UNSUPERVISED APPROACHES FOR DETECTING AND FORECASTING FRAGILE COUNTRIES

Rachit Dev

Mid-Thesis Report (MS Data Science)

November 2020

# 

# **TABLE OF CONTENTS**

DEDICATION v

ACKNOWLEDGEMENTS vi

ABSTRACT vii

LIST OF FIGURES xv

LIST OF ABBREVIATIONS xvi

CHAPTER 1: INTRODUCTION 9

1.1 Background of the Study 9

1.2 Problem Statement 10

1.3 Aim and Objectives 11

1.4 Research Questions (IF ANY) 11

1.5 Scope of the Study 11

1.6 Significance of the Study 11

1.7 Structure of the Study 12

CHAPTER 2: LITERATURE REVIEW 13

2.1 Introduction 13

2.2.1 K-Means Clustering 13

2.2.2 Hierarchical Clustering 13

2.2.3 Time Series with Long Short-Term Memory (LSTM) 14

2.2.4 Why LSTM and not ARIMA (Time-Series)? 14

2.3 Discussion 14

2.4 Summary 14

CHAPTER 3: RESEARCH METHODOLOGY 15

3.1 Introduction 15

3.1.1 Data Explanation 15

3.1.1.1 Cohesion Indicators 16

3.1.1.2 Economic Indicators 17

3.1.1.3 Political Indicators 18

3.1.1.4 Social Indicators 20

3.1.1.5 Cross Cutting Indicators 20

3.1.2 Tools Required 21

3.1.3 Data Preparation 21

3.1.4 K-Means 22

3.1.5 Hierarchical Clustering 23

3.1.6 Time Series 24

3.1.7 Long Short-Term Memory (LSTM) 25

3.2 Novel Methodology of K-Means and Hierarchical Clustering 27

3.2.1 Data Preparation (K-Means and Hierarchical) 27

3.2.2 Validation Data Preparation 27

3.2.3 K-Means Data Modelling 27

3.2.4 K-Means Validation and Confusion Matrix 28

3.2.5 Hierarchical Approach 28

3.2.6 Hierarchical Approach Validation and Confusion Matrix 28

3.3 Novel Methodology for Time Series with LSTM 29

3.3.1 Time Series Data Preparation 29

3.3.2 LSTM Approach 30

3.3.3 Plotting and Observing Forecasted Trend 30

3.4 Expected Outcome 30

3.5 Summary 30

REFERENCES 31

APPENDIX A: RESEARCH PLAN 35

APPENDIX B: RESEARCH PROPOSAL 36

**DEDICATION**

To my brother Parichit for forcing me to study further, wife Shobhna for financing it, letting me study and to all my friends who believed in me.

**ACKNOWLEDGMENTS**

I would like to thank my supervisor Dr. Manoj Jayabalan and Mrs. Aayushi Verma for their consistent support and guidance during the running of this research. Furthermore, I would like to thank the Fund for Peace team for providing the fragile state index data. I would also like to acknowledge the UpGrad team for providing a great platform for the working professionals to study further.

**ABSTRACT**

We live in a world where we are facing conflict / riot news occurring in all the parts of the world. Though the impact is minor in majority of conflicts but we cannot get away with the major conflicts that occur in the certain parts of the world. There are many factors which contribute towards the occurrence of the conflict and we should try to predict the instability of the country by using machine learning tools and the relevant data. Machine Learning enables us to notice movements inside a country and counter with the right economic, political, and developmental authorizations by the government of the respective country and avoid clash or total governmental breakdown. Our inspiration is to capture and infer these movements on an impressive scale and construct a model that can show the fragility of a country. There are millions of people who lost their lives due to these conflicts and by predicting them we can raise the alarm to the particular authorities or the citizens and their lives might be saved. We hope that by applying machine learning techniques we can predict the conflicts which might occur in the countries that are prone to it.

# 

# **LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure Number | Description | Page Number |
| Figure 3.1 | Cohesion Indicators | 15 |
| Figure 3.2 | Economic Indicators | 16 |
| Figure 3.3 | Political Indicators | 18 |
| Figure 3.4 | Social and Cross-Cutting Indicators | 19 |
| Figure 3.5 | K-Means Algorithm Flowchart | 22 |
| Figure 3.6 | Hierarchical Clustering Flowchart | 23 |
| Figure 3.7 | Time Series Forecasting Flowchart | 24 |
| Figure 3.8 | LSTM Network | 25 |
| Figure 3.9 | Forget Gate | 26 |
| Figure 3.10 | Input Gate | 26 |
| Figure 3.11 | Output Gate | 26 |
| Figure 3.12 | Data Preparation and Pre-Processing | 27 |
| Figure 3.13 | Validation Data Preparation | 27 |
| Figure 3.14 | Determination of Number of Clusters | 27 |
| Figure 3.15 | Validation of K-Means Model | 28 |
| Figure 3.16 | Hierarchical Clustering | 28 |
| Figure 3.17 | Validation of Hierarchical Model | 29 |
| Figure 3.18 | Block Diagram to generate datasets based on countries | 29 |
| Figure 3.19 | Block Diagram to extract country-wise dataset | 29 |
| Figure 3.20 | Block Diagram of LSTM Forecasting | 30 |
|  |  |  |

# 

# **LIST OF ABBREVIATIONS**

FFP…………....... Fund for Peace

FSI…................... Fragile State Index

ARIMA………… Autoregressive, Integrated, Moving Averages

LSTM…………... Long Short-Term Memory

CI-CD…………... Continuous Integration, Continuous Development

**CHAPTER 1**

**INTRODUCTION**

In an exceptionally interconnected world, pressures on one fragile state can have grave repercussions for that state and its kin, yet additionally for its neighbors and different states most of the way over the globe. Since the end of the Cold War, various states have exploded into mass brutality coming from internal clash. A portion of these emergencies rise up out of ethnic strains; some are civil wars; others assume the type of revolutions; and many ends up in complex humanitarian crises.

Separation points can arise between different groups, characterized by language, religion, race, ethnicity, nationality, class, caste, tribe or territory of origin. Tensions can result into struggle through an assortment of conditions, for example, rivalry over assets, predatory or fractured leadership, corruption, or unresolved group complaints. The explanations behind state fragility are intricate yet not unpredictable. It is fundamentally significant that the worldwide community comprehend and intently monitor the conditions that add to fragility — and be prepared to make the essential moves to manage the hidden issues or in any case mitigate the damaging effects.

To have important early warning signs, and effective policy reactions, evaluations must go past specific field knowledge, narrative case studies and anecdotal evidence to recognize and grasp on expansive social patterns. A blended methodology coordinating qualitative and quantitative information sources is expected to set up examples and patterns. With the correct information and investigation, it is conceivable to distinguish issues that might be stewing beneath the surface. Leaders need admittance to this sort of information to execute successful strategies.

* 1. **Background of the Study**

Internal wars have not only become more persistent, but also more commonplace (Blattman and Miguel, 2010). It is not surprising, therefore, that there have been numerous studies and analyses of conflict across both the political science and economics disciplines that try to understand conflict. Such studies are more models of conflict than theories of conflict since there is a piecemeal aspect to them. This becomes apparent when one reads some of the comprehensive and definitive works on conflict (Sambanis (2000); Blattman and Miguel (2010); Coyne and Mathers (2011)) among others. All of these works are organized in ways that make it evident that there is not one, or even a few, theories of conflict under whose umbrella one can organize the existing theoretical and empirical works, whether it be predation versus production, greed versus grievance, or rational versus Irrational models of conflict. Each of these demarcations speaks more to the modeling framework adopted than the theories they present. The task, then, of presenting a comprehensive account of the existing models for predicting conflict is at best unwieldy and at worst challenging.

The Fragile States Index (FSI) produced by The Fund for Peace (FFP), is an advanced tool in featuring not just the ordinary problems that all states experience, yet additionally in recognizing when those problems are exceeding a states' ability to deal with those problems.

By featuring appropriate weaknesses which add to the danger of state fragility, the Index, the socio-politico-economics framework and the data analysis tools whereupon it is constructed makes political risk assessment and early warning of conflict available to leaders and the masses.

The quality of the FSI is its capacity to distil a huge number of snippets of data into a structure that is significant just as effectively understandable and instructive. Every day, FFP gathers a huge number of reports and data from around the globe, enumerating the current social, monetary and political problems faced by each of the 178 nations.

Use of machine learning tools in order to predict the socio-politico-economic problems of a particular country was first found in the research paper “Predicting High-Risk Countries for Political Instability and Conflict” by Blair Huffman, Emma Marriott, April Yu (2016). They were also using the same dataset but currently it has been updated by almost 4 years. Furthermore, they didn’t use the time series as well as the deep learning methods in order to predict the fragility of a state in future.

For validation, an article from Wikipedia “List of Riots” will be used. It’s a comprehensive list of major riots that has happened across the world in each and every country. Countries which has political dictatorship like Russia, China, North Korea etc. where news of such events doesn’t come out but still, we may have some glimpse of agitation and death count which is recorded in this list.

* 1. **Problem Statement**

We are currently living in a society which is extremely polarized. Whether its about the polarization among the religions in India, where Hindu supremacy is on the rise since 2014 or the polarization in United Kingdom where Brexit or No Brexit debate polarized the whole country or the closed / open economy debate that polarized America. These kind of thinking gives rise to the political tensions. The leaders of such countries create non progressive agendas among the public to divide them in factions, divide their votes and then try to win the parliamentary seats at any cost. This kind of strategies make countries unstable in social sense if not in economic sense. People start to hate each other for the reasons which may not have even existed or they existed in a distant past. Such states with so much amount of polarization involved are actually becoming unstable such polarity goes beyond social things and is involved politically as well as economically.To measure such things FFP came up with a concept of measuring these things on yearly basis since 2006. The quantization of such things is there in our dataset. Our research involves in clustering such countries which are high on fragility or they are becoming unstable whether its politically or socially or economically. Our research will also try to forecast the fragility of a country on such data using deep learning (LSTM).

* 1. **Aims and Objectives**

Our aims in this research are:

* To create models which will be able to predict socio-political-economically unstable or conflict/riot/civil war prone countries using the FSI dataset.
* Check for the variables which are responsible for the conflict to happen in a country.
* The forecasting of the fragile/unstable countries.

Our objectives in this research are:

* The usage of clustering machine learning algorithms and time series with deep learning algorithm in order to predict and forecast politically unstable or civil war prone countries using the FSI dataset.
* To measure the performance of various clustering machine learning algorithms being used and choose the best one for prediction.
* To forecast fragile countries using the time series with LSTM.
  1. **Research Questions**

1. Can Machine Learning techniques predict the future conflicts that might occur in any country across the world?
2. How well will an unsupervised learning technique be able to segment the countries on the basis of conflict occurrences?
3. How oversampling technique supports the analysis of time series for better prediction?
4. How LSTM can forecast the fragility of a country fed with time series data?
   1. **Scope of the Research**

The scope of our research is limited to the clustering and validation of those clustered countries also in the second section i.e. time series with LSTM it will be very difficult to forecast the fragility of each and every country. It could be done on a CI-CD server where each and every country’s projections and forecast could be done overnight with few automation steps. In our research we will be doing a case study on the claim made by IMF that Bangladesh will overtake India’s economy by 2025. In this case the forecasting of India’s economy will have to be on the decline and Bangladesh’s on uprise.

* 1. **Significance of the Research**

The applications of such estimation are phenomenal and have a direct impact on our social, political and economic well-being. If the country where we live in is basically on decline on such measures, chances are that it is becoming fragile. If the leadership is not putting any effort in resolving such matters its better to find another country for our home where we can live safely.

Our research can forecast the future of a country on fragility. If the leadership of that country is responsible enough to believe that data, they can take appropriate measures in mitigating those issues. But if the leadership is not taking any responsibility then the people will have to decide to stay there or leave the country.

This research can also be used by the major finance companies who provide loans to such countries so that they can progress. Such companies can deny the loans to fragile countries if their fragility index is going higher and correct measures are not being taken up by their respective governments. The multinational companies (MNCs) can project their business growths in such countries and may halt / continue their investment plans by checking our research.

* 1. **Structure of the Study**

The structure of the thesis is as follows. Chapter 1 presents the background of the research in in collecting and quantizing the fragility data of all the countries in the world, discusses the problem statements. The study aims and objectives are discussed in section 1.3. Section 1.4 presents the scope of the study to the body of knowledge. The significance of the study is provided in section 1.5.

Chapter 2 presents the necessary literature review and highlights the finalization of the research methodology steps, feature fusion approaches, related research publications and techniques followed by the research which has already been carried out using similar parameters with the gap areas clearly indicated. 2.2.1 and Section 2.2.2 presents the clustering research work previously done on the similar dataset or with similar aim. Section 2.2.3 talks about the various research publications and techniques in the area of time series with LSTM. Section 2.2.4 presents the discussion about the difference between the ARIMA and LSTM approaches. Section 2.3 discusses about finalization of research methodology based on the previous research papers. Section 2.4 is the summary of the literature review is discussed and concluded.

Chapter 3 discusses about the research methodology, the proposed framework as well as the techniques being used. Section 3.1 discusses about the introduction of our research methodology, explains the data variables in depth, discusses about the requirements to perform this research, explains the algorithms and techniques being used in this research. Section 3.2 discusses about the novelty in our research using clustering algorithms, it’s about the kind of clustering techniques and steps we are using which is novel in this research. Section 3.3 discusses about the novelty in our research using time-series with LSTM, it’s about the kind of steps which we are using are novel in this research. Section 3.4 discusses about the expected outcomes from our research. Section 3.5 discusses about the summary of the whole research methodology.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Introduction**

This topic where the data science is being used as tool in segregating and forecasting the countries is actually niche. There are not many research papers found with the same subject although some related works in the same or different fields is found. The related works in different fields is opted on the basis of the approach which is similar to our research methodology. Old research papers such as Sambanis (2000), Blattman and Miguel (2010), and Coyne and Mathers (2011) hadn’t used any kind of data science but had only presented the modelling framework rather than the theories they present. The research conducted by Blair Huffman, Emma Marriott and others (2016) is the base paper we are using in this research. Other researches by Atin Basuchaudhary and others (2018, 2019) is mostly about the state failures and terrorism. They tried to predict if terrorism roots are found in one state then will it become a terrorist state later on or the factors that lead to the failure of state.

In our research we are using the FSI data which was previously used by Blair Huffman, Emma Marriott and others (2016). Their research was mostly based on predicting the conflict or civil war in the countries while in our research we are not predicting but trying to project the conflict in the near future. They used the FSI data as well as the world bank data for their research although we are using only the FSI data in our research. They also didn’t try to validate the model using the historic data but we are doing it in our research using the list of riots (Wikipedia, 2020).

**2.2.1 K-Means Clustering**

The only previous research on this subject and data that dealt with this algorithm was conducted by Blair Huffman and others (2016). K-means was used in the clustering algorithm only and not the hierarchical. They used Support Vector Machine (SVM) methods other than the clustering algorithms to predict conflict / civil war in the countries, they heavily depended upon the split test data for validation of their respective models and used the total score of FSI indices rather than getting the real data and validating their model.

Atin Basuchaudhary and others discussed in their book “Predicting Hotspots” about K-Means clustering technique but they didn’t use it in their journal papers. They used tree-based classifications to classify the countries as failed states. In the works of Jack A. Goldstone and others (2009) they didn’t discuss any algorithm in forecasting the failed state at all.

**2.2.2 Hierarchical Clustering**

In all the research papers that have been covered so far, none of them have ever used this clustering algorithm to classify the countries according to their fragility parameters. We are using it on the basis of the coding part so far done. We found that the ratio of the fragile to non-fragile countries as classified by both the clustering algorithms were a bit different. So, the clustering algorithm that will be giving us better confusion matrix parameters should be used in the future too.

**2.2.3 Time Series with Long Short-Term Memory (LSTM)**

Previously Alaa Sagheer and Mostafa Kotb in their research paper (2018) used LSTM on the time series data of production of crude oil from the oil fields. They were able to forecast properly what would be the productions of those oil fields in near future. Similarly, Yu-Xi Wu and others in 2019 published a research paper to forecast the oil prices in the near future.

Taking the above-mentioned papers as base, we are going ahead with the strategy of time series data fed to LSTM network and get the predictions for the future dates. Alaa Sagheer and Mostafa Kotb used a strategy of converting the data into stationary data first and then using the LSTM on it while Yu-Xi Wu didn’t try that. We will try to check if the stationary data is better for forecasting or the normally transformed data is enough for good forecasts.

**2.2.4 Why LSTM and not ARIMA (Time Series methodology)?**

Sima Siami-Namini and others published their comparative research in 2018 based on the traditional time series strategy ARIMA and LSTM predictions. LSTM outperformed ARIMA models from huge margin in terms of RMSE (root mean square error). So, it will be a waste of time to go ahead with ARIMA model first and then try LSTM, we will directly jump to LSTM model once our time series dataset is ready. Also, ARIMA model requires data to be stationary in nature and for making a data stationary we require at least 30 data points and in our dataset we have only 15, so ARIMA can’t be used in that case too.

**2.3 Discussion**

The subject of socio-politico economics fused with data science is relatively new. Very few papers were being found where a subject of socio-politico economics is being researched using data science. Atin Basuchaudhary and others (2018) in their book ‘Predicting Hotspots’ discussed about how to approach with socio-politico economics subject with data science but still their research work as per their published journal is in quiet nascent stage. Also, the dataset being used by them is quite different as compared to our research. We hope in near future there will be more accurate and concise datasets available publicly about the socio-politico economics of all the countries in the world.

There were some research papers that discussed about the freedom and democracy as the progressive traits of a country, while studying them was interesting but including other factors in our dataset wasn’t required as FSI data (variable explained in research methodology) somehow has that information contained in the variables. So, complicating the dataset more will not be required.

**2.4 Summary**

After going through the previous research papers and other technical articles we decided to move ahead with both the clustering algorithms in order to classify fragile and non-fragile countries. We also decided to move ahead with the approach of time series with LSTM in order to forecast the fragility of a country. Traditional time series approaches have been outperformed by the deep learning algorithm (LSTM) in terms of forecasting the data with less root mean square error (RMSE).

**CHAPTER 3**

**RESEARCH METHODOLOGY**

**3.1 Introduction**

Our research aims at the study of the fragile state index data which can be called as the observed events quantized so that statistical and time series methods can be used on them to describe, explain and predict the development indexes of the countries which provides us the information about the fragility of it. In this research the both quantitative and qualitative methods are being used to explain the current condition and predict the future conditions of a particular country.

Following are the techniques of machine learning that we will use to fulfill our aims and objectives:

1. K-means and Hierarchical Clustering
2. Time Series Analysis with Long Short-Term Memory (LSTM)

K-means and Hierarchical clustering will be used to verify the past events according to the FSI data and the riots data while Time Series Analysis and LSTM combination will be used to observe the future trends of a particular country which will give us an idea towards its fragility.

**3.1.1**  **Data Explanation**

The FSI data [2] has 12 indicators and we will be using the data from the year 2006 to 2020. The explanation of all the indicators is as follows:



Figure 3.1: Cohesion Indicators (Indicators, n.d.)

**3.1.1.1 Cohesion Indicators**

**C1: Security Apparatus:** The Security Apparatus Indicator thinks about imbalance inside the economy, regardless of the real presentation of an economy. For instance, the Indicator takes a look at basic imbalance that depends on society, (for example, racial, ethnic, strict, or other personality gathering) or dependent on training, monetary status, or locale, (for example, metropolitan provincial gap).

The Indicator thinks about real imbalance, yet additionally impression of disparity, perceiving that view of monetary imbalance can fuel complaint as much as possible, support shared pressures or nationalistic way of talking. Further to estimating financial disparity, the Indicator additionally accepts into account the open doors for society to advance their monetary position, for example, through admittance to business, instruction, or employment preparing with the end goal that regardless of whether there is financial imbalance present, how much it is basic and fortifying.

**C2: Factionalized Elites:** This indicator thinks about the fracture of state foundations along ethnic, class, group or race just as and brinksmanship and gridlock between administering elites. It additionally factors the utilization of jingoistic radical way of talking by administering elites, frequently as far as patriotism, xenophobia, collective irredentism or of common unity (e.g., "ethnic cleansing" or "safeguarding the religion"). In extraordinary cases, it very well may be illustrative of the nonattendance of authentic initiative broadly acknowledged as speaking to the whole population. This pointer estimates power battles, political rivalry, political advances, and where decisions happen will factor in the validity of discretionary cycles (or in their nonattendance, the apparent authenticity of the decision class).

## C3: Group Grievance: This Indicator centers around divisions and factions between various group of people in the public eye – especially divisions dependent on social or political abilities – and their part in admittance to administrations or assets, and consideration in the political cycle. These groups may likewise have a recorded past, where wronged other groups refer to shameful acts of the past, now and again returning hundreds of years, that impacts and shapes that group's function in the public space and associations with different groups. This set of experiences may thus be formed by examples of genuine or saw atrocities or "violations" submitted with clear exemption against other groups. These groups may likewise feel abused on the grounds that they are denied self-governance, self-assurance or political freedom to which they accept they are entitled. The Indicator additionally looks about where explicit groups are singled out by state specialists, or by prevailing groups, for abuse or suppression, or where there is public accusing of other groups accepted to have gained riches, status or influence "misguidedly", which may show itself in the rise of searing way of talking, for example, through "disdain" radio, pamphleteering, and cliché or nationalistic political discourse.

## 

Figure 3.2: Economic Indicators (Indicators, n.d.)

**3.1.1.2 Economic Indicators**

## E1: Economic Decline and Poverty: This Indicator considers factors identified with monetary decay inside a nation. For instance, the Indicator takes a look at examples of reformist financial decrease of the general public overall as estimated by per capita income, Gross National Product, joblessness rates, swelling, efficiency, obligation, destitution levels, or business disappointments. It additionally considers unexpected drops in product costs, exchange income, or unfamiliar venture, and any breakdown or downgrading of the public cash. This Indicator further looks about the reactions to financial conditions and their results, for example, outrageous social difficulty forced by monetary importance programs, or saw expanding group differences. This Indicator is centered around the proper economy – just as unlawful exchange, including the medication and illegal exploitation, and capital flight, or levels of violation and unlawful exchanges, for example, tax evasion or fraud.

## E2: Uneven Economic Development: This Indicator indicates about imbalance inside the economy, regardless of the real exhibition of an economy. For instance, the Indicator takes a look at auxiliary imbalance that depends on public, (for example, racial, ethnic, strict, or other character gathering) or dependent on training, financial status, or locale, (for example, urban rural gap). The Indicator indicates us about real imbalance, yet in addition view of disparity, perceiving that impression of financial disparity can fuel complaint as much as possible, strengthen shared strains or nationalistic manner of speaking. Further to estimating financial disparity, the Indicator additionally accepts into account the open doors for public to improve their monetary status, for example, through admittance to business, instruction, or occupation preparing with the end goal that regardless of whether there is financial imbalance present, how much it is public oriented and strengthening.

## E3: Human Flight and Brain Drain: This Indicator thinks about the monetary effect of human removal (for financial or political reasons) and the outcomes this may have on a nation's turn of events. From one perspective, this may include the willful resettlement of the working class – especially financially profitable portions of the population, for example, business visionaries, or gifted specialists, for example, doctors – because of monetary disintegration in their nation of origin and the expectation of better open doors farther abroad. Then again, it might include the constrained removal of experts or learned people who are escaping their nation because of real or dreaded oppression or restraint, and explicitly the monetary effect that uprooting may unleash on an economy through the loss of gainful, talented expert work.

## 

## Figure 3.3: Political Indicators (Indicators, n.d.)

## 3.1.1.3 Political Indicators

## P1: State Legitimacy: This Indicator considers the representativeness and transparency of government and its relationship with its public. The Indicator takes a look at the populace's degree of trust in state organizations and measures, and surveys the impacts where that certainty is missing, showed through mass public showings, continued common noncompliance, or the ascent of equipped insurgencies. In spite of the fact that the State Legitimacy pointer doesn't really make a judgment on fair administration, it considers the respectability of races where they happen, (for example, boycotted races), the idea of political advances, and where there is a nonattendance of majority rule decisions, how much the legislature is illustrative of the number of inhabitants in which it oversees. The Indicator considers receptiveness of government, explicitly the receptiveness of administering elites to straightforwardness, responsibility and political portrayal, or alternately the degrees of degradation, profiteering, and underestimating, abusing, or in any case barring resistance groups. The Indicator additionally considers the capacity of a state to practice essential capacities that inference a populace's trust in its administration and organizations, for example, through the capacity to gather duties.

## P2: Public Services: This Indicator alludes to the presence of fundamental state works that serve the individuals. From one viewpoint, this may incorporate the arrangement of fundamental administrations, for example, security, education, water and electricity, transport, and internet. Then again, it might incorporate the state's capacity to secure its residents, for example, from psychological warfare and brutality, through saw compelling policing. Further, even where fundamental state capacities and administrations are given, the Indicator further considers to whom – regardless of whether the state barely serves the decision-making elites, for example, security organizations, presidential staff, the national bank, or the appeasing assistance, while neglecting to give equivalent degrees of administration to the overall people, for example, country versus metropolitan populaces.

## The Indicator likewise considers the level and support of general foundation to the degree that its nonappearance would contrarily influence the nation's real or possible turn of events.

## P3: Human Rights and Rule of Law: This Indicator considers the connection between the state and its public to the extent that principal common liberties are secured and opportunities are monitored and regarded. The Indicator takes a look at whether there is inescapable maltreatment of legitimate, political and social rights, including those of people, groups and establishments (for example badgering of the press, politicization of the legal executive, inward utilization of military for political finishes, suppression of political adversaries). The Indicator likewise looks about flare-ups of politically propelled (rather than criminal) brutality executed against regular folks. It additionally takes a glance at variables, for example, denial of fair treatment reliable with global standards and practices for political detainees or nonconformists, and whether there is current or developing tyrant, oppressive or military guideline in which established and majority rule foundations and cycles are deferred or controlled.

## 

## Figure 3.4: Social and Cross-Cutting Indicators (Indicators, n.d.)

## 3.1.1.4 Social Indicators

## S1: Demographic Pressures: This Indicator considers pressures upon the state getting from the public itself or the earth around it. For instance, the Indicator estimates populace pressures identified with food gracefully, admittance to safe water, and other life-continuing assets, or wellbeing, for example, pervasiveness of sickness and pandemics. The Indicator looks about segment qualities, for example, pressures from elite groups development rates or slanted public appropriations or pointedly different paces of public development among competing different groups, perceiving that such impacts can have significant social, financial, and political impacts.

## Past the public, the Indicator additionally considers pressures originating from catastrophic events (tropical storms, quakes, floods or dry season), and weights upon the populace from ecological dangers.

## S2: Refugees and IDPs: This Indicator gauges the weight upon states brought about by the constrained dislodging of enormous networks because of social, political, natural or different causes, estimating removal inside nations, just as exile streams into others. The marker estimates displaced people by nation of Asylum, perceiving that public inflows can squeeze public administrations, and can in some cases make more extensive compassionate and security encounters for the accepting state, if that state doesn't have the adjustment limit and sufficient assets. The Indicator likewise gauges the Internally Displaced Persons (IDP) and Refugees by nation of starting point, which predicts interior state pressures because of brutality, natural or different factors. These measures are considered inside the setting of the state's general public (per capita) and human growth index, and after some time, perceiving that a few IDPs or exiles for instance, may have been removed for extensive amount of time.

## 3.1.1.5 Cross-Cutting Indicators

## X1: External Intervention: This Indicator thinks about the impact and effect of outer interveners in the working, especially security and financial system of a state. From one viewpoint, this indicator centers around security parts of commitment from outside interveners, both undercover and unmistakable, in the inner issues of a state in danger by governments, armed forces, insight administrations, other groups, or different elements that may influence the overall influence (or goal of a contention) inside a state.

## Then again, External Intervention additionally centers around monetary commitment by outside interveners, including multilateral associations, through huge scope advances, improvement ventures, or unfamiliar guide, for example, progressing spending support, control of accounts, or the board of the state's financial strategy, making financial reliance. Outside Intervention likewise considers philanthropic mediation, for example, the organization of a worldwide peacekeeping mission.

## Our Dataset has 178 countries (2020) before that there may be lesser countries as some countries got split for example Sudan got split into Sudan and South Sudan since 2011.There is a list of riots [1] which will be used for checking the efficiency and also for checking the predictive power of the machine learning method used for the unsupervised learning.

**3.1.2 Tools Required**

**Hardware Requirement:**

* Processor: i5/i7
* RAM: 8/16 GB

**Software Requirements:**

* Operating System: Windows 10 or Ubuntu Linux
* Python Environment / Tools: Anaconda 3
* Python Libraries: Numpy, Scipy, Pandas, Keras etc.
* Text Editors

**3.1.3 Data Preparation**

Before performing any analysis using the Machine Learning algorithms the data needs to be prepared in a particular format, so that good results can be obtained using them. For example, the data is supposed to be “Gaussian Like” or its distribution of each variable should be normal in nature. These are the steps which should be performed on the data to create a dataset that should be analysis ready. They are:

* 1. Check the raw data for the number of variables that are numeric or category based.
  2. Separate the numeric and category-based variables.
  3. Check for the distribution of the numeric variables if they are not normally distributed, apply a scalar that can make them, for example: PowerTransformer in python SciKit-Learn.
  4. Create Dummy variable for the categorical variables.
  5. Combine all the numeric and dummy variables to make a complete dataset.

In our data we have 15 files from the year 2006 to 2020, some algorithms like K-means will be using the files individually and create the respective datasets on the other hand for time series analysis we will be combining all the datasets from the year 2006 to 2020 and timestamp each row data for further analysis.

The oversampling of Time Series might be required, we will be using the resampling to oversample the number of data points in the time series. Later on scaling the data using PowerTransformer to make the data more “gaussian like” so that analysis and prediction can be done properly.

After completing the above steps our data is ready for analysis and we can use the Machine Learning algorithms further for our research.

**3.1.4 K-Means**

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster. We will be using Hopkins method as well as elbow curve to find the best number of clusters. Once we are able to label the data with respective to their cluster identity, we might be experimenting with the different values of cluster for example 2 for countries: conflict free and conflict countries, 3: for developed, developing and under-developed countries. Later on, we can check about the countries which are developed / developing and still have chances of conflict as the under-developed countries have high chances of conflict as they are poor, low on education and healthcare, they also are the victims of poor governance. With every year change in the cluster label of country we can check for the instability occurring in it.

This is also going to be an unsupervised way of learning the conflict occurring countries. There won’t be any sort of prediction in this approach, it is more about grouping the countries and checking on the basis of data how well they are performing. The countries differentiated as the conflicted ones can be compared from the list of Riots (Wikipedia 2020), and the confusion matrix can be made. The novelty present in this approach is that we are comparing our clusters formed with the real data of riots (Wikipedia 2020).

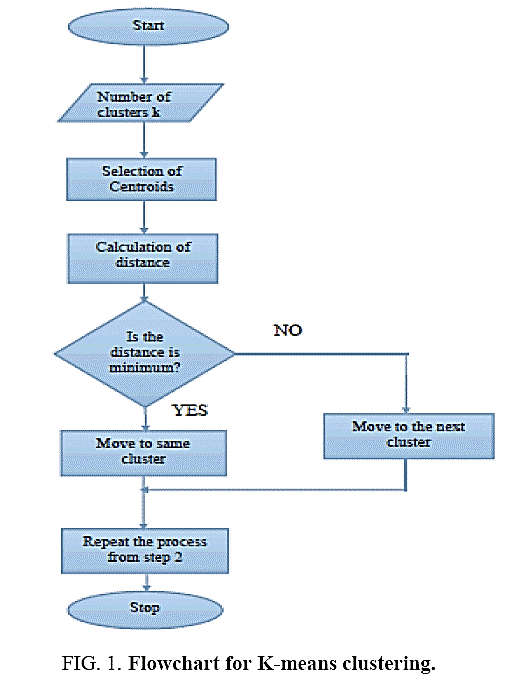


Figure 3.5: K-Means Algorithm Flow Chart (Palamera et al., 2011)

**3.1.5 Hierarchical clustering**

Hierarchical cluster analysis or HCA is a strategy for bunch examination which tries to manufacture a hierarchy of clusters. Here the countries are separate entities till we apply the algorithm and start combining them in the respective clusters. We will be trying the bottom up approach also known as the agglomerative clustering where in the end we get the dendrograms and we get to know in which category the country actually falls. We might be experimenting with the different values of cluster for example 2 for countries: conflict free and conflict countries, 3: for developed, developing and under-developed countries. Later on, we can check about the countries which are developed / developing and still have chances of conflict as the under-developed countries have high chances of conflict as they are poor, low on education and healthcare, they also are the victims of poor governance. With every year change in the cluster label of country we can check for the instability occurring in it.

This is also going to be an unsupervised way of learning the conflict occurring countries. There won’t be any sort of prediction in this approach, it is more about grouping the countries and checking on the basis of data how well they are performing. The countries differentiated as the conflicted ones can be compared from the list of Riots (Wikipedia 2020), and the confusion matrix can be made. The novelty present in this approach is that we are comparing our clusters formed with the real data of riots (Wikipedia, 2020).

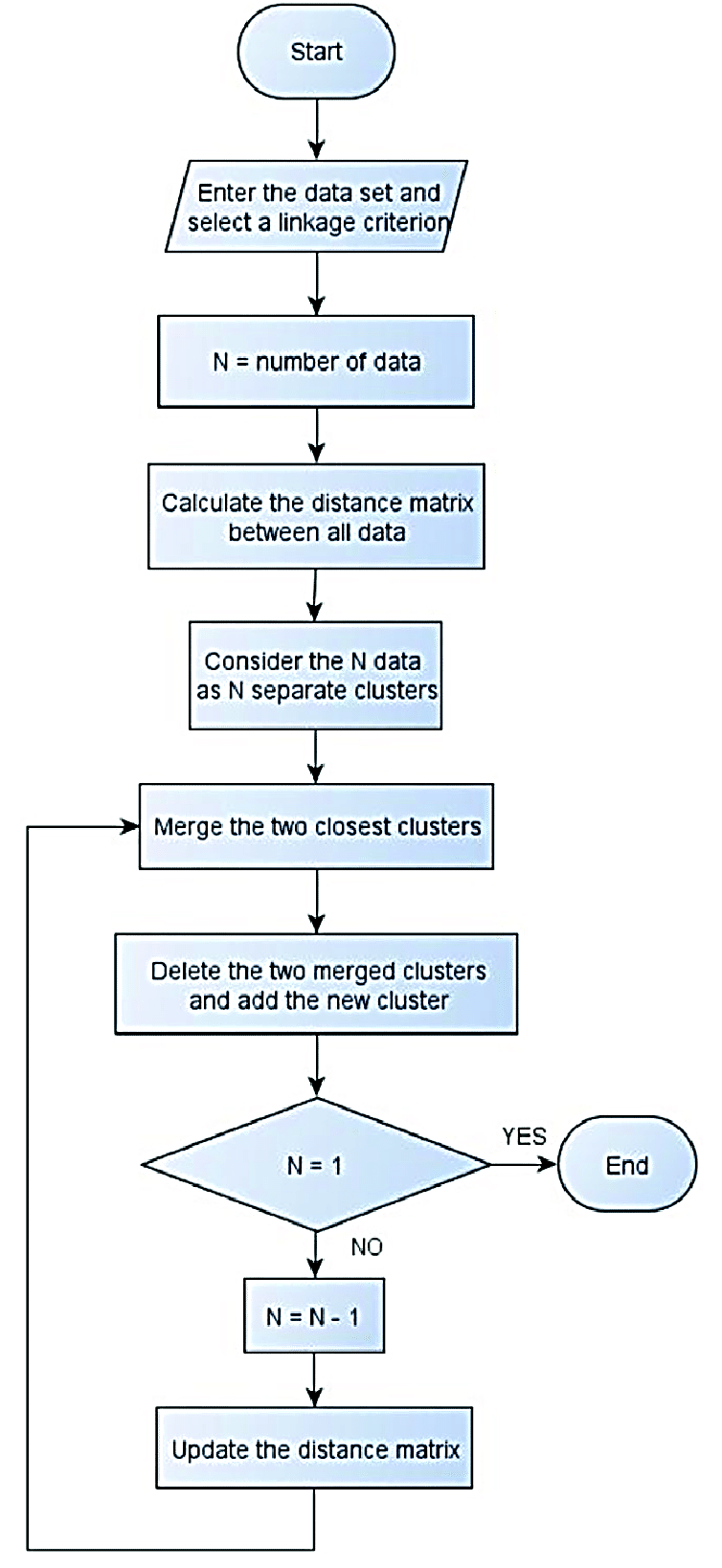


Figure 3.6. Hierarchical Clustering Flow Chart (Palamera et al., 2011)

We are using both the approaches and will be validating both of them as the number of countries ratio with respect to cluster identities will be different, also we will be differentiating them on the basis of their validation performance.

**3.1.6 Time Series**

Time series analysis contains techniques for examining time series data so as to extricate significant insights and different attributes of the information. Time series forecasting is the utilization of model to foresee future qualities dependent on previously observed qualities. We will take the data with respect to each and every country from 2006-2020. So, every country will have 15 rows of data on which we can firstly perform the Exploratory Data Analysis and try to convert them to Time Series. If 15 data points are going to be really less for the forecasting part, we can try to resample them using oversampling techniques.

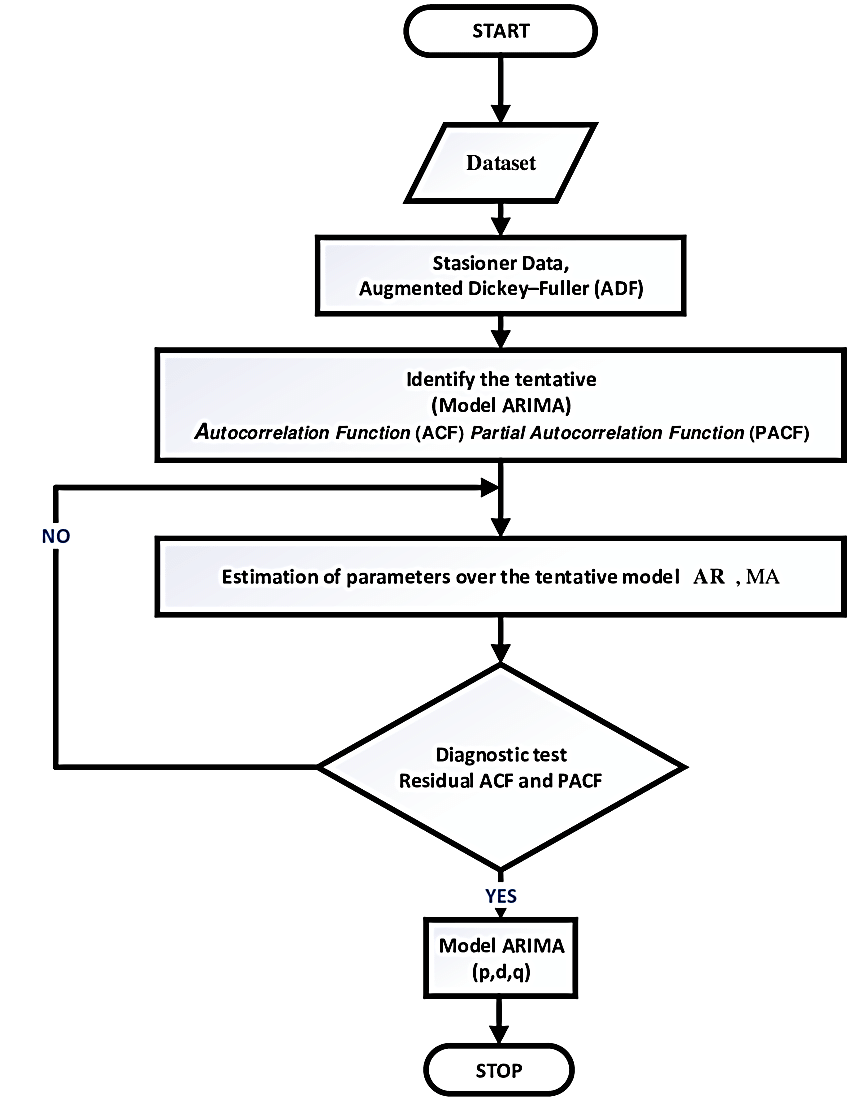


Figure 3.7: Time Series Forecasting Flow Chart (Palamera et al., 2011)

**3.1.7 Long Short-Term Memory (LSTM)**

Long Short-Term Memory (LSTM) is a unique type of Recurrent Neural Network (RNN) capable of learning long-term dependencies, which is useful for certain types of prediction that require the network to retain information over longer time periods.

RNNs are designed to take care of a sequence of events that occur in succession, with the understanding of each event based on information from previous events. The architecture of RNNs restricts its long-term memory capabilities, which are limited to only remembering a few sequences at a time. Consequently, the memory of RNNs is only useful for shorter sequences and short time-periods.

LSTMs are designed to overcome the vanishing gradient problem and allow them to retain information for longer periods compared to traditional RNNs. LSTMs can maintain a constant error, which allows them to continue learning over number of time-steps and backpropagate through time and layers.

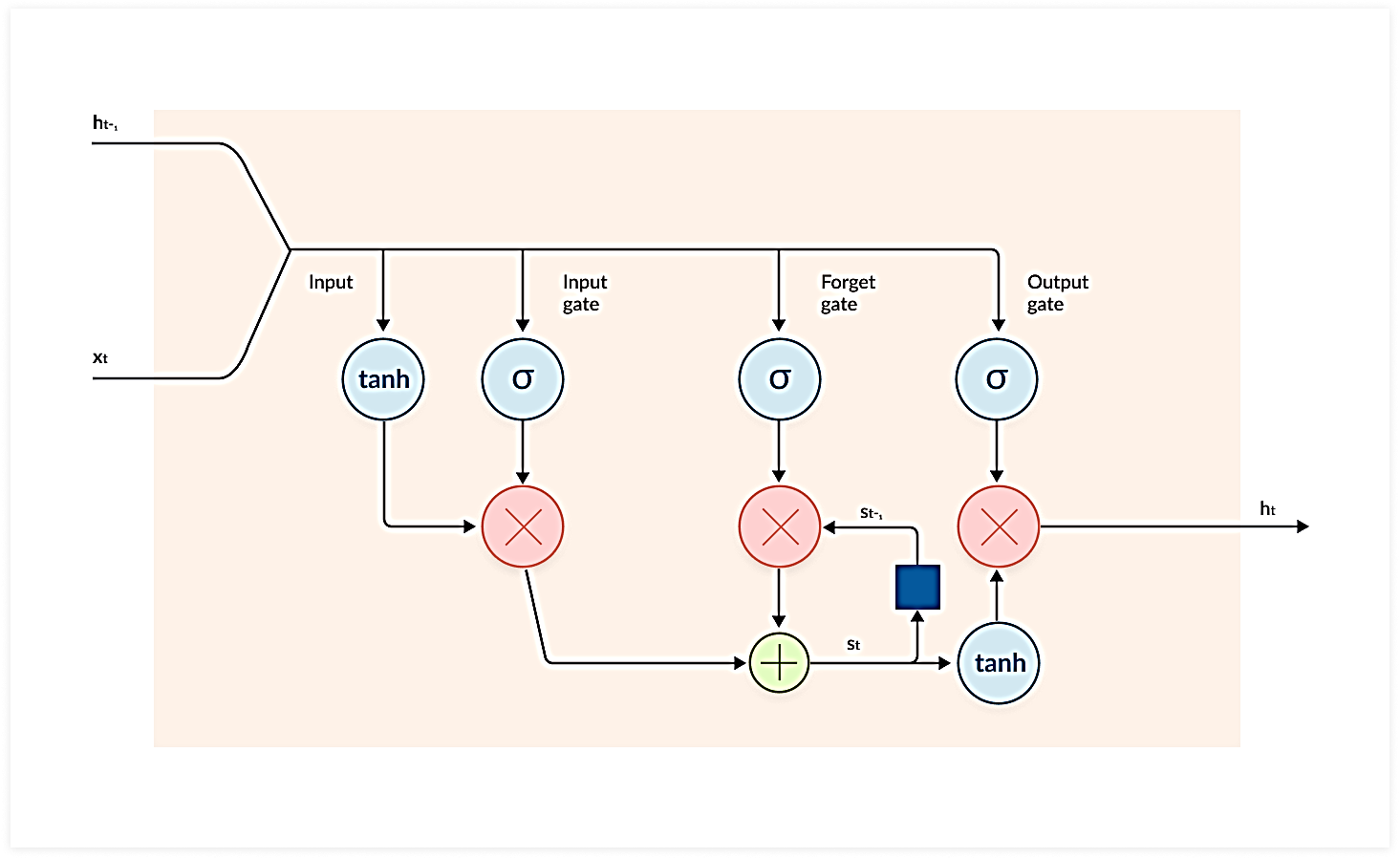


Figure 3.8: LSTM Network (MissingLink.ai, 2020)

As seen in the diagram above, LSTMs use gated cells to store information outside the regular flow of the RNN. With these cells, the network can manipulate the information in many ways, including storing information in the cells and reading from them. The cells are individually capable of making decisions regarding the information and can execute these decisions by opening or closing the gates.

The chain-like architecture of LSTM allows it to contain information for longer time periods, solving challenging tasks that traditional RNNs struggle to or simply cannot solve.

The three major parts of the LSTM include:

* **Forget gate**: This gate removes information that is no longer necessary for the completion of the task. This step is essential for optimizing the performance of the network.

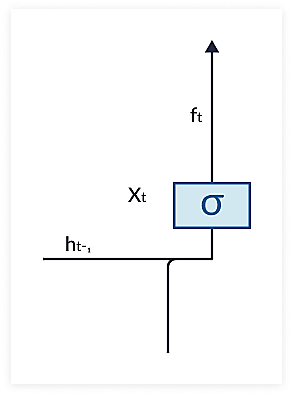


Figure 3.9: Forget Gate (MissingLink.ai, 2020)

* **Input gate**: This gate is responsible for adding information to the cells.

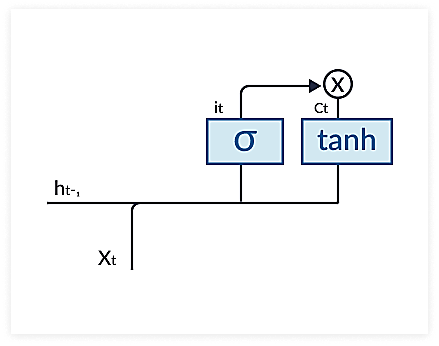


Figure 3.10: Input Gate (MissingLink.ai, 2020)

* **Output gate**: This gate selects and gives necessary output information.

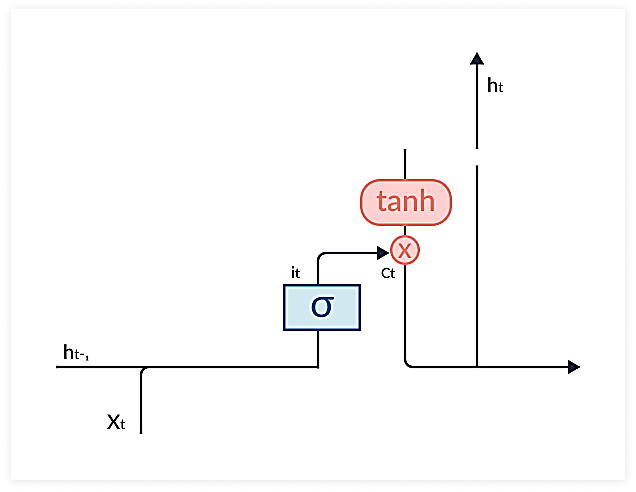


Figure 3.11: Output Gate (MissingLink.ai, 2020)

**3.2 Novel Methodology for K-Means**

The whole approach can be summarized as the following:

**3.2.1 Data Preparation (for K-Means and Hierarchical Approaches)**

* + 1. Combine all the data from 2006 to 2020 years.
    2. Combine the 2 columns i.e. country and year as Country\_Year.
    3. Scale the data using the PowerTransformer.



Figure 3.12: Data Preparation and Preprocessing

**3.2.2 Validation Data Preparation**

* + 1. Grab the data for the year 2006 to 2020 from the list of riots (Wikipedia 2020).
    2. Clean the data and make the data in the form of Country\_Year and mark all the entries as 1 in the column name Actual value. Till date this data has not been used for the analysis as per my research, so it’s a novelty.



Figure 3.13: Validation Data Preparation

**3.2.3 K-Means Data Modelling:**

* + 1. Now we continue with the Hopkins approach. We calculate the Hopkins measure on the dataset.
    2. Then we get the number of suggested clusters from the elbow curve.
    3. Then we mark all the countries on the column cluster values with their respective cluster values filled in the column.



Figure 3.14: Determination of Number of clusters

**3.2.4 K-Means Validation and Confusion Matrix**

* + 1. We perform a full outer join on the column Country\_Year on both the datasets i.e. the FSI data with cluster values and the riots data we got from the Wikipedia.
    2. We will get a lot of nan values in the actual column of the combined dataset we can fill them with values 0.
    3. Now we can make confusion matrix according to the predicted (cluster values) and the actual values from the Wikipedia page. According to the confusion matrix we can calculate accuracy, specificity, sensitivity and other scoring parameters, we can also draw the ROC too.

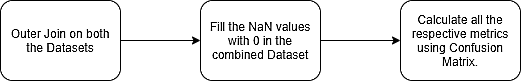
****

Figure 3.15: Validation of the K- Means Model

**3.2.5 Hierarchical Approach**

* + 1. Similarly, we can go ahead with the Hierarchical clustering, we will create the dendrograms using single and complete linkage methods and will check how many clusters are coming up from algorithm.
    2. Then we will mark the respective countries with the cluster ids.

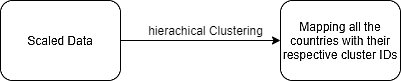


Figure 3.16: Hierarchical Clustering

**3.2.6 Hierarchical Approach Validation and Confusion Matrix**

* + 1. We perform a full outer join on the column Country\_Year on both the datasets i.e. the data we got above and the riots data we got from the Wikipedia.
    2. We will get a lot of nan values in the actual column of the combined dataset we can fill them with values 0.
    3. Now we can make confusion matrix according to the predicted (cluster values) and the actual values from the Wikipedia page. According to the confusion matrix we can calculate accuracy, specificity, sensitivity and other scoring parameters, we can also draw the ROC too.

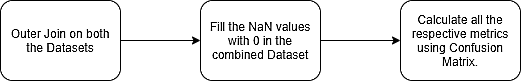
****

Figure 3.17: Validation of the Hierarchical Model

**3.3 Novel Methodology for Time Series with LSTM:**

The whole methodology can be summarized as following:

**3.3.1 Time-Series Data Preparation**

* + - 1. We will start with the data processing, combining all the data of the respective year.
      2. Then we generate the datasets according to the country.



Fig 3.18: Block Diagram to generate Datasets based on countries

* + - 1. Then we pick up one country’s dataset and sort the rows according to the years 2006 to 2020.
      2. Then we will use the time series resample method to make the yearly data quarterly. This will produce more datapoints in the respective dataset and will be good for the time series analysis.
      3. Now we will scale the data, we will experiment with many types of scalers and see which one will the best results. The scalers we are going to experiment are standard, min-max, max-abs and PowerTransformer.



Fig 3.19: Block Diagram to prepare an individual Dataset of a country

**3.3.2 LSTM Approach**

* + - 1. Create an LSTM (Recurrent Neural Network) model using the entire dataset there is no split involved here as we will be using the whole dataset to generate the future data points.
      2. Provide the future time values so that the LSTM network can predict the values of the respective columns on those time values.
      3. Generate the future data points on the given future time values for all the columns
      4. Use the inverse scaler to get the unscaled data for the whole country’s dataset.
      5. Create a new column of total points which will add all the other 12 column values in its value.

**3.3.3 Plotting and observing forecasted trend**

* + - 1. Plot the total column on the time series.
      2. If the trend of the plot goes up the country is destabilizing and it its going down the is getting strengthened / stabilizing.

****

Fig 3.20: Block Diagram of LSTM forecasting

**3.4 Expected Outcome**

**K-means / Hierarchical Linking**: We will be using this unsupervised learning algorithm to create the clusters of the countries which are prone to riot/conflict. We can check the algorithm performance by matching it with the real data.

**Time Series with LSTM**: Once we create the model which can predict the future trends, we will know if the trend is going up the country is getting strengthened and if the trend is going down, we will know that the country is getting unstable.

**3.5 Summary**

The research methodology has novelty related to the verification of the past events with respect to the cluster identities generated by the K-means and Hierarchical clustering approaches. On the other hand, the time series with LSTM has not been used before in projecting the fragility of a country. The K-means and Hierarchical approach will provide us with the confusion matrix and hence we can calculate the respective measures of the efficiency of the algorithm. Time Series with LSTM will give us the trend towards the fragility of a particular country.

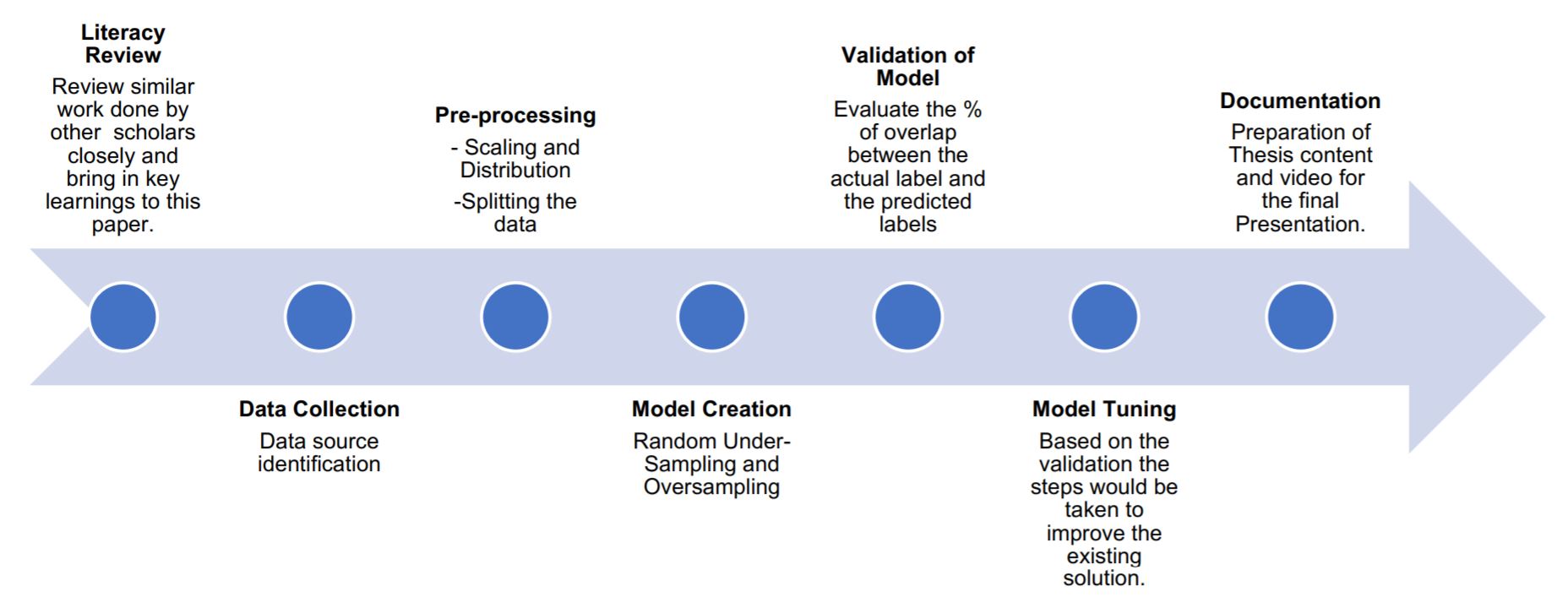
**References**

1. Wikipedia 2020 *List of riots*, Accessed on 4th Oct 2020. <https://en.wikipedia.org/wiki/List_of_riots#2001%E2%80%932009>
2. Fund for World Peace 2020. *Fragile States Index [Online]*, Accessed on 4th Oct 2020 <http://ffp.statesindex.org>
3. Blair Huffman, Emma Marriott and others Stanford University (2016). *Predicting High-Risk Countries for Political Instability and Conflict* <http://cs229.stanford.edu/proj2014/Blair%20Huffman,%20Emma%20Marriott,%20April%20Yu,%20Predicting%20high-risk%20countries%20for%20political%20instability%20and%20conflict.pdf>
4. Wikipedia 2020 *K-means clustering*, Accessed on 4th Oct 2020. <https://en.wikipedia.org/wiki/K-means_clustering>
5. Wikipedia 2020 *Time Series*, Accessed on 4th Oct 2020. <https://en.wikipedia.org/wiki/Time_series>
6. Iwana, Brian & Uchida, Seiichi. (2020). *Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher*. <https://www.researchgate.net/publication/340805406_Time_Series_Data_Augmentation_for_Neural_Networks_by_Time_Warping_with_a_Discriminative_Teacher/citation/download>
7. Wikipedia 2020, *List of wars and anthropogenic disasters by death toll*, Accessed on: 4th of Oct 2020. <https://en.wikipedia.org/wiki/List_of_wars_and_anthropogenic_disasters_by_death_toll>
8. Vision of Humanity, *Predicting Civil Conflict*, Accessed on 4th of Oct 2020. <http://visionofhumanity.org/economists-on-peace/predicting-civil-conflict-can-machine-learning-tell-us/>
9. Missinglink.ai 2020, *Long Short-Term Memory Networks,* Accessed on 14th Nov 2020. <https://missinglink.ai/guides/neural-network-concepts/deep-learning-long-short-term-memory-lstm-networks-remember/>
10. Collier, Paul, and Nicholas Sambanis. “Understanding Civil War: A New Agenda.” *The Journal of Conflict Resolution*, vol. 46, no. 1, 2002, pp. 3–12. *JSTOR*, www.jstor.org/stable/3176236. Accessed 28 Nov. 2020.
11. Blattman, Christopher, and Edward Miguel. “Civil War.” *Journal of Economic Literature*, vol. 48, no. 1, 2010, pp. 3–57. *JSTOR*, www.jstor.org/stable/40651577. Accessed 28 Nov. 2020.
12. Coyne, Christopher & Mathers, Rachel. (2010). Rituals: An economic interpretation. *Journal of Economic Behavior & Organization*. 78. 74-84. 10.1016/j.jebo.2010.12.009.
13. S. Siami-Namini, N. Tavakoli and A. Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, 2018, pp. 1394-1401, doi: 10.1109/ICMLA.2018.00227.
14. Alaa Sagheer, Mostafa Kotb, Time series forecasting of petroleum production using deep LSTM recurrent networks, Neurocomputing, Volume 323, 2019, Pages 203-213, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2018.09.082. (<http://www.sciencedirect.com/science/article/pii/S0925231218311639>)
15. Yu-Xi Wu, Qing-Biao Wu, Jia-Qi Zhu, Improved EEMD-based crude oil price forecasting using LSTM networks, *Physica A: Statistical Mechanics and its Applications*, Volume 516, 2019, Pages 114-124, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2018.09.120. (<http://www.sciencedirect.com/science/article/pii/S0378437118312536>)
16. Basuchoudhary, Atin & Bang, James. (2018). Predicting Terrorism with Machine Learning: Lessons from “Predicting Terrorism: A Machine Learning Approach”. *Peace Economics, Peace Science and Public Policy*. 24. 10.1515/peps-2018-0040.
17. Bang, James & Shughart II, William & Basuchoudhary, Atin. (2019). Predicting State Failure: Different Pathways into the Abyss.
18. JOUR, Goldstone, Jack, Epstein, David, Gurr, Ted, Lustik, Michael, Marshall, Monty, Ulfelder, Jay, Woodward, Mark, 2010, A Global Forecasting Model of Political Instability, 10.1111/j.1540-5907.2009.00426.x, American Journal of Political Science
19. From medium.com on “Predicting High-Risk Countries for Political Instability and Conflict using Tkinter” by Ravi Suthar. 28 Nov 2020, <https://medium.com/@udaan.ravi/predicting-high-risk-countries-for-political-instability-and-conflict-using-tkinter-cdb456d0908>
20. Usanov, Artur N. and Sweijs, Tim, Models Versus Rankings: Forecasting Political Violence (March 2017). Available at SSRN: https://ssrn.com/abstract=2930104. or <http://dx.doi.org/10.2139/ssrn.2930104>
21. Klaus Gründler, Tommy Krieger, Democracy and growth: Evidence from a machine learning indicator, European Journal of Political Economy, Volume 45, Supplement, 2016, Pages 85-107, ISSN 0176-2680, https://doi.org/10.1016/j.ejpoleco.2016.05.005., (<http://www.sciencedirect.com/science/article/pii/S0176268016300222>)
22. MUELLER, H., & RAUH, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review,* *112*(2), 358-375. doi:10.1017/S0003055417000570
23. Marcio Salles Melo Lima, Dursun Delen, Predicting and explaining corruption across countries: A machine learning approach, Government Information Quarterly, Volume 37, Issue 1, 2020, 101407, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2019.101407., (<http://www.sciencedirect.com/science/article/pii/S0740624X19302473>)
24. From towardsDataScience.com on “Time-Series Prediction Beyond Test Data” by Andrej Baranovskij, 28th Nov 2020. <https://towardsdatascience.com/time-series-prediction-beyond-test-data-3f4625019fd9>
25. From towardsDataScience.com on “LSTM to Predict Stock Prices — Time-Series Data” by Sarit Maitra, 28th Nov 2020. (<https://towardsdatascience.com/recurrent-neural-network-to-predict-multivariate-commodity-prices-8a8202afd853>)
26. Acemoglu, Daron, and James A. Robinson. 2001. “A Theory of Political Transitions.” American Economic Review 91 (4): 938–63.
27. Acemoglu, Daron, and James A. Robinson. 2012. Why Nations Fail: The Origins of Power, Prosperity, and Poverty. New York: Crown Publishing House.
28. Akerlof, George. 1979. “The Market for “Lemons”: Quality Uncertainty and the Market Mechanism.” Quarterly Journal of Economics 84 (3): 488–500.
29. Anderton, Charles H., and John R. Carter. 2009. Principles of Conflict Economics: A Primer for Social Scientists. Cambridge: Cambridge University Press.
30. Bang, James T., Atin Basuchoudhary, John David, and Aniruddha Mitra. 2016. “Predicting Aggregate Terror Risk: A Machine Learning Approach.” Working Paper.
31. Basuchoudhary, Atin, and Ghislain Dutheil de la Rochere. 2016. “Greed or Grievance, A Contextual Reconciliation?” Working Paper. Basuchoudhary, Atin, and J. Hentz. 2016. “The Conflict Dynamics of Warlords, Reformists, and the State: An Analytic Narrative.” Working Paper. Basuchoudhary, Atin, and Tinni Sen. 2016. “Toward a Predictive Theory of Civil Conflict: Path Dependence, Commitment, and Bargaining Failure.” Working Paper.
32. Bazzi, Samuel and Christopher Blattman. 2014. “Economic Shocks and Conflict: Evidence From Commodity Prices.” American Economic Journal: Macroeconomics 6 (4): 1–38. http://dx.doi.org/10.1257/mac.6.4.1.
33. Beard, Mary. 2010. “State Capacity, Conflict and Development.” Econometrica 78 (1): 1–34. ———. 2015.
34. SPQR: A History of Ancient Rome. New York, London: Liveright Publishing Corporation, A Division of W.W. Norton and Co. Besley
35. Timothy J., and Torsten Persson. 2008. “The Incidence of Civil War: Theory and Evidence.” National Bureau of Economic Research (NBER) Working Paper 14585 39.
36. Blomberg, S. Brock, and Gregory D. Hess. 2002. “The Temporal Links Between Conflict and Economic Activity.” Journal of Conflict Resolution 46 (1): 74–90.
37. Buchanan, James M. 1975. The Limits of Liberty: Between Anarchy and Leviathan. Chicago: University of Chicago Press.
38. Burke, Marshall B., Edward Miguel, Shanker Satyanath, John A. Dykema, and David B. Lobell. 2009. “Warming Increases Risk of Civil War in Africa.” National Academy of Sciences 106 (49): 20670–74.
39. Center For Systemic Peace. 2016. Global Conflict Trends: Assessing the Qualities of Systematic Peace, July 13, http://www.systemicpeace.org/conflicttrends.html.
40. Charness, Gary, and Matthew Rabin. 2002. “Understanding Social Preferences With Simple Tests.” Quarterly Journal of Economics 117 (3): 817069.
41. Chassang, Sylvain, and Gerard Padri-i-Miquel. 2009. “Economic Shocks and Civil War.” Quarterly Journal of Political Science 4 (3): 211–28.
42. Choi, Jung-Kyoo, and Samuel Bowles. 2007. “The Coevolution of Parochial Altruism and War.” Science 318 (5850): 637–40.
43. Ciccone, Antonio. 2011. “Economic Shocks and Civil Conflict: A Comment.” American Economic Journal: Applied Economics 3: 215–7.
44. Clark, Greg. 2007. A Farewell to Alms. Princeton, NJ: Princeton University Press.
45. Collier, Paul, and Anke Hoeffler. 1998. “On Economic Causes of Civil War.” Oxford Economic Papers 50 (4): 563–73. ———. 2004.
46. “Greed and Grievance in Civil War.” Oxford Economic Papers 56: 563–95.
47. Collier, Paul, and Nicholas Sambanis. 2002. “Understanding Civil War.” Journal of Conflict Resolution 46 (1): 3–12.
48. Collier, Paul, Anke Hoeffler, and Dominic Roehner. 2009. “Beyond Greed and Grievance: Feasibility and Civil War.” Oxford Economic Papers 61: 1–27.
49. Congleton, Roger D. 1995. “Ethnic Clubs, Ethnic Conflict, and the Rise of Ethnic Nationalism.”
50. In Nationalism and Rationality, by Albert Breton, Gianluigi Galeotti, Pierre Salmon, and Ronald Wintrobe, 71–97. Cambridge: Cambridge University Press.
51. Coyne, Christopher J., and Rachel L. Mathers. 2011. The Handbook of the Political Economy of War. Cheltenham, UK: Edward Elger. Cramer, Christopher. 2002. “Homo Economicus Goes to War: Methodological Individualism, Rational Choice, and the Political Economy of War.” World Development 30 (11): 1845–64.
52. Cross-National Time-Series data archives (CNTS). 2014. Dal Bó, Ernesto, and Pedro Dal Bó. 2011. “Workers, Warriors, and Criminals: Social Conflict in General Equilibrium.” Journal of the European Economic Association 9 (4) 646–77.
53. Dal Bo, Ernesto, and Robert Powell. 2009. “A Model of Spoils Politics.” American Journal of Political Science 53 (1): 207–22. Dal Bo, Pedro, and Guillame R. Frechette. 2011. “The Evolution of Cooperation in Infinitely Repeated Games.” The American Economic Review 101: 411–29.
54. Darley, John M., and Joel Cooper. 1998. Attribution and Social Interaction: The Legacy of Edward E. Jones. Washington, DC: American Psychological Association Press.
55. Desch, Michael C. 2008. Power and Military Effectiveness: The Fallacy of Democratic Triumphalism. Baltimore, MD: Johns Hopkins University Press.
56. Elbadawi, Ibrahim, and Nicholas Sambanis. 2000a. “External Intervention and the Duration of Civil Wars.” World Bank Policy. ———. 2000b. “Why Are There So Many Civil Wars in Africa? Understanding and Preventing Violent Conflict.” Journal of African Economies 93 (3): 244–69.
57. McBride, Michael, and Stergios Skaperdas. 2007. “Explaining Conflict in L Income Countries: Incomplete Contracting in the Shadow of the Future.” In Institutions and Norms in Economic Development, by Mark Gradstein, and Kai A. Konrad, 141–162. Cambridge and London: MIT Press.
58. McDermott, Rose. 2004. “The Feeling of Rationality: The Meaning of Neuroscientific Advances for Political Science.” Perspectives on Politics 2 (4): 691– 706.
59. Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. “Economic Shocks and Civil Conflict: An Instrumental Variables Approach.” Journal of Political Economy 112 (4): 725–53.
60. Mueller, Dennis C. 2003. Public Choice III. Cambridge and New York: Cambridge University Press. Murdoch, James C., and Todd Sandler. 2002. “Economic Growth, Civil Wars, and Spillovers.” Journal of Conflict Resolution 46 (1): 91–110.
61. Atin Basuchoudhary, James T. Bang, Tinni Sen and John David (2018), Predicting Hotspots: Using Machine Learning to Understand Civil Conflict. Published by Lexington Books An imprint of The Rowman & Littlefield Publishing Group, Inc.
62. Palamara, Federica & Piglione, Federico & Piccinini, Norberto. (2011). Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases. Safety Science. 49. 1215-1230. 10.1016/j.ssci.2011.04.003.
63. Indicators [online image] (n.d.) Available from: https://fragilestatesindex.org/indicators/ [Accessed: 30th November, 2020]

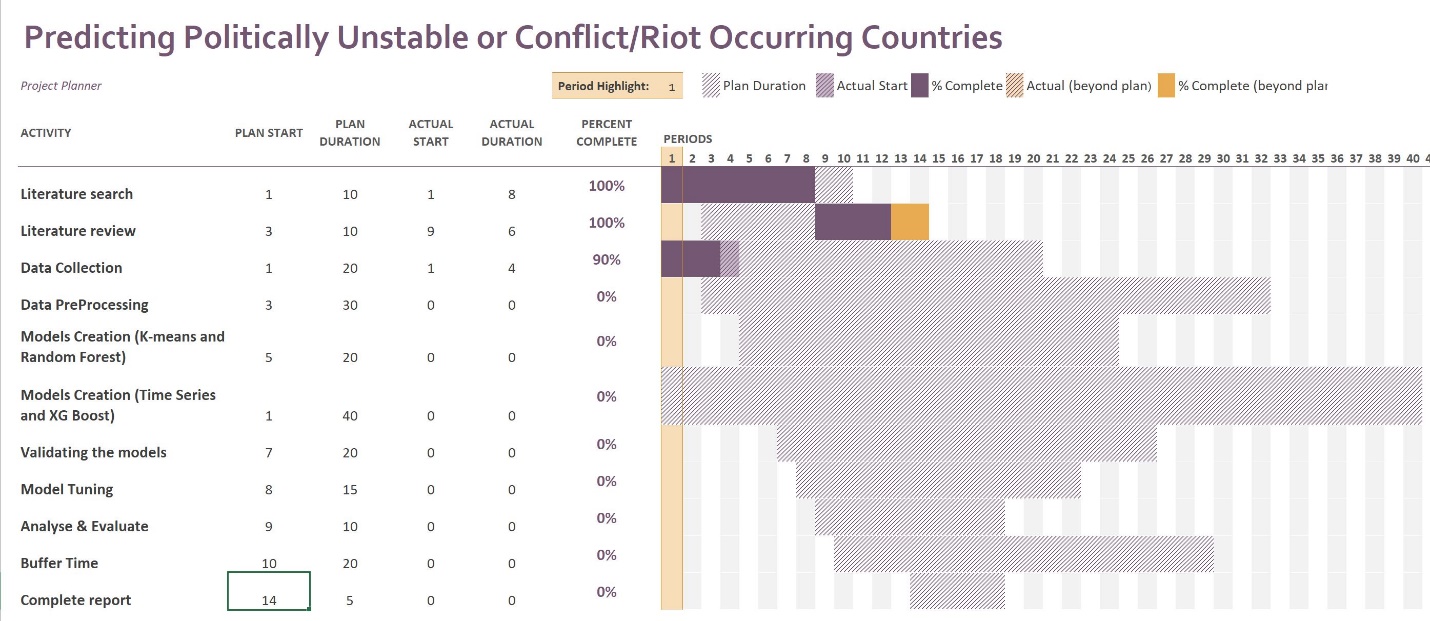
**Appendix A**

**Research Plan**

**Analysis Plan for components**

****

**Gantt Chart**



**Appendix B**

**Research Proposal**

**Rachit Dev**

Master of Data Science Research Proposal

Liverpool John Moore’s University

**Predicting Politically Unstable or Conflict/Riot Occurring Countries**

**Abstract**

We live in a world where we are facing conflict / riot news occurring in all the parts of the world. Though the impact is minor in majority of conflicts but we cannot get away with the major conflicts that occur in the certain parts of the world. There are many factors which contribute towards the occurrence of the conflict and we should try to predict the instability of the country by using machine learning tools and the relevant data. Machine Learning enables us to notice movements inside a country and counter with the right economic, political, and developmental authorizations by the government of the respective country and avoid clash or total governmental breakdown. Our inspiration is to capture and infer these movements on an impressive scale and construct a model that can show the fragility of a country. There are millions of people who lost their lives due to these conflicts [8] and by predicting them we can raise the alarm to the particular authorities or the citizens and their lives might be saved. We hope that by applying machine learning techniques we can predict the conflicts which might occur in the countries that are prone to it.

**Table of Contents**

1. [Abstract](#Abstract) 1

2. [Introduction](#Introduction) 2

3. [Background and Related Research](#Background) 2

4. [Research Questions](#ResearchQuestion) 2

5. Details of the Research Project 2

* [Aims and Objectives](#Aims)
* [Data Explanation](#data)

6. [Research Methodology](#Methodology) 9

7. [Outcomes](#Outcome) 13

8. [Requirements / Resources](#Requirement) 13

9. [Research Plan](#Timetable) / Timetable 14

[References](#Reference) 15

**Introduction**

Riots, civil war, armed protests, terrorists’ attacks are something which is problem all over the world. Thousands of people lose their lives in such events and these things can be avoided if we can predict them. These things occur due to various factors which we will be discussing in the data’s variable section. Predicting those countries where such conditions might arise is the sole aim of this research and we will be using machine learning as well as time series forecasting to predict it.

**Background and Related Work**

Various literatures related to prediction of politically instable / riot prone countries have been published already. Much of this published work discusses how techniques such exploratory data analysis, K-means, SVM, SMO, and SMO Regression be applied to predict conflict occurring countries. Many of these literatures are available to public usage. [3][9]

There are some NGOs and non-profit companies like Fund for Peace (Fragile State Index parent company) [2] who generates socio-economic-political indices for the all the prominent countries in the world and NGOs like Vision of Humanity [9] who do analysis on the data available worldwide and create heatmaps on the basis of the analysis performed.

There is a list of riots being maintained on Wikipedia [1] and lives lost due to he political and war activities [8]. These datasets are quite useful in understanding the inhumane aftermath of conflicts/riots/civil war situations.

**Research Question**

1. Can Machine Learning techniques predict the future conflicts that might occur in any country across the world?
2. How well will an unsupervised learning technique be able to segment the countries on the basis of conflict occurrences?
3. How oversampling technique supports the analysis of time series for better prediction?

**Details of the research project**

**Aims and Objectives**

Our aims in this research are to create models which will be able to predict politically unstable or conflict/riot/civil war prone countries using the FSI dataset [2], check for the variables which are responsible for the conflict to happen in a country. Our objectives here include the usage of various Machine Learning algorithms in order to predict politically unstable or civil war prone countries using the FSI dataset. We will be measuring the performance of various machine learning algorithms used and will be choosing the best one for prediction. We will be using K-Means, Time Series Analysis, Random Forest and XGBoost as the main algorithmic approach towards our problem. Model evolution measures used for predicting conflict-based countries includes Accuracy or Detection rate, True positive rate or Sensitivity, True negative rate or Specificity, False positive rate, ROC, Cost and F1-measure.

**Data Explanation**

The FSI data [2] has 12 indicators and we will be using the data from the year 2006 to 2020. The explanation of all the indicators is as follows:



Fig-1: Cohesion Indicators

**COHESION INDICATORS**

**C1: Security Apparatus:** The Security Apparatus Indicator thinks about imbalance inside the economy, regardless of the real presentation of an economy. For instance, the Indicator takes a look at basic imbalance that depends on society, (for example, racial, ethnic, strict, or other personality gathering) or dependent on training, monetary status, or locale, (for example, metropolitan provincial gap).

The Indicator thinks about real imbalance, yet additionally impression of disparity, perceiving that view of monetary imbalance can fuel complaint as much as possible, support shared pressures or nationalistic way of talking. Further to estimating financial disparity, the Indicator additionally accepts into account the open doors for society to advance their monetary position, for example, through admittance to business, instruction, or employment preparing with the end goal that regardless of whether there is financial imbalance present, how much it is basic and fortifying.

**C2: Factionalized Elites:** This indicator thinks about the fracture of state foundations along ethnic, class, group or race just as and brinksmanship and gridlock between administering elites. It additionally factors the utilization of jingoistic radical way of talking by administering elites, frequently as far as patriotism, xenophobia, collective irredentism or of common unity (e.g., "ethnic cleansing" or "safeguarding the religion"). In extraordinary cases, it very well may be illustrative of the nonattendance of authentic initiative broadly acknowledged as speaking to the whole population. This pointer estimates power battles, political rivalry, political advances, and where decisions happen will factor in the validity of discretionary cycles (or in their nonattendance, the apparent authenticity of the decision class).

## C3: Group Grievance: This Indicator centers around divisions and factions between various group of people in the public eye – especially divisions dependent on social or political abilities – and their part in admittance to administrations or assets, and consideration in the political cycle. These groups may likewise have a recorded past, where wronged other groups refer to shameful acts of the past, now and again returning hundreds of years, that impacts and shapes that group's function in the public space and associations with different groups. This set of experiences may thus be formed by examples of genuine or saw atrocities or "violations" submitted with clear exemption against other groups. These groups may likewise feel abused on the grounds that they are denied self-governance, self-assurance or political freedom to which they accept they are entitled. The Indicator additionally looks about where explicit groups are singled out by state specialists, or by prevailing groups, for abuse or suppression, or where there is public accusing of other groups accepted to have gained riches, status or influence "misguidedly", which may show itself in the rise of searing way of talking, for example, through "disdain" radio, pamphleteering, and cliché or nationalistic political discourse.

## 

Fig 2: Economic Indicators

**ECONOMIC INDICATORS**

## E1: Economic Decline and Poverty: This Indicator considers factors identified with monetary decay inside a nation. For instance, the Indicator takes a look at examples of reformist financial decrease of the general public overall as estimated by per capita income, Gross National Product, joblessness rates, swelling, efficiency, obligation, destitution levels, or business disappointments. It additionally considers unexpected drops in product costs, exchange income, or unfamiliar venture, and any breakdown or downgrading of the public cash. This Indicator further looks about the reactions to financial conditions and their results, for example, outrageous social difficulty forced by monetary importance programs, or saw expanding group differences. This Indicator is centered around the proper economy – just as unlawful exchange, including the medication and illegal exploitation, and capital flight, or levels of violation and unlawful exchanges, for example, tax evasion or fraud.

## E2: Uneven Economic Development: This Indicator indicates about imbalance inside the economy, regardless of the real exhibition of an economy. For instance, the Indicator takes a look at auxiliary imbalance that depends on public, (for example, racial, ethnic, strict, or other character gathering) or dependent on training, financial status, or locale, (for example, urban rural gap). The Indicator indicates us about real imbalance, yet in addition view of disparity, perceiving that impression of financial disparity can fuel complaint as much as possible, strengthen shared strains or nationalistic manner of speaking. Further to estimating financial disparity, the Indicator additionally accepts into account the open doors for public to improve their monetary status, for example, through admittance to business, instruction, or occupation preparing with the end goal that regardless of whether there is financial imbalance present, how much it is public oriented and strengthening.

## E3: Human Flight and Brain Drain: This Indicator thinks about the monetary effect of human removal (for financial or political reasons) and the outcomes this may have on a nation's turn of events. From one perspective, this may include the willful resettlement of the working class – especially financially profitable portions of the population, for example, business visionaries, or gifted specialists, for example, doctors – because of monetary disintegration in their nation of origin and the expectation of better open doors farther abroad. Then again, it might include the constrained removal of experts or learned people who are escaping their nation because of real or dreaded oppression or restraint, and explicitly the monetary effect that uprooting may unleash on an economy through the loss of gainful, talented expert work.

## 

## Fig 3: Political Indicators

## POLITICAL INDICATORS

## P1: State Legitimacy: This Indicator considers the representativeness and transparency of government and its relationship with its public. The Indicator takes a look at the populace's degree of trust in state organizations and measures, and surveys the impacts where that certainty is missing, showed through mass public showings, continued common noncompliance, or the ascent of equipped insurgencies. In spite of the fact that the State Legitimacy pointer doesn't really make a judgment on fair administration, it considers the respectability of races where they happen, (for example, boycotted races), the idea of political advances, and where there is a nonattendance of majority rule decisions, how much the legislature is illustrative of the number of inhabitants in which it oversees. The Indicator considers receptiveness of government, explicitly the receptiveness of administering elites to straightforwardness, responsibility and political portrayal, or alternately the degrees of degradation, profiteering, and underestimating, abusing, or in any case barring resistance groups. The Indicator additionally considers the capacity of a state to practice essential capacities that inference a populace's trust in its administration and organizations, for example, through the capacity to gather duties.

## P2: Public Services: This Indicator alludes to the presence of fundamental state works that serve the individuals. From one viewpoint, this may incorporate the arrangement of fundamental administrations, for example, security, education, water and electricity, transport, and internet. Then again, it might incorporate the state's capacity to secure its residents, for example, from psychological warfare and brutality, through saw compelling policing. Further, even where fundamental state capacities and administrations are given, the Indicator further considers to whom – regardless of whether the state barely serves the decision-making elites, for example, security organizations, presidential staff, the national bank, or the appeasing assistance, while neglecting to give equivalent degrees of administration to the overall people, for example, country versus metropolitan populaces. The Indicator likewise considers the level and support of general foundation to the degree that its nonappearance would contrarily influence the nation's real or possible turn of events.

## P3: Human Rights and Rule of Law: This Indicator considers the connection between the state and its public to the extent that principal common liberties are secured and opportunities are monitored and regarded. The Indicator takes a look at whether there is inescapable maltreatment of legitimate, political and social rights, including those of people, groups and establishments (for example badgering of the press, politicization of the legal executive, inward utilization of military for political finishes, suppression of political adversaries). The Indicator likewise looks about flare-ups of politically propelled (rather than criminal) brutality executed against regular folks. It additionally takes a glance at variables, for example, denial of fair treatment reliable with global standards and practices for political detainees or nonconformists, and whether there is current or developing tyrant, oppressive or military guideline in which established and majority rule foundations and cycles are deferred or controlled.

## 

## Fig 4: Social and Cross-Cutting Indicators

## SOCIAL INDICATORS

## S1: Demographic Pressures: This Indicator considers pressures upon the state getting from the public itself or the earth around it. For instance, the Indicator estimates populace pressures identified with food gracefully, admittance to safe water, and other life-continuing assets, or wellbeing, for example, pervasiveness of sickness and pandemics. The Indicator looks about segment qualities, for example, pressures from elite groups development rates or slanted public appropriations or pointedly different paces of public development among competing different groups, perceiving that such impacts can have significant social, financial, and political impacts. Past the public, the Indicator additionally considers pressures originating from catastrophic events (tropical storms, quakes, floods or dry season), and weights upon the populace from ecological dangers.

## S2: Refugees and IDPs: This Indicator gauges the weight upon states brought about by the constrained dislodging of enormous networks because of social, political, natural or different causes, estimating removal inside nations, just as exile streams into others. The marker estimates displaced people by nation of Asylum, perceiving that public inflows can squeeze public administrations, and can in some cases make more extensive compassionate and security encounters for the accepting state, if that state doesn't have the adjustment limit and sufficient assets. The Indicator likewise gauges the Internally Displaced Persons (IDP) and Refugees by nation of starting point, which predicts interior state pressures because of brutality, natural or different factors. These measures are considered inside the setting of the state's general public (per capita) and human growth index, and after some time, perceiving that a few IDPs or exiles for instance, may have been removed for extensive amount of time.

## CROSS-CUTTING INDICATORS

## X1: External Intervention: This Indicator thinks about the impact and effect of outer interveners in the working, especially security and financial system of a state. From one viewpoint, this indicator centers around security parts of commitment from outside interveners, both undercover and unmistakable, in the inner issues of a state in danger by governments, armed forces, insight administrations, other groups, or different elements that may influence the overall influence (or goal of a contention) inside a state. Then again, External Intervention additionally centers around monetary commitment by outside interveners, including multilateral associations, through huge scope advances, improvement ventures, or unfamiliar guide, for example, progressing spending support, control of accounts, or the board of the state's financial strategy, making financial reliance. Outside Intervention likewise considers philanthropic mediation, for example, the organization of a worldwide peacekeeping mission.

## Our Dataset has 178 countries (2020) before that there may be lesser countries as some countries got split for example Sudan got split into Sudan and South Sudan since 2011.

## There is a list of riots [1] which will be used for the supervised learning and also for checking the predictive power of the machine learning method used for the unsupervised learning.

**Research Methodology**

We will be using various techniques of Machine Learning to fulfill our aims and objectives. Those techniques are:

1. K-means
2. Time Series Analysis (if oversampling works)
3. Random Forest
4. XGBoost

**Data Preparation**

Before performing any analysis using the Machine Learning algorithms the data needs to be prepared in a particular format that good results can be obtained using them. For example, the data is supposed to be “Gaussian Like” or its distribution of each variable should be normal in nature. These are the steps which should be performed on the data to create a dataset that should be analysis ready. The steps for data preparation can be represented using the following steps:

* 1. Check the raw data for the number of variables that are numeric or category based.
  2. Separate the numeric and category-based variables.
  3. Check for the distribution of the numeric variables if they are not normally distributed, apply a scalar that can make them, for example: PowerTransformer in python SciKit-Learn
  4. Create Dummy variable for the categorical variables.
  5. Combine all the numeric and dummy variables to make a complete dataset.

In our data we have 15 files from the year 2006 to 2020, some algorithms like K-means will be using the files individually and create the respective datasets on the other hand for time series analysis we will be combining all the datasets from the year 2006 to 2020 and timestamp each row data for further analysis.

Class imbalance will be the problem for the supervised learning and for that oversampling techniques like ADASYN and SMOTE can be used to handle that. In case of Time Series analysis we will not go for class imbalance as it about projecting the future data only. Predictions will be done by Random Forest and XGBoost. Although if oversampling of Time Series will be required, we may use Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher [6].

After completing the above steps our data is ready for analysis and we can use the Machine Learning algorithms further for our research.

**K-Means**: Hierarchical clustering (hierarchical cluster analysis or HCA) is a strategy for bunch examination which tries to manufacture a hierarchy of clusters. Here the countries are separate entities till we apply the algorithm and start combining them in the respective clusters. We will be trying the bottom up approach also known as the agglomerative clustering where in the end we get the dendrograms and we get to know in which category the country actually falls. We might be experimenting with the different values of cluster for example 2 for countries: conflict free and conflict countries, 3: for developed, developing and under-developed countries. Later on, we can check about the countries which are developed / developing and still have chances of conflict as the under-developed countries have high chances of conflict as they are poor, low on education and healthcare, they also are the victims of poor governance. With every year change in the cluster label of country we can check for the instability occurring in it. This is also going to be an unsupervised way of learning the conflict occurring countries. There won’t be any sort of prediction in this approach, it is more about grouping the countries and checking on the basis of data how well they are performing. The countries differentiated as the conflicted ones can be compared from the list of Riots.

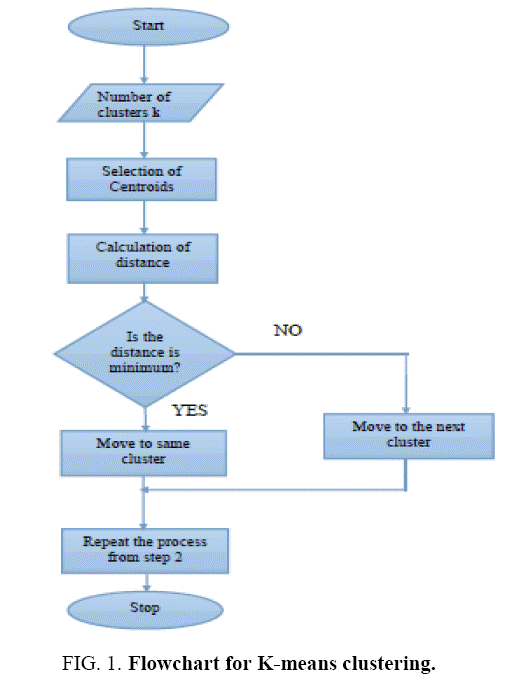


Fig 5: K-Means Algorithm on Flow Chart

**Time Series**: Time series analysis contains techniques for examining time series data so as to extricate significant insights and different attributes of the information. Time series forecasting is the utilization of model to foresee future qualities dependent on previously observed qualities. We will take the data with respect to each and every country from 2006-2020. So, every country will have 15 rows of data on which we can firstly perform the Exploratory Data Analysis and try to convert them to Time Series. If 15 data points are going to be really less for the forecasting part, we can try to oversample them using oversampling techniques [6]. Once we have the required number of samples for each country, we will try to make a model by splitting the data in train and test and see how well it performs for each country. This will be more over a kind of analysis and forecasting the future of a country exercise. If the time series graphs will be trending upwards that means the things are going to be in a bad shape and this is our novel approach.

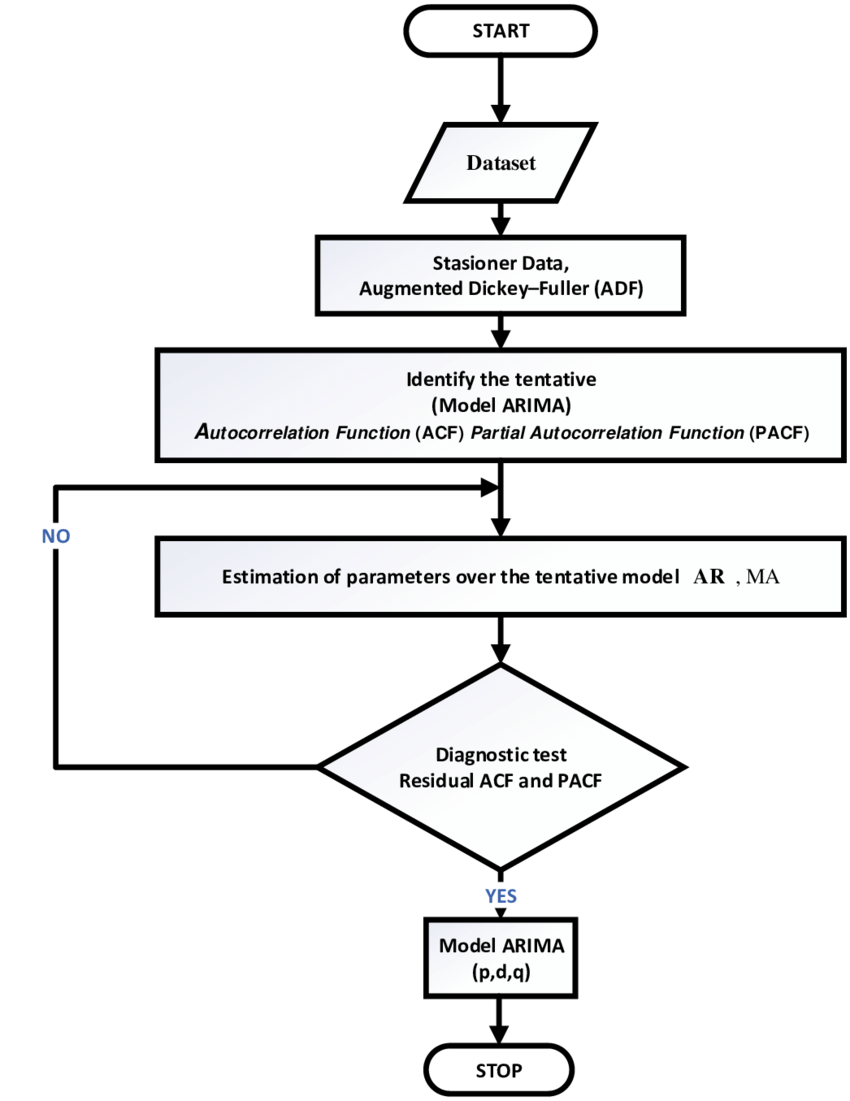


Fig 6: Time Series Forecasting Algorithm Flow Chart

**Random Forest**: Random Forests or random decision forests are an ensemble learning strategy for classification, regression and different assignments that work by developing a large number of decision trees at preparing time and yielding the class that is the method of the classes (classification) or mean prediction (regression) of the individual trees. Here for this problem we will be training the model using the supervised way, we will label the countries where riots/conflict has occurred using the list from Wikipedia [1]. Then, we will try to predict outcomes on the test set for just 3 years say 2017 to 2020. Then we will calculate all the performance matrices using the confusion matrix for the efficiency of the model.

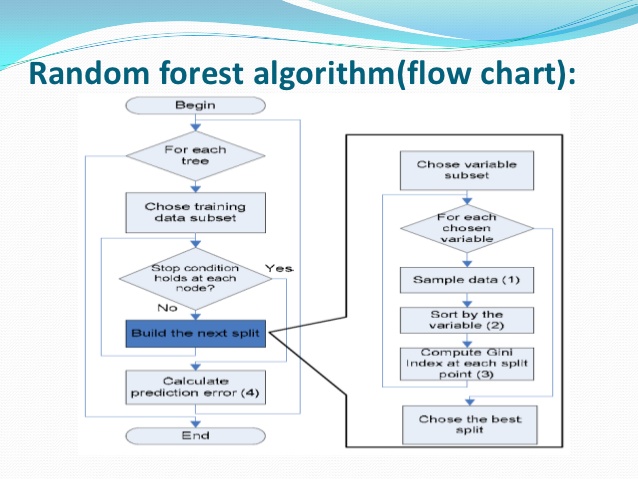
****

Fig 7: Random Forest Algorithm Flow Chart

**XGBoost:** XGBoost is a popular machine learning algorithm and is widely used for classification problems. XGBoost stands for Extreme Gradient Boosting, so far it has outperformed all the other algorithms representing statistical machine learning (not the neural networks). We will be using this similar to the Random Forest for our problem.

**Outcome**

**K-means**: We will be using this unsupervised learning algorithm to create the clusters of the countries which are prone to riot/conflict. We can check the algorithm performance by matching it with the real data (riot data[1]).

**Time Series**: Once we get the model which can predict the future and performs well on the test data, we can get data points for the future and later we predict using those points if a conflict can happen or not using Random Forest or XGBoost.

**Random Forest**: After creating the model we will be checking it on the confusion matrix criterion.

**XGBoost**: Similar to Random Forest we will be checking its performance on confusion matrix criterion.

**Risks or contingency plan**

I might skip one approach in the methodology if the time is limited. Time Series is going to be a cumbersome and highly time-consuming exercise, if there is less time, I might have to skip that part.

**Resource Requirements**

A machine with i5/i7 processor, 16 GB RAM and 4 GB graphics card is sufficient for this type of machine learning research. The data size is not that huge for a requirement of an array of graphics card. The resources mentioned above would be sufficient. In software and utilities part the requirement is windows 10 / Linux (Ubuntu latest version) operating system with Anaconda 3 installed on it. The input data is the 15 years of data from website of fragile state index and it is from 2006 -2020.