# Analysis of Vehicular Crash Data

**Dylan Schroers, Vaidehi Vaishnav**

**ABSTRACT** Road safety remains a pressing concern for communities worldwide. As vehicular populations grow and transportation networks become increasingly complex, the challenge of ensuring safe roads intensifies. This report delves into a comprehensive analysis of crash data, with a primary focus on predicting injury severity. Through rigorous data cleaning, preprocessing, and application of advanced data mining techniques, the study seeks to unearth patterns and insights that can inform targeted safety interventions. By bridging the gap between empirical data and actionable insights, this research aims to contribute meaningfully to the ongoing discourse on road safety and pave the way for evidence-based strategies that prioritize injury prevention and mitigation.

## I. INTRODUCTION

In an era characterized by rapid urbanization and technological advancements, the importance of road safety cannot be overstated. Roads, serving as lifelines of modern societies, facilitate economic activities, connect communities, and enable mobility. However, this connectivity comes with inherent risks, as evidenced by the alarming statistics of road crashes and their associated human toll. Every crash, beyond the immediate physical damage, represents shattered lives, families torn apart, and communities grappling with loss.

Recognizing the multifaceted nature of road safety, this report embarks on a journey to dissect crash data, unraveling the underlying factors that contribute to injury severity. The endeavor is not merely academic; it is a quest for actionable insights that can catalyze meaningful change. By harnessing the power of data, we aspire to illuminate the dark corners of road safety, identifying risk factors, understanding their interplay, and devising strategies to mitigate their impact.

The stakes are high, but so is the potential for positive change. Through a meticulous analysis of crash data, this report seeks to empower policymakers, safety advocates, and communities with the knowledge and tools needed to foster safer roads. In doing so, we reaffirm our collective commitment to a future where every journey is a step towards safety, and every road, a testament to our unwavering dedication to preserving human life and well-being.

## II.BACKGROUND/RELATED WORK

The study of road safety has been a focal point for researchers, policymakers, and safety advocates for decades. Numerous studies have delved into understanding the intricate dynamics of road crashes, attempting to discern patterns, causes, and repercussions. Here's a more detailed exploration:

### *A. Historical Perspective*

Over the years, the evolution of road safety measures has seen various interventions, from improved vehicle design and safety features to infrastructure changes aimed at minimizing the impact of collisions. The foundation of these interventions often rests on empirical evidence derived from crash data analysis.

### *B. Factors Influencing Crash Severity*

While the occurrence of crashes is multifaceted, certain determinants consistently emerge as significant contributors to the severity of injuries. These can range from driver behavior, vehicle speed, road conditions, to the types of vehicles involved. Previous research has extensively studied these factors, often highlighting their interplay and cumulative effect.

### *C. Data Mining in Road Safety*

With the advent of advanced computational techniques and the proliferation of data collection systems, data mining has emerged as a powerful tool in the realm of road safety. By sifting through vast datasets, data mining techniques can uncover hidden patterns, correlations, and predictive models, offering a more nuanced understanding of crash dynamics.

### *D. Predictive Analytics*

Beyond understanding past patterns, recent research has also ventured into predictive analytics. By harnessing historical crash data and incorporating real-time variables such as weather conditions, traffic volume, and time of day, predictive models can forecast potential crash hotspots or periods of heightened risk. Such insights can be invaluable

for preemptive safety interventions.

### E. Policy Implications

The insights derived from crash data analysis have profound implications for policymaking. By understanding the root causes and contributing factors of severe crashes, policymakers can formulate targeted interventions, allocate resources more efficiently, and collaborate with stakeholders to implement comprehensive road safety strategies.

### F. Awareness Campaigns

Beyond the realm of policymaking, crash data analysis also plays a pivotal role in shaping public awareness campaigns. By highlighting prevalent causes of severe injuries and fatalities, awareness campaigns can be tailored to address specific behaviors or risk factors, fostering a culture of safety and responsibility among road users.

## III.METHODOLOGY

The methodology section delineates the systematic approach adopted to analyze the crash data, ensuring robustness, accuracy, and reliability in the findings presented. A methodical process is paramount to drawing meaningful insights from complex datasets, especially when the objective is to inform critical safety measures and interventions.

### A. Data Collection and Preliminary Exploration

The study commenced with the collection of comprehensive crash data, sourced from reliable repositories to ensure data integrity and completeness. A preliminary exploration was conducted to familiarize with the dataset's structure, identifying potential challenges and opportunities. This phase also involved an initial assessment of missing values, outliers, and inconsistencies, setting the stage for subsequent data preprocessing steps.

### B. Data Cleaning and Preprocessing

Recognizing the imperatives of data quality, a rigorous cleaning and preprocessing regimen was employed. This involved:

- Addressing missing or erroneous entries through imputation or deletion strategies, ensuring data completeness without compromising its integrity.
- Standardizing data formats, particularly date and time fields, to facilitate uniform analysis.
- Encoding categorical variables, transforming them into a format conducive for analytical modeling, while preserving the intrinsic information they encapsulate.

### C. Feature Selection and Extraction

Given the dataset's complexity, feature selection emerged as a pivotal step to enhance model performance and interpretability. Through a judicious selection process, features deemed irrelevant or redundant were pruned, retaining only those with significant predictive power and relevance to injury severity.

Additionally, feature extraction techniques were employed to derive new variables that could potentially amplify the predictive capabilities of the models. This involved the creation of composite variables and transformations, capturing nuanced relationships and patterns embedded within the data.

### D. Model Development and Evaluation

With the refined dataset at hand, the focus shifted towards model development, leveraging advanced data mining techniques tailored to the problem's intricacies. Decision tree classifiers and ensemble methods, including Random Forests, were employed to model the relationship between the selected features and injury severity, offering both predictive accuracy and interpretability.

Model evaluation was conducted using rigorous validation protocols, encompassing metrics such as accuracy, precision, recall, and F1-score. This holistic assessment ensured that the models not only performed well on the training data but also demonstrated robustness and generalizability on unseen datasets, reinforcing their utility in real-world applications.

### E. Interpretation and Insights Generation

Beyond predictive performance, the models were scrutinized to extract actionable insights, elucidating the factors most influential in determining injury severity. This interpretive phase was pivotal, bridging the analytical findings with pragmatic interventions, thereby transforming data-driven revelations into tangible safety initiatives and awareness campaigns.

## IV.RESULTS (EVALUATIONS)

### A. Decision Tree and Random Forest

The provided Python code conducts a Decision Tree classification analysis on vehicle crash data using the scikit-learn library. First, the feature columns are extracted

from the 'hotdf' DataFrame, and the target variable 'Injury Severity' is obtained from 'finaldf'. The dataset is then divided into training (70%) and testing (30%) sets using the train_test_split function, ensuring reproducibility through a specified random seed (random_state=56):

```
hotAttNames = hotdf.columns.values.tolist()
X = hotdf[hotAttNames]  # Features
y = finaldf['Injury Severity']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3,
random_state=56)
```

Subsequently, a Decision Tree classifier is instantiated with the DecisionTreeClassifier class, and the model is trained on the training set.

```
clf = DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)
```

Predictions are then generated for the test set using the trained classifier, and the accuracy of the model is assessed by comparing the predicted values (y_pred) with the actual values from the test set (y_test) using the accuracy_score function from the metrics module.

```
y_pred = clf.predict(X_test)
print("Accuracy:",
metrics.accuracy_score(y_test, y_pred))
```

Building upon this foundation, the implementation of a Decision Tree classifier has offered valuable insights into predicting injury severity within the context of vehicle crash data. However, as data science endeavors often involve an iterative process of refinement, there exists an opportunity to enhance predictive performance further. A seamless progression involves transitioning from a single Decision Tree classifier to a Random Forest classifier, a more advanced ensemble learning technique.

The provided Python code establishes a function named `random_forest` designed to implement a Random Forest classifier using the scikit-learn library. This function expects two arguments, `X` and `y`, symbolizing the features and labels of a given dataset. The data is then divided into training and testing sets, allocating 27% of the data for testing while ensuring that the stratification based on labels is maintained through the use of `train_test_split`. Subsequently, the features undergo standardization using `StandardScaler` for both the training and testing sets. A Random Forest classifier is instantiated with 200 trees using `RandomForestClassifier`, and this model is then trained on the standardized training data.

```
model =
RandomForestClassifier(random_state=42,
n_estimators=200)
model.fit(X_train_scaled, y_train)
```

Predictions are made on the standardized test set (`X_test_scaled`), and the model's performance is assessed using various metrics such as accuracy, confusion matrix, and classification report, all of which are facilitated by scikit-learn functions. The results, including the accuracy with two decimal places, the confusion matrix, and the classification report, are then printed to provide a comprehensive overview of the model's performance.

```
print(f"Accuracy: {accuracy:.2f}")
print("\nConfusion Matrix:")
print(conf_matrix)
print("\nClassification Report:")
print(class_report)
```

It is important to note that for the code to run successfully, the actual dataset (`X` and `y`) should be defined before invoking the `random_forest` function. The comment `# Random Forest` implies that this code snippet is part of a larger script or program where Random Forest is being employed for a specific task.

### *B. Test Size and Accuracy*

The output from the Random Forest classifier evaluation provides a holistic view of the model's performance on a given dataset. The reported accuracy of 79% is indicative of the ratio of correctly predicted instances to the total dataset size.

```
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

The classification report further dissects the model's performance by providing precision, recall, and F1-score for each class.

```
print("\nMacro Average:")
print(f"Precision: {macro_precision:.2%}, Recall:
{macro_recall:.2%}, F1-score: {macro_f1:.2%}")
print("\nWeighted Average:")
```

The macro and weighted averages in the classification report contribute to a comprehensive assessment. The macro average, considering equal weight to each class, reveals precision, recall, and F1-score around 22%, 20%, and 18%, respectively. In contrast, the weighted average, which accounts for class imbalances, yields more representative overall metrics. The weighted precision, recall, and F1-score average at 65%, 79%, and 70%, respectively, providing a balanced evaluation of the model's performance across all classes.

### *C. Evaluation and Reflection*

Upon reviewing the Random Forest classifier, it's pivotal to delve beyond accuracy alone. Assessing precision, recall, and F1-score for individual classes, alongside macro and

weighted averages, offers nuanced insights. This prompts a deeper exploration of specific misclassification patterns and the context of acceptable performance levels.

Considering the achieved accuracy urges a more comprehensive understanding of why certain instances were misclassified. Examining performance in different classes provides valuable cues for improvement tailored to specific areas. Contextualizing the application's needs against achieved metrics is crucial for setting realistic performance expectations.

In terms of trade-offs, balancing precision and recall could be pivotal based on the application's requirements. Potential enhancements span diverse avenues like feature engineering, hyperparameter fine-tuning, tackling class imbalances, and exploring ensemble methods. This detailed assessment, encompassing various metrics and avenues for improvement, is pivotal for honing the Random Forest model's efficacy in its unique application context.

## V.CONCLUSION

The culmination of this analysis offers a multifaceted perspective on crash data, emphasizing the profound implications for road safety, policy formulation, and community awareness. The insights gleaned from this study serve as a clarion call, highlighting both the challenges and opportunities in mitigating the severity of injuries resulting from vehicular incidents.

### *A. Key Findings and Implications*

The analysis elucidated several pivotal factors contributing to injury severity, ranging from environmental conditions and collision types to driver behavior and temporal influences. These findings not only corroborate previous research but also introduce novel insights, underscoring the multifactorial nature of crash dynamics.

The implications of these findings are manifold. For policymakers and safety advocates, the identified factors offer a roadmap for targeted interventions, enabling the design of tailored safety measures and educational campaigns. By understanding the nuanced interplay between various variables, stakeholders can prioritize initiatives that promise maximal impact, fostering a holistic approach to road safety.

### *B. Limitations and Future Directions*

While the analysis provides a robust framework for understanding injury severity determinants, it is not without limitations. The reliance on retrospective crash data inherently limits the scope of causal inference, necessitating caution in extrapolating findings to broader contexts. Moreover, the absence of real-time data integration and spatial analyses warrants future exploration, offering avenues for more dynamic and geospatially informed insights.

Furthermore, the rapid evolution of automotive technologies and transportation paradigms underscores the need for ongoing research, ensuring that safety interventions remain aligned with contemporary challenges and innovations. Future endeavors could also delve deeper into the socio-economic dimensions of crash dynamics, exploring disparities and vulnerabilities that may transcend traditional analytical boundaries.

### *C. Concluding Remarks*

In conclusion, this analysis serves as a seminal contribution to the discourse on road safety, blending data-driven methodologies with actionable insights to foster a safer, more resilient transportation ecosystem. By unraveling the intricate web of factors influencing injury severity, this study paves the way for a renewed commitment to evidence-based interventions, fostering collaboration, innovation, and community engagement in the pursuit of a shared goal: safeguarding lives on our roads. As we navigate the complexities of modern mobility, the insights derived from this analysis resonate as a testament to the transformative power of data, illuminating pathways towards a future where road safety is not just an aspiration but a tangible reality.

## VI.WORKLOAD DISTRIBUTION

| **Vaidehi Vaishnav** | **Dylan Schroers** |
|---|---|
| Preprocessing | Data Collection |
| Feature Selection | Model Training |
| Report Compilation | |

## VII. INSTRUCTIONS TO RUN THE CODE

1. Ensure you have the required libraries installed: pandas, seaborn, matplotlib, scikit-learn, and imbalanced-learn.
2. Download the dataset ('CrashData.csv') and ensure it's in the same directory as the code.
3. Run the provided code in a Jupyter Notebook or any Python environment.