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

A Hamoye Capstone Project by Team Theano



### 

### 

### 

### **INTRODUCTION**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### A crime is a form of violence or illegal act performed by perpetrator(s) against another in order to cause harm to person or property which is punishable by the authority. Directly or indirectly, crime affects people's lives. It is a major variable that affects the development of a country. Crimes in African nations keep rising which span from a range of violent and non-violent actions by political agents, including governments, rebels, militias, identity groups, political parties, external actors, rioters, protesters and civilians. The rate and fatalities of these crimes brought about Crime Analysis.

### Crime analysis is a systematic approach that identifies & analyses patterns and trends in crime. This analysis includes, exploring & detecting crimes and identifying the relationships between these crimes with criminals. This process aims to extract meaningful information from a large dataset and to assist law enforcement agents to tackle and reduce crimes.

**PROBLEM STATEMENT**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. To analyse crime rates and fatalities in various African countries, as well as the variations among different actors and actor groups and their contributions to total crimes and fatalities recorded.

2. 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.

3. Build a crime analytics and forecasting tool that assesses fatalities threat level based on disorder type, actor, and possible crime events, leveraging insights from historical data to inform preemptive policy decisions.

**ABOUT THE DATASET**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The Armed Conflict Location and Event Data Project is designed for disaggregated conflict analysis and crisis mapping. This dataset codes the dates and locations of all reported political violence and protest events in dozens of developing countries in Africa.

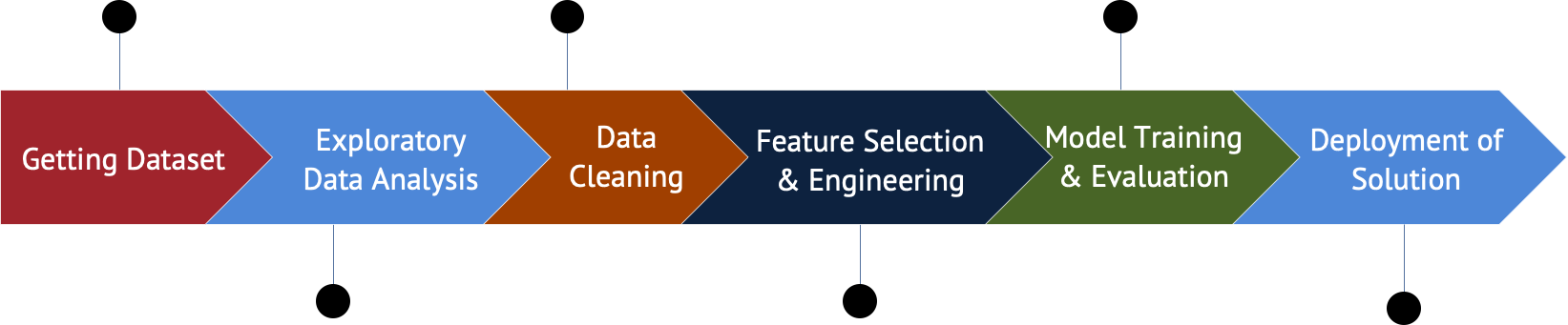
Political violence and protest includes events that occur within civil wars and periods of instability, public protest and regime breakdown. The dataset covers all African countries from 1997 to March 31st, 2023.

These crime data 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

**PROCESS FLOW**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



### GETTING DATASET

### The dataset was obtained directly from ACLED, an encoded dataset generated over a period of year (1997 to March 31st 2023.) detailing conflicts in African nations for 25+ years. All these are recorded in 315940 rows and 31 columns.

### EXPLORATORY DATA ANALYSIS

On exploratory data analysis, the following insights were generated:

1. Battles (an event type) contributed the most to fatalities in a crime.
2. State Forces as the main actors in the crime account for 39.4% of fatalities recorded in Africa between 1997 and 2023 March 31st.
3. The trend of number of crimes commited per year, showed that up to 2009, the number of crimes recorded per year didn’t really vary but from 2009 upwards, there is a steady upward trend.
4. We identified that the top 3 countries in terms of crime rate were Somalia, Nigeria and the Democratic Republic of Congo all having a record of 25000+ crimes.
5. Botswana and Comoros are among the bottom countries in terms of crime rate recording under 120 crimes.
6. Actor type 1 which represents State Forces, has the most contribution to number of fatalities in both cases.
7. Actor types 2, 3 and 4 which are rare in the countries with least number of fatalities recorded are common in the high fatality countries.
8. We can then assume that State Forces, Rebel Groups, Political Militias and Identity Militias are responsible for crimes incurring high number of fatalities.
9. As regards Angola which had relatively low number of crimes recorded but the highest number of fatalities, we observed that majority of the crimes that occured there were battles which from the dataset generally always resulted in a high number of fatalities.

### Chart, line chart Description automatically generatedRecord of crime per year

### DATA CLEANING

### We employed the use of some functions like shape, head, describe, info, columns to ascertain the core feature and its relatives. The dataset has 315940 records with over 60% null values. So, to ensure that our data doesn’t have bias, we took a closer look at the missing values per column:

Missing Values on some columns

* ASSOC\_ACTOR\_1 - 231997
* ASSOC\_ACTOR\_2 - 253699
* TAGS - 255561
* ACTOR2 - 86136
* ADMIN3 - 161453
* ADMIN2 - 2451
* ADMIN1 - 2 (Corresponding locations - Gulf of Guinea & Coast of Benin)

### Dropping Values

Looking at the amount of missing values, we definitely cannot drop rows, we then choose to drop columns with large amounts of null & missing values. They are: ASSOC\_ACTOR\_1, ASSOC\_ACTOR\_2, etc.

On the two (2) null of ADMIN1 (Gulf of Guinea & Coast of Benin), we replace them with the most closely related of the said ADMIN1 (Bayelsa & Oueme respectively).

### FEATURE SELECTION AND ENGINEERING

### 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 number 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.

**Fatalities** was binned into **seven label encoding**  in order to train the model:

* ‘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

### MODEL TRAINING AND EVALUATION

### The XGBClassifier algorithm from xgboost was used to train the model on the pre-processed 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 ‘class\_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.

### 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, majority fell to the next level in the range of fatalities, which means our model generalised well on the dataset.

Graphical user interface, application

Description automatically generated

### 

### DEPLOYMENT OF SOLUTION

We built a crime analytics and forecasting tool that assesses the threat level of fatalities based on disorder types, actors, and possible crime events. The tool leverages insights from historical data to inform pre-emptive policy decisions.

This tool was built as a web app using Streamlit and deployed using Streamlit Cloud. The dataset referenced by the web app was hosted on the cloud using Azure Blob Storage. All components of the web app can be found in the GitHub repo.

Link to Web App -<https://bit.ly/africa-crime-forecasting-tool>

Link to GitHub repo - [Team-Theano-Capstone-Project](https://github.com/Sammybams/HamoyeAI-Team-Theano-Capstone-Project)

# **SUMMARY AND RECOMMENDATIONS**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

We successfully provided a solution to our defined problem statement.

A pre-processed dataset was used to train an xgboost model that effectively predicted the target variable (fatalities).

A crime analytics and forecasting tool was built to assess the threat level of fatalities based on disorder types, actors, and possible crime events. The tool leverages insights from historical data to inform pre-emptive policy decisions.

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 analysing crime to make informed policy decisions.