iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5319

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Predicting Success of Terrorist Attack and Extent of its Economic impact using Data Mining

Shashwat Vaibhav
Department of Information Technology, Maharaja Agrasen Institute of Technology,New Delhi, India

Abstract: ***Terrorism has been and continues to be one of the most sordid problems to deal with and a perpetual threat to humankind. As the analysis have been more data centric in today's world, data mining is surely one of the tools that can be used to interpolate information for future extrapolation and analysis. Despite the emergence of huge chunks of data, the amount of structured data is present sparsely. The Global terrorism database (GTD), maintained by University of Maryland is the biggest source of unstructured data related to Terrorism since 1972. This paper uses several aspects of data science and data mining to predict the success of a terrorist attack and the extent of property damage leading to the impact on economy. The study focuses on various regions of the world comprising of North America, Middle East, North Africa and South Asia. Comparison of the Prediction Model on original data and the oversampled data using SMOTE has been done.***
Keywords: Data Mining, GTD, Prediction Model, Oversampled data, SMOTE

I. INTRODUCTION

The definition and impact of terrorism needs no allocution. The technological advancements have led to an increase in the efficiency and efficacy of the terrorist organizations. And coming into terms with that, the analysis of their activity has to be smarter, accurate and precise. Thus, the impact and importance of data science and its several paradigms comes into the picture. The steps followed along the lines of data mining using its various tools and procedure helps get us valuable insights and pave the way for future extrapolation and prediction that might and will help in the fight against the terrorism.

The Global Terrorism Database (GTD) is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland [1].

The GTD contains information of over 150,000 terrorist attacks sub grouped on the basis of 135 attributes or features [1].

This paper shows the study on the terrorism database and predicting the success of the terrorist attack and the economic impact by predicting the extent of property damage, using classification algorithms such as Decision Tree and Random Forest.

The study comprises of these following regions:

A. North America (region code- 1)
B. South Asia (region code- 6)
C. Middle East and North Africa (region code- 10)

The paper also shows the effect of Synthetic Minority Oversampling Technique (SMOTE) on the results of the classification models.

II. DATA PREPARATION AND TOOLS

A. Tools

The Python programming language provides numerous libraries and tools for scientific computing, data science and statistical analysis. It can help parse and read several kinds of file formats. The data exploration and visualization is quite efficient in python's development environment such as Jupyter. Its plethora of libraries for statistical analysis, regression modelling and predictive classifier are useful and efficient.

B. Data Preparation

The Global Terrorism Database (GTD) is available publicly to search, browse, and download on its website [1].

The GTD consists of the terrorism incidents across the globe from 1970 to 2018. This database consists of records apportioned across 135 attributes. These attributes or columns are either real valued or nominal. Of these 135, attributes relevant to our purpose were selected. These have been categorized into two types Real Valued (R) and Categorical (C):

1) Iyear, imonth, iday, country, region, latitude, longitude, nkill (R).
2) Extended, success, suicide, attacktype1, weaptype1, multiple, individual, property, propextent (C).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

After loading the dataset into a dataframe, the heatmap for the attributes of the database showed that it consisted of attributes with missing or undefined values (Figure 1).

These missing values filled with appropriate values. For example, *nkill* was filled with mean of all the values of the column. Some of the records for which it was not feasible to fill any value, were discarded as they would not provide any help in exploration and prediction.

The dataset was further sub-divided into three for region number 1, 6 and 10. Each region were separately studied as the factors and scenarios leading to the incidents were quite different from other regions.

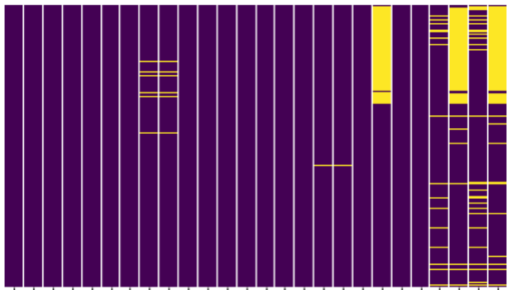


Figure 1: Heatmap of dataset. Yellow lines showing Missing values.

C. Analysis and Visualization.

The success rates and number of terror incidents from 1970 to 2018 in these 3 regions were found to be:

Table 1

Region	Success Rate	No. of Incidents
North America	0.83	3579
South Asia	0.87	48266
Middle East & North Africa	0.87	53110

The number of attacks associated with type of attack from the year 1970 to 2018 for these three regions can be seen in the plots (Figure 2, Figure 3, Figure 4).

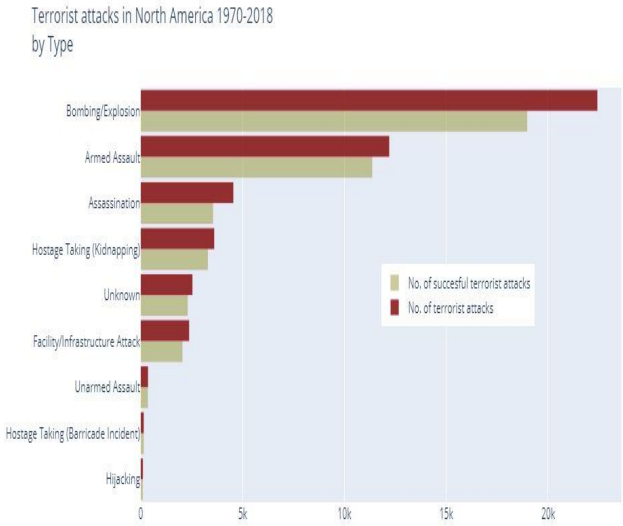


Figure 2: North America

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

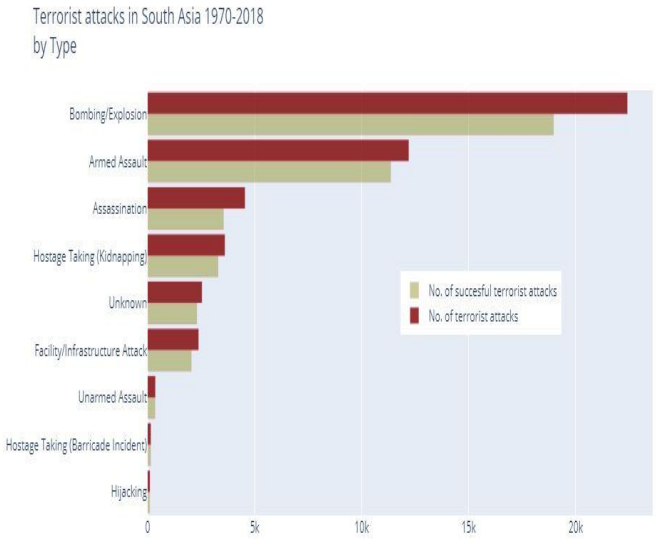


Figure 3: South Asia

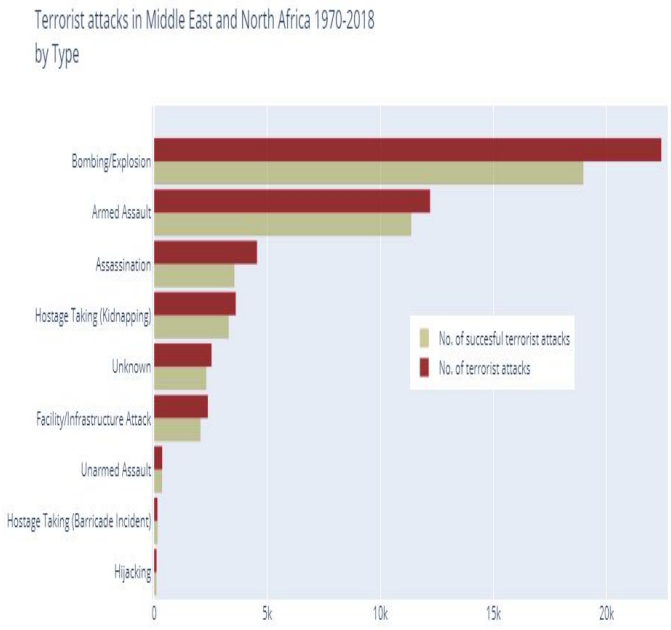


Figure 4: Middle East & North Africa

III. BUILDING PREDICTION MODELS

A. Classification Algorithm used

The classification algorithm used for the purpose were:

1) Decision Tree

2) Random Forrest

The main reason behind using these classifiers over others such as *Naïve Bayes* was the dataset comprised of too many factors and instances. For that reason, *Decision Trees* and *Random Forests* were optimal choice.

Also Decision tree suffers from *overfitting*, thus Random forest proved to be the best choice for classification.

B. Predicting Success of an Attack

The Decision Tree classifier when used on the dataset showed signs of overfitting with varying maximum depth of the tree.
For example, while predicting the success for South Asia Region, the Confusion matrices for Decision Tree classifier model for max-depth equal to 3, 4 and no depth were as follows:

```
Confusion Matrix with max depth =3:
  TN     FP
  FN     TP

[[ 541  822]
 [  59 9200]]
```

Figure 5: Decision Tree with max depth =3

```
Confusion Matrix with max depth =4:
  TN     FP
  FN     TP

[[ 597  766]
 [  78 9181]]
```

Figure 6: Decision Tree with max depth =4

```
Confusion Matrix with no max depth:
  TN     FP
  FN     TP

[[ 827  536]
 [ 597 8662]]
```

Figure 7: Decision Tree with no max depth

Clearly False Negative has increased and False positive has decreased, True Negatives have also increased from previous Decision Tree models. Thus, Decision Tree model suffers from overfitting.
Hence the use of Random Forest classifier was eminent as it performs better than Decision Trees and does not suffer that much from overfitting.
The confusion matrix for the Random Forest classifier on the same data came out to be:

```
Confusion Matrix with Random Forest Classifier:
  TN     FP
  FN     TP

[[ 769  594]
 [ 131 9128]]
```

Figure 8: Random Forest Classifier

C. Predicting Extent of Damage Leading to Impact on Economy.

The damage caused by a particular terrorism incident can cause damage to livelihood, properties and other financial sectors.
For every incident, the extent of damage was categorized into 3 classes which are as follows:

Table 2

CLASS	EXTENT	VALUE
1	Catastrophic	Above $ 1 billion
2	Major	Between $ 1 million and $ 1 billion
3	Minor	Below $ 1 million

After this categorization, Classifier algorithms were applied to predict the extent of the damage caused by the particular incident.

D. SMOTE and why is it not useful with Random Forest Classifier.

SMOTE stands for Synthetic Minority Oversampling Technique [2].

SMOTE is used to handle class imbalance problem. For example, most of the incidents fall into either class 2 or class 3 of the property damage extent (Table 2). So, class 1, having small number of instances, suffers from biasing.

The SMOTE algorithm generates dummy instances of the minority class and thus a balance is maintained to avoid biasing to a large extent [2].

Since Random Forest classifier consists of many Decision Trees which make greedy choice at each step and Random Forest classifier selects the best performing tree among them, it does not suffer from class imbalance. This is evident from the fact the accuracy of the classification model before and after applying SMOTE on training data, does remain same.

For example, the classification reports for the model before and after application of SMOTE are as follows:

	precision	recall	f1-score	support
1.0	1.00	0.05	0.10	38
2.0	0.82	0.85	0.84	2451
3.0	0.95	0.95	0.95	8036
accuracy			0.92	10525
macro avg	0.93	0.62	0.63	10525
weighted avg	0.92	0.92	0.92	10525

Figure 9: Classification Report for Random forest classifier for property extent damage on Middle East & North Africa without applying SMOTE

	precision	recall	f1-score	support
1.0	0.35	0.18	0.24	38
2.0	0.79	0.89	0.83	2451
3.0	0.96	0.93	0.95	8036
accuracy			0.92	10525
macro avg	0.70	0.67	0.67	10525
weighted avg	0.92	0.92	0.92	10525

Figure 9: Classification Report for Random forest classifier for property extent damage on Middle East & North Africa after applying SMOTE

IV. RESULTS

A. North America (region code- 1)

The success rate of a terrorist attack in North America is around 83%. Number of terrorist incidents from 1970 to 2018 has been 3579 (Table 1).

The accuracy of prediction models using different classifiers to predict the success of an attack in North America is summarized as (Table 3):

Table 3

MODEL	ACCURACY
Decision Tree (max. depth=3)	85%
Decision Tree (max. depth=3)	85%
Decision Tree (max. depth=3)	82%
Random Forest	88%

The data for North America was not sufficient to predict the extent of property damage as it had too many missing and unknown values.

B. South Asia (region code- 6)

The success rate of a terrorist attack in South Asia is around 87%. Number of terrorist incidents from 1970 to 2018 has been 48266 (Table 1).

The accuracy of prediction models using different classifiers to predict the success of an attack in South Asia is summarized as (Table 4):

Table 4

MODEL	ACCURACY
Decision Tree (max. depth=3)	91%
Decision Tree (max. depth=3)	91%
Decision Tree (max. depth=3)	88%
Random Forest	92%

The accuracy of Random Forest Classifier with and without SMOTE to predict the extent of damage in South Asia was found to be (Table 5):

Table 5

MODEL	ACCURACY
Random Forest (without SMOTE)	92%
Random Forest (with SMOTE)	92%

C. Middle East and North Africa (region code- 10)

The success rate of a terrorist attack in Middle East and North Africa is around 87%. Number of terrorist incidents from 1970 to 2018 has been 53110 (Table 1).

The accuracy of prediction models using different classifiers to predict the success of an attack in Middle East and North Africa is summarized as (Table 6):

Table 6

MODEL	ACCURACY
Decision Tree (max. depth=3)	92%
Decision Tree (max. depth=3)	92%
Decision Tree (max. depth=3)	89%
Random Forest	93%

The accuracy of Random Forest Classifier with and without SMOTE to predict the extent of damage in Middle East and North Africa was found to be (Table 7):

Table 8

MODEL	ACCURACY
Random Forest (without SMOTE)	92%
Random Forest (with SMOTE)	92%

D. Terrorism Incidents over the years in these regions

The incidents of terrorism over the years in North America, South Asia and Middle East & North Africa is shown in Figure 10, Figure 11 and Figure 12 respectively.

The number of incidents in North America has been sparse and few as compared to other regions.

The South Asia, Middle East and North Africa have seen large number of attacks over the years (1970 - 2018).

The number of terrorist attacks has seen a rise as the years has passed by. Middle East Asia, North Africa and the South Asia Region has suffered the most in terms of the number of total casualties.

In terms of the countries suffering the most, It can be conjectured that the Underdeveloped or developing countries are the most prone and vulnerable parts of the world.

The smaller number of attacks in North America does show the security and proactive measures taken to prevent any mishaps caused due to terrorism.

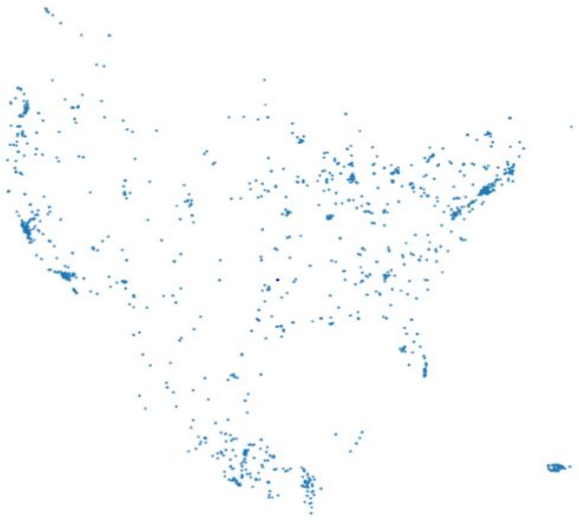


Figure 10: Terrorism in North America over the years

Figure 11: Terrorism in South Asia over the years

Figure 12: Terrorism in Middle East & North Africa over the years

V. SHORTCOMINGS

Although the paper predicts the success of a terrorist attack and the economic impact by predicting the extent of property damage, It does not give estimate of exact value of the property damage.

The models used for the prediction could have given better results if the data across every region had been combined.

It also does not consider Direct and Indirect costs associated with the attack across several sectors such as tourism, aviation and finance.

VI. CONCLUSIONS & FUTURE SCOPE

The success of a terrorist attack and the extent of property amage has been predicted using classisification models such as Decision Tree and Random Forest.

Random forest provided the most accurate result and its comparison with the result after applying SMOTE showed that random forest is not affected that much by class imbalance problem. It does not suffer from overfitting as compared to Decision Tree.

The Middle East and North African regions have suffered the most. 17 out of 20 attacks have been successful in South Asia, Middle East and North Africa.

This project can be further explored and improved by taking into account the Global Terrorism Index provided by Institute of Economics and Peace studies (IEP) [4]. The Direct and Indirect costs associated with attack can be further explored and the impact of the terrorism on the GDP can be predicted.

The Dark Web which is the platform for all kinds pf malicious activities could be mined for information related to Terror manifesto, illegal arms supply, networks of sleeper agents.

Sentiment Analysis of popular social networking platform, web scrapping and crawling to extract information would provide richer dimensions to the project.

REFERENCES

[1] "National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland, The Global Terrorism Database (GTD) [Data file]. Retrieved from https://www.start.umd.edu/gtd", October 2019

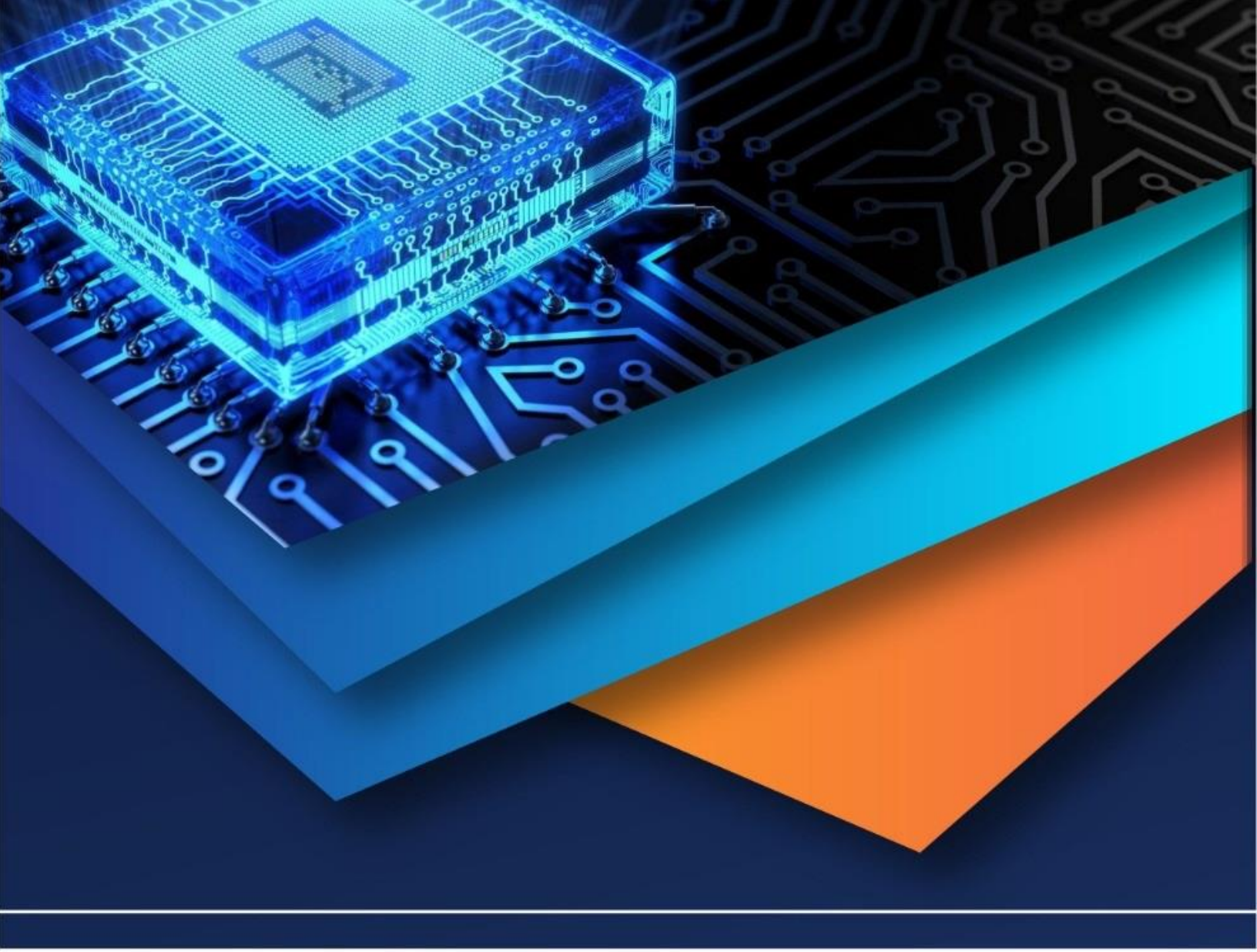
[2] Vivek Kumar, Manuel Mazzara, Angelo Messina, JooYoung Lee, “A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks Prevention and Prediction for Combating Terrorism”, https://arxiv.org/abs/1901.06483, January 2019

[3] Varun Teja Gundabathula, V. Vaidhehi, “An Efficient Modelling of Terrorist Groups in India using Machine Learning Algorithms”, http://www.indjst.org/index.php/indjst/article/view/121766, April 2018

[4] A. Sachan, D. Roy, “TGPM: Terrorist group prediction model of counter terrorism”, https://www.ijcaonline.org/archives/volume44/number10/6303-8516, April 2012

[5] Kumar., V., Zinovyev., R., Verma., Tiwari., P.: Performance Evaluation of Lazy And De-cision Tree Classifier: A Data Mining Approach for Global Celebrity‘s Death Analysis. IEEE Xplore: In International Conference on Research in Intelligent and Computing in Engineering (RICE), pp 1-6, 2018. DOI: 10.1109/RICE.2018.8509082 (2018).

[6] Ahsan S, Shah A. Data Mining, Semantic Web and advanced information technologies for fighting terrorism, International Symposium on Biometrics and Security Technologies, 2008



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY