Gun Violence Prediction Using Machine Learning Models (LLMS)

Sarthak Sethi

***Abstract*—This paper tries to explore various machine learning methods that can effectively forecast possible patterns or trends related to gun violence in the United States. In this work, K-Nearest Neighbors, Logistic Regression, Decision Tree, and Support Vector Machine are employed in the forecasting of gun violence incidents based on historical incident data. Accordingly, the pre-processing of data, selection of features, and model evaluation aim at enhancing predictive accuracy and identifying the relevant factors of the trend of gun violence.**

Keywords: Gun violence, machine learning, predictive analytics, KNN, Logistic Regression, Decision Tree, SVM, data preprocessing, feature selection, model evaluation.

I. Introduction

GUN violence in the United States is a critical public health and safety concern. At the end of September 2024, there had been over 385 mass shootings-a little more than one each day. The predictive modeling performed in this space will yield useful insights into possible risk factors and also proactively inform public policy measures toward community safety. This project uses a dataset of U.S. mass shootings through prior years to find patterns and make predictions by training machine learning models on the data; it is hoped that the output might provide a better understanding of what factors-both core and peripheral-lay at the roots of gun violence.

II. Proposed Methodology

This research adopts a multi-phase approach to predict gun violence trends using the following steps:

- **Data Collection:** The dataset, obtained from Kaggle, includes information on U.S. mass shootings up to October 2024, with details like incident IDs, dates, locations, victim counts, and suspect information.
- **Data Preprocessing:** The dataset underwent cleaning to handle missing values and irrelevant columns. The data was split into training and testing sets to ensure reliable model evaluation.
- **Model Selection and Training:** Various machine learning algorithms were chosen for their suitability in classification tasks:
  - **K-Nearest Neighbors (KNN):** Effective for pattern recognition, providing baseline performance.
  - **Logistic Regression:** Useful for binary classification, offering insights into factors influencing gun violence.
  - **Decision Tree:** Provides interpretable decision-making paths for identifying critical features.
  - **Support Vector Machine (SVM):** Useful for complex data relationships, adding robustness to predictions.
- **Evaluation Metrics:** Models were evaluated using accuracy, precision, recall, and F1 score to assess prediction performance.

III. Challenges and Solutions

- **Data imbalance:** can decrease the overall precision of the models; it typically means fewer instances of incidents with higher severities in most situations. Some possible techniques for overcoming this, such as resampling and the generation of synthetic data, are being explored.
- **Data Quality Issues:** So, partial records needed preprocessing in order to improve the data quality for providing better input to the model.
- **Performance Trade-offs:** While the trade-off for high accuracy is consumptive computational resource use, this presents an unending challenge of trade-off between the two. Hyperparameter tuning and cross-validation were used to optimize the performance.

IV. Related Work

Several studies have focused on predictive models related to public safety. Our work adds to this research literature, both in its focus on gun violence and in the range of multiple algorithms compared, as well as the breadth of the dataset for the purpose of prediction accuracy.

V. Results and Analysis

Each model was compared by its performance based on accuracy, precision, recall, and F1 score. Performance comparison was done by the following graphs: KNN, logistic regression, decision tree, and SVM.

Accuracy Comparison: It depicts the accuracy of each model, showing the percentage of correct predictions.

Precision Comparison: Shows the precision of the various models in order to show how well a model is able to minimize its false positives.

Recall Comparison: Conveys the recall for each of the models, something which is very crucial in finding the high-risk incidents.

F1 Score Comparison: Combining the above-mentioned Precision and Recall, it gives an overall performance measure.

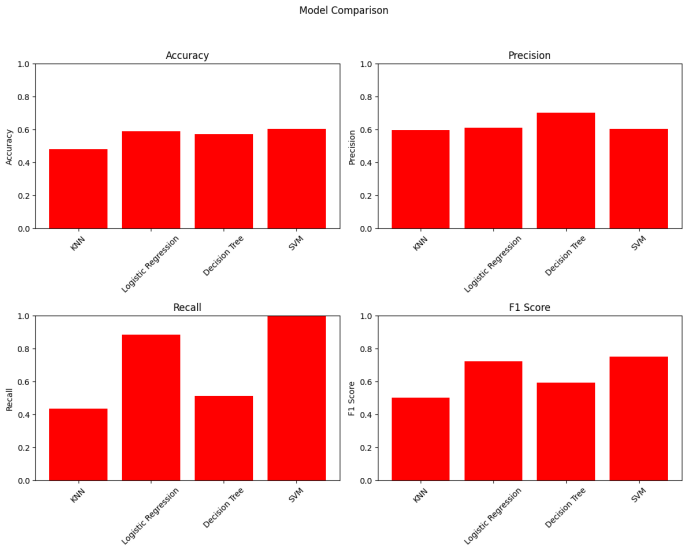

Fig. 1. Graphed out

Initial results show that Logistic Regression and SVM have high recall and are thus appropriate for identifying high-risk patterns. High Precision values in the Decision Tree model support high-confidence predictions. These results will help in choosing the models when applying them to real-life gun violence prediction applications.

VI. Conclusion

This work is a comparative study of different machine learning models in predicting gun violence trends in the United States. Preliminary findings from this study show that although predictive modeling has several limitations, it's worth pursuing for some useful proactive interventions in future. Improvement of preprocessing techniques, search for ensemble methods, and augmentation of the dataset for increased robustness are targeted efforts in the future.