**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, 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, 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.

# Exploratory Data Analysis

To identify trends in accident occurrence and severity, multiple EDA visualizations were conducted. The key findings include:

# 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.

# 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.