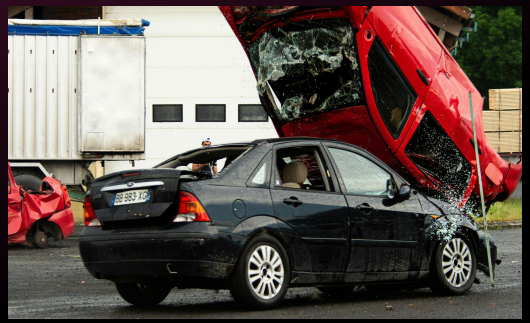
Course: **Data Mining** (CSCI 5502)

Mining Crash Patterns

Spatio-Temporal Analysis to Identify Accident Hotspots, Uncover Risk Factors, and Predict Severity



Professor: **Qin Lv**
TAs: **Mohsena Ashraf, Julia Romero**

THE TEAM

Hima Varshith Reddy Paduru CSCI 5502

Sai Gautham Ghanta CSCI 5502

Venkateswarlu Mopidevi CSCI 5502

INTRODUCTION

Context:
Road traffic accidents are a leading cause of fatalities and injuries, creating a major challenge for public safety and urban planning. The **Fatality Analysis Reporting System** (FARS) provides decades of nationwide fatal crash data (1975 - 2023), offering a strong foundation for data-driven safety research.

Why Spatio-Temporal Analysis?
Crash risk is shaped by spatio-temporal patterns - accidents cluster in specific locations and times, and are influenced by **weather, lighting, road type, and driver behavior**. While traditional statistical methods capture correlations, they often fail to provide localized, predictive, and actionable insights.

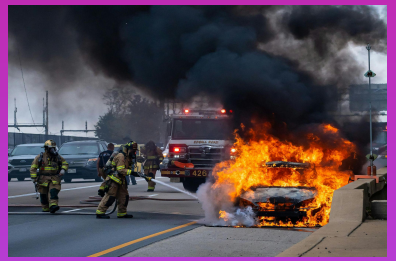
Approach:
This project applies data mining techniques such as **clustering, association rule mining, and predictive modeling** to uncover hidden crash patterns, identify high-risk factors, and move toward severity prediction.

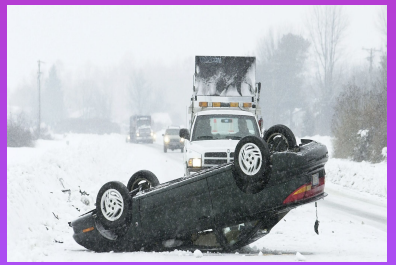
Scope and Purpose:
This project will primarily utilize Colorado crash data from the (FARS) across past 5 decades, with a focus on detecting hotspots and analyzing associated risk factors. Concentrating on Colorado offers a dataset that is both manageable in scope and sufficiently comprehensive to capture long-term crash trends.

Should the findings prove limited in scope, or if project capacity permits, the analysis may be expanded by integrating additional datasets mentioned below, or by extending the study to use the entire US dataset or other selected states. The purpose is to generate insights that are interpretable and actionable, supporting policymakers, emergency responders, and planners in making smarter, safety-focused decisions.

Additional Datasets for Future Integration:

- CRSS to include non-fatal crashes for multi-class severity prediction.
- NOAA climate data for weather-related crash analysis.
- HPMS to incorporate traffic exposure for fair hotspot detection.





RELATED WORK

Wrong-Way Fatal Crashes by Local vs Non-Local Drivers (2024)

→ A recent study used FARS with data-mining and association rules to compare contributing factors in wrong-way crashes by local and non-local drivers. Results showed that non-local drivers and certain roadway features were strongly associated with fatal wrong-way crashes, motivating analysis of driver and vehicle profile interactions in our project.

→ Reference Link: https://www.mdpi.com/2673-7590/4/3/47

Vehicle Age, Safety Tech, and Crash Severity (2025)

→ Zhang et al. analyzed FARS and showed that older vehicles without modern safety/ADAS are linked to higher fatality risk using multivariable logistic regression. This highlights the importance of vehicle characteristics, motivating the inclusion of vehicle age, body type, and safety technologies in crash severity prediction models.

→ Reference Link: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2833617

Daylight Saving Time and Fatal Crash Outcomes (2025)

→ Researchers used multilevel hierarchical modeling on U.S. fatal outcomes to show that crash risks shift with daylight availability, with small but significant decreases during Daylight Saving Time (DST) and variation across regions. These findings support time- and lighting-aware forecasting of high-risk periods.

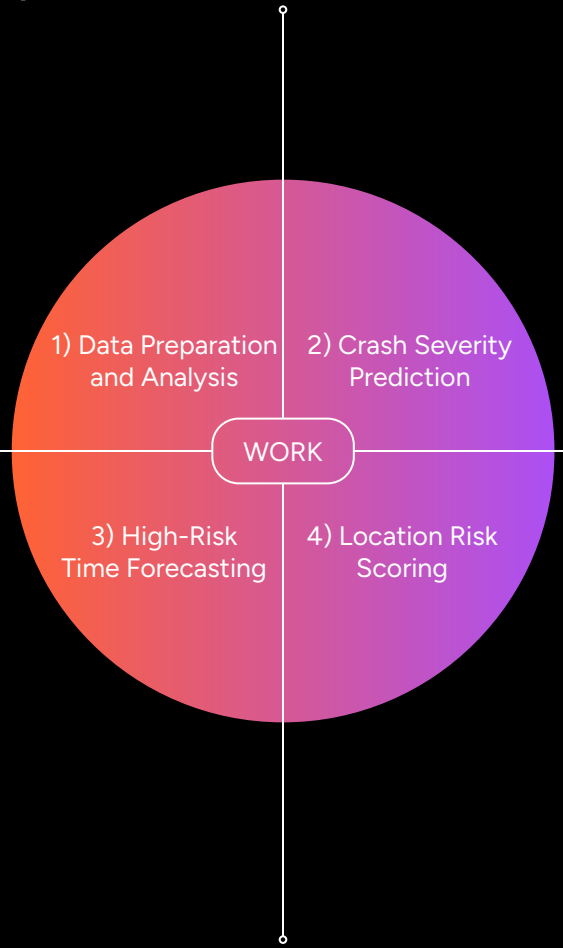
→ Reference Link: https://www.mdpi.com/2673-7590/5/3/102

Fatal Pedestrian Crashes at Intersections via Association Rules (2021)

→ Das et al. applied Apriori association-rule mining to pedestrian crash data and identified interpretable patterns such as night + no crosswalk + arterial road → higher pedestrian deaths. This demonstrates how association rules can extract actionable patterns for safety planning and informs our location risk scoring approach

→ Reference Link: https://www.sciencedirect.com/science/article/abs/pii/S0001457521003377

PROPOSED WORK (PART 1)



- → Use FARS crash data (five decades, nationwide).
- → Core tables: Accident (~30–40k/year), Vehicle (~50–70k/year), Person (~60–90k/year).
- → Clean and standardize codes (weather, lighting, roadway).
- → Verify locations and add features: time-of-day, season, rural/urban, road type, driver/vehicle factors.
- → Optional: enrich with NOAA (weather), CRSS (non-fatal crashes), HPMS (traffic exposure).

- → Build models: Logistic Regression, Random Forest, XGBoost.
- → Predict outcomes: fatal, severe injury, minor injury.
- → Key features: vehicle age, safety tech, driver demographics, restraint use, weather, lighting, road type.
- → Stratify by lighting conditions (daylight, dark with/without streetlights).
- → Use SHAP values and feature importance for interpretability.

- → Apply clustering and time-series methods (K-means, seasonal decomposition).
- → Identify high-risk hours, days, and seasons.
- → Use association rules to find interpretable patterns (e.g., winter nights + snow → higher severity).
- → Support forecasting of dangerous time windows.

- → Detect crash hotspots in Colorado using DBSCAN and KDE.
- → Compute severity-weighted risk scores for intersections and corridors.
- → Tie association rules to locations for interpretable risk insights.

PROPOSED WORK (PART 2)

05 Driver and Vehicle Factor Impact

- Analyze driver factors: age, impairment, distraction.
- Analyze vehicle factors: type (SUV, motorcycle, truck), age, safety features.
- Model interactions (e.g., driver age × vehicle age, lighting × vehicle type).
- Use regression and tree-based models for analysis.
- Present insights with SHAP values and partial dependence plots.

06 Visualization and Dashboard

- Build interactive dashboard to integrate results.
- Hotspot maps with filters (year, weather, roadway).
- Temporal trend charts by hour, day, season.
- Risk factor summaries from rules and models.
- Decision tools for ranking high-risk intersections/corridors.
- Ensure outputs are accessible for policymakers and planners.

EVALUATIONS

Data Splits

- Train on older years, validate on mid-years, and test on recent years.
- Simulates real-world forecasting with past → future prediction.

Crash Severity Prediction

- Metrics: Accuracy, F1-score, ROC-AUC.
- Check feature importance (weather, lighting, vehicle age, etc.).

High-Risk Time Forecasting

- Use clustering quality scores (e.g., silhouette) to validate patterns.
- Compare discovered time patterns with known trends (holidays, night-time risk).

Location Risk Scoring

- Compare predicted hotspots with real crash maps.
- Top-k evaluation: check how many future crashes happen at top-ranked locations.

Driver and Vehicle Factor Impact

- Use regression and tree-based models to measure effects of driver/vehicle factors.
- Provide clear visual explanations (graphs, tables).

Baselines & Practical Validation

- Crash frequency only (ignoring severity).
- Average risk by hour/day without modeling.
- Compare with past traffic safety studies.
- Ensure outputs (maps, graphs, rules) are easy for policymakers and planners to interpret.

PROJECT PLAN AND MILESTONES

Milestone 1

Data Collection and Cleaning

Sep 21 - Oct 4

Week 1:
Gather the FARS data and learn what information is available (crashes, vehicles, people).

Week 2:
Clean the data so it is consistent, remove errors, and prepare it for analysis.

Milestone 2

First Analysis and Hotspot Detection

Oct 5 - Oct 25

Week 3: Combine the data into one set and look at simple statistics (trends by year, crashes by type).

Week 4: Start finding accident "hotspots" - places where crashes happen more often.

Week 5: Look at how crashes change over time (seasons, years) and make first maps to show results.

Checkpoint (Oct 28): Share cleaned data, first findings, and hotspot maps.

Milestone 3

Risk Factors and Predictions

Oct 29 - Nov 23

Week 6: Look for common conditions that lead to crashes (for example, night + bad weather).

Week 7: Study how patterns repeat at certain times (like weekends or holidays).

Week 8: Build simple models to see if we can predict how severe a crash might be.

Week 9: Test the models and see how well they work.

Milestone 4

Results, Dashboard, and Final Report

Nov 24 - Dec 4

Week 10:
Create visuals (maps, charts, summaries) to clearly show the results.

Week 11:
Write the final report and prepare the presentation. Submit the report on Dec 2 and present on Dec 4.

THANK YOU