**Adrianna Ndubi**

**Traffic Stops Analysis Report**

**1. Introduction**

Traffic stops are one of the most common interactions between law enforcement and the public. Analyzing traffic stop data provides insights into policing practices, decision-making processes, and potential disparities in law enforcement activities. This report documents an analysis conducted on a traffic stop dataset, leveraging Python and machine learning tools to classify outcomes and identify key predictors. The goal is to derive actionable insights and promote evidence-based decision-making.

**2. Tools and Technologies Used**

The analysis was conducted in Python, leveraging the following tools and libraries:

* Pandas: For data cleaning and manipulation.
* NumPy: For numerical computations.
* Scikit-learn: For preprocessing, model building, and evaluation.
* Matplotlib & Seaborn: For data visualization and exploratory analysis.
* Jupyter Notebook: For an interactive and iterative analysis environment.

**3. Data Preparation**

Steps Taken:

1. Data Cleaning:

* Missing values in key columns were identified and handled using imputation methods where appropriate, ensuring data integrity.
* Irrelevant columns (e.g., identifiers with no analytical relevance) were dropped to focus on predictors influencing outcomes.

2. Feature Transformation:

* Categorical variables were encoded into numerical representations using techniques like one-hot encoding and label encoding.
* Numerical variables were standardized (scaled to a mean of 0 and a standard deviation of to ensure compatibility with machine learning models.

3. Train-Test Split:

* + The dataset was split into training (80%) and testing (20%) subsets to allow unbiased evaluation of model performance.
* Data preparation steps ensure that the dataset is both clean and optimized for machine learning models. Standardization prevents dominance by features with larger ranges, and splitting the dataset allows for rigorous model evaluation.

**4. Modeling Approach**

Models Implemented:

1. Logistic Regression:

* + Logistic regression was selected for its interpretability and simplicity, making it an effective tool for understanding the relationship between predictors and outcomes. interpretable results.

2. Random Forest Classifier:

* + An ensemble method combining multiple Decision Trees, known for its robustness against overfitting and superior generalization ability.

**Model Tuning:**

* + GridSearchCV was used for hyperparameter optimization. Parameters like the number of estimators, maximum depth, and minimum samples per leaf were systematically tuned for the Random Forest model to enhance its performance.

**5. Results and Evaluation**

The Random Forest model outperformed the Logistic regression model in terms of accuracy and other key metrics. Below is a summary of the findings:

Key Metrics:

* + Accuracy: Random Forest achieved a higher accuracy than the Logistic regression.
  + Precision and Recall: The Random Forest model balanced precision (minimizing false positives) and recall (capturing true positives effectively).
  + F1 Score: Indicated the harmonic mean of precision and recall, showcasing the model's overall robustness.

**Visual Insights:**

1. Confusion Matrix: Visualized the model's classification performance, highlighting true positives, true negatives, false positives, and false negatives.

2. Feature Importance Plot: Showed the contribution of individual features to model predictions. Predictors like location, time of day, and driver demographics ranked among the most important.

**6. Limitations and Recommendations**

Limitations:

1. The analysis was limited by the available features, potentially overlooking nuanced patterns.

2. Imbalanced class distributions (e.g., certain outcomes occurring far more frequently than others) may have influenced model performance.

Recommendations:

1. Advanced Feature Engineering:

* + Incorporate domain-specific transformations or generate interaction terms to uncover hidden patterns.

2. Bias Mitigation:

* + Consider implementing fairness-aware algorithms or conducting fairness evaluations to minimize bias in predictions.

**7. Conclusion**

* This analysis demonstrates the potential of machine learning in understanding traffic stop outcomes. The Random Forest model emerged as a robust classifier, offering interpretable insights into key predictors. Future work should focus on enriching the dataset with additional features, exploring advanced modeling techniques, and addressing ethical considerations to ensure fairness and transparency in analysis.