Final Report (CPE-695)

Yijia Qi, Nishanth Mudkey, and Tianyu Sheng

I. Introduction

This project is motivated by publicly available open-source crime data sets. In recent years, researchers have extensively used machine learning and data mining techniques in the crime analysis arena to extract association rules, frequent patterns, clusters, or correlations. Machine learning and prediction models have been used to classify patterns to predict future crime variables. This report is inspired by machine learning's usability in crime prediction and classification. The following sections are covered in this report — a detailed explanation of the problem statement, an overview of the related work, a description of the data set, detailed algorithmic implementations, comparison of results, future research directions, and conclusion.

A. Problem Statement

The Global Terrorism Database [1], [2] used in the current project is a terrorism incident database maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland. Exploratory data analyses are conducted and machine learning models are built to classify and predict multiple terrorism related variables.

1. To understand the global terrorism database.

In the current attempt, data visualization techniques are used to better understand the global terrorism database (GTD) and its inherent crime variables.

2. To implement classification and prediction models using the global terrorism database (GTD).

In the current project, four different algorithms are implemented for classification and prediction of terror success and casualties. The success of a terrorist event and the casualties (number killed + number wounded) involved are two significant terror attack outcomes.

"Success of a terrorist strike is defined according to the tangible effects of the attack. It is not judged in terms of the larger goals of the perpetrators. The definition of a successful attack depends on the type of attack - the key question is whether or not the attack took place. If a case has multipleattack types, it is successful if any of the attack types are successful, with theexception of assassinations, which are only successful if the intended target is killed." $

Casualties are reported using multiple variables in the GTD. This analysis includes numeric variables *nkill (Total Number of Fatalities) and nwound (Total Number of Injured).*

Yijia Qi (e-mail: yqi12@stevens.edu)
Nishanth Mudkey (e-mail: nmudkey@stevens.edu)
Tianyu Sheng (e-mail: tsheng2@stevens.edu)

TABLE I
PREDICTOR VARIABLES SUCCESS AND CASUALTIES

Predictor Crime Variable	Description	Notation
Success	Success of a terrorist strike is defined according to the tangible effects of the attack. Success is not judged in terms of the larger goals of the perpetrators.	1 = "Yes" 0 = "No"
Casualties	Numeric variables nkill (Total Number of Fatalities) and nwound (Total Number of Injured) added together.	***nkill***, ***nwound*** are numbered as reported. They are added together at each sequence to create the new numeric variable ***ncasualties***.

nkill — *This field stores the number of total confirmed fatalities for the incident. The number includes all victims and attackers who died as a direct result of the incident [2] .*

nwound — *This field records the number of confirmed non-fatal injuries to both perpetrators and victims. It follows the conventions of the "Total Number of Fatalities" field described above [2] .*

By feature crossing, a new predictor variable **ncasualties** is created using existing numeric variables nkill and nwound above as *df['ncasualties'] = df['nkill'] + df['nwound'].*

II. Background

A. Description of the data-set

Based on the Final Project Proposal Report and Mid-stage report, a comprehensive data set description is presented below.

Data and Codebook: The current GTD is a culmination of "several phases of data collection efforts" that relied on publicly available information sources including but not limited to media articles, electronic news archives, existing data sets, books and journals etcetera. Multiple data collection agencies have contributed to the GTD since the 1970s until today. The GTD is subject to continuing quality control tests and stabilization efforts that ensure data consistency. Inclusion Criteria and Variables [2] provides an overview on the data collection methodology, definition of terrorism and other inclusion criteria used, and other determinations concerning the GTD. It also describes the variables in the GTD and interprets the possible values of the variables — variable categories include the GTD ID, incident date, incident location, incident information, attack information, target/victim information, perpetrator information, perpetrator statistics, claims of responsibility, weapon information, casualty information, consequences, kidnapping/hostage taking information, additional information, and source information.

License Agreement Restrictions: The GTD was downloaded from the official website. Relevant communication details were provided electronically on a web form and an official email with an End User License Agreement provided a link to download the database in .xlsx format. Important restrictions and limitations are added to the appendix for reference.

The following are some analytical details extracted from the database,

- The global terrorism database (GTD) consists of 191464 rows × 135 columns; there are columns with the following data types in the data-set: int64, float64, object (or mixed data types).

```
<class 'pandas.core.frame.DataFrame'>
   RangeIndex: 191464 entries, 0 to
   191463 Columns: 135 entries,
   eventid to related dtypes: float64
   (53), int64(24), object(58) memory
   usage: 197.2+ MB
```

Code Snippet 1. df.info()

- Each column of the data-set represents a different crime variable. Database variables in the GTD cover different information about terrorist events such as — GTD event ID and Date, Incident Information, Incident Location, Attack Information, Weapon Information, Target/Victim Information, Perpetrator Information, Casualties and Consequences, Additional Information and Sources.
- The database contains a total of 14549883 missing values. Using pandas, a python data analysis library, columns with missing values were identified and the relative frequency of missing (missing values per column) was calculated.

Other important observations about the global terrorism database (GTD),

- Data collected in the GTD is of the highest quality
- The GTD is a relatively self-contained database of information
- Global terrorism data in the GTD is labeled with consistently
- The database covers terror events during the years 1970 to 2018
- Two major variables types, numerical and categorical, are included in the data set

A table with all the data variables (135) is added to the appendix for reference.

B. Related work

Prediction of terrorism activities is a popular topic in machine learning. Researchers use a variety of algorithms to build models and strive to find an algorithm with better performance in predicting related events. The following paragraphs talk about similar work carried out before.

1) GTD : [3] Uses data exploratory analysis and classification models, decision trees, and random forests to visualize terror data and predict possible terrorist attacks respectively. Prediction results show that Middle East & North Africa and South Asia are prone to future terror attacks with higher probabilities. Results also show that bombs and explosives have a higher probability of use in future attacks followed by armed assault. Classification algorithms used and implemented (Decision Tree and Random Forest) produce almost the same probabilistic results with 75.45% to 90.414% assertiveness. In [4], researchers analyze the GTD data-set and predict crime variables such as attacker groups and success along with an additional variable to understand the influence of weather on crime. Algorithms SVM (support vector machine), Random Forest, and Logistic Regression are implemented and compared to arrive at the conclusion that different algorithms suit different types of prediction problems.

Researchers in [5] focus on the specific use of Convolutional Neural Networks (CNNs) for long-term time series prediction of terrorist event data. They compare the performance of classical prediction methods, i.e., Naïve estimators, Averaging and Smoothing, Linear Regression, Auto-regressive Moving Average Models for time-series prediction of crime data with that of CNNs. Results demonstrate that CNNs make a reasonable tool for uni-variate long-term prediction of terror events. A similar study conducted in [6] compares the implementation of five different algorithms (Single-layer Neural Network, Five-layer Deep Neural Network, and three traditional machine learning algorithms, i.e., Logistic Regression, SVM, and Naïve Bayes) to predict different factors that contribute to terrorist events. Results illustrate that Deep Neural Networks (DNNs) exhibit a superior performance in comparison with remaining algorithms.

2) Other Crime Data-sets: Other crime-data studies also provide some insights for the current project. Work carried out in [7] using the Chicago crime data-set mainly revolves around predicting the types of crime which may happen if the location is known in advance. Researchers built machine learning models using algorithms such as K-Neighbors Algorithm, Gaussian NB, Multinomial NB, Bernoulli NB, SVC, and Decision Tree Classifier and found that k-nearest neighbors algorithm (k-NN) performs with the highest amount of accuracy (78.70%) for the given problem. Another research study [8] based on the Chicago crime data-sets concludes that the RandomForestClassifier algorithm performs the best in predicting "Per Capita Violent Crimes". It also notes that some common features with high importance scores (such as the number of people below the poverty line, the percentage of people who cannot speak English, and the number of people present in urban areas, etcetera) can be good indicators to predict future crimes. Another paper [9] explores the utilization of data mining techniques to detect crime patterns in Cheltenham by exploiting information about the percentage of different crime incidents. Researchers draw correlations between different areas and crime types; they use Naïve Bayes algorithm to determine areas that are vulnerable to crime.

It is clear that similar works have been carried out before to build machine learning models to predict patterns and trends of crime incidents. Machine learning research on crime datasets has deepened our understanding of crime by enabling researchers to derive significant insights using several algorithmic techniques. As mentioned in the Final Project Proposal Report, "segregating and correlating different data variables has enabled researchers to draw significant insights; pattern classification and regression have enabled prediction of future crime rates with useful accuracies."

3) Other Sources: The project also uses research sources [10], [11], [12], [13], [14], [15], [16]. Current trends and leading-edge methods and algorithms used in GTD analysis and prediction have been explored fully. In addition, some public Kaggle examples [17], [18], [19], [20], [21] act as implementation examples for similar data exploratory analyses conducted before on the same database (GTD).

C. Tools used

- **Google Colaboratory** is a web-based interactive computational environment for creating Jupyter notebook documents.
- **Pandas** is a Python software library that is used for data manipulation and analysis. It offers data structures and operations to manipulate numerical tables.
- **NumPy** is a Python software library used for manipulating large, multidimensional arrays.
- **Matplotlib** is a plotting library for the Python programming language and its mathematical extension NumPy. **Seaborn** is a Python data visualization library based on Matplotlib.
- **Scikit-learn** is a standard machine learning library in Python featuring classification, regression, and clustering algorithms.
- **Folium** is a Python library used for visualizing spatial data in an interactive manner.

III. Methodology

A. Data Set Preparation & Preprocessing

1) Data Set Preparation: The global terrorism database (GTD) was downloaded from the official website [1] in (.xlsx) format. The file was first converted to (.csv) format and then used for implementation. The data downloaded was already processed, labeled and clean.

2) Preprocessing: A standard methodology was used to preprocess data for algorithmic implementation in the following order — select feature columns (either by intuition or random selection or by identifying features with high importance scores), treat missing values, and encode data. Two different implementations using different feature selection, missing value treatment, and data encoding methods were employed. Table 1 provides a brief summary. Please find more details about the implementation algorithms used in section 3.3 below.

NOTE: Exploratory Data Analysis or Data Visualization did not require any preprocessing as the data downloaded was already labeled and clean.

TABLE II
PREPROCESSING TECHNIQUES USED FOR DIFFERENT IMPLEMENTATION METHODS

Technique	Implementation 1	Implementation 2
Feature Selection	Random	High Importance Scores
Missing Value Treatment	Scikit-learn's SimpleImputer	Pandas & NumPy
Data Encoding	OneHotEncoder	LabelEncoder

B. Data Visualization

Several data visualization plots were generated to understand the global terrorism database (GTD). Matplotlib and seaborn were predominantly used to generate plots. Presented below are a set of plots that try to make sense of different data variables related to different data sections of the database including but not limited to GTD event ID and Date, Incident Information, Incident Location, Attack Information, Weapon Information, Target/Victim Information, Perpetrator Information, Casualties and Consequences, Additional Information and Sources. Plots generated attempt to comprehensively encapsulate, explore, and extract inferences from the data available. Please find additional data visualizations in the appendix and in the Python file (.ipynb) submitted with this report.

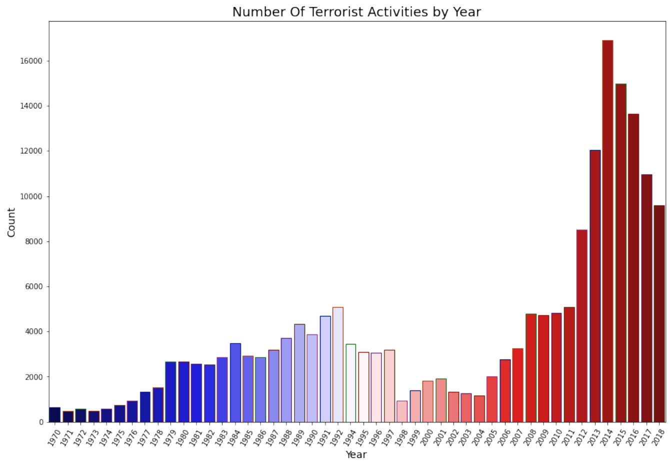

Fig. 1. Number of Terrorist Activities by Year

Inference: Terrorism has only increased through the years.

- 1970 - 1992: increase in terror activity
- 1994 - 2004: fluctuating activity
- 2004 - 2014: steep rise in terror activity
- 2015 - 2018: decrease in terror activity in relation to the period (2004 - 2014)

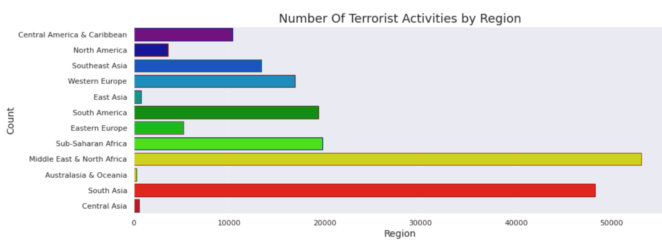

Fig. 2. Number of Terrorist Activities by Region

Inference: The following regions are prone to terror the most,

- Middle East & North Africa
- South Asia
- Sub-Saharan Africa
- South America

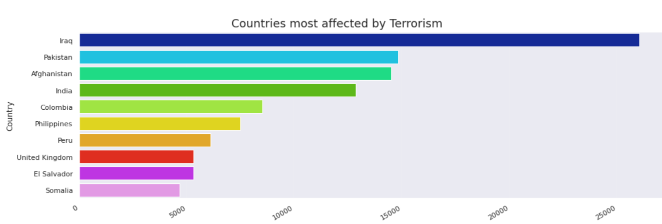

Fig. 3. Countries most affected by Terrorism

Inference: The following countries have suffered the most,

- Iraq
- Pakistan
- Afghanistan
- India
- Colombia

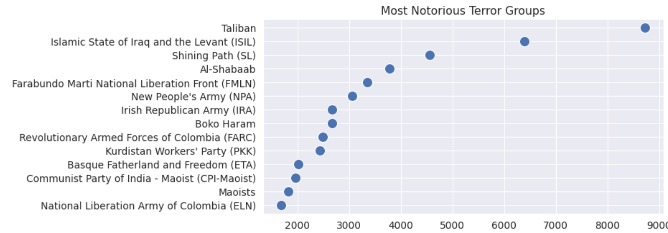

Fig. 4. Most Notorious Terror Groups

Inference: The following are the most notorious terrorist organizations,

- Taliban
- Islamic State of Iraq and the Levant (ISIL)
- Shining Path (ISL)
- Al-Shabaab
- Farabundo Marti National Liberation Front (FMLN)
- New People's Army
- Irish Republican Army (IRA)

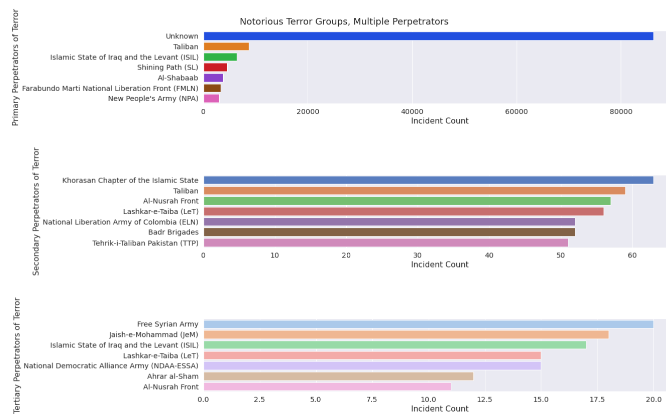

Fig. 5. Notorious Terror Groups, Multiple Perpetrators

Background: These plots apply individually (standalone) and also when responsibility for an attack is attributed to more than one perpetrator. Primary perpetrators of terror are terror groups that are most notorious for organizing a terror attack. Secondary and tertiary perpetrator groups are terror groups that either aided or contributed to an attack.

NOTE: Multiple perpetrator group attributions do not necessarily indicate that perpetrator groups collaborated to execute an attack. This could represent competing attributions, competing claims of responsibility, competing accusations, or a combination of these.

Inference: Most notorious groups of terror,

- Taliban
- Islamic State of Iraq and the Levant (ISIL)
- Shining Path (ISL)
- Al-Shabaab
- Farabundo Marti National Liberation Front (FMLN)
- New People's Army (NPA)
- Khorasan Chapter of the Islamic State
- Al-Nusrah Front
- Lashkar-e-Taiba (LeT)

- Badr BrigadesNational Liberation Army of Colombia
- National Democratic Alliance Army (NDAA-ESSA)

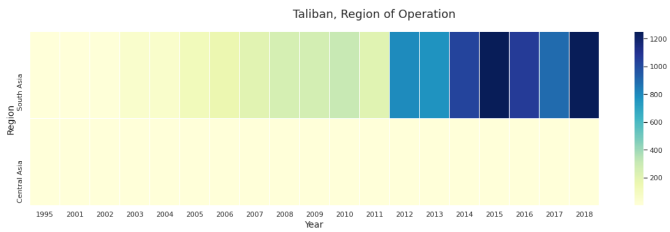

Fig. 6. Taliban, Region of Operation

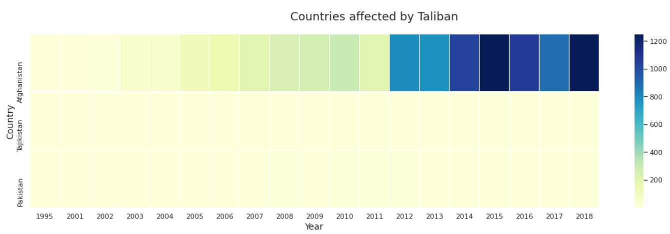

Fig. 7. Countries affected by Taliban

Inference: Taliban operates in South Asia. It predominantly wages terror in the country of Afghanistan.

NOTE: Similar heatmaps were generated for other regions.

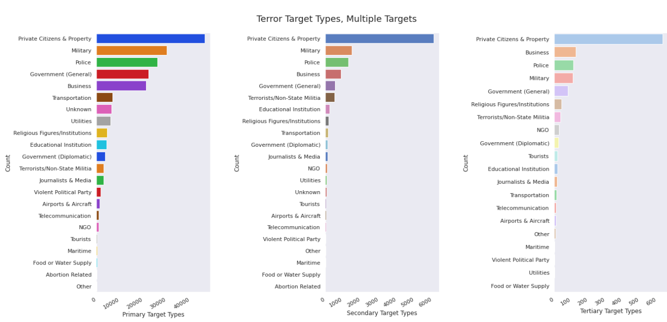

Fig. 8. Terror Target Types, Multiple Targets

Background: The target/victim type field captures the general type of target/victim. When a victim is attacked specifically because of his or her relationship to a particular person, such as a prominent figure, the target type reflects that motive. For example, if a family member of a government official is attacked because of his or her relationship to that individual, the type of target is "government." This variable consists of 22 different categories.

Inference: Primary Terror Targets,

- Private Citizens & Property
- Military
- Police
- Government (General)
- Business
- Transportation

Secondary and Tertiary Terror Targets are second and third target types in terror attacks or incidents. The field target type contains information on both intended targets and incidental bystanders, and therefore, intentionality should be carefully considered in each case.

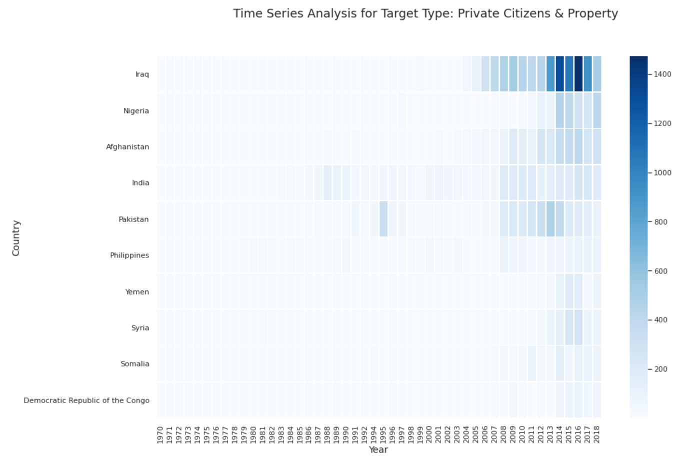

Fig. 9. Time Series Analysis for Target Type: Private Citizens & Property

Inference: Private Citizens & Property in the following countries suffered the maximum amount of damage,

- Iraq
- India
- Pakistan
- Nigeria
- Afghanistan

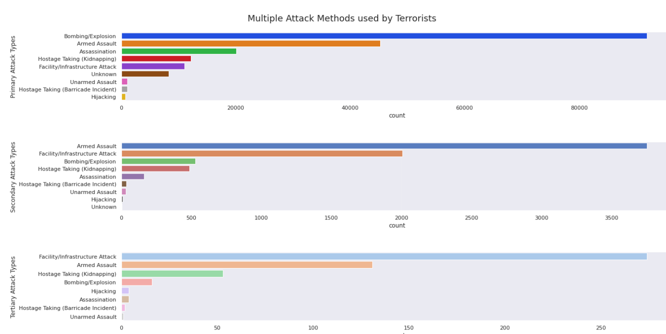

Fig. 10. Multiple Attack Methods used by Terrorists

Background: This field captures the general method of attack and often reflects the broad class of tactics used. It consists of nine categories, given below. Up to three attack types can be recorded for each incident. Typically, only one attack type is recorded for each incident unless the attack is comprised of a sequence of events. When multiple attack types may apply, the most appropriate value is determined based on the hierarchy below.

Attack Type Hierarchy:

1) Assassination
2) Hijacking
3) Kidnapping
4) Barricade Incident
5) Bombing/Explosion
6) Armed Assault
7) Unarmed Assault
8) Facility/Infrastructure Attack

9) Unknown

Inference: As can be inferred from the plots above, the most popular attack types used by terrorists are (in descending order),

- Bombing/Explossion
- Armed Assault
- Assassination
- Hostage Taking (Kidnapping)
- Facility/Infrastructure Attack
- Unarmed Assault

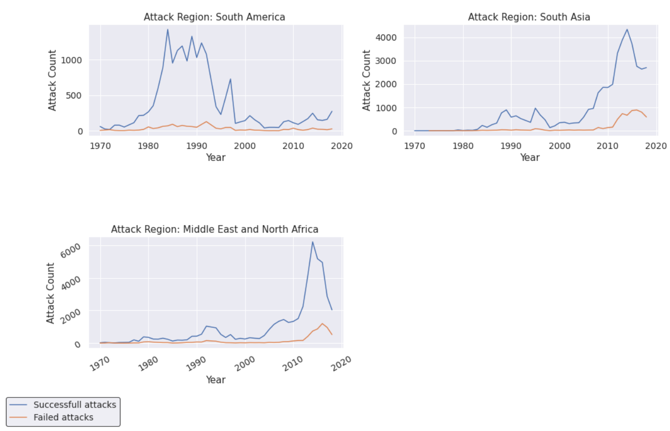

Fig. 11. Trends in Success and Failure by Regions

Inference: Plots show no clear trend through time. South Asia and Middle & North Africa display a strong increase in terror activity from the year 2005 and beyond. South America has a similar increase during the years 1970 to 1994.

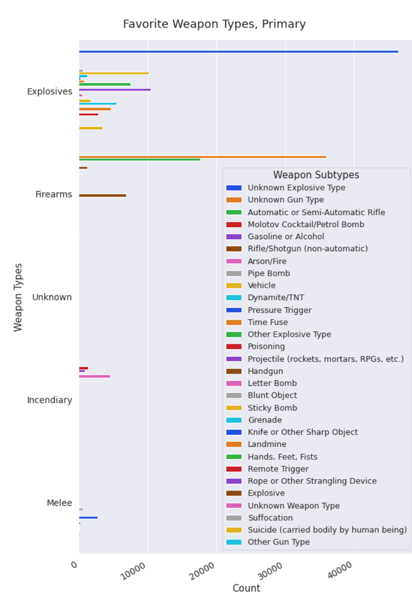

Fig. 12. Favorite Weapon Types, Primary

Inference: The plots above, i.e., Favorite Weapon Types (Primary, Secondary, and Tertiary) are self-evident. For each Weapon Type (Primary, Secondary, and Tertiary), weapon subtypes have been depicted cumulatively on the Y-axes; also the count for each weapon subtype is depicted in the X-axes.

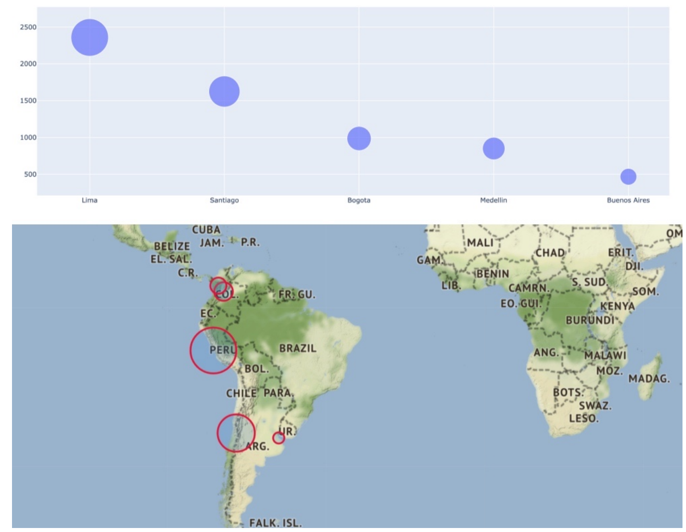

Fig. 13. Top 5 cities in South America who has seen the terrorist acts the most

Inference: Highly affected cities in South America,

- Lima
- Santiago
- Bogota
- Medellin
- Buenos Aires

NOTE: Similar graphs were generated for the top two terror-stricken regions, Middle East & North Africa and South Asia and it was found that Baghdad, Mosul, Istanbul, Kirkuk, and Beirut and Karachi, Quetta, Kabul, Peshawar, and Srinagar were the most affected cities in those regions respectively.

C. Preliminary Results — Algorithms Selected & Implemented

1) Implementation 1:

a) Preprocessing Method Used:

- 10 feature columns ["iyear", "extended", "region", "country", "attacktype1", "targtype1", "weaptype1", "nperps", "nkill", "nhostkid"] were used against the predictor column "success"; feature columns were selected at random — intuition or common sense
- Missing values were filled using Scikit-learn (SimpleImputer) and data arrays were transformed accordingly
- Data was encoded using OneHotEncoder from Scikit-learn

b) RandomForestRegressor:

- A RandomForestRegressor was used with default parameters to train the model
- The model achieved a relatively low accuracy of 36.64%
- No attempts were made to tune the hyperparameters as the computational capacity (time and cost) required to implement the RandomForestRegressor algorithm on the GTD was relatively expensive

c) RandomForestClassifier:

- A RandomForestClassifier was used with default parameters to train the model
- The model has 92.14% (approximate) accuracy
- A classification report was generated

Classification Table:

	precision	recall	f1-score	support
0	0.73	0.48	0.58	4353
1	0.94	0.98	0.96	33940
accuracy			0.92	38293
macro avg	0.84	0.73	0.77	38293
weighted avg	0.91	0.92	0.91	38293

Accuracy: 0.9213955553234272

Fig. 14. Classification Report — RandomForestClassifier

- r-squared value was calculated as 0.2227
- Mean absolute error was calculated as 0.0783
- Mean squared error was calculated as 0.0783
- A RandomGridSearch was conducted (for estimators 10, 20, 30, 40, 50, 60, 70, 80, and 90) to identify the best number of estimators for the model as 60 estimators
- Model accuracies were highest for 60, 70, 90, 100 estimators

d) kNN — k-nearest neighbors algorithm:

- k-nearest neighbors algorithm was used with default parameters to train the model
- The model has 91.44% (approximate) accuracy
- A classification report was generated

Classification Table:

	precision	recall	f1-score	support
0	0.68	0.46	0.55	4353
1	0.93	0.97	0.95	33940
accuracy			0.91	38293
macro avg	0.81	0.72	0.75	38293
weighted avg	0.91	0.91	0.91	38293

Accuracy: 0.914396887159533

Fig. 15. Classification Report — kNN

- r-squared value was calculated as 0.1504
- Mean absolute error was calculated as 0.0856
- Mean squared error was calculated as 0.0856
- An attempt was made to check other k-NN neighbor ranges but the process took beyond three hours and had to be terminated. GridSearchCV is a computationally expensive process for bigger databases such as the GTD.

e) NN — Multi-layer Perceptron:

- A multi-layer perceptron algorithm was used with two hidden layers, SGD (stochastic gradient descent), and logistic sigmoid function to train the model
- The model has 92.09% (approximate) accuracy
- A classification report was generated

- r-squared value was calculated as 0.2146
- Mean absolute error was calculated as 0.0791
- Mean squared error was calculated as 0.0792
- No attempts were made to tune the hyperparameters

Classification Table:

	precision	recall	f1-score	support
0	0.82	0.39	0.53	4353
1	0.93	0.99	0.96	33940
accuracy			0.92	38293
macro avg	0.87	0.69	0.74	38293
weighted avg	0.91	0.92	0.91	38293

Accuracy: 0.9208732666544799

Fig. 16. Classification Report — Neural Network

2) Implementation 2:

a) Preprocessing Method Used:

- 19 feature columns were used against the predictor column "has_casualties"; feature columns were selected at random and were then put to test to calculate feature importance.
- Missing values were filled using Numpy
- Data was encoded using LabelEncoder from Scikit learn

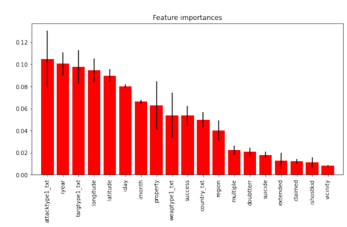

Fig. 17. Features plot

The features plot shows the significance of each feature on predicting whether there will be casualties or not. Spatio-temporal variables seem to be very dominantly present. Time variables will not be included as the model may be supplied with meaningless inputs into the future (as it is only applicable until 2018). To avoid overfitting the model, only features with an accuracy score at or beyond 0.05 are kept.

In all the models below the top 9 features ['attacktype1_txt', 'targtype1_txt', 'longitude', 'latitude', 'property', 'weaptype1_txt', 'success', 'country_txt', 'region'] were used for implementation.

b) RandomForestRegressor:

- The RandomForestRegressor algorithm was used with default parameters to train the model
- The model achieved an accuracy of 54.83%
- No attempts were made to tune the hyperparameters as the computational capacity (time and cost) required to implement the RandomForestRegressor algorithm on the GTD was relatively expensive

c) RandomForestClassifier:

- The classifier was used with n_estimators=20 to train the model
- The model has approximately 85.32 % accuracy
- A Classification Report was generated

Classification Table:

	precision	recall	f1-score	support
0	0.81	0.82	0.82	22991
1	0.88	0.87	0.88	34449
accuracy			0.85	57440
macro avg	0.85	0.85	0.85	57440
weighted avg	0.85	0.85	0.85	57440

Accuracy: 0.8532381615598886

Fig. 18. Classification Report — RandomForestClassifier

d) kNN — k-nearest neighbors algorithm:

- The kNN algorithm was used with default parameters to train the model
- The model achieved an accuracy of 76.96%
- A classification report was generated
- r-squared value was calculated as 0.0421
- Mean absolute error was calculated as 0.2304
- Mean squared error was calculated as 0.2304

Classification Table:

	precision	recall	f1-score	support
0	0.73	0.68	0.70	23134
1	0.79	0.83	0.81	34306
accuracy			0.77	57440
macro avg	0.76	0.76	0.76	57440
weighted avg	0.77	0.77	0.77	57440

Accuracy: 0.7696030640668524

Fig. 19. Classification Report — kNN

e) NN — Multi-layer Perceptron:

- A multi-layer perceptron algorithm was used with two hidden layers, SGD (stochastic gradient descent), and logistic sigmoid function to train the model
- The model has 59.72% (approximate) accuracy
- A classification report was generated
- Warning (Error)
- r-squared value was calculated as -0.6743
- Mean absolute error was calculated as 0.4027
- Mean squared error was calculated as 0.4027

Classification Table:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	23134
1	0.60	1.00	0.75	34306
accuracy			0.60	57440
macro avg	0.30	0.50	0.37	57440
weighted avg	0.36	0.60	0.45	57440

Fig. 20. Classification Report — NN

IV. Conclusion & Comparison of Results

- Data visualizations generated provide conclusive inferences that can be used to report significant findings
- We can conclude from the table 3 on the right that the RandomForestClassifier algorithm works best to predict terror success and also the casualties from the GTD database
- Implementation of some algorithms for tuning the hyper-parameters can be computationally expensive and time-consuming
- Multilayer perceptrons can be tuned in a variety of ways to achieve the required results but the process may be tedious if done manually and computationally expensive if automated; the same is the case with the kNN algorithm with the exception of the changing parameter being the number of nearest neighbors or k-neighbors.
- There is scope for implementing more algorithms and also for extending and improving the current models

A. Future Research

- It is concluded from this project that the scope of the current results is limited. Data exploratory analysis can be further elaborated and extended and more machine learning algorithms can be adopted and implemented to build efficient models. In addition, different feature variable can be used from the GTD to study other crime variables.
- Also, for data visualization, more animated maps of terrorist activities can be generated to have a better observation of the trend and pattern of the places where terror events occur. With the prediction of casualties, we can classify different magnitudes of casualties as hundreds or thousands and further investigate the relationship between different attack types, weapon types, terror groups with the number of people who are injured. Specific terror groups can be studied, especially whose terrorism activities are reducing in the recent years, to see what factors may affect the decrease of activities. In future researches, correlations among variables, terrorist activities with media impact or weather conditions or other terrorism-affecting data sets can be analyzed in unison to predict broader aspects of terrorism in general.

TABLE III
Comparison of Results

Algorithm	Implementation1 Accuracy	Implementation2 Accuracy
RandomForestRegressor	36.64	54.83
RandomForestClassifier	92.14	85.32
k-NN	91.44	76.96
Multilayer Perceptron	92.09	59.72

Appendix A

GLOBAL TERRORISM DATABASETM END USERLICENSE AGREEMENT

RESTRICTIONS. Licensee agrees NOT to:
a. Publicly post or display any part of GTD without express written permission by Licensor other than quoting the GTD in Licensee's reports, papers, or analyses;
b. Sell, license, sublicense or otherwise distribute any part of the GTD to third parties whether for cash or any other considerations or none;
c. Modify, hide, delete or interfere with any notices that are included in the GTD;
d. Use the GTD to draw conclusions about the official legal status or criminal record of an individual, or the status of a criminal or civil investigation;
e. Interfere with or disrupt the GTD, its website or servers and networks connected to the GTD website;
f. Interfere with or disrupt the Licensor's business or any website of the Licensor involved in the marketing and/or distribution of the GTD or;
g. Use robots, spiders, crawlers, automated devices and similar technologies to screen-scrape the site or to engage in data aggregation or indexing of the GTD other than to produce search results on a publicly available search engine.

LICENSEE'S RESPONSIBILITIES:
a. All information sourced from the GTD should be acknowledged and cited as follows: "National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland. (2019). The Global Terrorism DatabaseTM."
b. Licensee agrees to acknowledge any copyrightable materials with a copyright notice "Copyright University of Maryland 2019".
c. Any modifications Licensee make to the GTD for published analysis must be clearly documented and must not misrepresent analytical decisions made by START.

Appendix B

MORE VISUALIZATIONS

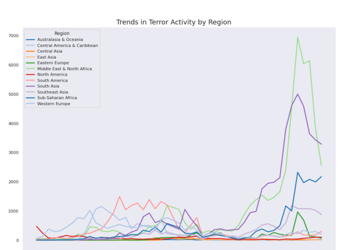

Fig. 21. Trends in Terror Activity by Region

Inference: This plot reinforces inferences drawn from the previous ones and provides the following insights in addition,

- South America dominated the word of terror during the period (1980 - 1993); it saw a steep decrease in terror activity following this period. This is an interesting trend because the region has managed to reduce terror activity over the years.
- Middle East & North Africa and South Asia see a sharp, amplified increase in terror activity during the period (2000 - 2014)
- Sub-Saharan Africa sees periods of fluctuating terror over the years with a sharp increase (on average) in terror during the period (2005 - 2018)

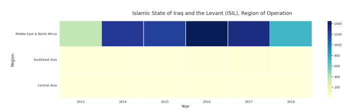

Fig. 22. Islamic State of Iraq and the Levant (ISIL), Region of Operation

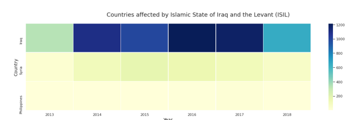

Fig. 23. Countries affected most by Islamic State of Iraq and the Levant (ISIL)

Inference: Islamic State of Iraq and the Levant (ISIL) operates in Middle East & North Africa. It predominantly wages terror in Iraq and Syria.

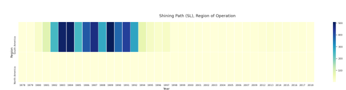

Fig. 24. Shining Path (ISL), Region of Operation

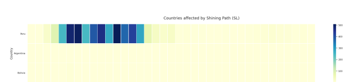

Fig. 25. Countries affected most by Shining Path (ISL)

Inference: Shining Path (ISL) operates in South America. It predominantly wages terror in Peru.

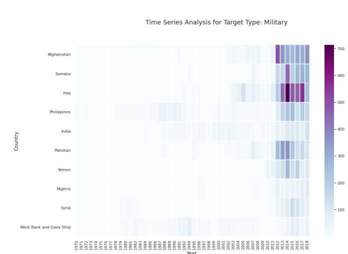

Fig. 26. Time Series Analysis for Target Type: Military

Inference: Military groups targeted the most by terror groups belong to the following countries,

- Afghanistan
- Iraq

- Somalia
- Philippines
- Pakistan
- India
- Yemen

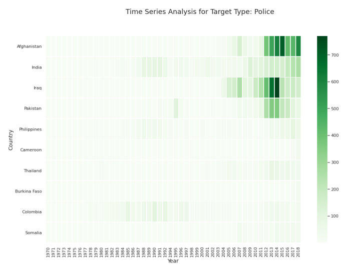

Fig. 27. Time Series Analysis for Target Type: Police

Inference: Police groups targeted the most by terror groups belong to the following countries,

- Afghanistan
- Iraq
- Pakistan
- India
- Colombia

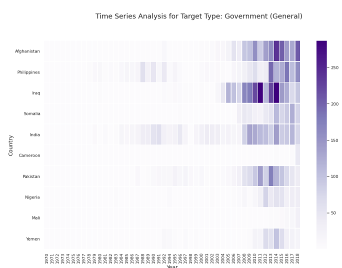

Fig. 28. Time Series Analysis for Target Type: Government (General)

Inference: Governments of the following countries suffered the most damage due to terrorism,

- Afghanistan
- Philippines
- Iraq
- Yemen
- Somalia

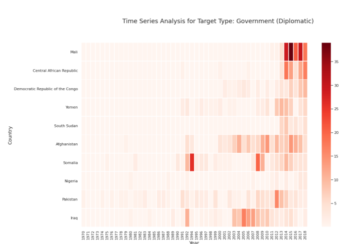

Fig. 29. Time Series Analysis for Target Type: Government (Diplomatic)

Inference: Attacks carried out against foreign missions, including embassies, consulates, etc. in the following countries make them the most vulnerable to terror attacks on Government personnel or property etc.

- Mali
- Central African Republic
- Democratic Republic of the Congo
- Yemen
- Somalia
- South Sudan

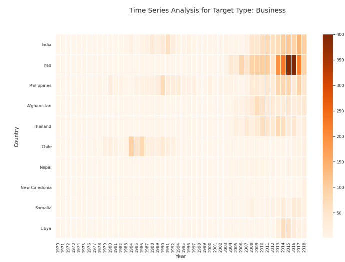

Fig. 30. Time Series Analysis for Target Type: Business

Inference: Businesses in the following countries suffered the most due to terrorism,

- India
- Iraq
- Philippines
- Afghanistan
- Thail
- Chile

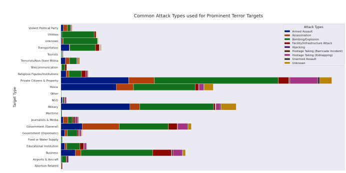

Fig. 31. Common Attack Types used for Prominent Terror Targets

Inference: The plot is self-evident. For example, it can be inferred that the three most popular attack types on "Private Citizens & Property" are,

- Bombing/Explosion
- Armed Assault
- Hostage taking (Kidnapping)

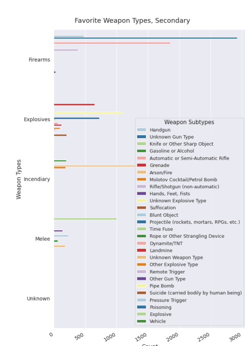

Fig. 32. Favorite Weapon Types, Secondary

APPENDIX C

CRIME DATA VARIABLES

Sr.No	Column	Name (Meaning)	Data Type	GTD Variable
1	A	GTD ID or Incident ID	Numeric Variable	**eventid**
2	B	Year	Numeric Variable	**iyear**
3	C	Month	Numeric Variable	**imonth**
4	D	Day	Numeric Variable	**iday**
5	E	Approximate Date	Text Variable	**approxdate**
6	F	Extended Incident?	Categorical Variable	**extended**
7	G	Date of Extended Incident Resolution	Numeric Date Variable	**resolution**
8	H	Country	Categorical Variable	**country**
9	I	Country	Categorical Text Variable	**country_txt**
10	J	Region	Categorical Variable	**region**
11	K	Region	Categorical Text Variable	**region_txt**
12	L	Province/Administrative Region/State	Text Variable	**provstate**
13	M	City	Text Variable	**city**
14	N	Latitude	Numeric Variable	**latitude**
15	O	Longitude	Numeric Variable	**longitude**
16	P	Geocoding Specificity	Categorical Variable	**specificity**
17	Q	Vicinity	Categorical Variable	**vicinity**
18	R	Location Details	Text Variable	**location**
19	S	Incident Summary	Text Variable	**summary**
20	T	Inclusion Criteria 1	Categorical Variable	**crit1**
21	U	Inclusion Criteria 2	Categorical Variable	**crit2**
22	V	Inclusion Criteria 3	Categorical Variable	**crit3**
23	W	Doubt Terrorism Proper?	Categorical Variable	**doubtterr**
24	X	Alternative Designation	Categorical Variable	**alternative**
25	Y	Alternative Designation	Categorical Text Variable	**alternative_txt**
26	Z	Part of Multiple Incident	Categorical Variable	**multiple**
27	AA	Successful Attack	Categorical Variable	**success**
28	AB	Suicide Attack	Categorical Variable	**suicide**
29	AC	First Attack Type	Categorical Variable	**attacktype1**
30	AD	First Attack Type	Categorical Text Variable	**attacktype1_txt**
31	AE	Second Attack Type	Categorical Variable	**attacktype_2**
32	AF	Second Attack Type	Categorical Text Variable	**attacktype2_txt**
33	AG	Third Attack Type	Categorical Variable	**attacktype3**
34	AH	Third Attack Type	Categorical Text Variable	**attacktype3_txt**
35	AI	Target/Victim Type 1	Categorical Variable	**targtype1**
36	AJ	Target/Victim Type 1	Categorical Text Variable	**targtype1_txt**
37	AK	Target/Victim Subtype 1	Categorical Variable	**targsubtype1**
38	AL	Target/Victim Subtype 1	Categorical Text Variable	**targsubtype1_txt**
39	AM	Corporate Entity/Government Agency Targeted	Text Variable	**corp1**
40	AN	Specific Target/Victim	Text Variable	**target1**
41	AO	Nationality of Target/Victim	Categorical Variable	**nalty1**
42	AP	Nationality of Target/Victim	Categorical Text Variable	**nalty1_txt**
43	AQ	Second Target/Victim Type	Categorical Variable	**targtype2**
44	AR	Second Target/Victim Type	Categorical Text Variable	**targtype2_txt**
45	AS	Second Target/Victim Subtype	Categorical Variable	**targsubtype2**
46	AT	Second Target/Victim Subtype	Categorical Text Variable	**targsubtype2_txt**
47	AU	Name of Second Entity	Text Variable	**corp2**
48	AV	Second Specific Target/Victim	Text Variable	**target2**
49	AW	Nationality of Second Target/Victim	Categorical Variable	**natlty2**
50	AX	Nationality of Second Target/Victim	Categorical Text Variable	**natlty2_txt**
51	AY	Third Target/Victim Type	Categorical Variable	**targtype3**
52	AZ	Third Target/Victim Type	Categorical Text Variable	**targtype3_txt**
53	BA	Third Target/Victim Subtype	Categorical Variable	**targsubtype3**
54	BB	Third Target/Victim Subtype	Categorical Text Variable	**targsubtype3_txt**
55	BC	Name of Third Entity	Text Variable	**corp3**
56	BD	Third Specific Target/Victim	Text Variable	**target3**
57	BE	Nationality of Third Target/Victim	Categorical Variable	**natlty3**
58	BF	Nationality of Third Target/Victim	Categorical Text Variable	**natlty3_txt**
59	BG	Perpetrator Group Name	Text Variable	**gname**
60	BH	Perpetrator Sub-Group Name	Text Variable	**gsubname**
61	BI	Second Perpetrator Group Name	Text Variable	**gname2**
62	BJ	Second Perpetrator Sub-Group Name	Text Variable	**gsubname2**
63	BK	Third Perpetrator Group Name	Text Variable	**gname3**
64	BL	Third Perpetrator Sub-Group Name	Text Variable	**gsubname3**
65	BM	Motive	Text Variable	**motive**
66	BN	First Perpetrator Group Suspected/Unconfirmed?	Categorical Variable	**guncertain1**
67	BO	Second Perpetrator Group Suspected/Unconfirmed?	Categorical Variable	**guncertain2**
68	BP	Third Perpetrator Group Suspected/Unconfirmed?	Categorical Variable	**guncertain3**
69	BQ	Unaffiliated Individual(s)	Categorical Variable	**individual**
70	BR	Number of Perpetrators	Numeric Variable	**nperps**
71	BS	Number of Perpetrators Captured	Numeric Variable	**nperpcap**
72	BT	Claim of Responsibility	Categorical Variable	**claimed**
73	BU	Mode for Claim of Responsibility	Categorical Variable	**claimmode**
74	BV	Mode for Claim of Responsibility	Categorical Text Variable	**claimmode_txt**
75	BW	Second Group Claim of Responsibility?	Categorical Variable	**claim2**
76	BX	Mode for Second Group Claim of Responsibility	Categorical Variable	**claimmode2**
77	BY	Mode for Second Group Claim of Responsibility	Categorical Text Variable	**claimmode2_txt**
78	BZ	Third Group Claim of Responsibility?	Categorical Variable	**claim3**
79	CA	Mode for Third Group Claim of Responsibility	Categorical Variable	**claimmode3**
80	CB	Mode for Third Group Claim of Responsibility	Categorical Text Variable	**claimmode3_txt**
81	CC	Competing Claims of Responsibility?	Categorical Variable	**compclaim**
82	CD	Weapon Type	Categorical Variable	**weaptype1**
83	CE	Weapon Type	Categorical Text Variable	**weaptype1_txt**
84	CF	Weapon Sub-type	Categorical Variable	**weapsubtype1**
85	CG	Weapon Sub-type	Categorical Text Variable	**weapsubtype1_txt**
86	CH	Second Weapon Type	Categorical Variable	**weaptype2**
87	CI	Second Weapon Type	Categorical Text Variable	**weaptype2_txt**
88	CJ	Second Weapon Sub-Type	Categorical Variable	**weapsubtype2**
89	CK	Second Weapon Sub-Type	Categorical Text Variable	**weapsubtype2_txt**
90	CL	Third Weapon Type	Categorical Variable	**weaptype3**
91	CM	Third Weapon Type	Categorical Text Variable	**weaptype3_txt**
92	CN	Third Weapon Sub-Type	Categorical Variable	**weapsubtype3**
93	CO	Third Weapon Sub-Type	Categorical Text Variable	**weapsubtype3_txt**
94	CP	Fourth Weapon Type	Categorical Variable	**weaptype4**
95	CQ	Fourth Weapon Type	Categorical Text Variable	**weaptype4_txt**
96	CR	Fourth Weapon Sub-Type	Categorical Variable	**weapsubtype4**
97	CS	Fourth Weapon Sub-Type	Categorical Text Variable	**weapsubtype4_txt**
98	CT	Weapon Details	Text Variable	**weapdetail**
99	CU	Total Number of Fatalities	Numeric Variable	**nkill**
100	CV	Number of US Fatalities	Numeric Variable	**nkillus**
101	CW	Number of Perpetrator Fatalities	Numeric Variable	**nkillter**
102	CX	Total Number of injured	Numeric Variable	**nwound**
103	CY	Number of U.S. injured	Numeric Variable	**nwoundus**
104	CZ	Number of Perpetrators injured	Numeric Variable	**nwoundte**
105	DA	Property Damage	Categorical Variable	**property**
106	DB	Extent of Property Damage	Categorical Variable	**propextent**
107	DC	Extent of Property Damage	Categorical Text Variable	**propextent_txt**
108	DD	Value of Property Damage (in USD)	Numeric Variable	**propvalue**
109	DE	Property Damage Comments	Text Variable	**propcomment**
110	DF	Hostage or Kidnapping Victims	Categorical Variable	**ishostkid**
111	DG	Total Number of Hostages/ Kidnapping Victims	Numeric Variable	**nhostkid**
112	DH	Number of U.S. Hostages/ Kidnapping Victims	Numeric Variable	**nhostkidus**
113	DI	Hours of Kidnapping/ Hostage Incident	Numeric Variable	**nhours**
114	DJ	Days of Kidnapping/ Hostage Incident	Numeric Variable	**ndays**
115	DK	Country That Kidnappers/ Hijackers Diverted to	Text Variable	**divert**
116	DL	Country of Kidnapping/ Hijacking Resolution	Text Variable	**kidhijcountry**
117	DM	Ransom Demanded	Categorical Variable	**ransom**
118	DN	Total Ransom Amount Demanded	Numeric Variable	**ransomamt**
119	DO	Ransom Amount Demanded from U.S. Sources	Numeric Variable	**ransomamtus**
120	DP	Total Ransom Amount Paid	Numeric Variable	**ransompaid**
121	DQ	Ransom Amount Paid By U.S. Sources	Numeric Variable	**ransompaidus**
122	DR	Ransom Notes	Text Variable	**ransomnote**
123	DS	Kidnapping/Hostage Outcome	Categorical Variable	**hostkidoutcome**
124	DT	Kidnapping/Hostage Outcome	Categorical Text Variable	**hostkidoutcome_txt**
125	DU	Number Released/Escaped/Rescued	Numeric Variable	**nreleased**
126	DV	Additional Notes	Text Variable	**addnotes**
127	DW	First Source Citation	Text Variable	**scite1**
128	DX	Second Source Citation	Text Variable	**scite2**
129	DY	Third Source Citation	Text Variable	**scite3**
130	DZ	Data Collection	Text Variable	**dbsource**
131	EA	International-Logistical	Categorical Variable	**INT_LOG**
132	EB	International-Ideological	Categorical Variable	**INT_IDEO**
133	EC	International-Miscellaneous	Categorical Variable	**INT_MISC**
134	ED	International-Any of the above	Categorical Variable	**INT_ANY**
135	EE	Related Incidents	Text Variable	**related**

REFERENCES

[1] SMART, University of Maryland, "GTD|Global Terrorism Database." [Online]. Available: https://start.umd.edu/gtd/

[2] National Consortium for the Study of Terrorism and Responses to Terrorism (START), "CODEBOOK: INCLUSION CRITERIA AND VARIABLES," pp. 01–65, October 2019.

[3] E. L. H. Uriarte, A. Roman-Gonzalez, and A. Alva, "Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 04, 2020. [Online]. Available: DOI:10.14569/IJACSA.2020.0110474

[4] P. Agarwal, S. M, and C. S, "Comparison of machine learning approaches in the prediction of terrorist attacks," in *Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, August 2019.

[5] A. K. Jain, C. Grumber, P. Gelhausen, I. Häring, A. Stolz *et al.*, "A Toy Model Study for Long-Term Terror Event Time Series Prediction with CNN," *European Journal for Security Research* , 2020. [Online]. Available: https://doi.org/10.1007/s41125-019-00061-w

[6] M. I. Uddin, N. Zada, and F. Aziz, "Prediction of Future Terrorist Activities Using Deep Neural Networks," *Complexity* . [Online]. Available: https://doi.org/10.1155/2020/1373087

[7] A. Bharati and D. S. RA.K, "Crime Prediction and Analysis Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 05, no. 09, pp. 1–6, 2018.

[8] P. Yerpude and V. Gudur, "Predictive Modelling of Crime Dataset using Data Mining," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 07, no. 04, pp. 01–16, 2017. [Online]. Available: 10.5121/ijdkp.2017.7404

[9] R. Kumar and B. Nagpal, "Analysis and prediction of crime patterns using big data," *International Journal of Information Technology*, pp. 1–7, 2018. [Online]. Available: https://doi.org/10.1007/s41870-018-0260-7

[10] K. Singh, A. S. Chaudhary, and P. Kaur, "A Machine Learning Approach for Enhancing Defence Against Global Terrorism," *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–5, 2019.

[11] O. Kounadi, A. Ristea, A. Araujo, and M. Leitner, "A systematic review on spatial crime forecasting," *Crime Science*, vol. 9, no. 1, pp. 1–22, 2020. [Online]. Available: 10.1186/s40163-020-00116-7;https://dx.doi.org/10.1186/s40163-020-00116-7

[12] G. Saltos and M. Cocea, "An Exploration of Crime Prediction Using Data Mining on Open Data," *International Journal of Information Technology & Decision Making*, vol. 16, no. 05, pp. 1155–1181, 2017. [Online]. Available: 10.1142/s0219622017500250;https://dx.doi.org/10.1142/s0219622017500250

[13] J. Svobodová and J. Koláček, 2018.

[14] R. Berk, *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, ser. SpringerBriefs in Computer Science, S. Zdonik *et al.*, Eds. Springer, 2012.

[15] J. Semmelbeck and C. Besaw, "Exploring the Determinants of Crime-Terror Cooperation using Machine Learning," *Journal of Quantitative Criminology*, vol. 36, no. 3, pp. 527–558, 2020. [Online]. Available: 10.1007/s10940-019-09421-0;https://dx.doi.org/10.1007/s10940-019-09421-0

[16] G. Lafree, "The global terrorism database: accomplishments and challenges," *Perspect. Terrorism*, vol. 4, no. 1, pp. 2003–2016, 2010.

[17] [Online]. Available: https://www.kaggle.com/daveianhickey/how-to-folium-for-maps-heatmaps-time-analysis

[18] [Online]. Available: https://www.kaggle.com/idjtech/denver-crime-forecasting

[19] *Terrorism Around The World by*, vol. 316.

[20] [Online]. Available: https://www.kaggle.com/shelars1985/can-we-build-resilience-against-terrorism

[21] [Online]. Available: https://www.kaggle.com/yannisp/sf-crime-analysis-prediction