Quest Journals
Journal of Software Engineering and Simulation
Volume 9 ~ Issue 4 (2023) pp: 40-46
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Using XGBoost to Study the Impact of Firearm Laws on Gun Violence

Ronald Feng
[1]*(Rye Country Day School)*
Reham Islam
[2]*(Computer Science Department, Rye Country Day School, US)*
Corresponding Author: Ronald Feng

ABSTRACT: *In recent years, the number of gun violence incidents and victims has increased dramatically in the United States. Unfortunately, there is a lack of studies towards understanding gun violence and finding ways to reduce and prevent it. In this paper, XGBoost, a cutting-edge machine learning technique used for binary classification and prediction, was employed to identify the most important firearm laws for predicting daily gun incidents. The XGBoost algorithm identifies the top 10 most important firearm laws which affect the daily gun incidents, with a high classification accuracy of 87%. Based on our literature review, this paper is the first study to use artificial intelligence and machine learning techniques to analyze gun violence data and identify the most important firearm laws for reducing and preventing daily gun incidents. This research is useful as the foundation for future studies focused on the impact of firearm laws and their connection to gun violence. The approaches used for this study could inspire the further assessment of gun violence through advanced data science techniques.*
KEYWORDS: *XGBoost, machine learning, gun violence, firearm laws*

***R**eceived 01 Apr., 2023; **R**evised 10 Apr., 2023; **A**ccepted 12 Apr., 2023 © The author(s) 2023.*
P**ublished with open access at **www.questjournals.org

I. INTRODUCTION

Gun crime has become a national public health and social epidemic in the US. In recent years, incidents of gun violence and victims have increased dramatically in the United States. On average, there are more than 37,000 deaths and 78,000 nonfatal injuries every year [1]. In 2022 alone, there have been 695 mass shootings causing 762 deaths and 2,902 injuries [2]. Gun violence impacts three million children each year, resulting in death, injury, and lasting trauma [3]. Statistically speaking, gun violence in the US is at least seven times greater than in the United Kingdom, Germany, France, Italy, Canada, Australia, Japan, and South Korea [4]. Clearly, gun violence in the US warrants special attention. Many policies and laws have been implemented and modified in the US for years trying to reduce gun crimes. However, there are several significant questions about gun violence for which researchers are still seeking answers [5]. Among them, the most important one is perhaps which firearm policies, if any, work best to prevent and reduce gun violence.

A few prior studies have been conducted on this subject. Rocque et al. applied regression analyses on mass shootings by state and year and concluded permit laws and large capacity magazine bans were related to fewer mass shootings and fewer victims [6]. Mass shootings only consist of a small portion of all gun crimes. The research does not study the impact of firearm laws on other types of gun violence. Another study used handgun background check data to estimate the association between CBC policies and changes in background check rates for firearm acquisition in Oregon and Washington [7]. The study is based on data observation. More sophisticated analysis should be applied to study the impact of firearm laws. Doucette et al. employed statistical methods to examine the average effect (pooled, cross-sectional, time-series analysis) and the state-specific effect (random effects meta-analysis) of right-to-carry (RTC) firearm laws on firearm workplace homicides (WPHs) in the United States from 1992 to 2017 [8]. All these research studies either focus on specific firearm laws or certain gun crimes. There is also a comprehensive review of gun policy studies by the RAND Organization [9]. They reviewed eighteen classes of gun policies and eight outcomes induced by those policies at the state level. Their research, based on scientific evidence, presents conclusions that can be drawn on various gun policies and their societal effects. Nevertheless, on many occasions, RAND could not identify specific research that could provide concrete evidence concerning a gun policy's effects. In summary, it is statistically significant that ten out of eighteen gun policy categories affect four out ten outcomes. For the remaining eight policies, no statistically important scientific evidence indicates their effectiveness. But the result doesn't really suggest that gun policies are ineffective, but rather that the current research effort is not sufficient to draw adequate or definitive conclusions.

In recent years, machine learning techniques have gained popularity among many researchers to uncover key insights in a wide variety of research fields. Machine learning algorithms are aimed to produce a model that can be used to perform classification, prediction, or estimation [10]. Machine learning primarily works with a huge amount of data. The ML techniques use feature engineering techniques to create and extract features from big datasets [11]. The features are fed into computer algorithms to train the algorithm and detect patterns that are not easily identifiable otherwise. In addition, the ML algorithm can be used to extract the optimal features affecting the objective variables or categories.

The goal of this paper is to find out if there is any impact of firearm laws on gun violence. This study merged gun violence and state firearm law data. A cutting-edge machine learning model, XGBoost (Extreme Gradient Boosting), was trained and tested by the merged data. This study successfully predicted the occurrences of gun incidents from various gun law provisions. The feature importance of the XGBoost model identified the most effective firearm laws to reduce gun violence incidents. Overall, the XGBoost model achieved a relatively high prediction accuracy of 87% in the testing set. This research provides new inspiration to apply advanced data science approaches to understanding gun violence. Hopefully, it will shed some light on how to reduce the number of gun crimes and their severity.

II. DATA COLLECTION

Reliable, detailed, and complete gun violence and firearm laws data are critical for studying gun crime incidents and the effects of firearms policies. It is challenging to access and retrieve clean and comprehensive gun crime and firearm law data that contains adequate information. In this research, a large and complete data set was created by downloading, combining, and cleaning the data from several sources. Our data set consists of about 10 years of state-level time series data on gun violence incidents and firearm laws in the United States.

The gun violence data used in this paper was downloaded from Gun Violence Archive (GVA) [12] by James Ko [13] and Emmanuel Werr [14]. Formed in 2013, GVA is a nonprofit organization that aims to collect accurate and comprehensive information about gun-related violence in the U.S. and then post and disseminate it online. This dataset has 29 columns and 472,820 rows, containing specific information regarding gun violence incidents from 2013 to May 2022.

Firearm law data was retrieved from a State Firearm Laws project [15] on Kaggle. The State Firearm Laws project provides information about the federal regulations of firearms from 1991 to 2017, for all 50 U.S. states. The firearm laws from 2018 to 2021 were supplemented by downloading data from Rand.org [16]. In Jan 2016, RAND launched the Gun Policy in America initiative to provide information for public discussion and support on the effects of gun laws. In order to complete the research, data was cleaned and compiled on 133 provisions of firearm laws in all 50 states through the years. The law data from RAND excludes the District of Columbia. There are 14 categories and multiple sub-categories under each category.

III. METHODOLOGY

Figure 1 shows the flow chart of the research methodology. The process consists of three parts – data downloading, data processing and XGBoost modeling.

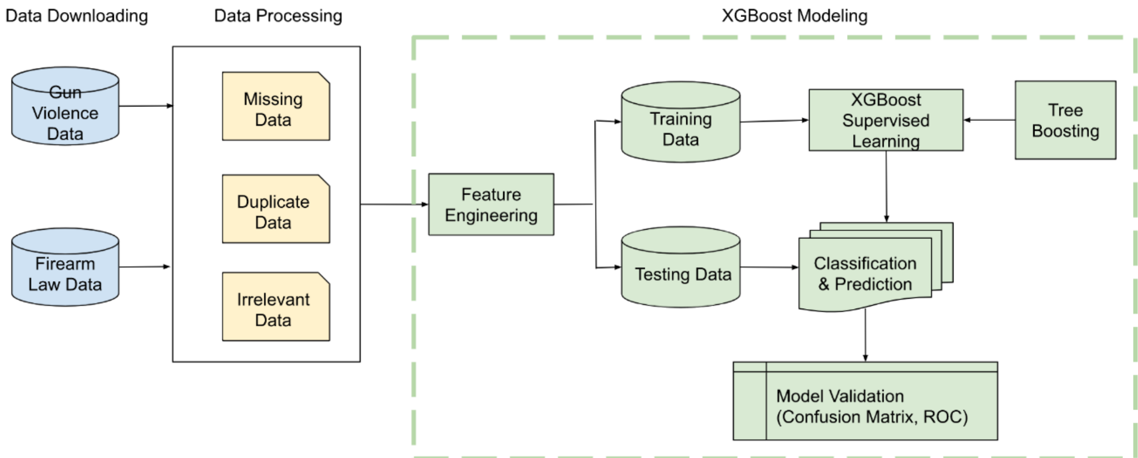

Figure 1: Methodology Flow Chart

This study combined gun violence data and firearm law data to a comprehensive dataset. Data was processed to clean the missing values, remove duplicate values, and fix irrelevant or incorrect data. The first step of developing our machine learning model is to merge the datasets. Due to the fact that our gun violence incidents are on a daily basis, while our firearm law data are on yearly basis, the 2 datasets were merged based on the assumption that the laws don't change during the year. Our assumptions are valid because laws should not change drastically in one year.

Next, feature engineering techniques were used to obtain a set of new features of gun violence incidents. Our new features include counts of gun incidents and sums of people killed and injured per state on a daily, monthly, and yearly basis.

After the new features were created, the featured variables and target variables were further split into a training set and a testing set. The training set was fed into the XGBoost model. The tree-boosting algorithm of the XGBoost generates prediction and classification reports. The testing data was then fed to validate the model result.

3.1 XGBoost Modeling

XGBoost (eXtreme Gradient Boosting Decision Tree), proposed by Chen et al. [17], is an optimized distributed gradient boosting algorithm designed to improve the running speed and accuracy of the original boosting algorithm. Since its introduction in 2014, XGBoost has become one of the leading machine learning models for regression, classification and ranking problems due to its fast speed and high-performance [18]. Thus, we expect XGBoost to perform well for our classification task.

By combining multiple learning models, XGBoost algorithm can achieve a strong generalization ability to obtain good modeling effects [19]. The XGBoost algorithm comprises many decision tree iterations. Through a large number of iterations, the classifier in each iteration is trained based on the prediction results of the previous classifier. XGBoost can be represented as:

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \rightarrow (1)$$

$f_k(x)$ represents the predicted value obtained after inputting sample x into the tree, y represents the prediction result. The XGBoost generally runs these steps:

$$\hat{y}^{(0)} = 0$$
$$\hat{y}^{(1)} = \hat{y}^{(0)} + f_1(x) \rightarrow (2)$$
$$\dots$$
$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + f_t(x)$$

Where $\hat{y}^{(t)}$ represents the prediction result of the t_{th} iteration, $\hat{y}^{(t-1)}$ represents the prediction result of the previous $t-1_{th}$ iteration, and $f_t(x)$ represents the newly added prediction function in each round. The goal of prediction is to make $\hat{y}$ as close as possible to the true value. The objective function is as follows：

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l(y, \hat{y}^t) + \sum_{k=1}^{t} \Omega(f_k) \rightarrow (3)$$

$$\Omega(f_k) = \gamma^T + \frac{1}{2}\lambda||\omega||^2 \rightarrow (4)$$

In equation (3), the first term $l(y, \hat{y}^t)$ represents the error or loss function, which measures the difference between the predicted y and the target y. The second term $\Omega(f_k)$ represents the regularization item, which defines the complexity of the tree to prevent the model from overfitting. Equation (4) is the mathematical expansion of the regularization item. γ is a parameter to control the number of T nodes, λ is the parameter that controls the weight of the leaf node. At the end of the iterations, the XGBoost algorithm will assign each factor an importance score and generate an overall accuracy score for the whole prediction process.

Figure 2 shows how the XGBoost model works. The model consists of many decision trees created in order. Each decision tree predicts the target variable Y from variable X. The residual of the prediction from each decision tree will feed into the next decision tree to adjust the prediction error. The process will iterate until the objective function result meets the requirement.

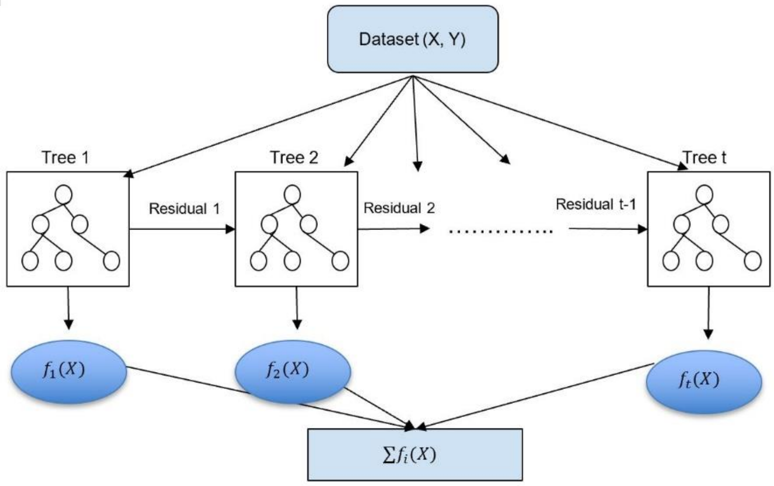

Figure 2: XGBoost Model

3.2 Model Evaluation

In this research, daily gun incidents were divided into 2 categories by 80% vs 20% rule: low occurrence (80%) - the number of daily incidents equal to or less than 6, and high occurrence (20%) - the number of daily incidents greater than 6. There were 91,912 records falling into the low occurrence category and 21,490 records contained in the high occurrence category. The chosen cutoff number 6 is reasonable because the number of daily gun incidents should be small. However, the two classes of gun crimes by this cutoff are unbalanced. In this case, a single accuracy score was not reliable enough to measure the model performance. Thus, more metrics were required to evaluate the overall performance of the XGBoost algorithm.

The performance of the machine learning models for classification problems can be measured through the confusion matrix [20]. A confusion matrix summarizes the number of correct and incorrect predictions made by a classifier [21], which provides insight into where the classification model is correct as well as what types of errors it is making [22]. A confusion matrix for binary classification is shown Figure 3. Below are definitions of TP, FP, FN, and TN.

- **TP: True Positive** - The actual and predicted values are both positive.
- **FP: False Positive** - The actual values are negative but falsely predicted as positive. Also known as Type I Error.
- **FN: False Negative** - The actual values are positive but falsely predicted as negative. Also known as Type II Error.
- **TN: True Negative** - The actual and predicted values are both negative.

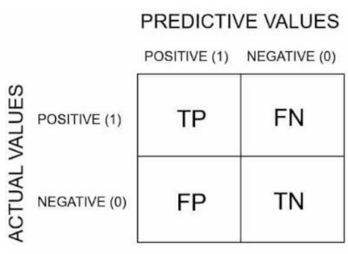

Figure 3: Confusion Matrix

Certain performance metrics can be calculated in a confusion matrix to evaluate the performance of a classification model [20]. A typical classification report includes four metrics (accuracy, precision, recall, and F1 score). Accuracy indicates the model's overall correctness, which is calculated as the ratio of the correct predictions to the total number of predictions (equation (5)). A high accuracy metric means a good prediction. Precision measures the prediction of a specific category. Precision is defined as the ratio of the number of true positive classes to the total predicted positive classes (equation (6)). The precision score should be high to indicate an accurate prediction. Recall shows the model's capability of detecting a specific category. Recall is calculated as the ratio of true positive classes to all positive classes (equation (7)). A high recall represents a good prediction. When the dataset is unbalanced, accuracy may not be a good measure. Instead, the F1-score is a better measure than accuracy. F1-score is defined as the harmonic mean of precision and recall (equation (8)). F1-score ranges from 0 to 1. A high F1-Score means a good predictive model and classification. The classification of binary outcomes can also be observed through a receiver operating characteristic (ROC) curve [23]. A ROC curve plots the false positive rate of predictions against the true positive rate. The area under the ROC curve (AUC) measures the degree of separability, which delineates the model's effectiveness to distinguish among different categorical groups. A higher AUC score represents a better model prediction for binary classes.

$$Accuracy = \frac{TP + TF}{TP + TN + FP + FN} \rightarrow (5)$$

$$Precision = \frac{TP}{TP + FP} \rightarrow (6)$$

$$Recall = \frac{TP}{TP + FP} \rightarrow (7)$$

$$F_1 - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \rightarrow (8)$$

IV. RESULTS & DISCUSSION

The analysis and modeling in this study are implemented using Python language programming. The simulation platform is Google Colaboratory with 1Tesla T4 GPU, Intel(R) Xeon(R) CPU @ 2.30GHz, 12.6 GB RAM, and 108 GB Disk Space; the programing packages include Python 3.8, xgboost0.9, seaborn 0.11.2 , sklearn 1.0.2, pandas1.3.5, numpy1.21.6, matplotlib3.2.2 and others.

To find out the relationship between gun crime and firearm laws, the XGBClassifier() class was employed to fit the classifier to the training set and make predictions on the test set. The feature variables (X) were defined as the firearm laws and the target variable (Y) were defined as the per-state daily gun incidents. The target variable, the daily gun incidents, was categorized as low occurrence (6) and high occurrence (>6) as described in Section 3.2. The data set was split into training and test sets with 80% to 20% ratio, allowing training the XGBoost model and examination of accuracy of prediction.

Figure 4 plots the top 10 features ranked by their importance related to daily gun incidents. The feature importance is produced by the XGBoost classifier and ranked on the "Gain" factor. According to the results from the XGboost model, there exists a high correlation between firearm laws and rates of gun violence. Among all firearm laws, *Opencarryl* (No open carry of long guns is allowed in public places unless the person has a permit) is the most important law to prohibit the number of gun incidents, followed by *age21handgunpossess* (Purchase of handguns from licensed dealers restricted to age 21 and older), *expartedating* (Restriction on gun ownership for persons possessing close relationships with dangerous people), etc. These top 10 features are under 6 categories - 4 in Possession regulations, 2 in Domestic violence, and 1 each in Child access prevention, Prohibition for high-risk gun possession, Immunity, and Background check. Our research results indicated that possession regulation is the most important regulation category related to gun violence reduction because there are 4 possession regulation firearm laws in the top 10 features. Additionally, the first two most important features all belong to possession regulations.

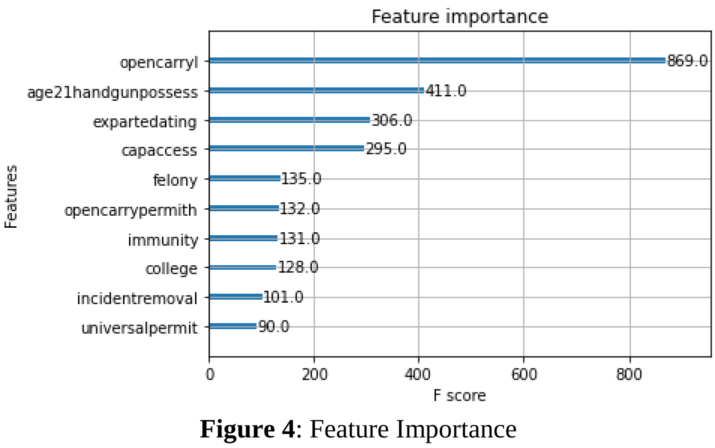

Figure 4: Feature Importance

The prediction accuracy score from our XGBoost fit is 0.87, which indicates the model has a high degree of accuracy. Figure 5 depicts accuracy, precision, and f-score. Category 0 represents daily gun incidents less than 6, and 1 represents daily gun incidents greater than 6. The precision score is 92%, the recall score is 92%, and the f1-score is 92% for category 0. Our scores for each of these categories are all high, so our prediction result is quite accurate. The category 1 has a precision score of 65%, the recall score of 66%, and the f1-score of 65%. These numbers also suggest a relatively accurate prediction. Figure 6 displays the XGBoost confusion matrix. The prediction accuracy (the upper left box and the lower right box) is very high compared to the wrong predictions. As shown in Figure 7, the ROC curve from XGBoost displays the superiority in classifying high vs. low daily gun incidents occurrences. The correspondent AUC score is 0.9, representing an accurate prediction for the classification of gun incidents by XGBoost.

```
Best AUC Score: 0.9001923810865555
              precision    recall  f1-score   support

           0       0.92      0.92      0.92     18413
           1       0.65      0.66      0.65      4268

    accuracy                           0.87     22681
   macro avg       0.79      0.79      0.79     22681
weighted avg       0.87      0.87      0.87     22681
```

Figure 5: Classification Report

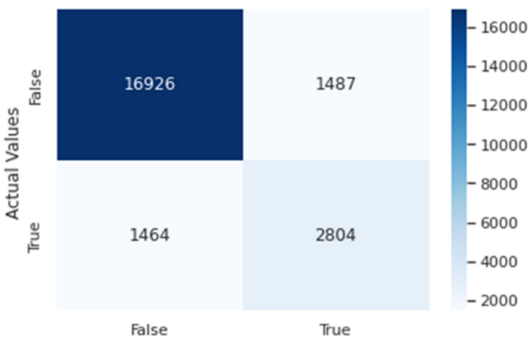

Figure 6: Confusion Matrix

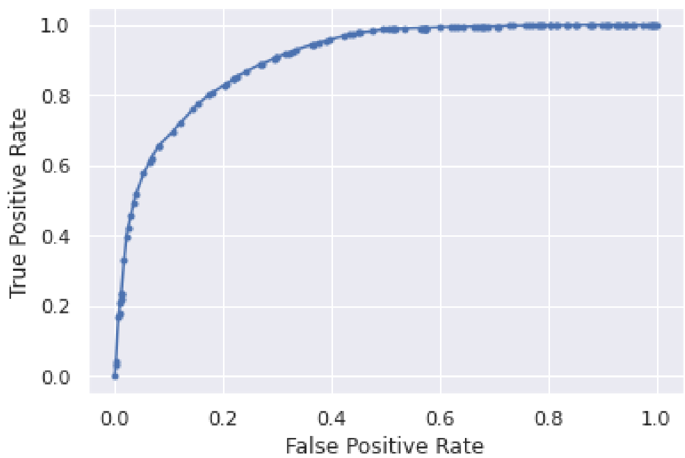

Figure 7: ROC Curve

V. CONCLUSION

Gun violence is rampant in the United States and is a serious public safety and health concern. Scientific research on what firearm laws work best to reduce gun violence significantly lags. This paper presents a new approach to identifying the most important firearm laws for predicting daily gun incidents. This research project is the first study to use machine learning algorithms to identify the most important firearm laws used for predicting daily gun incidents. Data cleaning and feature engineering were conducted to create features to feed our machine-learning prediction model. Then, XGBoost, a powerful machine learning technique for classification and prediction, was employed to identify the most important features for predicting daily gun incidents. The XGBoost algorithm had a high classification accuracy of 87% and selected the top 10 most important features which affected daily gun incidents, such as *opencarryl* (open carry long gun), *age21handgunpossess* (age 21 to possess handgun), and *expartedating* (restriction on gun possess of persons with a close relationship to dangerous people), etc. Furthermore, XGBoost model performance was validated through evaluation metrics. The model evaluation results showed the XGBoost model had good classification and prediction performance with an AUC score of 0.9. Altogether, this research offers promising new insights, thereby could serve as the foundation for future works on studying the impact of firearm laws on gun violence. Moreover, these approaches are likely to inspire the further assessment of gun violence through advanced data science techniques.

In the future, machine learning study will be extended to study mass shootings because of their significant adverse impact on American society. Understanding the impact of various gun policies on mass shootings is extremely important in order to prevent them from occurring in the future. Finally, this study discovered that the majority of gun violence happens on school campuses. Future researchers should use this groundbreaking approach to determine the causes of these tragic events and to offer viable solutions to prevent them from happening again.

REFERENCES

[1]. Galea S et al., Priorities in Recovering from a Lost Generation of Firearms Research. American Journal of Public Healths, 2018. 8(7)
[2]. List of Mass Shootings in the United States in 2022. https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2022
[3]. Katherine Fowler et al, Childhood Firearm Injuries in the United State. Pediatrics, 2017 140 (1).
[4]. Philip Alpers and Marcus Wilson., Global Impact Of Gun Violence Firearms, Public Health and Safety. http://www.gunpolicy.org/firearms/region
[5]. Brad Plumer, The Biggest Questions about Gun Violence That Researchers Would Still Like to See Answered. https://www.vox.com/platform/amp/2016/6/15/11934246/gun-violence-america-questions
[6]. Michael Rocque et al., Policy Solutions to Address Mass Shootings. National Institute of Justice, 2021 322.304951: p. 891-921
[7]. Castillo Carniglia et al., Effect on Background Checks of Newly-enacted Comprehensive Background Check Policies in Oregon and Washington: A Synthetic Control Approach. Injury Epidemiology, 2019 6(45).
[8]. Doucette ML et al., Right-to-Carry Laws and Firearm Workplace Homicides: A Longitudinal Analysis (1992-2017). AM Journal of Public Health, 2019 109 (12): p. 1747-1753
[9]. Samantha Cherney et al., Development of the RAND State Firearm Law Database and Supporting Materials. RAND Corporation, 2022
[10]. Stuart J. Russell and Peter Norvig., Artificial Intelligence: a modern approach. 3rd ed. Pearson, 2009.
[11]. Tara Rawat and Vineeta Khemchandani. Feature Engineering (FE) Tools and Techniques for Better Classification Performance, 2019.
[12]. GVA Site, Gun Violence Archive 2022. https://www.gunviolencearchive.org
[13]. James KO, Gun Violence Data, Comprehensive record of over 260k US gun violence incidents from 2013-2018. https://www.kaggle.com/datasets/jameslko/gun-violence-data
[14]. Emmanuel F. Werr,.Gun Violence Incidents in the United State, https://www.kaggle.com/datasets/emmanuelfwerr/gun-violence-incidents-in-the-usa
[15]. Jonathan Bouchet., Firearm Regulations in the U.S, https://www.kaggle.com/code/jonathanbouchet/firearm-regulations-in-the-u-s/report
[16]. Rand.org.The Gun Policy in America Initiativ.e https://www.rand.org/research/gun-policy.html
[17]. Tianqi Chen and Carlos Guestrin, XGBoost: A Scalable Tree Boosting System. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco California USA August 13 - 17, 2016
[18]. Liang H et al. XGBoost: An Optimal Machine Learning Model with Just Structural Features to Discover MOF Adsorbents of Xe/Kr. AACS Omega 2021 6(13), pp. 9066–9076
[19]. Bing Liu et al., Application of RR-XGBoost combined model in data calibration of micro air quality detector. Scientific Reports 2021 11(1), pp. 2045–2322
[20]. Ajay Kulkarni, Deri Chong, and Feras A.Batarseh., Data Democracy. Elsevier Inc., 2020. Chap. 3.1.
[21]. David M. W. Powers. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation. Journal of Machine Learning Technologies, 2011 2(1) p. 37–63.
[22]. Stephen V. Stehman. Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment 2011 62(1) p. 77–89.
[23]. P. Bradley, The Use of the Area under ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, 1997 30(7) p. 1145–1159.