Project Proposal: Predicting Common Causes of Vehicle Collisions with Causal Analysis of Environmental Factors

- **Group 5**

Motor vehicle collisions are one of the most serious public safety issues worldwide, caused by a mix of behavioral, environmental, and infrastructural factors. Understanding *why* these crashes occur is vital for developing preventive strategies. This project aims to predict the most probable causes of crashes—such as distracted driving, speeding, or mechanical failure—based on contextual data like location, time, and vehicle type. In addition, we will conduct a causal analysis to evaluate how external factors like weather and lighting conditions influence crash frequency and severity.

Unlike prior studies that focus on predicting crash frequency or severity, our project addresses the multi-factor nature of crash causation, treating it as a multi-label classification problem where each incident can have multiple simultaneous contributing causes. The inclusion of causal inference methods adds an analytical depth, allowing us not only to predict what causes crashes but also to quantify how external environmental conditions amplify those risks.

Dataset

We will use the Motor Vehicle Collisions (Crashes) dataset, publicly available from the U.S. Government's open data portal.
Dataset placeholder link: **https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes**

The dataset contains millions of crash records spanning multiple years and includes variables such as date, time, borough, geographic coordinates, number of persons injured or killed, vehicle type, and multiple "contributing factor" fields (e.g., driver inattention, unsafe speed, alcohol involvement, brake failure). It also records lighting and weather conditions, making it ideal for both predictive modeling and causal assessment. The dataset's size and diversity ensure statistical robustness and suitability for graduate-level analysis. For our model, we will be using data from 2022, 2023 and 2024 to train our model and the data from 2025 for testing.

Methodology

Our approach is divided into two interconnected parts:

1. **Prediction of Common Crash Causes (Primary Task)**
   Each record may list multiple causes. We will model this using multi-label classification, where each crash can belong to several cause categories simultaneously.
    - Preprocessing: Handle missing or duplicate records, encode categorical features, and balance underrepresented cause labels using oversampling or SMOTE.
    - Feature Engineering: Include temporal features (hour, weekday, rush hour flag), spatial clustering, and vehicle/driver attributes.
    - Algorithms:
        - Gradient Boosted Trees (XGBoost, LightGBM) for structured tabular prediction.
        - Classifier Chains to model interdependence between multiple causes.

        - Multi-Output Neural Network with sigmoid activations for joint cause prediction.
    - Evaluation Metrics: Hamming Loss, Macro and Micro F1-Scores, Subset Accuracy, and per-label ROC-AUC.

2. **Causal Analysis of Environmental Effects (Secondary Task)**
   To analyze how lighting and weather impact crash likelihood, we will perform causal inference using:
    - Propensity Score Matching (PSM) to create comparable treated (e.g., poor lighting) and control (e.g., daylight) groups.
    - Causal Forests to estimate heterogeneous treatment effects.
    - Difference-in-Differences (DiD) where temporal data permits, to validate results across time.
      This analysis will quantify how external factors influence crash frequency and severity, complementing the predictive results.
3. In addition to predictive modeling and causal analysis, we will incorporate **Explainable AI (XAI)** techniques—specifically **LIME (Local Interpretable Model-Agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)**—to interpret why our model predicts certain causes for each collision. These methods will help identify which environmental or behavioral features (e.g., lighting, weather, or driver factors) most strongly influence the model's decisions. By visualizing feature importance at both global and local levels, we will make our predictions transparent and trustworthy, ensuring stakeholders can understand not just *what* the model predicts, but *why*. This interpretability layer will strengthen the project's practical value by providing actionable, evidence-backed insights for policy and infrastructure improvements.

Expected Outcomes

We expect to build a robust multi-label classification model capable of accurately identifying the most likely causes behind vehicle crashes. The causal component will produce interpretable metrics quantifying the additional risk imposed by factors such as darkness or adverse weather. Together, these results will provide actionable insights for transportation authorities—such as identifying high-risk behaviors under specific conditions or guiding improvements in infrastructure and driver awareness programs.

Evaluation will rely on predictive accuracy metrics for the classification model and Average Treatment Effect (ATE) estimates with confidence intervals for causal analysis.

Feasibility and Significance

The project is feasible within the quarter timeframe, leveraging a well-structured and extensive dataset and applying advanced supervised and causal algorithms. The dual structure—predictive modeling of crash causes combined with causal interpretation of environmental impacts—ensures that the project meets the expectations of a graduate-level data science course in both scope and sophistication. The outcome will represent a practical and research-relevant contribution to the field of road safety analytics.