ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Conflict Analytics:

A predictive model to forecast violent conflicts in Ethiopia for improved early warning systems

A Thesis Submitted to the School information science of Addis Ababa University in Partial Fulfilment of the Requirements for the Degree of Master of Information Science and system (information science)

By

Meheret Takele Mandefro

June 2023

Addis Ababa

Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONSCIENCE
SCHOOL OF INFORMATION SCIENCE

Conflict Analytics:

A predictive model to forecast violent conflicts in Ethiopia for improved early warning systems

A Thesis Submitted to the School information science of Addis Ababa University in Partial Fulfilment of the Requirements for the Degree of Master of Information Science and systems (information science)

BY

Meheret Takele Mandefro

Advisor:-

Dr.Wondwossen Mulugeta (PhD.)

June 2023,

Addis Ababa, Ethiopia

DECLARATION

I declare that this thesis is my original work and has not been submitted as a requirement for a degree in any other university.

Name: Meheret Takele Mandefro

Signature:_____

Date: _____

This Thesis has been submitted for examination with my approval.

Name: Dr.Wondwossen Mulugeta

Signature:_____

Date: _____

Name and Signature of Members of the Examining Board

Name	Title	Date	Signature
Wondwossen Mulugeta (PhD)	Advisor	_____	_____
Martha Yifiru (PhD)	Examiner	_____	_____
Solomon Teferra (PhD)	Examiner	_____	_____

ACKNOWLEDGEMENTS

First of all, I would like to be thankful to my almighty God and His Mother Saint Marry for giving me strength and hope in the entirety of my life.

I would also want to thank my adviser, Dr.Wondwossen Mulugeta, for his insightful guidance and suggestions. His tips helped me in staying on track with my research and deliver it on time.

I'd like to thank my family, the special goes to Ababi and Mami, for constantly encouraging me and checking the status of my thesis work. I would like to extend thanks and appreciation to Tsega Takele, Amanuale Takel and Enatu who were constantly present beside me throughout my study. I would also like to convey my heartfelt love and gratitude to Yeshiwas Degu who supported and encouraged me to accomplish my study on time.

I am also grateful to Mekelle University for supporting my expenditures and providing me with a paycheck. Furthermore, I am appreciative to the Ministry of Education for giving me all of the necessary papers to start the program.

Finally, I'd want to thank my teachers and classmates for their welcoming and encouraging attitudes during the past two years.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

OAU	Organization of African Unity
AU	African Union
ACLED	Armed Conflict Location & Event Data Project
RF	Random Forests
GBT	Gradient Boosting Trees
SVM	Support vector
GSDRC	Governance and Social Development Resource Centre
GDP	Gross Domestic Product
UNDP	United Nations Development Program
PRS	Political Risk Services
ICRG	International Country Risk Guide
EIU	Economist Intelligence Unit
SDGs	Sustainable development goals
AI	Artificial intelligence
IGAD	Intergovernmental Authority on Development
CEWS	Conflict Early Warning System
ViEWS	Violence Early Warning System
CAR	Central African Republic
DRC	Democratic Republic of the Congo
DT	Decision Tree
NB	Naïve Bayes
ERD	Experimental research design
CRISP-ML	Cross Industry Standard Process for machine learning
EPO	Ethiopian Peace Observatory
FFP	Fund for Peace
EIU	Economic Intelligence Unit
UNPD	United Nations Population Division
UNDP	United Nations Development Program
WGI	Worldwide Governance Indicators
HDI	Human Development Index
HIEF	Historical Index of Ethnic Fractionalization
ML	Machine Learning
MSE	mean square error
TP	True Positive
TN	True Negative
FP	False Negative
FN	False Negative
EDA	Exploratory Data Analysis

TABLE OF CONTENTS

Contents **Page number**

Abstract

In the context of an increasing number of violent political conflicts across nations and societies, understanding and predicting violent conflicts that lead to significant economic, social and humanitarian consequences is both an academic interest and a moral obligation. Research on conflict prediction is critical to offer input for policymakers to see potential conflicts and devise strategies for conflict early warning. However, the application of data analytics tools to analyze the dynamics of violent conflicts in Ethiopia has hardly existed. With the motivation to fill this knowledge gap, this research aims to develop a model using supervised machine learning algorithm that can best forecast violent political conflicts in Ethiopia in terms of the dominant conflict types/categories and regions at risk of conflict incidents. Methodologically, this research has employed an experimental research design and adopted the CRISP- ML framework. Predictive analytics tools, as well as three algorithms (random forest, gradient boosting and Gaussian naive Bayes), are used. Open-source software called Jupyter Notebook is used for analysis. The research combined past and recent conflict incidents data with political, economic, social and environmental data from 2007 to 2022. After collecting data on both independent and dependent variables from open databases of research institutes and international organizations, models were built, compared and assessed using a new dataset of variable values projected for the next five years (2023-2028). The finding shows that the Gradient Boosting machine learning algorithm has a better performance than Random Forest and Gaussian naïve Bayes in predicting the location and types/categories of conflict classes individually. In predicting types of violent conflicts, while Gradient Boosting has a testing accuracy of 57%, both the Random Forest and Gaussian naïve Bayes has 55%. Yet, in terms of predicting the location of conflict incidents, all have 48% testing accuracy. However, when both type/category and location of conflict incidents were considered together the predictive performance of the selected algorithms declined to 30% testing accuracy. In conclusion, the performance of conflict-predicting models is determined by whether the classes (type and locations) are merged and predicted concurrently or separately and independently. The best performing algorithm(57%), such as the Gradient Boosting machine learning algorithm, predicts that Ethiopia will continue to be at risk of violent conflicts for the coming five years, 2023- 2028. The location and type of violent political conflicts shift yearly. With its original findings, this research can make valuable contributions to policy-making and academic fields, such as Information Science, Conflict Studies and others.

Keywords: *Predicative analytics, machine learning algorithms, conflict prediction, conflict early warning, Ethiopia*

CHAPTER ONE: INTRODUCTION

1.1. Background of the Research

Violent political conflicts are common experiences in different spatial (Africa, Asia, Europe, Latin America, etc.) and economic zones (developing/developed countries). The type, level, severance, and causes of conflicts may vary among countries. Conflicts are dynamic; they might increase or decrease in time. The causes are multiple: economic, social, political, environmental etc. In many developing countries, violent political conflicts have a profound challenge to development and peace (Etefa, 2019; Mengistu, 2015). In most cases, the ramifications of conflicts move beyond national borders and affect other states and societies. African countries are not exceptional when it comes to violent political conflicts; they are affected by different forms of violent conflicts at different times (Avis, 2019)

Ethiopia is Africa's oldest nation. It is a home of people with different ethnic, religious, and cultural backgrounds. It is also the home of the African continental organization (African Union) and various international organizations. The earlier assumption of hosting such offices was partly due to the country's relative peace, security, and stability (Mehari, 2014). This is, however, not given. It is uncertain. Violent political conflict in Ethiopia could have a significant impact not only on the Ethiopian state but also on international and regional issues. Conflicts are affecting Ethiopia's role in regional peace and security and its image. Due to its position at the centre of the Horn of Africa, Ethiopia's conflict has a huge implication for the stability of the region. This might include the refugee crisis, displacement and proliferation of armed groups etc (Mehari, 2020).

To address the cycle of violence, a variety of intervention mechanisms are put in place by the Ethiopian state. To mention, in 2018 Ethiopian government established the Ministry of Peace aimed at building peace, preventing and resolving conflict, and establishing contact between the federal and regional states to coordinate and support developing regions. Besides, it aims to establish a governance structure that can guarantee the rule of law and promote peace. The ministry works in partnership with relevant government bodies, cultural and religious institutions and others to promote peace and respect among peoples of different religions and nations, nationalities and peoples of Ethiopia (FDRE proclamation no.1097/2018). The government has also established National Dialogue Commission to further strengthen efforts of achieving and maintaining peace and stability in the country (FDRE proclamation no.1265/2021). Despite questions about their effectiveness in preventing conflicts and addressing them once they occurred, the establishment of this structure is a positive step forward.

It is also important to see the complexity of violent conflicts in Ethiopia. Many factors cause violent conflicts, but they have different levels of influence on conflict formation (Berihu, 2021). Thus, research that targets understanding the most fundamental drivers of violent conflict in Ethiopia, like this MSc thesis, is very important in assisting state policies and programs. Understanding the causes of conflicts is very crucial to act up on the most appropriate means of mobilizing early and effective responses or interventions. Conflict mapping and analysis play an important role in serving as a source of information to gain an

understanding of the conflict contexts and to plan effective intervention mechanisms before violent conflicts and insecurity happened (ibid). Academic and research work on this topic mainly appears limited to social science such as experts, researchers and teachers in peace and security, law, federalism, and political science. There, however, remains little study conducted by researchers who have knowledge and skill in natural science such as data analytics and machine learning. Having this background in mind, this MSc research engages with interdisciplinary research, integrating information science (particularly data analytics and algorithm) methods with theories of peace and conflict, to produce policy-relevant output that can provide a predictive model for the violent internal conflict in Ethiopia to improve early warning systems. This research is aimed at developing a machine learning model that will help us to understand and predict violent conflicts happening in Ethiopia so that we can act upon it to improve future outcomes.

.

1.2. Statement of the problem

In today's Ethiopia violent conflict is a daily occurrence with different intensities, causes, and effects. The conflict dynamics in Ethiopia are complicated and it is increasing from time to time. The security situation in the country is deteriorating. In 2020, the global peace index ranked Ethiopia amongst the less peaceful countries than in 2019 and ranked 133 out of 163 global countries in this respect. The Global Peace Index has also shown that Ethiopia has low levels of Positive Peace; positive peace exists when the state achieves better freedom, security, development, justice etc. The Armed Conflict Location & Event Data Project (ACLED) and International Crisis Group also pointed out Ethiopia as the 1st and 2nd conflict to worryabout in the coming years.

The impact of violent conflicts on the Ethiopian state and society is unimaginable. Conflicts have caused massive damage to the country in terms of human loss, large-scale displacements, and injuries. For instance, according to the Global Report of Internal Displacement in 2019, Ethiopia was ranked among the three leading countries in the world in Internal Displacement because of conflict subsequently for two years. In addition, due to the conflicts Ethiopia's physical infrastructures, roads, telecommunication lines, bridges, and material capital has been disrupted leading to economic disorder in the country. According to (Geda, 2004) conflict is the main and key reason for Ethiopia's poverty and backwardness. It is also observed that due to conflicts we have witnessed a disrupted social capital and social interaction leading to a huge social disorder, including disruption of families, households and neighborhoods.

The ongoing efforts and institutional structures to address the violent conflicts in the country appear ineffective. There is a lack of scientific locally driven- knowledge about which salient factors drive internal violent conflicts in Ethiopia and under what circumstances they occur. Mostly, conflict prevention mechanisms and tools are external-driven and top-down approaches and the early warning responses depend mostly on external financial assistance. This poses difficulties in designing effective early warning and conflict prevention mechanisms, prioritizing issues, allocating human and financial resources, and adapting national policies and strategies to address violent conflicts in Ethiopia. In addition, the topic of which machine learning algorithm performs better to predict the nature of violent conflicts and regions at risk of conflict at the country level is under investigated, if not ignored.

1.3. Research Objectives

General objective: This research aims to examine the performance of different supervised machine learning algorithm to develop a predictive model that can forecast violent political conflicts in Ethiopia in terms of the dominant conflict types/categories and areas at risk of conflict incidents.

Specific objectives: This research aims to assess Gaussian naive Bayes' predictive performance compared with ensemble algorithms such as Random Forest and Gradient Boosting. It also seeks to draw the most salient drivers of violent conflicts based on their variable importance. Finally, it seeks to forecast violent political conflicts in Ethiopia for each of the coming 5 years, from 2023 to 2028, using a best-performed machine learning algorithm and tools.

1.4. Research questions

The research will address the following research questions:

1. Which machine learning algorithm has the better performance to predict location and type of violent conflicts in Ethiopia?
2. What are the most salient drivers of violent political conflict in Ethiopia which has the highest feature importance?
3. Which types or categories of conflict incidents are more likely to dominate, and where, over the next five years (2023-2027) based on the best-performing predictive model?

1.5. Scope and limitation

This research is limited to violent conflicts in Ethiopia. Particular focus will be given to identifying the drivers of violent conflicts that serve as input for conflict prediction. The data will be limited to the last fifteen years (2007-2022). It is a fact that data analytics research faces a serious lack of enough available data for public use. Even for this research, there is a shortage in data availability especially accurate conflict data from the Ethiopian government. In addition the indicator datasets collected are also limited to yearly time span that forces this research to use same factors for different conflicts happening in different locations in the same year.

1.6. Significance of the study

The research will have academic and policy significance. This research will add to the conversation of conflict dynamics in Ethiopia. It will provide empirical findings on the drivers and categories of violent conflicts in Ethiopia. Having the lack the treatment of this topic by data analytics researchers and experts, the research seeks to break new ground in this area. It will also shed light on the importance of data analytics in conflict prevention and conflict analysis research.

At the policy level, the goal is to shed some light on the most salient drivers of conflict, dominant conflict types and regions at risk of violent conflicts. This will provide state policymakers and other relevant organizations with a tool to help them develop conflict

early warning systems. Generally, the research will help improve policy and practice regarding conflict prevention and early warning; the research will improve the country's response to conflict and ensure peace and security by identifying the main causes of conflict, forecasting the main conflict categories and informing areas in risk of violent conflicts.

1.7. Organization of the thesis

The paper was organized into seven chapters.

- Chapter 1: presents the introduction part including the background of the study, statement of the problem, research questions, general and specific objectives of the study, limitations of the study, and significance of the study.
- Chapter 2: reviews the literature relevant to this research. It presents a review of relevant literature on conflict and data analytics and shows important research gaps that the research aims to address.
- Chapter 3: presents the methodology that is employed for this research which includes the research design and data analytics methods.
- Chapter 4: provides the exploratory analysis and preprocessing steps. It presents the different data sources that are used for this research and preprocessing techniques applied to the dataset.
- Chapter 5: presents model building and evaluation. It presents the different machine learning algorithms that are employed to build a predictive model with their performance results.
- Chapter 6: offers discuss and analysis of the major results and findings of the research.
- Chapter 7: concludes and offers recommendations to policymakers and future researchers in this area.

CHAPTER TWO: LITERATURE REVIEW

Introduction

This chapter presents a review of literature on conflict and data analytics and shows important research gaps that the research aims to address. The first section provides an understanding of the meaning of data analytics and is followed by an understanding of conflict, its main attributes and driving conditions for violent political conflicts. The chapter also offers a theoretical link between data analytics, conflict prediction and conflict early warning systems. The last section of this chapter reviews relevant empirical research related to data analytics and conflict.

2.1. Data Analytics: Concepts, Models and Tools

Data analytics is an emerging field that intends to use vast volumes of data to illuminate important issues in a range of industries, including politics, economics, conflict resolution, and health. This section reviews data analytics literature that is largely related to conflict prediction. Data analytics is a multitude of tasks of examining raw data such as collecting, organising, cleaning, splitting, refining etc. These raw data include text, graphics, streaming data, relational data etc. The goal of data analytics is to employ mathematical and statistical techniques to establish outcomes, make predictions and draw important analyses to inform policy and decisions (Ahmed and Pathan, 2019: 24). It, however, doesn't necessarily similar to data analysis, which is mainly about the interpretation of the data. Data analytics requires the application of data science and machine learning.

In his study, Watson (2014) discussed the different types of data analytics: (1) Predictive analytics: It is the prediction of what will happen in the future. The goal of this type of data analytics is to get information about future development, such as the demands of individuals or corporate clients. (2) Exploratory analytics is primarily concerned with "discovering previously unknown relationships in big data." Watson (2014:1251). This is critical for large firms with a large volume of client data to help discover patterns of consumer activity. (3) Prescriptive analytics: Unlike predictive analytics, which predicts what will happen, this sort of data analytics gives suggestions on what to do. As a result, it is solution focused.

There are several data analytics tools, models and methodologies that can be used by researchers, especially in understanding and offering conflict prediction. Machine learning is a dominant approach these days due to its ability to process large data and come up with certain patterns within the data. The most common machine learning algorisms that can be used by researchers include the following:

- **Decision Trees:** this refers to available "tools that use human-understandable graphs to process data to provide a classification or decision point". To construct the trees, it is possible to use categorical and numeric and categorical variables (Ahmed and Pathan, 2019: 24).
- **Neural networks:** are one type of machine learning algorithm that allows the classification and clustering of data which are in huge amounts. (Osisanwo et al., 2017)
- **Random Forests (RF):** An ensemble learning method for classification, regression and, other tasks that are operated by constructing many decision trees during training. They

output either the class mode or the mean (regression) of the individual trees. (Ahmed and Pathan, 2019: 24). The algorithm combines the results of several individual models into a single more powerful prediction. Different versions of ensemble learning techniques and processes have previously been applied successfully in conflict forecasting settings but, as described by several scholars (Ettensperger, 2022; Havelange, 2021; Helle, Negus, and Nyberg, 2018) the one that shows improved prediction ability is the random forest algorithm. Random Forests is applied in tree-based learning algorithms where many single models are created and combined. It also works well with small to medium data. The algorithm is currently one of the most popular machine learning techniques applied to conflict prediction tasks across a multitude of scientific disciplines. It constitutes an efficient approach to reducing uncertainty in forecasting endeavours and improves prediction performance by a significant degree (Ettensperger, 2022; Havelange, 2021; Perry 2013). Random Forest creates a forest of weak predictors first and combines individual results into a strong predictor during a later stage. Specifically, the method is based on the averaging of predictions from decision tree forests (Helle, Negus and Nyberg, 2018). It uses a process called "bootstrap aggregation" or short "bagging" to avoid overfitting and reduce variance in the statistical learning method. "Bagging" describes the process of generating randomized sub- samples of the original data distribution drawn with replacement to create independent predictions. After generating a sufficient amount of independent prediction models, the results are averaged for numerical prediction, or the majority outcome is used for classification. Part of its popularity is due to the relatively simple hyperparameter tuning and robust learning results within diverse settings. In recent publications, it is a technique well-suited for different selections of conflict prediction data (Ettensperger, 2022). It is chosen as a model algorithm for this research due to its high predictive power and its computational efficiency.

- **Gradient Boosting Trees (GBT):** An ensemble learning method used for regression and classification purposes. The difference between RF and GBT is the gradient-boosted tree models sequentially. In GBT, a series of trees are constructed and every tree attempts to right the errors of the preceding tree in the series. (Friedman, 2001:15) the algorithm builds classifiers out of a large number of small classifiers. The first tree needs to grow, and this will be followed by making the proceeding tree fit the residuals from the predictions of the first tree to reduce the prediction error. Gradient boosting classifiers combine a large number of trees, but do so sequentially, learning from previous estimates. Gradient boosting has the advantage of building shorter trees than random forests. It is also better in prediction than random forests. But unlike the random forest, it might take longer to train as a limitation of this method (Mark, et al.,2021)
- **Gaussian Naive Bayes:** Gaussian Naïve Bayes is an extension of the Naïve Bays classifier that is compatible with continuous features. This method is a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. (Hindawi, 2022).

- **Support vector**: “They are used for classification and regression analysis, for SVM maps values into high dimensional space where the hyperplane of maximum linear separation is determined”. (Ahmed and Pathan, 2019: 24).

There are also different tools available for researchers to use. For instance, Jupiter Notebook is a platform and open-source software used for research in conflict prediction. It uses advanced techniques, such as natural language processing and machine learning to identify conflict events and future prediction of patterns.

2.2. Understanding the term Conflict

After reviewing several articles on conflict, it is observed that there is a lack of providing a single and agreed definition of violent conflict among scholars and researchers. Some scholars such as Mark (2016) and William (2019) define violent conflicts and used them interchangeably with, civil war, war, intra-state conflicts, or inter-state conflicts. This understanding is perhaps relevant in capturing the different natures of conflicts. What is common in all conflict situations is a disagreement between actors over the perceived incompatible goals and aims.

Wallensteen and Axel (1994: 333-349) have categorized violent conflicts according to the number of casualties they result in per year. They denoted that 25 deaths per year related to battles are a mark to delineate violent political conflicts, where between 25 to 1000 are categorized as intermediate conflicts and those 1000 battle-related deaths more than 1000 are demarcated as war or major armed conflict. (Cited in Frère and Wilen, 2015) This research mainly focuses on the internal violent conflict at the country level, in contrast to more systemic or structural violence (global wars) or individual conflicts since the research aims to contribute to state policymakers' and experts' analysis.

Conflict analysis is an analytical tool and a process of analyzing and understanding violent political conflict (Siân, 2017). This involves understanding the background of the conflict, historical conditions or cycle of the conflict, the main drivers and causes, the actors participated in and the consequences of the conflict. According to Fisher, there are three types of conflict:

✓ Open conflict: which is very noticeable and cases are entrenched historically and politically

✓ Surface conflict: these are visible but are not necessarily deep-rooted nature

✓ Latent conflict: these are hidden, emerging and unnoticeable conflicts.

Both open conflict and surface conflict can be analyzed using conflict analysis tools. Analyzing latent conflict by identifying actors, causes and issues would help to reduce the cost of conflicts and wars. Conflict analysis is helpful for conflict prevention and resolution experts.

2.3. Theorizing Conflict Drivers and Indicators for research analysis

Conflict and disagreement are inevitable and they occurred all the time in all places. If they can manage and resolve, societies and states can have better change and development. Most of the time they are context-specific, multi-causal and multidimensional and can result from a combination of different factors, as stated by the Governance and Social Development Resource Centre (GSDRC) (2014), such as political and institutional factors, socioeconomic factors and also resource and environmental factors. Considering the relevance to this research, in a broader sense, political (Governance, State fragility and Democracy), social (unemployment, population growth) and economic (GDP per capita, human development) factors are presented here as the main conditions that drive violent conflicts to emerge. Specific elements of each driving factor are also discussed below.

2.3.1 Political: Governance, State Fragility and Democracy

Governance

Governance is widely discussed among policymakers and scholars. It broadly covers "rules, enforcement mechanisms, and organizations" (Elizabeth, et al., 2020). According to the World Bank, governance refers to how leadership uses power and authority to manage the available resources for socioeconomic development (Kaufmann et al, 2010). According to Daniel, Aart and Pablo (2001), the idea that 'governance matters' has become a global theme. This is the reason why 'governance' is set as important policy agenda for international non-governmental institutions, for instance, United Nations Development Program (UNDP). Over the last decade, promoting good governance has been a topic issue (Elizabeth, et al., 2020). This is because governance is related to peace and conflict across states and societies. According to Daniel Kaufmann, et al (2010) Good Governance, as opposed to bad governance can be observed by looking at the following issues**:**

- ✓ As a process: how leadership is elected/selected, mentored and state power transferred.
- ✓ Government capacity: This is about how effective, efficient, and structured is the government to adopt policies and implement them.
- ✓ Citizens and state respectfulness: it is about how citizens and leadership respect state institutions and structures.

To measure the process by which governments are selected, monitored and replaced two indicators have been developed which are**:** (1) ***Voice/accountability*****:** The question of how responsive government is helping to state-society relations. A better responsive government might experience less protest and less responsive government. It is also about whether elections are free and fair, how popular the government is and how long it stayed in power. For example, according to Political Risk Services (PRS), leaders who emerged through (and from the military) military means are not elected. They are less representative of society and have less accountability. This presents a risk of conflict. **(2) *Political stability*:** Assess political violence and its influence on governance. According to this indicator, the highest scores go to countries

with no armed opposition, and where the government does not indulge in arbitrary violence, direct or indirect. The lowest ratings go to civil war-torn countries. Intermediate ratings are awarded based on the threats to the government and business: whether the acts of violence have a political objective or not, whether violent groups represent a sizeable minority or not, how well organized these groups are and how much popular support they receive, how frequent the act of violence is, and whether they are geographically limited or not.

To measure the capacity of the government to effectively formulate and implement sound policies two indicators have been developed which are**:** (1) ***Government effectiveness*****:** refers to the ability to conduct and implement policies and programmes within the duration of the leadership tenure in office. This will depend on issues such as the type of governance, the cohesion of the government and governing party or parties, the closeness of the next election, the government's command of the legislature, and popular approval of government policies. This indicator is also connected with the Bureaucratic Quality of the government it measures institutional strength and quality of the civil service, assesses how much strength and expertise bureaucrats have and how able they are to manage political alternations without drastic interruptions in government services, or policy changes. Good performers have somewhat autonomous bureaucracies, free from political pressures, and an established mechanism for recruitment and training. (2) ***Regulatory quality:*** This indicator most importantly looks atthe investment profile of the country in which the government's attitude towards investment. Specifically, whether the government permits and promotes private sector development through policies and regulations including the risk to operations, taxation system, and repatriation and labour costs. (Kaufmann, et al, 2010)

To measure how the leadership and citizens respect state institutions the respect of citizens and the state the institutions that govern economic and social interactions among them two indicators have been developed which are: (1) ***Rule of law*****:** it is related to how law and order observed and maintained in the country including the legal systems, policy and court impartiality (Ibid). (2) ***Control of corruption*****:** This is related to how the government system corrupts. This is due to its impact on the political, social, and economic environments. It has an impact on government efficacy and may result in patronage and favouritism rather than meritocracy and talent, resulting in instability in the state structure. Financial corruption, in the form of demands for bribes in connection with import and export permits, exchange controls, tax assessments, police protection, or loans, is the most prevalent type of corruption encountered directly by businesses. This measure is also concerned with real or possible corruption in the form of patronage, nepotism, job reservation, "favour-for-favour", covert party finance, and suspiciously strong relationships between politicians and business. The greatest danger posed by corruption is that a big political scandal causes a popular reaction, culminating in the government's collapse or overthrow, as well as a major reorganization or restructuring of the country's political institutions. (International Country Risk Guide (ICRG))

State fragility

State fragility, which is related to a weak state, denotes the lack of the state's will or competence to execute essential functions (Ines, 2015). In general, it refers to any type of political, social, or economic instability (Javier and Sebastian, 2009), and a lack of institutions to address poverty, development, and ensure security and human rights. When the state is 'fragile,' it is unable to fulfil its responsibilities, undermining its legitimacy (Natalie, 2021). This endangers livelihoods as growing economic downturns and other crises erode human security and raise the prospect of violent conflict (Javier and Sebastian, 2009).

State fragility is difficult to measure. The Fund for Peace, a non-governmental research in Washington, measures state fragility. FFP promotes more peaceful and successful communities by creating better processes and wiser collaborations, with a clear focus on the intersection of human security and economic growth (Natalie, 2021). The FSI Index assigns a ranking to countries and identifies key challenges. This makes it easier to use the data and conduct in-depth analysis. The following table summarizes the details of the fragility index.

Table 1: Description of state fragility variables

Variable	***Description***
Security apparatus	Consists of how state security institutions, such as police, military, and inelegance perform their duty in the country and state borders. They are the ones who maintain the state's monopoly of force in contrast to other groups.
Factionalized Elites:	It considers how the state institutions are fragmented along ethnic, racial or religious lines, as well as between ruling elites. It measures how power struggles, competitiveness, political shifts, and the location of elections will affect the legitimacy of democratic processes.
Group Grievance:	Focus on how groups in a state feel and perceive the system regarding historical injustice, development benefits and political resignations.
Economy Decline	Its variable indicates the performance of the economy in terms of GDP, unemployment, poverty level inflation, and productivity.
Economic Inequality	It looks at horizontal and vertical economic inequality between and among different groups in a state or region (urban-rural).
Human Flight and Brain Drain	It is about human displacement and migration forced or not (for economic or political reasons/ legal or illegal) and its impact on the country
State legitimacy	Citizens' level of confidence and trust in the government in power is expressed through election, demonstration, civil disobedience, or armed struggle.
Public Services:	It refers to how citizens get access to basic state services such as health care, education, electricity, transport, etc. It also includes protection from the state of their personal, and economic well-being.
Human Rights	This is related to how far the state protects and provides rights for its citizens.
Demographic Pressures	Measures population pressures related to the food supply, access to safe water, and other life-sustaining resources, or health, such as the prevalence of disease and epidemics
External Intervention	These indicators measure the intervention of foreign states and non-state actors in the internal affairs of a state.

Democracy

Finding an agreed-upon definition of democracy is difficult. It is more convincing to explain that democracy is a system, a process, and a set of activities aimed at institutionalizing and establishing structures to defend rights and freedom. The simplest definition of democracy requires a majority consent government that protects minorities and holds free and fair elections (Paul and Kirsten, 2008). Democracy necessitates the observance of laws, principles, and norms, as well as activities based on due process of law. (Amy, Ronald, and Christian: 2011).

Democracy indexes are developed to measure the level and rank of state democratic practices. For instance, the Economist Intelligence Unit (EIU) in the UK, produces an index for state. (Laza, 2019). It was in 2006, the Democracy Index report first emerged. It has over 60 indicators grouped under electoral process and pluralism, civil liberties, functioning of government, political participation and civil liberties. (Laza, 2019)

Table 2: Categories of Democracy Index

Variable	Description
Electoral Process and Pluralism	All processes that led to an election such as registration, camping, debate, voting, counting of ballots, etc. It also includes the idea that whoever gets the majority wins.
Civil liberties:	Those protected rights and freedoms free from government intervention and interference such as freedom of religion, assembly, speech, etc.
Functioning of Government	The legislative, executive and implementation power of the government.
Political participation	This includes participating in elections and other political activities such as mobilizing political resources for election, giving political opinions etc
Political culture	It refers to the attitudes and ideas that give structure and purpose to a political process, as well as the underlying assumptions and regulations that govern conduct in the political system.

2.3.2. Social factors: Unemployment, Population Growth

Unemployment: Unemployment is often examined in terms of the number of human labour force available for and seeking employment. It is a challenge for many countries, including Ethiopia. The nature of unemployment is the movement of the labour force from rural to urban areas in search of better opportunities. A multitude of factors contribute to this development; such as the dispossession of farmers, violent conflicts, economic hardships, and aspirations for better pay jobs in cities. In rural areas, this creates struggles among the peasant community and crimes and violence in urban areas. (Belay et al, 2020)**.** The role of the state in the process is paramount. The expansion of large-scale agricultural farming, where technology is more important than human labour shifted both the land distribution and the labour force. This has created dissatisfaction with the state and its policies, which caused joblessness and economic inequality (Belay et al, 2020). ***Population growth:*** It is the increase in the number of people in the population. If a country doesn't have a stable population growth that can go head-

to-head with the population of the country the country will face a scarcity of resources and would fail into deep poverty.

2.3.3. Economic factors: GDP per capita, Human Development

GDP per capita: **This** indicator measures the levels of Gross Domestic Product (GDP) per capita. This can be calculated by dividing GDP by the population of the country. It is mostly employed as an economic indicator to see the total economic output relative to the number of the population of a country. It reflects changes in the total well-being of the population. Arise in GDP per capita represents the rate of income growth per person. It is a potent summary indicator of economic progress when used as a single composite indicator (Rafael and Ferraz (2015). It does not necessarily demonstrate economic sustainability, but it is an important metric for a country's economy (General Authority for Statistics,2016).

Human Development: In contrast to GDP, human development is more related to individuals or citizens of a country. The UNDP, the UN's agency, produces a yearly report to assess development goals (SDGs) progress at the country and global level (Human Development Report, 2022). The report ranks countries based on how far they went in realising human development issues such as gender equality, employment, security, etc. (Amie, 2011). Understanding the country's level of human development would enable us to understand how much economic welfare it provides to its citizens and to know future development areas that require attention (Milorad, 2019).

2.3.4. Environmental Factors: Precipitation and Agricultural Land

Precipitation: In a given year, levels of precipitation can be measured. It is key to economic development. If societies and communities face low levels of perception drought and famine are more likely to happen. This has an impact on stability and conflicts. **Agricultural Land:** Societies need to have arable agricultural land to sustain their life. Given its importance land is a cause for conflict when people compete for the control of land. But it is also connected to perception; if there is low precipitation the capacity of the land to cultivate crops or other agriculture actives would decline.

2.4. Data Analytics, conflict prediction and conflict early warning

According to Sharma, Bhanu Tokas and Leo (2021), data analytics is a method of applying quantitative and qualitative techniques to analyse data, aiming for valuable insights. They also described that with the help of data analytics, we can describe data (descriptive data analytics), predict data (predictive data analysis) and we can even recommend the best action to be taken (prescriptive data analysis). Nowadays data science is exerting a substantial influence on different industries including tasks that historically relied on human judgment (Anusua et al, 2019). One of the areas in which data analytics has important is the conflict early warning system. It is used in the fields of conflict, state failure, natural disasters, and humanitarian emergencies.

It is lately the application of data analytics introduced in conflict early warning systems in contrast to other fields. The origin of such systems dates back to the 1950s but the market for conflict early warning systems is massive and still growing. According to (Rinaldi et al, 2022) Due to the wide range and complex nature of conflicts the interest in data analytics to build mathematical models for conflict prevention/ conflict early warning has shown a gradual increase in the last decades. This kind of system (Conflict early warning system) is developed to prevent conflict before occurring and its importance lies in providing variable success to predict conflict trends, alert communities to risk, inform decision-makers, provide inputs to action strategies, and initiate a response to violent conflict. (Herbert and Debiel, 2009)

Recent developments in data science have given a new opportunity for governments to approach conflict prevention, shifting from a speculative to a data-driven approach. Two tendencies have been noted in the evolution of data analytics. The first was an advancement in techniques, which implies that studies combined both qualitative and quantitative approaches. Second, numerous regional organizations, particularly in Africa, have begun to create early warning systems, partially at the request of and with the support of donor organizations. As described above. The predictive capacities of the systems have greatly improved over the last two decades. These capacities make the conflict early warning systems result used in early preventive action (Herbert and Debiel, 2009). The main purpose of the Conflict Early Warning System is to anticipate and prevent conflict from turning violent. Several regional and sub-regional organisations have early warning mechanisms such as the African Union (AU) and the Intergovernmental Authority on Development in Eastern Africa (IGAD).

2.5. Empirical research related to data analytics and conflict

Several researchers have used data analytics tools, approaches and models to predict violent conflicts and offer early warning recommendations. One of the most important pieces of literature on this topic is developed by Chris Perry (2019). He integrated machine learning methodologies into conflict analysis and tested the added value and importance of this approach. Chris Perry claims that selecting proper machine learning approaches may bring significant gains in accuracy and performance, and argues that comprehensive models have greater predictive value than merely considering a previous outbreak of violence as the leading predictor of present violence (Perry, 2019). Havard et al (2021) evaluated 9 years of (2010–2018) conflict predictions model made by Hegre et al. (2013) in 2011. Using multiple metrics, they evaluated the ability of this study to predict observed conflicts. They contend that Hegre and others made important and relatively accurate predictions of armed conflict. In contrast to significant armed conflict, they found that the model fared poorer in forecasting low-level conflict incidence. Their research provides an important contribution to future research and signifies the utility of predictive models for both testing and developing theory.

Valeria, Andra-Stefania and Jakob (2018) have carried out a project known as ViEWS+ to expand the software functionality of the Violence Early Warning System. The aim was to predict the probability of armed conflicts in the 3 years in the future. This was to assist policy-makers with conflict prediction models in their aid and humanitarian work. Their predictions

use conflict data including variables like past conflicts, child mortality and urban density. They made an automatic variable selection tool to use more relevant variables only which helped them to save time and resources. They also made a comparison of prediction functions and selected the best model. Finally, they tested how parameter values affect the performance of the chosen functions. For the Classifier Comparison, they used Jupyter Notebook which serves to compare several different machine learning functions. This is done by performing several predictions with each function and comparing the final scores to one another. All functions use their default parameter settings. Random Forest (RF) has the highest F score and is selected as the function to be used for predictions (Helle et al, 2018). Finally, the new tools they created enhanced both the execution speed and the prediction accuracy of the system when compared to previous findings. They found it 9 times faster than it was previously, and its accuracy has increased by a factor of three. (Ibid).

Christy (2017) has conducted research on Conflict Early Warning Systems for UN peacekeeping operations. The UN is responsible for deploying peacekeeping missions to protect civilians in conflict areas. Its role might be affected by the lack of robust conflict early warning systems. Christy (2017) criticizes the UN's early warning system for being subjective and unsystematic in its ability to generate accurate forecasts across different areas of the peacekeeping hosting country. He attempted to investigate models that use machine learning approaches, such as the Logistic regression model, the Lasso model, and the Random Forest. He attempts to evaluate the models' performance by examining how well they predict violence against civilians events in three case countries in Africa that currently host UN peacekeeping missions: the Democratic Republic of the Congo (DRC), the Central African Republic (CAR), and Somalia. In addition, he examines the political supportability and administrative feasibility of a new data-driven system in comparison with their current system. His analysis shows that the machine learning models have the potential to add considerable value to the UN's current system. In the best cases, machine learning models predict conflict outbreaks at an accuracy rate of 90%. Furthermore, there was also support for data-driven early warning within UN mission headquarters if a new system is user-friendly and can reduce administrative strain within the mission. A new data-driven system using machine learning models would require additional analytical capacity, but the data collection and processing capabilities are already in place within UN peacekeeping missions. Finally based on the scholars analysis and recommendations, the UN has incorporated machine learning models into its existing early warning system. The lasso model performs best in all cases on predicting conflict occurrence. (Christy, 2017).

Musumba and others (2021) explored the conflict predictive performance of logistic mode with other supervised classification machine learning (ML) algorithms. Focusing on 48 African countries, using a grid-level dataset of 5928 observations they found that gradient tree boosting is the best algorithm when the performance metric is recall but the multilayer perceptron algorithm produces the best model when precision or F1 score is the selected metric. They argued that if the accurate algorithm is selected, the effects of post-conflict state reconstruction such as social and economic instabilities can be reduced.

Table 3: Summaries of empirical research

Author	Year	Title	Method/machine learning algorithm	Research focus region
Christy	2017	Improving Conflict Early Warning Systems for United Nations Peacekeeping	Lasso model	D. R. Congo, the Central African Republic, Somalia
Valeria Helle, Andra-Stefania Negus and Jakob Nyberg	2018	Improving armed conflict prediction using machine learning	Random Forest	European countries
Chris Perry	2019	Machine Learning and Conflict Prediction: A Use Case	naïve Bayes, random forest	Africa
Si Chen, Xuedong Cai, Bo Li, and Shenzhen Hou	2019	Community conflict prediction method based on spliced BiLSTM	bidirectional LSTM	
Shallcross, and Ahner	2020	Predictive models of world conflict:		Global
Ettensperger	2020	Comparing supervised learning algorithmsand artificial neural networks for conflict prediction:	regression trees, k-nearest neighbour, random forest and neural networks	Global
Musumba, Mark, Naureen Fatema, and Shahriar Kibriya	2021	Prevention Is Better Than Cure: Machine Learning Approach to Conflict Prediction in Sub-Saharan Africa	gradient boosting	African
Havard et al	2021	*Can we predict civil war?*	multinomial logit model	Global
Darius and orgen	2021	Predicting Conflictin Africa- Leveraging Open Geodata and Deep Learning for Spatio-Temporal Event Detection	deep learning models	Africa
Havelange and Corentin	2021	"Predicting armed conflicts: a machine learning approach"	random forest model	Global
Seblewongel Habtemicheal	2022	The predictive model to identify effective conflict resolution methods at the institutional level	Support vector machine	Ethiopia
Linke, Witmer and O'Loughlin	2022	Weather variability and conflict forecasts: Dynamic human-environment interactions in Kenya	Random forest	Kenya
Li and Chen	2022	A Distributed Task Scheduling Method Based on Conflict Prediction for Ad Hoc UAV Swarms	Performance Impact (PI) algorithm	Global
Bazzi et al.	2022	The promise and pitfalls of conflict prediction: evidence from Colombia and Indonesia	Bayesian Model Average	Colombia and Indonesia
Lindholm, Hendriks, Wills and Schön	2022	Predicting political violence using a state-space model	state-space model	---
Zhang and Abdel-Aty	2022	Real-Time Pedestrian Conflict Prediction Model at the Signal Cycle Level Using Machine Learning Models	Gradient Boosting	---
Neumann,S., Ahner, D. and Hill, R.R.	2022	Forecasting country conflict using statistical learning methods	K-means clustering	United States
Felix Ettensperger	2022	Forecasting conflict using a diverse machine-learning ensemble:	multi-model ensemble learning techniques	Egypt, Cameroon and Mozambique.
Håvard Hegre, Paola Vesco & Michael Colaresi	2022	Lessons from an escalation prediction competition	Random forest regression models	African
Fulvio A., Marcello C. & Stefano M.	2022	Forecasting change in conflict fatalities with dynamic elastic net	Dynamic Elastic Net	African
Iris Malone	2022	Recurrent neural networks for conflict forecasting	Recurrent neural networks	Egypt, Cameroon, and Mozambique.
David Randahl & Johan Vegelius	2022	Predicting escalating and de-escalating violence in Africa using Markov models	Hidden Markov models	Africa
Hannes Mueller & Christopher Rauh	2022	Using past violence and current news to predict changes in violence	Random forest regression	Africa

Si Chen et al. (2019) propose a conflict prediction approach on the bases of spliced bidirectional LSTM. By using the approach, they first utilized two bidirectional LSTMs to process word embedding and graphs to break the temporal dependency. This was followed, by weighting the hidden states of the two bidirectional LSTMs, the conflict prediction was made by the neural network. The final experimental results show that there is an improvement in the AUC value. Linke, Witmer, and O'Loughlin (2022) used random forest (CRF) for predicting local conflicts. Special attention was given to see whether environmental data (weather variability and vegetation health) help to enhance the performance of the random forest model with 29 demographic and contextual variables. They found that their prediction of the 2018 conflict in rural Kenya was not enhanced by adding environmental predictors. Using only environmental data for conflict prediction is the worst according to their finding. Economic, social, and political variables have a better predictive value than weather alone.

Felix Ettensperger (2022) adopted multi-model ensemble techniques for conflict prediction in 54 African counties for six months to the future. They combined six different models from tree-based, distinct data foundation and geographical selection models for a unified forecast-ing framework. After building the model they have evaluated the performance of the model. For the presentation and examination purposes finally from a total of 54 African countries they have forecasted the conflict intensity for Egypt, Cameroon, and Mozambique. Håvard, Paola and Michael (2022) aim to address conflict prediction by identifying 25 causalities using machine learning techniques for the next four years. The identified causes are merely based on social, demographic, economic, geographic and political national indicators measured every year. From those identified indicators they have identified the significance of each of the indicators. Finally, they applied logistic regression and a random forest model for classification purposes. But they obtain better performance with the random forest model.

Focusing on Ethiopia, Seblewongel Habtemicheal in her MSc thesis has researched a prediction model to identify effective conflict resolution methods. Her research focuses on the social conflict at the institutional level and aims to develop a machine learning model for the prediction of effective conflict handling or resolution methods in Technical Vocational Educational and Training (TVET) colleges. She used five machine learning algorithms: Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbours (KNN), and Logistic Regression (LR). The result shows a classification accuracy of SVM 92.52%, DT 92.34%, NB 92.25%, KNN 91.37% and LR 92.16%. SVM outperformed better than the other machine learning algorithms. SVM works better for categorical and high-dimensional data. Even though this research is relevant methodologically, it doesn't cover broader conflict dynamics in Ethiopia.

This research focuses specifically on violent conflicts in Ethiopia which are happening across multiple geographies and times unlike other research stated above which focuses on a different context. Most studies on conflict and data analytics (Valeria Helle, Andra-Stefania Negus and Jakob Nyberg, 2018, Iris Malone, 2022 and Bazzi et al., 2022) focuses on single class (spatial or temporal). But lessons from previous research in other contexts and states will be used and incorporated.

CHAPTER THREE: RESEARCH METHODOLOGY

Introduction

As a structured academic approach, research methodology facilitates the conduct of research activities in a manner that is consistent and repeatable (Mimansha and Nitinl, 2019). Developing a clear methodology helps to show how the researcher ended up with the results (Chinelo, 2016), and thus other researchers can easily adapt, follow or even challenge the methodology and produce the same or different research findings (Sam, 2019).

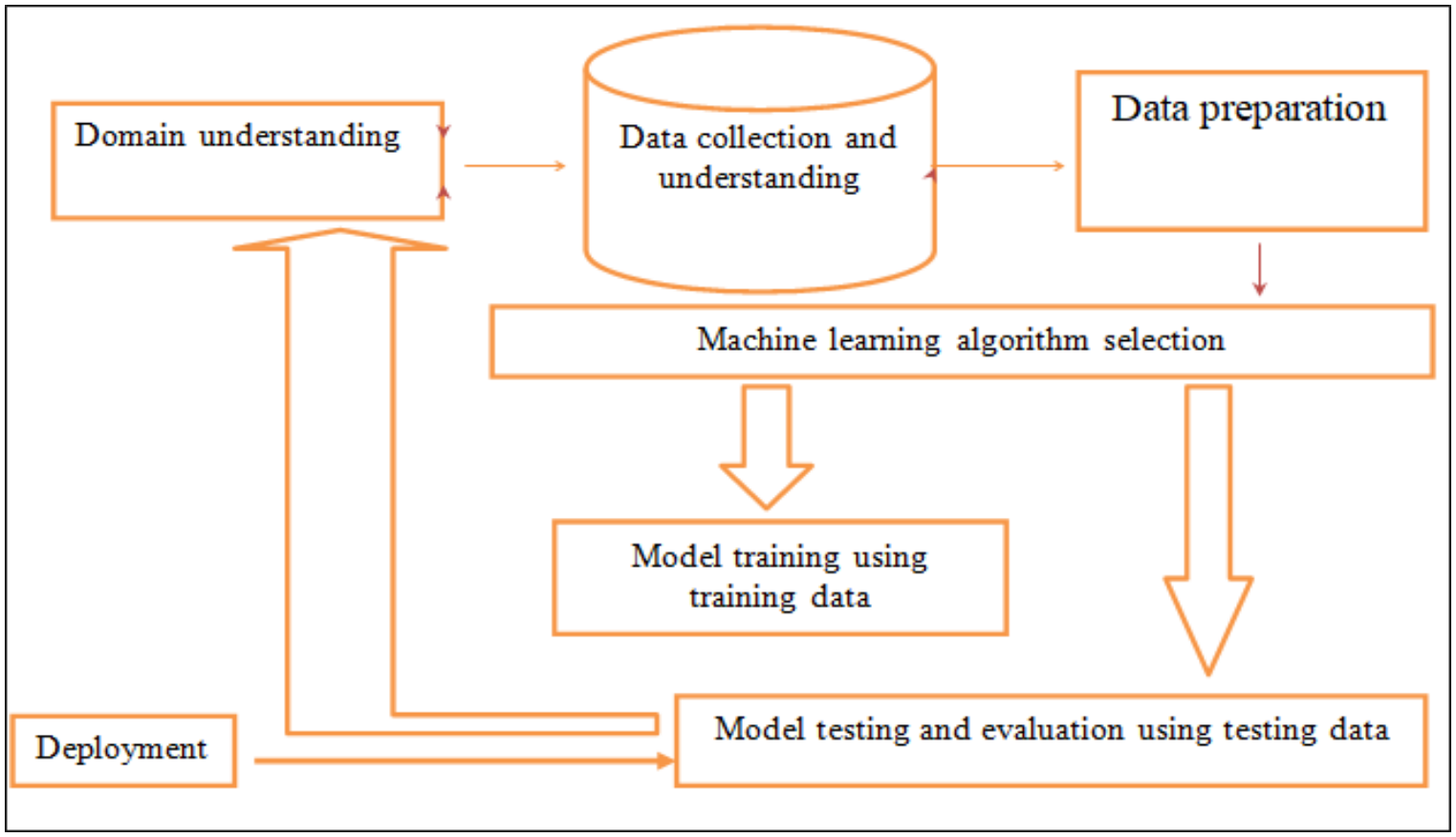
In data analytics research, like other areas, methodology plays a vital role, since the accuracy, credibility and reliability of the results are influenced by the methods and procedures employed. Conflict prediction and early warning system research, like this thesis, requires passing via a series of procedures starting from conducting a detailed literature review to understand the domain, identifying and analysing the list of conflict indicators and finally building, as well as modelling, an effective early warning system by predicting the internal violent conflict. This chapter presents a detailed discussion of the methodology employed in this research.

3.1 Research Design

The research employs a quantitative research design in general and an experimental research design (hereafter referred to as ERD) in particular. It is commonly observed that research in the field of machine learning mainly employs quantitative design since it involves the modelling of data and the use of statistical approaches and formulations (Özer, 2019). The choice of this research design in this study is due to its nature and value in offering the option to conduct an experiment, obtain results by gathering quantitative data and conduct statistical analysis. The research aims to observe the influence of an independent variable on the dependent one and build a machine learning model using different machine learning algorithms that will help to select the best model to develop an effective conflict early warning system.

The use of ERD is based on the strong assumption that it will provide an opportunity and a scientific tool to experiment with different machine learning algorithms and identify the best model that would effectively offer conflict prediction of country-specific violent conflicts taking Ethiopia as a case study. The analytics life cycle is mostly related to Cross Industry Standard Process for machine learning (CRISP-ML), a widely accepted methodology for building models. The CRISP-ML is a process model that serves as the base for a machine learning process. This scientific research tool is drawn from practical experience and literature and has proven to be general and stable (Stefan et al., 2021).

Figure 1: CRISP-ML used in this research project



Source: Source: Stefan Studer et al

As shown in the above figure, the CRISP-ML process model is expanded from Cross Industry Standard Process for Data Mining (CRISP-DM) (Ibid). What makes CRISP-ML more relevant and desirable is that the standard processes help to deal with multidisciplinary data analytics research, like this MSc research, as noted in the first chapter this research touches upon a multidisciplinary research area in the field of data analytics and peace/conflict research. The selected research design is considered more suitable for the research.

3.2. Domain understanding

Data science is a discipline that uses tools to model data, generates insights from data, and makes decisions based on the available data (Dirk et al, 2022). These tools can be applied to many academic fields such as engineering, law, medicine, finance, environment, peace and conflict etc. Data science research, therefore, requires an assessment of the general subject matter to which data science methods and tools are applied. Domain understanding refers to a general understanding of the background knowledge of the field or environment (Avrim, John and Ravindran, 2018). In this research, the domain in which the data science tool is applied is the peace and conflict field of study, and as discussed in the previous chapters, understandings about theoretical and conceptual aspects of peace and conflict are already brought into the research.

3.3. Method of Data Collection and Understanding

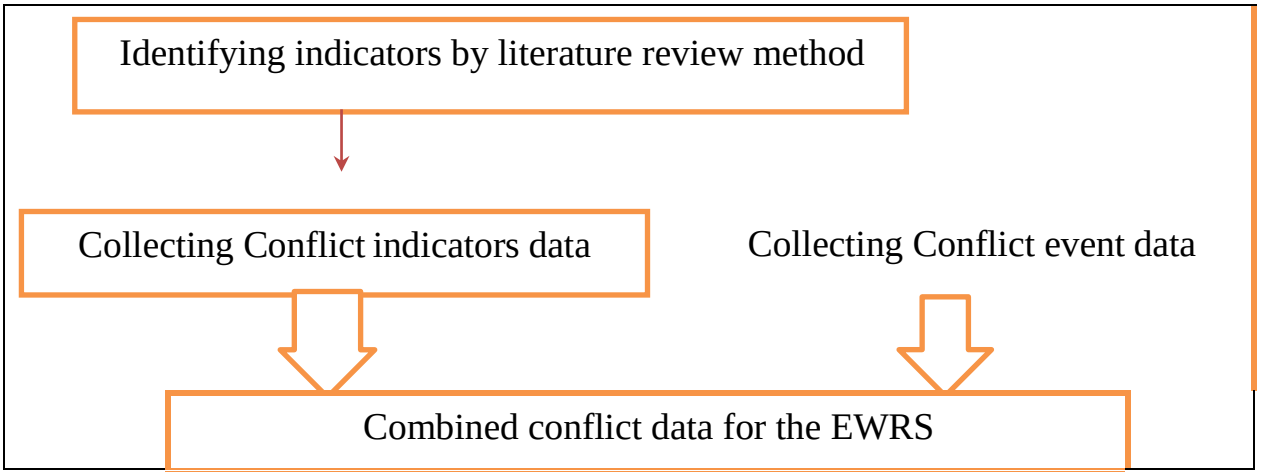
3.3.1. Method of Data Collection

A machine learning model learns from historical data, either primary or secondary. A wide pool of knowledge is created from which machines can learn and make decisions (Özer, 2019). This research has employed secondary data sources. The secondary data set was obtained from different data repositories.

As shown in Tables 4 and 5, databases which are run by different organizations are used as major sources of data. Those organizations include the Armed Conflict Location and Event Data Project (ACLED), Ethiopian Peace Observatory (EPO), World Bank, Fund for Peace (FFP), Economic Intelligence Unit (EIU), Transparency International, UNDP, United Nations Population Division (UNPD) and The Harvard Data Verse Repository.

The database of the Armed Conflict Location and Event Data Project (ACLED) is used to see and extract data about conflict trends in the last fifteen years (2007- 2022) and it will be presented and discussed using the form of a graph in the next chapter. The other sources of databases are used to collect data about the different variables which are taken as indicatorsof conflicts in light of the theoretical and conceptual understanding of the conflict. The process is illustrated in Figure 2.

Figure 2: Method of data preparation



Source: developed by the researcher based on literature review

Since data analytics research deals with a variety of data and uses various mathematical formulas to discover the best decision for a given situation using machine learning algorithms, the first step is to have adequate sources of data (Kauser and Acharjya, 2016). Based on this assumption, I have tried to find different data sources that could fit the research. Based on the research focus and literature reviewed, dependent variables and independent variables are identified. Conflicts which happened each year are considered the dependent variable. The conflict events data of Ethiopia is found in the Ethiopian Peace Observatory (EPO) database within the period (1997-2022). From the data source, the conflict event data is collected through internet downloading. The independent variables are those variables that provide conditions for violent conflicts to occur. Based on the literature reviewed, the drivers of conflict

are generally categorized into political, economic, social and environmental. This was followed by identifying key indicators related to each driver of conflict.

To identify drivers of conflict and specific indicators, books, articles, manuals and reports have been extensively assessed and incorporated into this research. During this process, the issue of validity was taken seriously. To ensure the credibility of the review, reference checks, conformability, and balance were given due attention. The literature review covered various materials that include program documents, project evaluation reports, research project publications, policy, periodic publications, and proceedings from academic and research institutions. Print and electronic materials are considered reliable sources of information in reporting conflict events. The literature review served to examine the history and causes of the conflict to date and physical and demographic features relevant to conflict analysis and to understand the national context including drivers of conflicts, and trends (spatial and temporal trends). After a detailed literature review process, and identification of the indicators of conflict, data was collected from different organizations' databases using internet search and downloading. Finally, a complete list of datasets was established.

3.3.2. Understanding the data

After completing the data collection process, the next most important phase is to understand the data itself. Two types of data are composed: the first data type is an indicator of conflict and numerical presented. The second data is violent conflict events data which is a categorical data type. The summary of the data collected from different repositories and their descriptions is shown in Table 4 and Table 5 below.

Table 4: Conflict events data

Data	***Temporal coverage***	***Data source***	***Description***
Ethiopian peace observatory (EPO) database	1997- 2022	The armed conflict location and Event data project (ACLED)	a disaggregated data collection, analysis, and crisis mapping project. It has data on, actors, areas, periods, locations, costs, and types of political violence across the world.

Table 5: Conflict indicators data description

Drivers of Conflict	*Indicators*	*Data*	*Temporal coverage*	*Data source*	*Description*
Political	Governance	Worldwide Governance Indicators (WGI)	1996 -2021	World Bank	A summary of the quality of governance in developed and developing countries. It uses data from institutes, think tanks, NGOs, international organizations, etc. Available at: http://info.worldbank.org/governance/wgi/#home
	State Fragility	Fragile State Index	2006- 2022	The Fund for Peace (FFP)	The fragile state index is focused on issues of violent conflict, state fragility, and securityand human rights. Available at: https://fragilestatesindex.org/data/
	Democracy	Democracy index	2006- 2021	Economic Intelligence Unit (EIU)	The democracy index combines information on the extent to which citizens can choose their political leaders in free and fair elections, enjoy civil liberties, prefer democracy over another political system, can and do participate in politics, and have a functioning government that acts on their behalf. Available at: https://ourworldindata.org/grapher/democracy-eiu?tab=chart&country=~ETH
Economic	GDP	World Developmt Indicators	1990- 2020	World Bank	The sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. Available at: https://ourworldindata.org/grapher/gdp-per-capita-worldbank?tab=chart&country=ETH
	Economic inequality	Income inequality	1995-2015	World Bank	The data measures income inequality using the Gini coefficient on a scale between 0 and 1, where higher values indicate greater inequality. https://ourworldindata.org/grapher/economic-inequality-gini-index?tab=chart&country=ETH
	Corruption	Corruption Perception index	2012- 2018	Transparency International	Yearly ranking of countries by their perceived levels of corruption. It uses expert and opinion surveys. https://ourworldindata.org/grapher/ti-corruption-perception-index?tab=chart&country=ETH
	Human Developmentt	Human Development Index	2000-2021	UNDP	a measure of achievements in human development: A long and healthy life, access to knowledge and a decent standard of living. https://ourworldindata.org/grapher/human-development-index?tab=chart&country=ETH
Social	Population growth	United Nations Population Record	1950- 2022	United Nations - Population Division	The data provides population growth by country yearly. Available at: https://ourworldindata.org/grapher/population?tab=chart&country=ETH

	Ethnic fractionalization	Historical Index of Ethnic Fractionalization Dataset (HIEF)	1945-2013	The Harvard Data Verse Repository	An extension of previous ethnic fractionalization indices. It helps to compare issues in ethnic fractionalization over time. Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2F4JQRCL
	Unemployment rate	World Development Indicators	1991- 2021	World Bank	The data refers to the share of the labour force that is without work but available for and seeking employment. https://ourworldindata.org/grapher/unemployment-rate?tab=chart&country=ETH
	Poverty	Poverty gap index	1995-2015	World Bank	The poverty gap index is the mean shortfall from the poverty line counting the non-poor as having zero shortfalls and expressed as a percentage of the poverty line. This data is adjusted for inflation and for differences in the cost of living between countries. Available at:https://ourworldindata.org/grapher/poverty-gap-index- at-190-int-per-day povcal? tab= chart
Environmental	precipitation		1991-2021	World Bank	The precipitation data contains complete sub-national aggregation as well as seasonal data for all precipitation variables and continue to update as new offerings are produced. Available at: https://climateknowledgeportal.worldbank.org/download
	Agricultural land (sq. km)		1991-2021	World Bank	This refers to the share of land area that is arable, crops and pastures use. But, it doesn't include wood or timber-covered lands. Available at: https://data.worldbank.org/indicator/AG.LND.AGRI.

3.3.3. Methods of Data Analysis

This section presents the method of data analysis; the steps that have been tracked to solve the problem, methods to apply exploratory data analysis, data pre-processing, the ratio between training data and testing data, choice of machine learning algorithms used in the model and the performance measures of the algorithms are discussed.

a. Exploratory data analysis

As Matthieu et al (2016) indicated the need to have a proper data exploration since it greatly impacts the accuracy of learning predictions. This is because in data exploration important aspects of a data set are brought into focus for further analysis by describing the data using statistical and visualization techniques. In this paper on the data exploration stage mainly visualization of dataset using different visualization techniques and correlating the data sets is carried out. This phase helps to get familiar with the data, understanding correlations in the data and identify the most salient drivers of violent conflict in Ethiopia. For the exploratory data analysis, the conflict event dataset and conflict indicators

dataset are both utilized (as described in Section 3.4). The main reason to employ two different datasets is the inadequacy of the conflict events data set to indicate the drivers and conditions for the conflict events. In the conflict event dataset, the emphasis was less on the drivers of conflict and more on the actors, the locations, and the types of conflicts. By combining conflict events and conflict indicators data sets, the research aims to have the full picture of the violent conflicts that occurred and their causes. Using Python, I run exploratory analyses on both sets of data.

b. Data Preprocessing and splitting of the data for model building

- **Data Preprocessing**

Data preprocessing is a fundamental stage of data analysis. It helps to describe any type of processing performed on raw data and in turn to prepare it for another data processing procedure (Ho and Chong, 2010). Its importance in machine learning is well emphasised in academic research because this stage can influence the performance of models and consequently influence the results, given the fact that machine learning algorithms learn from the data provided to them and use the acquired knowledge to make decisions. Therefore, data preprocessing is an essential stage not to be overlooked, regardless of the type of algorithms used in the research (supervised, unsupervised, or reinforcement learning) or the nature of the research itself (classification or regression) (Kotsiantis, Kanellopoulos and Pinte, 2006; Kamiran and Calders, 2011; Smola and Vishwanathan, 2008).

Figure 3: Data preprocessing techniques

Source: Prepared by the researcher based on concepts from Kamiran and Calders (2011)

Developing a working dataset requires several different processes applied to the available data (Kotsiantis, Kanellopoulos and Pinte, 2006). Using the existing literature on this method, as shown in Figure 3, this research uses four main data preprocessing techniques: data integration, data quality assessments, data cleaning and data transformation will be done by utilizing a programming language called Python (Python 3.7).

- ✓ **Data Integration:** Data integration is a vital part of the data preprocessing process. Having in mind that this stage could result in duplicate and inconsistent data, thus resulting in poor data model accuracy and speed, approaches like data quality assessments and Data cleaning are necessary to address the gaps and maintain data integrity. In this step first the independent variables are integrated using the merge python library. After that the merged independent variables are merged with the dependent variables using excel sheet.
- ✓ **Data quality assessments**: In this step of data preprocessing, the data was carefully examined and the researcher obtained an idea of its overall quality, consistency and relevance to the research in hand. The main goal of taking this step was to check if there were any mismatched data types, mixed data values, data outliers and missing data. By identifying those data anomalies my data was prepared for the other preprocessing steps.
- ✓ **Data cleaning:** Data cleaning is the practice of correcting or deleting incorrect, corrupted, improperly formatted, duplicate, or incomplete data from a dataset. Based on the definition, in this stage outliers were removed, missing values were replaced, noisy data were smoothed, and inconsistent data were corrected using data cleaning techniques.
- ✓ **Data Transformation:** Data transformation involves the process of turning the data into the proper format(s) that is appropriate for modelling. Strategies that are applied in this paper to enable data transformation include label encoding, data mapping, aggregation, normalization, feature selection and discretization.

- ➢ **Data preparation for model fitting**

Preparing the data for model fitting follows the acquisition of preprocessed data. It is for this purpose that the dataset-splitting technique has been used to train and test the models effectively with different selected algorithms.

Dataset Splitting: When machine learning algorithms are used to make predictions on data that was not used to train the model, the train-test split process is used to measure their performance (Borislava, 2021). It's a quick and simple technique to compare the performance of different machine learning algorithms for predictive modelling research. This technique can be used for any supervised learning technique and can be utilized for classification or regression tasks (Anita, Dávid and Károly, 2021). A dataset will be divided into two subsets. The training dataset is the first subset, which is used to fit the model. The second subset is not used to train the model; instead, the dataset's input element is given to the model which then makes predictions and compares them to the predicted values.

The goal of splitting the dataset is to estimate the machine-learning model's performance on additional data that was not used to train the model. This is how we want to use the model in the real world. Specifically, to fit it to existing data with known inputs and outputs, and then generate predictions for future examples where we do not have the expected output or goal values. The train-test procedure is appropriate when there is a sufficiently large dataset available. The train-test split evaluation technique is implemented in the scikit-learn Python machine learning toolkit. The program takes a loaded dataset as an argument and splits it into two subsets (Kamila, Pawluszek and Andrzej, 2020).

c. Machine learning algorithm selection

The study employs Predictive Analytics, which is the process of dealing with a variety of data and applying various mathematical formulas to discover the best decision for a given problem. It will use a supervised classification machine-learning technique (Rustagi and Goel, 2022). Classification methods create classes by examining already classified cases and inductively finding the pattern typical to each class. The process has two steps:: First, a model is constructed by analysing the conflict dataset tuples described by the attributes. Each tuple is assigned to a predefined class, as determined by one of the attributes, called the class label attribute. The conflict data tuples are analysed to build the model collectively from the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the dataset. Second, the model is used for classification. The holdout method is a simple technique that uses a test set of class-labelled samples. These samples are randomly selected and are independent of the training samples. Then the accuracy of a model on a given test set will be evaluated. To apply a supervised learning application, this study has adopted the Random Forest, Gradient Boosting and Gaussian Navies Bayes.

Model evaluation

Model evaluation is known as Performance Evaluation for the Predictive Model. This activity is responsible for describing the evaluation parameters of the designed model (Kroese, 2019). Evaluation of the model is made with the evaluation parameter that compares the number of the data, which is predicted correctly and incorrectly. The comparison was done between the data categorized by the proposed model system and that of the manually labelled (categorized) data (Kroese, 2019). In this research, the confusion matrix, accuracy, precision, recall, F-score and mean square error (MSE) is applied to evaluate the model.

- **Confusion matrix:** This matrix will evaluate the performance of the classification model on a testing dataset. According to Zheng (2015), the confusion matrix has four labels:
  - True Positives (TP): Cases when the classifier predicted the true and correct class was true.
  - True Negatives (TN): Cases when the model predicted false and the correct class was false.
  - False Positives (FP) (Type I error): Classes predicted true but the correct class was false.
  - False Negatives (FN) (Type II error): The classifier predicted false but the correct class was false
- **Accuracy:** it measures the number of all incorrectly classified samples divided by the total number of samples in the dataset. Accuracy has the best value of one and the worst value of zero (Powers, 2011).
- **Precision (P):** precision is the fraction or percentage of identified or retrieved instances that the classification algorithm considers important. High precision means that the majority of items labelled for instance positive indeed belong to the class positive and defined as the precision is characterized as the number of true positives isolated by the whole sum of true positives and false positives (Zheng, 2015; Powers, 2011).

- **Recall:** A recall is considered a measure of completeness, which is the level of positive examples that are marked as positive. The review of grouping is characterized as the number of true positives isolated by the all-out number of components that have a place with the positive classes (Zheng, 2015).
- **F-score:** F- Measure (F1 score) is defined as the harmonic means of precision and recall which is a measure that joins recall and precision into a single measure of performance (Zheng, 2015).

To sum up, to achieve the objectives stated in the first chapter, this research makes use of different machine learning algorithms to observe the influence of an independent variable on the dependent one that will help to select the best model using the model evaluation tools including confusion matrix, accuracy, precision and recall, that helps to develop an effective conflict early warning system focusing on Ethiopia.

CHAPTERFOUR: EXPLORATORYDATAANALYSIS AND PREPROCESSING

Introduction

In this research, different datasets from different repositories have been employed. In this section, these data sources will be explored and preprocessed for making the dataset compatible with the next data analytics process which is model building.

4.1. Exploring and preprocessing independent variables

As noted, this research follows the experimental design and specifically the Cross Industry Standard Process for machine learning (CRISP-ML). The Exploratory Data Analysis and the Data Preprocessing steps are illustrated below, and Python programming with the Jupiter Notebooks interface is used to implement and analyse the data. Since the research aims to train the best-performing model possible that can effectively and efficiently predict internal violent conflicts in Ethiopia, the first two steps such as exploratory analysis and data preprocessing very crucial to build the model. These two processes help to have well-explored and preprocessed data that a machine can understand and learn. These steps can greatly impact the accuracy of the model predictions because machine learning algorithms learn from the data provided to them and use the acquired knowledge to make decisions. Accordingly, detailed data exploratory data analysis and data preprocessing are crucial.

In this section, the steps followed to apply the EDA and preprocessing techniques including their results are described. Different datasets are used. The conflict indicator datasets are loaded, explored and preprocessed. The conflict indicators dataset consists of all indicators (N=26) that may contribute to violent conflict in Ethiopia, but it does not have the conflict events data.

4.1.1. Importing conflict indicators dataset and exploring independent variables

The main drivers of conflict are seen as political, economic, social and environmental. These selected conflict drivers have their variable indicators, the variable indicator under each indicator is considered the main indicator of conflict and merged to understand the possible causes of conflict. In the following sections, an exploratory analysis of each indicator's dataset is shown.

For analysis purposes, Jupyter Notebook is used. It is considered a tool using Python programing language because Scikit-learn has an environment that provides a state-of-the-art implementation of tools and an easy-to-use interface tightly integrated with the Python language. In the Jupyter Notebook, the first main important step is to import the most important libraries in Python using the import keyword. All the libraries that are important for exploratory analysis and preprocessing steps are imported.

❖ **The political driver of conflict**

The political driver of conflict is operationalised with three indicators such as governance, state fragility and democracy. For those indicators, we have collected data from different data repositories which have a period from 2007 to 2021.

Table 5: Summary of political drivers of conflict

Driver	Indicators	Temporal coverage	Data source
Political	Governance	2007-2021	World Bank
	State fragility	2007-2021	The Fund for Peace (FFP)
	Democracy	2007-2021	Economic Intelligence Unit (EIU)

Some indicators have specific indicators variables that will explain the indicators. Those indicator variables are very important to understand the major causes of conflict. Accordingly, the following sub-sections explore each indicator in detail using Python Jupiter notebook and Excel, as well as making use of the different activities.

A. Governance indicator

Governance indicator is mainly associated with peace and conflict dynamics across states and societies. It is related to the process and procedures of election, monitoring, and replacement of government and how effectively it implements sound policies. In this regard, as pointedout in the table below, the World Bank has developed different variable indicators, which can help me to identify the specific area to be addressed.

Table 6: Governance indicators developed

Variable	Variable indicators
Governance	Voice and Accountability (Estimate)
	Political Stability (Estimate)
	Government Effectiveness (Estimate)
	Regulatory Quality
	Rule of Law
	Control of Corruption

Source: World Bank

Importing the governance indicator data and visualizing it in a tabular form before preprocessing is performed as follows.

Figure 4: Top 5 governance indicator data

```
In [3]: df1.head(5)
```

Out[3]:

	Country	year	Voice and Accountability	Political Stability	Government Effectiveness	Regulatory Quality	Rule of Law	Control of Corruption
0	Ethiopia	2007	-1.20	-1.81	-0.47	-0.98	-0.66	-0.63
1	Ethiopia	2008	-1.31	-1.73	-0.44	-0.90	-0.71	-0.67
2	Ethiopia	2009	-1.30	-1.64	-0.56	-0.93	-0.83	-0.70
3	Ethiopia	2010	-1.34	-1.64	-0.48	-0.85	-0.81	-0.69
4	Ethiopia	2011	-1.36	-1.51	-0.51	-0.99	-0.74	-0.67

B. State Fragility indicator

State fragility is highly related to the lack of will or capacity of the state to perform its core functions. States which are in a fragile condition lack state structures, political will and/or capacity to provide the basic functions needed for their society. To measure this situation, the Fund for Peace (FFP) developed different variable indicators. The variable indicators are very important to understand the main indicator that will lead a country to be fragile and in turn face violent conflict. The variable indicators of state fragility indicator are listed in the table below.

Table 7: State fragility indicators developed

Variable	Variable indicators
	Security apparatus
	Factionalized Elites
	Group Grievance
	Economy
	Economic Inequality
	Human Flight and Brain Drain
	State Legitimacy
	Public Services
	Human Rights
	Demographic Pressures
	Refugees and IDPs
	External Intervention

Source: The Fund for Peace (FFP)

Importing the state fragility indicator data and visualizing it in a tabular form before preprocessing is performed as follows.

Fig-5- Top 5 state fragility indicator data

```
In [5]: df2.head(5)
```

Out[5]:

	Country	year	security apparatus	Factionalized Elites	Group Grievance	Economy	Economic Inequality	Human Flight and Brain Drain	State Legitimacy	Public Services	Human Rights	Demographic Pressures	Refugees and IDPs	External Intervention
0	Ethiopia	2007	7.5	8.9	7.8	8.3	8.6	7.5	7.9	7.0	8.5	9.0	7.9	6.7
1	Ethiopia	2008	7.5	8.9	7.8	8.0	8.6	7.5	7.9	7.5	8.5	8.9	7.5	7.3
2	Ethiopia	2009	7.5	8.8	8.2	7.7	8.8	7.7	7.9	8.2	8.5	9.4	8.0	7.6
3	Ethiopia	2010	7.8	9.0	8.6	7.4	8.5	7.5	7.7	8.1	8.7	9.2	7.8	7.9
4	Ethiopia	2011	7.9	9.0	8.4	7.7	8.2	7.2	7.5	8.4	8.5	9.1	8.2	8.1

C. Democracy indicator

Democracy indicates a set of practices and principles of governance and the rule of law including the principle of majority rule, the existence of free and fair elections, the protection of minorities and respect for basic human rights. Based on those features the Economic Intelligence Unit (EIU) develops an index called the Democracy Index.

Table 8: Democracy Index

Variable	Variableindicators
Democracy	Democracy Index

Source: Economic Intelligence Unit (EIU)

Importing the democracy indicator data and visualizing it in a tabular form before preprocessing.

Fig-6- Top 5 democracy index data

```
In [7]: df3.head(5)
```

Out[7]:

	Country	year	Democracy index
0	Ethiopia	2007	NaN
1	Ethiopia	2008	4.52
2	Ethiopia	2009	NaN
3	Ethiopia	2010	3.68
4	Ethiopia	2011	3.79

❖ Economical drivers of conflict

To look at the economic drivers of conflict, three indicators are identified: Gross Domestic Product (GDP), Economic inequality, Corruption and Human Development. For those indicators, we have collected data from different data repositories which have a time span from 2007-2021.

Table 9: Economical drivers of conflict

Driver	Indicators	Temporal coverage	Data source
Economical	GDP	2007-2020	World Bank
	Economic inequality	2007-2021	World Bank
	Corruption	2007-2018	Transparency International
	Human Development	2007-2021	UNDP

Source: Developed by the researcher using different data sources

The data on the Economic drivers of conflict are collected from global international organizations. These economic indicators are very important to understand the economic causes of conflict. Each indicator of the economic drivers of conflict is discussed below in detail.

A. Gross Domestic Product (GDP) per capita

The World Bank annually publishes different world development indicators including the GDP per capita of every country. This indicator is a basic economic indicator and is used to measure the level of total economic output relative to the population of a country. It reflects changes in the total well-being of the population. The indicator measures the levels of GDP per capita that are obtained by dividing GDP at current market prices by the population. For this research, we considered GDP as one of the economic indicators of violent conflict.

Table 10: GDP per capita

Variable	Variableindicators
World Development Indicators	Gross Domestic Product (GDP)

Source: World Bank

Importing the GDP indicator data and visualizing it in a tabular form before preprocessing.

Figure 7: The top 5 instances of GDP indicator data

```
In [9]: df4.head(5)
```

Out[9]:

	Country	year	GDP per capita
0	Ethiopia	2007	1008.135315
1	Ethiopia	2008	1086.699463
2	Ethiopia	2009	1150.206055
3	Ethiopia	2010	1259.022583
4	Ethiopia	2011	1360.938477

B. Economic inequality indicator

This indicator considers inequality within the economy. In this research, economic inequality is considered an indicator that can fuel grievance as much as real inequality and can reinforce communal tensions or nationalistic rhetoric. The World Bank annually generates income inequality data on a yearly base for every country.

Table 11: Income inequality data

Variable	Variableindicators
Economic inequality	Income inequality

Source: World Bank

Importing the economic inequality indicator data and visualizing it in a tabular form before preprocessing.

Figure: 8- The top 5 instances of Economic inequality indicator data

```
In [11]: df5.head(5)
```

Out[11]:

	Country	year	Economic Inequality
0	Ethiopia	2007	8.6
1	Ethiopia	2008	8.6
2	Ethiopia	2009	8.8
3	Ethiopia	2010	8.5
4	Ethiopia	2011	8.2

C. Corruption indicator

This indicator is highly related to the economic and financial environment. It reduces the efficiency of government and business by enabling people to assume positions of power through patronage rather than ability and introduces an inherent instability in the political system. The most potential corruption that distorts the economic and financial sectors are financial corruption, nepotism, job reservation, "favour-for-favour", secret party funding, and suspiciously close ties between politics and business. Due to the above reasons, we consider corruption as one of the economic conflict indicators.

Table 12: Corruption perception index

Variable	variable indicators
Corruption	Corruption perception index

Source: Transparency International

Importing the Corruption indicator data and visualizing it in a tabular form before preprocessing.

Figure 9: The top 5 instances of corruption indicator data

```
In [17]: df7.head(5)
```

Out[17]:

	Country	year	Corruption perception index
0	Ethiopia	2007	24
1	Ethiopia	2008	26
2	Ethiopia	2009	27
3	Ethiopia	2010	27
4	Ethiopia	2011	27

D. Human Development

Understanding the human development level in the country would enable us to understand how much the economic welfare of a country is developing, it gives the idea regarding areas of development which require improvement. The United Nations Development Programme (UN-DP) measures and scores every country's human development and developed an index called the human development index.

Table -13- Human Development Index data developed by UNDP

Variable	Variable indicators
Human Development	Human Development Index

Importing the Human development indicator data and visualizing it in a tabular form before preprocessing.

Fig-10- Top 5 instances of Human Development indicator data

```
In [13]: df6.head(5)
```

Out[13]:

	Country	year	Human Development
0	Ethiopia	2007	0.380
1	Ethiopia	2008	0.395
2	Ethiopia	2009	0.402
3	Ethiopia	2010	0.412
4	Ethiopia	2011	0.422

- **Social Drivers of Conflict**

To understand the social drivers of conflict, three indicators are identified: Population growth, and Unemployment rate. For these indicators, we have collected data from different data repositories from 2007 to 2021.

Table 14: Summary of social drivers of conflict

Driver	Indicators	Temporal coverage	Data source
Social	Population growth	2007-2020	United Nations - Population Division
	Unemployment rate	2007-2021	World Bank

Those social indicators are very important to understand the social causes of conflict. Based on that, I will explore each indicator of social drivers of conflict in detail.

A. Population growth

This indicator measures the increase in the number of the country's population. If a country doesn't have a stable population growth that can go head-to-head with the population of the country the country will face a scarcity in resources and would fail in deep poverty. We consider this kind of imbalance between the population growth and resource in the country would be one indicator of social conflict.

Table -15: Population growth data developed by United Nations - Population Division

Variable	Variable indicators
Population growth	United Nations Population Record

Importing the Population growth indicator data and visualizing it in a tabular form before preprocessing.

Figure 11- Top 5 instances of population growth indicator data

```
In [20]: df8.head(5)
Out[20]:
```

	Country	year	Population
0	Ethiopia	2007	81996184
1	Ethiopia	2008	84357104
2	Ethiopia	2009	86755584
3	Ethiopia	2010	89237800
4	Ethiopia	2011	91817936

B. Unemployment

This indicator generally refers to the share of the labor force that is without work but available for and seeking employment. In Ethiopia, like many other countries in the developing world, the indicator unemployment rate is a national challenge. The challenge has caused frustration among the youth who have experienced joblessness and persistent dependency forcing them to resort to informal migration and violence.

Table 16: Unemployment data developed by World Bank

Variable	Variable indicators
Unemployment	Unemployment rate

Source: World Bank

Importing the unemployment indicator data and visualizing it in a tabular form before preprocessing.

Fig-12- Top 5 instances of unemployment indicator data

```
In [22]: df9.head(5)
Out[22]:
```

	Country	year	Unemployment rate
0	Ethiopia	2007	2.435
1	Ethiopia	2008	2.406
2	Ethiopia	2009	2.380
3	Ethiopia	2010	2.339
4	Ethiopia	2011	2.312

- **Environmental drivers of conflict**

To explore the Environmental drivers of conflict, two indicators have been identified: precipitation and agricultural land. For these indicators, we have collected data from World Bank data repositories from 2007 to 2021.

Table 17: Environmental drivers of conflict

Driver	Indicators	Temporal coverage	Data source
Environmental	Precipitation	2007-2021	World Bank
	Agricultural land	2007-2021	World Bank

These environmental indicators are very important to understand the environmental causes of conflict. In the sense that it could predicate how environmental conditions contribute to violent conflicts.

A. Precipitation

Nowadays scholars suggest that there are strong links between climate and violent conflicts. The effects of climate change, such as changes in temperature and precipitation, can increase the likelihood and intensity of conflict and violence. Accordingly, we consider precipitation as one of the environmental indicators of conflict.

Table 18: Precipitation data

Variable	Variable indicators
Precipitation	Precipitation from the World Bank

Figure 13: Top 5 instances of precipitation indicator data

```
In [24]: df10.head(5)
Out[24]:
```

	Country	year	precipitation
0	Ethiopia	2007	854.84
1	Ethiopia	2008	851.90
2	Ethiopia	2009	770.19
3	Ethiopia	2010	859.61
4	Ethiopia	2011	834.67

B. Agricultural land

Today, communities are experiencing a scarcity of farmland, fragmented land, drainage and land-grabbing practices both by the state and private actors. This disruption of agricultural land causes farmland conflicts, food insecurity, and deterioration of land and infrastructure. For this reason, we consider agricultural land as one of the environmental indicators of conflict. Importing the Agricultural land indicator data and visualizing it in a tabular form before preprocessing are presented below.

Table 19: Agricultural land data developed by World Bank

Variable	Variable indicators
Agricultural land	Agricultural land (sq. km)

Figure 14: Top 5 instances of agricultural land indicator data

```
In [26]: df11.head(5)
Out[26]:
```

	Country	year	Agricultural land (sq. km)
0	Ethiopia	2007	350770.0
1	Ethiopia	2008	345130.0
2	Ethiopia	2009	349850.0
3	Ethiopia	2010	356830.0
4	Ethiopia	2011	363252.0

❖ Merging the conflict indicators data

The above section shows how the data set is loaded. The next step is to merge the different indicators of conflict into one conflict indicators data set. By merging the individual conflict data set we will consume the time to explore each data set individually. To merge the conflict data set, the pandas data frame merge () was used. The merge () method updates the content of two data frames by merging them. By using merge () we can specify the columns we wanted to join, or it also checks if there are common columns in the data frames that we are joining and join them on those columns. In the data set, I have two common columns which are called "country" and "year". By using the merge () pandas data frame the datasets have been merged into one data frame.

Figure 15: Top 5 instances of the merged conflict indicator data

	Country	year	Voice and Accountability	Political Stability	Government Effectiveness	Regulatory Quality	Rule of Law	Control of Corruption	security apparatus	Factionalized Elites	...	Refugees and IDPs	External Intervention	Democracy index
0	Ethiopia	2007	-1.20	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	...	7.9	6.7	NaN
1	Ethiopia	2008	-1.31	-1.73	-0.44	-0.90	-0.71	-0.67	7.5	8.9	...	7.5	7.3	4.52
2	Ethiopia	2009	-1.30	-1.64	-0.56	-0.93	-0.83	-0.70	7.5	8.8	...	8.0	7.6	NaN
3	Ethiopia	2010	-1.34	-1.64	-0.48	-0.85	-0.81	-0.69	7.8	9.0	...	7.8	7.9	3.68
4	Ethiopia	2011	-1.36	-1.51	-0.51	-0.99	-0.74	-0.67	7.9	9.0	...	8.2	8.1	3.79

❖ **Exploratory Data Analysis on the conflict indicator dataset**

The main idea behind exploratory analysis is to examine the data before building a model. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using statistical graphs and other visualization techniques. Specifically in EDA, any errors and outliers have been detected. As well as it has been tried to understand different patterns in the data. This process allows for understanding the data better before making any assumptions. The outcomes of EDA will help to understand the data and take decisions accordingly. The first step in the EDA is to know the total dimension of our dataset. To check the total shape of the data frame pandas shape method is used. The shape method will give the total shape of the data frame using the number of columns with the number of rows. Knowing the dimension of the data set will help me to perform data manipulation on my dataset.

Figure 16: Checking the total dimension of the merged indicator data frame

```
In [73]: merged_indicators.shape
Out[73]: (15, 28)
```

As presented in the above fig the data set has 28 columns and 15 rows. The row in the dataset contains the value of each indicator within the time span of 2007-2021 and 28 different indicators in the columns. To visualize and manipulate all the indicators that are in the merged data pandas columns method is used.

Figure 17: Names of columns of merged conflict indicator data

```
In [74]: merged_indicators.columns
Out[74]: Index(['Country', 'year', 'Voice and Accountability', 'Political Stability',
                'Government Effectiveness', 'Regulatory Quality', 'Rule of Law',
                'Control of Corruption', 'security apparatus', 'Factionalized Elites',
                'Group Grievance', 'Economy', 'Economic Inequality',
                'Human Flight and Brain Drain', 'State Legitimacy', 'Public Services',
                'Human Rights', 'Demographic Pressures', 'Refugees and IDPs',
                'External Intervention', 'Democracy index', 'GDP per capita',
                'Human Development', 'Corruption perception index', 'Population',
                'Unemployment rate', 'precipitation ', 'Agricultural land (sq. km)'],
               dtype='object')
```

As presented in the above figure, there are 27 indicators of conflict in the data set excluding the country column. By knowing every column name, it is possible to operate each column by its column name. But getting only the name of the column is not enough to understand the dataset; the info method can also be used to get general information about the data. In pandas, there is a method which is called info () that will help to get general information about the data frame. This includes the number of columns, column labels, column data types, memory usage, and the number of cells in each column (non-null values). In the below figure, byusing the info () method the general info about the data frame is visualized. As it can be observed in the figure on the "non-null count" column there are different values for the "democracy index", "GDP per capita" and "agricultural land" indicators which means there are null values on those columns.

Figure-18: General information about the merged indicator data frame

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15 entries, 0 to 14
Data columns (total 28 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Country                         15 non-null     object
 1   year                            15 non-null     int64
 2   Voice and Accountability        15 non-null     float64
 3   Political Stability             15 non-null     float64
 4   Government Effectiveness        15 non-null     float64
 5   Regulatory Quality              15 non-null     float64
 6   Rule of Law                     15 non-null     float64
 7   Control of Corruption           15 non-null     float64
 8   security apparatus              15 non-null     float64
 9   Factionalized Elites            15 non-null     float64
 10  Group Grievance                 15 non-null     float64
 11  Economy                         15 non-null     float64
 12  Economic Inequality             15 non-null     float64
 13  Human Flight and Brain Drain    15 non-null     float64
 14  State Legitimacy                15 non-null     float64
 15  Public Services                 15 non-null     float64
 16  Human Rights                    15 non-null     float64
 17  Demographic Pressures           15 non-null     float64
 18  Refugees and IDPs               15 non-null     float64
 19  External Intervention           15 non-null     float64
 20  Democracy index                 13 non-null     float64
 21  GDP per capita                  14 non-null     float64
 22  Human Development               15 non-null     float64
 23  Corruption perception index     15 non-null     int64
 24  Population                      15 non-null     int64
 25  Unemployment rate               15 non-null     float64
 26  precipitation                   15 non-null     float64
 27  Agricultural land (sq. km)      14 non-null     float64
dtypes: float64(24), int64(3), object(1)
memory usage: 3.4+ KB
```

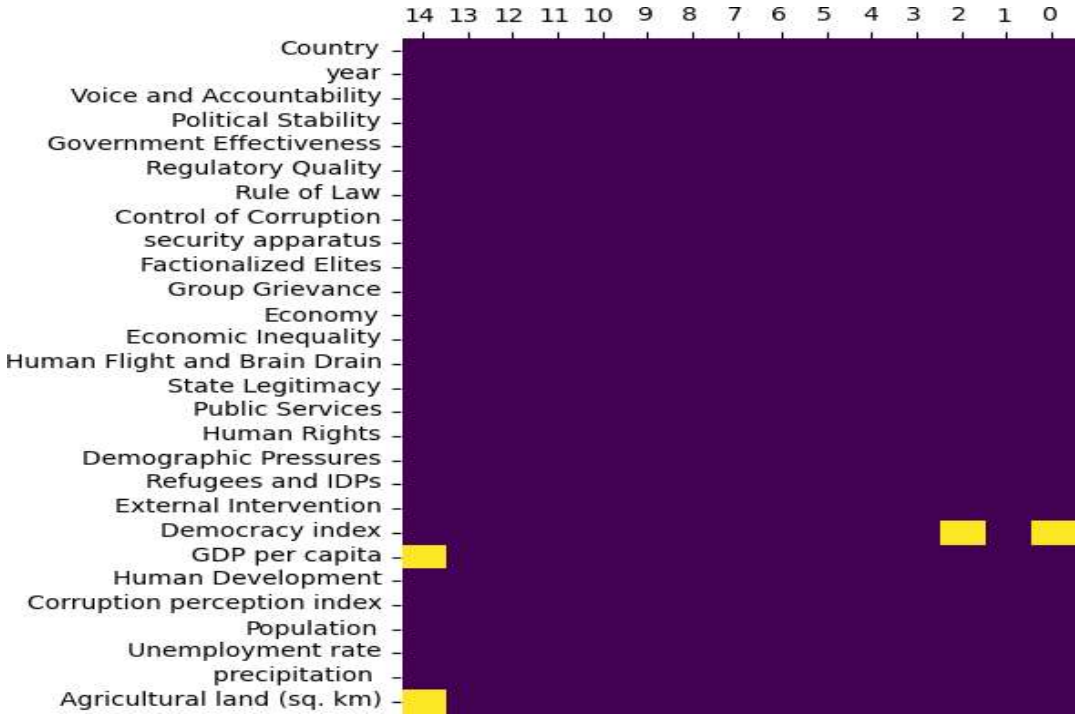
To get the exact amount of the null values on those columns pandas method can be used which is called isnull (). By using the isnull().sum() method the total amount of null values in each column can be determined.

Figure 19: Null values in the merged indicator data frame

```
In [70]: merged_indicators.isnull().sum()
Out[70]: Country                                0
         year                                   0
         Voice and Accountability               0
         Political Stability                    0
         Government Effectiveness               0
         Regulatory Quality                     0
         Rule of Law                            0
         Control of Corruption                  0
         security apparatus                     0
         Factionalized Elites                   0
         Group Grievance                        0
         Economy                                0
         Economic Inequality                    0
         Human Flight and Brain Drain           0
         State Legitimacy                       0
         Public Services                        0
         Human Rights                           0
         Demographic Pressures                  0
         Refugees and IDPs                      0
         External Intervention                  0
         Democracy index                        2
         GDP per capita                         1
         Human Development                      0
         Corruption perception index            0
         Population                             0
         Unemployment rate                      0
         precipitation                          0
         Agricultural land (sq. km)             1
         dtype: int64
```

Figure 18 shows that there are 2 null values in the "democracy index" column, one null value in the "GDP per capital" column and one null value in the "agricultural land" column. the null values can also be visualized using the sns.heatmap () method. The heatmap will visualize the data in a coloured matrix.

Figure 20: Missing value in the data frame using the heatmap



The other most important step in EDA is to detect outliers in the data set. In pandas, there is also a method called describe () that will return a description of the data in the data frame. The method helps to generate descriptive statistics that summarize the count, mean, standard deviation, minimum, maximum and percentiles to include in the output excluding NaN values. Most importantly mean is sensitive to outliers. For example, if the mean is so small when compared to the max value the max value is an outlier in our data.

Figure-21; Describing the first 13 columns in the data frame set

```
merged_indicators.describe()[['Voice and Accountability', 'Political Stability','Government Effectiveness', 'Regulatory Quality',
                              'Rule of Law','Control of Corruption', 'security apparatus', 'Factionalized Elites','Group Grievanc
                              'Economy', 'Economic Inequality','Human Flight and Brain Drain', 'State Legitimacy']]
```

	Voice and Accountability	Political Stability	Government Effectiveness	Regulatory Quality	Rule of Law	Control of Corruption	security apparatus	Factionalized Elites	Group Grievance	Economy	Economic Inequality	Human Flight and Brain Drain	St Legitima
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000
mean	-1.258667	-1.590667	-0.560000	-0.989333	-0.593333	-0.533333	8.060000	8.713333	8.560000	7.153333	7.466667	7.120000	7.840
std	0.123107	0.211473	0.090159	0.081981	0.143958	0.119323	0.373784	0.322638	0.468737	0.699864	0.902905	0.452296	0.503
min	-1.430000	-2.070000	-0.710000	-1.140000	-0.830000	-0.700000	7.500000	7.900000	7.800000	5.900000	6.200000	6.300000	7.100
25%	-1.325000	-1.705000	-0.620000	-1.045000	-0.695000	-0.650000	7.850000	8.650000	8.300000	6.700000	6.700000	6.700000	7.450
50%	-1.300000	-1.620000	-0.560000	-0.980000	-0.610000	-0.500000	8.100000	8.700000	8.600000	7.400000	7.300000	7.200000	7.900
75%	-1.185000	-1.455000	-0.475000	-0.930000	-0.470000	-0.435000	8.400000	8.900000	8.850000	7.700000	8.350000	7.500000	8.100
max	-1.040000	-1.280000	-0.440000	-0.850000	-0.390000	-0.360000	8.700000	9.200000	9.500000	8.300000	8.800000	7.700000	8.800

Figure-22- describing the last 13 columns in the data set

```
merged_indicators.describe()[['Public Services','Human Rights', 'Demographic Pressures', 'Refugees and IDPs','External Interventi
                              'Democracy index', 'GDP per capita','Human Development', 'Corruption perception index', 'Populatio
                              'Unemployment rate', 'precipitation ', 'Agricultural land (sq. km)']]
```

	Public Services	Human Rights	Demographic Pressures	Refugees and IDPs	External Intervention	Democracy index	GDP per capita	Human Development	Corruption perception index	Population	Unemployment rate	precipitation
count	15.000000	15.00000	15.000000	15.000000	15.000000	13.000000	14.000000	15.000000	15.000000	1.500000e+01	15.000000	15.000000
mean	8.326667	8.48000	9.306667	8.640000	7.940000	3.660000	1630.105953	0.448333	32.000000	1.002756e+08	2.473800	887.727333
std	0.510555	0.27826	0.268506	0.621978	0.492515	0.320884	432.477065	0.040293	4.675162	1.223799e+07	0.415036	55.822021
min	7.000000	7.90000	8.900000	7.500000	6.700000	3.300000	1008.135315	0.380000	24.000000	8.199618e+07	2.250000	770.190000
25%	8.250000	8.45000	9.100000	8.100000	7.750000	3.420000	1284.501556	0.417000	27.000000	9.052787e+07	2.287500	853.370000
50%	8.400000	8.50000	9.300000	8.700000	8.100000	3.680000	1600.906250	0.450000	33.000000	9.974677e+07	2.318000	905.520000
75%	8.650000	8.65000	9.450000	9.050000	8.200000	3.790000	1989.896484	0.484500	34.500000	1.096637e+08	2.393000	920.910000
max	8.900000	9.00000	9.800000	9.500000	8.700000	4.520000	2296.827393	0.498000	39.000000	1.202830e+08	3.694000	1001.380000

As shown in Figure 22, the describe () method gives statistical observation of our data. The example given above shows that comparing the max value with the mean is one method of detecting an outlier in the data set. As observed, the mean values in each column are not too small when compared with the max value. For this reason, by using the described method we can't identify any outliers in the data set.

4.1.2. Preprocessing the merged conflict indicator dataset

The preprocessing step also needs to deal with missing values and remove unnecessary columns.

A. Dealing with missing values

In the merged conflict indicator dataset, it is observed that there are missing values in the "democracy index", "GDP per capita" and "agricultural land" columns. If the missing values are not dealt with properly, it could present a risk of building a biased model which might result in incorrect outcomes and affect precision. There are two primary ways of dealing with missing values one is to delete the Missing values and the second is to impute the Missing Values. For this research, imputing the missing values using their mean is selected rather than deleting the values. This is because the conflict indicators data set is small in size and deleting the values may bias the model since these indicators are used as a common denominator for the model building at the end.

Figure-23 Filling the "democracy index" column in the data set

```
mean_value=merged_indicators['Democracy index'].mean()
```

```
merged_indicators['Democracy index'].fillna(value=mean_value, inplace=True)
```

Figure-24: Filling the "GDP per capita" column in the data set

```
mean_value=merged_indicators['GDP per capita'].mean()
```

```
merged_indicators['GDP per capita'].fillna(value=mean_value, inplace=True)
```

Figure-25: Filling the "Agricultural land" column in the data set

```
mean_value=merged_indicators['Agricultural land (sq. km)'].mean()
```

```
merged_indicators['Agricultural land (sq. km)'].fillna(value=mean_value, inplace=True)
```

Table 20 shows that missing values are checked using isnull().sum() and isnull().any() method. In any of the methods, no missing values have been identified in the dataset. In the first method, the summation of every column missing value is 0, which means there is no missing value. In the second method by using any () all the columns have a Boolean value of false, which means the method couldn't identify any missing value.

Table -20- checking missing values in the data

Checking for missing values

```
merged_indicators.isnull().sum()
```

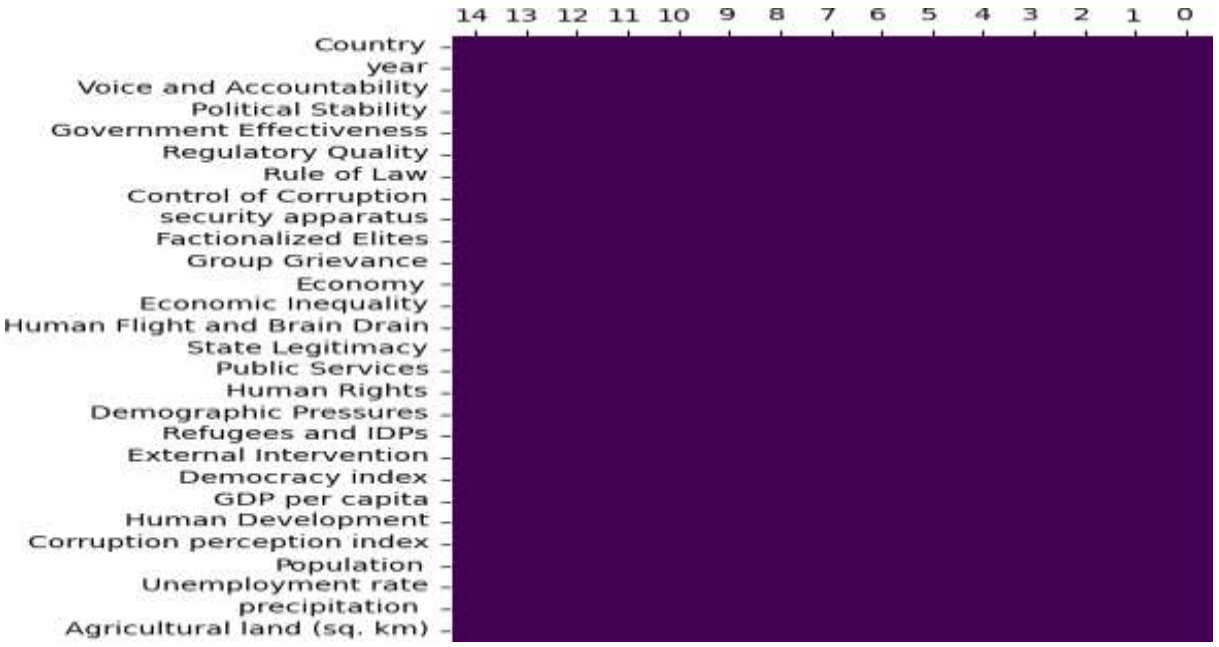
```
Country                          0
year                             0
Voice and Accountability         0
Political Stability              0
Government Effectiveness         0
Regulatory Quality               0
Rule of Law                      0
Control of Corruption            0
security apparatus               0
Factionalized Elites             0
Group Grievance                  0
Economy                          0
Economic Inequality              0
Human Flight and Brain Drain     0
State Legitimacy                 0
Public Services                  0
Human Rights                     0
Demographic Pressures            0
Refugees and IDPs                0
External Intervention            0
Democracy index                  0
GDP per capita                   0
Human Development                0
Corruption perception index      0
Population                       0
Unemployment rate                0
precipitation                    0
Agricultural land (sq. km)       0
dtype: int64
```

```
merged_indicators.isnull().any()
```

```
Country                          False
year                             False
Voice and Accountability         False
Political Stability              False
Government Effectiveness         False
Regulatory Quality               False
Rule of Law                      False
Control of Corruption            False
security apparatus               False
Factionalized Elites             False
Group Grievance                  False
Economy                          False
Economic Inequality              False
Human Flight and Brain Drain     False
State Legitimacy                 False
Public Services                  False
Human Rights                     False
Demographic Pressures            False
Refugees and IDPs                False
External Intervention            False
Democracy index                  False
GDP per capita                   False
Human Development                False
Corruption perception index      False
Population                       False
Unemployment rate                False
precipitation                    False
Agricultural land (sq. km)       False
dtype: bool
```

Figure 26 shows that there are no missing values in our dataset. In the next section, the unnecessary columns that are not important for prediction purposes are removed.

Figure-26: The dataset using the heatmap



B. Unnecessary columns

As described above, the data has one unnecessary column named “country”. It is because the data is collected for only Ethiopia as a country. It has the value “Ethiopia”. Based on that, the column country is removed to save storage space.

Figure-27: Removing the “country” column from the dataset

```
merged_indicators=merged_indicators.drop(['Country'], axis=1)
```

```
merged_indicators.columns
```

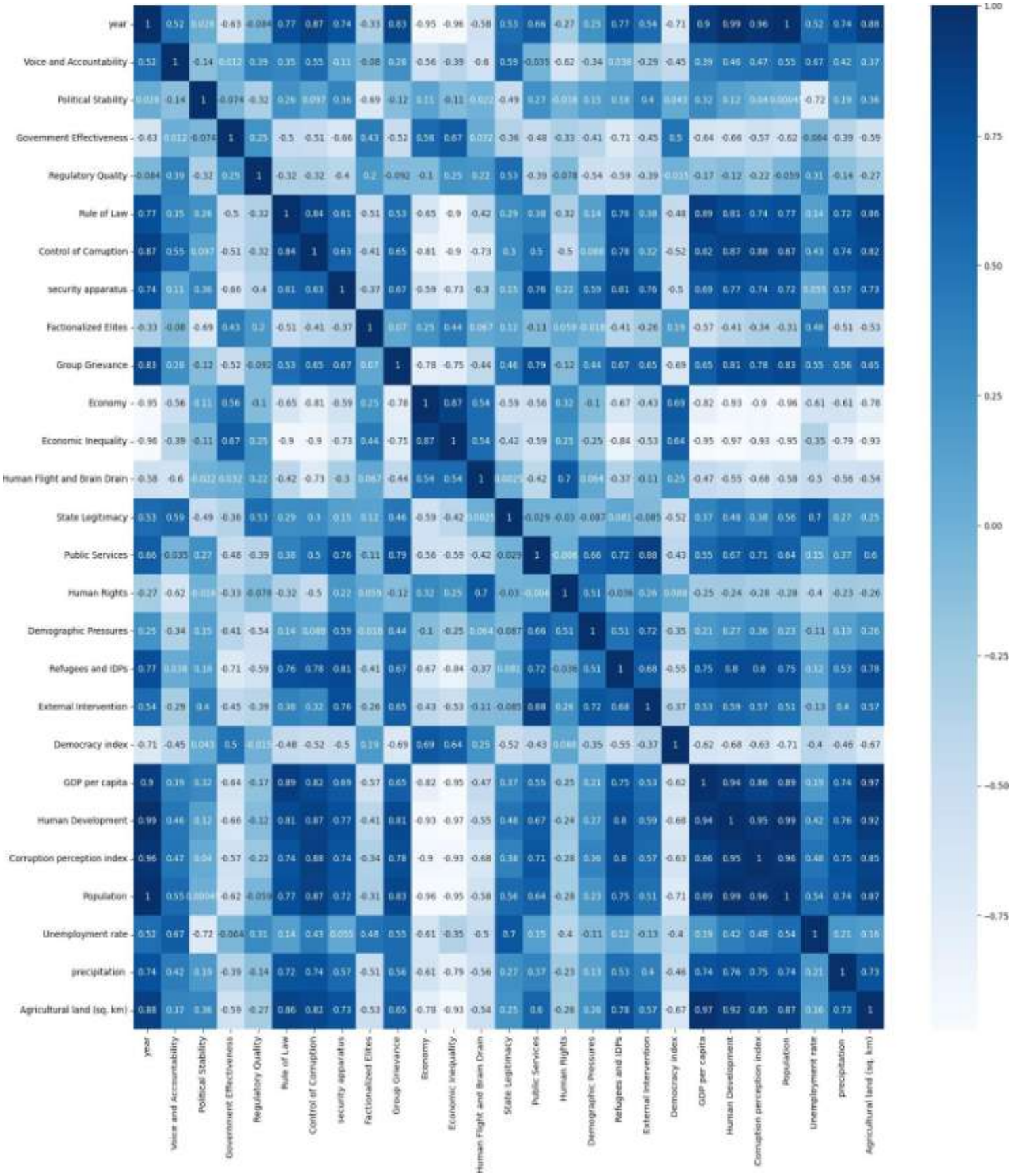
```
Index(['year', 'Voice and Accountability', 'Political Stability',
       'Government Effectiveness', 'Regulatory Quality', 'Rule of Law',
       'Control of Corruption', 'security apparatus', 'Factionalized Elites',
       'Group Grievance', 'Economy', 'Economic Inequality',
       'Human Flight and Brain Drain', 'State Legitimacy', 'Public Services',
       'Human Rights', 'Demographic Pressures', 'Refugees and IDPs',
       'External Intervention', 'Democracy index', 'GDP per capita',
       'Human Development', 'Corruption perception index', 'Population',
       'Unemployment rate', 'precipitation ', 'Agricultural land (sq. km)'],
      dtype='object')
```

❖ **Building a correlation matrix between conflictindicators**

The correlation matrix is a statistical technique used to evaluate the relationship between all possible pairs of values in our data set in a matrix format. By using the correlation matrix we can summarize a large data set and identify patterns by looking at which variable is more correlated to which variable. The matrix involves a rows and columns table of variables. Every cell contains a correlation coefficient within the value of -1 to 1. This helps to determine the relationship between variables: 1 is a strong relationship; 0 is a neutral relationship and -1 is a weak (not strong) relationship.

The independent variables help to predict the dependent variable. Before combining our independent variables, there is a need to build a correlation matrix to determine the correlation coefficients between the independent variables. By building the correlation matrix I will identify patterns and get a better understanding of what is the most important for my model.

Figure 28: Checking for the correlation coefficient in the merged indicator data frame


year
Voice and Accountability
Political Stability
Government Effectiveness
Regulatory Quality
Rule of Law
Control of Corruption
security apparatus
Factionalized Elites
Group Grievance
Economy
Economic Inequality
Human Flight and Brain Drain
State Legitimacy
Public Services
Human Rights
Demographic Pressures
Refugees and IDPs
External Intervention
Democracy index
GDP per capita
Human Development
Corruption perception index
Population
Unemployment rate
precipitation
Agricultural land (sq. km)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

As observed in the correlation matrix, most of the conflict indicators have a moderate positive relationship and a fairly strong positive relationship having a correlation coefficient between 0.5 and 0.9. The positive coefficients indicate that when the value of one variable increases, the value of the other variable also tends to increase producing an upward slope on a scatterplot. The matrix shows, most of the results lie between 0 and +1, which shows there is a relationship. But the points don't all fall on a line. As the correlation-coefficient approaches 1, the strength of the relationship increases and the data points tend to fall closer to a line. Based on the figure above, understanding that relationship is useful because the value of one variable can be used to predict the value of the other variable. For example, population and unemployment rates are correlated. If we observe an increase in population in a country, we can predict that the unemployment rate will also increase. But this does not imply causation. As described above the correlation between two variables only indicates that changes in one variable are associated with changes in the other variable. But, correlation does not mean that the changes in one variable cause the changes in the other variable. Sometimes it is clear that there is a causal relationship. For the population and unemployment column, it might make sense that population growth in a country might cause an increased unemployment rate if there is no economic growth in the country. However, in other cases, a causal relationship is not possible. So the correlation coefficient results explain the relationship between variables, not causality. But considering some independent variables are highly correlated with each other, further experimentation is done in the experimental analysis to identify the most important variables and remove the less irrelevant.

4.2. Exploring and preprocessing dependent variables

This research applies exploratory data analysis (EDA) and preprocessing of the conflict events dataset and presents their results. The conflict event datasets are loaded, explored and preprocessed. The dataset consists of conflict events that happen in Ethiopia, but it does not have the causes of the conflict events. The conflict event dataset consists of all conflict types that occurred in the past years (from 2007 to 2021) and the actors involved but without specific drivers of conflict indicated.

4.2.1. Importing and Exploring Conflict Event Dataset

For analysis purposes, the Jupyter Notebook is used. The imported Python libraries in the above sections are also used here to load the dataset and visualize the dataset.

Table 21: Summary of the conflict events dataset

Data	Columns	Description
ACLED Political Violence Events Data	ISO	A numeric code for each individual country
	EVENT_ID_CNTY	An individual identifier by number and country acronym
	EVENT_ID_NO_CNTY	An individual numeric identifier
	EVENT_DATE	The day, month and year on which an event took place
	YEAR	The year in which an event took place
	TIME_PRECISION	A numeric code indicating the level of certainty of the date code for the event
	EVENT_TYPE	The type of event
	SUB_EVENT_TYPE	The type of sub-event

ACTOR1	The named actor involved in the event

ASSOC_ACTOR_1	The named actor associated with or identifying ACTOR1
INTER1	A numeric code indicating the type of ACTOR1
ACTOR2	The named actor involved in the event
ASSOC_ACTOR_2	The named actor associated with or identifying ACTOR2
INTER2	A numeric code indicating the type of ACTOR2
INTERACTION	A numeric code indicating the interaction between types of ACTOR1 and ACTOR2
REGION	The region of the world where the event took place
COUNTRY	The country in which the event took place
ADMIN1	The largest sub-national administrative region in which the event took place
ADMIN2	The second largest sub-national administrative region in which the event took place
ADMIN3	The third largest sub-national administrative region in which the event took place
LOCATION	The location in which the event took place
LATITUDE	The latitude of the location
LONGITUDE	The longitude of the location
GEO_PRECISION	A numeric code indicating the level of certainty of the location code for the event
SOURCE	The source of the event report
SOURCE_SCALE	The scale (local, regional, national, international) of the Source
NOTES	A short description of the event
FATALITIES	The number of reported fatalities which occurred during the event

Source: Armed Conflict Location & Event Data Project (ACLED)

Importing the conflict events data and presenting it in a tabular form before preprocessing is performed as follows. All data collected from the ACLED data repository were imported to Anaconda data analysis software. To illustrate, the table below shows the topic 5 conflict events data.

Figure 29: The top 5 conflict events data

	ISO	EVENT_ID_CNTY	EVENT_ID_NO_CNTY	EVENT_DATE	YEAR	TIME_PRECISION	EVENT_TYPE	SUB_EVENT_TYPE	ACTOR1	ASSOC_ACTOR_1
0	231	ETH1053	1053	11-February-2007	2007	1	Battles	Armed clash	Gabra Ethnic Militia (Ethiopia)	NaN
1	231	ETH1054	1054	22-April-2007	2007	1	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	Muslim Militia (Somalia)
2	231	ETH1055	1055	25-April-2007	2007	1	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	NaN
3	231	ETH1056	1056	04-June-2007	2007	1	Battles	Armed clash	Military Forces of Ethiopia (1991-2018)	NaN
4	231	ETH1057	1057	07-June-2007	2007	1	Battles	Armed clash	Borana Ethnic Militia (Ethiopia)	Gabra Ethnic Militia (Ethiopia)

❖ Exploratory Data Analysis on the Conflict Events Dataset

This section explores the conflict events data imported to the analysis tool. To observe the total dimension of the dataset, the shape () pandas method was used. This technique is very crucial to apply further EDA and preprocessing. As shown below in Figure 30, the data set has 8261 rows and 29 columns.

Figure 30: Total dimension of the conflict event data frame

```
df.shape
```

```
(8261, 29)
```

The 8261 rows indicate the conflict events that happened in each year and the 29 columns indicate the different variables that are described in Table 1. Following that, all the column names loaded into the pandas data frame were visualized using the columns () pandas method, as seen in the following figure, for further manipulation of the data set and ensuring that all column values were loaded correctly.

Figure 31: Name of columns of conflict events data

```
df.columns
```

```
Index(['ISO', 'EVENT_ID_CNTY', 'EVENT_ID_NO_CNTY', 'EVENT_DATE', 'YEAR',
       'TIME_PRECISION', 'EVENT_TYPE', 'SUB_EVENT_TYPE', 'ACTOR1',
       'ASSOC_ACTOR_1', 'INTER1', 'ACTOR2', 'ASSOC_ACTOR_2', 'INTER2',
       'INTERACTION', 'REGION', 'COUNTRY', 'ADMIN1', 'ADMIN2', 'ADMIN3',
       'LOCATION', 'LATITUDE', 'LONGITUDE', 'GEO_PRECISION', 'SOURCE',
       'SOURCE_SCALE', 'NOTES', 'FATALITIES', 'TIMESTAMP'],
      dtype='object')
```

This process has offered the opportunity to manipulate all columns by their column names. A focus was also given to getting more information about the dataset using the info () method which informs about the number of columns, column labels, column data types, memory usage, and the number of cells in each column (non-null values).

Figure 32: General information about the conflict events data frame

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8261 entries, 0 to 8260
Data columns (total 29 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ISO                 8261 non-null   int64
 1   EVENT_ID_CNTY       8261 non-null   object
 2   EVENT_ID_NO_CNTY    8261 non-null   int64
 3   EVENT_DATE          8261 non-null   object
 4   YEAR                8261 non-null   int64
 5   TIME_PRECISION      8261 non-null   int64
 6   EVENT_TYPE          8261 non-null   object
 7   SUB_EVENT_TYPE      8261 non-null   object
 8   ACTOR1              8261 non-null   object
 9   ASSOC_ACTOR_1       2847 non-null   object
 10  INTER1              8261 non-null   int64
 11  ACTOR2              6525 non-null   object
 12  ASSOC_ACTOR_2       2626 non-null   object
 13  INTER2              8261 non-null   int64
 14  INTERACTION         8261 non-null   int64
 15  REGION              8261 non-null   object
 16  COUNTRY             8261 non-null   object
 17  ADMIN1              8261 non-null   object
 18  ADMIN2              8261 non-null   object
 19  ADMIN3              8261 non-null   object
 20  LOCATION            8261 non-null   object
 21  LATITUDE            8261 non-null   float64
 22  LONGITUDE           8261 non-null   float64
 23  GEO_PRECISION       8261 non-null   int64
 24  SOURCE              8261 non-null   object
 25  SOURCE_SCALE        8261 non-null   object
 26  NOTES               8261 non-null   object
 27  FATALITIES          8261 non-null   int64
 28  TIMESTAMP           8261 non-null   int64
dtypes: float64(2), int64(10), object(17)
memory usage: 1.8+ MB
```

As shown in Figure, by using the info() method on the non-null count different values were observed on the “Assoc_Actor_1”, “Actor2” and “Assoc_Actor_2” which demonstrates a missing value on these columns. In addition, on the data type column 2 float values, 10 int values and 17 object values were observed. This suggests that the data has to be preprocessed because machine learning models do not accept object and float type data rather than numeric. The missing values in each column were visualised using isnull().sum() pandas method.

Figure 33 shows that there are 5,414 missing values in the “Assoc_Actor_1” column, 1,736 missing values in the “Actor 2” column and 5,635 missing values in the “Assoc_Actor_2” column. It shows many missing values in these columns and suggests the need for much work to be done on the preprocessing step to deal with them because these missing values might bias the model at the end.

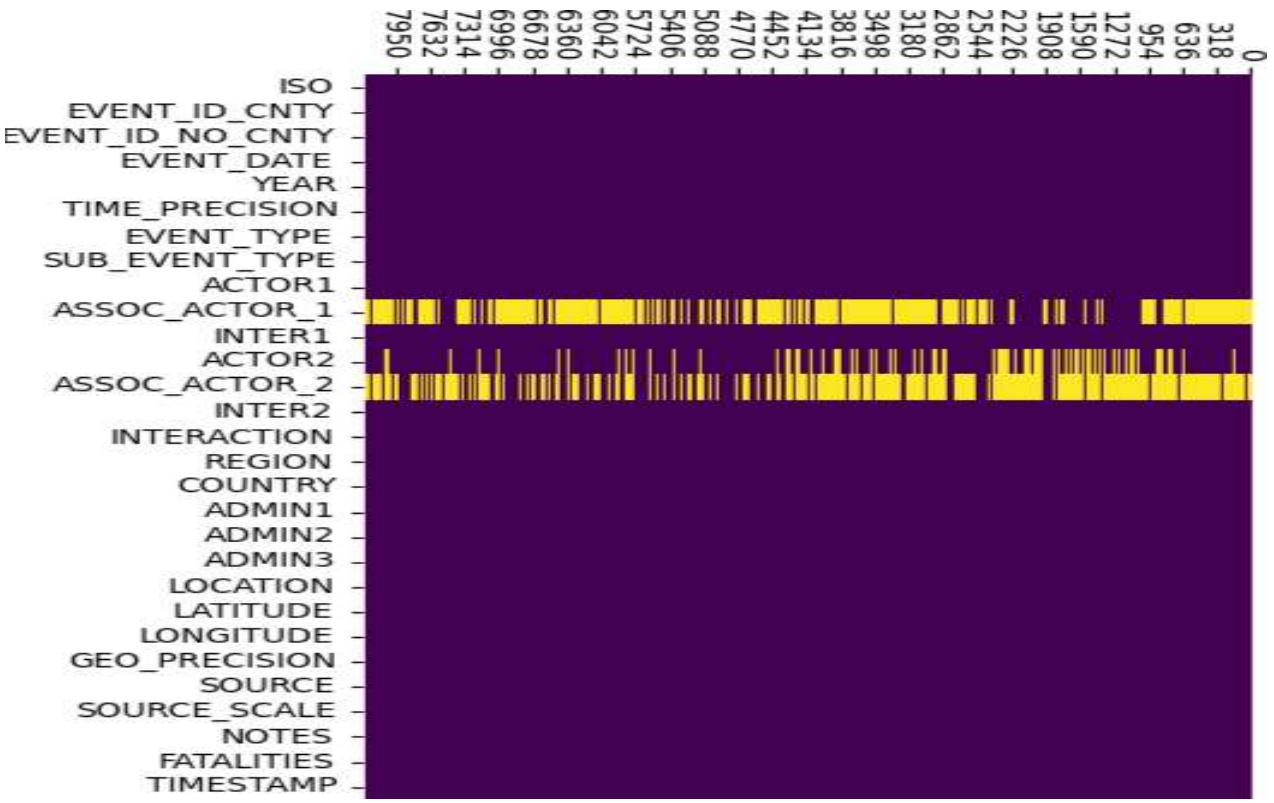
Figure 33: Null values in the conflict event data frame

```
df.isnull().sum()
```

ISO	0
EVENT_ID_CNTY	0
EVENT_ID_NO_CNTY	0
EVENT_DATE	0
YEAR	0
TIME_PRECISION	0
EVENT_TYPE	0
SUB_EVENT_TYPE	0
ACTOR1	0
ASSOC_ACTOR_1	5414
INTER1	0
ACTOR2	1736
ASSOC_ACTOR_2	5635
INTER2	0
INTERACTION	0
REGION	0
COUNTRY	0
ADMIN1	0
ADMIN2	0
ADMIN3	0
LOCATION	0
LATITUDE	0
LONGITUDE	0
GEO_PRECISION	0
SOURCE	0
SOURCE_SCALE	0
NOTES	0
FATALITIES	0
TIMESTAMP	0

To visualize the missing values, the sns.heatmap () method was used. The heatmap visualizes the data in a coloured matrix as shown below.

Figure 34: Visualizing missing value in the conflict event data frame using the heatmap



Certain issues need to be discussed in analysing violent political conflicts such as conflict event type, actors involved and the consequences of a given conflict event. The conflict data set used in this research has event type, the actors and fatalities. The latter helps to observe the consequences of a given conflict focusing on the number of people dead in a conflict incident. The limitation that needs to be uncovered is the difficulty in gaining adequate data on the consequences of a given conflict including the exact number of people who died. Some of the data shows that the "Fatalities" column has many "0" values. This however does not mean that conflicts occurred without impact of any kind such as displacement, infrastructure destruction or livelihoods disruption. An incident of violent political conflict is considered an incident consisting of at least two actors' engagement and/or the confrontation or disagreement resulting in at least more than one number of fatalities. In some data, the "Actor 2" column is missing. For this reason, the number of fatalities is seriously considered. The preprocessing step used these conditions to assess every conflict event data.

4.2.2. Preprocessing the conflict events dataset

EDA demonstrates certain problems that need to be preprocessed to be compatible with model building. This includes missing values and unnecessary columns. The following discussion illustrates that efforts were made to deal with them one by one.

A. Dealing with missing values

EDA shows that there are missing values in the "Assoc_Actor_1", "Actor 2" and "Assoc_Actor_2" columns. The two columns which are Assoc_Actor_1 and Assoc_Actor_2 have more than 50% missing values on their data. These two columns are not going to contribute to the learning of the model, rather, they will bias the model in a particular direction. For this reason, efforts were carried out to remove them. The actor 2 column is a very important variable in the dataset to determine the conflict events type. In cases where data values don't have a Second Actor, the number of fatalities was carefully examined to include or remove it. Below the total dimension of missing values in the "Actor 2" column is presented.

Figure 35: Checking the total dimension of missing values on the "Actor 2" column

```
miss_values.shape
```

```
(1734, 5)
```

Figure 36: The top 10 data only with missing value in the "Actor 2" column with "sub-event type", "Actor 2", "Fatalities" and "Notes" columns.

```
df2 = df[["SUB_EVENT_TYPE", "ACTOR1","ACTOR2", "FATALITIES", "NOTES"]]
```

```
miss_values = df2[df2['ACTOR2'].isnull()]
```

```
miss_values.head(10)
```

	SUB_EVENT_TYPE	ACTOR1	ACTOR2	FATALITIES	NOTES
39	Peaceful protest	Protesters (Ethiopia)	NaN	0	Participants in Great Ethiopian Run use the ev...
91	Peaceful protest	Protesters (Ethiopia)	NaN	0	Thousands protest hit and run charges against ...
172	Peaceful protest	Protesters (Ethiopia)	NaN	0	Opposition stages protest in Addis Ababa
177	Headquarters or base established	Ethiopian Unity and Justice Movement	NaN	0	Newly formed Ethiopian Unity and Justice Movem...
180	Peaceful protest	Protesters (Ethiopia)	NaN	0	Oromo students stage protest against Ethiopian...
223	Change to group/activity	ONLF: Ogaden National Liberation Front	NaN	0	Movement of forces: Ogaden rebels warn oil fir...
260	Peaceful protest	Protesters (Ethiopia)	NaN	0	Students protest river pollution
271	Peaceful protest	Protesters (Ethiopia)	NaN	0	Citizens gather to protest confiscation of lan...
327	Change to group/activity	ADFM: Amhara Democratic Force Movement	NaN	0	Change to armed group: New rebel group forms i...
353	Change to group/activity	Unidentified Armed Group (Eritrea)	NaN	0	Change to armed group: 8 Eritrean political or...

Figure 37: The final 10 data only with missing value in the "Actor 2" column with "sub-event type", "Actor 2", "Fatalities" and "Notes" columns

```
miss_values.tail(10)
```

	SUB_EVENT_TYPE	ACTOR1	ACTOR2	FATALITIES	NOTES
8078	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 23 October 2022, a protest against western ...
8079	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 23 October 2022, a protest against western ...
8080	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 23 October 2022, a protest against western ...
8081	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 23 October 2022, a protest against western ...
8082	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 23 October 2022, a protest against western ...
8113	Change to group/activity	Government of Ethiopia (2018-)	NaN	0	Security measures: On 31 October 2022, the adm...
8119	Peaceful protest	Protesters (Ethiopia)	NaN	0	On 1 November 2022, ethnic Tigray residents of...
8233	Change to group/activity	Government of Ethiopia (2018-)	NaN	0	Security measures: On 24 November 2022, SNNP r...
8255	Other	Civilians (Ethiopia)	NaN	0	Displacement: Around 30 November 2022, several...
8259	Other	TPLF: Tigray People's Liberation Front	NaN	0	Other: On 2 December 2022, TPLF Commanders ann...

As figures 36 and 37 indicate, the values in the conflict event data that has a missing value in the "Actor 2" column have 0 fatalities. Most of their sub-event type value is peaceful protests, security measures that have been made by the government, formation of new military groups and some kind of strategic development. Conflict event types without a second actor and having 0 fatalities were not considered. For example, the following table shows a second actor missing value and their sub-event type is "peaceful protest". To observe this, first creating a new data frame which has only peace full protest in their sub-event type from the missing values data frame was necessary and then followed by checking the total dimension of the data frame.

Figure 38: A new data frame on peaceful_ protest with missing values on "Actor 2"

```
peaceful_protest = miss_values[miss_values['SUB_EVENT_TYPE']== "Peaceful protest"]
```

```
peaceful_protest.head(5)
```

	SUB_EVENT_TYPE	ACTOR1	ACTOR2	FATALITIES	NOTES
39	Peaceful protest	Protesters (Ethiopia)	NaN	0	Participants in Great Ethiopian Run use the ev...
91	Peaceful protest	Protesters (Ethiopia)	NaN	0	Thousands protest hit and run charges against ...
172	Peaceful protest	Protesters (Ethiopia)	NaN	0	Opposition stages protest in Addis Ababa
180	Peaceful protest	Protesters (Ethiopia)	NaN	0	Oromo students stage protest against Ethiopian...
260	Peaceful protest	Protesters (Ethiopia)	NaN	0	Students protest river pollution

Figure 39- checking the total dimension of the peaceful_protest data frame

```
peaceful_protest.shape
```

```
(1500, 5)
```

The figures show that the "Actor 2" column has 1,734 missing values. As the variable determines the types of conflict, there is a need for further discussion on the missing values. By applying further exploratory analysis on the "Actor 2" column missing values, focusing on their sub-event conflict type values, it shows that there are 1,500 peace full protests and 234 values related to the formation of new military groups and strategical developments. For the fact that they are nonviolent in effect, and have zero fatalities, they cannot be considered as violent political conflicts. Data values that have null values on their second actor are considered irrelevant for our analysis and were removed from the data set.

Figure 40: Removing the "Actor 2" column missing values

```
df=df.dropna(subset=["ACTOR2"])
```

Figure 41: Checking for missing values on the "Actor 2" column

```
df['ACTOR2'].isnull().sum()
```

```
0
```

In this way, all values that are null on their "Actor 2" column were removed. After removing the null values on the specified column I checked for missing values using the isnull().sum()

pandas method. It resulted in a zero number of missing values on the "Actor 2" column. By correctly removing the values, we have cleaned the data from incidents that are not considered violent political conflicts. The other columns with missing values are "ASSOC_ACTOR_1" and "ASSOC_ACTOR_2". These columns are values that are related to the main actors of the incidents which are "Actor1" and "Actor2". In these two columns, we have 5,414 and 5,635 missing values consecutively. Having such a huge amount of missing value (more than 75%) on a given column would make the model unable to learn anything from the values. For this reason, they were removed.

Figure 42: Removing the "ASSOC_ACTOR_1" column

```
df=df.drop(['ASSOC_ACTOR_1'], axis=1)
```

Fig-43- Removing "ASSOC_ACTOR_2" column

```
df=df.drop(['ASSOC_ACTOR_2'], axis=1)
```

The process continued to observe the number of missing values on the dataset and to check if other missing values are not being removed.

Table 22: Checking missing values in the data

Checking for missing values

```
df.isnull().sum()
```

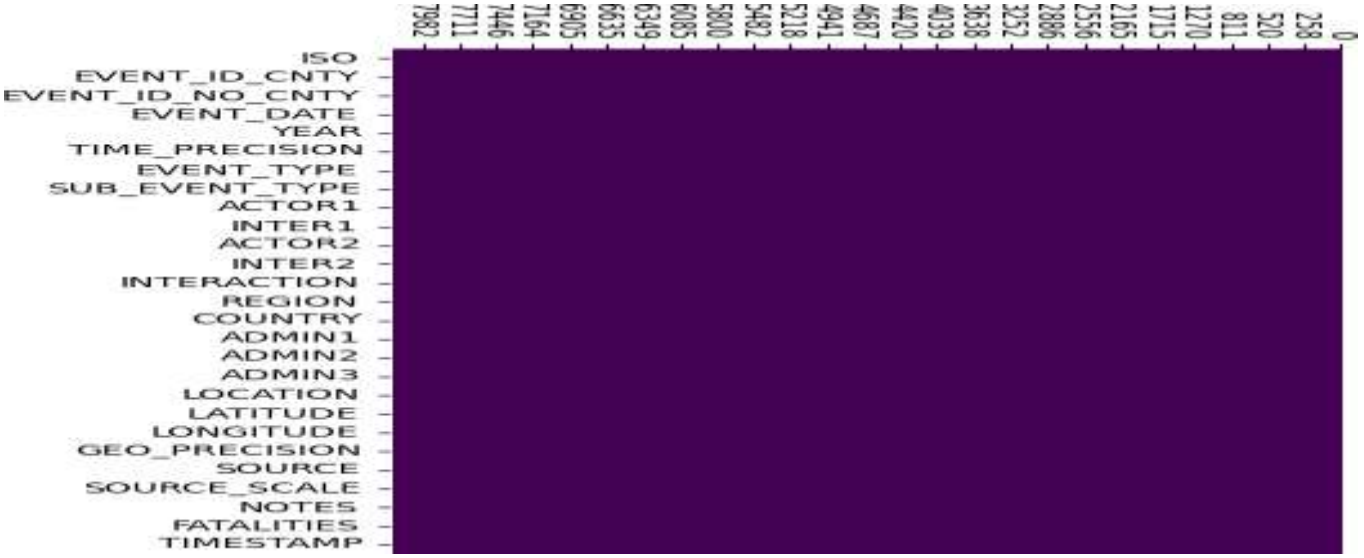
```
ISO                0
EVENT_ID_CNTY      0
EVENT_ID_NO_CNTY   0
EVENT_DATE         0
YEAR               0
TIME_PRECISION     0
EVENT_TYPE         0
SUB_EVENT_TYPE     0
ACTOR1             0
INTER1             0
ACTOR2             0
INTER2             0
INTERACTION        0
REGION             0
COUNTRY            0
ADMIN1             0
ADMIN2             0
ADMIN3             0
LOCATION           0
LATITUDE           0
LONGITUDE          0
GEO_PRECISION      0
SOURCE             0
SOURCE_SCALE       0
NOTES              0
FATALITIES         0
TIMESTAMP          0
dtype: int64
```

```
df.isnull().any()
```

```
ISO                False
EVENT_ID_CNTY      False
EVENT_ID_NO_CNTY   False
EVENT_DATE         False
YEAR               False
TIME_PRECISION     False
EVENT_TYPE         False
SUB_EVENT_TYPE     False
ACTOR1             False
INTER1             False
ACTOR2             False
INTER2             False
INTERACTION        False
REGION             False
COUNTRY            False
ADMIN1             False
ADMIN2             False
ADMIN3             False
LOCATION           False
LATITUDE           False
LONGITUDE          False
GEO_PRECISION      False
SOURCE             False
SOURCE_SCALE       False
NOTES              False
FATALITIES         False
TIMESTAMP          False
dtype: bool
```

As we can see in Table 22, using isnull().sum() and isnull().any() method it was assured that there are no missing values in the dataset. The process resulted in the removal of missing values in the "Actor 1" column, "ASSOC_ACTOR_1" and "ASSOC_ACTOR_2". As shown, the summation of every column missing value is 0; which means there is no missing value. Using any () method, all the columns have a Boolean value of false. This demonstrates that the method couldn't identify any missing value. Figure 44 shows the finalised data using the heatmap visualization technique.

Figure 44: Missing value on the conflict event dataset using the heatmap



B. Unnecessary columns

The above section shows that the data set has 27 columns. The objective of this research is to develop a model that can predict violent political conflicts in Ethiopia including spatial and temporal dimensions. In doing so, it is necessary to see all the columns in the data set and select the important and required dependent variables needed for further analysis and discard the unnecessary ones.

Figure 45: column names

```
df.columns
```

```
Index(['ISO', 'EVENT_ID_CNTY', 'EVENT_ID_NO_CNTY', 'EVENT_DATE', 'YEAR',
       'TIME_PRECISION', 'EVENT_TYPE', 'SUB_EVENT_TYPE', 'ACTOR1', 'INTER1',
       'ACTOR2', 'INTER2', 'INTERACTION', 'REGION', 'COUNTRY', 'ADMIN1',
       'ADMIN2', 'ADMIN3', 'LOCATION', 'LATITUDE', 'LONGITUDE',
       'GEO_PRECISION', 'SOURCE', 'SOURCE_SCALE', 'NOTES', 'FATALITIES',
       'TIMESTAMP'],
      dtype='object')
```

- ✓ **For predicting conflict event type:** To understand the type of each conflict incident, identifying the actors plays a crucial role. For this purpose, columns in the data set the

"Actor 1" column, "Actor 2" column, "Event type" column and "Sub event type" column are very important columns and have been selected.

- ✓ **For predicting the location:** The location of the event is also extremely crucial and will provide the whole picture of conflict prediction. The "Longitude" column, "Latitude" column, and "Location" column are required columns for predicting the exact place where the occurrences would occur. The most important column to predict location is the "Location" column but the "Longitude" column and "Latitude" column are considered for exploratory data analysis purpose to visualize the spacial distribution of conflict in the Ethiopia graphically.

To summarize, the aforementioned aspects are highly interconnected. Anticipating conflict or establishing a successful early warning system must include three major components: the type of event, the year in which the incident occurred, and the specific place where the incident occurred. In this regard, 8 columns that are critical to satisfying the event type, year, and location requirements are selected. This needs to be followed by eliminating the non-relevant columns that would not be used in the predictor model that will be developed. As demonstrated in the figures below, the extraneous columns are deleted, leaving only the crucial columns required for prediction.

Figure 46: Removing unnecessary columns

```
df=df.drop(['ISO', 'EVENT_ID_CNTY', 'EVENT_ID_NO_CNTY', 'EVENT_DATE', 'TIME_PRECISION', 'INTER1','INTER2', 'INTERACTION',
            'REGION', 'COUNTRY', 'ADMIN1','ADMIN2', 'ADMIN3','GEO_PRECISION', 'SOURCE', 'SOURCE_SCALE','FATALITIES',
            'TIMESTAMP'], axis=1)
```

Fig-47- visualizing the selected columns

```
df.columns
```

```
Index(['YEAR', 'EVENT_TYPE', 'SUB_EVENT_TYPE', 'ACTOR1', 'ACTOR2', 'LOCATION',
       'LATITUDE', 'LONGITUDE', 'NOTES'],
      dtype='object')
```

❖ Exploring the preprocessed conflict event dataset

The preprocessing approach applied in the data set is described in the preceding sections; this is followed by looking at some of the characteristics of the preprocessed data set.

Figure 48: Top 5 values in the preprocessed dataset

```
df.head(5)
```

	YEAR	EVENT_TYPE	SUB_EVENT_TYPE	ACTOR1	ACTOR2	LOCATION	LATITUDE	LONGITUDE	NOTES
0	2007	Battles	Armed clash	Gabra Ethnic Militia (Ethiopia)	Borana Ethnic Militia (Ethiopia)	Moyale	3.539	39.049	19 dead, breakdown unspecified
1	2007	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	Civilians (Ethiopia)	Abole	9.867	38.450	Oil exploration facility attacked. 65 Ethiopia...
2	2007	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	Civilians (Ethiopia)	Jijiga	9.350	42.800	Family of vitcim of earlier rebel attack, atta...
3	2007	Battles	Armed clash	Military Forces of Ethiopia (1991-2018)	ONLF: Ogaden National Liberation Front	Mustahil	5.244	44.732	4+ deaths
4	2007	Battles	Armed clash	Borana Ethnic Militia (Ethiopia)	Guji Ethnic Militia (Ethiopia)	Arero	4.750	38.817	On 7 June 2007, Borana Ethnic Militia (Ethiopi...

For conflict prediction, the first important thing is to predict the conflict events which are likely to happen. For predicting that, we have selected five columns which are "EVENT_TYPE",

“SUB_EVENT_TYPE”, “ACTOR1”, “ACTOR2” and “NOTES”. All the selected columns have an object data type and are not compiled for prediction purposes. (see Figure 49).

Figure 49: The unique values in the dataset

```
df.nunique(axis=0)

YEAR                16
EVENT_TYPE           6
SUB_EVENT_TYPE      22
ACTOR1             126
ACTOR2             142
LOCATION          1108
LATITUDE           903
LONGITUDE          888
NOTES             6000
dtype: int64
```

The values selected for forecasting the conflict events type have numerous distinct characteristics that are unsuitable for classification and comprehension. Furthermore, there are 6000 different values in the "NOTES" column, making it difficult for our model to learn anything from these values. But the notes columns somehow have helped to assess the nature of the conflict incident and to identify the conflict event type. The conflict events data were categorised into four categorical values using Excel to make them suitable for prediction. In this research the categorisation of the conflict events was based on conflict actor as well as conflict event type. The four categories are**:**

- ✓ **Inter-communal or inter-group conflict:** This conflict categorization includes conflict events occurring between two and more groups in Ethiopia that are organized along a shared communal identity such as ethnicity, religion, or livelihood (farmers and pastoralists).
- ✓ **Armed groups' violence against civilians:** The category includes any act of use of military force by armed groups against the civilian population and their livelihoods.
- ✓ **The battle between armed groups and the state:** This conflict type constitutes conflict between an organized armed group and the state's military forces or apparatus.
- ✓ **State violence against local communities:** This type includes violent military actions taken by the state military forces against local communities or civilians.

The conflict event type and actors involved as well as the notes columns are carefully observed in the process. In this process, all conflict events were categorised and tagged on the data set using Excel. As seen in the picture below, a column labelled "violent political conflict type" was added. Every conflict episode in this column was classified into one of the four categories. These categories can somehow show the kind of actors engaged in the incident as well as the sort of dispute that occurred at the time. By classifying the conflict event and the actors involved, the "Actor 1," "Actor 2," "Event_type," "sub_event_type," and "notes" columns were eliminated. As we strive to forecast the type of conflict occurrence, the values that are established replaced the full column indicated above. For instance, the "Armed groups violence against civilians," value indicates the sort of event that occurred, in which people were assaulted by armed forces. In these incidents, civilians and non-state armed groups are believed to be

involved. So, bytagging the conflict events with new conflict categories a values were created which are well descriptive and representative of the incidents.

Figure 50: The tagged conflict dataset

```
df_tagged.head(5)
```

	YEAR	EVENT_TYPE	SUB_EVENT_TYPE	ACTOR1	ACTOR2	violent political conflict type	LOCATION	LATITUDE	LONGITUDE	NOTES
0	2007	Battles	Armed clash	Gabra Ethnic Militia (Ethiopia)	Borana Ethnic Militia (Ethiopia)	Inter communal	Moyale	3.539	39.049	19 dead, breakdown unspecified
1	2007	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	Civilians (Ethiopia)	Armed groups violence against civilians	Abole	9.867	38.450	Oil exploration facility attacked. 65 Ethiopia...
2	2007	Violence against civilians	Attack	ONLF: Ogaden National Liberation Front	Civilians (Ethiopia)	Armed groups violence against civilians	Jijiga	9.350	42.800	Family of vitcim of earlier rebel attack, atta...
3	2007	Battles	Armed clash	Military Forces of Ethiopia (1991-2018)	ONLF: Ogaden National Liberation Front	Battle between armed groups and the state	Mustahil	5.244	44.732	4+ deaths
4	2007	Battles	Armed clash	Borana Ethnic Militia (Ethiopia)	Guji Ethnic Militia (Ethiopia)	Inter communal	Arero	4.750	38.817	On 7 June 2007, Borana Ethnic Militia (Ethiopi...

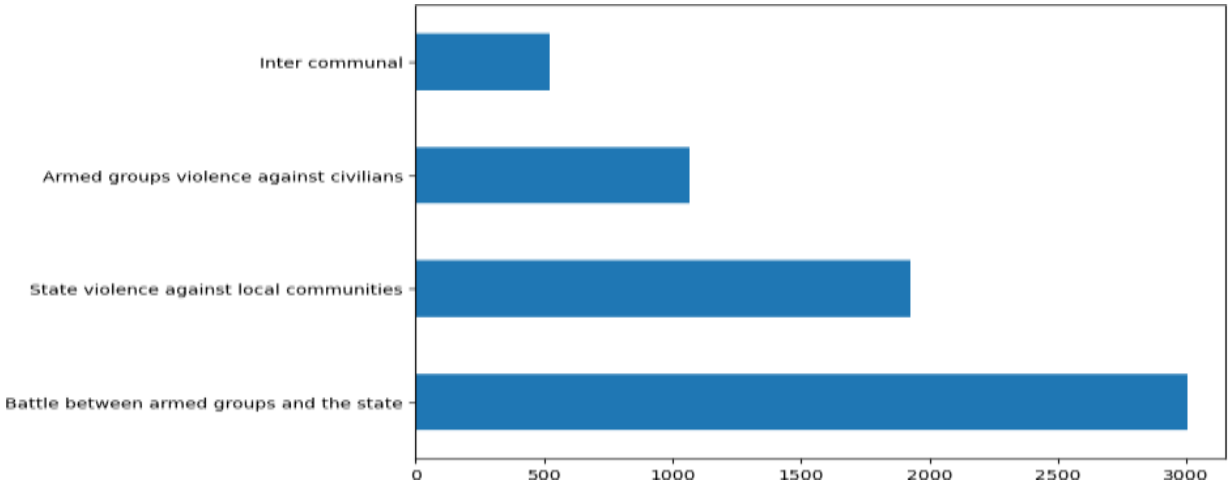
Figure 51: Unique values from the data set on the “violent political conflict type” column

```
df_tagged.violent_political_conflict_type.unique()
```

```
array(['Inter communal', 'Armed groups violence against civilians',
       'Battle between armed groups and the state',
       'State violence against local communities'], dtype=object)
```

Figure 51 above shows that there are four unique values on the added “violent_political_conflict_type” column. The number of each category is counted and described using the number of conflict incidents that happened in our country from 2007 to 2022. The bar graph below shows that among the four categories created for the conflict event, the "battle between armed groups and the state" type of incident took first place, followed by "state violence against local communities", "armed groups violence against civilians", and "inter-communal" incident types

Figure 52: The number of conflict incidents, 2007- 2022



The graph shows the majority of conflict incidents that occurred between 2007 and 2021 are violent political confrontations or clashes between the country's governmental forces and various non-state armed organizations at various times and places. Finally, it is important to delete the non-relevant columns.

Figure 53: Removing unnecessary columns

```
df_tagged=df_tagged.drop(['EVENT_TYPE',"SUB_EVENT_TYPE","ACTOR1","ACTOR2","NOTES"], axis=1)
```

```
df_tagged.columns
```

```
Index(['YEAR', 'violent_political_conflict_type', 'LOCATION', 'LATITUDE',
       'LONGITUDE'],
      dtype='object')
```

By removing the other unnecessary columns, now we are left with “year”, “violent conflict type”, “location”, “latitude” and “longitude” dependent variables that are important for the prediction purpose. These values are taken to predict when, where and what kind of violent political conflict will happen in the future.

Figure 54: The total dimension and the top 5 values of the tagged conflict events data

```
df_tagged.head(5)
```

	YEAR	violent_political_conflict_type	LOCATION	LATITUDE	LONGITUDE
0	2007	Inter communal	Moyale	3.539	39.049
1	2007	Armed groups violence against civilians	Abole	9.867	38.450
2	2007	Armed groups violence against civilians	Jijiga	9.350	42.800
3	2007	Battle between armed groups and the state	Mustahil	5.244	44.732
4	2007	Inter communal	Arero	4.750	38.817

```
df_tagged.shape
```

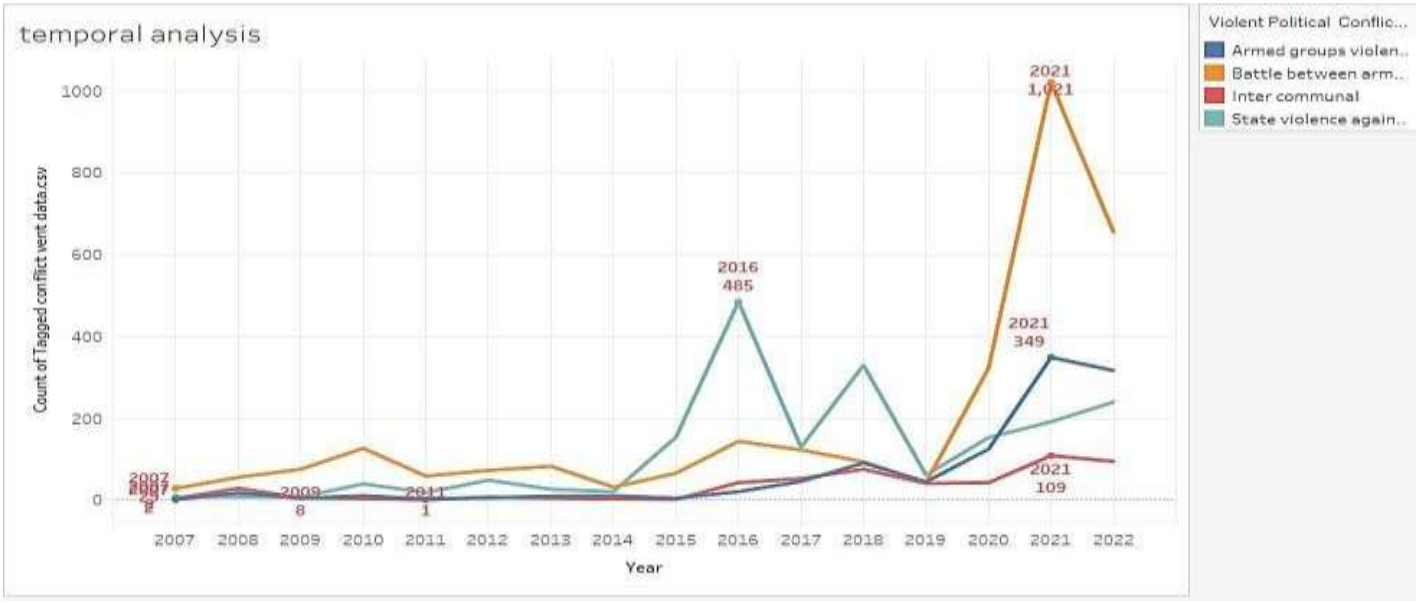
```
(6518, 5)
```

The final tagged data set comprises 6,518 rows and 5 columns. The number of rows represents the number of conflicts that occurred in each year from 2007 to 2021, and the number of columns represents the selected attributes that will be predicted. The data contains both time and space aspects and this is important to visualise the temporal and spatial dimensions.

❖ Exploring the temporal dimension of the tagged conflict event data

The temporal dimension helps to see change and continuities in the conflict landscape. It measures the natural frequencies of a given behaviour; which means how repeatedly it occurs in a given time. In this research, the temporal dimension of the data is related to how frequently a given conflict incident happens in a given year. For this purpose, two columns such as “YEAR” and “Violent political conflict type” are necessary. The first column indicates the temporality, from 2007 to 2021. The second, column shows the type of conflict incidents within the time frame. By combining these two columns, a full picture of the temporality or the frequency of the conflict incidents within the given period can be drawn.

Figure 55: Temporal dimension of violent political conflict in Ethiopia

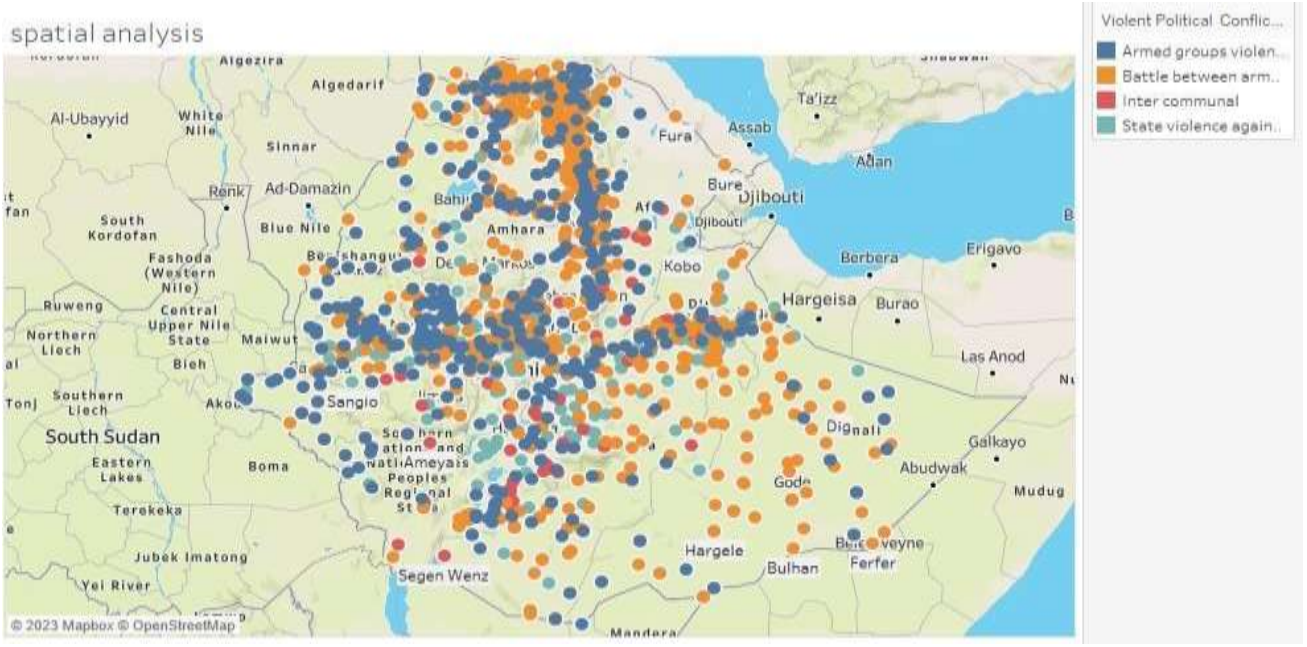


The figure shows the temporal dimension of the tagged conflict events dataset. The temporality of the dataset gives a better picture of the distribution across time. As shown the x-axis represents the "year" value and the y-axis demonstrates the "count of tagged conflict events". Plotting the line graph using these values gives the frequency of each conflict incident in each year. The 4 different conflict incident types are represented in different colours. The first conflict incident type that has the maximum value more than other incidents is "battle between an armed group and the state", which happened in the year 2007 and is represented in orange colour. The second conflict incident type that has maximum value is "State violence against civilians" in the year 2016. This is represented in the light blue colour. "Armed group violence against civilians" stood on the third-ranking by having the third maximum conflict incident type that happened in the year 2021 and represented in the blue-black colour in the line graph. Last but not least, the "inter-communal" conflict incident type stood having the smallest maximum value of conflict incidents in 2021 which is represented in red colour. The maximum values of three conflict incident types occurred in the year 2021, which has witnessed frequent conflicts in contrast to other years 2007-2021.

❖ **Exploring the spatial dimension of the tagged conflict event data**

The special dimension shows the geographic distribution of the dataset. It gives much deeper insights into the records in a given dataset by representing objects in a geographic coordinate system. In this research, the spatial dimension of the data is related to the location where the conflict incident occurred. For this purpose, four columns such as "Location", "Latitude", "Longitude" and "violent political conflict type" are important. The first three columns help to show the temporal dimensionality of the dataset and the last column shows the type of incident that happens in the given location. The combination of these four columns offers a full picture of the spatiality of the conflict incidents in Ethiopia as shown in the following figure.

Figure 56: The spatial dimension of the tagged conflict event data



The spatial dimension of the labelled conflict event dataset is depicted in Figure 56. We can visualize conflict episodes based on their location by viewing the spatiality of the dataset. I utilized the latitude and longitude values of each occurrence, together with their place name, to generate the figure. Using the "violent political conflict type" column, I noticed the four kinds of conflict occurrence types on the map by determining the exact region of the sites. By highlighting every conflict incidence on the map, it is possible to see where the majority of the conflict types occurred. The four different categories of conflict incidents are depicted in different colours on the map above. As seen in the graph, the conflict incidence category "Armed groups violence against civilians" represented in blue-black colour occurs primarily in the country's northern and western regions. The second event type, illustrated in orange, is "Battle between the state and armed groups." This conflict event type occurs mostly in the country's northern, western, and eastern regions. The third conflict type depicted in red is "intercommunal," which has occurred in modest numbers throughout the country, but the majority of them have occurred in Ethiopia's southwestern region. The final conflict category, displayed in light blue, is "state violence against civilians," which occurred mostlyin the country's central and western regions. In general, we can see from the geographical visualization that two conflict incident types, "armed group violence against civilians" and "battle between the state and armed groups," are more prevalent in Ethiopia, and that these conflict incident types have mostly occurred in the northern and western parts of the country.

4.3. Exploring and preprocessing merged dataset

This section presents the outcome of correlating dependent and independent variables and feature selection. Before that, it presents the process that led to merging the conflict indicators dataset and the conflict event data sets. The data sets have been explored and preprocessed separately in the previous chapters. Exploratory and preprocessing techniques are applied to the merged dataset.

4.3.1. Merging and exploring independent and dependent variables

Data analytics requires two main important variables to build a prediction model: independent and dependent variables. The first one is the set of variables that the model uses as a predictor feature. The second variable is the one that the model considers as the target variable. In this research to build the predictor model that serves as a CEWS tool, the features in the conflict indicators data are considered independent variables and the values in the conflict event data are considered dependent variables.

Table 23: Summary of the conflict indicators dataset and conflict events dataset

Data	Variables	
Conflict indicators	Voice and Accountability (Estimate)	Refugees and IDPs
	Political Stability (Estimate)	Democracy Index
	Government Effectiveness (Estimate)	GDP
	Regulatory Quality	Corruption
	Rule of Law	Human Development
	Control of Corruption	Population growth
	Security apparatus	Unemployment rate
	Factionalized Elites	Precipitation
	Group Grievance	Agricultural land
	Economy	Year
	Economic Inequality	
	Human Flight and Brain Drain	
	State Legitimacy	
	Public Services	
	Human Rights	
	Demographic Pressures	
Conflict event	Violent political conflict type	
	Year	
	Location	
	Latitude	
	Longitude	

As presented in the above table (56), the independent variables which are the conflict indicators have 27 variables and the dependent variable which is the conflict events data set has 4 variables. The conflict indicators dataset is yearly based, and for each year the indicators have a single value. The conflict events data set is also classified by year but for each year there is more than one or many conflict events that happened in each year. For this purpose, the conflict indicators in each year serve as a common denominator for the same year of conflict incidents. For example, the value of conflict indicator Economy in 2007, will serve as a common value for every incident that happen in 2007. For the analysis purpose, these two different data sets that have different dimensions are to build the predictor model, as follows.

Figure 57: Visualizing the top 5 values in the merged conflict data

	year	Voice and Accountability(Estimate)	Political Stability(Estimate)	Government Effectiveness(Estimate)	Regulatory Quality	Rule of Law	Control of Corruption	security apparatus	Factionalized Elites	Group Grievance
0	2007	-1.2	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	7.8
1	2007	-1.2	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	7.8
2	2007	-1.2	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	7.8
3	2007	-1.2	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	7.8
4	2007	-1.2	-1.81	-0.47	-0.98	-0.66	-0.63	7.5	8.9	7.8

❖ Exploratory Data Analysis on the merged dataset

To know the total dimension of the dataset, the shape () pandas method was used. As shown below, the data set has 6520 rows and 31 columns.

Figure 58: Total dimension of the merged data frame

```
df.shape
(6520, 31)
```

The 6520 rows indicate the values of the variables that are merged from the two data sets and the 31 columns indicate the different variables that are merged from the data sets. Following that, all the column names loaded into the pandas data frame were visualized using the columns () pandas method, as seen in the following figure, for further manipulation of the data set and ensuring that all column values were loaded correctly.

This process has offered the opportunity to manipulate all columns by their column names. A focus was also given to getting more information about the dataset using the info () method which informs about the number of columns, column labels, column data types, memory usage, and the number of cells in each column (non-null values).

Figure 59: Name of columns of merged data

```
df.columns
```

```
Index(['year', 'Voice and Accountability(Estimate)',
       'Political Stability(Estimate)', 'Government Effectiveness(Estimate)',
       'Regulatory Quality', 'Rule of Law', 'Control of Corruption',
       'security apparatus', 'Factionalized Elites', 'Group Grievance',
       'Economy', 'Economic Inequality', 'Human Flight and Brain Drain',
       'State Legitimacy', 'Public Services', 'Human Rights',
       'Demographic Pressures', 'Refugees and IDPs', 'External Intervention',
       'Democracy index', 'GDP per capita', 'Human Development',
       'Corruption perception index', 'Unemployment', 'Population',
       'precipitation ', 'Agricultural land (sq. km)', 'LOCATION', 'LATITUDE',
       'LONGITUDE', 'violent political conflict type'],
      dtype='object')
```

As shown in Figure 5, by using the info() method, on the non-null count the same values were observed which shows that there are no missing values on any of the columns due to the preprocessing steps applied to the individual data set. On the other hand, in the data type column, there are 26 float values, 3 int values and 2 object values. This suggests that the data has to be preprocessed because machine learning models do not accept object data types rather than numeric ones. To make sure that there is no missing value in each column and for checking the exact data type of each column, the count of missing values in each column and the data type of each column were visualised using isnull().sum() pandas method.

Figure 60: General information about the merged conflict data frame

```
RangeIndex: 6520 entries, 0 to 6519
Data columns (total 31 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   year                                  6520 non-null   int64
 1   Voice and Accountability(Estimate)    6520 non-null   float64
 2   Political Stability(Estimate)         6520 non-null   float64
 3   Government Effectiveness(Estimate)    6520 non-null   float64
 4   Regulatory Quality                    6520 non-null   float64
 5   Rule of Law                           6520 non-null   float64
 6   Control of Corruption                 6520 non-null   float64
 7   security apparatus                    6520 non-null   float64
 8   Factionalized Elites                  6520 non-null   float64
 9   Group Grievance                       6520 non-null   float64
 10  Economy                               6520 non-null   float64
 11  Economic Inequality                   6520 non-null   float64
 12  Human Flight and Brain Drain          6520 non-null   float64
 13  State Legitimacy                      6520 non-null   float64
 14  Public Services                       6520 non-null   float64
 15  Human Rights                          6520 non-null   float64
 16  Demographic Pressures                 6520 non-null   float64
 17  Refugees and IDPs                     6520 non-null   float64
 18  External Intervention                 6520 non-null   float64
 19  Democracy index                       6520 non-null   float64
 20  GDP per capita                        6520 non-null   float64
 21  Human Development                     6520 non-null   float64
 22  Corruption perception index           6520 non-null   int64
 23  Unemployment                          6520 non-null   float64
 24  Population                            6520 non-null   int64
 25  precipitation                         6520 non-null   float64
 26  Agricultural land (sq. km)            6520 non-null   float64
 27  LOCATION                              6520 non-null   object
 28  LATITUDE                              6520 non-null   float64
 29  LONGITUDE                             6520 non-null   float64
 30  violent political conflict type       6520 non-null   object
dtypes: float64(26), int64(3), object(2)
memory usage: 1.5+ MB
```

As the figure below shows, there are 0 missing values in each column. This shows that the preprocessing technique that has been applied in the individual data set has removed all the missing values effectively. It suggests that there is no work to be done for dealing with missing values. But based on the second result that illustrated the data types of each column, we have object values. According to the data types of variables, there is a need for much work to be done on the preprocessing step to deal with them.

Table 24: Null values in the merged conflict data frame

Checking for missing values and data types

```
df.isnull().sum()
```

```
year                                0
Voice and Accountability(Estimate)  0
Political Stability(Estimate)       0
Government Effectiveness(Estimate)  0
Regulatory Quality                  0
Rule of Law                         0
Control of Corruption               0
security apparatus                  0
Factionalized Elites                0
Group Grievance                     0
Economy                             0
Economic Inequality                 0
Human Flight and Brain Drain        0
State Legitimacy                    0
Public Services                     0
Human Rights                        0
Demographic Pressures               0
Refugees and IDPs                   0
External Intervention               0
Democracy index                     0
GDP per capita                      0
Human Development                   0
Corruption perception index         0
Unemployment                        0
Population                          0
precipitation                       0
Agricultural land (sq. km)          0
LOCATION                            0
LATITUDE                            0
LONGITUDE                           0
violent political conflict type     0
```

```
df.dtypes
```

```
year                                int64
Voice and Accountability(Estimate)  float64
Political Stability(Estimate)       float64
Government Effectiveness(Estimate)  float64
Regulatory Quality                  float64
Rule of Law                         float64
Control of Corruption               float64
security apparatus                  float64
Factionalized Elites                float64
Group Grievance                     float64
Economy                             float64
Economic Inequality                 float64
Human Flight and Brain Drain        float64
State Legitimacy                    float64
Public Services                     float64
Human Rights                        float64
Demographic Pressures               float64
Refugees and IDPs                   float64
External Intervention               float64
Democracy index                     float64
GDP per capita                      float64
Human Development                   float64
Corruption perception index           int64
Unemployment                        float64
Population                            int64
precipitation                       float64
Agricultural land (sq. km)          float64
LOCATION                             object
LATITUDE                            float64
LONGITUDE                           float64
violent political conflict type      object
```

Trend analysis with the final merged data

After merging the conflict events with the indicators, the trend that the conflict indicators show in line with the conflict incidents is discussed. As illustrated in Appendix I, every variable concerning conflict incidents has been visualized using the line graph. For clear visualization, every two indicators have been visualized with the conflict trend in one line graph. In each graph at the x-axis, the time interval is given, from 2007 to 2022 and on the y-axis the values of every variable can be found. All the values of the indicators are scaled down to the scale of 0-1 because of the high dimensionality in the data set. In addition, on each line graph, the maximum and minimum value of each indicator has been visualized.

As visualized in the line graph it can be observed that

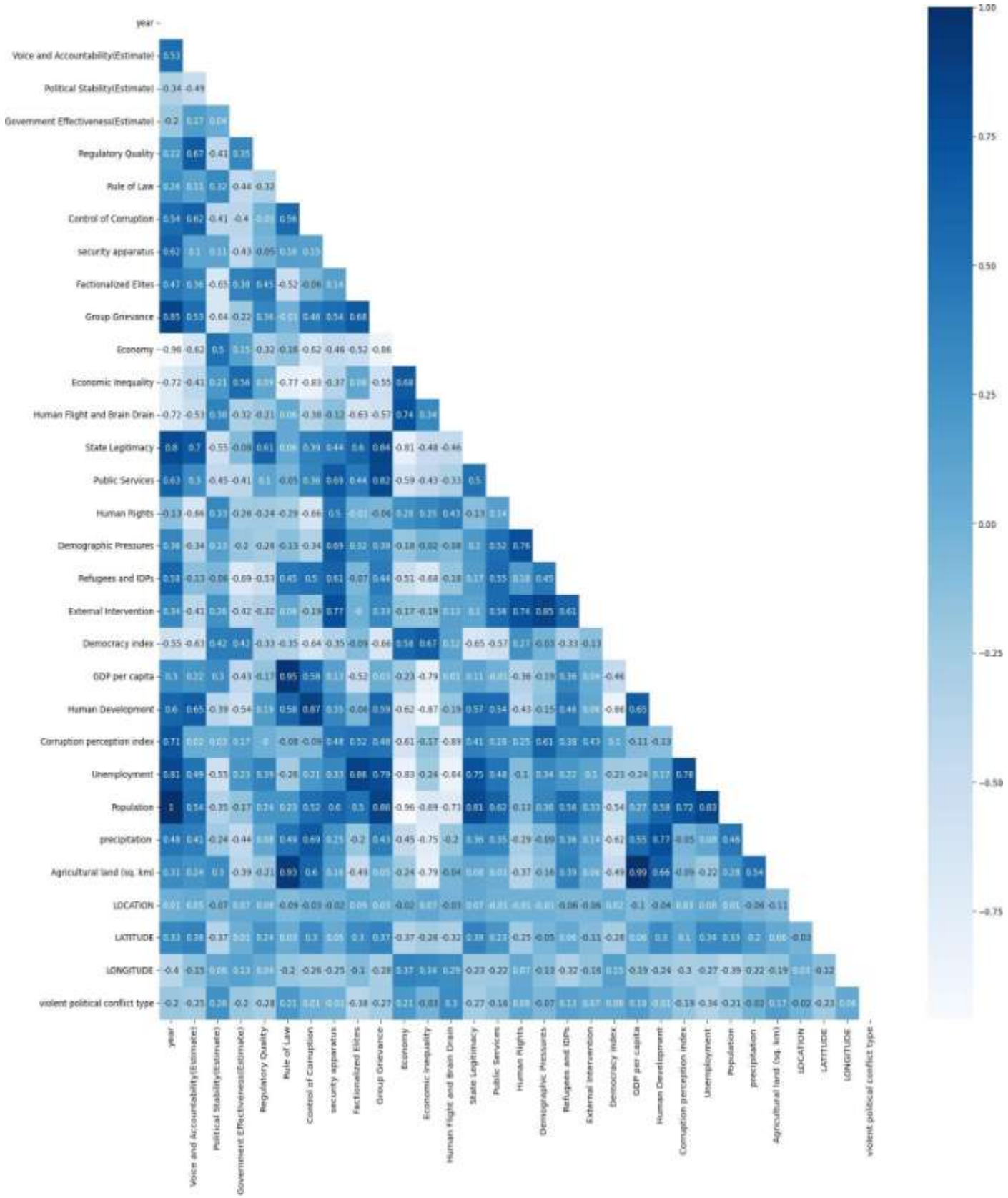
- ✓ Some indicators increase when the value of conflict incidents was also increasing and a decrease in their value when the values of conflict incidents decrease.
- ✓ Some indicators show an increase when the value of conflict incidents decreases and a decrease when conflict incidents increase

Based on the above-listed criterion, population, refugees and IDPs, human development, human flight and brain drain, group grievance, control of corruption, GDP per capita and democracy index have shown the above-stated trend with the conflict, at some time. At this point, it has to be noted that the indicators don't show a linear relationship with the dependent variable at the whole-time span. This suggests that the indicator might have a low positive or low negative correlation coefficient with the variable. To make sure about the positive correlation and negative correlation, the next section presents in detail a discussion by building the correlation matrix with every variable.

4.3.2. Building a correlation matrix

As a correlation matrix explains the linear association between each pair of variables, and it is constructed to visualize the association between dependent and independent variables by calculating their correlation coefficient and visualising it in a tabular form. The correlation matrix visualized below shows the correlation coefficients of each violent political incident indicator with the violent political incident that happened. All values have been correlated and their correlation coefficient value has been calculated. Each cell in the matrix shows the correlation between two variables. This helps to interpret the correlation coefficient between the dependent and independent variables. Violent political conflict type is the main dependent variable and the independent variables are the different indicators of conflict that serves as a predictor variable.

Figure 61 - correlation matrix for the final merged data set



Based on the results observed, the correlation coefficient values can be categorized into two groups: weakly positive correlation and weekly negative correlation. The positive correlation shows that when the value of one variable increases the other also increases and the negative correlation indicates when the value of one indicator increases the other decreases. The violent political conflict indicators such as voice and accountability, government effectiveness, regulatory quality, security apparatus, factionalized elites, group grievance, economic inequality, state legitimacy, public service, demographic pressure, human development, corruption index, unemployment, population and precipitation have a negative correlation with the dependent variables with different intensities. The rest conflict indicators political stability, rule of law, control of corruption, economy, human flight and brain drain, human rights, refugees and IDPs, external intervention, democracy index and GDP per capita show a positive correlation coefficient also with different intensities. Some indicators show a weekly positive and negative correlation with the dependent variable, but all of them with different intensities of relation with the variable.

4.3.3. Preprocessing the merged dataset

Based on the exploratory section applied above, two main issues need to be dealt with. First, the variables that have object data type should be changed into a compatible form of data type which is numeric values. Second, feature engineering will be applied to the data set to get the most important predictor variables to improve the final model's performance.

A. Data transformation

Data transformation is the process of taking raw data to transform it into data ready for analysis (Mallick 2021:5) Having this idea in mind, much of the data transformation has been done on the individual data set before it was merged such as dealing with missing values, removing unnecessary columns and checking for outlier values. Data set conversion or converting of categorical values into numerical values were carried out. As discussed above, in the final merged data set there are two object values in the dependent variables such as "Location" and "violent political conflict types". These variables have object data types that are not compatible with machine learning models. To deal with the categorical values, the label encoder encoding approach was used. Label encoding is converting each value in a column to a number. This Label Encoding in Python can be achieved using the sklearn Library, which provides a very efficient tool for encoding the levels of categorical features into numeric values. Label encoder encodes labels with a value between 0 and n_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier. By using the encoder, the categorical values in the data set were converted into numerical data as illustrated below.

Figure 62: encoding categorical variables in the data

```
from sklearn.preprocessing import LabelEncoder
enc = LabelEncoder()
enc.fit(df['violent political conflict type'])
df['violent political conflict type'] = enc.transform(df['violent political conflict type'])
```

```
from sklearn.preprocessing import LabelEncoder
enc = LabelEncoder()
enc.fit(df['LOCATION'])
df['LOCATION'] = enc.transform(df['LOCATION'])
```

After using the label encoder all the categorical values in the "Location" column and "violent political conflict type" are converted to numerical values as shown below.

Figure 63: Visualizing encoded categorical variables in the data

```
encoded.head(5)
```

	violent political conflict type	LOCATION
0	2	872
1	0	12
2	0	686
3	1	881
4	2	135

As seen, all the values on the columns are changed into numerical values according to their class values. Finally, by checking the data types of the columns and their unique values efforts were made to make sure that values are correctly encoded as shown in the below figures.

Figure 64: Checking the data type of the variables in the data

```
encoded.dtypes
```

```
violent political conflict type    int32
LOCATION                           int64
dtype: object
```

The data type of each column has been changed to integer values suitable for the model to read and understand. Another way to check the encoded values are correct is by checking the unique values in these columns. As discussed above, there are 4 groups/classes for the violent political incidents type and 1108 classes for the location value. Based on the unique values that are assigned to each value, the encoded values should also have the same encoded unique values.

Figure 65: Checking the unique values of the variables in the data within the two encoded columns

```
encoded.nunique(axis=0)
```

```
violent political conflict type       4
LOCATION                           1108
dtype: int64
```

The encoded columns also have the same unique value as the original data set. This shows that the encoder has perfectly encoded the object values found in the data set.

B. Feature selection

Predictive modelling such as classification or regression involves learning from examples or cases from the domain that characterize the problem that will be solved which is generally called data. The data that the model will learn typically cannot be raw data. One of the techniques to give well-preprocessed data to the model is by applying feature selection. Feature selection is the process that reduces the number of input variables when developing a predictive model (Caijie, 2017: 20). This approach is desirable because it reduces the number of input variables to both reduce the computational cost of modelling and, improve the performance of the model. (Charonyktakis, 2021: 5)

By using the merged conflict dataset, a predictive model has been developed. To give a well preprocessed data for the model applying feature selection was very crucial. Using a supervised machine learning algorithm to build the predictive model, supervised feature selection techniques were applied by using the target variable, such as methods that remove irrelevant variables. The technique that was applied to select features is known as intrinsic. Intrinsic methods use algorithms such as decision trees that perform automatic feature selection during training. (Brownlee, 2019: 15) Here an extra-tree classifier was applied to select features that have greater importance value. This method implements a Meta estimator that fits several randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The figure below shows the Python code that was applied to find the importance value of each of the independent variables.

Figure 66 - Calculating the feature importance of each of the independent variables

```
from pandas import read_csv
from sklearn.ensemble import ExtraTreesClassifier
array = df.values
X = array[:,0:27]
Y = array[:,27]
# feature extraction
model = ExtraTreesClassifier(n_estimators=10)
model.fit(X, Y)
print(model.feature_importances_)
```

```
[0.06422995 0.00875159 0.03887162 0.01627717 0.01223456 0.0354366
 0.04058535 0.01581593 0.03042887 0.07401756 0.03268323 0.15083284
 0.03866546 0.01088772 0.01447685 0.01533522 0.02067111 0.06956451
 0.02303313 0.03455159 0.03376862 0.04931439 0.04314599 0.04109126
 0.06450591 0.0053213  0.01550165]
```

The feature importance of each independent variable has been calculated using the extra-tree classifier. As observed in the figure the values are presented in the array form. Each feature value has to be ranked and related to its indicator name to understand the full picture and get the most relevant features.

Table 25: Ranking independent variables based on feature importance

Indicators	Feature importance	Rank
Economic Inequality	0.15083284	First
Group Grievance	0.07401756	Second
Refugees and IDPs	0.06956451	Third
Population	0.06450591	Fourth
Year	0.06422995	Fifth
Human Development	0.04931439	Sixth
Corruption perception index	0.04314599	Seventh
Unemployment	0.04109126	Eighth
Control of Corruption	0.04058535	Ninth
Political Stability	0.03887162	Tenth
Human Flight and Brain Drain	0.03866546	Eleventh
Rule of Law	0.0354366	Twelfth
Democracy index	0.03455159	Thirteenth
GDP per capita	0.03376862	Fourteenth
Economy	0.03268323	Fifteenth
Factionalized Elites	0.03042887	Sixteenth
External Intervention	0.02303313	Seventeenth
Demographic Pressures	0.02067111	Eighteenth
Government Effectiveness	0.01627717	Nineteenth
security apparatus	0.01581593	Twentieth
Agricultural land	0.01550165	Twenty-first
Human Rights	0.01533522	Twenty-second
Public Services	0.01447685	Twenty third
Regulatory Quality	0.01223456	Twenty fourth
State Legitimacy	0.01088772	Twenty-fifth
Voice and Accountability	0.00875159	Twenty-seventh
Precipitation	0.0053213	Twenty-eighth

Each indicator has been ranked based on its future importance value. The feature importance value of each variable is very close to each other, for example, the first indicator that is ranked in the first place has a feature importance value of 0.15083284 and the last indicator has a value of 0.0053213 feature importance. The feature importance value difference between these two indicators is around 0.05030071. Considering the differences in the feature importance values between indicators, deciding the number of features to consider is very challenging. To overcome this problem, in the next chapter when the predictive modelling was built, it was experimented with a greater number of features and tested with accuracy till it was reached no change in accuracy by adding extra features to the model.

CHAPTER FIVE: MODEL BUILDING AND EVALUATION

Introduction

In data analytics, there are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement. In this research, to develop a predictive model, a supervised type of machine learning algorithm is applied primarily due to the nature of the dataset. Supervised machine learning technique offers to discover the relationship between independent variables and dependent variables. But there are two main categories when dealing with this technique. The first is a regression that takes non-stop values for the target value and the second is a classification that takes class labels for the output variables. The latter category is applied due to the target classes in the dataset having categorical class labels. Classification is the process of predicting the class labels of data points. The application of Classification predictive modelling is not without reason. This approach works by approximating a mapping function (f) from input variables (X) to discrete output variables (y) finally it classifies each object in a set of statistics into one of a predefined set of lessons or groups. The main goal of this research is to train the best-performing predictive model possible using pre-processed data. In this section, the pre-processed data is used to build machine-learning models to predict conflict incident types. The label data is available for training and supervised machine learning models are used for modelling in thisstudy.

5.1. Splitting the dataset into Training and Test Sets

Train test split is a model validation process that allows to simulate how the model performs with new data. To apply this technique, the dataset is split into a training set and a testing set. The training set is used for training the model, and the testing set is used to test the model. This allows us to train the models on the training set and to test their accuracy on the unseen testing set. For training a model the dataset is initially split into two sections: 'Training data', and 'Testing data'. By classifying the dataset, the classifier model will be trained using a 'training dataset',' and followed by a test of the performance of the classifier on an unseen 'test dataset'. The classifier model is, thus, developed by using the training dataset. Subsequently, the performance of the model is going to be tested using the test dataset.

The main issue in splitting the data is how to decide the ratio of the training and testing dataset. This Splitting can be varied according to the dataset's shape and size. The dataset is split into the data-set ratio of 80:20. The ratio implies, 80% of the dataset will be used for the training purpose and 20% of the dataset will be used for testing the performance of the dataset. To split the dataset into a training and a testing set the train_test_split() function is used as visualized below.

Figure 67: Splitting the data set into training and testing

```
# splitting the data set in to training and testing dataset
from sklearn.model_selection import train_test_split
array = df.values
X = array[:,0:27]
Y = array[:,27]
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,
                                                    random_state=42)
```

As shown above, to split the dataset into training and testing set first the independent and dependent variables are stored in the variable "X" and "Y" consequently. After storing the data set into the variables each variable has been split into "X_train", "X_test", "Y_train" and "Y_test" to store the training and testing values of each variable. In the final merged conflict incidents dataset, the first 27 columns are stored in the X variable (independent variables) and the last 4 columns are stored in the Y variable (dependent variables). The whole dataset was split into training and testing datasets by using the train_test_split function from sklearn bypassing 0.2 as the test size and 42 as a random state value. By using the above function, the dataset was split into 80% for training and 20% for testing.

5.2. Building Classification Prediction Models

Among the various classification techniques or classifiers, three machine learning algorithms were selected for building the predictive model by taking into consideration the nature of the research and the dataset. The research dataset has multiple output variables or target classes known as "Location" and "Violent political conflict types". Since this research is dealing with multiple target values a machine learning algorithm that supports multiple targets was paramount. The approach selected for this purpose is the Multi-Output Classifier. It is a type of machine learning that predicts multiple outputs simultaneously. In Multi-Output classification, the model provides two or more outputs after making any prediction. Other types of classification algorithms are used for estimators which do not support multi-target classification independently such as Random Forest and Logistic regression. Since these algorithms couldn't handle multioutput classification, the algorithms need to be combined with the multi-output classifier to make predictions. The selected machine learning algorithms that are combined with Multi-Output Classifier include, Random Forest, Gradient Boosting and Gaussian Navies Bayes which are selected for the model building. These models are trained on the training dataset using fit() and tested using the testing dataset. Finally, a model that has the best performance is selected from the experiments and used to answer the research questions.

5.2.1. Building predictive model using Random Forest machine learning algorithm

Random forests machine learning algorithms are popular ensemble machine learning algorithms and are useful when there is a labelled target variable. It can be used for both classification and regression types of problems and the algorithm produces good predictions that can be understood easily. The opportunity to handle large datasets efficiently is its advantage. The random forest algorithm provides a higher level of accuracy in predicting outcomes than the decision tree algorithm. The algorithm first selects random samples from a given data or training set then the algorithm constructs a decision tree for every training data and subsequently voting will take place by averaging the decision tree. In the end, the algorithm selects the most voted prediction result as the final prediction result.

The first experiment has been done using the final merged dataset from the previous chapter. The random forest algorithm is combined with Multi-Output Classifier and is applied to develop a predictive model using the merged conflict dataset. The splitting of the dataset into the training and testing data helps to fit them into the machine learning algorithm using the training data., the random forest algorithm and Multi-Output Classifier is imported from

sklearn. After importing the algorithm, the training dataset has been fitted into the algorithm, and then a prediction is made using the test dataset. This was followed by evaluation using the test label dataset. Finally, the performance of the model is checked using different metrics that can evaluate multi-output or multi-labels. The performance matrices that are used for this research include Training accuracy, Mean Squared Error, Classification report and Confusion Matrix. Training accuracy measures how much the training dataset fits into the prediction model; the Mean squared error measures the most common loss function that measures the expected squared distance between what the model predictor predicts for a specific value and what the true value is; the Classification report displays the precision, recall, F1, and support scores for the model; and the Confusion matrix summarizes the performance of a classification model by comparing its predicted labels to the true labels. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of the model's predictions.

Figure 68: Building a predictive model using Random Forest machine learning algorithm

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(random_state=42)
multi_target_forest = MultiOutputClassifier(forest, n_jobs=2)
multi_target_forest.fit(x_train, y_train)
y_pred = multi_target_forest.predict(x_test)
print(multi_target_forest.score(x_train, y_train))
```

```
0.3065567484662577
```

As observed above, a predictive model using the multioutput classifier and random forest classifier machine learning algorithm is trained with the training data set and it shows 30% of training accuracy. The trained model has made a prediction, by using the X_test data that has been kept for evaluating the model without the class label values.

The first index in the array is for the “Location” and the second index is for “violent political conflict type”. The values are in numerical form because of the encoding that has been applied in the preprocessing technique. After finishing building the models, It can be decoded for interpretation purposes. Having the predicted value, the model can also be evaluated with different metrics for comparison purposes with other models.

Figure 69: Measuring the performance of the model using mean squared error

```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, multi_target_forest.predict(x_test))
print("The mean squared error (MSE) on test set: {:.4f}".format(mse))
```

```
The mean squared error (MSE) on test set: 9.2216
```

The above figure shows that the prediction model that is built using the random forest machine learning algorithm shows 9.216 MSE. Because the lower the MSE value the more accurate the model, other models that are going to be built need to be compared with this MSE value if they result in a lower MSE value.

Figure 70: Measuring the performance of the model using accuracy

```
import numpy as np
np.mean(np.all(y_test == y_pred, axis=1))
```

```
0.299079754601227
```

Since the accuracy score in the sklearn doesn't support multiple outputs for this research, the np.mean() function is used. By taking the concept of truth must equal pred, the model is evaluated using this approach. As the truth values (Y_test) and the predicted values (Y_pred) are two numpy arrays the case of truth == pred can be applied. This method creates a new array in which 1 indicates the corresponding values in truth and pred are equal and 0 otherwise. By taking the mean of all of these values, it shows how accurate the predictor is on average. This approach measures how many data points the model predicted correctly. The above model with random forest shows 29% accuracy which is a very low value. For better accuracy, it is also checked with accuracy metrics by taking each output or label (Y1 and Y2) individually.

Figure -71: Measuring the performance of the model using the accuracy of each label

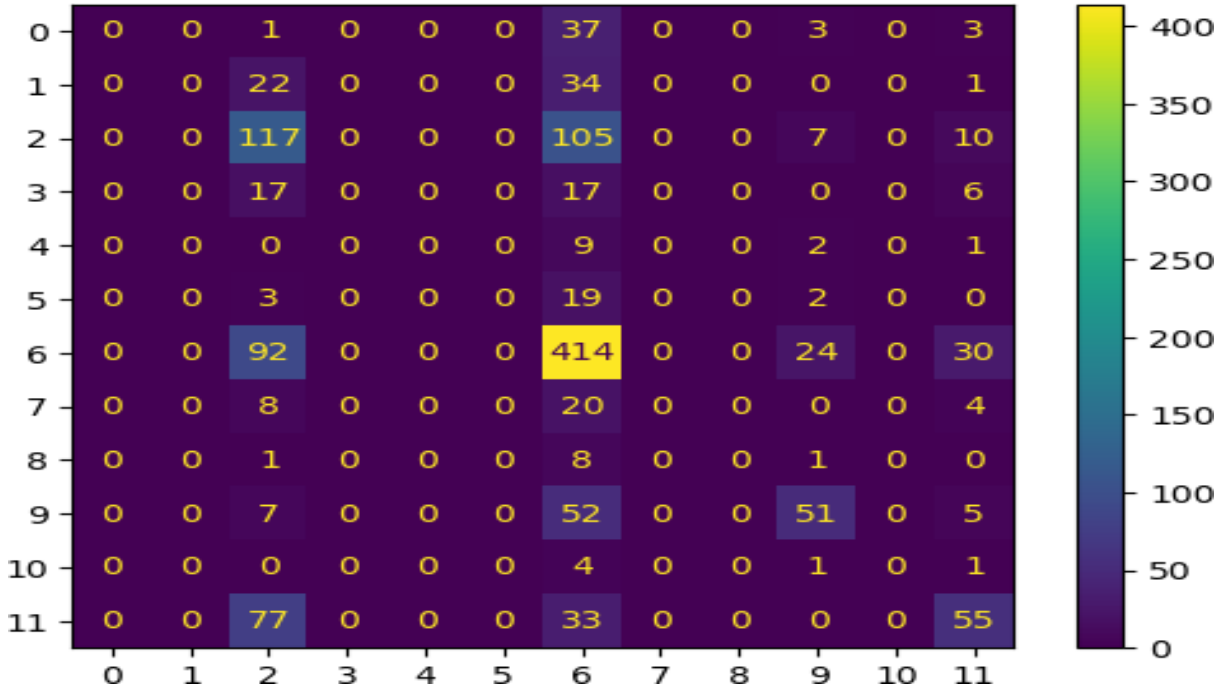
```
from sklearn.metrics import accuracy_score
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC1,AC2)
```

```
0.48849693251533743 0.557515337423312 8
```

The model built with Random Forest shows a better performance of accuracy having 48% for the location label and 55% for the violent conflict incident types. That shows that the model performed better in predicting the individual labels than predicting both at the same time. The other models are also evaluated using both mechanisms.

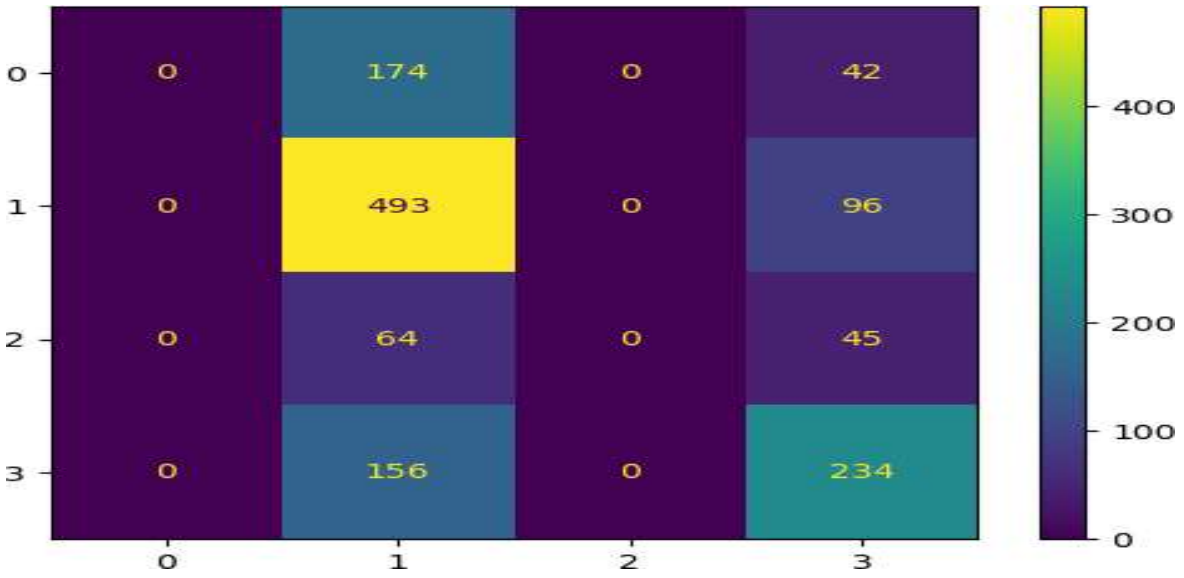
The next performance measure that is going to be applied is the confusion matrix. These metrics can also be applied to multi-class classification problems too. When dealing with multiclass prediction, the matrix needs to have more than two rows and columns depending on the number of labels that the model is tasked to predict in each class. Two confusion matrices have been built for each class (y1, y2) as visualized in the below figure.

Figure 72: Measuring the performance of the model using the confusion matrix for the first class



The above figure shows the confusion matrix for the first class that predicted the values of location for the violent conflicts. Each row in the above matrix represents the instances in the actual class, and each column represents the instances in a predicted class. In this matrix, the diagonal cells show correctly classified samples. So, the correctly classified (TP and TN) samples can quickly be visualized for each label and the off-diagonal cells show model errors (FP and FN). The matrix total results in 637 TP, TF values, 753 FP, FN values out of the total 1390 testing values. Now the confusion matrix for the violent political type class is going to be plotted below.

Figure73: Measuring the performance of the model using the confusion matrix for the first class



The confusion matrix for the second class, which is predicted values of violent political conflict types, resulted totally with 727 TP, TP values and 577 FP, FN values out of 1390 testing values. In this regard, the model correctly predicted the values of TP and TN. The final performance metrics applied to the predictive model is the classification report. This method also supports multi-class classification by measuring the precision, recall and f1 score of each class individually as presented in the figure below.

Figure 74: Measuring the performance of the model using the classification report for the first class

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	57
2.0	0.34	0.49	0.40	239
3.0	0.00	0.00	0.00	40
7.0	0.55	0.74	0.63	560
10.0	0.56	0.44	0.50	115
12.0	0.47	0.33	0.39	165
micro avg	0.49	0.54	0.51	1176
macro avg	0.32	0.33	0.32	1176
weighted avg	0.45	0.54	0.49	1176

As demonstrated above, the first class which is the location of incidents, the precision, recall and f1 score values of the class labels which have predicted values by the model have been calculated. The support column at the last column in the figure represents the number of values that are predicted for a specific label for a given class. Averagely the model results are 32%, 33% and 32% for precision, recall and f1-score metrics respectively. The second classification report is also calculated for the second class, which is the violent political conflict type and is presented below.

Figure 75: Measuring the performance of the model using classification report for the first class

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	216
1.0	0.56	0.84	0.67	589
2.0	0.00	0.00	0.00	109
3.0	0.56	0.60	0.58	390
accuracy			0.56	1304
macro avg	0.28	0.36	0.31	1304
weighted avg	0.42	0.56	0.48	1304

The classification report for the second class, which is the type of violent conflict incident types, the precision, recall and f1 score values of the class labels that have predicted values by the model is calculated. It resulted in 28%, 36% and 31% for precision, recall and f1-score matrices respectively.

5.2.2. Building predictive model using Gradient Boosting machine learning algorithm

Gradient boosting is another machine learning algorithm with a prediction speed and accuracy, especially with large and complex datasets. It has the benefit of minimizing the bias error of the model. This algorithm helps to build models sequentially and these subsequent models are more likely to reduce the errors of the previous model. This machine learning algorithm is selected due to the nature of the algorithm to capture complex patterns in thedata.

Like the above model building, the gradient boosting algorithm is also combined with Multi Output Classifier to make it capable of handling multioutput features and then applied to develop a predictive model using the merged conflict dataset. The Gradient Boosting classifier and Multi-Output Classifier have been imported from sklearn. After importing the algorithm, the training dataset has been fitted into the algorithm, and then a prediction has been made using the test dataset. After that, the prediction that has been made is evaluated using the test label dataset. Finally, by using the above performance metrics that are used to evaluate the above model are also applied to measure the performance of this model too.

Figure -76: Building a predictive model using Gradient Boosting machine learning algorithm

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
print(multi_target_forest.score(x_train, y_train))
```

```
0.3065567484662577
```

From the figure, a predictive model using the multi-output classifier and Gradient Boosting classifier machine learning algorithm was trained with the training data set and the result shows 30% of training accuracy similar to the random forest algorithm. The trained model has made a prediction, by using the X_test data that has been kept for evaluating the model without the class label values. The predicted values are visualized in the below figure in an array form.

The first index in the array is for the “Location” and the second index is for “violent political conflict type”. The values are in numerical form. After model building, it is decoded for interpretation purposes. The model is also evaluated and compared with othermodels.

Figure -78- Measuring the performance of the model using mean squared error

```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, multi_target_forest.predict(x_test))
print("The mean squared error (MSE) on test set: {:.4f}".format(mse))
```

```
The mean squared error (MSE) on test set: 9.2216
```

As described in the above figure the prediction model that is built using the Gradient Boosting classifier machine learning algorithm shows 9.216 MSE the same as the first model that has been built with the random forest algorithm. Because the first model and the second one have the same MSE value, the third model is compared with the above two models' value of MSE.

Figure -79: Measuring the performance of the model using accuracy

```
import numpy as np
np.mean(np.all(y_test == y_pred, axis=1))
```

```
0.3013803680981595
```

By using the np.mean() method, the performance of the model predicting the correct instance is measured. The model built with the Gradient Boosting Classifier shows 30% accuracy, which is a little increase in contrast to the random forest machine learning algorithm. This shows the overall performance of the model with the two classes. The model is also passed through and checked for its predicting performance of each class individually.

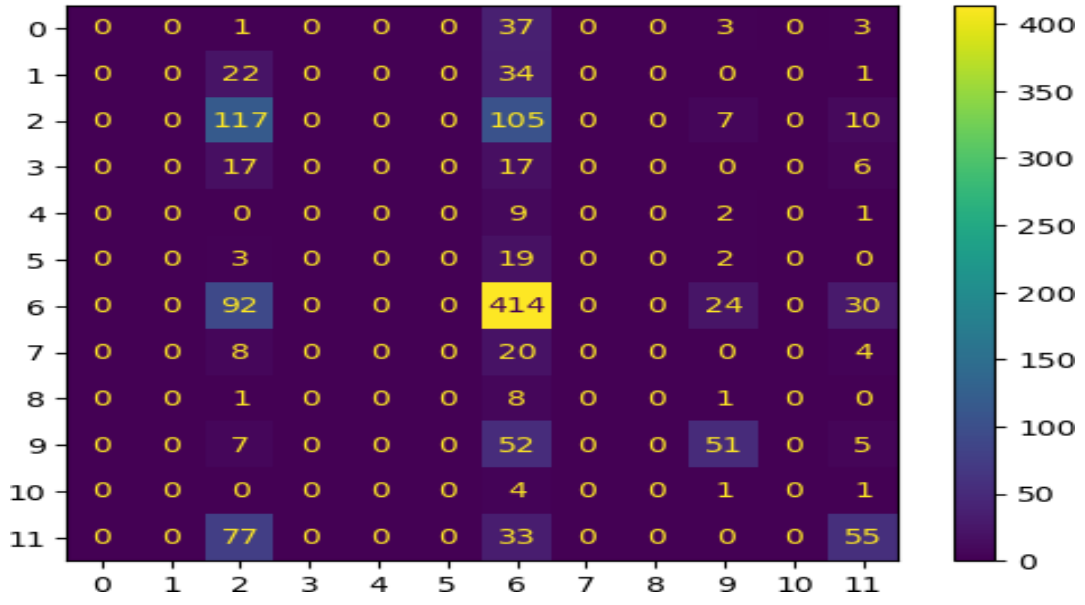
Figure -80- measuring the performance of the model using the accuracy of each label

```
from sklearn.metrics import accuracy_score
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC1,AC2)
```

```
0.48849693251533743 0.5575153374233128
```

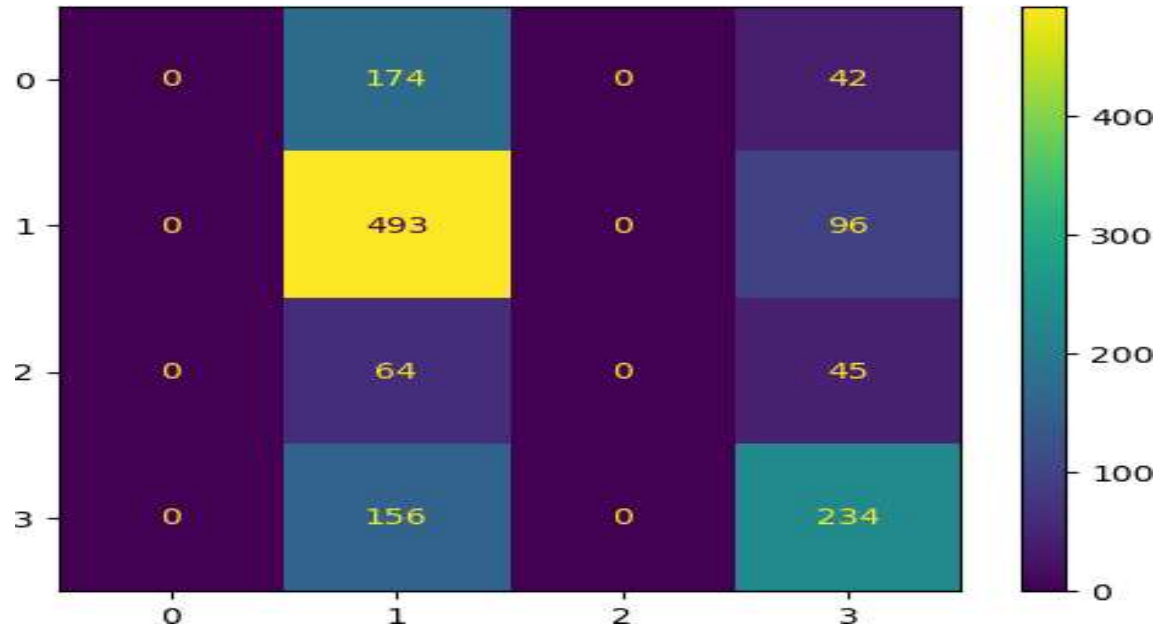
The model with Gradient boosting classifier shows an accuracy of 48% for the location label and 55% for the violent conflict incident types, similar to the above model. This shows how both models performed better predicting the individual labels than predicting both at the same time. The other performance measure applied is the confusion matrix. Like the above model, two confusion matrices have been built for each class (y1, y2) as visualized in the below figure.

Figure 81: Measuring the performance of the model using the confusion matrix for the first class



The figure shows the confusion matrix for the first class by using the second model built with a Gradient boosting classifier for the location of violent conflicts. The above matrix resulted in, 637 TP, TN values and 753 FP, FN values out of the total 1390 testing values. The model also predicts the same output for the second class compared with the above model. The confusion matrix plotted below shows the model-predicted values of violent political conflict types. The model resulted in 727 TP, TN values and 577 FP, FN values out of 1390 testing values.

Figure 82: measuring the performance of the model using the confusion matrix for the first class



The final performance metrics applied for the predictive model is the classification report. This method also supports multi-class classification by measuring the precision, recall and f1 score of each class individually as presented in the figure below.

Figure 83: Measuring the performance of the model using the classification report for the first class

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	57
2.0	0.34	0.49	0.40	239
3.0	0.00	0.00	0.00	40
7.0	0.55	0.74	0.63	560
10.0	0.56	0.44	0.50	115
12.0	0.47	0.33	0.39	165
micro avg	0.49	0.54	0.51	1176
macro avg	0.32	0.33	0.32	1176
weighted avg	0.45	0.54	0.49	1176

As demonstrated above, by visualizing the classification report for the first class (location of incidents) the precision, recall and f1-score values of the class labels which are predicted values by the model are calculated. Based on the report presented above, the model results with 32%, 33% and 32% respectively for the precision, recall and f1-score performance matrices. The

second classification report is calculated for the second class and presented in the below figure. The model also resulted in 28%, 36% and 31% for precision, recall and f1-score performance matrices respectively.

Figure -84: Measuring the performance of the model using classification report for the first class

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	216
1.0	0.56	0.84	0.67	589
2.0	0.00	0.00	0.00	109
3.0	0.56	0.60	0.58	390
accuracy			0.56	1304
macro avg	0.28	0.36	0.31	1304
weighted avg	0.42	0.56	0.48	1304

5.2.3. Building predictive model using Gaussian Naive Bayes machine learning algorithm

Gaussian Naïve Bayes (GNB) is one of the Naïve Bayes algorithms suitable for a dataset that has continuous features. The main reason for selecting a machine learning algorithm is depending on the type of dataset. This algorithm shows that each feature has an independent capacity for predicting the output variable. Finally, combining the prediction of all features returns the final prediction by calculating the probability of the dependent variable to be classified in each group.

Like the above model building, the GNB algorithm is also combined with Multi Output Classifier to make it capable of handling multioutput features. This is applied to develop a predictive model using the merged conflict dataset. The GNB classifier and Multi-Output Classifier are imported from sklearn. The training dataset was fitted into the algorithm, and then a prediction has been made using the test dataset. Evaluation of the model conducted using performance metrics.

Figure -85: Building a predictive model using GNB machine learning algorithm

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.naive_bayes import GaussianNB
GNB = GaussianNB()
multi_target_GNB = MultiOutputClassifier(GNB, n_jobs=2)
multi_target_GNB.fit(x_train, y_train)
y_pred = multi_target_GNB.predict(x_test)
print(multi_target_GNB.score(x_train, y_train))
```

```
0.2223926380368098
```

A predictive model using the multioutput classifier and GaussianNB classifier machine learning algorithm was trained with the training data set and it shows 22% of training accuracy. This is less than the other two models. Utilizing the X_test data that was used to evaluate the model without the class label values, the trained model generated a prediction. The predicted values are visualized in an array form.

The first index in the array is for the “Location” and the second index is for “violent political conflict type”. The values are in numerical form because of the encoding that has been applied in the preprocessing technique. After model building, it is evaluated using matrices for comparison purposes with other models.

Figure 86: Measuring the performance of the model using mean squared error

```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, multi_target_forest.predict(x_test))
print("The mean squared error (MSE) on test set: {:.4f}".format(mse))
```

```
The mean squared error (MSE) on test set: 9.2216
```

As described in the above figure, the prediction model built using the GaussianNB classifier machine learning algorithm shows 9.2216 MSE the same as the above two models. Since the three models have the same MSE value, other matrices are considered to measure the performances of the models.

Figure -88: Measuring the performance of the model using accuracy

```
import numpy as np
np.mean(np.all(y_test == y_pred, axis=1))
```

```
0.21779141104294478
```

The np.mean() method is used to measure the performance of the model to predict the correct instance. The model built with the GNB classifier shows 21% accuracy. This is lower than the above models' accuracy values (30% and 30% respectively). This might happen due to the data not having Gaussian data distribution or normal distribution. This value shows the overall performance of the model with predicting the two classes. The model is checked for the performance of predicting each class individually.

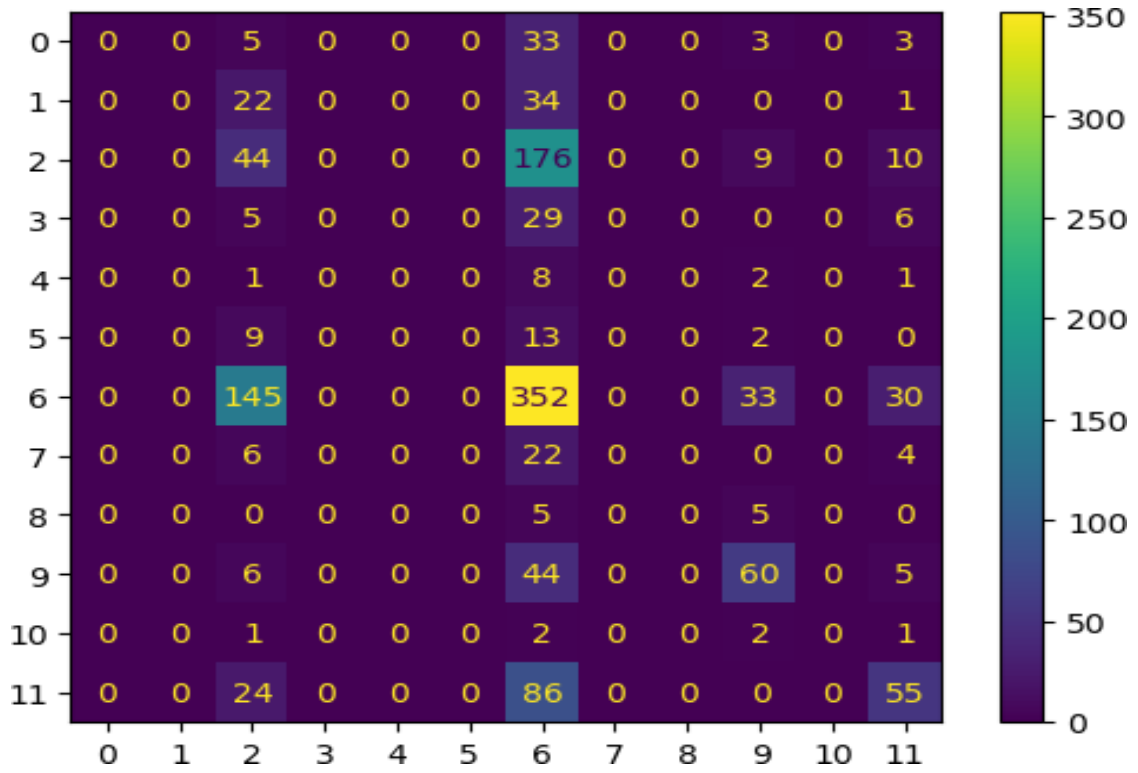
Figure -89: Measuring the performance of the model using the accuracy of each label

```
from sklearn.metrics import accuracy_score
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC1,AC2)
```

```
0.48849693251533743 0.5575153374233128
```

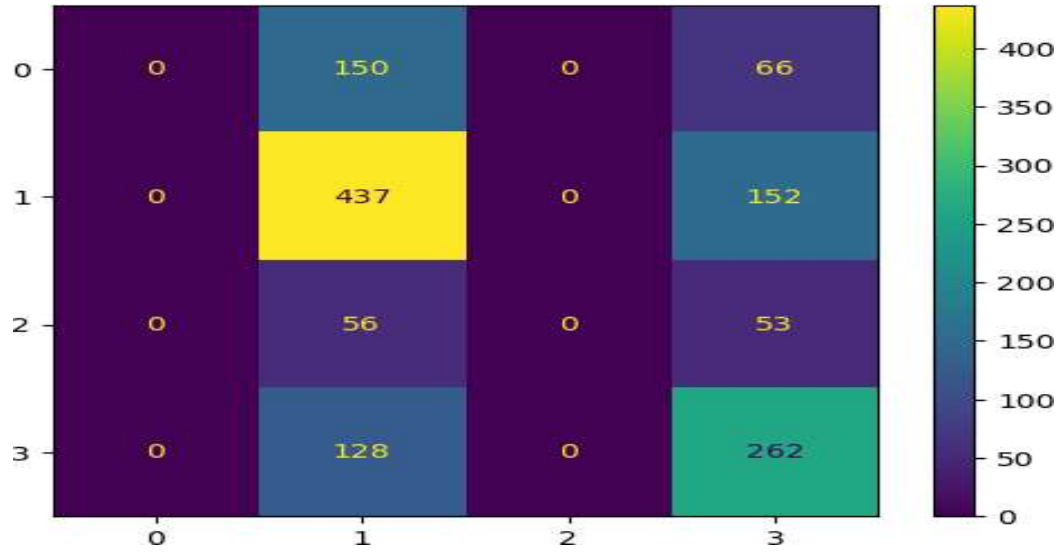
As shown in the above figure, the model with the GaussianNB classifier shows an accuracy of 48% for the location label and 55% for the violent conflict incident types. This shows all three models performed better when predicting the individual labels than predicting both at the same time. Like the above two models, two confusion matrices are built for each class (y1, y2) as visualized in the below figure.

Figure 90: Measuring the performance of the model using the confusion matrix for the first class



The above figure shows the confusion matrix for the first class using the third model built with the GaussianNB classifier, which is the predicted value of location for violent conflicts. The results that are shown on the above matrix are 511 TP, TN and 879 FP, FN values those values are experienced from the total 1390 testing values. Now the confusion matrix for the violent political type class is going to be plotted below.

Figure -91: Measuring the performance of the model using the confusion matrix for the first class



The above figure shows the confusion matrix for the second class, which is the predicted value of the violent political conflict type. As the result shows in the confusion matrix, the model built with GNB classifier resulted in 699 TP, TN values and 691 FP, FN values from the total of 1390 testing values. The final performance metrics applied for this predictive model is the classification report. This method supports multi-class classification by measuring the precision, recall and f1 score of each class individually as presented in the figure below.

Figure -92: Measuring the performance of the model using classification report for the first class

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	57
2.0	0.16	0.18	0.17	239
3.0	0.00	0.00	0.00	40
7.0	0.44	0.63	0.52	560
10.0	0.52	0.52	0.52	115
12.0	0.47	0.33	0.39	165
micro avg	0.39	0.43	0.41	1176
macro avg	0.27	0.28	0.27	1176
weighted avg	0.36	0.43	0.39	1176

According to the classification report plotted above the third model built with the Gaussian NB classifier resulted in an average of 27%, 28% and 27% precision, recall and f1-score respectively. The second classification report is calculated for the second class, which is the violent political conflict type. According to the report, the second class to the model resulted in 26%, 35% and 30% precision, recall and f1-score respectively. The report is presented as follows.

Figure -93: Measuring the performance of the model using the classification report for the second class

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	216
1.0	0.57	0.74	0.64	589
2.0	0.00	0.00	0.00	109
3.0	0.49	0.67	0.57	390
accuracy			0.54	1304
macro avg	0.26	0.35	0.30	1304
weighted avg	0.40	0.54	0.46	1304

5.3. Experimenting with Gradient boosting machine learning

In this section, experiments show Gradient-boosting machine learning algorithm has little better performance while measured using testing accuracy (30%). Using this algorithm, different experiments will be made by changing random states and tuning parameters. The first experiment will be done as follows by changing the random state and the parameter tuning.

Table -26: Experiments by changing the random state of the algorithm

Experiment-1: Performance of the model when random state=0

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier()
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-2: Performance of the model when random state=10

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=10)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-3: Performance of the model when random state=30

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=30)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-4: Performance of the model when random state=42

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

The next experiment is parameter tuning. Two types of parameters have to be tunned in the gradient boosting machine learning algorithm: tree-based parameters and boosting parameters. From the tree-specific parameters max_depth, min_samples_split, min_samples_leaf and max_features are tuned. From the boosting parameters learning_rate, n_estimators and subsample are tuned. For the experimenting a baseline model is built. The baseline model serves as a reference for comparing performance after tuning the model in a machine learning project. This model uses Tree specific parameters (max depth = 3, min samples split =2, min sample leaf =1 and max_features = sqrt) and bosting parameters (learning rate =0.1, n_estimators = 100 and subsamples = 1).

Figure 94– Building a baseline model

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score
GBC = GradientBoostingClassifier(learning_rate=0.1, n_estimators=100,max_depth=3, min_samples_split=2,
                                 min_samples_leaf=1, subsample=1,max_features='sqrt', random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

As shown in the figure above (26), the baseline model performs 48% and 57% for the two classes of the model respectively. The experimentation followed to identify the optimum values for each of the parameters of the model that can improve the model. Table 27 demonstrates the 5-fold cross-validation to identify the optimum values for the parameters in the model.

Table 27: Experiments by changing the parameters of the algorithm

Experiment-1: cross-validation to identify learning rate and n_estimator values

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import GradientBoostingClassifier
p_test1 = {'learning_rate':[0.15,0.1,0.05,0.01,0.005,0.001],
           'n_estimators':[100,250,500,750,1000,1250,1500,1750]}
tuning = GridSearchCV(estimator =GradientBoostingClassifier(max_depth=4, min_samples_split=2,
min_samples_leaf=1, subsample=1,max_features='sqrt', random_state=10),
                      param_grid = p_test1,n_jobs=4, cv=5)
tuning.fit(x_train,y_train)
tuning.cv_results_, tuning.best_params_
```

```
{'learning_rate': 0.005, 'n_estimators': 250}
```

Experiment-2: cross-validation to identify max_depth

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import GradientBoostingClassifier
p_test2 = {'max_depth':[2,3,4,5,6,7] }
tuning = GridSearchCV(estimator =GradientBoostingClassifier(learning_rate=0.005,n_estimators=250,
        min_samples_split=2, min_samples_leaf=1, subsample=1,max_features='sqrt', random_state=10),
            param_grid = p_test2, n_jobs=4, cv=5)
tuning.fit(x_train,y_train)
tuning.cv_results_, tuning.best_params_
```

```
{'max_depth': 2}
```

Experiment-3: cross validation to identify min_samples_leaf and min_samples_split

```
p_test3 = {'min_samples_split':[2,4,6,8,10,20,40,60,100], 'min_samples_leaf':[1,3,5,7,9]}

tuning = GridSearchCV(estimator =GradientBoostingClassifier(learning_rate=0.005, n_estimators=250,
                                    max_depth=2, subsample=1,max_features='sqrt', random_state=10),
            param_grid = p_test3, n_jobs=4, cv=5)
tuning.fit(x_train,y_train)
tuning.cv_results_, tuning.best_params_
```

```
'min_samples_leaf': 1, 'min_samples_split': 2}
```

Experiment-4: cross-validation to identify max_features

```
p_test5 = {'max_features':[2,3,4,5,6,7]}
tuning = GridSearchCV(estimator =GradientBoostingClassifier(learning_rate=0.005, n_estimators=250,
                max_depth=2, min_samples_split=2, min_samples_leaf=1, subsample=1, random_state=10),
param_grid = p_test5,n_jobs=4, cv=5)
tuning.fit(x_train,y_train)
tuning.cv_results_, tuning.best_params_
```

```
{'max_features': 2}
```

Experiment-5: cross-validation to identify sub_samples

```
p_test6= {'subsample':[0.7,0.75,0.8,0.85,0.9,0.95,1]}
tuning = GridSearchCV(estimator =GradientBoostingClassifier(learning_rate=0.005, n_estimators=250,
            max_depth=2, min_samples_split=2, min_samples_leaf=1,max_features=2 , random_state=10),
param_grid = p_test6,n_jobs=4, cv=5)
tuning.fit(x_train,y_train)
tuning.cv_results_, tuning.best_params_
```

```
{'subsample': 0.7}
```

After identifying the optimum parameters for the algorithm, evaluation was conducted using these parameters and compared with the baseline model.

Figure –95: Evaluation of the model using the identified parameter values

```
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score
GBC=GradientBoostingClassifier(learning_rate=0.005, n_estimators=250,max_depth=2, min_samples_split=2,
                               min_samples_leaf=1, subsample=0.7,max_features=2, random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5736196319018405
```

The last experiment applied to the gradient boosting machine learning algorithm is by changing the number of independent values that are ranked during future selection. Totally in the above models, there were 27 independent variables. These variables have been ranked according to their future importance in the above sections. Considering the different values of feature importance, the gradient boosting machine learning algorithm has to be tested by taking the different number of independent variables to get better performance on the model.

Table -28: Experiments by changing the number of independent variables

Experiment-1: Performance of the model using the top 10 independent variables

```
# using the top 10
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-2: Performance of the model using the top 15 independent variables

```
# using the top 15
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-3: Performance of the model using the top 20 independent variables

```
# using the top 20
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

Experiment-4: Performance of the model using all the 27 independent variables

```
# using the top 27
from sklearn.multioutput import MultiOutputClassifier
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(random_state=42)
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(x_test)
AC_Y1= accuracy_score(y_test[:,0],y_pred[:,0])
AC_Y2= accuracy_score(y_test[:,1],y_pred[:,1])
print(AC_Y1,AC_Y2)
```

```
0.4815950920245399 0.5766871165644172
```

5.4. Predicting conflict incident types and incident locations with gradient boosting machine learning algorithm

By using the built predictive model, a prediction was made on new data, which the model has never experienced, to predict the conflict incidents type/category and regions that are at high risk of conflict incidents.

The first step to making a prediction was to have suitable data that can be fitted to the model. The prediction model needs the value of independent variables as an input to make the prediction. The independent values are conflict indicators. Based on the experiments carried out using different numbers of independent variables, the model shows the same performance. Based on the experiment results, the top ten indicators were used to reduce the time for

processing. The indicators selected for prediction purposes are year, economic inequality, group grievance, refugees and IDPs, population, human development, corruption perception index, unemployment, control of corruption and political stability. Using the values of these indicators and the gradient boosting algorithm conflict incident types and locations are predicted for the next five years, 2023- 2028.

To get new data or values of indicators for prediction purposes, the past values of the indicators are used to predicate the values for the next year with a Vector Autoregressive Model (VAR). This model provides insight and uses the relationship between several variables to provide better forecasting results. By using this model, the next five-year values of the indicator will be forecasted.

Figure 94: Predicting indicators values

```
from statsmodels.tsa.vector_ar.var_model import VAR
model = VAR(endog=data)
model_fit = model.fit()
prediction = model_fit.forecast(model_fit.endog,steps=len(data))
```

```
model = VAR(endog=data)
model_fit = model.fit()
pridicted_indicators = model_fit.forecast(model_fit.endog, steps=5)
```

As the above figure shows, by using the VAR method from statsmodels the values for 10 indicators for the next year (2023) have been predicted. After all the information needed to predict the types of incidents and location are ready, the model built by using the gradient boosting algorithm is applied to predict the conflict types and conflict locations.

Figure 95: Building a time series forecast model VAR

```
from sklearn.multioutput import MultiOutputClassifier
from pandas import DataFrame
from sklearn.ensemble import GradientBoostingClassifier
GBC = GradientBoostingClassifier(max_depth=3,max_features=2,min_samples_split=2,n_estimators=2
multi_target_GBC = MultiOutputClassifier(GBC, n_jobs=2)
multi_target_GBC.fit(x_train, y_train)
y_pred = multi_target_GBC.predict(future_pridiction)
```

Finally, the prediction has been made using the above-built predictive model. A detailed description of the analysis and prediction will be presented in the next chapter.

CHAPTER SIX: RESULTS AND DISCUSSION

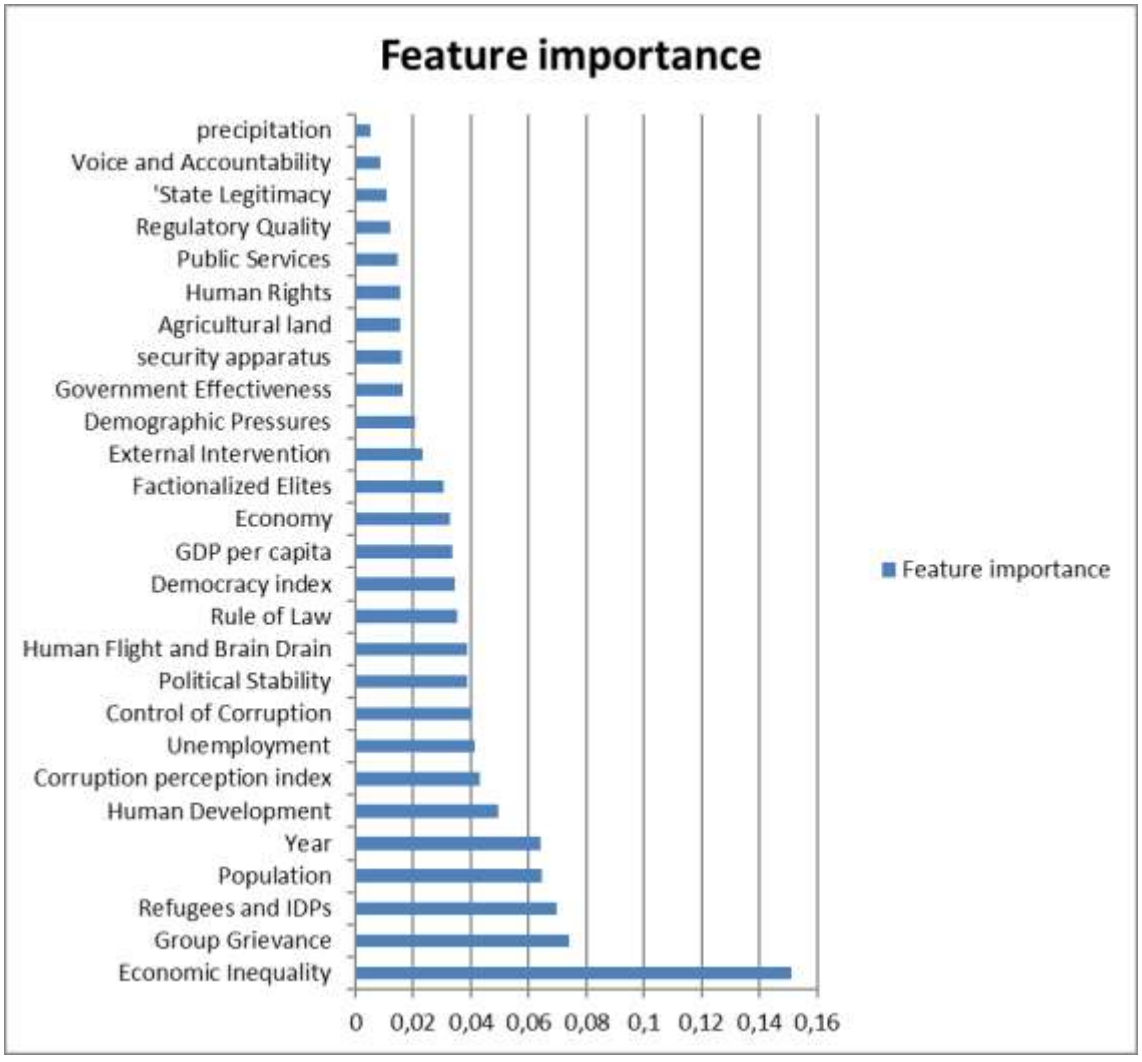
Introduction

This chapter presents the major results of this research. As shown in previous chapters, different analyses have been done to achieve the research objectives. The main important analyses that are performed are feature significance, model building and evaluation and finally predictive analysis to help guide conflict early warning system (CEWS).

6.1. Feature importance

Feature importance guides to select the most significant factors for conflict prediction A high feature importance score means the feature is more significant to the output variable. In this research, the extra tree classifier is applied to the given dataset to identify the columns in the dataset which have the most influence in determining the values of the dependent variables. By employing this approach, the research first computes the feature importance of each indicator used as the independent variable.

Figure -96: List of 25 variables used to predict conflict using the gradient boosting algorithm



Using their feature importance, independent variables are ranked accordingly. The results which are computed show that the feature importance values of each variable are very close to each other. Given the small differences between the feature importance values of indicators, all the independent variables are used for model building. But, after selecting the model that has a better performance it was experimented by using the different number of features to check for model performance change. The overall trend of feature importance values with their variable names is visualized in the above figure.

6.2. Model performance comparisons

This section analyses the results obtained from the evaluation of the performances of classifiers and experiments done in the study during model building and evaluation. In the above section, three different models are built using different machine learning algorithms. After building the models, the performances of the models are tested with different model performance metricises. These are training accuracy, mean squared error, classification report and confusion matrix. In light of the aim of the research, the objective was to identify the machine learning algorithm that shows a good performance. As shown in the table below, Random Forest, Gradient Boosting and Gaussian naïve Bayes are adopted to build a predictive model. By using these algorithms, three different models are built and tested with the performance matrices. The table below presents the summaries of results obtained from comparing the three built models using three different machine learning algorithms.

Table 29: Summaries of the performances of the three models

Machine learning algorithms	**Performance metrics**										
	Training accuracy	**Mean squared error (MSE)**	**Testing accuracy**	**Testing accuracy for individual classes**		**Precision**		**Recall**		**F1-score**	
				Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
Random Forest	30%	9.216	29%	48%	55%	32%	28%	33%	36%	32%	31%
Gradient Boosting	30%	9.216	30%	48%	57%	32%	28%	33%	36%	32%	31%
Gaussian Naive Bayes	22%	9.216	21%	48%	55%	27%	26%	28%	35%	27%	30%

As the table shows, the performances of the models built using the ensemble method such as Random Forest and Gradient Boosting have performed better compared to the Navies Bays Classifier. The results of the ensemble algorithms are the same in most of the metrics, but it showed a small difference in the testing accuracy in which Gradient Boosting has an improved testing accuracy.

The results also show that the predictive performance of models decreases when the number of classes increases from one to two (location and type of conflict incidents). When each class is considered separately, the model shows testing accuracy of 48% and 57% for location and type of conflict, respectively. When both type/category and location of conflict incidents were considered together and concurrently, the predictive performance of the selected

algorithms declined to 30% accuracy. This difference in accuracy has occurred because each class (conflict type and location) has a different number of categories (4 and 13). When the number of classes and classifications increases, the values have more room to move to different categories which makes the number of uncertainty (entropy) increase. When entropy increases the information gain of the model decreases. When the number of categories increases the information gained for the node to split in the algorithm decreases, and the model accuracy decreases at the same time.

The experimental analysis was conducted on Gradient Boosting, a selected model, to acquire a better performance: such as better prediction, minimized computational cost and reduced processing time. Three different experiments are applied to the model: changing the random state, parameter tuning and changing the number of independent variables. The below tables below present the results of each experiment.

Table -30: Experimental results by changing random state

Machine learning algorithms	**Experimental result**							
	Random state = 0		**Random state = 10**		**Random state = 30**		**Random state = 42**	
	Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
Gradient Boosting	48%	57%	48%	57%	48%	55%	48%	57%

Table -31- Experimental results by parameter tuning

Machine learning algorithms	**Baseline model testing accuracy**		**Tuned model testing accuracy**	
	learning_rate=0.1, n_estimators=100, max_depth=3, min_samples_split=2, min_samples_leaf=1, subsample=1, max_features= sqrt , random_state=42		learning_rate=0.005, n_estimators=250, max_depth=2, min_samples_split=2, min_samples_leaf=1, subsample=0.7, max_features=2, random_state=42	
	Y1	Y2	Y1	Y2
Gradient Boosting	48%	57%	48%	48%

Table -32: Experimental Results by changing the number of independent variables

Machine learning algorithms	**Experimental result**							
	Top 10 features		**Top 15 features**		**Top 20 features**		**Top 27 features**	
	Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
Gradient Boosting	48%	57%	48%	57%	48%	57%	48%	57%

The above tables show the experiments performed on the selected machine-learning algorithm. Three different experiments have been applied: These are changing the number of random states, parameter tuning and changing the number of independent variables.

Table 30 shows the results of the first experiment by changing the random state of the machine learning algorithm. The Random state hyperparameter controls the shuffling for train_test_split() due to the random sampling for the train-test split. The model is checked with different random state values 0, 10, 30 and 42 to observe improvements in the models' performance. The results demonstrate that altering the random state did not affect the performance. Keeping the random state constant from one of the above experiments' random state values will keep the model control the shuffling for train test split for future prediction.

Table 31 presents the second experiment: hyperparameter tuning. The hyperparameter tuning controls the learning process of the Gradient Boosting and determines the values of the model parameters which the model learns. First, a baseline model was built for reference purposes. Second, a cross-validation process was applied to the model to identify the optimum values for each parameter of the algorithm. Finally, the model was evaluated with the determined optimum value of parameters such as learning rate (0.005), n_estimators (250), max depth(2), min samples split(2), min samples leaf (1), sub sample (0.7), max features (2), random state(42). By giving the identified optimum values of the parameters, the model was tested for performance change. The experiment shows no change in the accuracy of the prediction. Despite this result, the identified values are considered for future prediction. This was to control the learning process and to determine the values of model parameters in which the Gradient Boosting algorithm learns much faster.

The third experiment was performed by changing the number of independent variables based on their feature importance. Using this approach, variables that have a significant influence can be determined. As shown in Table 32, the experiment using the top 10, 15 and 20 independent values didn’t illustrate any performance change. For this reason, it was preferable to use the small amount of independent variables (N=10), to minimize processing time and computational cost.

Based on the above results future prediction by gradient boosting algorithm will be made using learning rate = 0.005, n_estimators = 250, max depth = 2, min samples split = 2, min samples leaf = 1, sub sample = 0.7, max features = 2, random state = 42 as hyperparameter and the top 10 independent variables. This is aimed at getting the maximum performance of the Gradient-Boosting machine learning algorithm.

6.3. Predicting high-intensity conflict

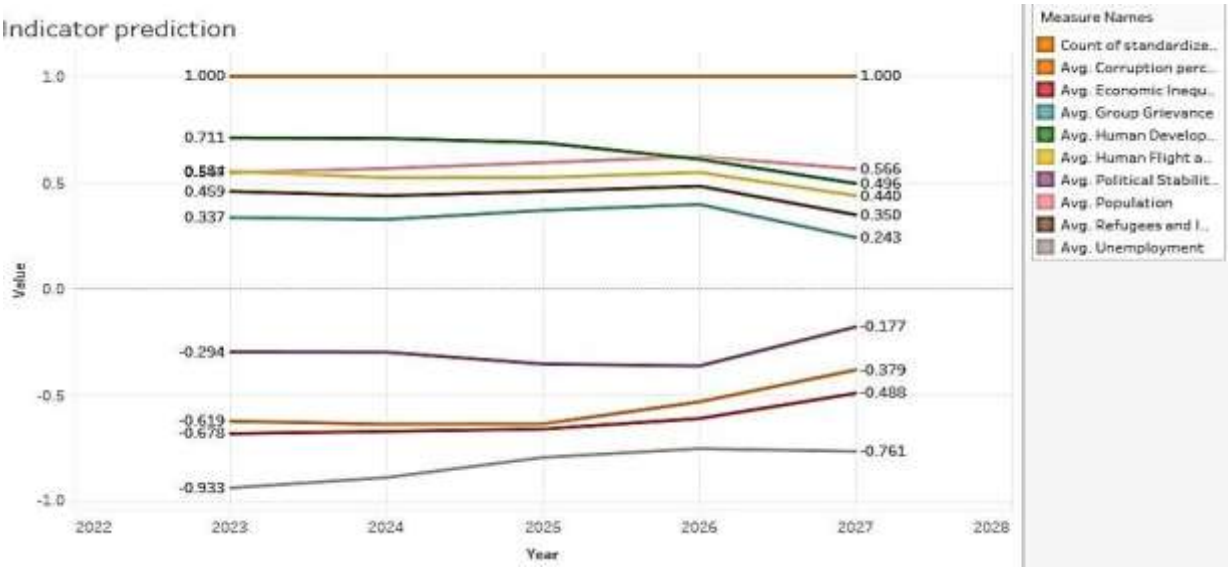
This section presents the findings of violent political conflict prediction for the coming 5 years using a model built by a gradient-boosting machine learning algorithm. For prediction purposes, the independent variables which are the top ten conflict indicators which have a high feature importance are used for the model to make a prediction. But, to predict the future values of these indicators, time series forecasting was made to get the values of the conflict indicators for the next five years (2023 -2028). For time series forecasting, the VAR algorithm was employed. Finally, the following values are acquired.

Table-33: Predicted values of conflict indicators, 2023- 2028

Conflict indicators	**Year**				
	2023	*2024*	*2025*	*2026*	*2027*
Economic Inequality	8.461343	9.863648	11.291701	17.648061	32.943897
Group Grievance	9.535551	9.100000	10.891575	6.535551	-5.514836
Refugees and IDPs	9.100000	-7.350635	1.377894	1.331221	-44.316596
Population	125729742.5	127195064.0	129141188.2	131181923.0	127071373.7
Human Development	0.401576	0.498000	0.260312	0.489000	-0.907813
Corruption perception index	78.947730	56.088695	58.357351	34.000000	470.724155
Unemployment	4.659630	9.831801	19.783581	2.318000	22.760393
Control of Corruption	-0.400000	2.227188	1.518534	-0.480000	4.888399
Political Stability (Estimate)	-2.070000	-1.760000	-7.890930	-1.280000	21.772049
Human Flight and Brain Drain	6.600000	-13.872671	-13.432198	7.000000	-49.069856

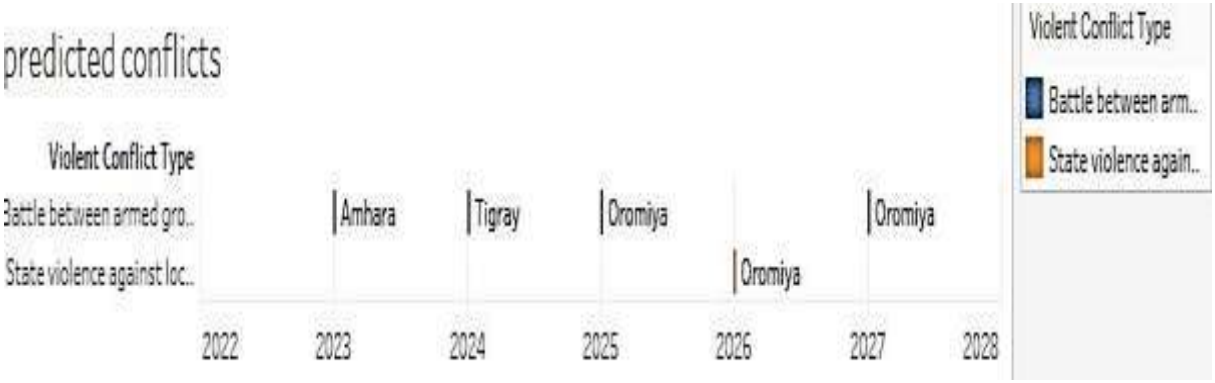
The above table shows the forecasted values for conflict indicators for the coming 5 years. The importance of time series forecasting using VAR algorism in achieving these values is very clear. The aim was by using these forecasted values to predict the conflict location and violent conflict types or categories which are more likely to dominate and occur during the forecasted values period. The line graph below presents the result of the trend of the forecasted indicator values.

Figure-97: The forecasted indicator values



After getting the forecasted indicator values, a prediction was made to see the location or regions with a high risk of experiencing violent political conflict each year. It has to be noted that this model has predicted locations or regions with a risk of high-intensity violent conflict and the most dominant conflict type predicated for each year. In addition, conflict incidents are dependent on the indicators that are forecasted each year. The figure below visualizes the locations and types of incidents predicted for each year. The model considers that the locations will experience more conflicts in contrast to other regions from 2023 through 2028.

Figure -98: The highest conflict incidents predicted,



The above figure shows the model predicated a single most dominant conflict category in each year and a single region that is most likely to experience the dominant conflict category in a particular year. Thus, three different locations face a risk of experiencing a high level of violent conflict incidents according to the model prediction. These locations are the Amhara, Tigray and Oromia regions of Ethiopia. While Amhara and Tigray (from the total 11 regions) regions will face a high risk of violent conflicts in 2023 and 2024 respectively, the Oromia region has a risk of experiencing violent conflicts between 2025 and 2028.

The model has also predicted the single most dominant and high-intensity violent political conflict in Ethiopia, among the four different categories of violent political conflicts. In this regard, the two categories such as state violence against civilians and battle between the state and armed groups, are predicated to be dominant from 2023 through 2028. The battle between armed groups and the state is predicted to be the dominant conflict category in 2023, 2024, 2025 and 2028. State violence against civilians is predicted to be a single high-intensity conflict category in 2026. The model shows a change in the dynamics of the conflict in which after the coming three years the nature of the conflict is predicted to shift to violence against civilians and in turn battles are predicted to emerge.

Looking at the location and types of violent conflict incidents the predication can be summarised as follows:

- ✓ The model predicted that in 2023 high-intensity conflict will likely emerge in the Amhara region and the most dominant conflict type or category is predicated to be a battle between armed groups and the state.
- ✓ The model predicted that in 2024 high-intensity conflict will shift to Tigray and the most dominant conflict type at this location is a battle between armed groups and the state.

- ✓ The prediction indicates that the Oromia region will be exposed to the country's largest violent conflict in 2025. The most common type of conflict in this region is between armed groups and the state.
- ✓ The model predicted that in 2026, the dominant conflict will continue to happen in Oromia and but the most dominant conflict type is likely to shift to state violence against civilians.
- ✓ According to the model prediction, Oromia will likely be the site of Ethiopia's major conflict in 2027, and the war between armed groups and the state will likely be the most common conflict type.

Two major findings and trends can be extracted from analysing figures (Figure 1 and Figure 2) that explain the shift of conflict categories' prominence. First, from 2023 to 2025 economic inequality has the lowest value (-0.678) and from 2025 onwards the value of economic inequality starts to increase. This value change is simultaneous with the shift of violent conflict from the Amhara and Tigray to the Oromia regions. This indicates how the conflict location prediction is dependent on the economic inequality indicator. Second, the shift in conflict types or categories in 2026 from the battle and violence against civilians coincided with an increase in two indicators values such as the corruption perception index and economic inequality. This marks the interdependence between these indicators and the prediction of conflict types or categories.

Another important finding that needs to be discussed is the issue of the performance of the model. As noted in the previous chapter related to model building, the performance of the model shows a huge difference when predicting both locations of incidents and types of incidents at the same time and when they are predicated separately or interdependently. The model has shown a better performance while predicting the two classes independently without considering the other predicted value. The following are the main findings: First when both location and conflict categories are concurrently considered for prediction, the model has a performance of 30% testing accuracy. Second, when measuring the performance of the model on the individual classes, the model accuracy to predict location is 48% and for conflict incidents types is 57%. The research, however, used the latter approach. This has resulted, as a conclusion, despite the shifts in the location and types of violent political conflicts, the gradient boosting machine learning algorithm model prediction shows that Ethiopia will remain at risk of violent conflicts. Among other regions, Amhara, Tigray and Oromia will continue to face risks of conflict.

CHAPTER SEVEN: CONCLUSION AND RECOMMENDATIONS

This research is informed by the idea that there is a research gap in applying data analytics models and tools in the context of Ethiopia's violent conflict prediction research. This is further supported by the fact that there is an increased incidence of violent conflicts across countries and societies. This needs to initiate academic interest and the moral obligation to understand and predict violent conflicts that lead to significant economic, social and humanitarian consequences. This research aims to develop a predictive model using a supervised machine learning algorithm that can best forecast violent political conflicts in Ethiopia in terms of the dominant conflict types/categories and regions at risk of conflict incidents. Specifically, it seeks to assess Gaussian naive Bayes' predictive performance compared with ensemble algorithms such as Random Forest and Gradient Boosting. It also seeks to draw the most salient drivers of violent conflicts based on their variable importance. Finally, it seeks to forecast violent political conflicts in Ethiopia for each of the coming 5 years, from 2023 to 2028, using a best-performed machine learning algorithm and tools. Methodologically, this research has employed an experimental research design and adopted the CRISP-ML framework. The research combined past and recent violent conflict data with political, economic, social and environmental data. Data on both independent and dependent variables were collected from open sources databases from international organisations and research institutes such as World Bank, UNDP, etc. Predictive analytics tools, as well as three algorithms (random forest, gradient boosting and Gaussian naive Bayes), are used. Open-source software called Jupyter Notebook is used for analysis.

Three main conclusions can be drawn from the research. First, the finding shows that using the extra tree classifier algorithm, the ten most salient drivers of violent conflicts in Ethiopia are the following with their feature importance value: Economic Inequality (0.15), Group Grievance (0.07), Refugees and IDPs (0.06), Population (0.064), Human Development (0.049), Corruption perception index (0.043), Unemployment (0.041), control of corruption (0.04), political stability (0.038) and human flight and brain drain (0.038). Their importance is also supported by the newly forecasted values using the GNB algorithm. This suggests the complexity and diversity of drivers of violent conflicts in Ethiopia mainly along economic, political and social factors.

Second, the Gradient Boosting machine learning algorithm has a better performance than Random Forest and Gaussian naïve Bayes in predicting the location and types/categories of violent conflicts individual classes. In predicting types of violent conflicts, while Gradient Boosting has a testing accuracy of 57%, both the Random Forest and Gaussian naïve Bayes has 55%. Yet, in terms of predicting the location of incidents of conflict, all have 48% testing accuracy. However, when both type/category and location of conflict incidents were considered together the predictive performance of the selected algorithms declined to 30% accuracy. This difference in accuracy has occurred because each class (conflict type and location) has a different number of categories (4 and 13). When the number of classes and classifications increases, the values have more room to move to different categories which makes the number of uncertainty (entropy) increase. When entropy increases the information gain of the model decreases. When the number of categories increases the information gained for the node to

split in the algorithm decreases, and the model accuracy decreases at the same time. The research argues that the performance of the conflict-predicting model depends on whether the classes (type and locations) are merged and predicated concurrently or when they are predicated separately and independently.

The third main finding shows continuity in violent political conflicts for the coming five years, 2023- 2028. The Gradient Boosting machine learning algorithm model predicts that Ethiopia will continue to be at risk of violent conflicts. The location and type of violent political conflicts shift yearly. Amhara, Tigray, and Oromia are expected to face dangers of conflicts, more than other areas. While Amhara and Tigray regions will face a high risk of violent conflicts in 2023 and 2024 respectively, the Oromia region has a risk of experiencing violent conflicts between 2025 and 2028. State violence against civilians and battles between the state and armed groups are predicated to be dominant from 2023 through 2028. The battle between armed groups and the state is predicted dominant conflict category in 2023, 2024, 2025 and 2028. State violence against civilians is predicted to be a high-intensity conflict category in 2026. The model shows a change in the dynamics of the conflict in which after three years the nature of the conflict is predicted to shift to violence against civilians and in the final year battles are predicted to emerge.

This research has policy and academic implications and importance. The following two main issues are put forward as a recommendation for future work. First, violent political conflicts are dynamic and unstable; they shift in terms of time, intensity, nature and actors involved. this provides important input for policymakers in their analysis and decision-making. This suggests that decision-makers and peacebuilders need to emphasise and engage with data analytics as a research technique and make use of its outputs to address violent conflicts adequately. Second, this research shows the importance of using different machine learning algorithm to select a better performing model. This suggests that future researchers working on conflict prediction using data analytics algorithms need to use, compare and contrast different algorithms (in addition to the one used in this research) to build conflict prediction models and provide reliable output.

References

Acharjya, D and Kauser, A. (2016). "A Survey on Big Data Analytics: Challenges, Open Research Issuesand Tools", *International Journal of Advanced Computer Science & Applications,* 7(2), 1-8.

Alemayehu, G., and Befekadu, D. (2005). "Conflict, Post-Conflict and Economic Performance in Ethiopia". *International Economic Association,* 125-126.

Amy, A. C., Ronald, I., & Christian, W. (2011). "Measuring effective democracy". *International Political Science Review,* 43-49.

Anusua, T., et al. (2019). "Using AI for Everyday Armed Conflict Analysis". *The Carter Centre,* 3-4

AU. (2019). Conflict Prevention and Early Warning. *Division of the AU Peace and Security Department.*

Bandara, et al., (2020). "Forecasting across time series databases using recurrent neural networks on groups of similar series", *Elsevier,* 140, 10-20.

Bazzi S., Blair R.A., Blattman C., Dube O., Gudgeon M., Peck R. (2022). "The promise and pitfalls of conflict prediction: evidence from Colombia and Indonesia", *NBER Working Paper,* 1-5.

Berihu Asgele Siyum(2021)." Underlying Causes of Conflict in Ethiopia: Historical, Political, and Institutional?". World Conference on social studies, Hungary, 13-15.

Blum, A., Hopcroft J., and Kannan, R. (2018). *Foundations of Data Science*. Cambridge University: London.

Chris, Perry (2019). "Machine Learning and Conflict Prediction: A Use Case". *International Journal of Security & Development,* 2(3), 7-9.

Christy, L. (2017). "Improving Conflict Early Warning Systems for United Nations Peacekeeping". *Masters thesis*

Cocodia, J., 2008, 'Exhuming Trends in Ethnic Conflict and Cooperation in Africa: Some Selected States', *African Journal on Conflict Resolution,* vol. 8, no. 3, pp. 9-26

Randahl and Vegelius (2022). "Predicting escalating and de-escalating violence in Africa using Markov models". *International Interactions,* 48(4), 597-613.

Daniel , K., Aart, K., & Pablo, Z.-L. (2001). Governance Matters. *Policy research working paper,* 40-42.

David Randahl & Johan Vegelius (2022). "Predicting escalating and de-escalating violence in Africa using Markov models". *International Interactions,* 48:4, 597-613.

Donald, B.E., and Ahmed, A. (2015). "Predictive Analytics". *IEEE Computer Society*,1-3.

Elizabeth et al. (2020). "Good Governance and the World Bank". *University of Oxford,* 4-12.

Ettensperger, Felix (2022). “Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging with multiple tree-based algorithms and variance-promoting data configurations”. *International Interactions,* 48 (4), 555-578.

Ettensperger F. (2020). "Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field"., *Qual Quant*, 54, 567–601.

Frère, M.S. and Wilen, N. (2015). “INFOCORE (Definitions: “Violent (conflict”. Bruxelles: &ULB”., 5-10.

Fulvio Attinà, Marcello Carammia & Stefano M. Iacus (2022) Forecasting change in conflict fatalities with dynamic elastic net, International Interactions, 48:4, 649-677.

Fund for peace. (2017). *Fragile state index methodologies.* Washington, D.C: the fund for peace.

Gaël Varoquaux, and Olivier Colliot (2019). “Evaluating machine learning models and their diagnostic value. In Olivier Colliot, *Machine Learning for Brain Disorders*, Springer, in Press. PP. 3-10.

General Authority for Statistics. (2016). *Gross Domestic Product Per Capita.* National income statistics.

Hannes Mueller & Christopher Rauh (2022) Using past violence and current news to predict changes in violence, International Interactions, 48:4, 579-596.

Havard Hegre et al. (2021). “Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand Up to Reality”. *International Studies Quarterly*, 1, 1–9.

Håvard Hegrea et al.(2019). “ViEWS: A political violence early-warning system. Peace research”, 56(2), 156.

Håvard Hegre, Paola Vesco & Michael Colaresi (2022). Lessons from an escalation prediction competition, International Interactions, 48:4, 521-554.

Havelange, Corentin (2021). “Predicting armed conflicts: a machine learning approach”. *Masters in Mathematical Engineering*, 50-67.

Helle, V., Negus, A.S., and Nyberg, J. (2018). “Improving armed conflict prediction using machine learning”. *International Interaction,* 48 (4), 20-25.

Herbert, W., & Tobias, D. (2009). “Conflict early warning and response”. *Crisis states* research centre, 2(49), 30-35.

Herbert, S. (2017). “Conflict analysis: Topic guide*”. Birmingham, UK: GSDRC, University of Birmingham*.

Ho and Chong. (2010). “Exploratory data analysis in the context of data mining and resampling”. *International Journal of Psychological Research*, 3(1), 9-22.

Human Development Report. (2022). *Human Development Report.* New York: United Nations Development Programme.

Igwenagu, Chinelo (2016). *Fundamentals of Research Methodology and Data Collection.* LAP LAMBERT Academic Publishing: Germany

Ines, A. R. (2015). "Defining and measuring state fragility: a new proposal". *The Annual Bank Conference on Africa,* (pp. 7-9). Berkeley.

Iris Malone (2022). "Recurrent neural networks for conflict forecasting". *International Interactions,* 48 (4), 614-632.

Ishfaque, M. et al (2022). "Use of Recurrent Neural Network with Long Short-Term Memory for Seepage Prediction at Tarbela Dam", *Energies*, 15(3123), 5-15.

Javier , F. M., & Sebastian , Z. (2009). "Measuring Fragility". *United Nations Development Program*

Kamila, Pawluszek, F., and Andrzej, B. (2020). "On the Importance of Train–Test Split Ratio". *Remote Sensing, 12*(18), 50-375.

Kamiran, F. and Calders, T. (2011). "Data preprocessing techniques for classification without discrimination". *SpringerLink,* 1-33.

Kaufmann, D. (2010). "The Worldwide Governance Indicators: Methodology and Analytical Issues". *the world bank,* 1-10.

Kotsiantis, B.S., Kanellopoulos, D. and Pinte, E.P. (2006). "Data Preprocessing for Supervised Learning". *International Journal of Computer Science*, 1(12),1-6.

Kroese, D.P., et al (2019). *Data Science and Machine Learning: Mathematical and Statistical Methods (1st ed.).* Chapman and Hall/CRC: New York.

Lake, D. A. and Rothchild, D., (1996), "Containing Fear: The Origins and Management of Ethnic Conflict", *International Security*, vol. 21, no. 2, pp. 41-75

Laza, K. (2019). "The Economist Intelligence Unit's index of democracy". *The Economist Intelligence Unit,* 2-5.

Lazcano, A., Herrera, P.J. and Monge, M. A. (2023). "Combined Model Based on Recurrent Neural Networks and Graph Convolutional Networks for Financial Time Series Forecasting". *Mathematics*, 11(224), 2-21.

Li J., Chen R. (2022). "A Distributed Task Scheduling Method Based on Conflict Prediction for Ad Hoc UAV Swarms", the *National University of Defence Technology*, 1-3.

Lindholm A., Hendriks J., Wills A., Schön T.B. (2022), "Predicting political violence using a state-space model", *Taylor & Francis Journals*, 48(4),759-777.

Linke A.M., Witmer F.D.W., O'Loughlin J. (2022). "Weather variability and conflict forecasts: Dynamic human-environment interactions in Kenya". *Political Geography*, 92, 1- 12.

M., and Goel, N. (2022). "Predictive Analytics: A Study of its Advantages and Applications". *International Research Journa*l, 12(1), 4-6.

Mark, C. (2016). "*Violent Conflicts in Africa: Towards a Holistic Understanding.* Swaziland", *World Journal of Social Science Research*.

Matthieu, K., et al (2016). "Exploratory Data Analysis". *Secondary Analysis of Electronic Health Records* [Internet],185-200.

McAlexander, Richard J. and Mentch Lucas (2020). "Predictive inference with random forests: A new perspective on classical analyses", *Research & Politics,* 7(1), 5-15.

Milorad, k. (2019). "human development index: concepts and Measurements*". New York: Human Development Report Office*.

Mimansha, P., and Nitinl, P. (2019). "Exploring Research Methodology". *International Journal of Research & Review,* 6(3), 45-55.

Musumba, Mark, Naureen Fatema, and Shahriar Kibriya. (2021). "Prevention Is Better Than Cure: Machine Learning Approach to Conflict Prediction in Sub-Saharan Africa" *Sustainability* 13 (13),1-18.

Natalie, F. (2021). "Fragile States Index Annual Report*".* Washington D.C: The Fund for Peace.

Neumann, S., Ahner, D. and Hill, R.R. (2022), "Forecasting country conflict using statistical learning methods", *Journal of Defense Analytics and Logistics*, Vol. 6 No. 1, 59-72. Özer, C. (2019). "A Research on Machine Learning Methods and Its Applications". *Journal of Educational Technology & Online Learning, Turkey,*1(3), 25-40.

Park, J., Dokkyun, Yi and Sangmin, Ji (2020). "Analysis of Recurrent Neural Network and Predictions". *Symmetry,* 12(615), 3-10.

Paul , S., & Kirsten, B. (2008). "The Everyday Democracy Index". 12-25.

Perry, C. (2013). "Machine Learning and Conflict Prediction: A Use Case". *International Journal of Security & Development*, 2(3),1-18.

Powers, D.M.W. (2011). "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation". *Journal of Machine Learning Technologies, South Australia*, 2(1), 37-45.

Racz, A., Bajusz, D., and Heberger, K. (2021). "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification". *Molecules,* 26 (1111), 13-16.

Rinaldi S, Gragnani A, Moro FN, Della Rossa F. (2022). "A theoretical analysis of complex armed conflicts". *PLoS ONE*. 17(3)

Sam, G. (2019). "Research Methodology and Research Method". *Open Journal of Business and Management,* (9)4, 30-43.

Sebastian von Einsiedel (2014). "Major Recent Trends in Violent Conflict". *United Nations University Centre for Policy Research Occasional Paper*.

Sharma, K., Bhanu Tokas, T., & Leo, A. (2021). "*Deep learning in big data and data mining".* (5-9).

Shallcross N.J., Ahner D.K. (2020), "Predictive models of world conflict: accounting for regional and conflict-state differences", The *Journal of Defense Modeling and Simulation*, 17 (3), 243-267

Siân, H. (2017). "Conflict Analysis: Topic Guide". GSDRC, University of Birmingham. 12-19

Si Chen, Xiaodong Cai, Bo Li, and Zhenzhen Hou. (2019). "Community conflict prediction method based on spliced BiLSTM", International Conference on Image and Video Processing, and Artificial Intelligence, 113211V.

Smola, A. and Vishwanathan, S.V.N. (2008). Introduction to Machine Learning. Cambridge University Press: United Kingdom.

Stefan, S., et al. (2021). "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology". Machine Learning and Knowledge Extraction, 1, 5-10.

UNDP. (2011). "The Human Development Index (HDI)". UNDP Human Development Report Office.

Valeria, H., Andra-Stefania, N., and Jakob, N. (2018). "Improving armed conflict prediction using machine learning". 22-27.

Vrigazova, B. (2021). "The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems". Business Systems Research, 12 (1), 228-242.

Wallensteen, Peter and K. Axell. (1994). "Conflict Resolution and the End of the Cold". Journal of Peace Research 31 (3). 333-349.

Wei, J. (2020). "Research on Machine Learning and Its Algorithms and Development". Journal of Physics: Conference Series, 3-6.

William Robert Avis (2019). "Current trends in violent, evidence and earning for development", UK Department for International Development., 2-5.

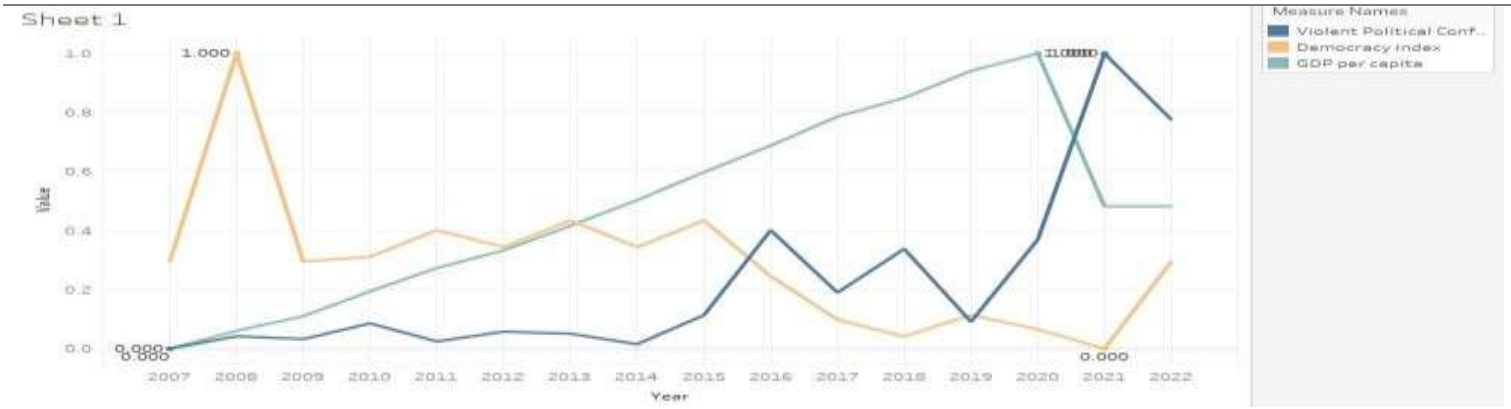
Zhang, G. Peter and Qi, Min (2005). "Neural network forecasting for seasonal and trend time series", European Journal of Operational Research, 160(2), 501-514.

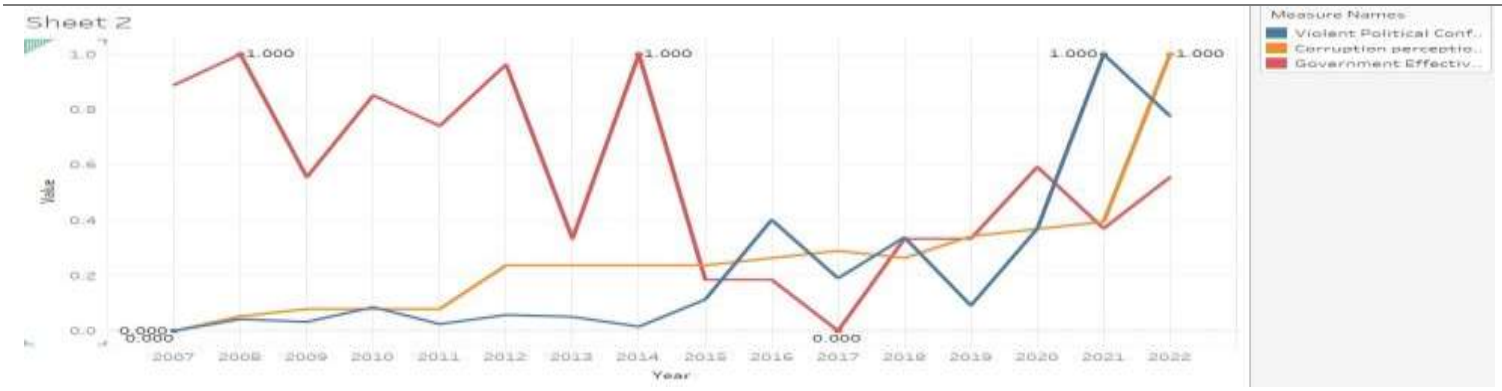
Zheng, A. (2015). Evaluating Machine Learning Models. O'Reilly Media: United States of America.

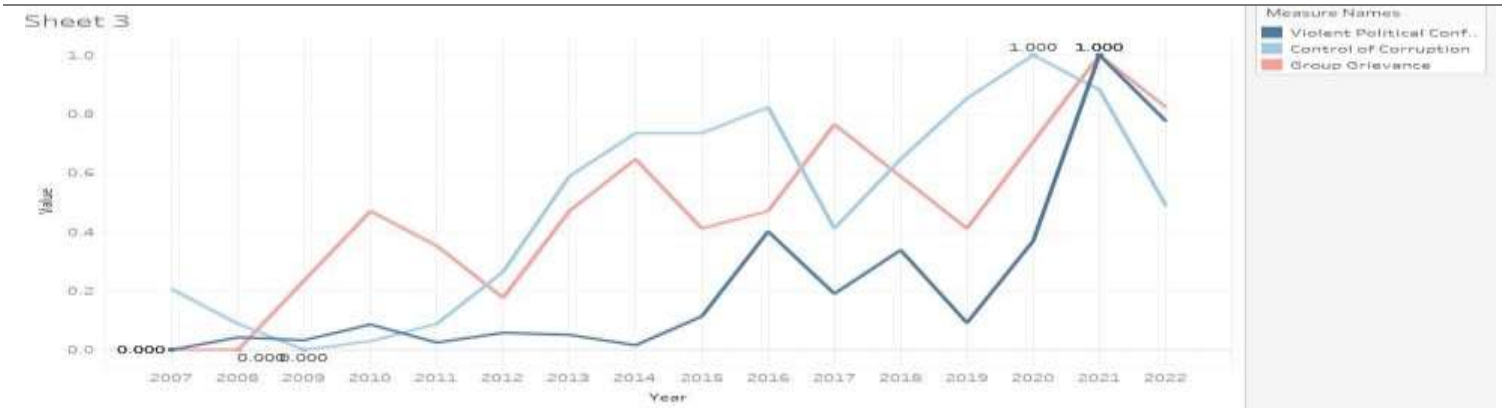
Zhang S., Abdel-Aty M. (2022). "Real-Time Pedestrian Conflict Prediction Model at the Signal Cycle Level Using Machine Learning Models," open journal of intelligent transportation systems, 5-10.

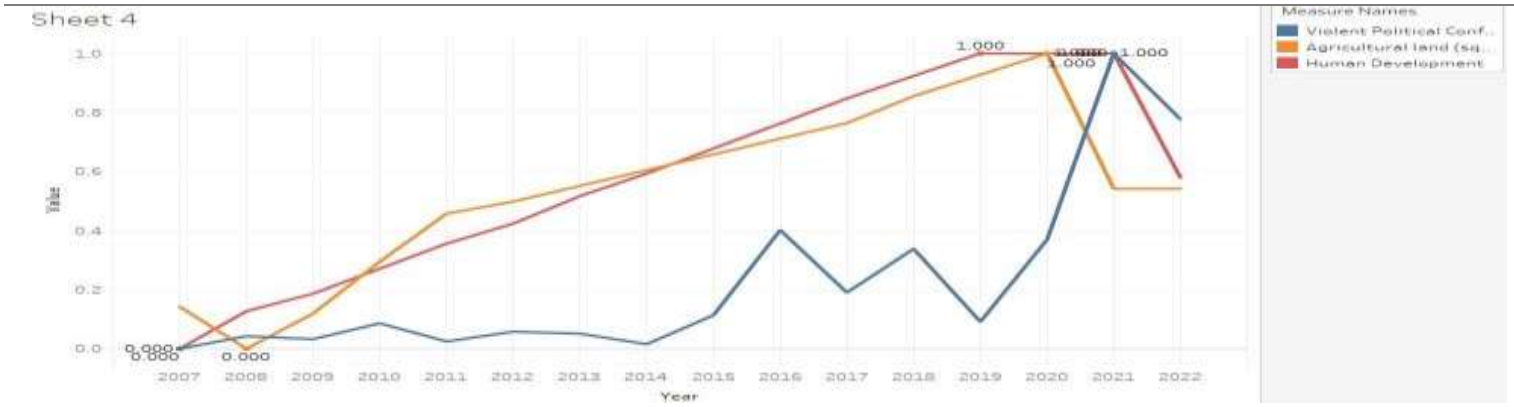
Appendix

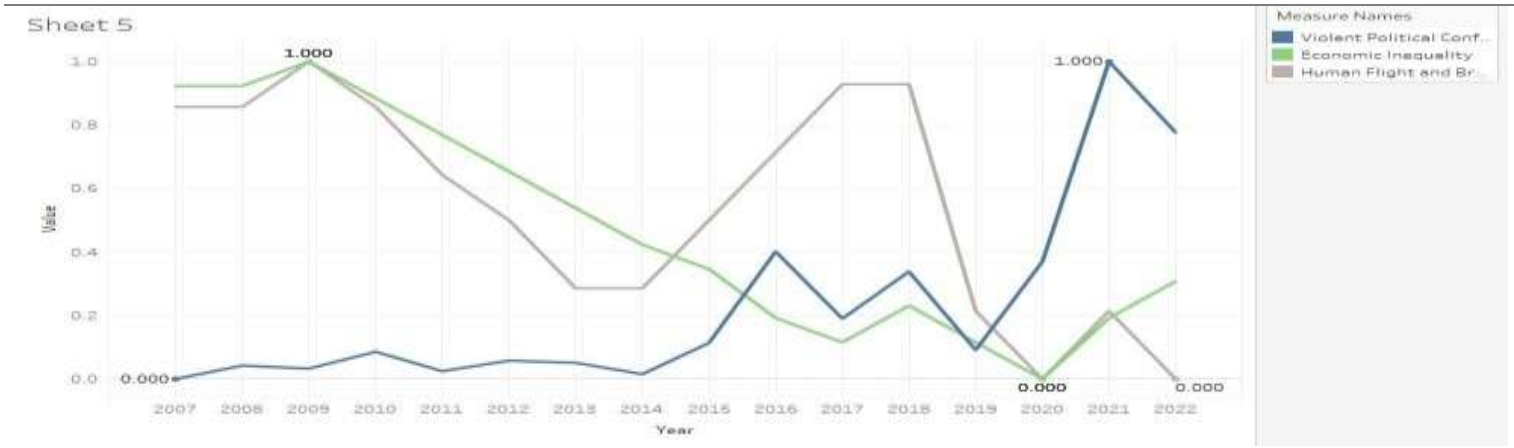
Table 34: visualizing the variables in a final merged conflict data

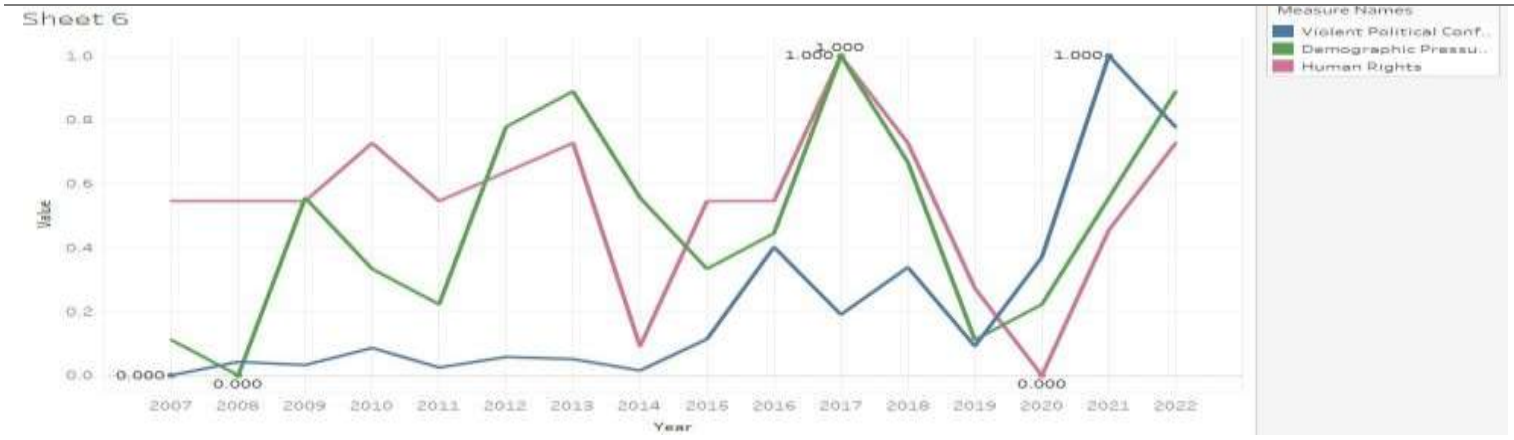


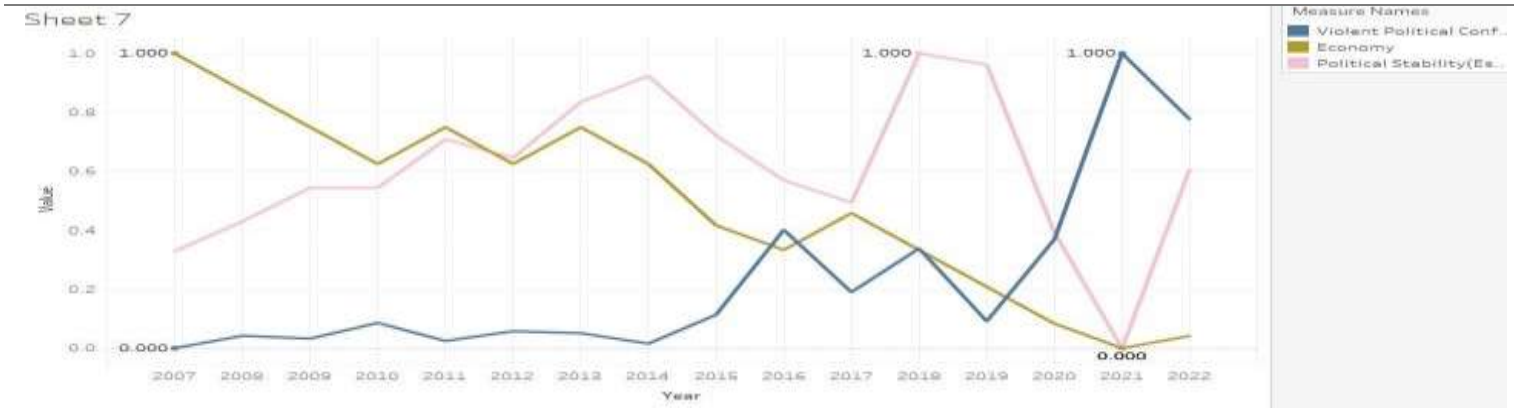


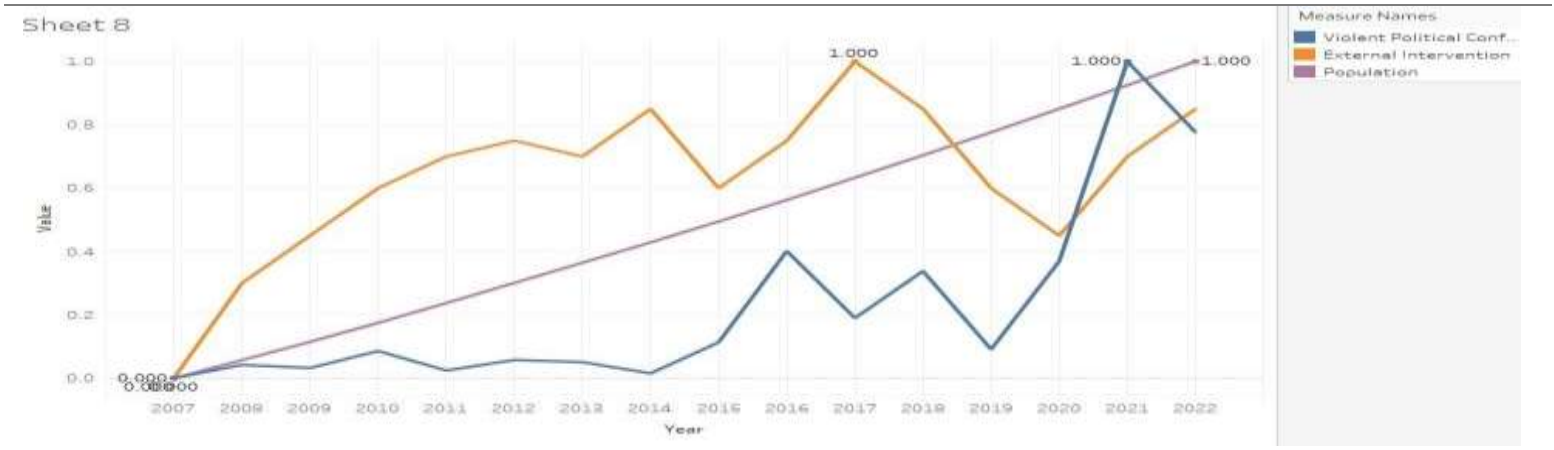


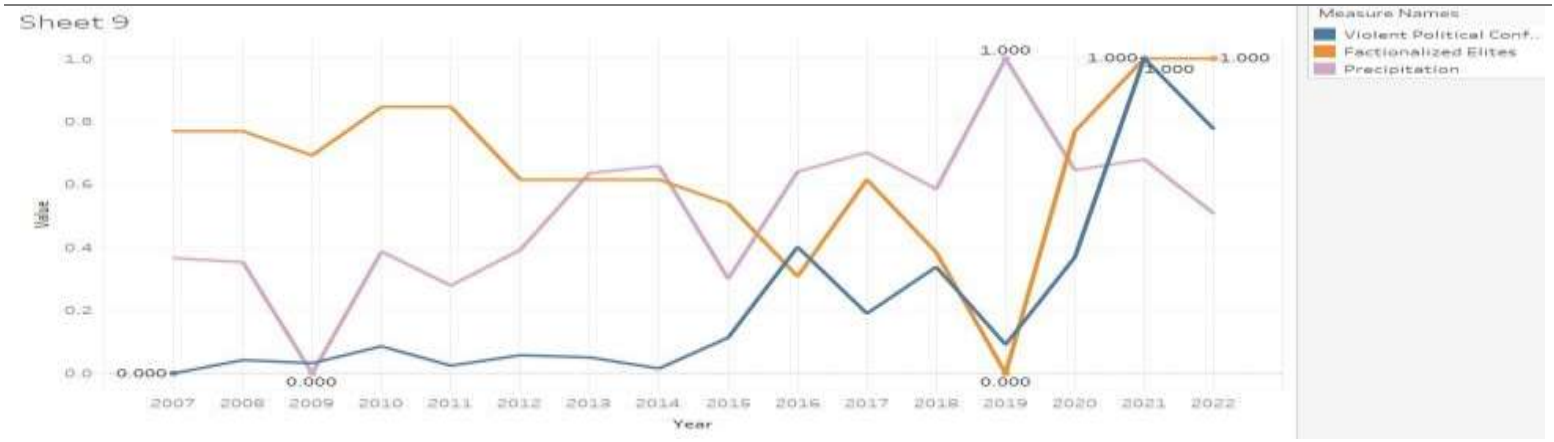

Sheet 4
Measure Names
Violent Political Conf..
Agricultural land (sq..
Human Development
Value
Year
1.000
0.000

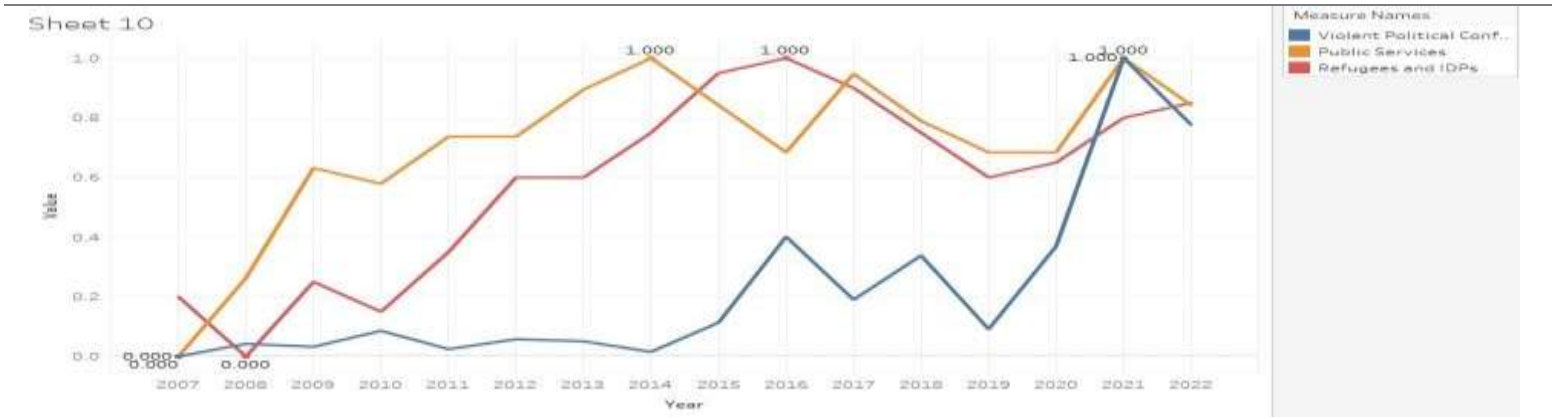

Sheet 5
Measure Names
Violent Political Conf..
Economic Inequality
Human Flight and Br..
Value
Year
1.000
0.000

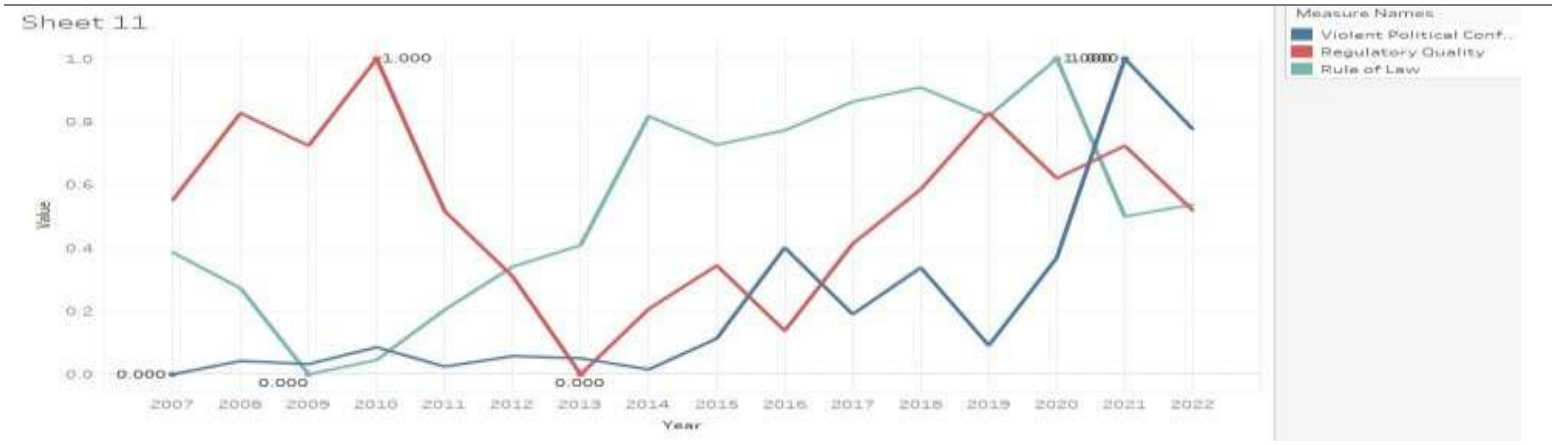

Sheet 6
Measure Names
Violent Political Conf..
Demographic Pressu..
Human Rights
Value
Year
1.000
0.000

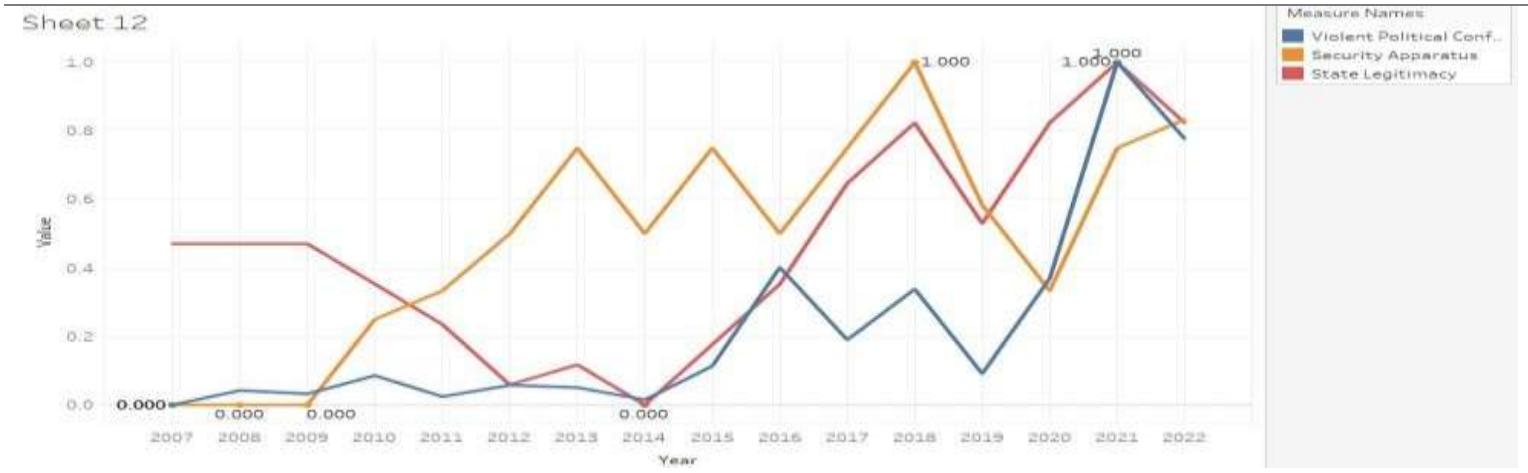

Sheet 7
Measure Names
Violent Political Conf..
Economy
Political Stability(Es..
Value
Year
1.000
0.000

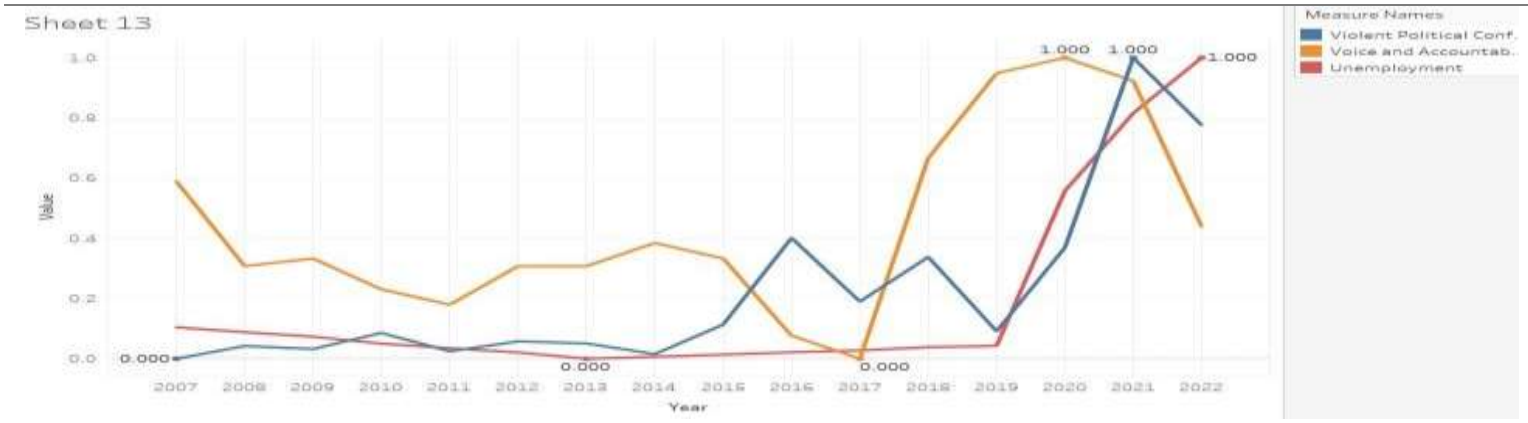

Sheet 8
Measure Names
Violent Political Conf..
External Intervention
Population
Value
Year
1.000
0.000


Sheet 9
Measure Names
Violent Political Conf..
Factionalized Elites
Precipitation
Value
Year
1.000
0.000


Sheet 10
Measure Names
Violent Political Conf..
Public Services
Refugees and IDPs
Value
Year
1.000
0.000


Sheet 11
Measure Names
Violent Political Conf..
Regulatory Quality
Rule of Law
Value
Year
1.000
0.000


Sheet 12
Measure Names
Violent Political Conf..
Security Apparatus
State Legitimacy
Value
Year
1.000
0.000


Sheet 13
Measure Names
Violent Political Conf.
Voice and Accountab..
Unemployment
Value
Year
1.000
0.000