**TITLE:**

**FORECASTING THE PROBABILITY OF CONFLICT ARISING IN AN AFRICAN NATION**

**Abstract**

Crime is a form of violence or illegal act done by a perpetrator against another person that can cause harm or property damage. African nations have since experienced different crimes in varying regions of the countries from 1997 - 2023 with Eastern Africa being the hotspot of crime and Southern Africa the least. Most challenging areas of the nation are mostly identified with crimes ranging from Battle, Violence against civilians, Explosions, Riots and Protests.

Machine Learning was utilized to perform Exploratory Data Analysis (EDA), it was observed that crime rate had rapidly increased from 2009 to 2023 (March 31st). The total crime committed in Q1 2023 is greater than the total crime committed in 2011. Somalia, Nigeria and Democratic Republic of Congo are top Africa countries with highest crime rate which are mostly battle induced, Sao Tome, Seychelles and Equatorial Guinea are least countries with low crime rate. Angola with the lowest crime rate of 3794 had the highest fatalities of 144098 in the whole Africa Nation.

Standard scaler and XGBOOST were used in training our model for evaluation, the model predicted an F1 score of 0.73 and accuracy of 0.77 which are very good for the size of our dataset.

The model can further be improved in the future leveraging deep learning algorithms.

Keywords: Machine Learning, XGBOOST, Standard Scaler, Deep Learning

1.0 **INTRODUCTION**

A crime is a form of violence or illegal act done by a perpetrator against another person that can cause harm or property damage and is punishable by the law of the governing state of authority in which the crime was carried out. Law authorities apply crime-solving techniques to take preventive measures but in many cases, they cannot deliver effective results (Dakalbab et al., 2022). The rate of crime is increasing from time to time. (ToppiReddy et al., 2018) described that crimes are the major social problems that can affect an individual's life and economic growth of the country. Crime affects peoples’ lives directly or indirectly. It is one of the major variables, which mainly affects the development of the country. Crime analysis is a systematic approach that identifies and analyzes patterns and trends in crime (Sathyadevan & Devan, 2014). This study focuses on the prediction of future occurrences of crime using machine learning techniques between the years 1997 and Q1 2023 shows the occurrence of varying crime types along with the frequency in African Nations., to answer the identified research questions and provide further research on the gaps that have been recognized in the field. We seek to broaden the opportunity for further research on crime investigation and machine learning and our motivation for this work was not to only publish research on previous and current studies on crime prediction, but to:

* To analyze crime rates and fatalities in various African countries, as well as the variations among different actors and the actor groups and their contributions to total crimes and fatalities recorded.
* To observe patterns in the trend in crime rate over the years from 1997 to 2023 (March 31st ) and the factors that are most dominant or differ the most in the top 3 and bottom 3 countries in terms of crime rate.
* Build a crime analytics and forecasting tool that assesses fatalities threat level based on disorder type, actor, and the possible crime events, leveraging insights from historical data to inform preemptive policy decisions.

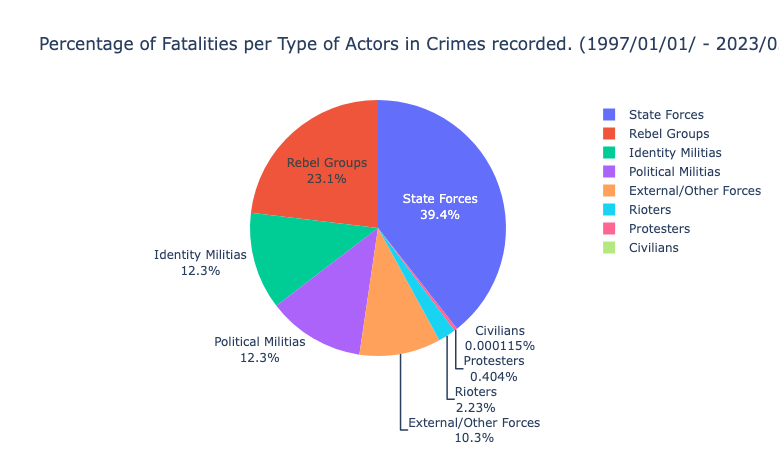


Fig 1: Fatalities per actor in crimes

2.0 **LITERATURE REVIEW**

In a review conducted by Falade et al. (2019), they focused on the use of data mining for crime prediction and concluded that it is a popular research topic due to the impact of crime on a nation's socio-economic development (Kounadi et al. 2020). They noted that the data used for crime prediction could be statistical reports of crimes in a specific area or region. They further suggested that crime prediction research would be beneficial to society by helping law enforcement agencies and governments to understand the multiple layers that contribute to the cause of a crime. This research could also assist governments in making better decisions to enhance the security and safety of their citizens, and adopt a more proactive approach towards the improvement of communities, resulting in a decrease in crime.

(Butt et al. 2020) focuses his research on detecting and predicting crime hotspots. The researchers emphasize that the abundance of data that is being collected and made publicly available has made it increasingly feasible to conduct further research in the field of crime and investigation. The availability of historical data has enabled the forecasting of future crimes, and the development of significant machine learning models is becoming an area of growing interest in order to identify different features related to crime prediction.

According to Falade et al. (2019), governments and law enforcement agencies are responsible for maintaining crime datasets, both past and present. Accurate predictions of future crimes can have significant positive effects on society and the economy, as noted by Butt et al. (2020). Ippolito et al. suggest that a crime detection system could notify users and make them aware of the crimes they are about to commit, including the likelihood of getting caught. The implementation of such a system in society could have significant benefits.

3.0 **METHODOLOGY**

A crime dataset that contains dates and locations of conflict events, specific types of events including battles, civilian killings, riots, protests and recruitment activities, events by a range of actors including rebels, governments, militias, armed groups, protesters and civilians, changes in territorial control, and reported fatalities can be analyzed  using the following methodology

**Data collection and Preparation**

In order to accomplish this research, we have used a crime record dataset from ACLED (Raleigh et. al 2010) from 1997- 2023 (March 31st ). We gathered 315940 records with 31 attributes. The attributes are battles, region, sub\_events, actors, event\_type, disorder type, civilian targeting and many others.

The collected data from ACLED is encoded. It contains a lot of missing values which can reduce the performance of the algorithm. The basic tasks of data preparation include attribute selection, data cleaning and transformation of data for modeling tools (Fig 2) (Endalew, 2017). The data is preprocessed using ML techniques to transform the raw data into an efficient, useful and meaningful form.

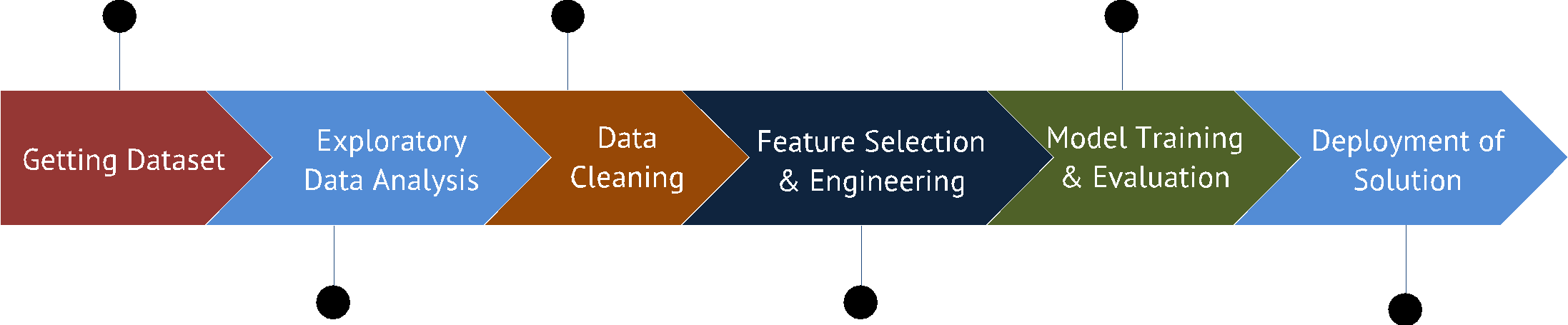


Fig 2: Process Flow Diagram

3.1 **Data Cleaning**

Our dataset contains categorical data. The recommended way to fill missing values for categorical data is using mode values (Fujikawa & Ho, 2002). Checking the missing values in some columns was over 200000 with the whole dataset having 300000+ rows. Dropping these can result in losing about 60% of the data. Actor 2 had over 80000 missing data (Fig 3). The reason for not replacing values but dropping columns instead is because we are dealing with geospatial data. Inter 1 describes actor 1, inter 2 describes actor 2 and we find balance giving a feature engineering to how the training can be improved.

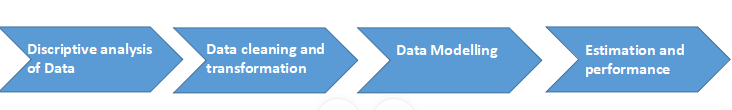


Fig 3: Proposal approach

3.2 **Data Analysis/ Exploration**

The entire dataset is visualized and analyzed using the pandas profiling library. And missing values were all dropped from the dataset before prediction. Seaborn and matplotlib were used for the purpose of visualization.

The data was studied to observe the top and bottom three (3) nations in Africa in terms of record of crimes to have occurred there. The top three Africa Nations are Somalia with about 40000 records of crimes followed by Nigeria with over 30000 and DRC in similar range (Fig 4) between 1997 to Q1 2023. Political violence was the top event in the top three countries with crime rate followed by Demonstrations, Strategic Developments and Political violence/development. Bottom three countries in Africa Nation with least crimes from 1997 to Q1 2023 are Seychelles, Sao Tome and Equatorial Guinea disorder type experienced are political violence, Strategic Developments and Political violence/development.

From the below plot in (Fig 6), it can be inferred that more crimes were committed between 2009 and 2023 of which Q1 2023 has the highest number of crimes yet. Early years of 1997 to 2008 had the lowest number of crimes which indicates that crimes are committed more frequently from the year 2009 to first quarter of 2023.

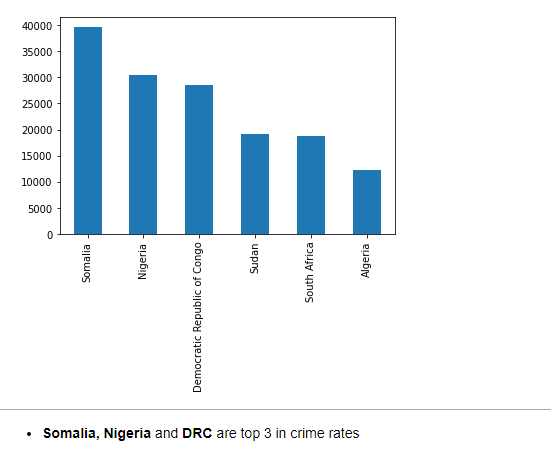


Fig 4: Top three Africa Nation with top crime rate

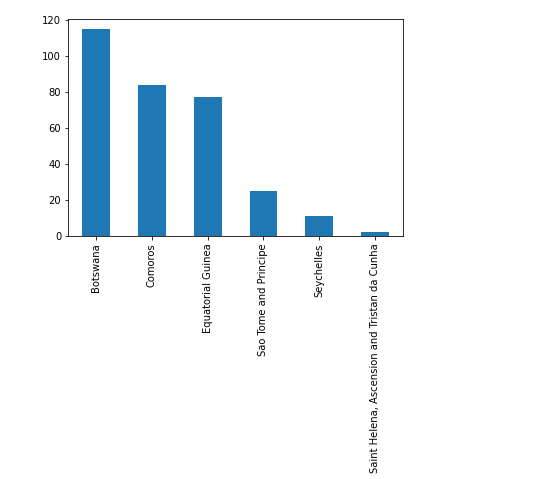


Fig 5: Bottom three Africa Nation with least crime rate

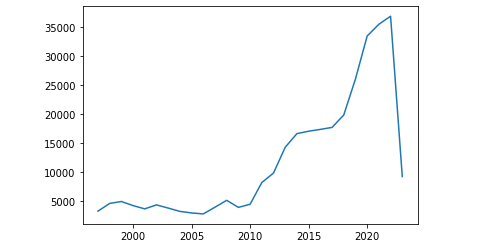


Fig 6: Crime distribution curve per year

3.3 **Data Visualization**

Data visualization is the means of conveying information to users in a graphical view. This can be in the form of graphs, maps, etc. It can be said to be a graphical representation of information for easy understanding and interpretations. The important information in the dataset is exploited to obtain good information about the African Nation crime dataset. This is done through the process of Exploratory Data Analysis (EDA). This enables us to see some useful trends in the dataset .Fatalities committed in African Nation per event types with Battle leading the chart with about 536107 events followed by Violence against civilians, Explosions, Riots, Protest and the least of the events Strategic developments.

Eastern Africa has a higher amount of fatalities, followed by Middle Africa and the least is Southern Africa. Eastern Africa countries show a higher fatalities rate.

3.4 **DATA MODELLING**

After the data has been cleaned, exploited and visualized, it is time to select the machine learning model. Selecting or choosing a machine learning model enables us to perform the work easily and also to obtain the right predictions we expect. The choice of the model (Fig 7) to use shall depend on the amount of data available for the purpose of the training and testing and prediction. It is also very paramount to note that the choice of model selection also depends highly on the problem to be solved. In machine learning a model is created in order to predict the outcome of the event such as predicting the crime rate. After the model is created, the performance of the model is calculated using the method called Train and Test.

**Training**

The *XGBClassifier* algorithm from *xgboost* was used to train the model on the preprocessed dataset. The target variable chosen was Fatalities, which was binned into seven categories/levels.

To address the class imbalance in the target variable, we used the ‘c*lass\_weight’* parameter from the ‘*utils’* module in the Scikit-learn library to compute sample weights based on the target variable's distribution. The generated sample weights were then passed into the *xgboost* model during fitting to improve the model's ability to handle the imbalance. This resulted in a trained model that can effectively predict the target variable with high accuracy.

4.0 **RESULT AND DISCUSSION**

For feature selection, columns with no missing data were chosen except for ADMIN1 which was manually filled after searching for the right value based on other columns. Selected features included event date, event type, sub event type, and actor1.

One hot encoding was performed on categorical features with fewer numbers of categories while label encoding was used for columns with a large number of categorical variables. Target variable (Fatalities) was binned into seven categories/levels. The training set was then scaled using standard scaler and dimensionality was reduced to 98 components from 121 using Principal Component Analysis.

Principal Component Analysis (PCA) was used as a statistical technique used to reduce dimensionality of a data set while preserving the maximum amount of information. PCA transforms a high dimensional dataset into a lower dimensional one (Fig 8) by projecting it into a new set of orthogonal axes referred to as principal components. PCA is useful when working with datasets that have many features or variables, and where it’s difficult to visualize the relationships between the variables.

PCA works by identifying the underlying structure of the dataset and identifying the principal components that explain the most variance in the data. These principal components are linear combinations of the original variables and each one is orthogonal to the others. Discarding the components that explain the least variance, it's possible to reduce the dimensionality of the data while preserving the maximum amount of information.

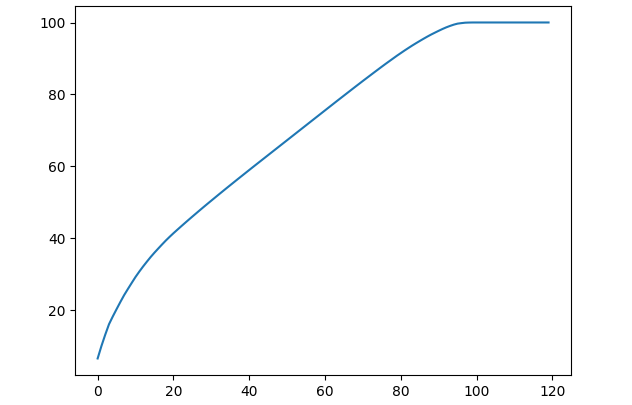


Fig 8: PCA dimensionality reduction curve

**Fatalities label encoding**

{'NO\_FATALITY': 6,

'1\_FATALITY': 2,

'2\_TO\_10': 3,

'11\_TO\_50': 1,

'51\_TO\_100': 5,

'101\_TO\_500': 0,

'501\_TO\_1350’: 4}

Evaluating the model, we got an f1-score of 0.70 and a precision of 0.77 which are very good for the size of the dataset. The model can further be improved in the future leveraging deep learning algorithms.

And for the predictions that were wrong, the majority fell to the next level in the range of fatalities, which means our model generalized well on the dataset.

**CONCLUSION**

Our solution provided a comprehensive analysis of crime rates and fatalities in various African countries, with insights into the contributions of different actors and actor groups. Our analysis also highlighted the patterns in the trend of crime rate over the years, and the factors that are most dominant in the top and bottom countries in terms of crime rate. Our findings underscored the importance of considering location, time, and actors when analyzing crime to make informed policy decisions.

**REFERENCE**

Butt UM, Letchmunan S, Hassan FH, Ali M, Baqir A, Sherazi HHR (2020) Spatio-temporal crime HotSpot detection and prediction: a systematic literature review. IEEE Access 8:166553–166574.

Dakalbab F, Abu Talib M, Elmutasim O, Bou Nassif A, Abbas S, Nasir Q (2022) Artificial intelligence & crime prediction: a systematic literature review. Soc Sci Humanit 6:100342

.

Endalew, A. (2017). Crime Analysis For The Case of Bahir Dar City Administration Using Data Mining Approach. Bahir Dar University.

.

Falade A, Azeta A, Oni A, Odun-ayo I (2019) Systematic literature review of crime prediction and data mining. Rev Comput Eng Stud 6(3):56–63.

Fujikawa, Y., & Ho, T. (2002). Cluster-based Algorithms for Filling Missing Values. In 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining,. Taiwan

Hajela G, Chawla M, Rasool A (2020) A clustering-based hotspot identification approach for crime prediction. Procedia Comput Sci 167:1462–1470.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. (2010). “Introducing ACLED-Armed Conflict Location and Event Data.” Journal of Peace Research 47(5) 651-660.

Sathyadevan, S., & Devan, M. S. (2014). Crime Analysis and Prediction Using Data Mining. First International Conference on Networks and Soft Computing, 406–412.

ToppiReddy, H., Saini, B., & Mahajan, G. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. Procedia Computer Science, 132(Iccids), 696–705. <https://doi.org/10.1016/j.procs.2018.05.075.>