A Predictive Study of Injury Outcomes in U.S. Fatal Traffic Accidents in 2023

DSAN 5300 Final Project

Xiangzhi Chen, Zenan Wang, Yanan Wu

Georgetown University

1st May, 2025

Table of Contents

I. Introduction

Traffic accidents remain a major public health issue, causing approximately 1.35 million deaths and up to 50 million non-fatal injuries globally each year (WHO, 2018). Serious injuries from fatal crashes often lead to long-term health consequences and significant societal costs, making it essential to better understand when and how these injuries occur. While many studies focus primarily on predicting crash fatalities, the patterns behind injury outcomes within fatal accidents deserve deeper exploration.

In this project, we aimed to investigate the injury side of fatal crashes by conducting a detailed exploratory data analysis (EDA) on a nationwide dataset. We examined how time patterns, spatial distributions, road characteristics, environmental conditions, and accident structures correlate with the number of injuries per incident. Building on these findings, we applied a range of statistical learning techniques: supervised methods including Support Vector Machines (SVM), LDA & QDA, Random Forests and Neural Network with hyperparameter tuning, and unsupervised methods such as correlation analysis, Principal Component Analysis (PCA), K-Means clustering, DBSCAN, and Gaussian Mixture Models (GMM).

Our goal is to uncover meaningful patterns that could help inform better traffic safety strategies, emergency response plans, and preventive interventions aimed at minimizing injury impacts.

II. Literature Review

Understanding the dynamics of injuries in fatal traffic accidents requires insights from both public health research and machine learning methodologies.

Traffic accidents remain a significant public health challenge worldwide, and even in high-income countries like the United States, they continue to cause considerable mortality and injury burdens (WHO, 2018). This persistent impact emphasizes the need for deeper exploration into the mechanisms of injury outcomes within fatal crashes, particularly in the U.S. context.

While many studies have historically focused on crash fatalities, some research has begun addressing injury-related outcomes. Noland (2003) emphasized that non-fatal injuries in vehicle crashes impose significant hidden costs on society, such as long-term healthcare burdens and loss of productivity. Rush hour patterns were identified by Adeyemi et al. (2021) as critical windows where injury risks sharply increase due to traffic congestion and driving behavior changes.

Beyond descriptive patterns, the application of machine learning techniques has gained traction in traffic accident research. Mohamed et al. (2013) demonstrated how unsupervised clustering methods, such as K-means, can uncover hidden structures within accident data, offering new angles to categorize and interpret crash patterns without relying on predefined labels. This inspires the use of clustering and dimensionality reduction techniques like K-Means and PCA in our project to better understand accident severity dynamics.

Meanwhile, AlMamlook et al. (2019) showed that supervised machine learning models, particularly Random Forest and Boosted Trees, outperform traditional logistic regression in predicting injury severity outcomes. Their findings emphasize the value of adopting ensemble methods not only for higher predictive accuracy but also for identifying influential factors contributing to injury risks.

Overall, prior literature supports two major points:

(1) Injury outcomes warrant independent exploration beyond fatalities, given their distinct patterns and societal impacts.

(2) Integrating traditional analytical approaches with advanced machine learning techniques offers a promising path toward uncovering deeper risk factors.

Building on these foundations, our project conducts a comprehensive exploratory data analysis and applies a combination of supervised and unsupervised learning methods to better understand injury patterns within fatal crash events.

III. Data Source and Preparation

Our analysis is based on the 2023 Fatality Analysis Reporting System (FARS) dataset, which compiles detailed records of all fatal traffic accidents across the United States. The FARS database provides rich contextual, geographic, environmental, and participant-level information for each crash.

To prepare the data for analysis, we first filtered the accident dataset to retain only fields relevant to accident context, location, vehicle and person counts, and road and environmental characteristics. Specifically, we kept the following variables:

- DAY_WEEK – Day of the week
- HOUR – Hour of crash
- STATE, COUNTY, LATITUDE, LONGITUD – Crash location
- ROUTE, FUNC_SYS, TYP_INT, REL_ROAD – Road design and type
- LGT_COND, WEATHER – Lighting and weather conditions
- PEDS, VE_TOTAL, PERSONS, FATALS – Basic structural indicators
- ST_CASE – Case ID (for merging purposes)

Special placeholder values indicating unknown or invalid entries (e.g., HOUR=99, COUNTY=999, LATITUDE=77.7777) were removed to ensure data quality and consistency.

Since the original accident records did not explicitly provide an injury count, we constructed a new variable, **NUM_INJURED**, by aggregating injury severities from the related person dataset. Specifically, we counted individuals involved in each case (ST_CASE) whose injury severity was classified as greater than minor. The result was then merged back to the main accident dataset, and missing values were treated as zero to reflect no recorded injuries.

The final cleaned dataset's nationwide coverage and structured fields made it highly suitable for both descriptive analysis and machine learning applications.

IV. Exploratory Data Analysis

We started by exploring the dataset to understand how different factors are associated with the number of injuries in fatal traffic accidents. Our EDA is organized into five parts: time, space, road characteristics, environment conditions, and accident structures.

a. Time Patterns

We first looked into how injuries vary across weekdays and hours.

For the weekday pattern (Figure 4.1), total injuries peaked on weekends, especially Saturday, while Tuesday showed the lowest counts. When focusing on average injuries per case, weekends remained high but the gap between days narrowed. The pattern across the week was roughly "high at both ends, low in the middle," suggesting that weekends pose greater risks both in volume and severity.

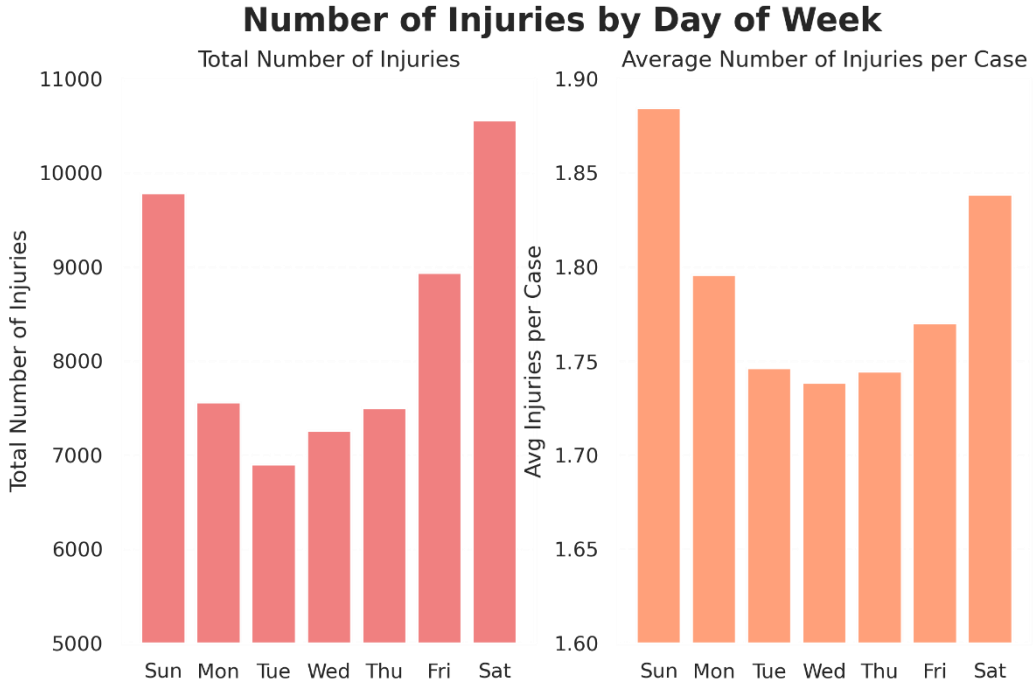

Figure 4.1

For the hourly pattern (Figure 4.2), total injuries followed a strong curve: relatively stable overnight, dropping slightly during early morning hours, then surging after noon and peaking in the late afternoon to evening before declining again at night. Average

injuries per case showed a clearer day-night trend. Average injuries rose steadily from early morning, peaked during afternoon hours, and fell again after evening, suggesting daytime traffic has a higher per-case injury risk.

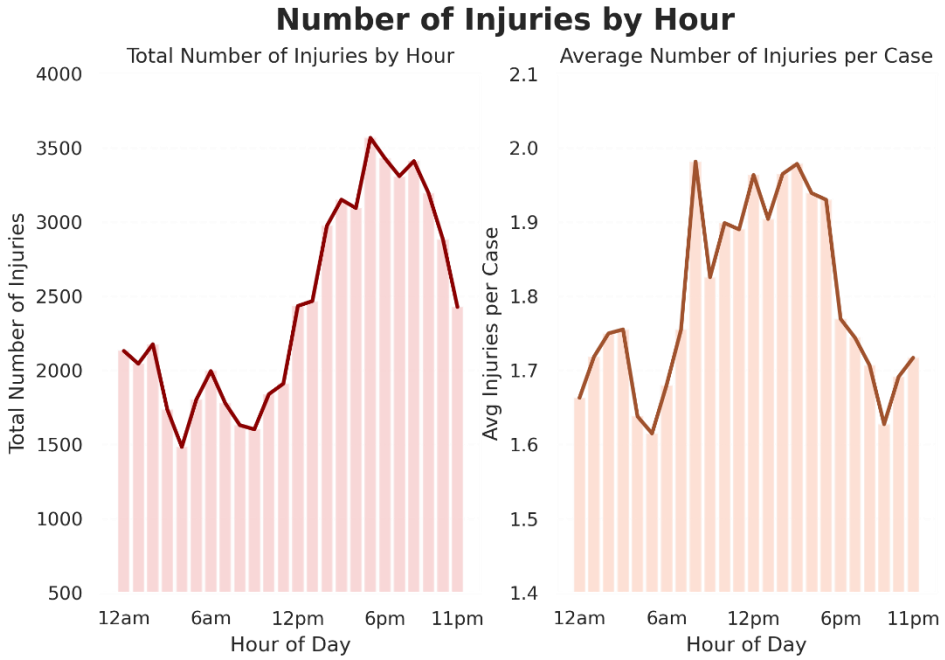

Figure 4.2

For a more detailed understanding, we built a heatmap (Figure 4.3) combining both day of week and hour of day. The results showed clear patterns: Monday and Tuesday mornings around 8 a.m. had distinct spikes in average injuries per fatal accident, reflecting weekday commuting pressure. In general, average injuries rose during daytime and dropped overnight across all days. Weekends had slightly more spread-out risks throughout the day, while weekdays showed sharper peaks during working hours.

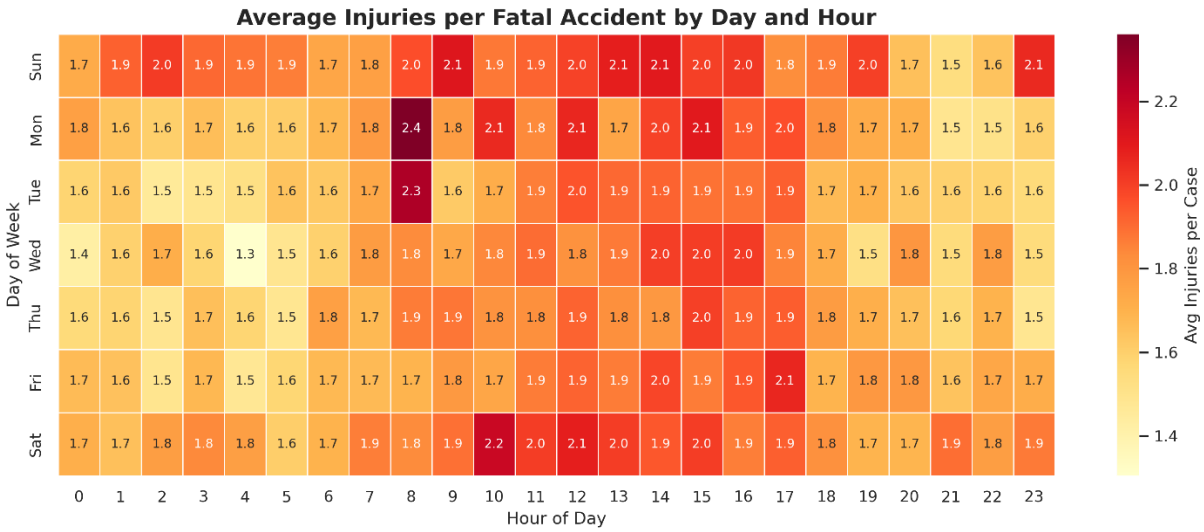

Figure 4.3

This view highlights how injury risks are not evenly distributed, but cluster around specific times depending on both the day and driving behavior.

b. Spatial Distribution

Next, we explored the geographic distribution of accidents.

The scatter plot of accident locations (Figure 4.4) showed clear density patterns across the U.S. Eastern and Southeastern regions appeared significantly more concentrated, especially around states like Florida and Georgia. Urban clusters were clearly visible, though the scatterplot only reflected accident locations, not injury counts.

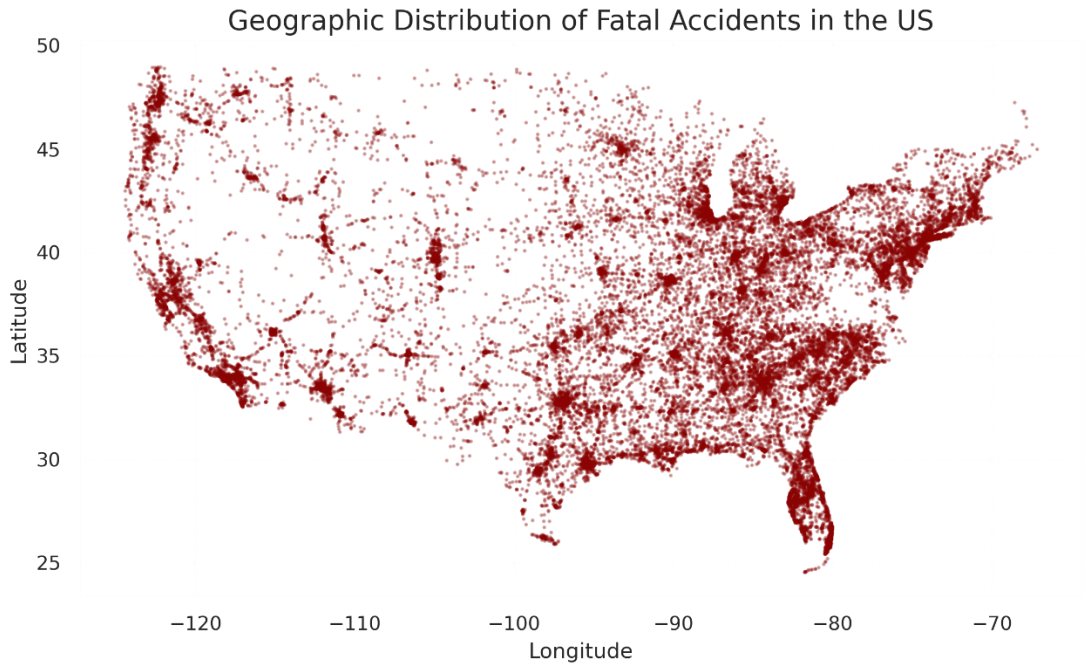

Figure 4.4

To better capture injury burden, we aggregated total injuries by state and presented the results in a choropleth map (Figure 4.5). Texas, Florida, and California stood out with the highest numbers, aligning with their population sizes and traffic volumes. To sharpen the comparison, we also plotted the top 10 states separately. The top three—Texas, Florida, and California—had significantly higher injury counts than the others. North Carolina and Georgia followed but at a visibly lower level. From sixth place

onward, differences between states became much smaller, suggesting that injury burden is heavily concentrated in just a few major regions.

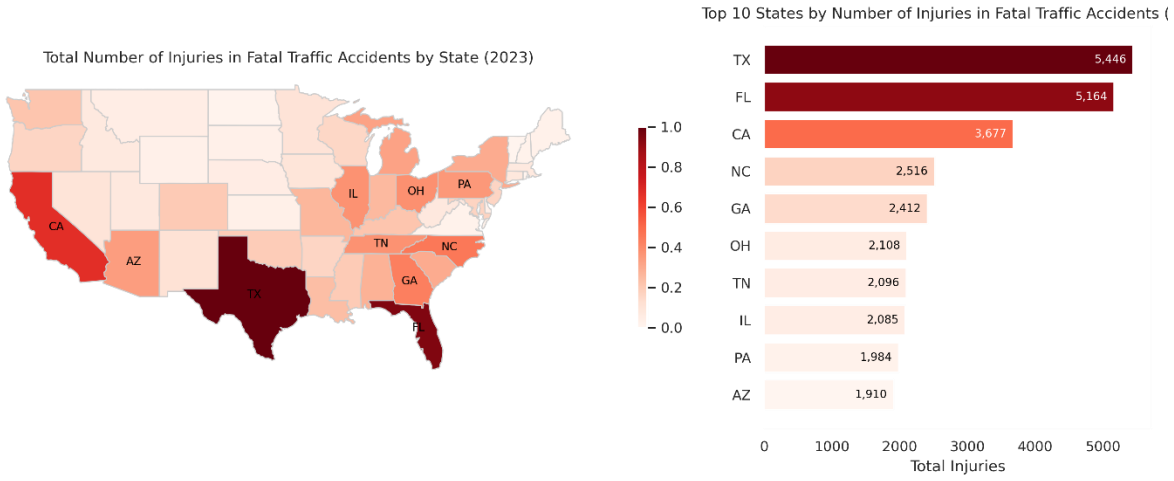

Figure 4.5

c. Road Characteristics

Next, we examined the influence of road characteristics on injury counts, focusing on intersection types and road relations.

From the intersection analysis (Figure 4.6), most injuries actually happened "Not at an Intersection," which makes sense given the vast majority of road segments are not at intersections. However, within intersection-related cases, "Four-Way Intersections" and "T-Intersections" were the most prominent. This suggests that more frequent intersection points may introduce extra risks.

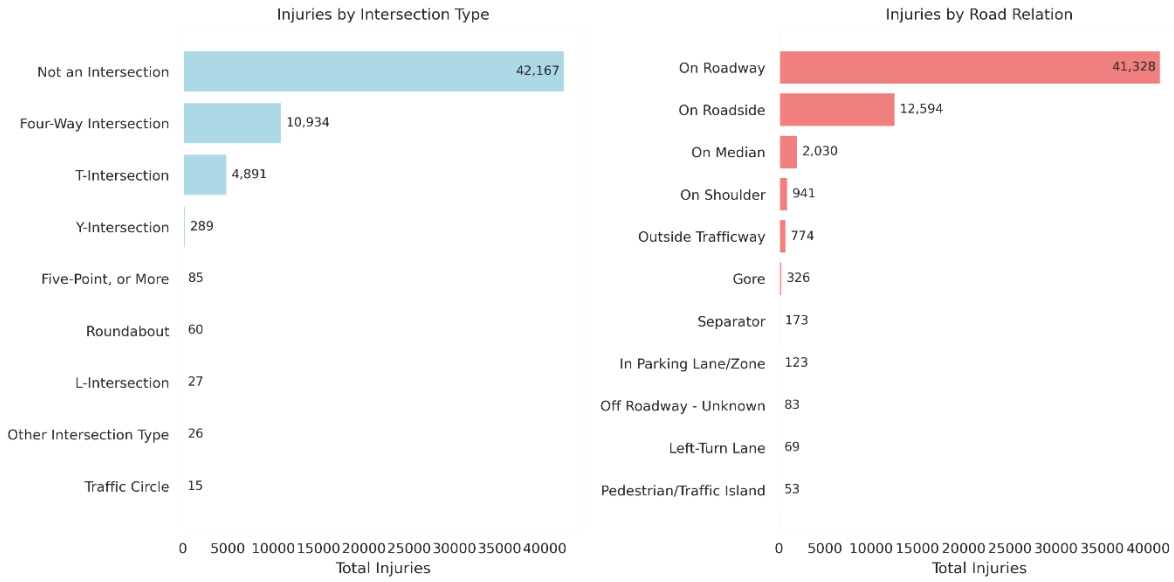

Figure 4.6

Looking at road relation, injuries overwhelmingly occurred "On Roadway," far exceeding other categories. But a closer look reveals that injuries on "Roadside" and "Median" areas also contributed non-trivial counts. These may reflect scenarios like vehicles leaving lanes or emergency maneuvers, hinting at potential secondary risks once drivers leave the main road structure.

d. Environmental Factors

We then examined how lighting and weather conditions impact injury numbers (Figure 4.7).

For lighting, as expected, most injuries happened under "Daylight" conditions. However, "Dark - Not Lighted" cases were notably high, even exceeding "Dark - Lighted" ones. This suggests that when accidents occur at night without sufficient street lighting, they tend to be more dangerous.

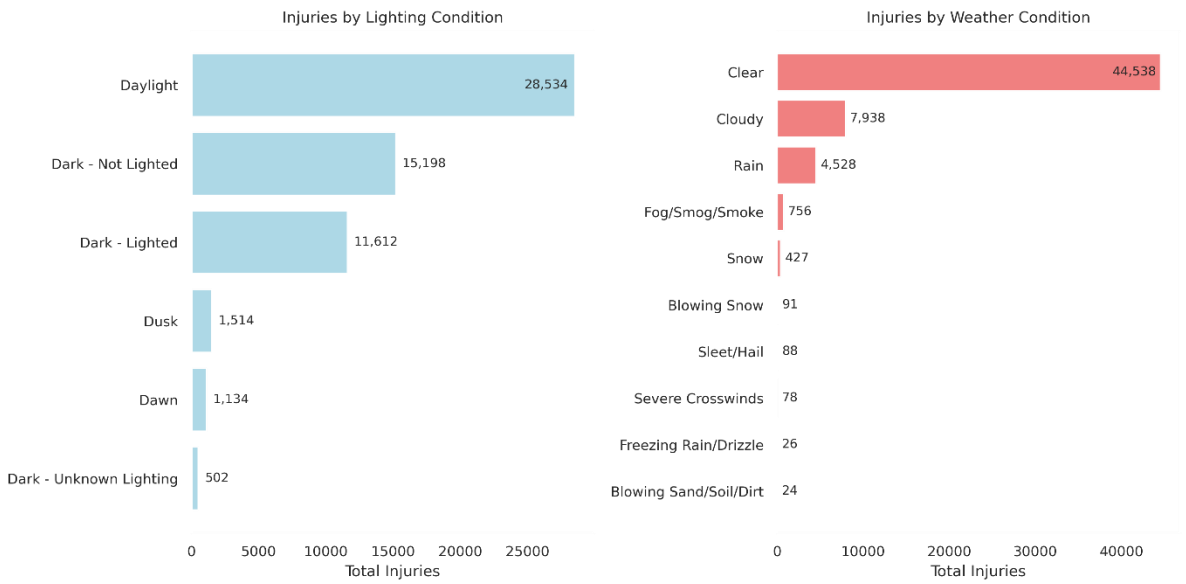

Figure 4.7

For weather, the majority of injuries still happened under "Clear" conditions, simply because most driving occurs on good weather days. Snow and other severe weather types had very few cases, but this may partly reflect lower traffic volumes during extreme conditions.

e. Accident Structural Relations

For accident structures, we first looked at the correlation matrix of several key variables (Figure 4.8). Number of Injuries showed a strong positive correlation with Number of Persons (0.79), which makes intuitive sense: larger accidents tend to involve more people and cause more injuries. Number of Vehicles also correlated moderately (0.43), while Fatalities had a smaller but still notable link (0.38). Pedestrian involvement appeared negatively correlated, suggesting that when pedestrians are involved, injury patterns might differ.

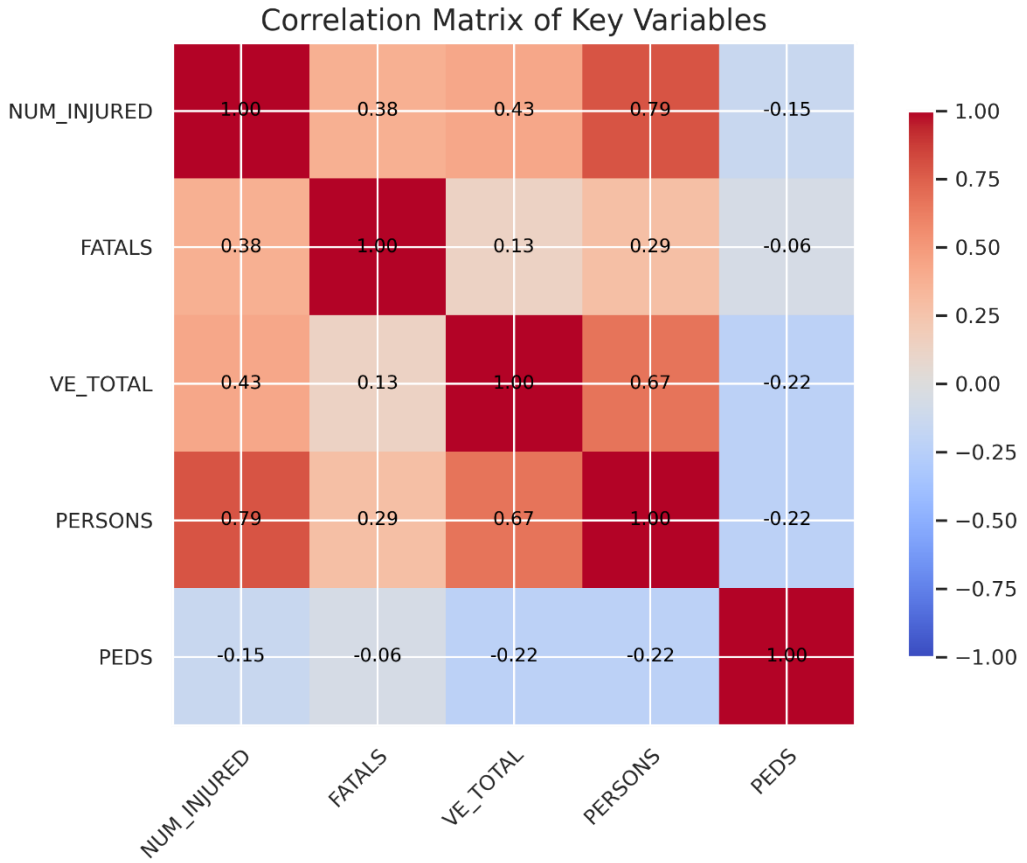

Figure 4.8

To explore further, we compared different severity classes against these structure-related variables (Figure 4.9). In general, as severity level increased (from Low to Medium to High), the median numbers of Persons, Vehicles, and Fatalities rose accordingly. The violin plots clearly showed heavier tails for the High (4+) group,

emphasizing that the most severe accidents are associated with higher counts of people, vehicles, and tragic outcomes.

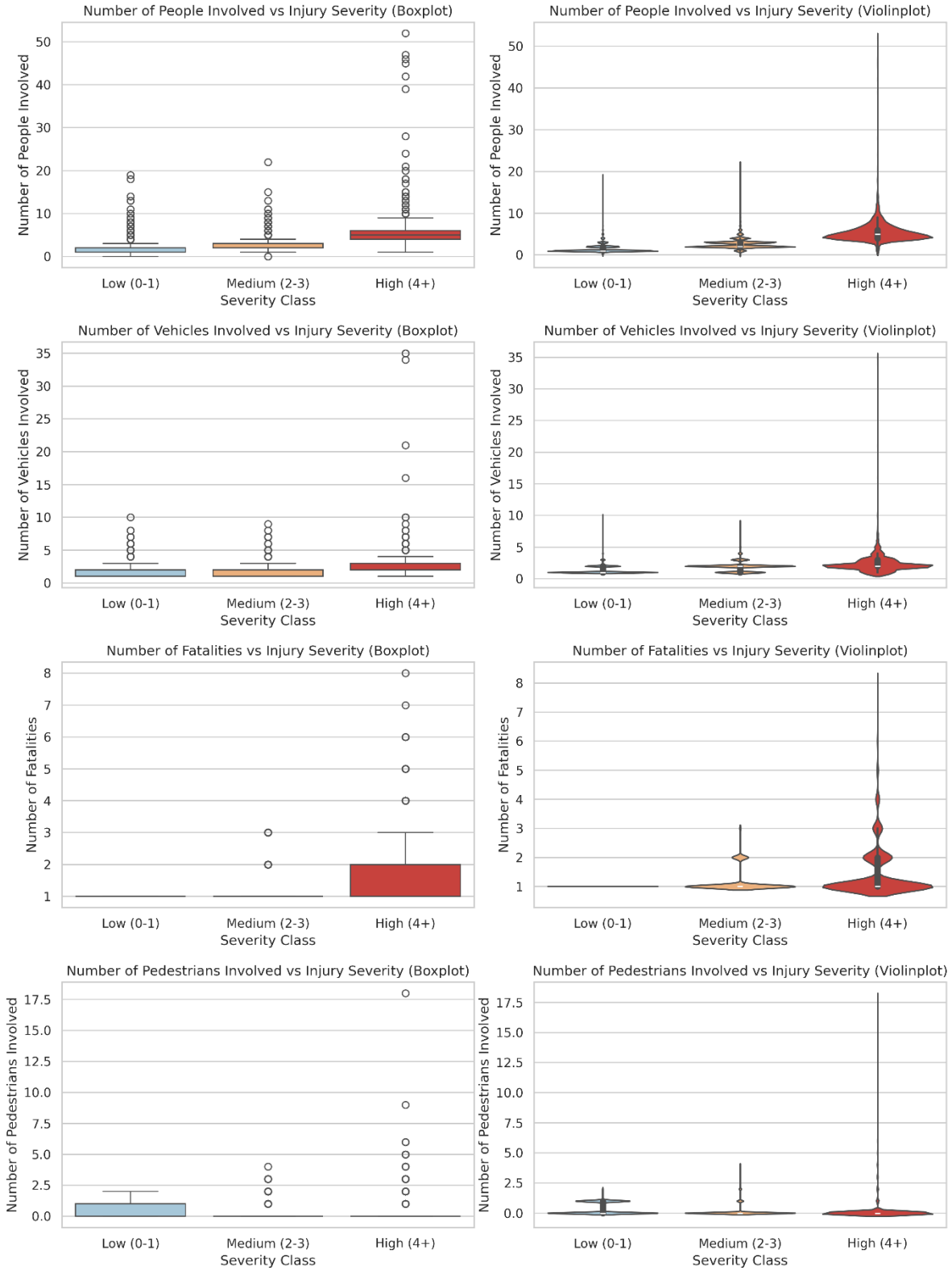

Figure 4.9

f. Summary

Through this multi-angle exploration, our EDA highlighted several key takeaways. Time of day and day of week showed strong patterns tied to commuting and leisure behaviors. Geographic distribution confirmed that injury burdens are heavily concentrated in a few large states. Road environment, lighting, and weather all subtly influenced injury numbers, while accident structure variables like persons and vehicles strongly shaped severity outcomes. These findings not only offered us a deeper understanding of traffic injury dynamics, but also helped guide the feature selection and modeling strategy in the following stages.

V. Unsupervised Learning: Clustering and Feature Selection

This part presents an analysis of accident data using unsupervised learning techniques, specifically clustering. The goal of the analysis is to uncover patterns and relationships in the data that could provide insights into accident characteristics, such as the number of injured individuals. We also applied feature selection to enhance the model's performance and focus on the most relevant predictors.

a. Objective

Building upon the insights uncovered through exploratory data analysis (EDA), we applied unsupervised learning techniques to identify patterns and groupings within fatal traffic accident data. The goal was to uncover hidden structures in the data that could provide valuable insights into the factors influencing injury severity in accidents.

The dataset includes various features such as time, location, weather conditions, number of vehicles and people involved, and crash-related features like lighting conditions and road configurations. We applied clustering and dimensionality reduction techniques to understand the underlying patterns in accident characteristics and group similar incidents together.

For this project, we used clustering methods to uncover natural groupings in the data. We also employed Principal Component Analysis (PCA) for dimensionality reduction to visualize these groups in 2D space and identify potential outliers.

b. Model Training and Evaluation

- **Clustering Methods Applied**

We applied KMeans, DBSCAN, and GMM clustering algorithms to identify underlying patterns in the accident data. The features used for clustering included accident-related variables such as time of day, weather conditions, vehicle count, and location.

1. Optimal Number of Clusters (K)

To determine the optimal number of clusters for KMeans, we used the Elbow Method, which showed that the optimal number of clusters was around 4. This was based on the sharp decrease in inertia followed by a flattening of the curve, indicating that the data could be effectively partitioned into four distinct groups.

The Elbow Method results confirmed the choice of 4 clusters for KMeans clustering, which was then validated by Silhouette Scores and used in further model evaluation.

2. KMeans Clustering

The KMeans algorithm was applied with 4 clusters based on the Elbow Method. This segmentation captured different accident patterns based on features like vehicle count, accident location, and time of day.

3. DBSCAN

DBSCAN was applied to identify dense regions of accidents and flag outliers (noise). The algorithm identified high-density areas with clear accident patterns, and the noise points (accidents that didn’t fit the patterns) were detected.

4. Gaussian Mixture Model (GMM)

GMM was used to model the data probabilistically, where each data point has a probability of belonging to multiple clusters. This allowed for more flexible and overlapping cluster assignments compared to KMeans.

- **Performance Evaluation**

To evaluate the quality of the clusters, the Silhouette Score was computed for each clustering model. This metric evaluates how similar each data point is to its own cluster compared to other clusters. A higher Silhouette Score indicates well-defined clusters.

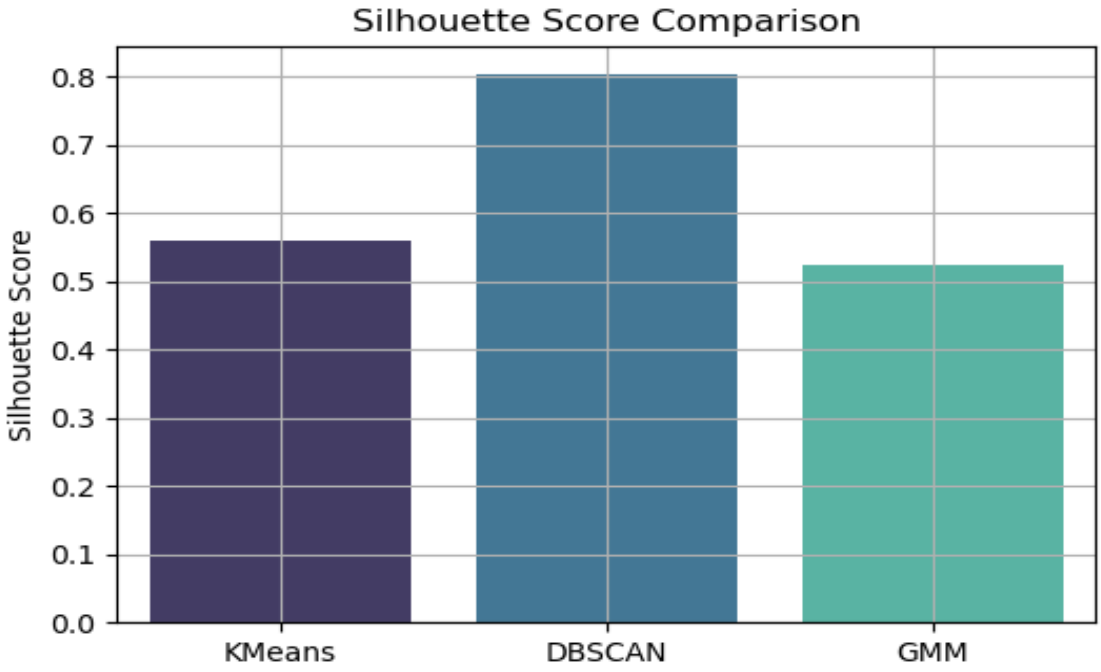

Figure 5.1

DBSCAN achieved the highest Silhouette Score, indicating the best clustering performance in terms of cohesion and separation.

c. Model Insights

- **KMeans Clustering**

The KMeans algorithm clustered the data into 4 distinct groups, and the clusters were visualized using PCA. The resulting clusters captured different accident patterns based on factors like vehicle count, accident location, and time of day.

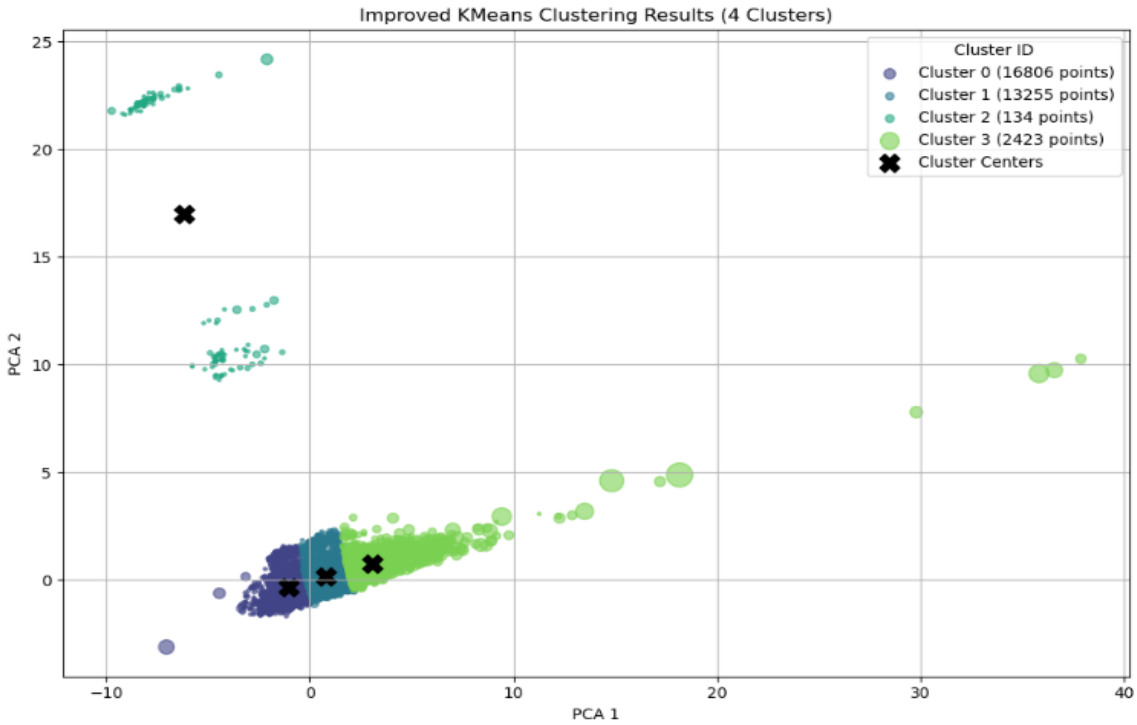

Figure 5.2: Example of Clustering

Visualizations

The KMeans clusters were visualized with PCA, showing how well the clusters were separated in 2D space. The cluster centers were also marked on the plot to identify the centroids of each group.

- **DBSCAN Clustering**

DBSCAN identified dense clusters of accidents and flagged some points as noise. The noise points represent accidents that did not fit into any of the larger clusters, likely due to rare or extreme accident characteristics.

Visualizations

The DBSCAN clusters were visualized with points marked in different colors based on their cluster label. Noise points were shown in a distinct color, which highlighted accidents that deviate from the typical patterns.

- **Gaussian Mixture Model (GMM)**

The GMM clustering visualization highlights four distinct groups in the dataset, with soft boundaries and probabilistic cluster memberships. The ellipses around each

cluster represent the spread of data, while the PCA plot reveals how these clusters are distributed in the 2D feature space. The visualization suggests that the dataset can be effectively segmented into these four groups, which can further be explored for deeper analysis or predictive modeling.

Visualizations

The GMM clustering results were visualized similarly to KMeans, but with the added benefit of showing the probabilistic nature of cluster assignments. Each point's membership to a cluster was represented by varying intensity or color, providing a softer boundary between clusters compared to KMeans.

d. Interpretation and Insights

- **Key Findings**

The clustering analysis revealed four distinct clusters based on accident characteristics. The clusters were derived from the KMeans algorithm after determining the optimal number of clusters using the Elbow Method. The clusters were visualized in PCA space, which helped to identify patterns in how accidents were grouped.

1. KMeans Clustering provided 4 distinct clusters, each representing a unique grouping of accidents based on similarities in the data.
2. DBSCAN identified high-density accident areas and flagged outliers as noise, offering further insights into accident zones that deviate from typical patterns.
3. GMM offered a probabilistic perspective of the clusters, showing the likelihood of each data point belonging to multiple clusters.

e. Limitations

- **Data Limitations**

1. The dataset was limited to fatal accidents only, and the dynamics of these accidents may differ from non-fatal or minor accidents.
2. The classification of injuries into binned categories might introduce noise, especially for boundary cases.

- **Missing Features**

    1. Potentially important features like seatbelt usage, vehicle speed, and driver impairment were not included in the dataset. Including these features could improve model performance and provide more meaningful insights.

f. Summary

The application of unsupervised learning models (KMeans, DBSCAN, and GMM) on fatal traffic accident data provided valuable insights into the patterns that influence injury severity. KMeans, DBSCAN, and GMM all identified meaningful clusters, with DBSCAN outperforming the others in terms of silhouette score.

The results of this analysis can help improve traffic safety measures and emergency response strategies, especially in high-risk regions or accident-prone times.

VI. Supervised Learning: Predicting Injury Counts in Fatal Crashes

a. Set up

Building upon the insights uncovered through exploratory data analysis (EDA) and unsupervised methods, we implemented several supervised learning models to predict the number of injuries in fatal traffic accidents. The objective was to develop a robust model that can identify key predictors of injury severity and provide accurate estimates based on contextual and structural crash variables.

The target variable in our supervised modeling was NUM_INJURED, which represents the number of individuals injured (including fatalities) in each fatal crash. Because NUM_INJURED is a discrete numerical value, we binned it into three major classes based on results from unsupervised machine learning models. For this project, we grouped injury counts into ordered bins (0–1, 2–3, 4+ injuries) to allow for more stable predictions and performance evaluations using accuracy and ROC curves.

We tested several supervised learning models, including:

1. Support Vector Machine (SVM)
2. Linear Discriminant Analysis (LDA)
3. Quadratic Discriminant Analysis (QDA)
4. Random Forest
5. Neural Network (Multi-Layer Perceptron Model)

Each model was trained on the same processed feature set derived from the FARS dataset, including variables related to time, location, number of vehicles and persons involved, lighting conditions, weather, and road configuration. Categorical variables were one-hot encoded, and missing or anomalous values were cleaned as described in the data preparation step. Note that here we employed under-sampling and SMOTE method to balance our input dataset. And we extract 5000 data points for each of the three classes, so in total of 15000 rows for the training set.

b. Model Training and Evaluation

To ensure a fair comparison across models, we split the dataset into training (80%) and testing (20%) subsets using stratified sampling to preserve injury count distributions. We built our model using sk-learn, and tuned the parameters using Grid-Search. We evaluated model performance using accuracy, F1-score, and ROC-AUC, the latter of which measures how well each model distinguishes between injury severity categories.

Models	Accuracy	Weighted F1-score	Macro F1-score
SVM	**0.77**	**0.78**	**0.75**
LDA	**0.72**	**0.73**	**0.71**
QDA	**0.73**	**0.74**	**0.70**
Random Forest	**0.80**	**0.81**	**0.77**
Neural Network	**0.80**	**0.80**	**0.77**

Table 5.1

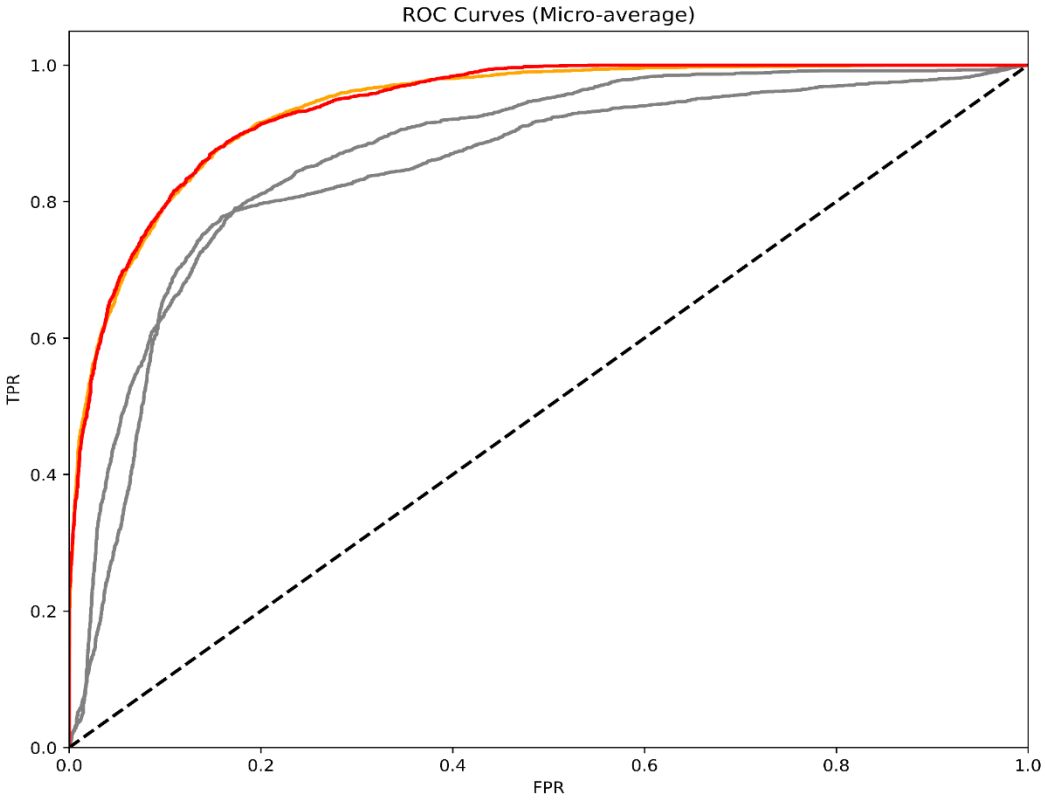

Figure 5.3

Both Random Forest and Neural Network achieved the highest accuracy of 0.80, outperforming traditional linear classifiers like LDA and QDA. The ROC curves (see poster) show that these two models also had superior area under the curve (AUC), indicating stronger classification power across different thresholds.

- **Random Forest**

The Random Forest model, an ensemble of decision trees, demonstrated strong performance due to its robustness against overfitting and ability to handle non-linear interactions among features. Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf were tuned using grid search with cross-validation.

Feature importance analysis from the Random Forest revealed the top predictors of injury counts: **Number of persons involved**, **Number of vehicles**, **Crash location (latitude/longitude)**.

The scale of the crash (vehicles and people involved) was the most influential determinant of injury severity. Interestingly, while environmental factors like weather and road surface had minor influence, location seemed to matter.

- **Neural Network**

Neural networks can capture complex, non-linear relationships in the data, though their interpretability is lower compared to tree-based models. Despite this, they performed exceptionally well, especially in distinguishing higher-injury cases, as reflected in their ROC curve performance.

The Neural Network model, structured as a feedforward multilayer perceptron (MLP), achieved comparable performance to Random Forest. After Grid-Search, we get a complex architecture with four hidden layers (256, 128, 64 and 32 neurons) and Tanh activations. The model was trained using the Adam optimizer with early stopping based on validation loss. We also tried to build upon this MLP to generate our own neural network. We added dropout layers and tuned L1 and L2 regularization, but it didn't result in a significantly better performance. So we just conclude on the MLP.

While both Random Forest and Neural Network models showed excellent accuracy, their use cases differ:

1. Random Forest offers more interpretability and is ideal for policy analysis or identifying risk factors.
2. Neural Networks might excel in real-time applications (e.g., embedded in vehicle systems or dispatch software) due to their flexibility and performance with large-scale inputs.

The consistent performance of both suggests that injury count prediction is feasible with machine learning tools, even with moderate feature engineering.

c. Interpretation and Insights

The predictive performance of our models confirmed that injury counts in fatal crashes are not random but influenced by specific, measurable factors.

Key findings include:

1. Scale matters most: Accidents with more vehicles and passengers are significantly more likely to result in higher injury counts. This seems intuitive, yet confirms the importance of controlling high-occupancy or multi-vehicle situations through safety measures like speed limits and driver monitoring.
2. Location matters: Geographic coordinates (especially certain counties and urban areas) were consistently predictive. This aligns with our EDA findings, suggesting that regional traffic infrastructure and emergency response systems may play a role in injury outcomes.
3. Time of day and lighting were moderate predictors, reinforcing our earlier observation that afternoon and evening crashes tend to result in more injuries.

On the other hand, weather conditions and some road features had little to no predictive value, contrary to common assumptions. This could suggest that most fatal accidents occur in standard conditions, where driver behavior and crash scale become the dominant drivers of injury count.

d. Limitations

Despite encouraging results, several limitations should be acknowledged:

1. The injury classification was based on manually grouped counts, which may introduce noise or obscure boundary cases.
2. The dataset only includes fatal accidents, which may have different dynamics from non-fatal or minor incidents. Thus, our findings are not directly generalizable to all crashes.
3. Some potentially important features (e.g., seatbelt usage, airbag deployment, vehicle speed, or driver impairment) were not included.

VII. Conclusion and Future Work

Our supervised learning models successfully predicted the number of injuries in fatal traffic accidents with high accuracy, achieving up to 0.80 accuracy using both Random Forest and Neural Network classifiers. The most influential predictors were

related to the scale of the accident (vehicles, persons) and geographic location, rather than environmental or road conditions.

These findings challenge some common assumptions — for instance, that weather or time of day are dominant risk factors — and suggest that prevention strategies should focus more on vehicle monitoring, occupant safety, and high-risk locations.

For future work, we recommend:

1. Integrating additional datasets for vehicle-level or driver-level details. Our models often result in lower precision, which means that there are some features that are not fully examined in our analysis.
2. Developing real-time prediction models that could be integrated into emergency response systems or autonomous vehicle algorithms.

VIII. Reference List

1. Adeyemi, O. J., Arif, A. A., & Paul, R. (2021). Exploring the relationship of rush hour period and fatal and non-fatal crash injuries in the US: A systematic review and meta-analysis. *Accident Analysis & Prevention*, *163*, 106462.
2. AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT)* (pp. 272-276). IEEE.
3. GeeksforGeeks. (n.d.). *Gaussian Mixture Model*. https://www.geeksforgeeks.org/gaussian-mixture-model/
4. Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety science*, *54*, 27-37.
5. Noland, R. B. (2003). Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. *Accident Analysis & Prevention*, *35*(4), 599-611.

6. Scikit-learn developers. (2021). *sklearn.metrics.silhouette_score*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

7. World Health Organization. (2019). *Global status report on road safety 2018*. World Health Organization.