 2020, the number of fatalities has increased [1]. In 2019, there was approximately one fatality in every 186 accidents, whereas in 2020 there was one fatality in every 134. With approximately 2.28 million people injured in 2021, vehicular safety is a growing concern.

Previous work modeling the road accident safety has leveraged logistic regression [2], decision tree, random forest, Tobit regression [3], and Bayesian spatial modeling [4], in an attempt to build a correlation and potentially predict future crashes. Each has focused on a small subset of data, some including heavy vehicles, bicycles, small sampling of data, or specific areas. In general, the findings did display some correlations between population, gender, and established infrastructure.

In this study, all traffic incidents in the state of Colorado from the years of 2007 to 2022 were compiled and analyzed. For each accident, up to eighty-four points of data were collected, including time of day, road conditions, location, severity, any impairment of driver, type of vehicle involved, and injuries. With the intent of finding any correlation to build a predictive model, all points were kept, resulting in approximately 1.6 million records to review.

In the past, the relative accuracy of model has been brought into question. Some found random forest to be the most accurate, in relation to tests against logistic regression. Linear regression, Tobit regression, and Bayesian modeling were also all claimed to be strongest. To build on those prior findings, this study tests four major modeling styles: ID3, Random Forest (RF), Naïve Bayes, and Logistic Regression (LR) in an attempt to find any correlations across those elements.

The intention of this analysis is to find correlations between the accidents datapoints to leverage for targeted awareness campaigns and to potentially explore additional training for drivers with the recognition of the problem area. This study could also be leveraged by insurance companies when calculating the premium charged to specific parties and confirms whether key indicators, such as age or weather, do lead to an increased likelihood of incident.

II. RELATED WORK

The groundwork of any research lies in a complete understanding of existing work in the field. In this section, we will do an extensive review and exploration of appropriate literature and research related to Machine Learning and its applications in monitoring vehicle traffic crashes and predicting future accidents.

Over the past few decades, the field of machine learning (ML) has seen significant changes. The emergence of big data science has also fueled the advancement of machine learning methodologies and techniques, which account for more nuances, improve the accuracy of our predictions, and paint a clear picture of the world in which we live. This overview serves the main goal of giving a broad overview of the research and approaches that have been conducted in the past and have helped advance our understanding of how to use machine learning (ML) to monitor vehicle traffic and predict future collisions.

Furthermore, this review aims to critically assess and summarize the conclusions, research methods, and theoretical foundations of the foundational works rather than just repeating previous studies. Through situating our findings within the larger framework of previous research and publications, we hope to identify patterns in motor vehicle collisions, clarify the typical issues that arise from collisions, and provide guidance on preventing accidents in the future. We have pointed out the main highlights of the research, if you want to check the whole research, please check the References page of the document.

Sun Z et al. (2020) conducted a detailed crash analysis of mountainous freeways, revealing challenges arising from the presence of numerous bridges, tunnels, and interchanges. Their findings underscored the higher incidence of rear-end crashes due to difficulties in managing inter-vehicle interactions in these terrains (1).

- The research was conducted in China but has mountains freeways and tunnels are like the freeways and tunnels in Colorado, USA.
- The reason we have decided to research this article is the abundance of bridges, tunnels, and interchanges in China's central and western areas. Frequent cross-section switching, rapid changes in longitudinal driving situations, and alternating light and dark sections are characteristics of these new mountainous roadways. Because of this, driving is a dynamic and challenging task overall.
- While the proportion of accidents with fixtures is the opposite of that of rear-end crashes, the proportion of rear-end crashes with the highest frequency on the entire road is around 10% greater than that on regular road sections. This suggests that drivers find it more challenging to manage vehicle-to-vehicle interactions on steep highways with high ratios of bridges to tunnels than on other road segments. Improper car-following conduct is the primary cause of the high crash incidence among crash causes.
- In this study, one crucial factor that must be considered is the severity of traffic accidents. A total of 1739 crashes—26 involving fatalities, 68 involving injuries, and 1,645 involving simply property damage—were examined in this study. 285 people were killed in these crashes. Russo F et al. (2014) emphasized the nuanced impact of driver gender and age on crash frequencies, highlighting specific scenarios where males or females exhibited higher risks. Their study underscored the need for targeted awareness campaigns tailored to diverse driver demographics and situations.

Russo F et al. (2014) worked out road safety from the perspective of driver gender and age as related to the injury crash frequency and road scenario. The study investigates the correlation between different demographics (gender and age) of drivers and the likelihood of being involved in injury crashes. It aims to understand how these factors interplay with different road scenarios, likely exploring scenarios such as urban, rural, highway, etc (2).

- This study has 8 years of data analyzed.

- These studies show that the combination of driver age and gender further refines the contribution of those variables to collisions.
- The authors' suggestion is that awareness campaigns should be targeted at a specific audience and not be generic.
- The study says that the process of validation was successful. It has been discovered that, depending on the driving situation and the gender of the other drivers engaged in the incident, men and women are involved in crashes at different frequencies. The risk of head-on/side collisions increased for only female drivers on dry road surfaces during daylight hours on tangent segments; the risk of single-vehicle crashes increased for only male drivers on wet road surfaces during daylight hours on circular curves; and the risk of head-on/side collisions increased for both male and female drivers on dry road surfaces during daylight hours on circular curves.

Patil S et al. (2012) employed nested logit models to analyze crash severities, particularly focusing on accounting for underreported crashes. Their research highlighted the significance of considering unreported incidents to avoid biased estimations of the effects of explanatory variables (3).

- The nominal logit (NL) and multinomial logit (MNL) model specifications are used in the crash injury severity data analysis for various reasons. These specifications are more flexible in terms of capturing the effects of independent variables when compared to the ordered response probit (OP) models.
- The authors came to the conclusion that estimations of the effects of the explanatory variables and their elasticities may be markedly skewed if the underreporting is ignored. For instance, they discovered that when estimating an OP model using conventional estimation methodologies, the influence of environmental conditions, the use of safety restraints, and the driver's gender may be over- or under-estimated.
- The study leveraged an additional variable for unreported crashes and grouped the bottom severity levels together due to ambiguity of classification

Anderson J & Hernandez S (2017) delved into heavy-vehicle crash rate analysis using Tobit regression, identifying contrasting effects of lower speeds on crash rates for heavy versus smaller vehicles (4).

- The study isolates the crash data to analyze large vehicles.
- The authors used Tobit Regression to analyze vehicle crash-rate.
- The study found that lower speed decreases crash rates on heavy vehicles, but the crash rate increases on smaller vehicles.

Qu Y et al. (2022) leveraged association rule mining for crash analysis, unveiling intrinsic relationships between contributing factors and accidents, revealing correlations based on road conditions, driver characteristics, and environmental factors (5).

- Association rule mining (ARM) has been proposed for crash analysis. Since ARM much outperforms conventional modeling techniques and can expose the underlying links between contributing elements and accidents without making assumptions, it is frequently employed in the field of traffic safety.
- With a total of 1080 samples, the summary results provide information on the event severity of baseline, crash, and near-crash occurrences. It is a much more compact dataset.
- The study compares crash and near crash events to identify which features to be identified and used for the analysis. It measures five parameters such as Antecedent, Support, Confidence, Lift and Conviction to predict the likeliness of a crash.
  - The key findings are summarized as follows:
    - (1) <u>Based on Roads:</u> (a) Undivided roadways are more likely associated with crash events, especially IV—low-risk tire strike events. In contrast, divided

roadways are more likely associated with near-crash events. It is assumed that a median strip or barrier could prevent crashes.

(b) Roads with less than 2 lnes are highly correlated with crash events, especially II—police- reportable events. Roads with 2-7 lanes are highly correlated with near- crash events or lower-severity crash events. Wider roadways are recommended to reduce the frequency and severity of crash events.

(c) Crash events mainly occur on level roads, whereas near-crash events mainly occur on straight roads. However, this factor is only related to C/NC events in combination with other factors.

- (2) <u>Based on Driver:</u>

(a) Female drivers have a low correlation with low-severity crash events, whereas male drivers have a high correlation with severe crash and near-crash events.

(b) Young drivers have a higher likelihood of being involved in crash events, whereas middle-aged and older drivers show a stronger association with near-crash events. However, the driver's age is not highly correlated with the severity of crash events.

(c) Crash events occur more likely when the drivers' estimated average annual mileage during the past five years is less than 10,000 miles. Near-crash events are more likely to occur when the drivers' average annual mileage during the past five years is greater than 15,000 miles. It is assumed that drivers with more driving experience have a safer driving style.

(d) Performing secondary tasks is highly correlated with crash events (especially the II—police-reportable crash events and III—minor crash events) and near-crash events

(e) Improper behavior is linked to crash events, whereas driver impairments are not.

Both factors are not strongly correlated with the severity of crash events

(f) The number of traffic violations or crash records is not strongly correlated to the frequency of C/NC events. However, drivers with one crash record during the past five years are more likely to be associated with I—most severe events.

(g) Minor visual-spatial disabilities are not strongly correlated with crash events but are strongly correlated with I—most severe events. It is assumed that minor visual- spatial disabilities do not affect driving significantly. However, during a crash event, visual- spatial disabled drivers may have problems handling the situation; thus, the crash event is typically more severe

- (3) <u>Based on Environment:</u>

(a)Crash events occur more likely in free flow traffic, and near-crash events are more likely in stable or unstable/forced flow. The results suggest that a higher traffic density keeps drivers alert, preventing crashes

(b)Crash events are more likely in sections with no traffic control or controlled intersections. However, these factors do not affect the severity of crash events

(c)Residential or business/industrial areas have a higher correlation with C/NC events than other areas. More traffic safety precautions should be considered in these areas

Chen MM & Chen MC (2020) compared logistic regression, decision tree, and random forest models for road accident severity prediction, concluding that random forest exhibited superior accuracy in their analysis (6).

- The authors say that Logistic Regression is the most popular.
- According to the study the nonparametric categorization and regression tree (CART) approach has garnered more attention in recent years for its application to transportation-related issues.
- The study found that Random had better accuracy based on input factor variables.

Huang H et al. (2010) proposed a Bayesian spatial model to account for county-level variations in crash risk, considering factors like daily vehicle miles traveled and population, challenging inconsistent findings in prior studies due to diverse geographical influences (7).

- This study proposes a Bayesian spatial model to account for county-level variations of crash risk in Florida by explicitly controlling for exposure variables of daily vehicle miles traveled and population.
- A series of studies by the authors extensively investigated the effect of various infrastructure changes on traffic-related fatalities and crashes at different levels of dis-aggregate spatial units, specifically, 50 states in the United States,102 counties in Illinois, and 8,414 census wards in England. In general, they found that some improvements in road infrastructure have led to increased crashes and fatalities.

Further insights were gained from predictive risk analysis frameworks applied by researchers in China (Predicting Future Driving Risk of Crash-Involved Drivers) utilizing machine learning classifiers to categorize high and low-risk drivers based on extensive longitudinal data (8).

- Conducted in China but have worked on driving risk analysis that can be used to categorize high-crash-risk drivers with seven years of data (2011-2017).
- They have used common classifiers like Random Forrest to assess the accuracy of high-risk and low-risk drivers.

Bhowmik T et al. (2021) presented an econometric approach for modeling crash frequency analysis by crash type and severity, offering a comprehensive methodological framework to analyze various count variables related to crashes (9).

- The authors used 28000 crash data samples from the Central Florida Region.
- The study is trying to find
  - a) common unobserved factors simultaneously affecting crash counts of different crash types.
  - b) common unobserved factors simultaneously affecting crash severity proportions of different crash types.
  - c) common unobserved factors that simultaneously impact crash counts and severity proportions by different crash types.

Additionally, national crash statistics from authoritative sources provide foundational information essential for understanding and contextualizing the broader landscape of road safety issues (10).

This collective body of research contributes to an enhanced understanding of the multifaceted elements influencing road safety, crash severity, and risk assessment across varied contexts, laying a foundation for further investigation and policy interventions in this critical domain.

The literature pertinent to vehicle traffic crashes contains a diverse array of studies, methodologies, and approaches. To facilitate a structured review and better comprehension, we have organized the related work into several distinct methodologies used by various authors to conduct their own research.

- The authors developed the road scenario-based discretization method to separate and distinguish different types of vehicle traffic crashes. Three distinct mountainous freeways, including tunnels, bridges, interchanges, and freeway service roads, have been taken into consideration by the writers. The width of entrance/exit locations, traffic patterns, and sight of oncoming cars are varied in each of the circumstances. This provides us with a distinct viewpoint on how to interpret the crash rate on each piece of road and compare it to other crash categories.

- Develop Safety Performance Functions: The goal of this study was to identify patterns in the frequency of car accidents on two-lane roads. The authors recommended remedies to lower the injury crash and offer structural measures on collision-prone sections of roads after researching human factors and diverse road scenarios.
- Analysis of accident severities using nested logit model: In their study, crash injury severity can be examined using this method. Numerous research has been conducted on different collision types and crash probability. However, research on the likelihood of a crash's severity level occurring is scarcer.

Every reference we looked at for this research had a great approach to the issue and offered improved features and accuracy from several global data sources. Numerous writers have examined a sizable dataset of different car accidents from a particular area and offered insightful analysis on the reasons behind the majority of the mishaps.

To provide more accurate predictions for car accidents that happen in the US, our approach uses crash data spanning ten years (2013 to 2022). Using data from Colorado, we are projecting and analyzing car crashes. It is possible that reference research papers lack a substantial dataset that spans nearly ten years for analysis; hence, subtleties in data classification and visualization may be overlooked. Due to the low traffic density at the time of covid pandemic, numerous state and local governments have used the opportunity to construct or improve their road infrastructure. Additionally, our statistics will consider the most recent patterns in vehicle crashes brought on by newly constructed or upgraded roadways in the mentioned states.

We want to use the same classifiers that are employed in most reference research publications in our analysis of the most recent crash data. Their degrees of accuracy and insights will serve as a roadmap for us as we conduct our study in order to deliver more insightful findings and feedback.

In order to address the issue of vehicle crash-related deaths and significant injuries caused to the drivers and passengers of the cars, this study can be utilized to assess all of the recent crashes that have occurred in the state of Colorado. This study can also be used to enhance certain road segments that are more likely to be involved in accidents because of their design, the state of the road, or the drivers' lack of vision. We can create better action plans to reduce collision events in all the accident-prone locations as we collect and assemble more data on this subject. We are confident that our research will serve as a basis for future investigations into the causes of crashes worldwide.

III. EXPERIMENT

This is a pivotal section that meticulously details the methodological approach undertaken in this comprehensive study. This segment is crucial as it bridges the gap between theoretical framework and practical application, offering readers a transparent view into the research process and techniques employed to analyze traffic crash data.

The focus shifts from the overarching goals and theoretical underpinnings of the study to the concrete steps taken in data handling, analysis, and model development. It encompasses a series of well-defined modules, each tailored to address specific aspects of the data lifecycle, from acquisition and preprocessing to analysis, modeling, and visualization.

Each module within this experiment is crafted with meticulous attention to detail, employing best practices in data science and machine learning, and is supported by robust coding practices and advanced analytical tools.

Data Acquisition Module

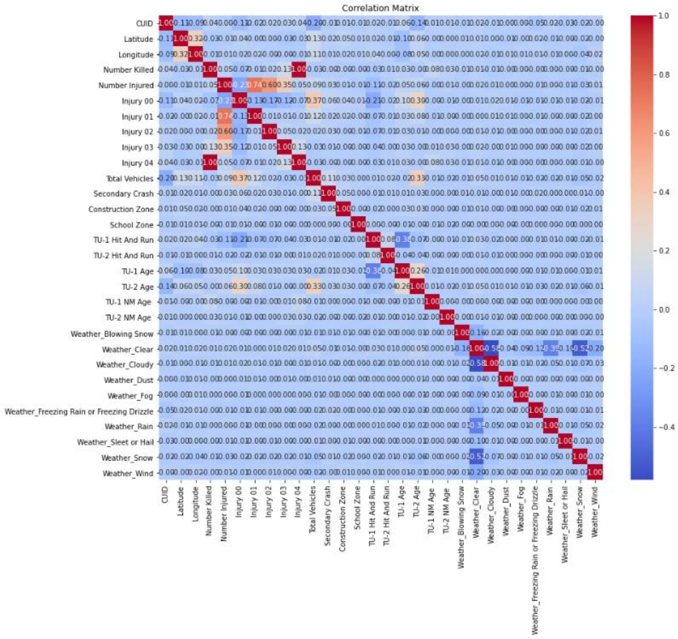
The foundation of the study is laid in the Data Acquisition Module, where a comprehensive setup process is undertaken to ensure robust data handling and analysis capabilities. Key libraries such as Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn are installed, equipping the research with a versatile tooklit for data manipulation, statistical computation, machine learning algorithms, and visualization. The establishment of global variables, including crucial data paths, creates a uniform environment, ensuring consistency and accuracy in data handling across the entire project. Furthermore, the meticulous definition of data columns standardizes the nomenclature and formats used, facilitating clearer understanding and smoother data processing throughout subsequent phases of the project.

Data Preprocessing Module

In the Data Preprocessing Module, a rigorous process of refining the raw data is undertaken. This stage is crucial for ensuring the quality and reliability of the data before it undergoes detailed analysis. The module starts with the loading of data, focusing on an extensive range of crash details such as severity, weather conditions, time, and location. Initial cleaning tasks include the elimination of duplicate records and addressing missing values, foundational steps critical for maintaining data integrity. Numerical data is approached with a strategy of median value imputation for missing entries, while categorical data is treated with mode imputation, ensuring that the dataset remains robust and representative. This meticulous preprocessing sets the stage for accurate and insightful analysis.

Data Analysis Module

The Data Analysis Module represents the core of the study, where deep dives into the dataset unveil significant patterns and correlations. This module employs statistical analysis to extract descriptive insights from key variables like 'SEVERITY' of the accident, 'WEATHER' conditions, 'LIGHTING' conditions, and road 'CONDITION'. Understanding the skewness and distribution of these variables provides a foundational understanding of the data characteristics. The module then progresses to pattern detection and correlation analysis. Tools like pair plots and heatmaps are employed to visually represent relationships between variables, offering intuitive insights into potential correlations. The heatmap, in particular, serves as a powerful tool to visually discern the strength and direction of relationships between multiple variables, enabling us to pinpoint significant factors that could influence traffic accident trends.



Model Development Module

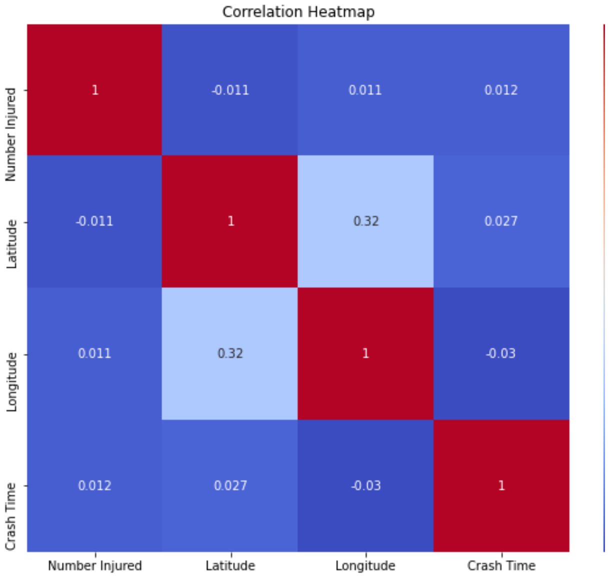
The Model Development Module is where theoretical insights translate into practical applications. In this module, the processed dataset undergoes a critical phase of transformation, where it is sampled and split into features and target variables, setting the stage for predictive modeling. This phase sees the applicationof four major machine learning models: ID3(Decision Tree), Random Forest, Naïve Bayes, and Logistic Regression. Each model serves a specific purpose, offering diverse perspectives on the data. The transformation of categorical variables into dummy variables and the meticulous handling of date-time columns ensure that the models receive well-structured and relevant input. The training and evaluation of these models are conducted with rigor, using accuracy scores and classification reports as key performance indicators. This module is pivotal in determining the potential of machine learning in predicting and understanding traffic accident trends.

Visualization Module

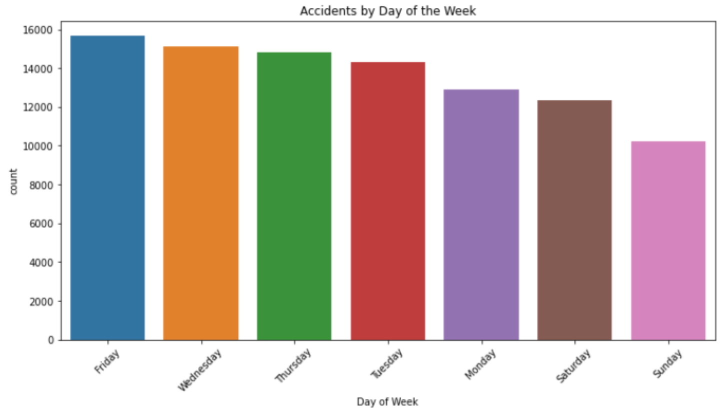
The Visualization Module plays a critical role in translating complex data into understandable and insightful visual representations. This module focuses on a variety of aspects of accident data, offering a multi-dimensional view of the factors influencing road safety. Through the use of correlation heatmaps, scatter plots, count plots, and histograms, the module explores the severity of accidents based on diverse factors like location, weather, and time. It delves into temporal patterns, highlighting trends across days of the week and months, and identifies regions or road segments that are particularly prone to accidents. The module also sheds light on the common types of accidents for different vehicle categories and examines the impact of weather and lighting conditions on accident frequency and severity. These visualizations are not just tools for analysis but also powerful means of communication, making the data accessible and comprehensible to a broader audience, including policymakers and the general public.

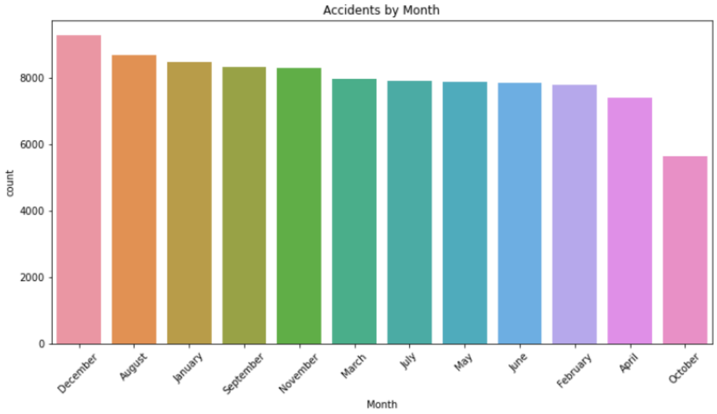
Visualizations of some correlations:

- **Heatmap to display the severity of Accidents based on Location, Weather, and Time**

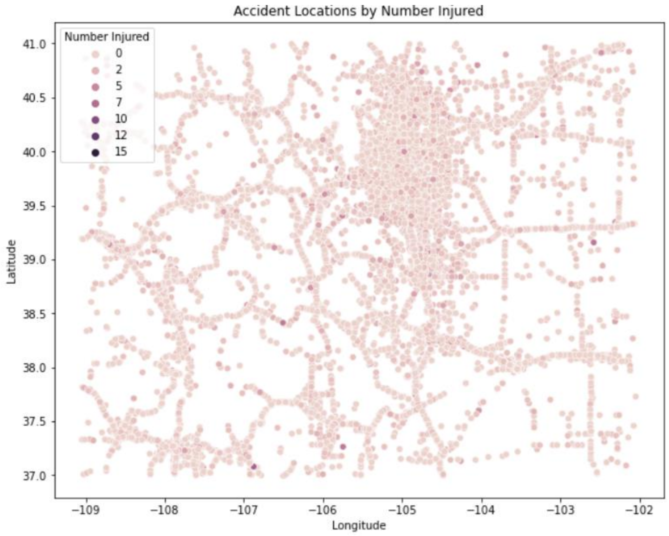


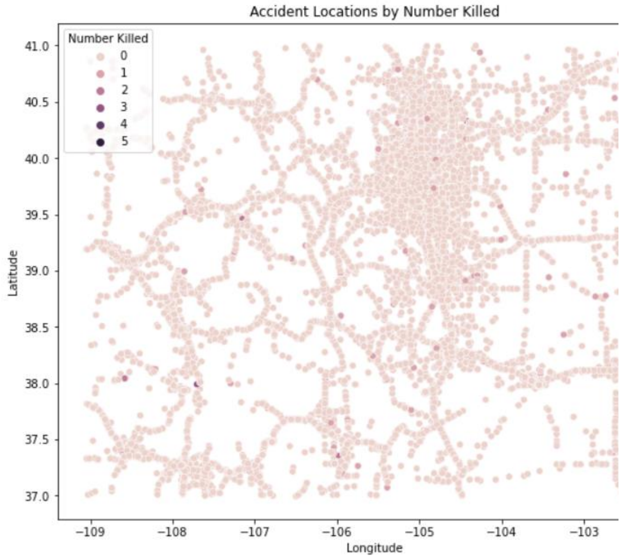
- Temporal Patterns in Accident Occurences



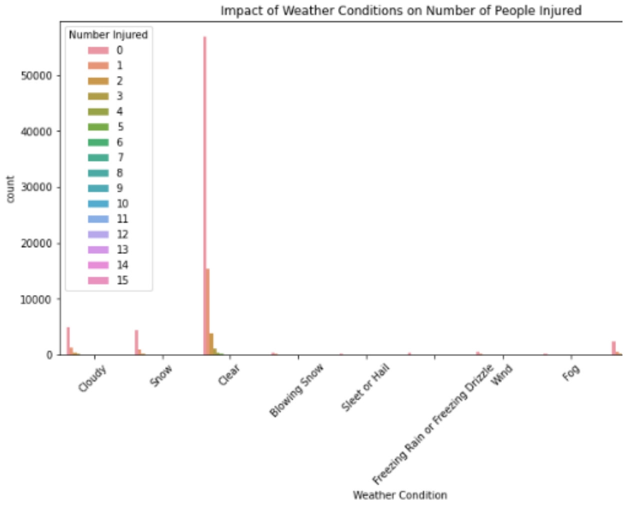


- Plotting Regions or Road Segments Prone to Accidents

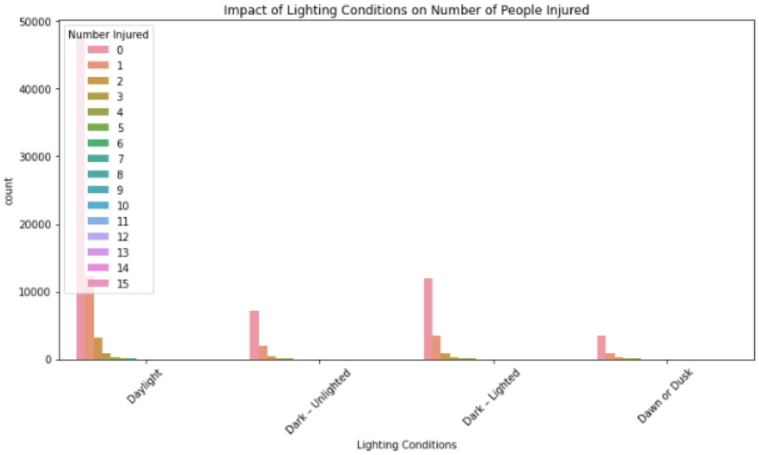




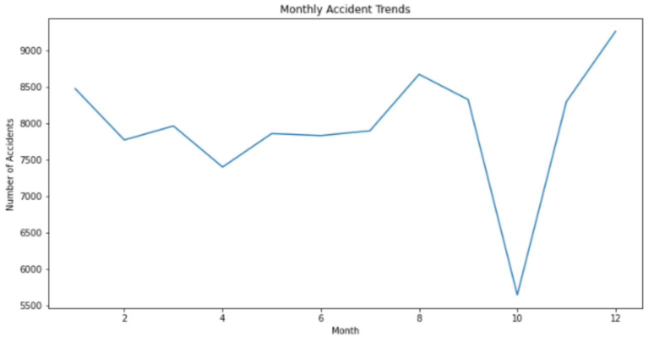
- Effect of Weather Conditions on Accident Frequency and Severity

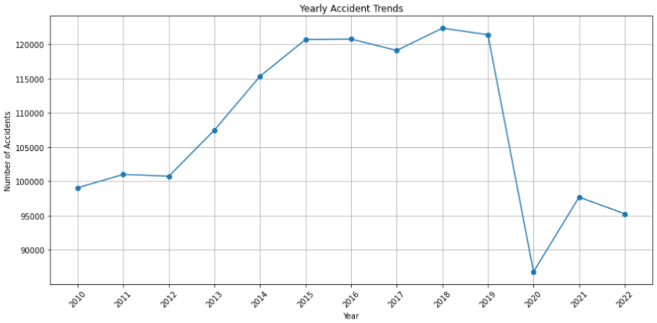


- Effect of Lighting Conditions on Accident Frequency and Severity

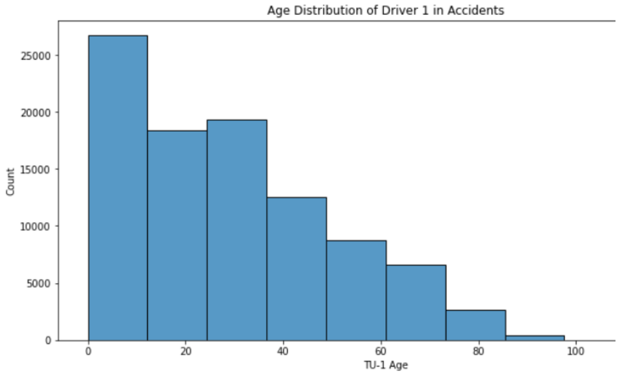


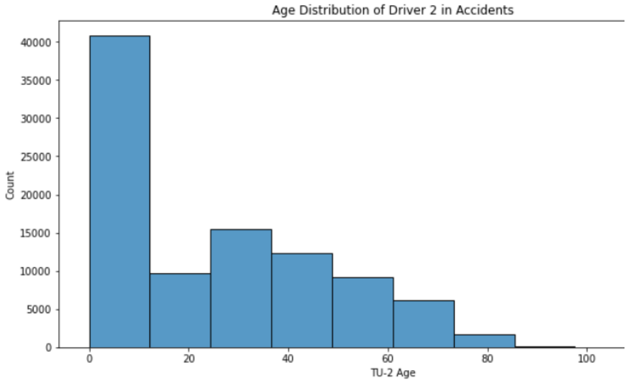
- Long-term Trends in Accident Data

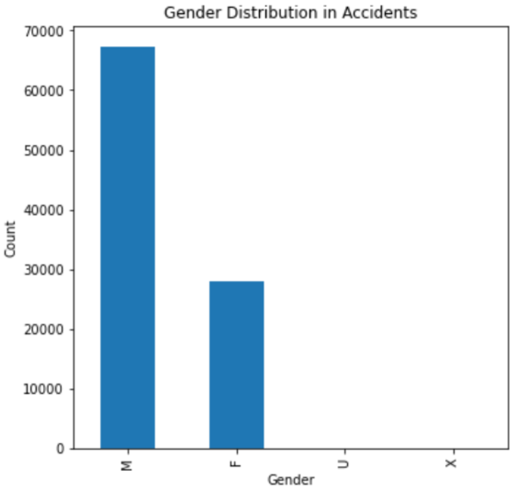




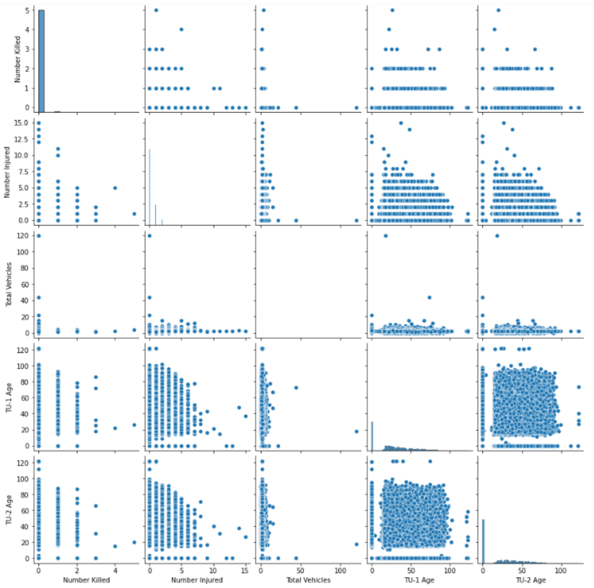
- Demographic Patterns in Accident Data







- A pairplot to show relationships between multiple variables



IV. CONCLUSION

We conducted a thorough review of traffic collision data from Colorado, for this study. We gathered and evaluated data on a variety of characteristics, including time of day, road conditions, location, severity, driver impairment, vehicle type, and injuries, using a dataset spanning 2010 to 2022. The primary goal was to find correlations and trends by employing four important machine learning models: ID3, Random Forest (RF), Nave Bayes, and Logistic Regression (LR).

The inspiration for this research originated from patterns found by the National Highway Traffic Safety Administration (NHTSA), in which the number of deaths increased despite a drop in the overall number of reported collisions. Previous research on road safety included various models, including logistic regression, decision tree, random forest, Tobit regression, and Bayesian spatial modeling. This research, however, frequently concentrated on certain data subsets or locations.

Our study broadened the scope by using a big dataset from Colorado, recognizing the need for a more comprehensive knowledge of the factors impacting traffic accidents. The machine learning models were chosen to give a variety of viewpoints on the data, and the research attempted to contribute to existing debates over the relative accuracy of various models.

The related work section included a thorough evaluation of current literature, grouping works into approaches such as road scenario-based discretization, safety performance functions, and nested logit model analysis of collision severities. This placed our findings into the larger framework of machine learning applications in traffic collision prediction.

The methodological approach was outlined in the experiment section, beginning with data gathering, preprocessing, analysis, model construction, and visualization. Pandas, NumPy, Scikit-Learn, Matplotlib, and Seaborn were used because of their versatility in data processing, statistical computation, machine learning methods, and visualization.

The experiment's lessons underlined the significance of thorough data preparation to preserve data integrity, the value of carefully selecting machine learning models depending on the nature of the data, and the usefulness of visualization in conveying complicated information.

In summary, by examining a broad dataset through the lens of machine learning models, our work adds to a better understanding of traffic accident patterns. The findings, which are not expressly stated in this overview, attempt to shed light on traffic incident correlations and trends. The lessons gained and future directions presented give insights for future study and advancements in traffic accident prediction and mitigation.

The tools chosen for this study included:

1. Pandas and NumPy: These libraries were essential for data manipulation, handling, and preprocessing due to their efficiency in working with structured data.

2. Scikit-Learn: Used for implementing machine learning models, Scikit-Learn provided a versatile toolkit for model development, training, and evaluation.

3. Matplotlib and Seaborn: These visualization libraries were crucial for creating insightful plots and graphs, aiding in the interpretation and communication of results.

Lessons Learned

1. Data Preprocessing is Key: Rigorous data preprocessing is essential for ensuring the quality and reliability of the dataset, influencing the effectiveness of subsequent analyses.

2. Model Selection Matters: Careful consideration is needed when selecting machine learning models, as the nature of the data may favor certain models over others.

3. Visualization Enhances Communication: Visualization is a powerful tool for conveying complex information in a comprehensible manner, benefiting both analysis and communication.

Team Contributions

Our team collaborated effectively, with each member contributing to various aspects of the project:

- Kurt Cushman: 25% - Led the data preprocessing, analysis and documentation phases.

- Sai Srikar Emani: 25% - Focused on model development and evaluation; code explanation video.

- Saurav Avinash Changde: 25% - Spearheaded the literature review and related work section.

- Harshitha Sai Addepalli : 25% - Took charge of the experiment design and visualization.

Overall, our team worked cohesively to ensure a thorough and well-rounded exploration of traffic crash data using machine learning methodologies.

V. FUTURE WORK

While the current study used machine learning models to give significant insights into traffic collision data, there are numerous opportunities for further research and enhancement:

- Integration of Additional States and Time Periods: Updating the study to incorporate data from additional states and a longer period may improve the conclusions' generalizability. Multiple areas may display distinct patterns driven by local circumstances, and a longer timescale may reveal changing tendencies.
- Feature Engineering and Selection: Exploration of sophisticated feature engineering approaches and a rigorous feature selection procedure might improve

model performance. Identifying the most significant aspects and modifying input attributes may help to improve forecast accuracy.

- Ensemble Approaches: Looking into ensemble approaches like stacking or boosting might be advantageous. Combining the predictions of different models may result in enhanced performance, particularly when capturing varied data patterns.
- Fine-Tuning Model Hyperparameters: A complete tuned hyperparameter method for every machine learning model may improve its performance. Fine-tuning the parameters assists in improving the models for certain dataset features.
- Temporal Analysis: A complete temporal analysis, which includes seasonality and long-term trends, may show patterns that emerge over time. This might entail investigating the effect of various months, seasons, or years on the frequency and severity of traffic accidents.
-

VI. REFERENCES

(1) Sun Z, Liu S, Li D, Tang B, Fang S. Crash analysis of mountainous freeways with high bridge and tunnel ratios using road scenario-based discretization. PLoS One. 2020 Aug 10;15(8):e0237408. doi: https://doi.org/10.1371/journal.pone.0237408. PMID: 32776981; PMCID: PMC7416955.

(2) Russo F, Biancardo SA, Dell'Acqua G. Road safety from the perspective of driver gender and age as related to the injury crash frequency and road scenario. Traffic Inj Prev. 2014;15(1):25-33. Doi: https://doi.org/10.1080/15389588.2013.794943. PMID: 24279963.

(3) Sunil Patil, Srinivas Reddy Geedipally, Dominique Lord, Analysis of crash severities using nested logit model—Accounting for the underreporting of crashes, Accident Analysis & Prevention, Volume 45, 2012, Pages 646-653, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2011.09.034.

(4) Anderson, J., & Hernandez, S. (2017). Heavy-Vehicle Crash Rate Analysis: Comparison of Heterogeneity Methods Using Idaho Crash Data. Transportation Research Record, 2637(1), 56-66. https://doi-org.aurarialibrary.idm.oclc.org/10.3141/2637-07

(5) Yansong Qu, Zhenlong Li, Qin Liu, Mengniu Pan, Zihao Zhang, "Crash/Near-Crash Analysis of Naturalistic Driving Data Using Association Rule Mining", *Journal of Advanced Transportation*, vol. 2022, Article ID 6562649, 19 pages, 2022. https://doi.org/10.1155/2022/6562649

(6) Chen M-M, Chen M-C. Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. *Information*. 2020; 11(5):270. https://doi.org/10.3390/info11050270

(7) Huang, H., Abdel-Aty, M. A., & Darwiche, A. L. (2010). County-Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling. Transportation Research Record, 2148(1), 27-37. https://doi-org.aurarialibrary.idm.oclc.org/10.3141/2148-04

(8) Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework. https://www.mdpi.com/1660-4601/16/3/334

(9) Tanmoy Bhowmik, Shamsunnahar Yasmin, Naveen Eluru, A New Econometric Approach for Modeling Several Count Variables: A Case Study of Crash Frequency Analysis by Crash Type and Severity, Transportation Research Part B: Methodological, Volume 153, 2021, Pages 172-203, ISSN 0191-2615, https://doi.org/10.1016/j.trb.2021.09.008.(https://www.sciencedirect.com/science/article/pii/S0191261521001818)

(10) Table of national crash statistic (https://cdan.dot.gov/tsftables/National%20Statistics.pdf).