**Traffic Accident Analysis Using Machine Learning Algorithms**

Jacob Edwards, Atul Prasad, Santosh Kumar  
Shiley-Marcos School of Engineering, University of San Diego  
AAI 500: Probability and Statistics for Artificial Intelligence  
Dr. Dallin Munger  
February 18, 2025

# Abstract

Traffic accidents remain a significant public safety concern, leading to injuries, fatalities, and economic losses. This study utilizes machine learning algorithms, specifically logistic regression and random forest, to analyze and predict accident trends across California, Virginia, New York and Florida. Data preprocessing and exploratory data analysis (EDA) were conducted to identify key accident factors. Our findings indicate that weather conditions, time of day, and road features play crucial roles in accident severity. The results of this study can aid policymakers and traffic safety authorities in implementing effective interventions.

# Background

Traffic accidents have been a persistent issue in the United States, with over 36,000 fatalities reported in 2019 alone, according to the National Highway Traffic Safety Administration (NHTSA). Between 2016 and 2023, accident trends have fluctuated due to various factors, including road infrastructure improvements, changes in driving behavior, and external influences such as weather. California, New York, Virginia, and Florida are among the states with the highest accident rates, making them ideal case studies for machine learning-based analysis. This study aims to leverage data from these states to uncover accident patterns and propose data-driven safety recommendations.

# Data Cleaning & Preprocessing

The raw dataset underwent a rigorous cleaning process to ensure data quality and consistency. The following steps were taken:

1. Handling Missing Values: Columns with more than 5% missing data were removed, while those with minor missing values were imputed or dropped selectively.
2. Removing Redundant Columns: Fields such as 'Nautical\_Twilight', 'Astronomical\_Twilight', and 'Airport\_Code' were eliminated as they did not contribute meaningful insights.
3. Standardizing Time Formats: The 'Start\_Time' and 'End\_Time' columns were converted into datetime format for accurate temporal analysis.
4. Duplicate Removal: Any duplicate entries in the dataset were identified and removed.
5. Feature Engineering: New variables such as 'Hour of Day' and 'Day/Night Classification' were derived to improve the effectiveness of machine learning models.

For our statewide EDA comparison, the code begins with a function, load\_data(file\_id), which is responsible for loading and preprocessing the dataset. This function downloads data from a Google Drive. The dataset is then read into a Pandas DataFrame with minimal memory usage to enhance performance. The Start\_Time and End\_Time columns are converted into datetime format, and any rows where these conversions fail are removed to ensure data consistency. This data preparation step is crucial for ensuring that time-based analyses, such as accident trends by hour and month, operate on clean and standardized data.

# Exploratory Data Analysis

Several functions within the code perform exploratory data analysis (EDA) by visualizing accident trends based on various factors:

1. Severity Distribution: plot\_accidents\_by\_severity(data, ax) generates a bar plot showing the number of accidents for each severity level (Matplotlib Development Team, 2023).
2. City and County Analysis: plot\_accidents\_by\_city(data, ax) and plot\_accidents\_by\_county(data, ax) highlight the top 20 cities and counties with the highest accident counts.
3. Day vs. Night Analysis: plot\_accidents\_by\_category(data, category\_column='City', top\_n=10) introduces a categorical comparison of accidents occurring during the day and night.
4. Time-Based Analysis: plot\_accidents\_by\_hour\_of\_day(data, ax) and plot\_accidents\_by\_month(data, ax) illustrate accident trends across different hours of the day and months of the year.
5. Road Condition Impact: plot\_comparing\_by\_accident\_by\_road\_condition(data, ax) examines the effect of specific road features, such as traffic signals and crossings, on accident frequency.
6. Weather Conditions: plot\_weather\_condition(data, ax) identifies the top 10 most common weather conditions during accidents.
7. Visibility Impact: plot\_accidents\_by\_visibility(data, ax) categorizes accidents based on visibility ranges.

These visualizations provide an initial understanding of how accidents are distributed across time, location, and environmental conditions.

**Key Graphs for Analysis**

The most important graphs for analysis are selected based on their ability to highlight key differences across states and identify critical risk factors.

Accident Trends by Hour of Day (New York, California, Florida)

* This graph shows peak accident times, which appear consistent across states, with a notable increase during morning and evening rush hours. However, the intensity of these peaks varies, with California experiencing a more significant rise in accidents during late afternoon hours.

Accident Trends by Month (New York, California, Florida)

* Seasonal trends reveal that accident numbers increase in certain months. While New York shows higher accident frequencies in winter months, Florida exhibits a more balanced distribution across the year. This highlights potential weather-related influences.

Effect of Road Features on Accidents (New York, California)

* This graph identifies intersections, traffic signals, and crossings as primary accident locations. New York shows a higher impact of traffic signals, whereas California has a significant number of accidents near stop signs.

Top 10 Weather Conditions (Florida, California)

* A graph and chart with green lines

  AI-generated content may be incorrect.Interestingly, most accidents occur in fair weather rather than adverse conditions. This suggests human factors, such as speeding or distraction, play a greater role than poor weather.

# Model Selection & Analysis

To predict accident severity, logistic regression and random forest algorithms were implemented. Logistic regression provides interpretability in understanding variable contributions, while random forest captures complex relationships through ensemble learning. The models were evaluated using accuracy, precision, recall, and F1-score. The random forest model demonstrated higher classification accuracy, while logistic regression allowed better interpretability of individual factors influencing accident severity.

# Conclusion & Recommendations

This study highlights the significant factors contributing to accident severity, including time of day, weather conditions, and road features. Based on these insights, the following recommendations are proposed:

1. Enhanced Traffic Regulations: Increased law enforcement and stricter speed regulations in high-risk areas.

2. Infrastructure Improvements: Installation of smart traffic control systems at junctions and crossings.

3. Public Awareness Campaigns: Educating drivers on the impact of weather conditions and visibility on road safety.

**Analysis Summary**

* The time-based trends show that rush hours are the most dangerous times, emphasizing the need for improved traffic management during these periods.
* The seasonal variation suggests that weather conditions influence accident frequency in northern states more than in warmer regions.
* Road features highlight high-risk areas such as intersections, where better traffic control measures could be implemented.
* The weather impact analysis indicates that accidents are not solely caused by poor conditions, meaning driver behavior is a critical factor that should be addressed through public safety campaigns.

# References

References will include relevant academic papers, datasets, and articles used for the study.

# Appendix

The appendix will contain the Jupyter Notebook outputs from model training and analysis.