**Project Proposal**

**Course Code: 5340**

**Course Title: Discovery and Learning with Big Data**

**Submitted By:**

**Group-9**

**Gauri Mahalle**

**Venkata Krishna Garipalli**

**Manjusha Pohar**

**Roshith Nalmas**

**Shivani Podila**

**Submitted To:**

**Prof. Orhan Zeynep**

**Date of Submission: 18th November 2024**

**Project Name**: Analyzing and Predicting Traffic Violations in Montgomery County

**Purpose and Story:**

This project aims to gain insight into traffic violations that occur within Montgomery County to address road safety issues, identify the most effective measures to use in addressing violations, and inform policy-making. The goals of the project are to investigate temporal and spatial violation patterns, forecast outcomes from the contextual attributes and indicate the areas or time frames of the highest risk. It exists to help local authorities make more effective decisions as to where resources are best allocated and to support the targeting of awareness campaigns. Furthermore, the project aims to predict further violations and relate them to environmental or infrastructural aspects for a safer traffic management system that will utilize data.

Below is the information regarding the dataset

* **Dataset Name** – Traffic Violations
* **Source –** <https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q/about_data>
* **Size of dataset** – 1970000 (Rows) 43(Columns)

**Dataset Summary:**

* SeqID: Unique identifier for each traffic violation record.
* Date Of Stop: Date when the traffic stop occurred.
* Time Of Stop: Time when the traffic stop occurred.
* Agency: Law enforcement agency responsible for the stop.
* SubAgency: Subdivision of the agency responsible for the stop.
* Description: Description of the traffic violation committed.
* Location: Location of the traffic stop.
* Latitude: Latitude coordinates of the stop location.
* Longitude: Longitude coordinates of the stop location.
* Accident: Indicates if the traffic stop was related to an accident.
* Belts: Indicates if seat belts were used during the violation.
* Personal Injury: Indicates if the stop involved personal injury.
* Property Damage: Indicates if property damage occurred during the stop.
* Fatal: Indicates if the incident involved fatalities.
* Commercial License: Indicates if the driver had a commercial license.
* HAZMAT: Indicates if hazardous materials were involved in the stop.
* Commercial Vehicle: Indicates if a commercial vehicle was involved in the stop.
* Alcohol: Indicates if alcohol was involved in the violation.
* Work Zone: Indicates if the violation occurred in a work zone.
* Search Conducted: Indicates if a search was conducted during the stop.
* Search Disposition: Outcome of the search if conducted.
* Search Outcome: Result of the search during the stop.
* Search Reason: Reason for conducting a search during the stop.
* Search Reason For Stop: Explanation for why the stop warranted a search.
* Search Type: Type of search conducted during the stop.
* Search Arrest Reason: Reason for arrest if a search was conducted.
* State: State where the traffic stop occurred.
* Vehicle Type: Type of vehicle involved in the traffic stop.
* Year: Year of manufacture of the vehicle.
* Make: Manufacturer of the vehicle involved in the stop.
* Model: Model of the vehicle involved in the stop.
* Color: Color of the vehicle involved in the stop.
* Charge: Legal charge associated with the violation.
* Article: Legal article under which the charge falls.
* Contributed To Accident: Indicates if the violation contributed to an accident.
* Race: Driver’s race as identified by the officer.
* Gender: Driver’s gender as identified by the officer.
* Driver City: City of the driver’s residence.
* Driver State: State of the driver’s residence.
* DL State: State where the driver’s license was issued.
* Arrest Type: Type of arrest conducted, if applicable.
* Geolocation: Combined latitude and longitude coordinates of the stop.

**Target Variable:**

* Violation Type: Type of violation committed by the driver.

**Data Limitation:**

* Some variables might be missing like the latitude and longitude of violation, or the time at which the stop was made; this just creates a gap in the data set especially when doing spatial/time series analysis.
* It means that the aggregation of violation information in the dataset may contain reporting bias: some districts appear to contribute more violations than others, and some violation types may be under-represented because of the local enforcement policies or data input discrepancies
* Geolocation coordinates might have been slightly imprecise since some records were missing or had zero values in latitude and longitude; this would have a major impact on geospatial analyses and the identification of areas of increased violations.

**Project Objectives and Non-Trivial Questions**

**Objectives :**

* Investigate demographic patterns in violation data to confirm fair enforcement of traffic laws and detect potential biases.
* Evaluate the correlation between violations and factors like accidents, alcohol involvement, and work zones to recognize high-risk conditions.
* Develop predictive models to recast the type of violation based on various input features, including location, time, vehicle type, and driver characteristics.

**Non-Trivial Questions:**

* How does the distribution of violation types vary across different times of day or days of the week?
* Can violation types be used to identify patterns contributing to more severe accident outcomes?
* Do specific violation types tend to occur more frequently in particular geographic locations, such as near schools, work zones, or high-traffic areas?
* What violation types are most frequently associated with searches and arrests, and how do these actions correlate with the type of violation?

**Analytical Tools and Techniques**

* **Exploratory Data Analysis (EDA):** To uncover potential relationships between different violations, locations, and various patterns within the data
* **Machine Learning Models:** Logistic regression, decision trees, and ensemble methods.
* **Model Evaluation:** Using metrics like accuracy, precision, recall
* **Visualization Tools:** Tableau, Power BI for interactive dashboards