Predictive Analysis of School Shooting Incidents

Rajeshwar Vempaty
Computer and Information Science
Fordham University
New York City, United States
rvempaty@fordham.edu

Samanta Ujjwal
Computer and Information Science
Fordham University
New York City, United States
usamanta@fordham.edu

***Abstract*— This paper presents an in-depth exploration and analysis of a comprehensive dataset on school shooting incidents, with the ultimate goal of developing a predictive model to determine the likelihood of victim occurrence in such events. Leveraging a rich dataset encompassing various aspects of school shootings, including incident specifics, perpetrator details, and school characteristics, we employ advanced data analysis and machine learning techniques to uncover critical patterns and factors. The primary objective is to enhance our understanding of these tragic incidents and contribute to preventive strategies. The analysis reveals significant insights into the correlations between various features and the presence of victims, forming the foundation for a robust predictive model that could be instrumental in guiding policy and security measures in educational institutions.**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. INTRODUCTION

The increasing frequency and impact of school shootings in recent years have necessitated a deeper understanding of these incidents to aid in the development of effective prevention and response strategies. This project is centered around analyzing a detailed school shooting dataset, encompassing a range of variables from incident specifics to perpetrator profiles. The primary aim is to construct a predictive model capable of assessing the likelihood of victim occurrence in school shooting incidents. School shootings pose a significant threat to public safety and the well-being of students and educators. Understanding such incidents' dynamics and contributing factors is crucial for implementing effective preventive measures. This study delves into various aspects of school shootings, including the nature of the incidents, characteristics of the perpetrators, and environmental factors, to extract meaningful patterns and insights.

II. OBJECTIVES

A. Comprehensive Analysis

To conduct an exhaustive exploratory data analysis (EDA) on the school shooting dataset, uncovering critical trends, patterns, and relationships within the data.

B. Predictive Modelling

To develop and validate a machine learning model that accurately predicts the presence of victims in school shooting incidents, based on the insights gained from the EDA.

C. Insights and Recommendations

To provide data-driven insights and recommendations that could inform policymaking and security protocols in educational institutions, aimed at preventing such tragedies.

III. RELATED WORK

A police department can identify offenders most likely to be connected to gun violence in the future by using the VOID (Violent Offender Identification Directive) [10] tool for risk assessment. VOID forecasts 103 people will use guns for violent crimes in 2013 based on a historical dataset of 200,000 past offenders who used firearms by December 2012. Each case chosen at random by the police department's crime experts is given a varied weight by the tool. In terms of prediction accuracy, VOID outperformed the generalized boosted models and optimized logistic regression in identifying individuals involved in gun violence.

An ML model was employed by Heller et al. [16] to forecast the likelihood that a victim will be shot in the near future. The ML model was trained and tested using 644,000 victimization records from the Chicago Police Department. The algorithm had a 13% success rate in predicting victims who were actually shot within the next 18 months out of 500 individuals at the greatest predicted risk.

The mass shooting tragedy in Sandy Hook elementary school (Connecticut, US) in 2012 received unprecedented coverage from both public and social media. Varghese et al. collected over 700,000 tweets from people in the US about this incident [2]. The tweets' dataset was used to train various ML models, including random forest, bagged tree, boosted tree, and support vector machine, to analyze the anti-gun and pro-gun sentiment in different states in the US. It was found that both the anti-gun and pro-gun opinion rates were high on the incident day; however, anti-gun sentiment fell behind after a few days, while the pro-gun sentiment remained elevated for longer.

IV. METHODOLOGY

The methodology of our study is designed to systematically dissect and analyze the multifaceted nature of school shooting incidents. It encompasses a comprehensive workflow that begins with acquiring a detailed dataset and culminates in evaluating predictive models. This process is carefully structured to thoroughly examine the data, allowing for a deep understanding of the various dimensions that characterize school shooting events. The methodology is segmented into distinct

phases: beginning with data sourcing and overview, followed by meticulous data cleaning and preprocessing, an in-depth exploratory data analysis (EDA), judicious feature selection, rigorous model building, and concludes with a detailed evaluation of the model's performance. Each phase of this workflow is crucial, contributing to the overarching aim of developing a robust and insightful predictive model. The subsequent subsections describe the activities and analytical techniques employed in each phase, reflecting our commitment to a methodical and data-driven approach in unraveling the complexities of school shootings.

A. Data Source and Overview

Our machine learning project, focused on predicting the outcomes for victims of school shooting incidents, utilizes data from the "**Public K-12 School Shooting Database**" [1]. This comprehensive dataset, from 1966 to September 2023, includes detailed information on incidents, shooter profiles, victim details, and weapon types. As a publicly available resource, its use is limited to educational and research purposes, and we have adhered to these terms. Necessary permissions for our research were obtained from k12ssdb@gmail.com, and all inquiries about data usage should include the requester's name, affiliation, and intended data application. The K-12 School Shooting Database documents every instance of gun-related incidents in American schools, including firing, brandishing, or hitting school property with bullets. It covers a wide range of incidents, irrespective of the number of victims, time, or reason. The database is developed using a comprehensive methodology that includes merging multiple data sources, detailed research, and reliability scoring. The data includes dates, locations, details of incidents, and is cross-referenced with primary sources for accuracy. The database is designed to provide a broad perspective on school shootings, aiding in detailed analysis and informed decision-making[2].

The database meticulously documents every gun-related incident in American schools, including firing, brandishing, or striking school property with bullets. It encompasses a broad spectrum of incidents, regardless of victim count, timing, or motive. Developed through a robust methodology, it merges various data sources and includes reliability scoring. The dataset is divided into four tabs:

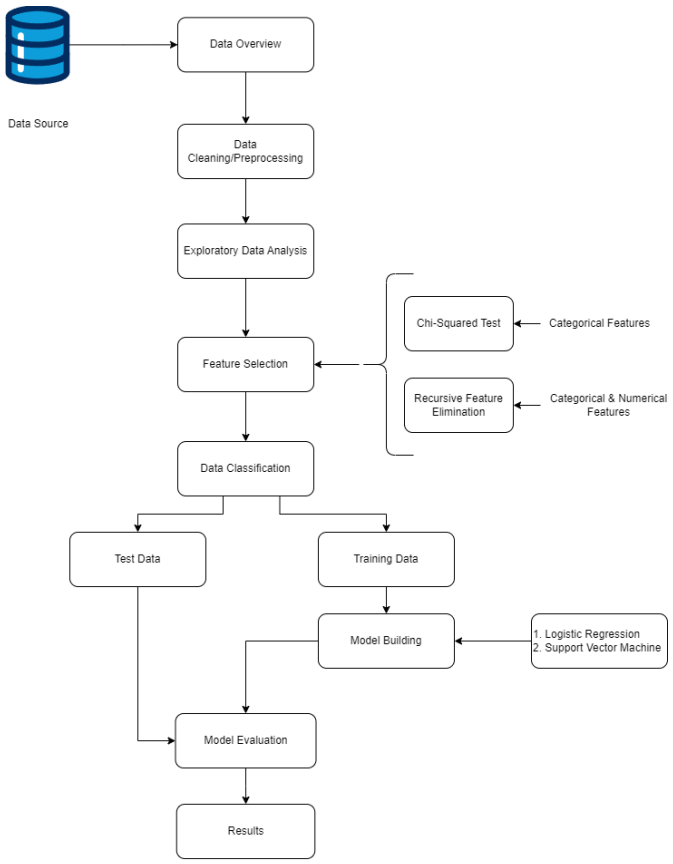
i. Incidents Tab: Each of the **2,514 incidents** is uniquely identified and includes specific dates, school names, and locations. The data covers a wide geographical and temporal span, offering a comprehensive view of the trends over the years.

ii. Shooter Tab: Details about the shooter(s) linked to each incident. This information is vital for understanding the demographics and backgrounds of individuals involved in school shootings.

iii. Victim Tab: Information about the victims, including multiple entries for incidents with several victims. This critical information provides a somber perspective on the impact of these incidents.

iv. Weapon Tab: Data on weapons used or possessed during each incident, which is crucial for analyzing the meaning of these incidents.

Figure 1 Basic Framework of this Paper



B. Data Cleaning and Pre-Processing

In our analysis, the Incident tab is the master sheet, encompassing unique identifiers and comprehensive details for each incident, this sheet also acts as One-Many relationships with other sheets. This tab is pivotal in understanding the context and specifics of each school shooting. The other tabs – Shooter, Victim, and Weapon – provide expanded details about where multiple shooters, victims, or weapons were involved in individual incidents. These tabs are linked to the master sheet through incident identifiers, enabling a holistic view of each event.

To ensure the integrity and reliability of our analysis, we loaded each sheet individually for meticulous examination. Our initial step involved a thorough assessment of missing values across different parameters.

In our study, we meticulously cleaned the 'Shooter' tab from the 'Public K-12 School Shooting Database' to prepare it for machine learning analysis. The process began with renaming columns such as 'Age', 'Gender', and 'Race' to more descriptive titles like 'Shooter_Age', 'Shooter_Gender', and 'Shooter_Race', enhancing clarity. We conducted a thorough examination of the value distribution across these columns, addressing missing and

inconsistent data. The age data was categorized into 'Teen', 'Adult', 'Old_Age', or 'Unknown', while missing gender information was marked as 'Unknown', and 'Transgender' and 'Multiple' entries were reclassified as 'Other'. Racial categories were streamlined by combining several minority groups into 'Other/Minority' and addressing missing data with 'Unknown'. School affiliations were grouped into broader categories, and less common affiliations were labeled as 'Other' or 'Unknown'. The outcomes of the shooters were re-categorized into broader groups such as 'Fled', 'Apprehended/Killed', 'Suicide/Attempted Suicide', etc., with a catch-all 'Other/Unknown' category for ambiguous cases. Missing data in 'Shooter_Outcome', 'Shooter_Died', and 'Shooter_Injury' columns were carefully filled with 'Other/Unknown' and 'No_Injury' as required. Post-cleaning, a reevaluation of the value counts ensured consistency and accuracy. Lastly, we aggregated the shooter data at the incident level, creating a concise summary for each incident, which included the number of shooters and a consolidated view of their profiles.

Following the shooter data preparation, we applied a detailed data cleaning methodology to the 'Victim' sheet of the dataset. The process started by renaming columns for better specificity, such as 'Injury' to 'Victim_Injury', 'Gender' to 'Victim_Gender', and so on. We standardized the 'Victim_Injury' column by replacing 'None' with 'No_Injury' for consistency. The 'Victim_Gender' was categorized into 'Male', 'Female', or 'Unknown', to accommodate all gender data accurately. School affiliations of victims were grouped into broader categories like 'Student/Former_Student', 'Staff/Former_Staff', and 'Family/Intimate', among others, to streamline this aspect of the data. In the case of missing or unclear affiliations, 'Unknown' was used as a default category. Age categorization followed a similar approach with the shooters, with ages classified as 'Teen', 'Adult', 'Old_Age', or 'Unknown', based on the provided age data. Missing racial data was filled with 'Unknown' to maintain data completeness. After these steps, we conducted a thorough assessment of the cleaned columns to ensure data integrity and consistency. Finally, similar to the shooter data, we aggregated victim data at the incident level, summarizing key attributes and counting the number of victims per incident, providing a comprehensive view of the victim demographics for each recorded shooting.

The final phase of our data preparation involved the 'Weapon' tab of the dataset. We started by removing less relevant columns like 'Weapon_Caliber' and 'Weapon_Details' to focus on the most pertinent information. The 'Weapon_Type' column was then meticulously categorized. Standard weapon types such as 'Handgun', 'Rifle', and 'Shotgun' were retained, while multiple weapon instances were classified under 'Multiple_Weapons'. For entries lacking clear data, such as 'No Data', 'NaN', and 'Unknown', we used 'Unknown/Not_Specified' as a category, ensuring consistency in weapon classification. Our cleaning process also involved identifying and handling duplicated incident IDs. Following these steps, we aggregated the weapon data at the incident level. This aggregation included a special handling for 'Multiple' types, which were reclassified as 'Multiple_Weapons', providing a clearer understanding of the incidents involving various weapons. The final step in our data preparation was merging all cleaned tabs – Shooter, Victim, and Weapon – with the Incident tab. This merger was executed using a left join on the 'Incident_ID', effectively consolidating all relevant data into a comprehensive dataset ready for analysis in our machine learning framework.

Once the data tabs were merged with the master 'Incident' sheet using the Incident ID, we conducted a thorough cleaning of the remaining features. This began with a check for missing values, where we calculated the total and percentage of missing data for each column. Columns with a significant amount of missing data or deemed less relevant, such as 'Accomplice_Narrative', 'Number_News', and 'No_of_Victims', were dropped. For other columns, we filled missing values with 'Unknown' or replaced specific entries with more generalized categories for consistency. For instance, 'Active_Shooter_FBI', 'Media_Attention', 'Involves_Students_Staff', 'Gang_Related', 'Victim_Injury', and several others were handled in this manner. We also categorized various attributes like 'Targets', 'No_of_Shooters', 'Situation', 'Time_Period', and 'Location' using custom functions to ensure a uniform and meaningful classification. The categorization process extended to columns such as 'Shots_Fired', 'School_Level', and 'Location_Type', considering different criteria for each. In cases like 'Duration_min', we input missing values with the median. Finally, we standardized the time of day for the 'First_Shot' and filled any missing values in key shooter columns. This meticulous cleaning and categorization process was crucial for ensuring the integrity and reliability of our data for subsequent analysis in our machine learning project.

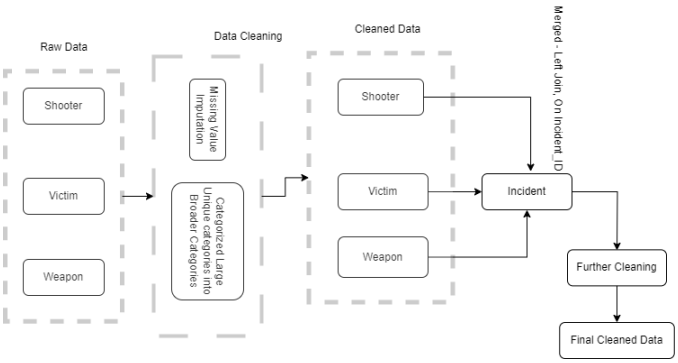


Figure 2 Data Cleaning Process

C. Exploratory Data Analysis

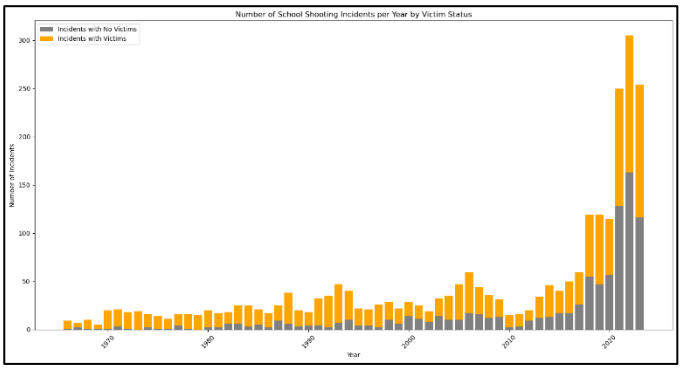
The Exploratory Data Analysis (EDA) conducted in this study serves as the foundational pillar of our research, providing critical insights into the underlying structure and characteristics of the dataset on school shooting incidents. EDA is an essential step in any data science project as it allows for an in-depth understanding of the data's nuances and complexities. This process involves a series of steps aimed at uncovering patterns and dependability on target features, which are crucial for the subsequent modeling phase. This section was structured into several distinct sections as follows,

i. Temporal Analysis

In this section, we delved into the temporal patterns of school shootings, examining trends and fluctuations over time. Key

aspects such as the frequency of incidents across years, months, and days were analyzed to discern any significant temporal trends. This analysis helped in understanding whether certain times were more prone to such incidents.

Figure 5 No of School Shooting Incidents per year by Victim Status



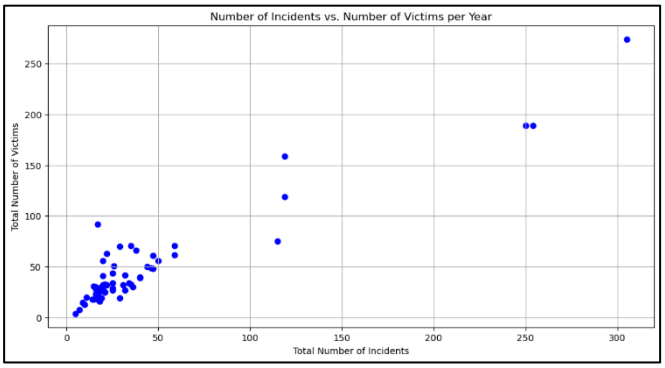
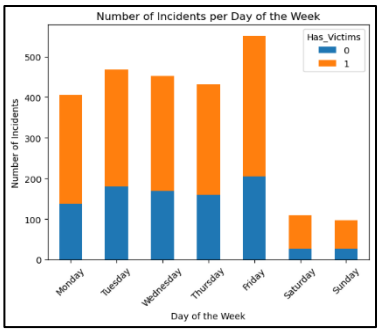


Figure 4 Total No of Victims per year V/s Total No of Incidents

In Figure 3, we present a stacked bar chart that illustrates the frequency of school shooting incidents over time, differentiated by the presence or absence of victims. The X-axis represents the years, while the Y-axis quantifies the number of incidents. The data is segmented to delineate incidents resulting in victims, whether wounded or killed, from those with no casualties. A notable observation is the significant uptick in the number of incidents starting from the year 2020, it coincides with a period of global disruption, suggesting potential underlying social dynamics worth investigating.

In Figure 4, It shows the relationship between the Number of Incidents per year and Total Number of Victims per year, it is clear from the chart that there is a linear relationship between the two features. Isn't it obvious also, as the number of incidents increases the people getting killed or wounded will increase.

Our analysis indicates a clear seasonal pattern in school shooting incidents, with **September** experiencing the highest incidence, likely due to the start of the school year, and the summer months of June and July recording the lowest, coinciding with school vacations. Weekday incidents significantly outnumber those on weekends, reflecting the operational days of schools. These trends are critical for strategic planning in school safety measures.



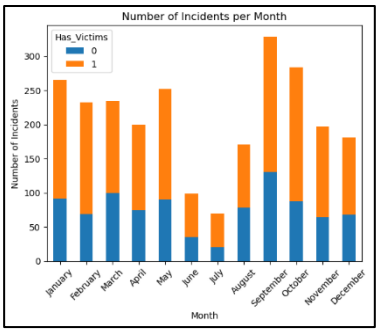
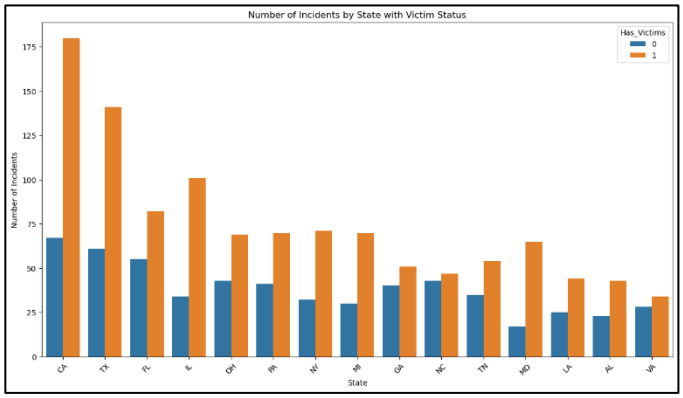


Figure 3 Number of Incidents Per Month and Day of the Week

ii. Geographical Analysis



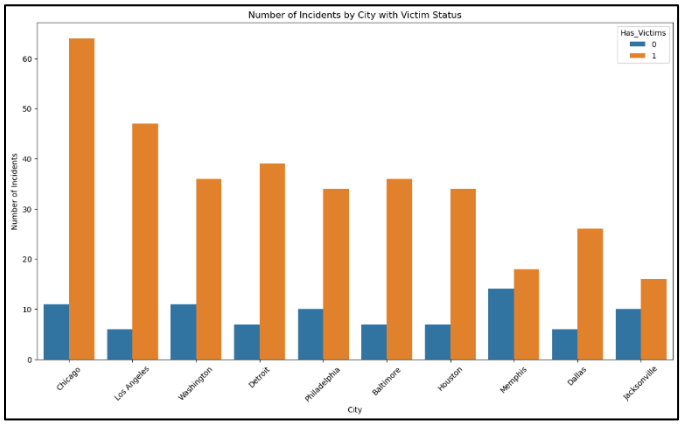


Figure 6: State and City Wise Incidents with Victim Status

Our investigation reveals a geographical concentration of school shooting incidents within certain U.S. states and cities. Approximately 25% of the recorded incidents occur in **California, Texas, Florida, and Illinois**, with a significant portion of these events resulting in casualties. Urban areas, particularly major metropolitan cities such as Chicago, Los Angeles, Washington, and Detroit, are **observed to have a higher occurrence of such incidents. This** trend is further substantiated when examining the number of victims affected, with Chicago, Los Angeles, and Philadelphia frequently experiencing events with casualties, with the notable inclusion of Cokeville. These findings highlight the importance of urban-centric

preventive measures and the need for focused research on mitigating factors contributing to school shootings in these high-incidence locations.

iii. School Analysis

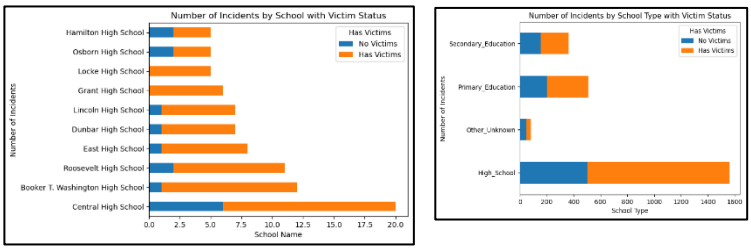


Figure 7 School Wise Incidents

In our dataset, Central High School has the highest occurrence of school shootings, with approximately 20 reported incidents, most of which resulted in casualties. High Schools are the most frequently impacted, accounting for around 64% of all incidents. However, the analysis does not indicate a direct correlation between the frequency of incidents at a school and the number of victims affected. Notably, only about 11 schools have experienced more than five incidents. A significant outlier in the dataset is Cokeville Elementary School, which, despite recording a single incident, had approximately 74 victims, either wounded or killed. This data point is a stark deviation from the normal.

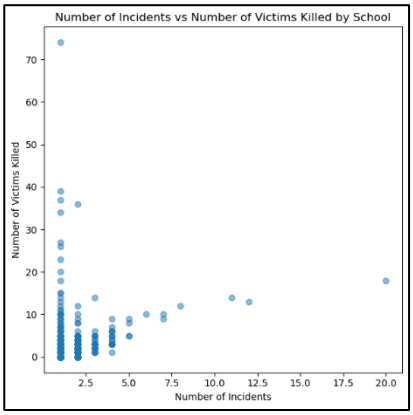


Figure 8 Number of Incidents Vs Total Victims by School

iv. Perpetrator Analysis

In our comprehensive analysis of the shooter profiles, it was observed that the majority of the incidents were perpetrated by males, predominantly within the Teen and Adult age groups. Notably, over 1000 incidents involved a student or Former Student as the shooter. Furthermore, the data indicates that most incidents were carried out by a single shooter. Regarding the motive, approximately 1000 incidents were attributed to conflict-related causes, while around 400 incidents were linked to criminal activities. This insight into the demographic and motivational aspects of the shooters is crucial for understanding the underlying factors of these incidents and developing effective preventative strategies.

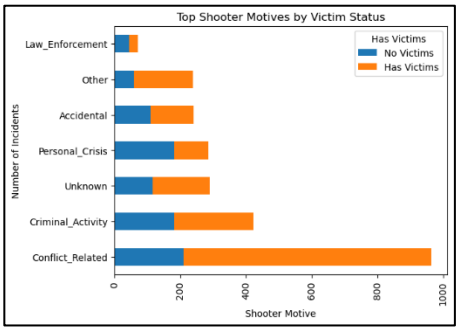


Figure 9 Shooter Motives

v. Weapon Analysis

Most of the instances we observed there was use of Handgun, there were less than 200 instances where multiple weapons were used.

vi. Victim Analysis

In the detailed analysis of victim demographics, it was observed that male victims are more prevalent, with a notable number among them also identified as victims. The predominant age group is teenagers, with adults following closely. Regarding their roles, a significant proportion of the victims were students or former students. Limited data availability on the victims' racial backgrounds restricts further commentary in this area.

D. Feature Selection

Before feature selection, we encoded the 39 categorical features in our dataset to prepare them for machine learning analysis. With over 25 features having more than three categories, we used count encoding to convert these into numerical values based on category frequency. For features with up to three categories, label encoding was applied, assigning distinct integers to each category. This encoding, carried out using specialized functions, transformed the categorical data into a machine learning-compatible format, preserving its integrity for further analysis, including feature selection and modeling.

Feature selection plays a pivotal role in predictive modeling, especially when dealing with datasets that have a large number of variables. In this study, we employed two robust feature selection techniques to identify the most relevant features for our predictive models: the Chi-Squared Test and Recursive Feature Elimination (RFE) using Random Forest.

a. Chi-Squared Test

The Chi-Squared test is a statistical test used to determine if there is a significant association between two categorical variables. In the context of feature selection for machine learning, it's used to test whether a categorical feature is independent of the target variable. The null hypothesis for the Chi-Squared test states that no relationship exists (independence) between the feature and the target.

$$\chi\text{^}2 = (\sum(O - E)\text{^}2)/E$$

Where O represents the observed frequency, and E is the expected frequency under the assumption of independence. The expected frequency for each cell in a contingency table is calculated as

$$E_{rc} = \frac{(Row_r Total) X (Column_c\ Total)}{Total\ Samples}$$

In our study, the Chi-Squared test was employed to assess the relationship between categorical features and the target variable. This was efficiently conducted using a library function that abstracts the detailed calculations involved. The function internally constructs contingency tables for each feature, computes the Chi-Squared statistic by comparing observed frequencies against expected frequencies, and then calculates the corresponding p-values. These p-values are pivotal in our analysis, as a low p-value indicates a significant association between a feature and the target variable, suggesting that the feature is not independent and could be influential in our predictive models. Hence, features with the lowest p-values were prioritized, reflecting their strong relationship with the outcome of interest. This streamlined approach enabled a robust and efficient feature selection process, integral to refining our model.

b. Recursive Feature Elimination

In this study, Recursive Feature Elimination (RFE), integrated with the Random Forest algorithm, was utilized to identify key predictive features from our dataset. RFE, a method known for its efficacy in reducing feature dimensionality, operates by iteratively constructing models and eliminating the least significant features. We initiated RFE with a comprehensive set of features, employing the Random Forest model due to its robustness and inherent feature importance ranking capability. The process involved iterative training of the model, where in each iteration, the feature deemed least significant was removed. This elimination continued until a predetermined subset of features was achieved. The efficacy of each subset was rigorously evaluated, allowing us to ascertain the most impactful features. The application of RFE, facilitated by a library function, streamlined this process, efficiently distilling the feature set to those most consequential for our predictive analysis. This approach not only refined the model's performance but also enhanced interpretability by isolating the most relevant predictors, thereby striking an optimal balance between model simplicity and predictive accuracy.

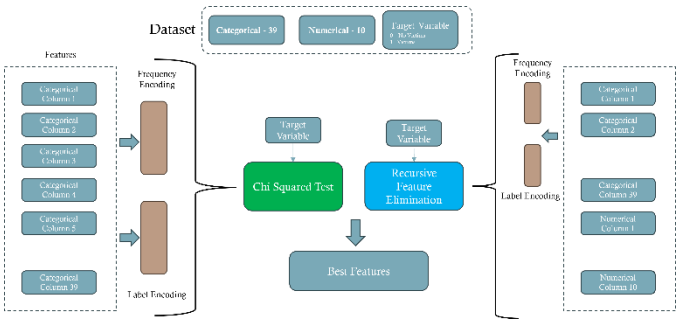


Figure 10 Framework for Feature Selection

E. Model Building

After feature selection, our next methodology involves experimenting with both Logistic Regression and Support Vector Machine (SVM) models, each with two variations: one using all available features and the other using only the best-selected features. This approach allows us to compare the effectiveness of feature selection in enhancing model performance. Logistic Regression and SVM are employed due to their suitability for binary classification problems and their ability to handle high-dimensional data.The dataset was split into two parts: the training set and the test set. This split is crucial for evaluating the model on unseen data, thereby providing an unbiased assessment of its performance. Typically, the data is divided with 70-80% allocated for training and the remaining 20-30% for testing. This ratio ensures adequate data for model learning while reserving a sufficient amount for evaluation.

Logistic Regression:

In our study, we first employed a Logistic Regression model, a widely used statistical method for binary classification problems. The logistic regression function is defined as:

$$\text{Logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 X_1 + \cdots \beta n Xn$$

Where,

p is the probability of the presence of the characteristic of interest. The coefficients $\beta_0, \beta_1, . \beta_n$ are estimated from the training data. The model was trained on our dataset and evaluated on a test set. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, ROC-AUC, and the Confusion Matrix. These metrics provide a comprehensive view of the model's performance, balancing both the error rate and the ability to correctly identify positive instances.

Support Vector Machine:

Following the logistic regression, we implemented a Support Vector Machine (SVM) with a polynomial kernel. SVM is a powerful and versatile classification algorithm, particularly effective in high-dimensional spaces. In its basic form, SVM attempts to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. The kernel trick is used to transform the data and then based on these transformations, find an optimal boundary between the possible outputs. In our study, the polynomial kernel was utilized. The polynomial kernel is defined as

$$K(x_i, x_j) = (\gamma \langle x_i, x_j \rangle + r)_d$$

where K is the kernel function, x are data points, γ is the scale parameter of the kernel, r is the independent term, and d is the degree of the polynomial. The SVM model was similarly evaluated using metrics such as Accuracy, Precision, Recall, F1-Score, ROC-AUC, and the Confusion Matrix. These metrics facilitate an understanding of the model's effectiveness in classifying data, with ROC-AUC providing insight into its discriminatory ability.

F. Model Evaluation

In the realm of machine learning, particularly for classification problems, the efficacy of models is evaluated using a suite of metrics, each offering unique insights into various aspects of performance. Accuracy serves as a primary indicator, reflecting the overall rate of correct predictions. Precision and Recall, respectively, measure the model's exactness in predicting positive classes and its effectiveness in identifying true positive instances.

The F1-Score, as the harmonic mean of Precision and Recall, provides a balanced metric in cases of uneven class distribution. Complementing these is the ROC-AUC, which assesses the model's discriminatory ability between classes, with higher values indicating superior distinction. Additionally, the Confusion Matrix offers a detailed breakdown of the model's predictions, categorizing them into true positives, true negatives, false positives, and false negatives. Together, these metrics present a comprehensive evaluation of a model's classification capacity, underscoring its precision, reliability, and overall predictive quality.

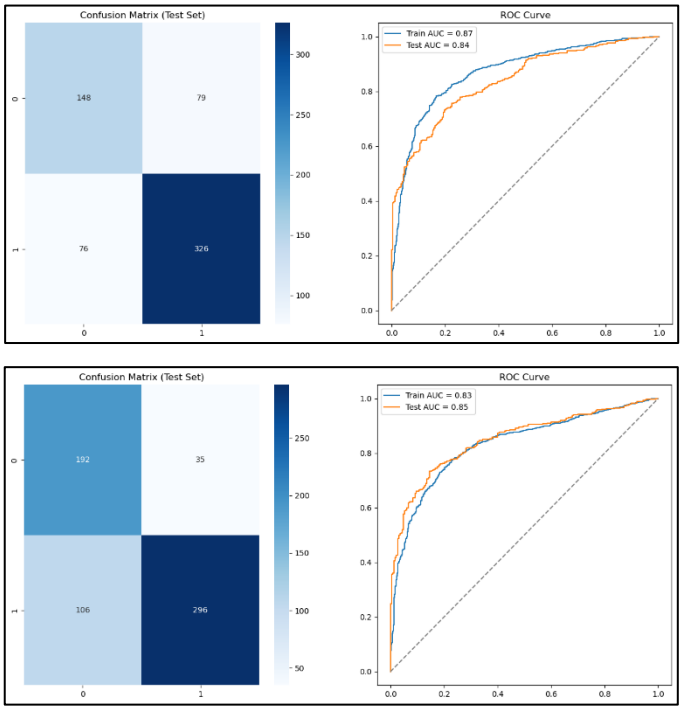
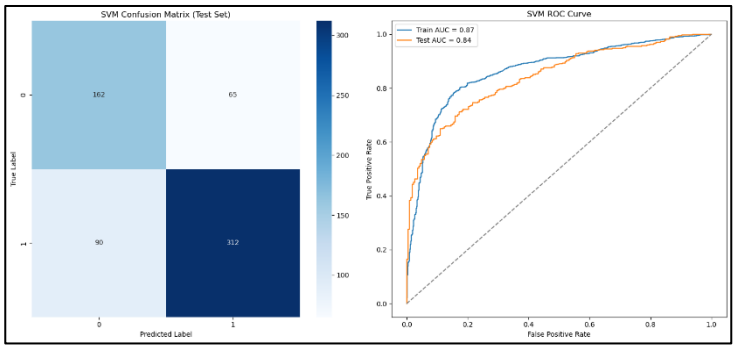
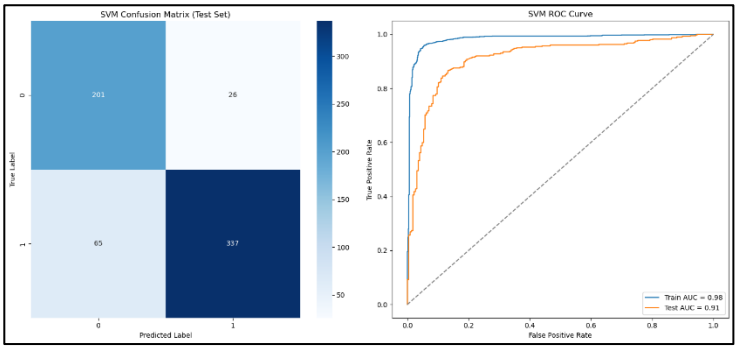
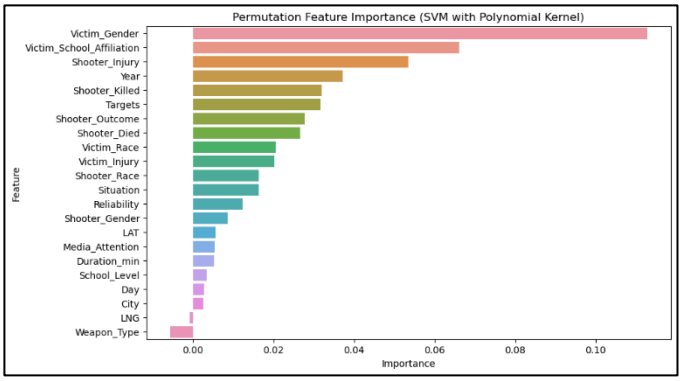


Figure 11 : Confusion Matrix and ROC-AUC Curve for Logistic Regression







Support Vector Machines (SVMs) do not inherently provide feature importance scores. For linear SVMs, feature weights can indicate importance, but for non-linear SVMs, such as those with polynomial kernels, methods like permutation feature importance are used. This technique involves shuffling each feature in the dataset and observing the impact on model performance. Libraries like scikit-learn offer tools for implementing permutation feature importance, which is model-agnostic and widely applicable across different machine learning algorithms.

It is important to note that permutation feature importance measures the increase in the model's prediction error after permuting the feature. A higher value means that scrambling the feature's values increases model error, thereby indicating that the model relied on the feature for the prediction.

These insights can inform further model refinement and feature engineering. For instance, features with low importance could potentially be removed to simplify the model without

substantially affecting performance. Conversely, features with high importance could be the focus of further analysis or feature engineering to enhance model predictions.

V. Results

In this section, we see how well our model performed on the unseen test dataset. Below are the performance metrics with respect to the test dataset.

Logistic Regression		
Metric	**Before Feature Selection**	**After Feature Selection**
Accuracy	75%	78%
Precision (Class 0)	66%	64%
Precision (Class 1)	80%	89%
Recall (Class 0)	65%	85%
Recall (Class 1)	81%	74%
F1-Score (Class 0)	66%	73%
F1-Score (Class 1)	81%	81%
ROC-AUC (Test)	84%	85%

Support Vector Machine		
Metric	**Before Feature Selection**	**After Feature Selection**
Accuracy	75%	86%
Precision (Class 0)	64%	76%
Precision (Class 1)	83%	93%
Recall (Class 0)	71%	89%
Recall (Class 1)	78%	84%
F1-Score (Class 0)	68%	82%
F1-Score (Class 1)	80%	88%
ROC-AUC (Test)	84%	91%

VI. Conclusion

The conducted experiments encompassed four distinct modeling scenarios, comprising Logistic Regression and Support Vector Machine (SVM) algorithms, each with two feature-set variations. The SVM leveraging selected features emerged as the most proficient model, manifesting an 88% F1-Score and an approximate ROC-AUC of 0.95. In contrast, the SVM with all features registered an F1-Score of 80% and an ROC-AUC close to 0.90.

Comparatively, Logistic Regression models exhibited divergent tendencies based on the feature sets employed. The model utilizing all features attained an 81% F1-Score, whereas the version with selected features also reached an 81% F1-Score but with a notable precision improvement from 80% to 89%. ROC-AUC estimations from ROC curves indicated a marginal superiority of the all-features Logistic Regression model (approximately 0.88) over the selected-features variant (approximately 0.85).

These outcomes underscore the pivotal role of feature selection in the SVM algorithm's capacity to enhance predictive accuracy. Moreover, the discernible discrepancy in performance metrics between models with all features versus those with selected features reaffirms the necessity for judicious feature selection in SVM applications.

Future research may benefit from a deeper exploration into optimizing feature selection methods, potentially integrating ensemble techniques to further amplify the predictive acumen of SVM models in complex classification landscapes.

References

[1] Riedman, David. ""K-12 School Shooting Database"" (2023).

[2] https://k12ssdb.org/methodology-1.

[3] [3] Kyleigh Cummings, “Down the Barrel of School Shootings”,California State University, Fullerton.

[4] [4] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5] [5] K. Elissa, “Title of paper if known,” unpublished.

[6] [6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[7] [7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] [8] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[9] [9] Catherine N. Rasberry, “Morbidity and Mortality Weekly Report”, center for Disease control and prevention.

[10] [10] Wheeler, A.; Worden, R.; Worden, R.; Silver, J. The Accuracy of the Violent Offender Identification Directive (VOID) Tool to Predict Future Gun Violence.

[11] [11] R. L. Plackett, “Karl Pearson and the Chi-Squared Test”, Vol. 51, 1983, International Statistical Institute.

[12] [12] Serpil Ustebay, Zeynep Turgut, M.ali Aydin, Intrusion Detection System With Recursive Feature Elimination.

[13] [13] Sara B. Heller, Benjamin Jakubowski, Zubin Jelveh, Max Kapustin, Machine Learning Can Predict Shooting Victimization Well Enough To Help Prevent It, National Bureau Of Economic Research.

[14] [14] R Khemchandani, Suresh Chandra, et al. Twin support vector machines for pattern classification. IEEE Trans. pattern analysis and machine intelligence, 29(5), 2007.

[15] [15] Yitian Xu, Xianli Pan, Zhijian Zhou, and et al. Structural least square twin support vector machine for classification. Applied Intelligence.

[16] [16] Heller, S.B.; Jakubowski, B.; Jelveh, Z.; Kapustin, M. Machine Learning Can Predict Shooting Victimization Well Enough to Help Prevent It; Working Paper 30170; National Bureau of Economic Research: Cambridge, MA, USA, 2022.

[17] Sugen Chen, Xiaojun Wu, and Renfeng Zhang. A novel twin support vector machine for binary classification problems. Neural Processing Letters.

[18] [18] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Networks.

[19] [19] Vojtech Franc and V´aclav Hlav´ac. Multi-class support vector machine. In IEEE Trans. Pattern Recognition, volume 2, pages 236–239, 2002.

[20] [20] Lei Chen, Sule Gunduz, and M. Ozsu. Mixed type audio classification with support vector machine. In IEEE, ICME, pages 781–784, 2006.

[21] [21] Huiqin Chen and Lei Chen. Support vector machine classification of drunk driving behaviour. International Journal of Environmental Research and Public Health, 14(1), 2017

[22] [22] King, Gary, and Langche Zeng. 2001b. “Logistic Regression in Rare Events Data.” Political Analysis, 9: 137–163.

[23] [23] Christian Westphal, Logistic Regression for Extremely Rare Events: The Case of School Shootings, University of Marburg - School of Business & Economics; University of Marburg.

[24] [24] Wang, N.; Varghese, B.; Donnelly, P.D. A machine learning analysis of Twitter sentiment to the Sandy Hook shootings. In Proceedings of the 2016 IEEE 12th International Conference on e-Science (e-Science), Baltimore, MD, USA, 23–27 October 2016; pp. 303–312.

[25] [25] Liu, D.; Sasha Dong, Z.; Qiu, G. Exploring the contagion effect of social media on mass shootings. Comput. Ind. Eng. 2022, 172, 108565.