Analyzing Driver Behavior and Crash Factors Using Machine Learning

NAME: KRISHNA VAMSI UPPALA

Table of Contents

1. Introduction

This project is oriented on analyzing Crash Reporting—Driver Data with the use of machine learning to determine the causes of traffic accidents and enhance the level of traffic safety. The project was selected because it is useful to society to minimize the number of accidents and deaths on the roads, which is a goal of public security. It provides a chance to use modern machine learning approaches on real-life data and, therefore, make progressive predictive estimates and recommendations concerning the quality of roads and potential accident rates.

1.1 Research questions

1. What are the most common weather conditions reported during crashes, and how do they vary across collision types?
2. How does driver distraction correlate with the extent of vehicle damage reported in crashes?
3. Can we build a predictive model to classify whether a driver is at fault in a crash based on the road, weather, and driver behavior variables?
4. Using a gradient-boosting algorithm, which factors are the most significant predictors of crash injury severity?

2.0 The Data

The Crash Reporting – Drivers Data was obtained from data.gov, compiled on November 10, 2020, and updated on September 20, 2024. It was collected through the Automated Crash Reporting System (ACRS) of the Maryland State Police and covers reports of the Montgomery County Police, Gaithersburg Police, Rockville Police, and Maryland-National Capital Park Police. We obtained the data by downloading it from data.gov, which records driver-related incidents in

Maryland. The dataset is accessible via this link: https://catalog.data.gov/dataset/crash-reporting-drivers-data.

The data set contains 184,521 observations and 39 variables that describe features of crash reports. These variables can be categorized into categorical, integer, and float.

Categorical Variables (32 variables):

1. Report Number: Unique identifier for each report.
2. Local Case Number: Local identifier for cases.
3. Agency Name: Name of the reporting agency.
4. ACRS Report Type: Type of report filed.
5. Crash Date/Time: Date and time of the crash incident.
6. Route Type: Type of roadway involved.
7. Road Name: Name of the road where the crash occurred.
8. Cross-Street Name: Name of the nearest cross-street.
9. Off-Road Description: Description of off-road incidents.
10. Municipality: Local government jurisdiction.
11. Related Non-Motorist: Information on non-motorist involvement.
12. Collision Type: Nature of the collision.
13. Weather: Weather conditions during the crash.
14. Surface Condition: Condition of the road surface.
15. Light: Lighting conditions at the time of the crash.
16. Traffic Control: Type of traffic control present.
17. Driver Substance Abuse: Information on driver substance use.
18. Non-Motorist Substance Abuse: Information on substance use by non-motorists.

19. Person ID: Unique identifier for individuals involved.

20. Driver At Fault: Indicates if the driver was at fault.

21. Injury Severity: Severity of injuries reported.

22. Circumstance: Circumstances surrounding the crash.

23. Driver Distracted By: Distractions affecting the driver.

24. Drivers License State: State issuing the driver’s license.

25. Vehicle ID: Unique identifier for each vehicle involved.

26. Vehicle Damage Extent: Extent of damage to the vehicle.

27. Vehicle First Impact Location: Location of the first impact.

28. Vehicle Body Type: Type of vehicle body.

29. Vehicle Movement: Movement of the vehicle before the crash.

30. Vehicle Going Direction: The direction the vehicle was traveling.

31. Driverless Vehicle: Indicates if a driverless vehicle was involved.

32. Parked Vehicle: Information on parked vehicle involvement.

33. Vehicle Make: Manufacturer of the vehicle.

34. Vehicle Model: Specific model of the vehicle.

35. Location: Geographic location of the crash.

Integer Variables (2 variables):

36. Speed Limit: Speed limit on the road where the crash occurred.

37. Vehicle Year: The year of manufacture of the vehicle is involved.

Float Variables (3 variables):

38. Latitude: Latitude coordinate of the crash location.

39. Longitude: Longitude coordinate of the crash location.

The dataset provides essential information concerning the motor vehicle drivers involved in accidents on Montgomery County and local roads. It involves all aspects of details about traffic accidents, including different factors like weather conditions, road conditions, and driver's behaviors. This data is essential for recognizing traffic safety trends in urban and suburban territories and for making decisions appropriate for local authorities and police. Studying such occurrences will help the stakeholders determine patterns and come up with measures to improve road safety and, thereby, minimize the rates of accidents.

3.0 Literature review

3.1 Introduction

This literature review is to briefly discuss previous work done in the field of traffic safety, crash data, and the application of machine learning methods in predicting and analyzing crash factors. These areas are reviewed to define the background of the current project, which will, in turn, help in making the required findings on the best approach to evaluating crash data. Previous research is useful in revealing the ways of the project's application as well as in ensuring that the optimal method for improving traffic safety and reducing the number of accidents has been applied.

3.2 Traffic Safety and Crash Analysis

The quantitative study undertaken by Ellison et al., (2015) aimed at identifying trends in traffic safety about driver behaviour and the environment factor. The researchers collected quantitative data on the crash and qualitatively surveyed the drivers. The sample included 5,000 traffic accidents recorded in the police reports from five years in the metropolitan region. The study showed that risky driving behaviour, including speeding and driving under the influence, greatly enhanced crash proneness, especially when weather conditions are adverse, such as rainy or foggy weather. The study revealed that vehicle types, including SUVs, were common in accidents during these unfavourable conditions and thus require specific interventions and awareness.

Zheng (2012) aimed at determining the correlation between traffic flow attributes and crash frequencies. The researcher used a quantitative research method to gather traffic volume and crash data from the Department of Transportation databases for two years. His analysis was done by regression modelling to determine the relationship between different factors. The findings portrayed that an increased traffic flow led to a higher risk of accidents, especially at the junctions. The study concluded that cyclists and pedestrians were factors that affected crash rates, and this should be considered in traffic safety planning.

3.3 Predictive Modeling in Traffic Safety

Yang et al., (2022) aimed to develop machine learning-based predictive models to predict the crash risk in intersections. The authors analyzed the data on 10,000 crash reports from the city traffic databases for three years using decision trees and logistic regression. As from the decision tree algorithm, risk levels of crashes were well classified by traffic volume, road geometry and driver behaviour. On the other hand, the logistic regression model painted a probability of crashes in intersections. The findings in this study indicated that the decision tree model was superior in the high-risk prediction of intersections as compared to the logistic regression model. This study underlined the necessity of applying more than one algorithm to enhance the predictive characteristics. However, it admitted some weaknesses regarding the quality of data and the updating and upgrading of the models for greater validity.

Wu et al. (2024) sought to estimate the injury severity in crashes using random forest classification in combination with gradient boosting. This study utilized data from 15,000 crash incidents collected from the National Highway Traffic Safety Administration database; the data comprises demographic factors, environmental factors, and vehicle factors. The random forest approach was demonstrated to be robust to missing values, and it was determined that factors that impact the level of injury include the type of vehicle and weather conditions. The gradient boosting algorithm improved the prediction of the model by adding interaction terms of variables. These models were proved to make reasonable predictions of the injury severity, while some problems still existed, such as how to deal with the bias of data history that may influence the results.

3.4 Exploratory Data Analysis (EDA) in Traffic Studies

The research done by Jalayer et al. (2018) had the main objective of carrying out EDA to identify the relationships between weather factors and crash severity. The authors selected 12,000 traffic accident reports from a large urban population, concentrating on weather conditions, time of day, and crash outcome variables. Scatter plots and heat maps were used to represent the data trends and the patterns that showed that there was a strong relation between crash severity and rain and snow. The research established that crashes during poor weather conditions are more lethal as compared to crashes under any other weather condition and thus prompted a debate on how to improve the signs to prevent traffic accidents under poor weather conditions and public awareness during adverse weather conditions.

Kmet & Macarthur (2006) aimed to analyze the correlation between road types and accident rates using EDA. The researchers used a sample of 18,000 crash reports compiled by the type of road: highways, rural, and urban streets. The technique used bar charts and box plots to present the accident frequencies by type of road. The research found that the roads in the urban areas had the highest number of accidents, most especially during rush hours, while the rural areas had more severe accidents than any other areas. The analysis revealed the need for the definition of specific actions on traffic control and the further study of the potential of the infrastructural development in the cities as well as EDA enhancement for traffic safety enhancement.

3.5 Machine Learning Applications in Crash Data Analysis

Santos et al. (2022) performed a study with the primary objective of using a machine-learning algorithm to classify the degree of traffic crash injuries. The researchers used a sample of 15000 crash reports in the context of urban environment, climate, age and sex of the driver of the vehicle and type of vehicle. The method used was gradient boosting machines (GBM) to classify outcomes by the level of the injury and the model was assessed for accuracy, precision, and recall. It also pointed out that the developed GBM model achieved a high accuracy of 87% as compared to the statistical models. This success led to the conclusion that it is possible to further apply the presented machine learning methods to model the other higher-level characteristics of crash data and to improve the predictive traffic safety analysis.

Mamdoohi and Miller-Hooks (2022) employed clustering to study the crash rates to determine the high crash zones. The rationale for this research was to use K-means clustering to classify crashes with the aim of segmenting crashes by place, time and state of the environment. Crash types, weather conditions, and road conditions were the attributes documented concerning the 20,000 traffic accidents. By using clustering, the researchers were in a place to measure some of the areas that received more accidents, which frequently occurred at night and during unfavourable weather conditions. The findings suggested that the proposed ML approach could detect the relationship between the variables and the structure of data, which will be beneficial for the traffic safety authorities to plan their resource availability and use particular measures that can reduce crash occurrences.

3.6 Challenges and Limitations in Crash Data Analysis

Deficiencies and challenges of traffic crash data analysis have been noted in the reviewed literature, particularly with regard to data accuracy. Several of the research studies have pointed out that the reliability of crash data and data comprehensiveness are key concerns as far as predictive models are concerned. For example, in several cases, there were missing values in several datasets due to unobserved driver behaviour or environmental factors at the time of the accident. Consequently, there has been no full data set, and this has called for data imputation techniques that, if not well managed, can lead to bias (Pampaka et al., 2016). The issues of feature selection are another challenge: the researchers are faced with the challenge of choosing which variables should be included in the models without crowding the models while at the same time ensuring that the models do not remove significant variables that can have a big impact on the outcome of a crash.

Challenges were observed in terms of technical and methodological approaches when applying machine learning algorithms to crash data. Several of the works described difficulties related to model interpretability, particularly when employing high-complexity models, including gradient boosting. These models, as realistic as they are, can sometimes complicate the relationship between features, hence complicating the ability of the practitioner to make some useful inference on the results obtained. The process of training and checking a model for model validity may be very time consuming and requires so much computational resources and this may not be affordable to small research groups or agencies (Himanen et al., 2019). These problems simultaneously cause validity concerns and increase the need for attention to data cleaning, model reporting and the potential use of more interpretable techniques in the future.

4. Methods

4.1 Tools used

To perform this detailed analysis of Crash Reporting—Driver Data, Jupyter Notebook, an interactive coding environment, and Python, a powerful programming language, were employed. We found using Jupyter Notebook quite effective for writing well-arranged and commented code along with data visualizations in a single file to enhance data analysis and presentation throughout the process. Python has a variety of libraries, such as Pandas for data handling, Matplotlib, and Seaborn for data representation, and Sklearn for machine learning. These mentioned tools were effectively applied in the data cleaning process, data pre-processed, categorical data encoding, model construction and feature relevance.

4.2 Techniques

4.2.1 Data importation

To begin the analysis, major libraries required in handling, visualizing and modelling data were imported. For data management and handling, the pandas library was used while for numerical computation the numpy was used. Visualization techniques from Seaborn and Matplotlib were useful in visualizing data which is very essential in identifying patterns in the data as well as in the right interpretation of insights (Tariq et al., 2023). XGBoost provided better options for the implementation of the gradient-boosting model and the enhancement of its performance, and it provided the feature importance analysis option. For the preparation of the data for the machine learning, other libraries such as sklearn.model_selection for the splitting of data and LabelEncoder for encoding of categorical variables were useful.

4.2.2 Data Reading

The required dataset "Crash_Reporting_-_Drivers_Data.csv" was read and loaded into the environment from using the read_csv function from pandas and stored in a DataFrame. This method was helpful in looking at the raw data and getting an overall idea of what the data looks like and some of the first features of the data.

4.2.3 Data Preprocessing

In the preprocessing step, the data was cleaned and transformed for analysis and modelling and included data cleansing, filtering and data transformation.

4.2.3.1 Data cleaning

In the strategy of dealing with the missing values, the data was analyzed by the type of the column. Missing values in numeric columns were imputed by mean to keep numerical data consistent and reduce any possible bias; for categorical columns missing values were replaced by the most frequent value to keep categorical data consistent.

To remove any possibility of bias in the results due to repeated data entries, df.duplicated().any() was used to search for and eliminate dups. Duplicate handling made it possible to achieve data uniqueness to the distribution of the dataset (Li et al., 2021).

Certain columns were then pre-selected in order to keep only the records of interest. For example, records with 'Weather' factors such as 'Rain,' 'Cloudy,' and 'Clear' were excluded, together with records that contain imprecise values for 'Collision Type,' such as 'OTHER' and 'UNKNOWN.' This filtered dataset gave a rather specific foundation for detailed analysis without outside interference from unnecessary data.

4.2.3.2 Exploratory Data Analysis

In the EDA step, relationships within the data were investigated by using graphs pertinent to the research questions in an effort to identify trends.

Research Question 1: What are the most common weather conditions reported during crashes, and how do they vary across collision types?

To identify the most common weather conditions during crashes and their variance across collision types, the dataset was grouped by 'Weather' and 'Collision Type,' and entries with low counts were excluded for clarity. Using Seaborn’s barplot(), the analysis displayed the frequency of crashes under different weather conditions, segmented by collision types.

Research Question 2: How does driver distraction correlate with the extent of vehicle damage reported in crashes?

To examine the impact of driver distractions on vehicle damage, irrelevant data from 'Vehicle Damage Extent' and 'Driver Distracted By' columns were filtered out. The dataset was grouped to show counts for combinations of damage extent and distraction type, which was then visualized with a bar plot.

4.2.4 Data Encoding

To apply machine learning to the data, the categorical data in the features under consideration and the target variable ‘Driver At Fault’ were encoded numerically. These categorical variables were transformed into integers for ease of interpretation by the machine learning algorithms through label encoding. This encoding process was useful for training and testing the dataset as all features needed to be in a format that was understandable by machine (Pfülb, 2022).

4.2.5 Build a Predictive Model to Classify Whether a Driver is at Fault

To develop the model, the XGBoost classifier was used to classify driver fault due to its effectiveness when working with big data. After the data was divided into training and test sets, the training set was used to train the model. The model then predicted fault classifications on the test data, and a confusion matrix was generated to assess its classification accuracy visually.

4.2.6 Using a Gradient-Boosting Algorithm to Identify Significant Predictors of Crash Injury Severity

A Gradient-Boosting Classifier (XGBoost) was applied again to identify key predictors of crash injury severity. The target variable here was 'Injury Severity,' and feature selection was critical; only relevant columns were retained. Post encoding, the data was split into training and test sets, and the model was trained on injury severity classification. Feature importance scores from the XGBoost model were analyzed, isolating the top five predictive factors, which were visualized to highlight their relative impact on crash injury outcomes.

5.0 Preliminary Results

5.1 Research question 1: What are the most common weather conditions reported during crashes, and how do they vary across collision types?

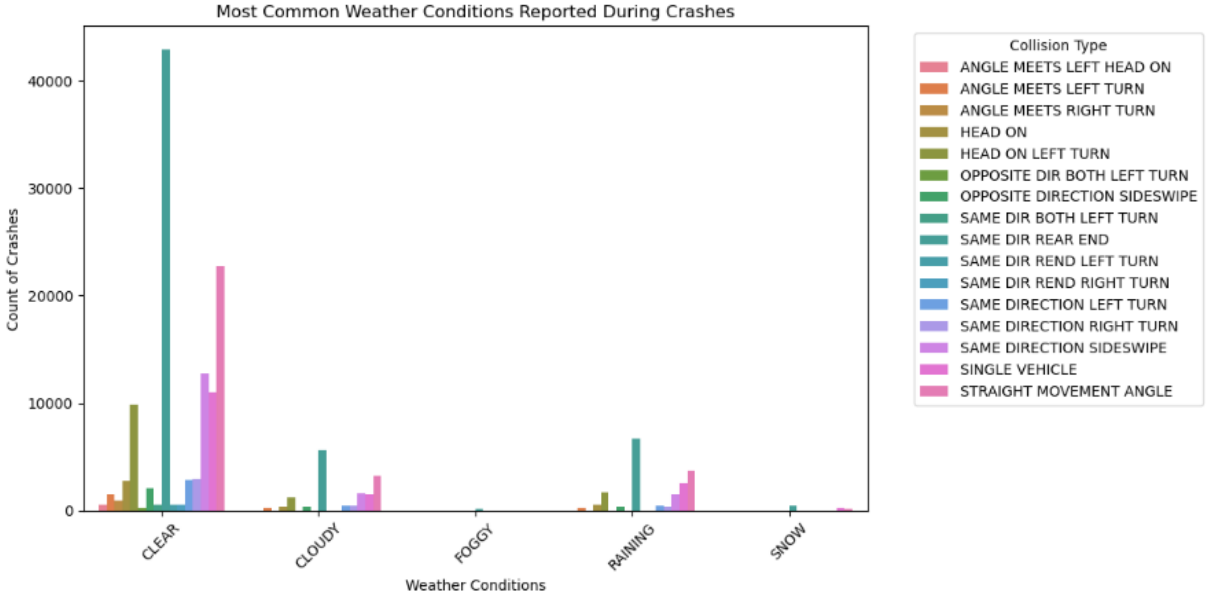


Figure 1: Most Common Weather Conditions Reported During Crashes

The results reveal that clear weather conditions are the most frequently reported during crashes, with the "Same Direction Rear End" collision type being the most common, totalling 42,932 incidents. This is followed by "Straight Movement Angle" collisions, also in clear weather, with 22,766 cases. Among cloudy conditions, the highest number of crashes occurred in "Same Direction Rear End" collisions (5,619), while "Straight Movement Angle" and "Single Vehicle" incidents were also prominent, with 3,272 and 1,536 cases, respectively. Raining conditions saw "Same Direction Rear End" collisions as the most frequent, with 6,715 occurrences, while "Straight Movement Angle" and "Single Vehicle" crashes followed at 3,723 and 2,601, respectively. In snowy weather, "Same Direction Rear End" collisions remained predominant, with

426 incidents, whereas "Single Vehicle" and "Straight Movement Angle" types were recorded at 225 and 200 crashes.

5.2 Research Question 2: 2. How does driver distraction correlate with the extent of vehicle damage reported in crashes?

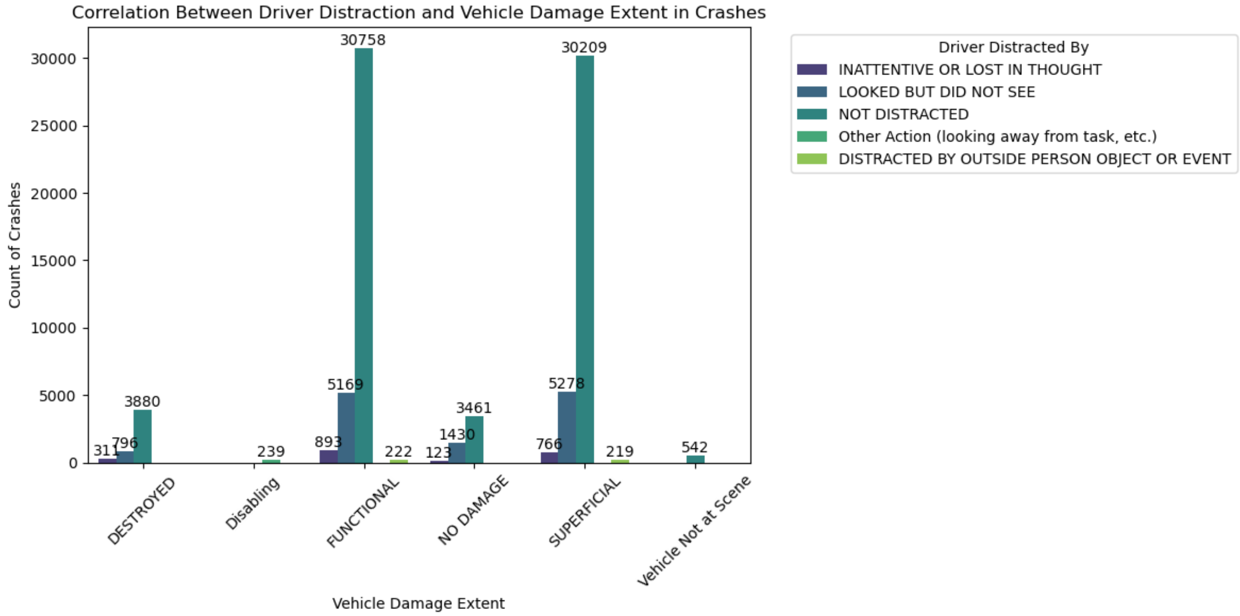


Figure 2: Correlation between Driver Distraction and Vehicle Damage Extent in Crashes

The results show a correlation between driver distraction and the extent of vehicle damage in crashes, with distinct frequencies across distraction types. "Destroyed" vehicles most commonly involved drivers who were not distracted, totalling 3,880 cases, followed by "Looked But Did Not See" at 796 incidents. For "Functional" damage, the highest frequency was also among drivers not distracted, with 30,758 cases, while 5,169 cases involved drivers who "Looked But Did Not See." In "Superficial" damage incidents, drivers not distracted represented the majority at 30,209 cases, followed by "Looked But Did Not See" at 5,278 cases. "No Damage" crashes were likewise more frequent among non-distracted drivers, totalling 3,461, with fewer cases for drivers who "Looked But Did Not See" at 1,430.

5.3 Research Question 3: Can we build a predictive model to classify whether a driver is at fault in a crash based on the road, weather, and driver behavior variables?

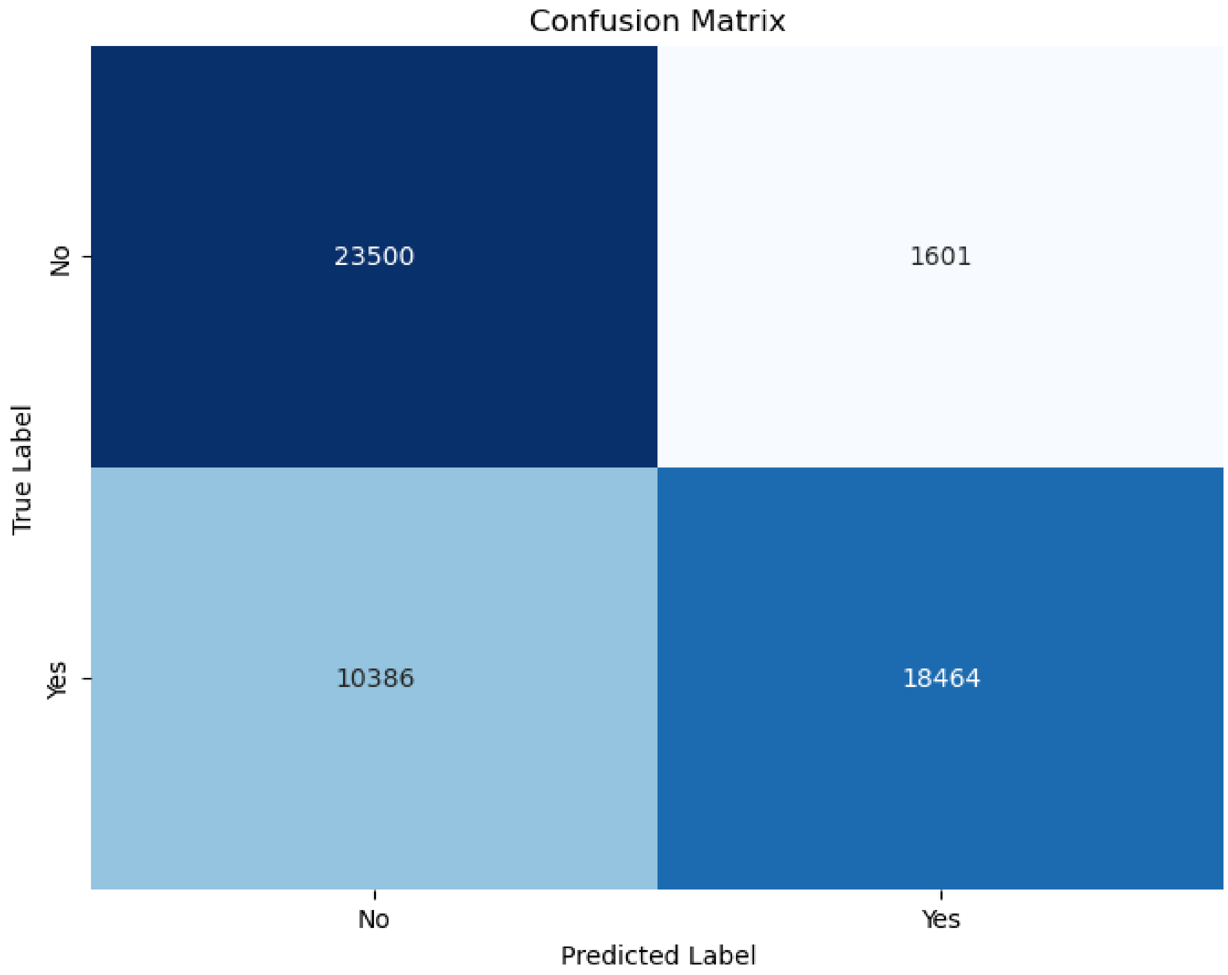
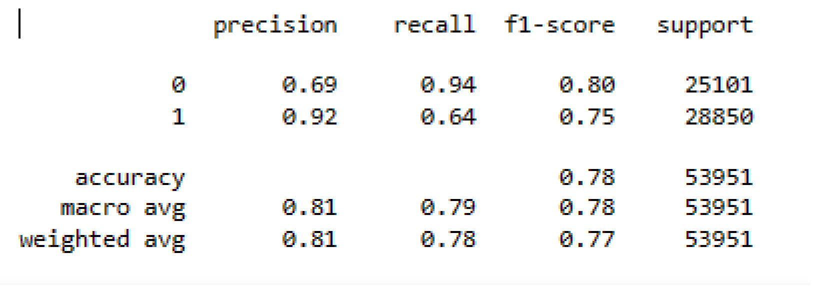


Figure 3: Confusion matrix

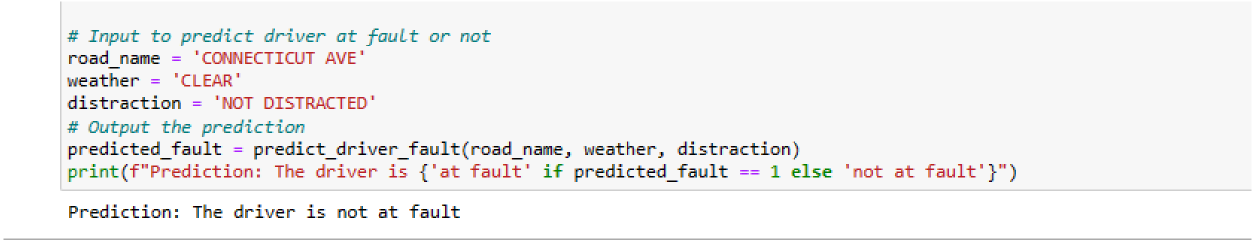


```
              precision    recall  f1-score   support

           0       0.69      0.94      0.80     25101
           1       0.92      0.64      0.75     28850

    accuracy                           0.78     53951
   macro avg       0.81      0.79      0.78     53951
weighted avg       0.81      0.78      0.77     53951
```

Figure 4: Classification report



```
# Input to predict driver at fault or not
road_name = 'CONNECTICUT AVE'
weather = 'CLEAR'
distraction = 'NOT DISTRACTED'
# Output the prediction
predicted_fault = predict_driver_fault(road_name, weather, distraction)
print(f"Prediction: The driver is {'at fault' if predicted_fault == 1 else 'not at fault'}")
```

```
Prediction: The driver is not at fault
```

Figure 5: prediction of whether the driver is at fault or not.

The predictive model developed to classify driver fault in crashes based on road, weather, and driver behavior variables yielded a confusion matrix revealing 23,500 true negatives and 16,01 false positives, indicating a robust performance in identifying non-fault cases. Among the 28,850 instances of drivers at fault, the model correctly identified 18,464 true positives, leading to an overall accuracy of 78%. The classification report highlights a precision of 0.69 and a recall of 0.94 for non-faulty drivers, while for drivers at fault, it shows a precision of 0.92 and a recall of 0.64. The macro average precision stands at 0.81, with a recall of 0.79 and an F1-score of 0.78. The weighted averages further reinforce these metrics, showing precision and recall scores of 0.81 and 0.78, respectively.

Based on the goodness of the model, the prediction of whether the driver is at fault or not was made using ‘CONNECTICUT AVE’ as the 'Road Name' feature entry, 'CLEAR' as the 'Weather' feature entry, and 'NOT DISTRACTED' as the 'Driver Distracted By' feature entry. The outcome was that the driver was not at fault as shown in figure 5.

5.4 Research Question 4: Using a gradient-boosting algorithm, which factors are the most significant predictors of crash injury severity?

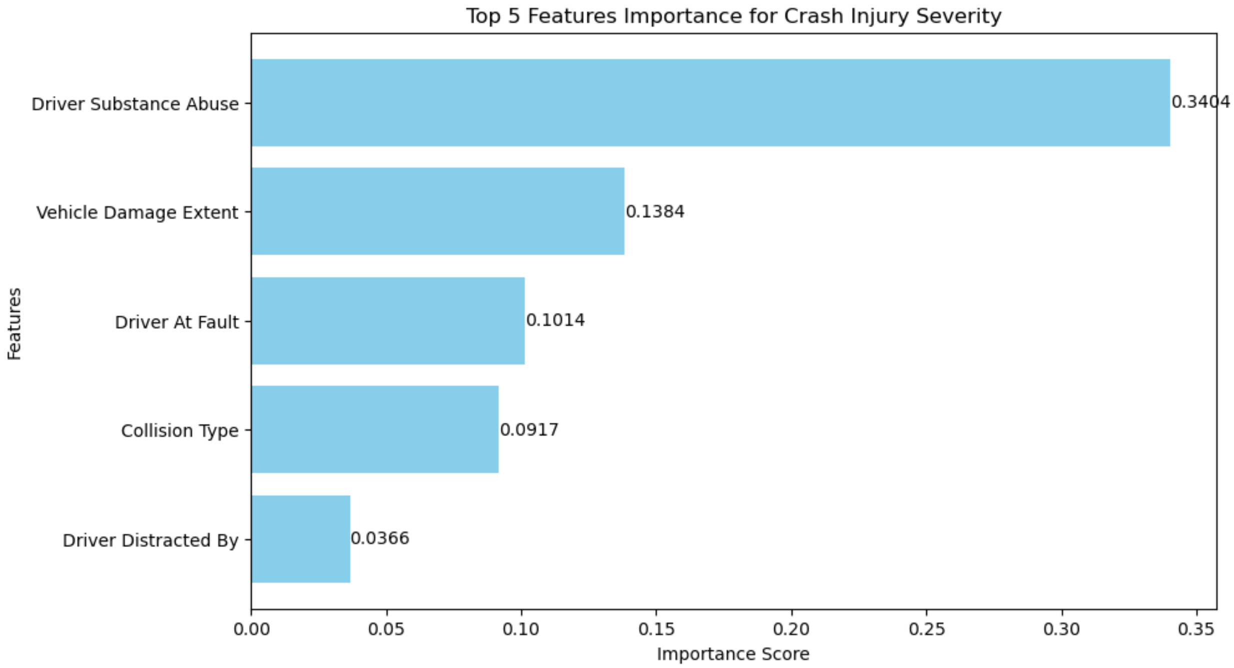


Figure 6: Top 5 Features Important for Crash Injury Severity

The analysis of crash injury severity using a gradient-boosting algorithm identified five significant predictors. The most critical factor was driver substance abuse, with an importance score of 0.3404, indicating a strong correlation with injury severity. Vehicle damage extent emerged as the second most important predictor, with an importance score of 0.1384, suggesting that greater damage may lead to more severe injuries. The third predictor, the driver at fault, had an importance score of 0.1014, highlighting the relevance of fault in determining injury outcomes. The collision type was identified as a significant factor with an importance score of 0.0917, indicating variations in injury severity based on the nature of the collision. Driver distracted by was the least influential of the top five features, with an importance score of 0.0366, suggesting a lesser impact on injury severity compared to the other factors.

6.0 Discussion/ Interpretation

The results of this research indicate a complex relationship between factors in traffic incidents, particularly concerning weather conditions and collision types. Clear weather emerged as the predominant condition under which most crashes occurred, particularly for the "Same Direction Rear End" collision type. This suggests that while clear weather is typically perceived as safe driving conditions, the sheer volume of traffic on roads may contribute to a higher incidence of these types of collisions. The persistence of "Straight Movement Angle" crashes in clear conditions further indicates that even in favorable weather, driver behavior or road characteristics may play critical roles in accident frequency. In contrast, adverse weather conditions, such as rain and snow, still showed "Same Direction Rear End" collisions as prevalent, emphasizing that certain collision types remain consistently problematic regardless of weather variations.

The analysis of driver distraction reveals a significant correlation with the extent of vehicle damage. Interestingly, the data indicate that a substantial number of crashes with "Destroyed" vehicles involved drivers who were not distracted, which raises questions about the underlying factors contributing to such severe outcomes. The data highlight that non-distracted drivers were involved in the majority of crashes across all damage extents, suggesting that factors other than distraction, such as speed, vehicle safety features, or even the nature of the impact, might be at play. Conversely, while the number of crashes for drivers who "Looked But Did Not See" is notably lower, their involvement in significant damage cases indicates a potential gap in driver awareness and situational assessment, which could warrant further investigation.

In evaluating the predictive model's performance concerning driver fault classification, the results suggest that the model successfully differentiates between fault and non-fault cases, achieving an overall accuracy of 78%. The relatively high precision and recall for identifying drivers at fault indicate the model's effectiveness in recognizing true positives in fault situations. However, the lower recall for non-faulty drivers implies potential challenges in identifying all instances accurately, which may be crucial for liability assessments. The balanced performance metrics, both macro and weighted averages, demonstrate that the model maintains reliability across varying data instances, which is essential for its application in real-world scenarios.

The gradient-boosting algorithm showed that driver substance abuse contributes most to the crash injury severity, highlighting the importance of driver behaviour in affecting safety impacts. As a result, having a high importance score it means that working with substance abuse could ensure significant improvements in decreasing the rates of severe injuries. The relatively high degree of attention paid to vehicle damage extent suggests that it, too, has value in predicting the severity of an injury; more damage may imply a more serious injury, which may be attributed to a higher force of impact. The fact that the driver is at fault and the collision type strengthens the view that the circumstances of a crash determine injury severity, whereas driver distraction has a less marked impact. Such a differentiation encourages analysing the connections between these factors and the possibility of special approaches to enhance road safety in general.

7.0 Recommendations

To address the prevalence of "Same Direction Rear End" collisions in clear weather conditions, road safety agencies should enhance driver education and awareness campaigns aimed specifically at promoting safe driving behaviors in clear weather. Given that clear weather accounts for the majority of reported crashes, an implementation strategy could involve targeted public service announcements and community outreach programs emphasizing the importance of maintaining safe following distances and adjusting speeds appropriately, even when conditions appear ideal. This initiative could benefit public safety by reducing the number of preventable accidents that occur during clear weather, ultimately leading to safer roadways.

In light of the correlation between driver distraction and vehicle damage severity, insurance companies and vehicle manufacturers should collaborate to develop advanced driver assistance systems (ADAS) that actively monitor and mitigate distractions. An effective implementation strategy would involve integrating technology that alerts drivers when they are likely to be distracted, such as through visual or auditory notifications. By addressing this issue proactively, the benefits would extend beyond mere accident prevention; improved vehicle safety features could lead to reduced insurance premiums and enhanced consumer trust in-vehicle technologies, creating a safer driving environment overall.

To leverage the predictive model's ability to classify driver faults, law enforcement agencies should incorporate these advanced algorithms into their traffic accident investigations. An implementation strategy could include training officers to use the predictive model when analyzing crash data, allowing for more accurate assessments of fault. The benefits of this approach would include streamlined investigations, clearer liability determinations, and potentially reduced legal disputes arising from traffic incidents.

8.0 Conclusion

In conclusion, this research underscores the complex interplay between weather conditions, driver behavior, and collision types, revealing that even under clear conditions, "Same Direction Rear End" collisions are prevalent, suggesting a need for targeted driver education. The correlation between driver distraction and vehicle damage severity highlights the potential benefits of implementing advanced safety technologies to mitigate distractions. The effectiveness of the predictive model in accurately classifying driver fault opens avenues for its integration into law enforcement practices, promoting fairer accident assessments. Future work should focus on refining the predictive algorithms to enhance their accuracy further and exploring the development of comprehensive safety interventions that address both driver education and technology adoption, ultimately aiming to reduce crash incidence and severity on our roadways.

References

Ellison, A. B., Greaves, S. P., & Bliemer, M. C. (2015). Driver behaviour profiles for road safety analysis. Accident Analysis & Prevention, 76, 118-132. https://doi.org/10.1016/j.aap.2015.01.009

Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-driven materials science: status, challenges, and perspectives. Advanced Science, 6(21), 1900808. https://doi.org/10.1002/advs.201900808

Jalayer, M., Zhou, H., & Das, S. (2018). Exploratory analysis of run-off-road crash patterns. In Data Analytics for Smart Cities (pp. 183-200). Auerbach Publications. https://www.taylorfrancis.com/chapters/edit/10.1201/9780429434983-8/exploratory-analysis-run-road-crash-patterns-mohammad-jalayer-huaguo-zhou-subasish-das

Kmet, L., & Macarthur, C. (2006). Urban–rural differences in motor vehicle crash fatality and hospitalization rates among children and youth. Accident Analysis & Prevention, 38(1), 122-127. https://doi.org/10.1016/j.aap.2005.07.007

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., & Zhang, C. (2021, April). Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 13-24). IEEE. https://doi.org/10.1109/ICDE51399.2021.00009

Mamdoohi, S., & Miller-Hooks, E. (2022). Identifying the impact area of a traffic event through k-means clustering. Journal of big data analytics in transportation, 4(2), 153-170. https://doi.org/10.1007/s42421-022-00060-9

Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research & Method in Education, 39(1), 19-37. https://doi.org/10.1080/1743727X.2014.979146

Pfülb, B. (2022). Continual Learning with Deep Learning Methods in an Application-Oriented Context. arXiv preprint arXiv:2207.06233. https://doi.org/10.48550/arXiv.2207.06233

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research, 80, 254-269. https://doi.org/10.1016/j.jsr.2021.12.007

Tariq, D. T. H. S., & Aithal, P. S. (2023). Visualization and Explorative Data Analysis. International Journal of Enhanced Research in Science, Technology & Engineering, 12(3), 11-21. https://dx.doi.org/10.2139/ssrn.4400256

Wu, Z., Misra, A., & Bao, S. (2024). Modeling pedestrian injury severity: a case study of using extreme gradient boosting vs random forest in feature selection. Transportation research record, 2678(1), 1-11. https://doi.org/10.1177/03611981231170014

Yang, Y., He, K., Wang, Y. P., Yuan, Z. Z., Yin, Y. H., & Guo, M. Z. (2022). Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods. Physica A: Statistical Mechanics and Its Applications, 595, 127083. https://doi.org/10.1016/j.physa.2022.127083

Zheng, Z. (2012). Empirical analysis on relationship between traffic conditions and crash occurrences. Procedia-Social and Behavioral Sciences, 43, 302-312. https://doi.org/10.1016/j.sbspro.2012.04.103