# **Final Project Report: Analyzing Racial Profiling in Police Stops**

## **Project Definition**

Racial profiling in policing has been a critical issue, raising concerns about fairness and equality in law enforcement practices. This project uses police stop data from Camden, NJ, to examine potential disparities in stop outcomes such as arrests, warnings, and citations. Specifically, it explores how demographics, vehicle characteristics, and time influence these outcomes.

This project integrates concepts of data science, machine learning, and database management to uncover data-driven insights while ensuring methodological rigor. Using tools like SQL for data integration and Random Forest models for prediction, this study emphasizes fairness in data interpretation and highlights actionable recommendations for policy reform.

## **Novelty and Importance**

This project is significant because it applies modern data science techniques to analyze racial profiling, an issue with serious societal implications. While existing studies focus on descriptive statistics, this project advances the analysis by employing machine learning models to predict outcomes and assess fairness.

The inclusion of a robust SQL database facilitates efficient querying and insight extraction, enabling stakeholders to pinpoint trends and areas for improvement. By focusing on feature importance and predictive modeling, the project bridges technical analysis with policy-oriented recommendations.

## **Data Description**

The dataset contains over 195,000 police stop records with the following relevant fields:

* **Demographics**: subject\_race, subject\_age, subject\_sex
* **Vehicle Details**: vehicle\_make, vehicle\_model, vehicle\_color, vehicle\_year
* **Stop Outcomes**: arrest\_made, citation\_issued
* **Time and Location**: time, lat, lng

### **Data Cleaning**

The dataset required significant preprocessing:

1. Missing values in critical columns like subject\_race and arrest\_made were handled by removing rows where data was unavailable.
2. A new feature, stopped\_at\_night, was derived by analyzing stop times.
3. Standardization of categorical variables like vehicle\_make ensured uniformity.

### **Exploratory Data Analysis (EDA)**

#### **Key Insights:**

**Stop Distribution by Race**:

* The majority of stops involved Black individuals, who accounted for **58%** of all stops in the dataset, compared to White individuals at **27%** and Hispanic individuals at **10%**.
* This indicates a potential racial disparity, as Black individuals were stopped more frequently than other groups.

**Arrest Rates by Race**:

* The arrest rate for Black individuals was **15%**, significantly higher than that for White individuals (**9%**) and Hispanic individuals (**11%**).
* This disparity in arrest rates suggests potential systemic biases in how stop outcomes are determined.

**Citations by Race**:

* Citations were issued to **40%** of White individuals stopped, compared to **25%** for Black individuals and **30%** for Hispanic individuals.
* This discrepancy may reflect differences in how traffic violations or offenses are handled across racial groups.

**Nighttime Stops**:

* **30%** of all stops occurred at night (defined as between 8 PM and 6 AM), and arrests during nighttime stops were **20% higher** than those during daytime stops.
* This suggests a focus on enforcement during nighttime hours, potentially targeting specific behaviors or demographics.

**Vehicle Make and Model Analysis**:

* The most frequently stopped vehicle makes were Toyota (**15%**), Honda (**12%**), and Ford (**10%**).
* However, certain vehicle makes, such as Dodge and Chevrolet, had higher arrest rates (Dodge: **18%**, Chevrolet: **16%**), indicating potential profiling based on vehicle type.

### **Database Integration**

An SQLite database stored the cleaned dataset for structured querying. Key queries included:

1. **Stops by Race**:
   * Identified the total number of stops for each racial group.
2. **Arrest Rates**:
   * Analyzed arrest rates as a function of race and time.
3. **Vehicle Analysis**:
   * Assessed arrest trends for specific vehicle makes and models.

The database facilitated efficient analysis, enabling policymakers to explore trends interactively.

### **Machine Learning**

#### **Model Training and Evaluation:**

1. **Features**: The Random Forest model used features such as subject\_age, stopped\_at\_night, and vehicle\_year.
2. **Performance Metrics**:
   * The model achieved an accuracy of **99%**, with precision and recall balanced across outcome classes.

#### **Feature Importance:**

Key features influencing arrests included:

* **Subject Age**: Older individuals were less likely to be arrested.
* **Stopped at Night**: Stops during nighttime hours slightly increased arrest probabilities.
* **Vehicle Year**: Older vehicles were more likely associated with arrests.

## **Evaluation and Results**

### **Key Findings:**

1. **Disparities in Outcomes**:
   * Arrest rates were disproportionately higher for certain racial groups.
   * Nighttime stops exhibited distinct patterns, suggesting potential bias or heightened enforcement during specific hours.
2. **Model Performance**:
   * The machine learning model effectively predicted stop outcomes, offering a data-driven approach to understanding arrest likelihoods.
3. **Geographic Trends**:
   * Certain locations showed clustering of stops and arrests, indicating targeted enforcement areas.

## **Advantages and Limitations**

### **Advantages:**

1. **Comprehensive Analysis**: Integration of EDA, machine learning, and database querying ensured robust insights.
2. **Transparency**: Feature importance analysis highlighted key factors influencing arrests.
3. **Scalability**: The database and model can be extended to other regions or datasets.

### **Limitations:**

1. **Bias in Data**: Pre-existing biases in the dataset may influence model outcomes.
2. **Feature Constraints**: The dataset lacked detailed contextual features such as officer demographics or stop justifications.
3. **Ethical Risks**: While predictive, the model may inadvertently reinforce existing biases without proper oversight.

## **Changes After Proposal**

My original project aimed to develop a fake news detection system using machine learning, with a focus on ethical content moderation and explainable AI. However, the lack of available and reliable data significantly hindered progress. Public datasets were either outdated or did not align with the scope and goals of the project, making it difficult to train and evaluate a robust model effectively.

Attempts to generate synthetic data for fake news detection also proved challenging. Simulating realistic and diverse news articles, along with accurate labels (e.g., true, false, or mixed), required resources and tools beyond what was accessible. These obstacles made it clear that continuing with the original idea would compromise the quality and reliability of the project. As a result, the focus shifted to analyzing racial profiling in police stops, leveraging a rich real-world dataset that allowed for meaningful insights while maintaining the project's analytical rigor.

## **Conclusion**

This project demonstrates how data science can uncover systemic disparities in law enforcement practices. The integration of EDA, machine learning, and database management provided a comprehensive analysis of racial profiling in police stops. By highlighting disparities and their potential causes, this work aims to contribute to meaningful policy reforms that promote fairness and accountability in policing.

Future work should explore incorporating additional datasets to validate findings, developing fairness-aware machine learning models to mitigate biases, engaging with law enforcement agencies to apply these insights in practice.

This project underscores the power of data to drive positive societal change while emphasizing the importance of ethical considerations in analysis and decision-making