KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

THE FATAL MIX: ALCOHOL, SPEED, AND NO SEATBELT

Manu Johnsen – Fall 2025, Charles Lane – Spring 2026

Advisors: Holly Deal, Dr. Nina Grundlingh, Junjun Huo

INTRODUCTION

This project uses data mining to look at U.S. crash fatality data and find important factors that are linked to fatal injuries. The Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA) showed that there were 37,654 fatal crashes and 40,901 deaths in 2023. The goal of this study is to learn more about the behavioral and environmental factors that lead to roadway injuries by looking at age, alcohol use, restraint use and travel speed in relation to injury severity.

METHODS

Feature Selection: ANOVA tests were performed to identify the most influential predictors, keeping only those that were statistically significant and had meaningful application potential.

Correlation Analysis: A correlation bar chart summarized the strength and direction of relationships between key predictors and injury severity.

Data Visualization: Multiple visual tools were used to investigate trends and patterns

Modeling: Several models were tested to determine the best predictive family, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and LinearSVC.

Model Evaluation: Recall, ROC, and Precision-Recall (PR) curves were used to evaluate performance, while confusion matrices measured class-wise prediction accuracy.

Feature Importance: A Random Forest feature importance plot identified which variables contributed most strongly to fatality outcomes.

Injury Status Distribution: The multiclass injury categories were simplified to emphasize the most practical distinctions — No Injury, Minor Injury, Severe Injury and Fatal Injury. Stratified sampling was applied to maintain proportional class balance across training, validation, and test sets, ensuring fair model evaluation.

LINKS

Manu-linkedin

Charles-linkedin

GitHub-Codes

Figure 1: Injury Severity Distribution

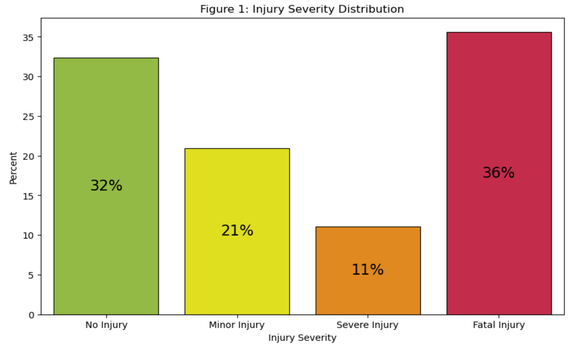


Figure 2: Injury Level Distribution by Alcohol/Seatbelt Group (All Known Values)

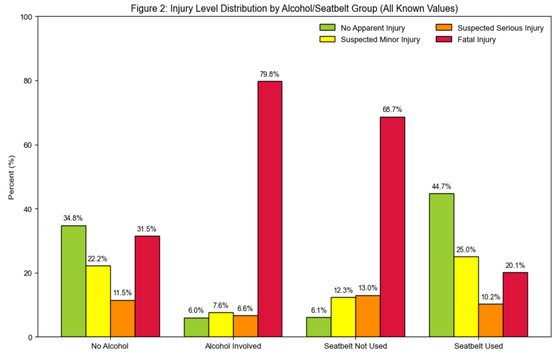


Figure 3: Random Forest Feature Importances

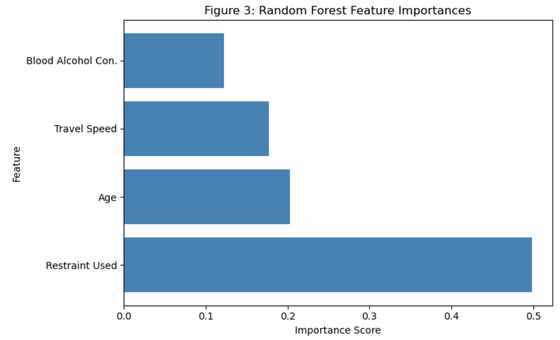


Figure 4: Restraint Used Percentage by Age Groups

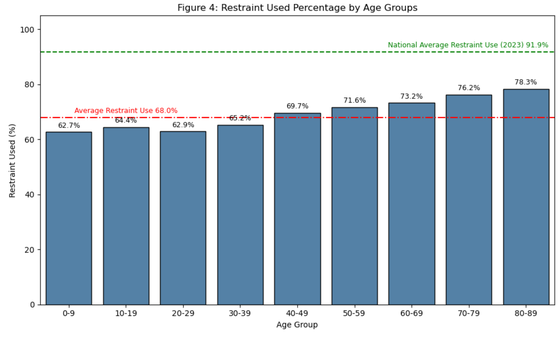


Figure 5: Normalized Confusion Matrix (Test Set)

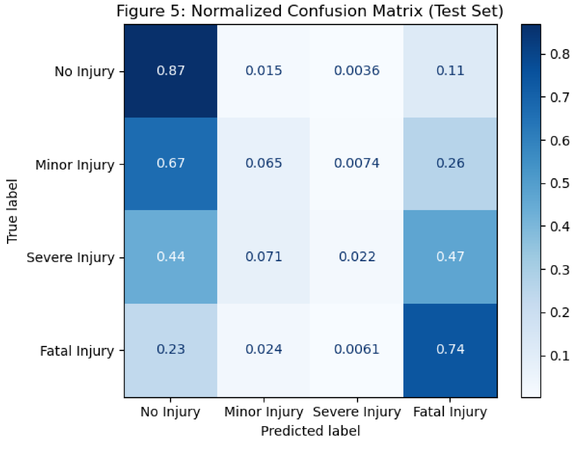


Figure 6: Travel Speed by Injury Severity

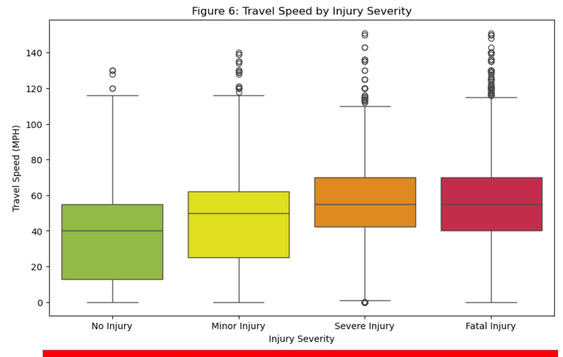


Figure 7: One-vs-Rest Model Evaluation

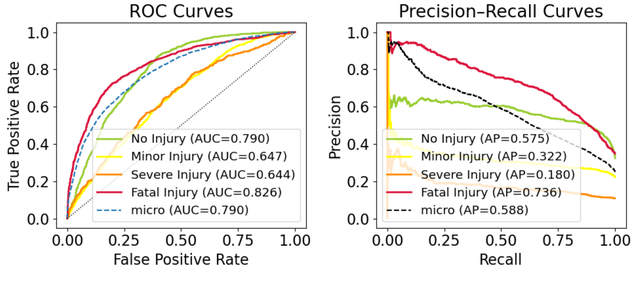
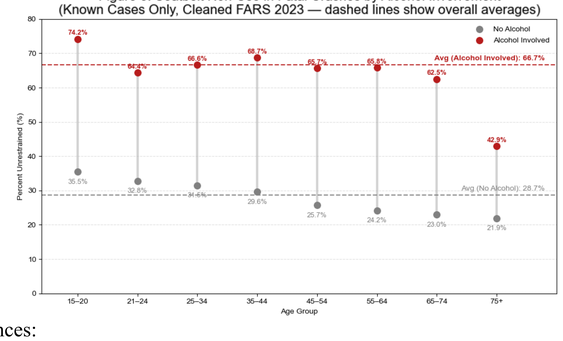


Table 1. Fatality Rates by Alcohol and Seatbelt Use

Behavior Combination	Total Cases	Fatality Rate (%)
Alcohol + No Seatbelt	1611	89.4
Alcohol + Seatbelt Used	804	60.4
No Alcohol + No Seatbelt	7531	64.2
No Alcohol + Seatbelt Used	18587	18.3

Figure 8: Seatbelt Non-Use in Fatal Crashes by Alcohol Involvement
(Known Cases Only, Cleaned FARS 2023 — dashed lines show overall averages)



RESULTS

Random Forest Model

- Overall (Micro-averaged) Metrics
- ROC-AUC: 0.790 — moderate discrimination
- PR-AP: 0.588 — weak average precision
- Recall: 0.554 — weak overall true positives

Class-wise (One-vs-Rest) Performance

- No Injury: ROC-AUC=.790, PR-AP=.575, Recall=.638
- Minor: ROC-AUC=.647, PR-AP=.322, Recall=.065
- Severe: ROC-AUC=.644, PR-AP=.180, Recall=.022
- Fatal: ROC-AUC=.826, PR-AP=.736, Recall=.740

Top Predictor

Restraint Use emerged as the most influential predictor (importance = 0.485), contributing more than twice as strongly as Age (importance = 0.212). This indicates that whether a driver was restrained had the greatest impact on predicting fatality outcomes.

DISCUSSION

Feature Influence: It is notable that Blood Alcohol Content contributed less to the decision-making in the Random Forest model than the other factors. One possibility is the relatively few results that were positive.

Although it might seem like common sense, deaths still occur every year due to variables under our control like wearing a seat belt and driving an appropiate speed. To combat this, more driver training and education could be needed before a potential driver gets their license.

Class Imbalance: Due to No Injury and Fatal Injury being a higher percentage of the data set, the model lost sensitivy to the minority groups. This could be acceptable depending on the application as our primary evaluation metric was Recall for Fatal Injury class. Our reasoning is that prediciting a Fatality and being incorrect is better than predicting No Injury and being incorrect.

Limitations: This dataset is restriced to high-injury crashes, meaning every crash had a minimum of one fatality. Although this gives us insight into this specific group, we cannot make observations about drivers at large.

Future Research: One promising direction would be to include crashes that have Injury Severity data but dont require at least one fatality. Severe Injuries can be just as life changing as Fatal Injuries and could broaden or generalize our insights.

Another direction that overlaps this research is determining the fault of the crash. While this dataset intentionally omits any determination of fault, there are insights to be gained by looking at things like cell phone usage or weather.

References:

National Highway Traffic Safety Administration. (2016, November 14). Fatality Analysis Reporting System (FARS). NHTSA. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars