UNSUPERVISED APPROACHES FOR DETECTING AND FORECASTING FRAGILE COUNTRIES

Rachit Dev

A thesis submitted in fulfillment of the
requirements for the award of the degree of
MASTER OF SCIENCE IN DATA SCIENCE

LIVERPOOL JOHN MOORES UNIVERSITY (LJMU)

February 2021

TABLE OF CONTENTS

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. One approach is to cluster the countries of the whole world and segregate the fragile to the stable ones. The other approach is the more intensive to a particular country i.e., the time series analysis of the indicators which define the stability or fragility of a country. If a particular country is falling behind in those indicators it means its fragility is on the rise and the country may break down in near future. From this research we expect that we can project the internal situation of a particular country by time series analysis. This study can be helpful for the citizens worldwide to know the performance of their country in various indicators and can be useful for the governments if they would like to know the reality and update themselves for betterment of their country and its citizens. The citizens can use this research as factual way to change their governments if it is not performing well on these indicators.

LIST OF FIGURES

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

w.r.t…… with respect to

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.

1.1 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 done by (Blair Huffman et al., 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” is used. It’s a comprehensive list of major riots that has happened across the world in each and every country. Countries which have political dictatorship like Russia, China, North Korea etc. where news of such events doesn’t come out (rsf.org, 2021) but still, we may have some glimpse of agitation and death count which is recorded in this list.

1.2 Problem Statement

We are currently living in a society which is extremely polarized. Whether it's about the polarization among the religions in India (carnegieendowment.org, 2020), where Hindu supremacy is on the rise since 2014 or the polarization in United Kingdom where Brexit or No Brexit debate (reuters.com, 2020) polarized the whole country or the closed / open economy debate that polarized America (pewresearch.org, 2020). 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 it's politically or socially or economically. Our research will also try to forecast the fragility of a country on such data using Facebook's Prophet Library in Python.

1.3 Aim and Objectives

Our aim in this research is to create models which will be able to predict socio-political-economically unstable or conflict/riot/civil war prone countries using the FSI dataset. Also, to check for the variables which are responsible for the conflict to happen in a country and to forecast the fragile/unstable countries.

Our objectives in this research are:

- The usage of clustering machine learning algorithms and time series with Facebook's Prophet 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 Facebook's Prophet.

1.4 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 Facebook's Prophet can forecast the fragility of a country fed with time series data?

1.5 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 Facebook's Prophet 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 India, its economy and the other factors contributing to its declination in recent past.

1.6 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 or maybe indirectly involved in the declining events for their own personal benefits, then it's a matter of time for that country to be ruined.

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.

1.7 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 Prophet. Section 2.2.4 presents the discussion about the difference between the ARIMA and Prophet 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 Prophet, 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 4 discusses about the implementation with the use of python programming and its machine learning libraries. Section 4.1 discusses about the introduction and how implementation was done in brief. Section 4.2 discusses about the dataset preparation for the 2 clustering algorithms as well as validation data set and time series dataset. Section 4.3

discusses about the clustering algorithms implementation in detail. Section 4.4 discusses about the implementation of time series data for a particular country using Facebook's prophet library. Section 4.5 summarizes about the whole chapter and gives a small introduction for the next chapter.

Chapter 5 is based on the results and evaluation of this whole study. Section 5.1 discusses about the introduction to the results and evaluation. Section 5.2 is based on the exploratory data analysis of the clustering algorithms. Section 5.3 discusses on the exploratory data analysis of the time series of individual indicators for the country India. Section 5.4 discusses about the output produced by the clustering model. Section 5.5 discusses about the Prophet model output on the time series data in other words it's the forecasting of the individual indicators of the country India. Section 5.6 summarizes the whole chapter 5.

Chapter 6 is based on the conclusions and recommendations based on this study. Section 6.1 is the introduction to the chapter. Section 6.2 discusses and concludes the whole study. Section 6.3 discusses about the contribution of this study to the whole society and to the world. Section 6.4 discusses about the future works of this study and how this study can be taken forward further.

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; Coyne and Mathers, 2011) hadn't used any kind of data science but had only presented the modelling framework, their work was more of theories presentation. The research conducted by (Blair Huffman et al., 2016) is the base paper we are using in this research. Other researches by (Atin Basuchaudhary et al., 2018, 2019) are 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 this research the FSI data is used which was previously used by (Blair Huffman et al., 2016). Their research was mostly based on predicting the conflict or civil war in the countries while in our research we are not just 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 et al., 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 et al. (2018) 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 et al. 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. In analysis it was 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 gave us better confusion matrix parameters or other ratios should be used in the future too.

2.2.3 List of Riots (Wikipedia, 2020) and why Riots can be called as a main point of a fragile country?

The various social, economic and political factors constitute in building of the unrest of a society. Riots is one measure where the unrest among the people is very high due to the frustration of lying low in terms of social, economic and political factors. Although riots can be induced by the political leaders too where they are being brainwashed continuously stating that the other group of people is the culprit behind their distress. Although those particular problems also arise due to the lack of education, money and lifestyle. Lack of such things are also being accounted in our FSI dataset by FFP. So, riots can be considered as a good measure of fragility. From the list of riots, we are creating a dataset of past events to measure the performance of our clustering algorithm.

2.2.4 Time Series with Facebook's Prophet

Previously (Weytjens et al., 2019) used forecasting tools like ARIMA, LSTM and Prophet to predict the cash flow. Although LSTM outperformed ARIMA as well as Prophet in forecasting the cash flow but we have used Prophet only because it's easy to use and can be applied in a short span of time.

2.2.5 Why Facebook's Prophet and not ARIMA (Time Series methodology)?

(Yenidoğan et al., 2018) used Prophet and ARIMA and compared their results in their research on Bitcoin's value prediction. Prophet outperformed ARIMA. Prophet was accurate with more than 95% accuracy while ARIMA was accurate up to 68% only.

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 et al. (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 prophet library in order to forecast the fragility of a country. Traditional time series approaches have been outperformed by the Facebook's Prophet library 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:

a. K-means and Hierarchical Clustering
b. 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 (Fund for Peace, 2020) 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 (Wikipedia, 2020) 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 scale the data normally so that analysis and prediction can be done properly. Machine Learning predictions work better on normally distributed data.

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).

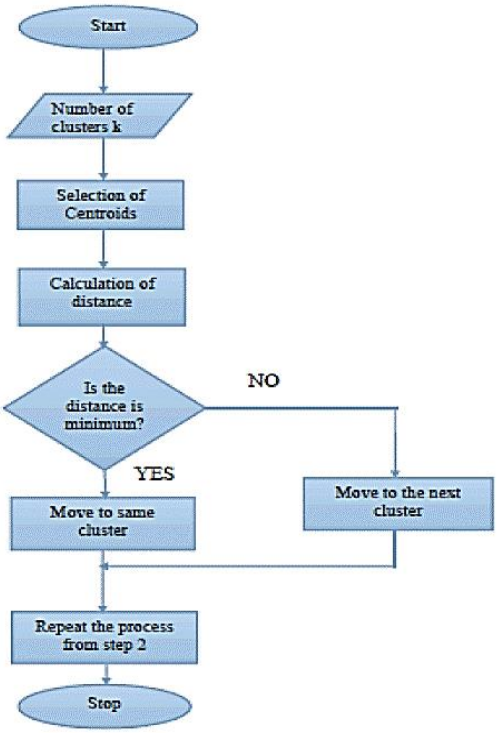

Figure 3.5: K-Means Algorithm Flow Chart (Palamera et al., 2011)

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).

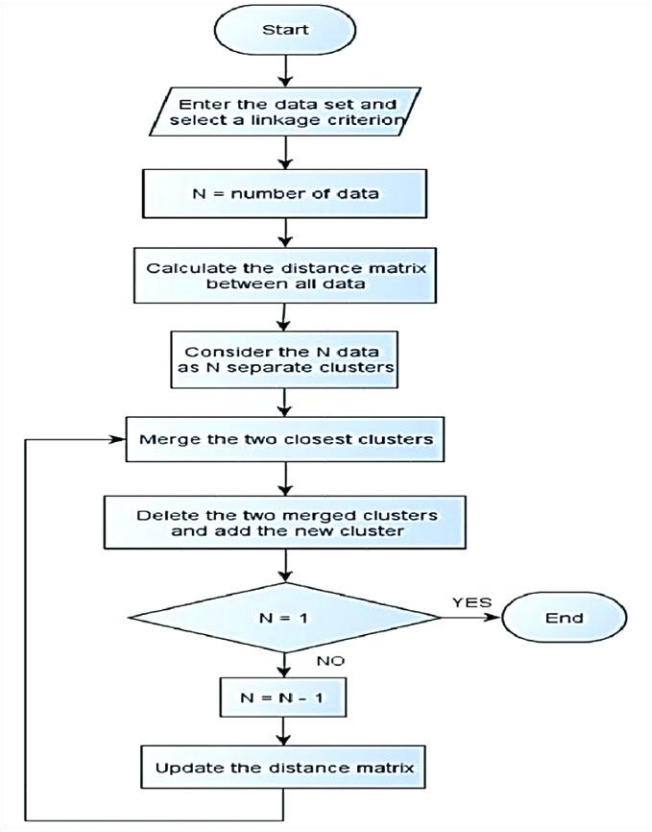

Figure 3.6. Hierarchical Clustering Flow Chart (Palamera et al., 2011)

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.

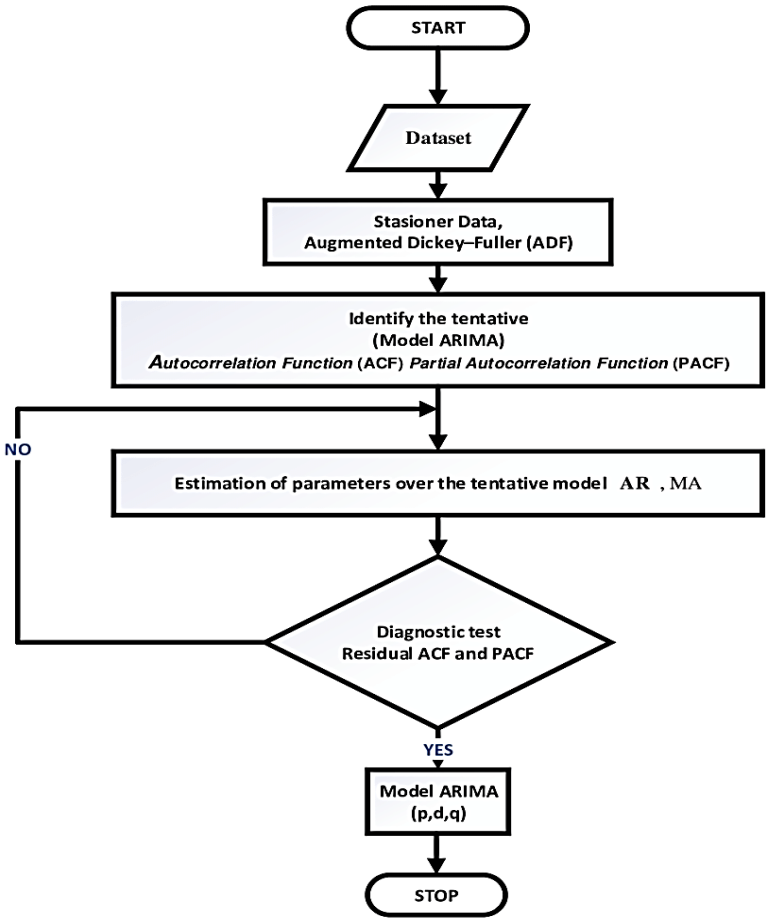

Figure 3.7: Time Series Forecasting Flow Chart (Palamera et al., 2011)

3.1.7 Facebook's Prophet

At the point when an estimating model doesn't run as arranged, we need to have the option to tune the parameters of the technique with respect to the particular issue within reach. Tuning these models requires an intensive comprehension of how the basic time arrangement models work. The originally input parameter to auto-ARIMA, for example, are the maximum orders of the differencing, the auto-regressive components, and the moving average segments. A typical examiner won't realize how to change these requests to dodge the conduct and this is the sort of ability that is difficult to get and scale.

The Prophet bundle gives intuitive parameters which are not difficult to tune. Indeed, even somebody who needs aptitude in determining models can utilize this to make significant forecasts for an assortment of issues in a business situation. The Prophet bundle gives intuitive parameters which are not difficult to tune. Indeed, even somebody who needs aptitude in determining models can utilize this to make significant forecasts for an assortment of issues in a business situation.

Prophet's Forecasting Model

We use a time series model which has three main components: trend, seasonality, and holidays. They are all represented in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \qquad (3.1)$$

- **g(t)**: Component for the piecewise linear or logistic growth curve for modelling non-periodic changes in time series.
- **s(t)**: periodic changes (for example: weekly/yearly seasonality).
- **h(t)**: effects of non-working days (user provided) with irregular schedules.
- **ϵ(t)**: error term accounts for any unusual changes not accommodated by the model.

Utilizing time as a regressor, Prophet is attempting to fit a few straight and nonlinear elements of time as parts. Demonstrating irregularity as an added substance part is a similar methodology taken by outstanding smoothing in Holt-Winters procedure. We are, as a result, outlining the anticipating issue as a curve-fitting activity instead of taking a gander at the time sensitive reliance of every perception inside a time series.

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 3.8: Data Preparation and Preprocessing

3.2.2 Validation Data Preparation

4. Grab the data for the year 2006 to 2020 from the list of riots (Wikipedia 2020).
5. 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.

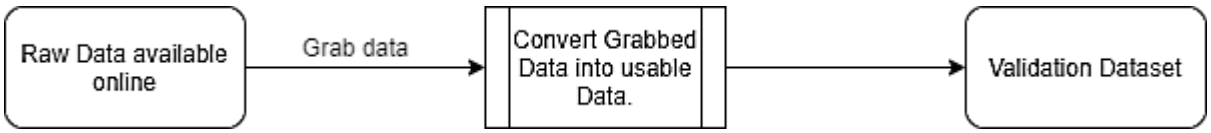

Figure 3.9: Validation Data Preparation

3.2.3 K-Means Data Modelling:

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

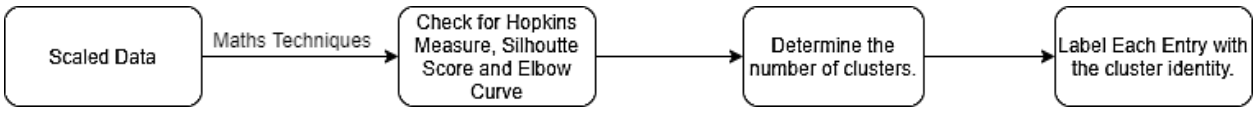

Figure 3.10: Determination of Number of clusters

3.2.4 K-Means Validation and Confusion Matrix

9. 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.
10. We will get a lot of nan values in the actual column of the combined dataset we can fill them with values 0.
11. 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.

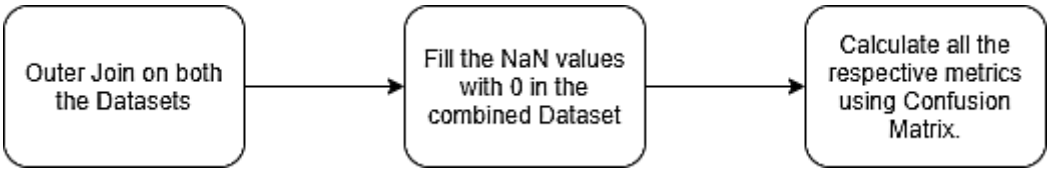

Figure 3.11: Validation of the K- Means Model

3.2.5 Hierarchical Approach

12. 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.
13. Then we will mark the respective countries with the cluster ids.

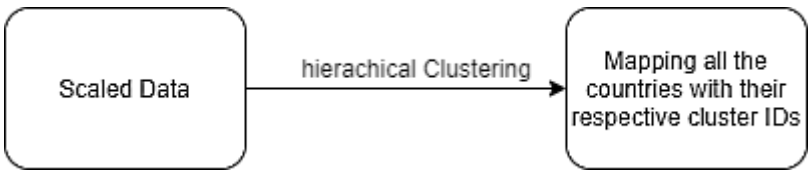

Figure 3.12: Hierarchical Clustering

3.2.6 Hierarchical Approach Validation and Confusion Matrix

14. 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.
15. We will get a lot of nan values in the actual column of the combined dataset we can fill them with values 0.
16. 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.

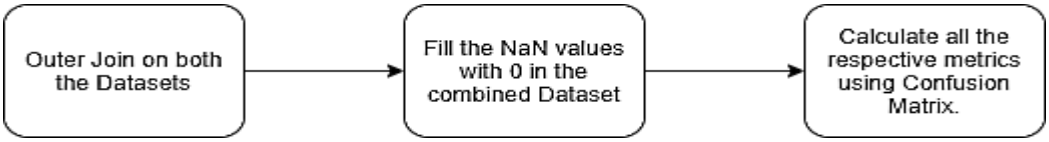

Figure 3.13: Validation of the Hierarchical Model

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 3.14: Block Diagram to generate Datasets based on countries

3. Then we pick up one country's dataset and sort the rows according to the years 2006 to 2020.

4. 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.
5. 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.

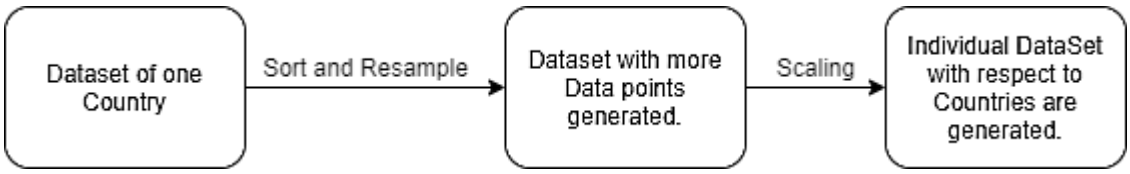

Fig 3.15: Block Diagram to prepare an individual Dataset of a country

3.3.2 Facebook's Prophet Approach

6. Create a Prophet 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.
7. Provide the future time values so that the Prophet model can predict the values of the respective columns on those time values.
8. Generate the future data points on the given future time values for all the columns
9. Use the inverse scaler to get the unscaled data for the whole country's dataset.
10. 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

11. Plot the total column on the time series.
12. If the trend of the plot goes up the country is destabilizing and it its going down the is getting strengthened / stabilizing.

Fig 3.16: Block Diagram of Prophet 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 Prophet: 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 Prophet 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 Prophet will give us the trend towards the fragility of a particular country.

CHAPTER 4

IMPLEMENTATION

4.1 Introduction

Fund for Peace (FFP) has provided us the fragility data of the respective countries on the yearly basis. In other words, we have Microsoft excel sheets on the yearly basis which contains the fragile state data. In order to use that data in clustering algorithm as well as time series Prophet based forecasting, we have to prepare it according to the format on which we are doing our research. If we do not prepare our data well the algorithms might not work or may provide erroneous results. Data preparation is one of the most important steps to perform any research.

We are using 2 cluster-based implementations i.e., K-means and Hierarchical clustering. Both of them have a little bit of difference in implementation and results are also a bit different, we have compared the results for both of them later.

In case of time series forecasting the data has to be timestamped. So, the timestamped data is being used to project the future values of the variables of the country on which the research is being performed.

4.2 Dataset Preparation

There are different kind of data to be given as input to the different type of models. For example, timestamped data is a requirement for the time series analysis but in case of clustering algorithms it's not a requirement. Let's dive into the various kind of data preparations for the particular algorithms / models.

4.2.1 Clustering Dataset Preparation

The list of files which contains the yearly data of fragile state index for most of the countries across the world is our data for clustering. Our first step is to combine that data together. We do it in python by creating the Pandas dataset of all the excel files and then combining them together. The below mentioned points with code snapshots prove the way it is done. All the code is there in appendix B.

1. **Importing the data:**

```
In [1]: import pandas as pd

In [2]: df2006 = pd.read_excel('fsi-2006.xlsx', engine='openpyxl')
        df2007 = pd.read_excel('fsi-2007.xlsx', engine='openpyxl')
        df2008 = pd.read_excel('fsi-2008.xlsx', engine='openpyxl')
        df2009 = pd.read_excel('fsi-2009.xlsx', engine='openpyxl')
        df2010 = pd.read_excel('fsi-2010.xlsx', engine='openpyxl')
        df2011 = pd.read_excel('fsi-2011.xlsx', engine='openpyxl')
        df2012 = pd.read_excel('fsi-2012.xlsx', engine='openpyxl')
        df2013 = pd.read_excel('fsi-2013.xlsx', engine='openpyxl')
        df2014 = pd.read_excel('fsi-2014.xlsx', engine='openpyxl')
        df2015 = pd.read_excel('fsi-2015.xlsx', engine='openpyxl')
        df2016 = pd.read_excel('fsi-2016.xlsx', engine='openpyxl')
        df2017 = pd.read_excel('fsi-2017.xlsx', engine='openpyxl')
        df2018 = pd.read_excel('fsi-2018.xlsx', engine='openpyxl')
        df2019 = pd.read_excel('fsi-2019.xlsx', engine='openpyxl')
        df2020 = pd.read_excel('fsi-2020.xlsx', engine='openpyxl')
```

Figure 4.1: Data import

Here, Pandas library has been imported and the yearly data frames of the respective files have been created.

2. **Cleaning the data:** All the unwanted columns from the data frames which are of no use are removed or deleted.

```
In [3]: del df2019['Change from Previous Year']
        del df2019['Unnamed: 17']
        del df2019['Unnamed: 18']
        del df2020['Change from Previous Year']
        del df2020['Unnamed: 17']
        del df2020['Unnamed: 18']
        del df2020['Unnamed: 19']
        del df2020['Unnamed: 20']
        del df2020['Unnamed: 21']
        del df2020['Unnamed: 22']
        del df2020['Unnamed: 23']
        del df2020['Unnamed: 24']
        del df2020['Unnamed: 25']
        del df2020['Unnamed: 26']
        del df2020['Unnamed: 27']
        del df2020['Unnamed: 28']
        del df2020['Unnamed: 29']
        del df2020['Unnamed: 30']
        del df2020['Unnamed: 31']
        del df2020['Unnamed: 32']
        del df2020['Unnamed: 33']
        del df2020['Unnamed: 34']
        del df2020['Unnamed: 35']
        del df2020['Unnamed: 36']
        del df2020['Unnamed: 37']
        del df2020['Unnamed: 38']
```

Figure 4.2: Removing unwanted columns

3. **Combining the data:** The data is now cleaned. The next step is to combine the data.

```
In [20]: dataframes = [df2006, df2007, df2008, df2009, df2010, df2011, df2012, df2013, df2014, df2015, df2016,
                       df2017, df2018, df2019, df2020]

In [21]: fsi6_20_df = pd.concat(dataframes)

In [22]: fsi6_20_df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2633 entries, 0 to 177
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Country                         2633 non-null   object
 1   Year                            2633 non-null   datetime64[ns]
 2   Rank                            2633 non-null   object
 3   Total                           2633 non-null   float64
 4   C1: Security Apparatus          2633 non-null   float64
 5   C2: Factionalized Elites        2633 non-null   float64
 6   C3: Group Grievance             2633 non-null   float64
 7   E1: Economy                     2633 non-null   float64
 8   E2: Economic Inequality         2633 non-null   float64
 9   E3: Human Flight and Brain Drain 2633 non-null  float64
 10  P1: State Legitimacy            2633 non-null   float64
 11  P2: Public Services             2633 non-null   float64
 12  P3: Human Rights                2633 non-null   float64
 13  S1: Demographic Pressures       2633 non-null   float64
 14  S2: Refugees and IDPs           2633 non-null   float64
 15  X1: External Intervention       2633 non-null   float64
dtypes: datetime64[ns](1), float64(13), object(2)
memory usage: 349.7+ KB
```

Figure 4.3: Concatenated Data and its information

In the above-mentioned snippet, a list of all the year-based data frames has been created and in the next step concatenation of all the data frames is completed. At last, the data information is printed, the data contains 16 columns and 2633 rows.

4. **Plotting the heat map:** The heat map which is one of the steps in exploratory data analysis, heat map is created to check how much the variables are connected to each other. Although all the variables are being used in our predictions but we should check the heat map.

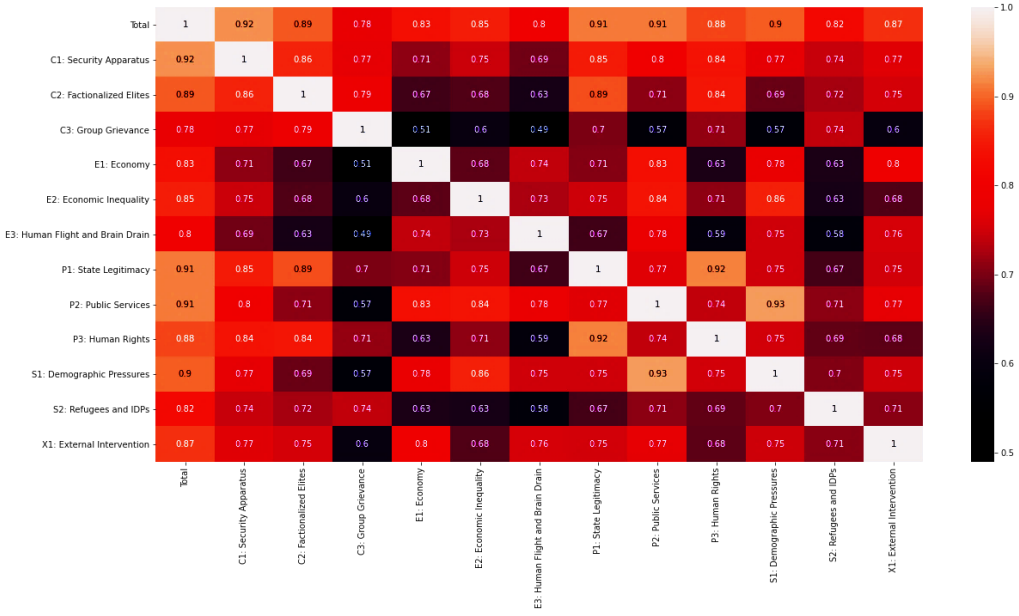

Figure 4.4: All variables Heat Map

The total column is nothing but the total of all the other columns for a particular column and we can drop it in our further steps.

5. **Combining country and year column and dropping the unwanted columns:** Year and country columns are combined using python command and then the removal of unwanted columns (not required for analysis) is performed. The result is as per the below mentioned snapshots.

Country_Year
sudan_2006
congo democratic republic_2006
cote d'ivoire_2006
iraq_2006
zimbabwe_2006

Figure 4.5: Joined Countries with Years

Country and year columns are combined together so the removal of the individual columns is performed. Total column is just the combined value of all the other columns and Rank is not required for our analysis hence, both of them are removed.

6. **Scaling using PowerTransformer:** The dataset contains only the positive values, the box-cox method to transform the data has to be applied. Only numeric data is transformed using this scalar.

```
In [51]: col = df_cluster.select_dtypes(include=['float64']).columns
         col

Out[51]: Index(['C1: Security Apparatus', 'C2: Factionalized Elites',
                'C3: Group Grievance', 'E1: Economy', 'E2: Economic Inequality',
                'E3: Human Flight and Brain Drain', 'P1: State Legitimacy',
                'P2: Public Services', 'P3: Human Rights', 'S1: Demographic Pressures',
                'S2: Refugees and IDPs', 'X1: External Intervention'],
               dtype='object')

In [52]: scaler = PowerTransformer(method = 'box-cox')
         df_cluster[col] = scaler.fit_transform(df_cluster[col])
```

Figure 4.6: Scaling of Data

Now, the data is ready to be applied on the clustering algorithms.

4.2.2 Wikipedia Dataset Preparation

The article called as the list of riots on Wikipedia contains our validation data. To extract this data from the webpage, spacy, Wikipedia, pandas and regular expressions library of python has been used. The whole code is attached in the appendix C of this report. The important steps are being discussed below:

1. **Grabbing the data from Wikipedia.org:** By using the Wikipedia API for Python, grabbing of the text data contained in the article is possible. Grabbing of the data using the API is done and is saved in a variable. The data is split according to the lines.
2. **Cleaning the data:** Cleaning of data is performed by checking the format of the article and in the near future if the format of the article changes then the code will require updating.
3. **Extracting Countries and Year information from the data using Natural Language Processing (NLP):** en_core_web_sm library is used to extract the keywords from the text data available to us. Extraction of year and country names is done using this library.
4. **Cleaning the text data after extraction using regular expressions:** Regular expressions library of python is used to clean the data of all the columns. There are characters like "[], ()" which comes along when using the python library en_core_web_sm.
5. **Manually filling the missing values after excel file generation:** There are missing values left in the columns of country if a country name is not found in the related text. There is a reference to the article in the text. We have to manually open the article to find the related countries with respect to that particular riot and update the excel sheet.

4.2.3 Time Series Dataset Preparation

1st to 4th points as per the stanza 4.2.1 are repeated in this section too. The whole code is attached in the appendix D of this report. After those procedure steps are completed, the next steps are as follows (Starting with number 5 as 1-4 steps are common):

5. **Creating a list of all the countries existing from 2006 to 2020:** Some new countries are formed in 15 years of time span, some gets renamed. A list is created using the python command which contains all the unique names of all the countries in past as well present in a 15-year time span.
6. **Creating a dictionary with respect to country names as keys and its values as the data frame of all variables:** A large dictionary is created where the keys are the country names and the values are the data frames related to them with each and every year. From this dictionary all the variables data can be accessed using the country name as a key of that particular country. This technique is very useful for app building related to this concept.
7. **Creating the data frame of a particular country and setting the year values as index:** In this step a new data frame is created to save a particular country's data with respect to the yearly information. The index of that data frame is the year values.

Now, the data is ready for forecasting.

4.3 Clustering Algorithm Implementations

As the dataset for implementation for clustering algorithms is now ready, we will continue our analysis using 2 clustering algorithms i.e., K-means and Hierarchical (Agglomerative).

4.3.1 Pre-Requisites for K-means Clustering

Hopkins measure and elbow curves are the basic requirements of the K-means clustering approach. Appendix B contains all the code for how these things are applied.

Hopkins Statistic Measure: It is a way to measure the cluster inclination of a dataset. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

In our analysis the dataset scored 0.88455 on Hopkins measure, so this dataset is cluster able.

Silhouette score: Silhouette score provides us the number of clusters of a particular dataset. This is done by plotting a graph between the score and the number of clusters identified by

the algorithm. In our case, the score reached its peak on 2 clusters. Now those 2 clusters can be identified as fragile and stable countries.

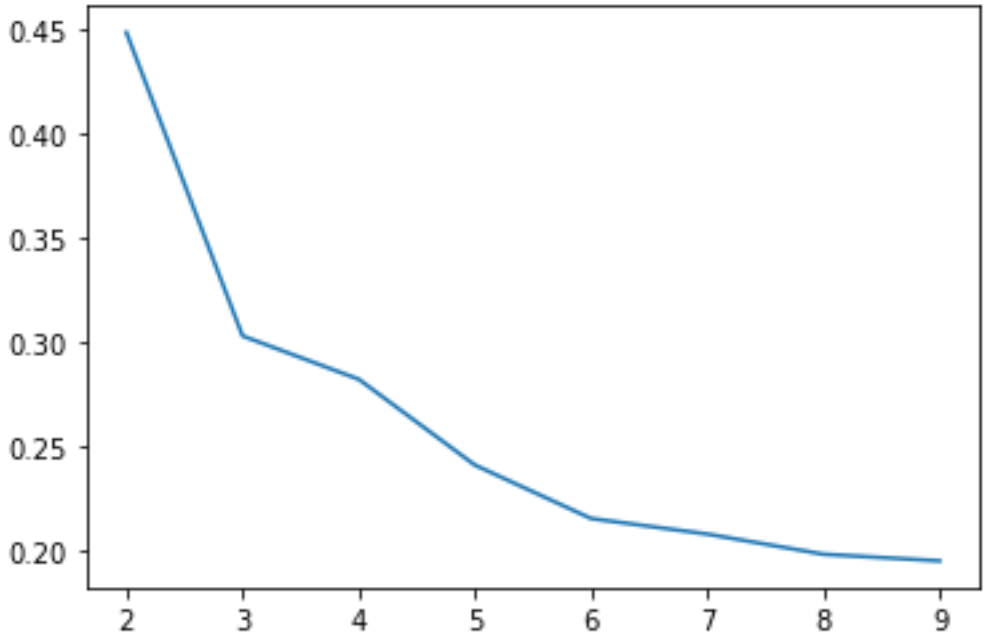

Figure 4.7: Silhouette Score reaches its peak on 2 clusters

Elbow Method: It's a heuristic method to determine the number of clusters in a dataset. In our analysis the number of clusters found were 2.

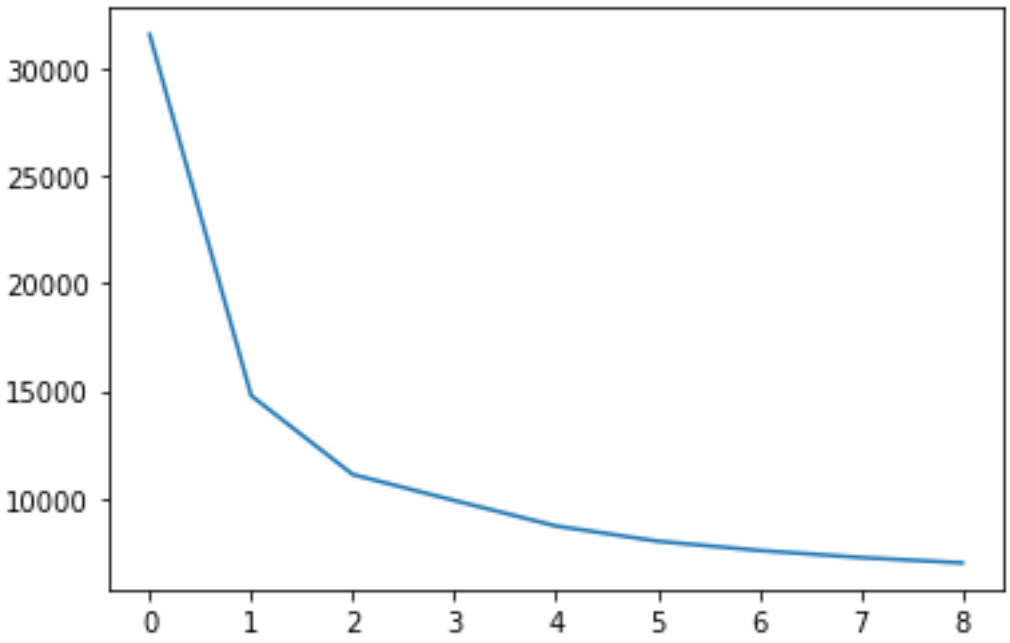

Figure 4.8: The elbow curve smoothens after 2 value.

This curve shows that after 2 value of cluster the curve goes smoothly so the optimal value for the clusters is 2.

4.3.2 K-means clustering

We fit our dataset in the K-means clustering algorithm with 50 iterations and cluster value 2 (as suggested by the above-mentioned methods). The count of the clustered countries w.r.t year (Country_Year) are, for cluster ID 0 it is 1805 and for cluster ID 1 it is 828.

Number of fragile countries are more so cluster 0 is assigned to the fragile countries and 1 is assigned to the stable countries. More results will be discussed in the results section of this report where the data from list of riots (Wikipedia, 2020) has been used for the validation.

4.3.3 Pre-Requisites for Hierarchical Clustering

Complete Linkage Method: This method is used to determine the number of clusters. While applying this method the number of clusters were found to be 2. These clusters can be assigned as the stable and fragile countries.

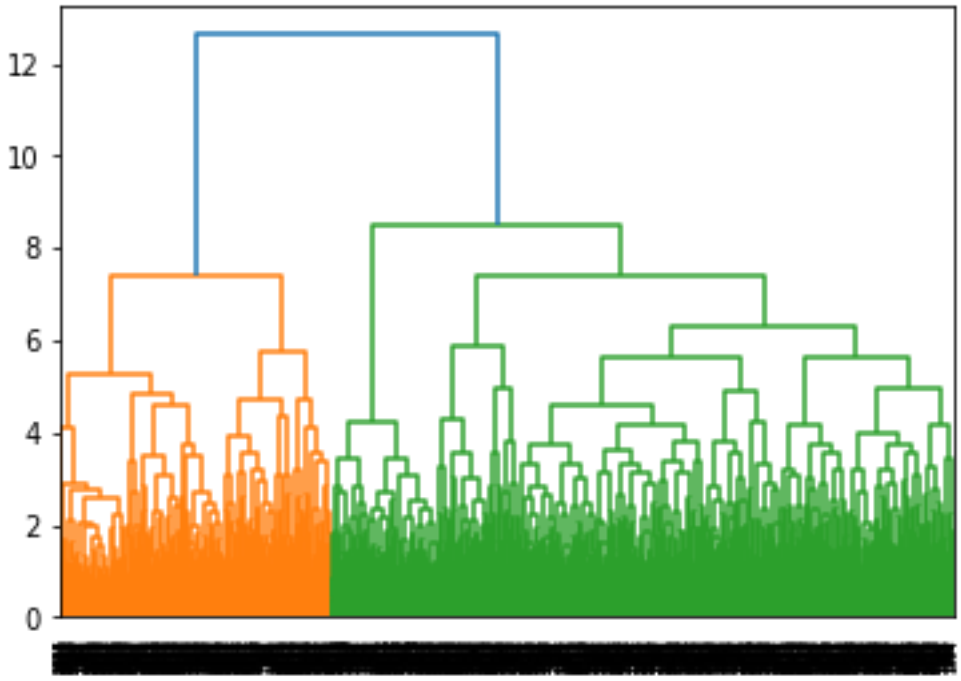

Figure 4.9: Complete Linkage Method clearly shows 2 clusters

4.3.4 Hierarchical Clustering

We fit our dataset in the Hierarchical clustering algorithm. The count of the clustered countries w.r.t year (Country_Year) are, for cluster ID 0 it is 1830 and for cluster ID 1 it is 803.

Number of fragile countries are more so cluster 0 is assigned to the fragile countries and 1 is assigned to the stable countries. More results will be discussed in the results section of this report.

4.4 Forecasting with Prophet, Implementation

The time series data set is ready for implementation. Here the country is India, which we have chosen for analysis. The whole code for this particular exercise can be found in appendix D of this report. The step-by-step procedure is as follows:

1. **Create a dictionary for all the countries**: Our first step is to create a dictionary where the keys are of unique countries and the values is the dataset of all the years with respect to that country. Once we have that dictionary, we can call a country data from that dictionary to perform time series analysis on all the variables.
2. **Setting the index on time and plotting the respective variables**: The index of this dataset is set as value of the year and then the graphs are plotted. If the trend of a particular variable is going upwards that means the situation in the domain of a country is on a decline and if the trend is going downwards then the situation is getting better of a country. More about this is discussed in the results section
3. **Using Facebook's Prophet for projection**: For projecting the values of the variables in the future we have to create the datasets of each and every variable with respect to time. Then we have to rename those columns as ds and y, this is because the column names are hardcoded in prophet library. Then we fit the respective datasets in the prophet model. We have to create the future placeholders in other words the time length on which the projections are to be done. We are taking them as 5 years. So, for the years 2021-25 the projections will be made by the prophet model. At last, we project the variables on future placeholders and the projection is done for all the variables. The projections results are discussed more in the results section of this report.

4.5 Summary

In this chapter we discussed about the implementation of our thesis. This is the main chapter which involved lot of effort in terms of coding and time. From dataset preparation for different algorithms to getting the results was tedious task. This part of thesis was the most cumbersome in some sense as many methods were applied and failed and only those which worked were finally kept. The appendices B, C and D contains all the code which helps us in determining the fragility/stability of a country/s. We have our results ready, lets jump to chapter 5 to discuss in details about how to interpret the results, which clustering algorithm performed better and how the projection of variables of India are doing.

CHAPTER 5

RESULTS AND EVALUATION

5.1 Introduction

As the implementation part of the research is complete. The interpretation of results on the basis of research questions and some exploratory data analysis is being discussed in this section related to clustering as well as time series forecasting. The comparison between 2 clustering algorithms is also included here.

5.2 Exploratory Data Analysis (Clustering)

Once the clusters are created, we try to plot the graphs of the countries with respect to the various variables. Some were tried in this research to get insights:

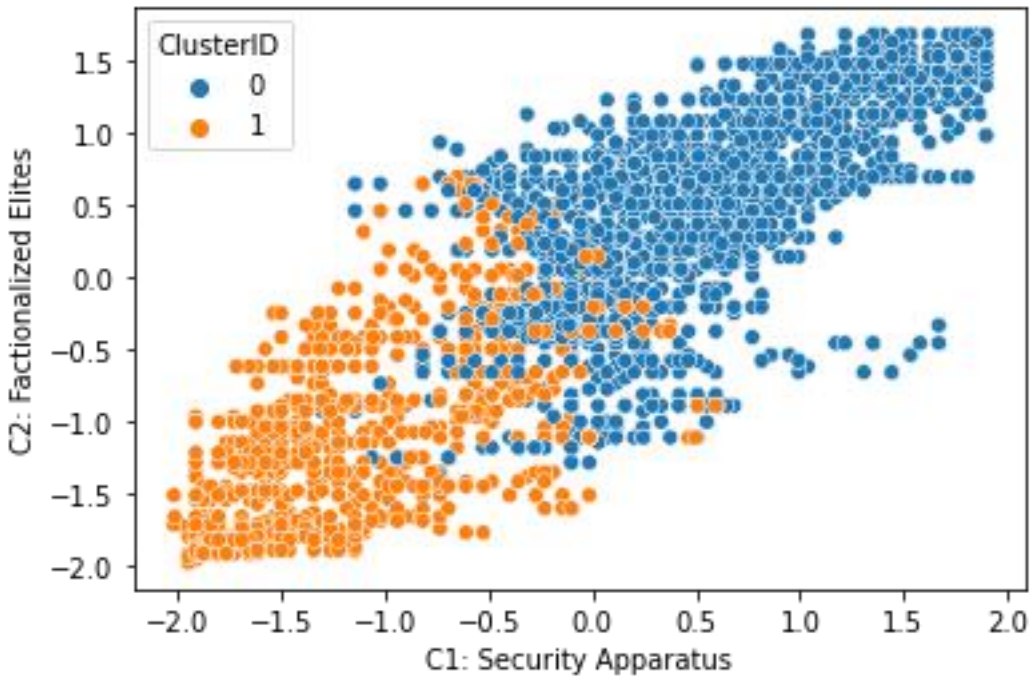

Figure 5.1: Plot of cluster IDs on C1 and C2 indicators

In the Figure 5.1 the comparison on Factionalized Elites and Security Apparatus is done. We can clearly see that the fragile countries represent the 0 cluster(blue). Higher the numbers for security apparatus, higher goes the factionalized elites. Some blue points can be seen in the orange cluster that means those particular countries might have higher values in the other variables that is why we can some overlapping. Countries which are being plot on the low values are highly stable countries.

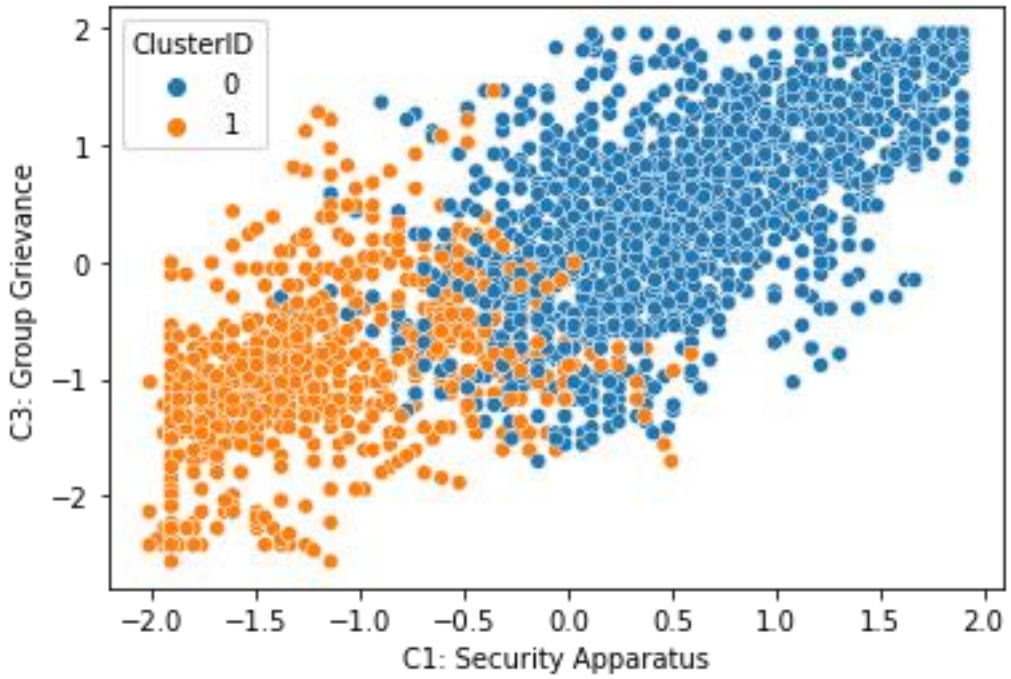

Figure 5.2: Plot of cluster IDs on C1 and C3 indicators

In the figure 5.2 the comparison on Group Grievance and Security Apparatus is done. We can clearly see that the fragile countries represent the 0 cluster(blue). Higher the numbers for security apparatus, higher goes the Group Grievance. Some blue points can be seen in the orange cluster (1) that means those particular countries might have higher values in the other variables that is why we can some overlapping.

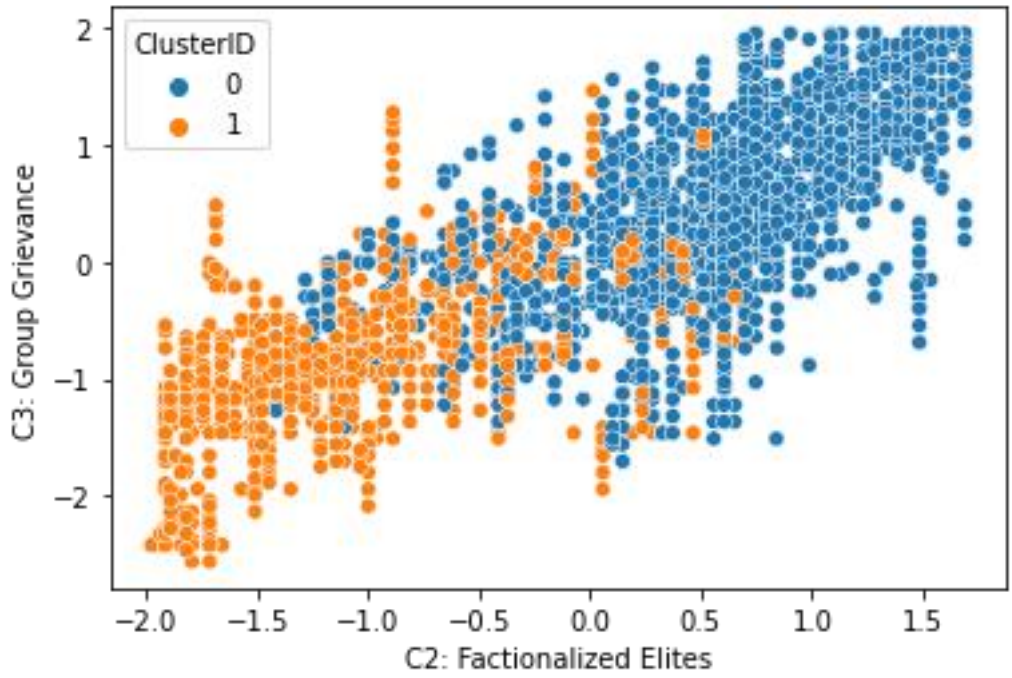

Figure 5.3: Plot of cluster IDs on C2 and C3 indicators

In the figure 5.3 the comparison on Group Grievance and Factionalized elites is done. We can clearly see that the fragile countries represent the 0 cluster(blue). Higher the numbers for Factionalized elites, higher goes the Group Grievance. Some blue points can be seen in the orange cluster (1) that means those particular countries might have higher values in the other variables that is why we can some overlapping. In some stable countries the group grievance problem is quite evident for example the Maori Problem in New Zealand, the native people of New Zealand called as Maoris are still outclassed by the European settlement.

5.3 Exploratory Data Analysis (Time Series)

Before projecting the values, we perform the time series analysis on the FSI data, keeping all the values of a particular country and time as the index. For country India it came out to be below mentioned.

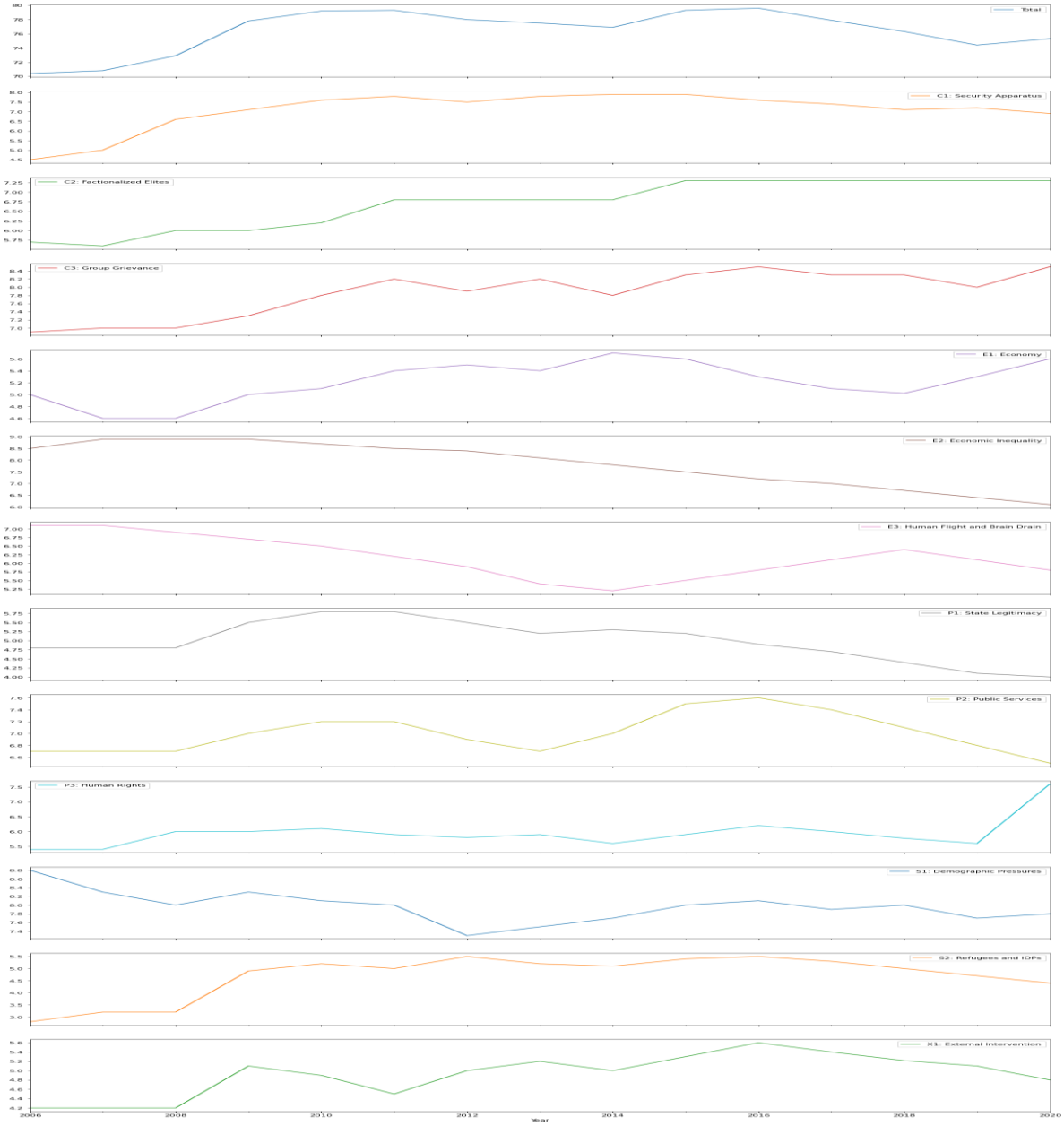

Figure 5.4: Time Series Analysis of all the indicators for India

All the variables according to their current trend is being discussed below:

1. **Total**: Clearly, we can see that current trend is going upwards, this means that current condition of the country is not going in good direction. As this variable is the total of all the other variables, the degradation of it means that the fragility of the country is on the high.
2. **C1: Security Apparatus**: The trend is almost straight that means things are neither going in good way or the bad way according to this particular variable.

3. **C2: Factionalized Elites**: In the recent past things were bad enough and now this variable is consistent. Elite factions are still there and are consistent.
4. **C3: Group Grievance**: In this variable too, the trend is going upwards, it means that the things are deuterating.
5. **E1: Economy**: This is one of the major factors of any country and clearly its worsening. The job losses of a country occurs when this factor deuterate. The trend also says that its continuously worsening which is not a good sign.
6. **E2: Economic Inequality**: According to recent trend, India is doing better in this particular variable although according to the recent Oxfam report the country's 73% wealth has gone into the hands of 1% of the population, which proves that somewhat FSI data needs tweaking or updating.
7. **E3: Human Flight and Brain Drain**: The recent trend shows that intelligent or the highly qualified people are not leaving the country as before, it maybe due to the loss of affordability of foreign citizenship due to the rise in the academic fee of the professional colleges.
8. **P1: State Legitimacy**: This variable is doing good according to the recent trend.
9. **P2: Public Services**: This variable is doing good according to the recent trend.
10. **P3: Human Rights**: This variable is completely on the verge of being destroyed. It is one of the major factors to reside in a country. A country with bad human rights condition will mean that the value of human life is not the first priority in its society.
11. **S1: Demographic Pressures**: This variable is slightly worsening according to the recent trends.
12. **S2: Refugees and IDPs**: This variable is doing better according to the recent trend and it's a good sign for the refugees to enter into India.
13. **X1: External Intervention**: India is doing good in this variable although according to the recent news the Chinese intervention in India is taking place, maybe the latest data of 2021 will reflect that too.

Overall, the condition of the country India is actually deuterating.

5.4 Clustering Model Output

The verification of the clusters is being done by the List of Riots page data from Wikipedia. The entries from the Wikipedia page are being treated as the actual values while the values coming from the cluster algorithms is being treated as the predicted values. Our approach was to create a confusion matrix for both the algorithms and calculate their accuracies. Homogeneity, Completeness, V-Measure as well as Fowlkes-Mallows scores were also calculated in order to check the performance metrics of our analysis.

Accuracies of Clustering Algorithms:

```
                precision    recall  f1-score   support

           0.0       0.00      0.00      0.00         0
           1.0       1.00      0.70      0.82       368

      accuracy                           0.70       368
     macro avg       0.50      0.35      0.41       368
  weighted avg       1.00      0.70      0.82       368

[[  0   0]
 [111 257]]
Accuracy Score on k-means:   0.6983695652173914
```

The accuracy score of K-means is found to be 0.6984

```
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         0
         1.0       1.00      0.70      0.83       368

    accuracy                           0.70       368
   macro avg       0.50      0.35      0.41       368
weighted avg       1.00      0.70      0.83       368

[[  0   0]
 [109 259]]
Accuracy Score on the Hierarchical set:   0.7038043478260869
```

The accuracy Score of the Hierarchical clustering is found to be 0.704

Both the algorithms gave pretty similar accuracy results but Hierarchical clustering gave the result slightly better.

Homogeneity: Each cluster contains members of a single class.

Completeness: All members of a given class are assigned to the same cluster.

V-Measure: The harmonic mean of the above 2 scores.

Although the homogeneity score of both algorithms came out to be perfect i.e., 1 but the completeness and V-measure scores were bad i.e., 0 or nearly 0. This is also because the actual class has only positive values (1) and no negative value i.e., 0.

Fowlkes-Mallows scores: It's a tool to measure similarity between 2 clusters. This tool is being used here to measure the similarity between the outputs of both the algorithms. The scores are follows:

Fowlkes-Mallows scores

```
In [124…  metrics.fowlkes_mallows_score(labels_true, labels_predKM)
Out[124…  0.7599690882171529
In [125…  metrics.fowlkes_mallows_score(labels_true, labels_predHC)
Out[125…  0.762847548298654
```

The results of both the algorithms are very similar according to the Fowlkes-Mallows scores.

5.5 Facebook's Prophet Model Output

Each and every variable is projected with the Prophet model, a discussion on the forecast of each and every variable is being done in this section. Few points relating to the socio-politico-economics impact on a country is also being discussed here with respect to each and every variable.

1. Total: It's a dependent variable and is the total of all the other variables being used for the analysis. The projection and trend plotted by the Facebook's Prophet library is below mentioned:

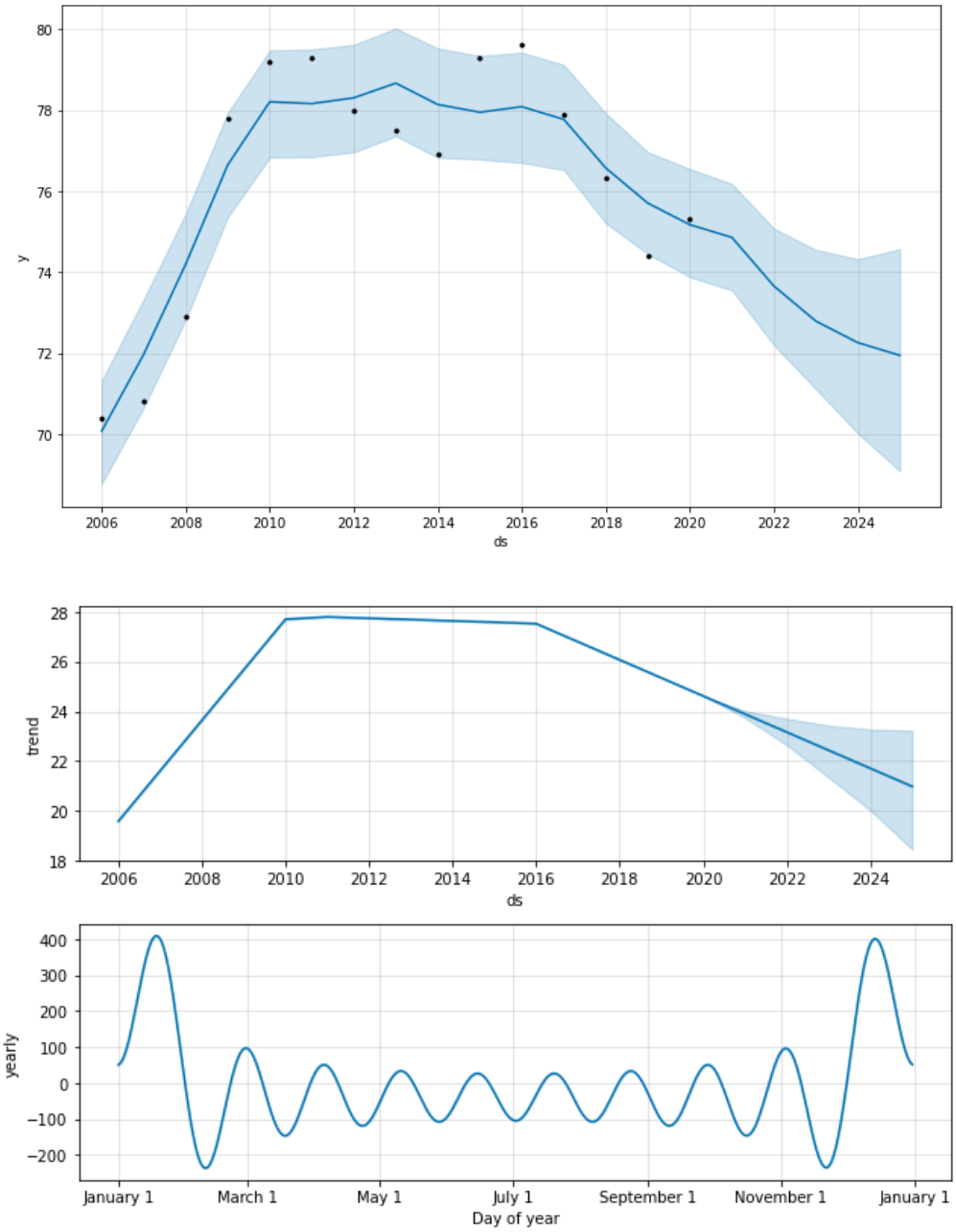

Figure 5.5: Plot and Trend of "Total"

From the figure 5.5, clearly the trend shows that it's the betterment of the country being projected in the near future by this model. But There is problem in this projection i.e., its trend shows that the overall performance of the country is doing good but the other variables on which this variable is dependent are mostly deuterating. This problem can be mitigated and will be discussed more in the future works section.

2. C1: Security Apparatus:

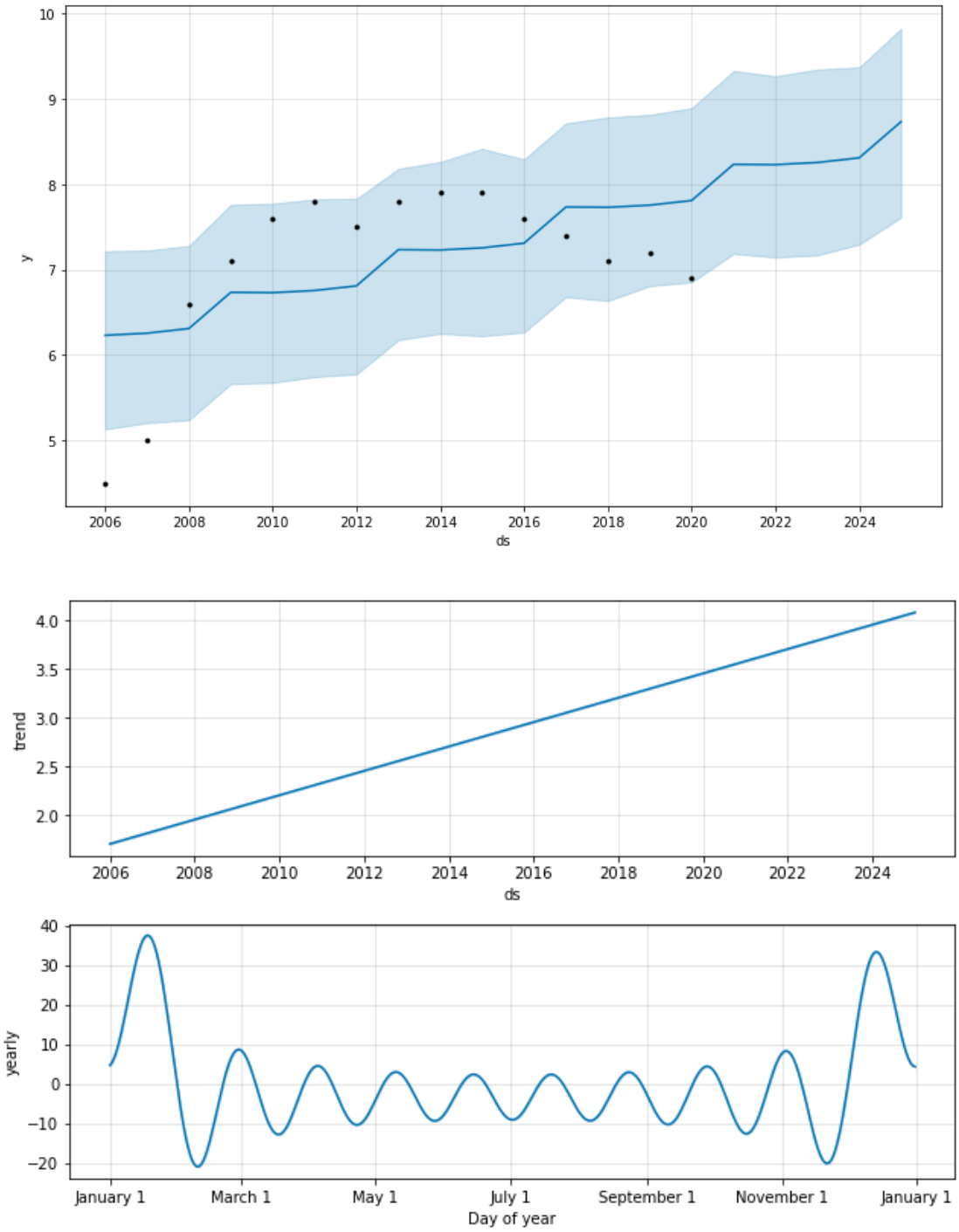

Figure 5.6: Plot and Trend of C1 indicator

From the figure 5.6, clearly the trend is going upwards. It means that the variable and the condition of the country associated with it is on a decline. Also, as per the graph it is clear that the declination of this variable is continuous since 2006.

3. C2: Factionalized Elites:

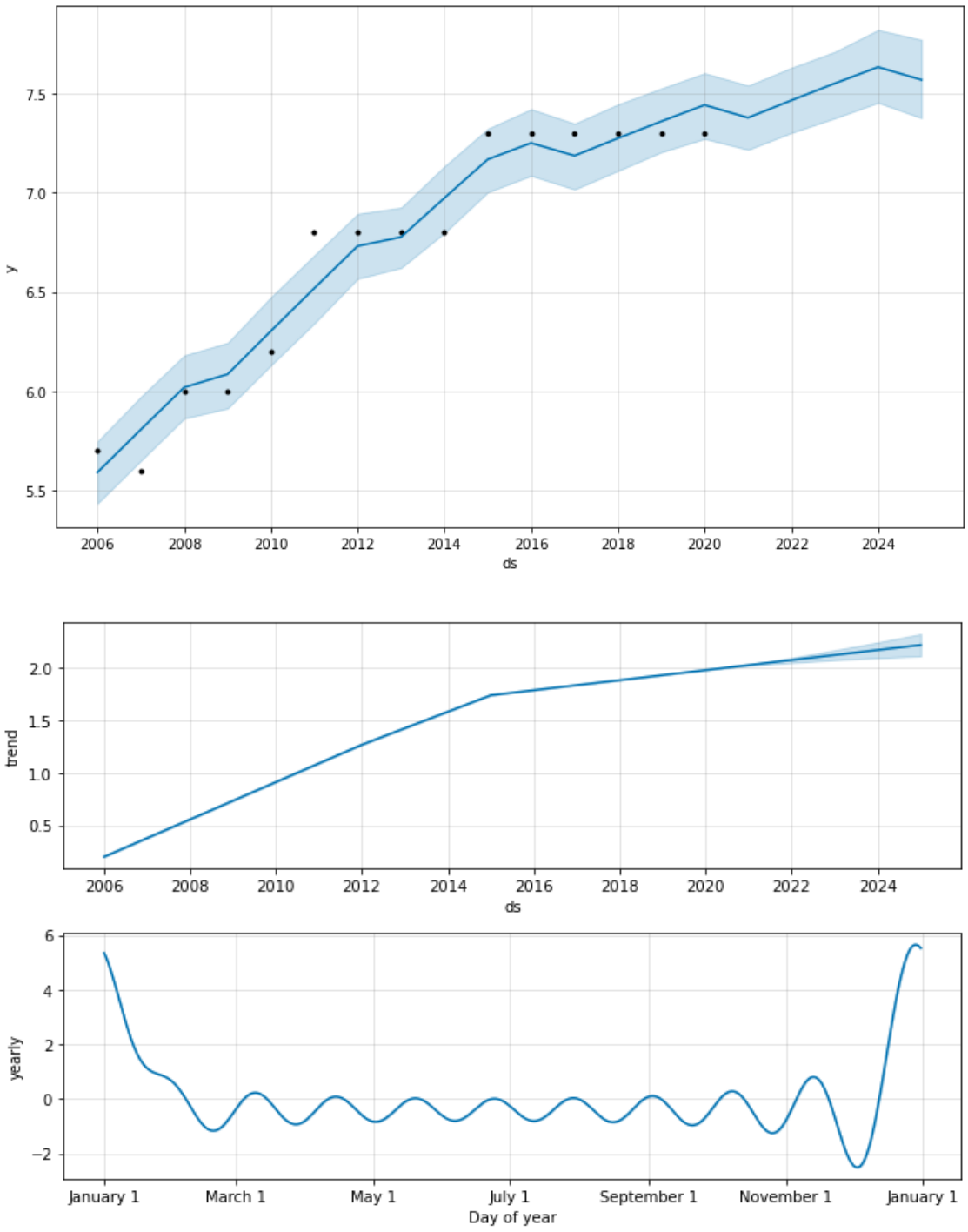

Figure 5.7: Plot and Trend of C2 indicator

From the figure 5.7, clearly the trend is going upwards. It means that the variable and the condition of the country associated with it is on a decline. The variable might start become better after 2024.

4. C3: Group Grievance:

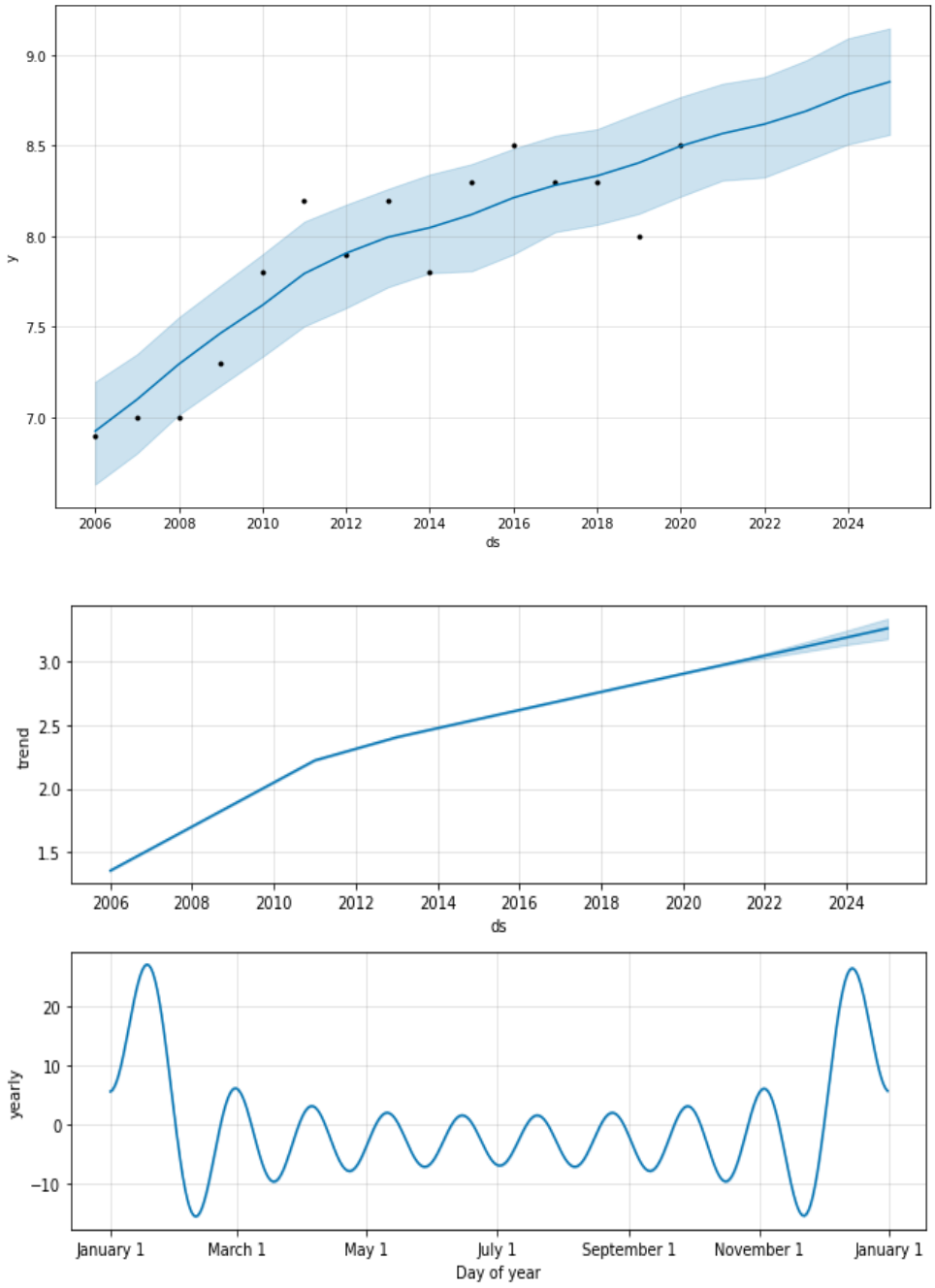

Figure 5.8: Plot and Trend of C3 indicator

From the figure 5.8, clearly the trend is going upwards. It means that the variable and the condition of the country associated with it is on a decline. Even the projections are providing no chance of the things getting better for this variable.

5. E1: Economy:

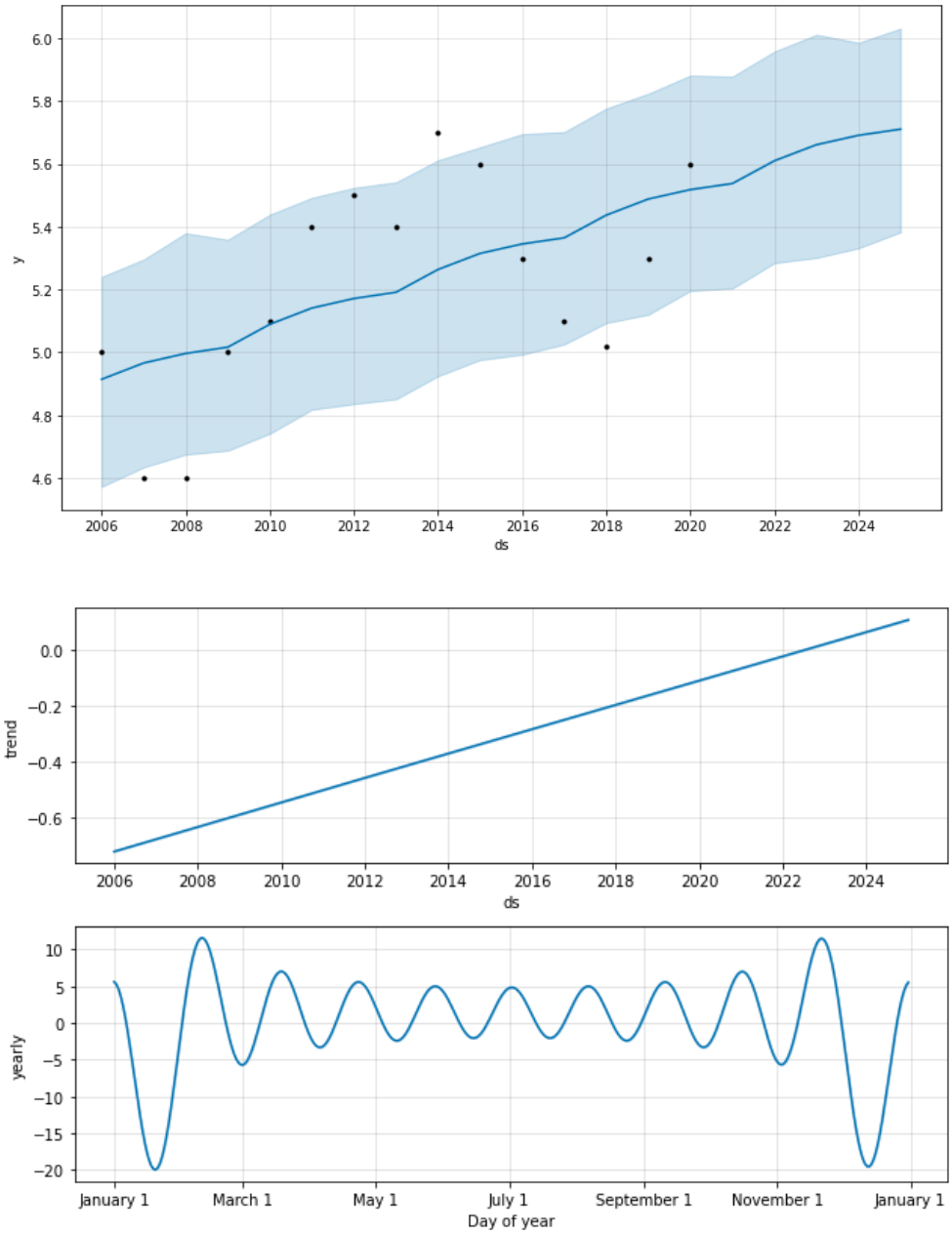

Figure 5.9: Plot and Trend of E1 indicator

This is one of the main criteria of a country the economy. Continuous decline has been seen in the past trends from the figure 5.9 and it's still going to further decline in the near future. No prediction of getting better is seen.

6. E2: Economic Inequality:

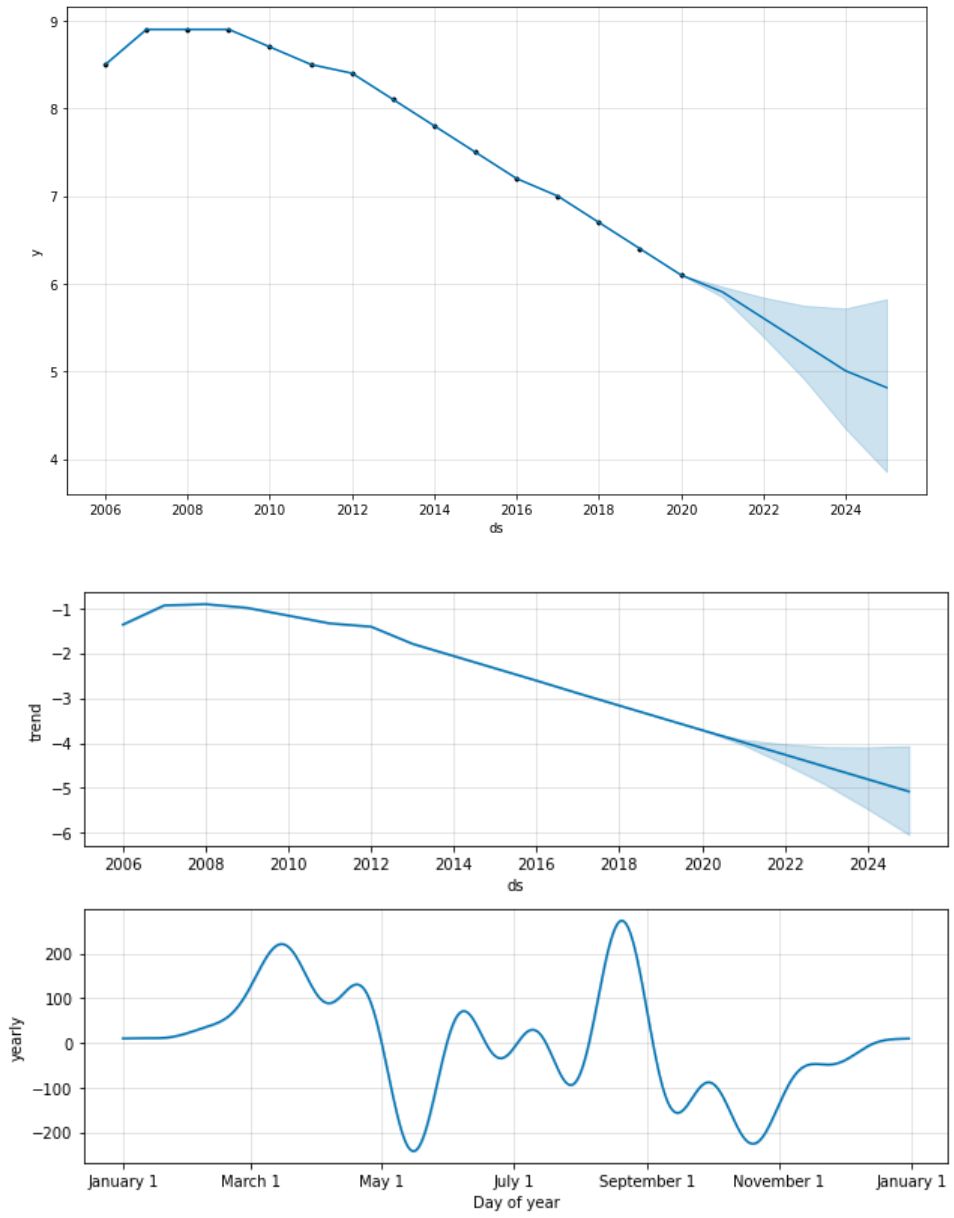

Figure 5.10: Plot and Trend of E2 indicator

Although according to the figure 5.10 it is being projected that the country will be doing better in this variable but as discussed previously according to the recent Oxfam report (Oxfam.org, 2018) on economic inequality amongst Indian is that the 73% wealth of the whole country has gone in to the hands of 1% of the top rich people. That's one of the contrasts of this report and FSI needs to update their data on this particular indicator.

7. E3: Human Flight and Brain Drain:

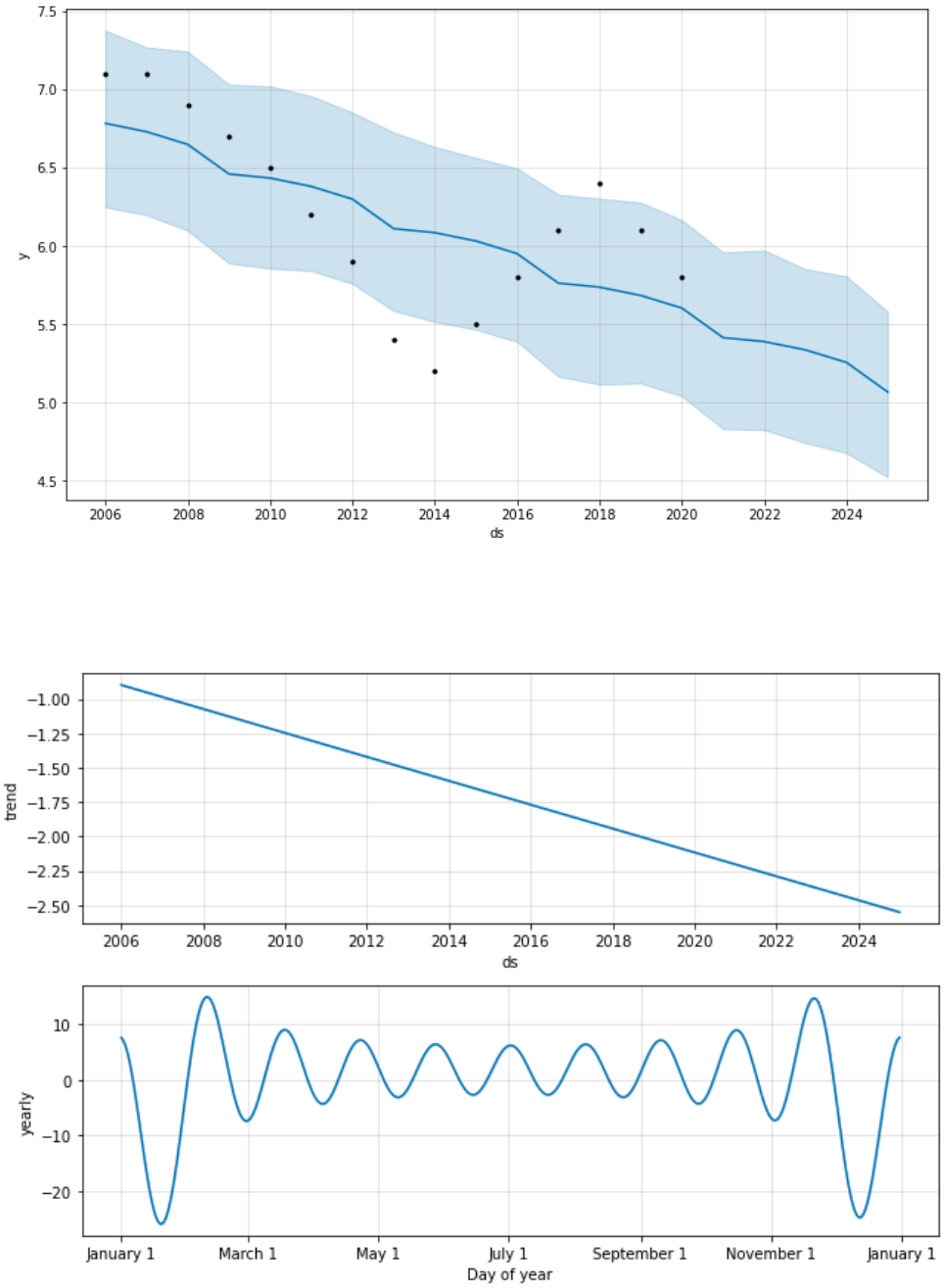

Figure 5.11: Plot and Trend of E3 indicator

We can clearly see that the trend is going downwards from the figure 5.11. It means that the country is doing good in this indicator as lesser number of people are going to leave the country recently and in near future. Although according to recent media reports of Indian news agencies (Business Standard, 2019) it was being told that super wealthy people especially businessmen are leaving the country.

8. P1: State Legitimacy:

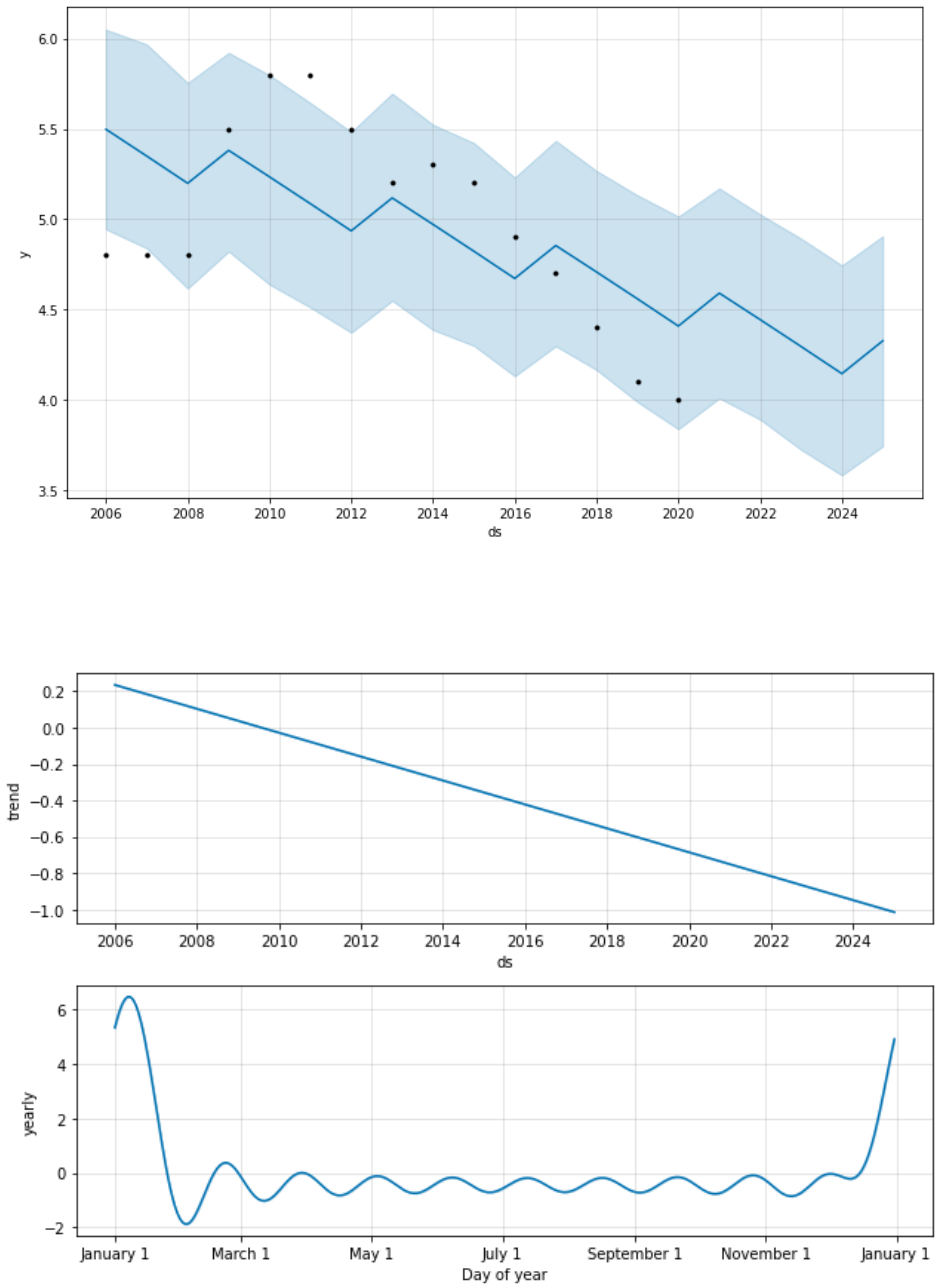

Figure 5.12: Plot and Trend of P1 indicator

The state legitimacy is going down according to the data points and will be doing better in near future as interpreted from the figure 5.12

9. P2: Public Services:

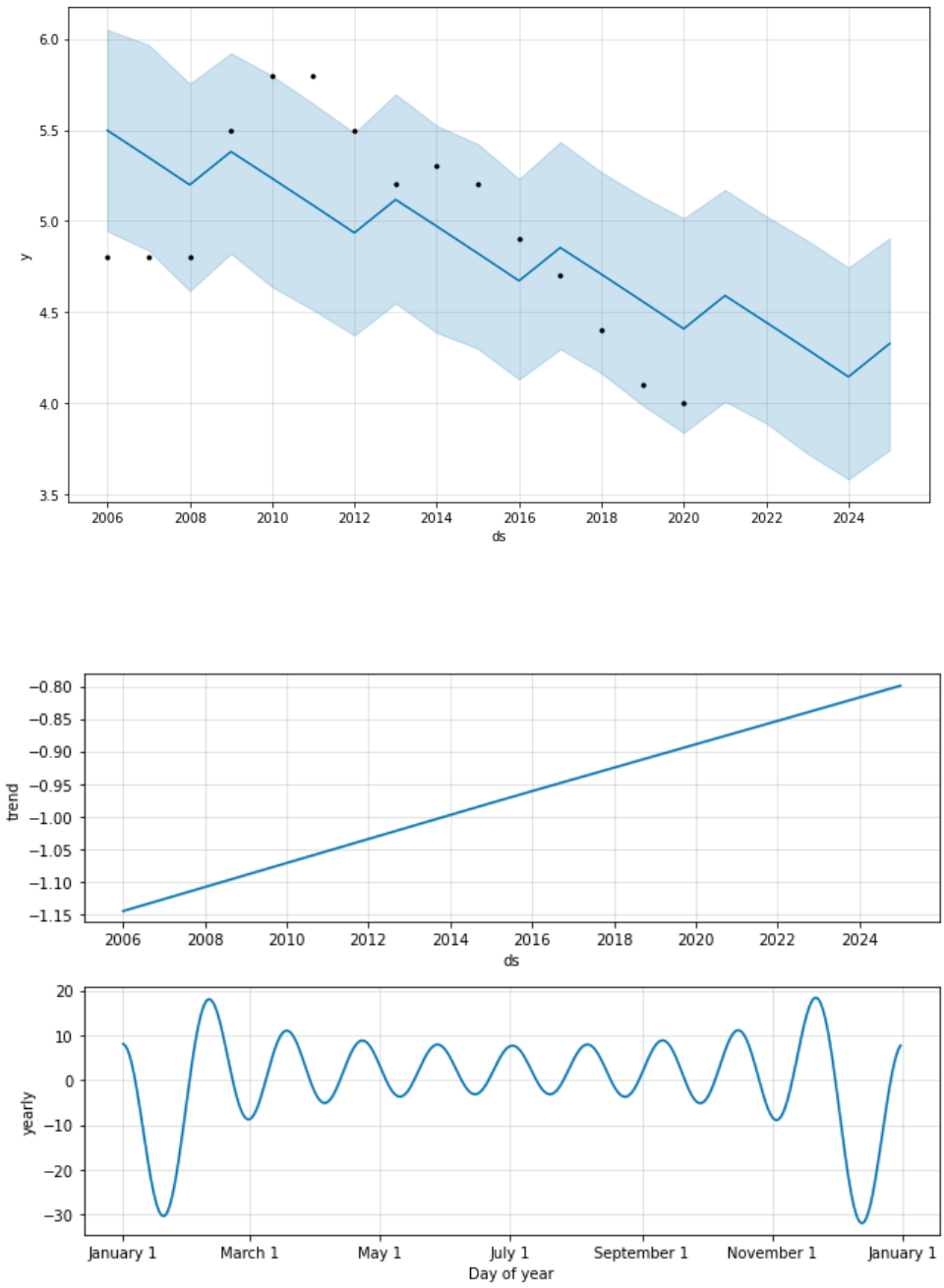

Figure 5.13: Plot and Trend of P2 indicator

The public services are going down according to the trend in figure 5.13 and will be on a decline in near future.

10. P3: Human Rights:

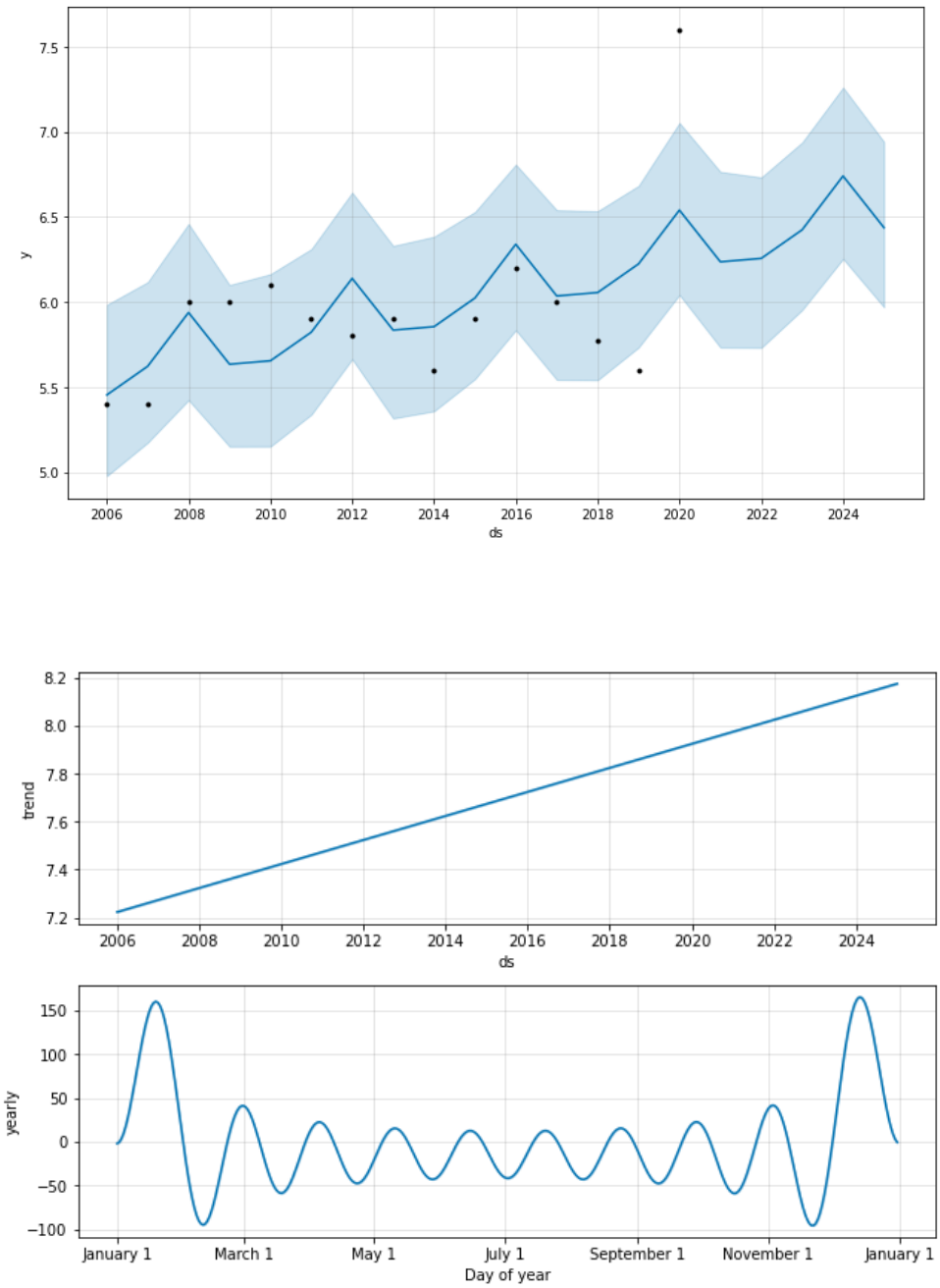

Figure 5.14: Plot and Trend of P3 indicator

Current trend is on a decline and also might not work well as projected by the model as there is a steep decline currently as interpreted from the figure 5.14

11. S1: Demographic Pressures:

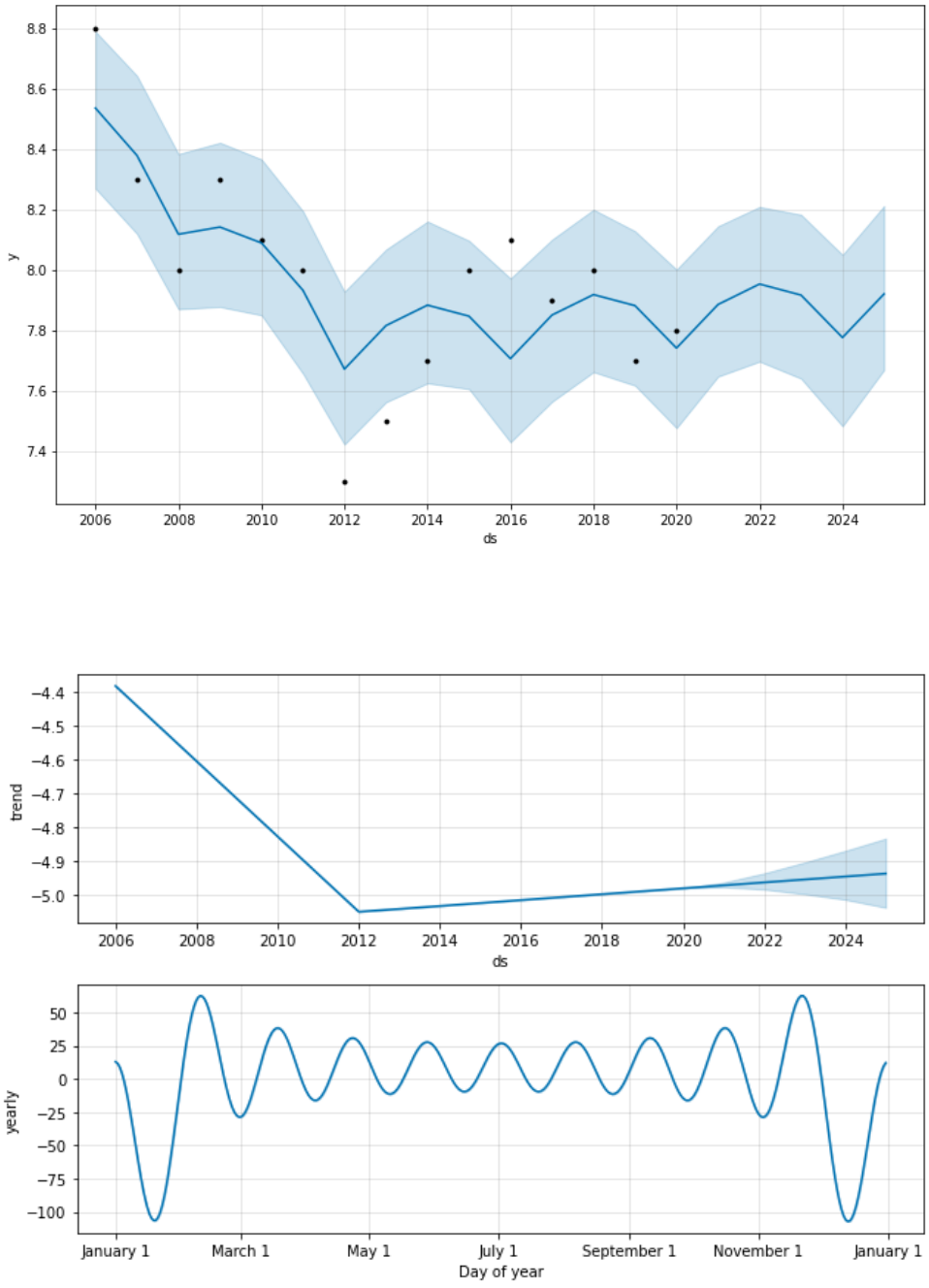

Figure 5.15: Plot and Trend of S1 indicator

From the figure 5.15 the projected trend is going upwards which means that the condition in this variable is also declining but not as steeply as human rights.

12. S2: Refugees and IDPs:

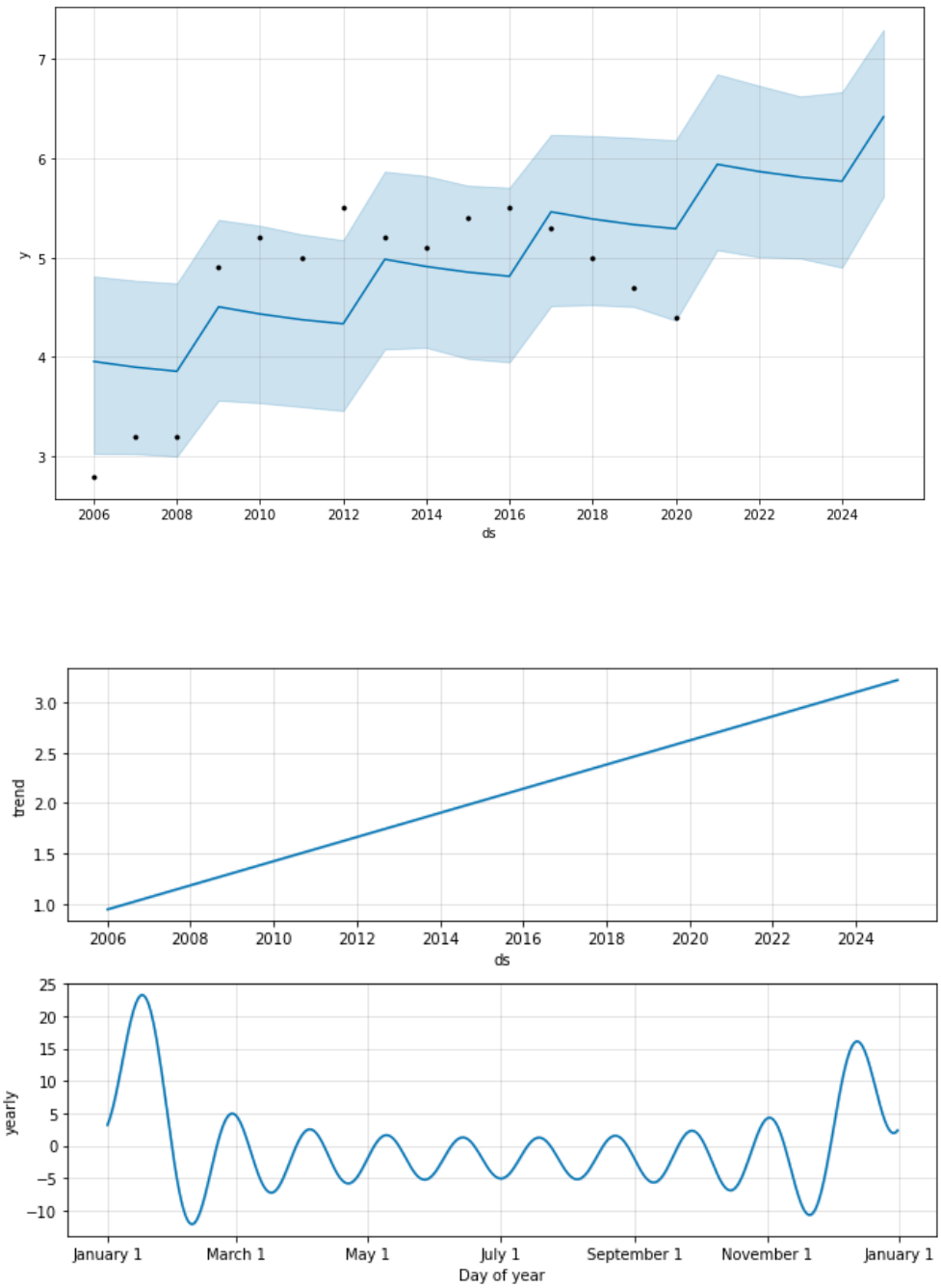

Figure 5.16: Plot and Trend of S2 indicator

Clearly in this variable too the trend is going higher that means a sharp decline in this indicator is predicted in near future from the figure 5.16. The Refugee condition in India will be pretty bad by this trend.

13. X1: External Intervention:

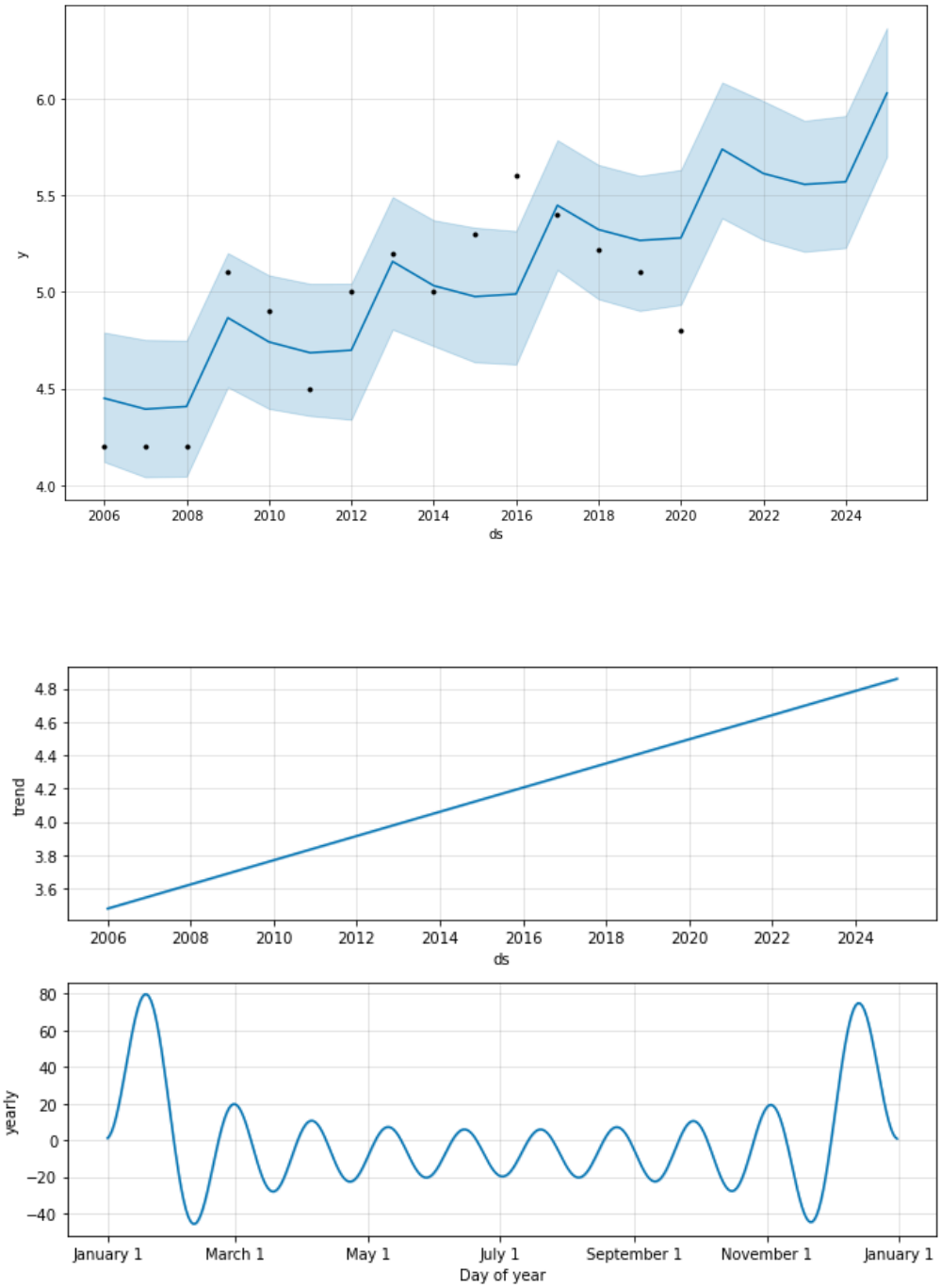

Figure 5.17: Plot and Trend of X1 indicator

The trend proves that the external intervention of India is becoming a problem from the figure 5.17. In near future too things are looking really bleak in this indicator too.

5.6 Summary

In this section we compared the accuracy of the 2 clustering algorithms and how Hierarchical clustering was slightly better than the K-means on accuracy measures. Also, the comparison on other scores were done where we found that both the clustering algorithms were quite similar in nature.

In case of forecasting with Facebook's Prophet library the projections were being made on India and all of its variables associated with it. The projections were clearly giving the evidence of declination of the country as a whole.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

In this chapter, a discussion on reality vs prediction / forecast is discussed. A conclusion is being made keeping the reality in mind that how much machine learning models can help in predicting / projecting the reality for a country on the basis of fragility and stability. In the contributions section a discussion is done on how machine learning can be a great tool for predicting and forecasting socio-politico-economical condition of a country. If something could have been better in this research is being discussed in the future works section of this chapter. Future works also discusses about the next level research in the same topic.

6.2 Discussion and Conclusion

With an accuracy of 70% the prediction of clustering algorithms was good enough to predict the fragility of all the countries in the 15 years of time span. Mostly under developed and developing countries falls in the category of fragile countries and developed countries are mostly stable. Our second research question “How well will an unsupervised learning technique be able to segment the countries on the basis of conflict occurrences?” is answered by this approach.

India was taken as a test subject for this research as it’s a developing country and riots are commonly seen and read about in the media reports. The clustering algorithm also predicted India as a continuously fragile country for straight 15 years. In the time series forecasting of all the variables of country “India”, there were few discrepancies found with respect to the reality. In the variable E2: Economic Inequality Drain the Oxfam report (Oxfam.org, 2018) which clearly said that economic inequality is on the rise and they backup their claim with proper data and research, FSI needs to update this data according to the reality. In the variable E3: Human Flight and Brain there are media reports in India that the super wealthy people of this country are fleeing and taking citizenships of other countries and they claim it with proper data (business standard, 2019), FSI needs to update this data according to the

reality. The “Total” indicator which is a dependent variable showed a wrong trend because all the data was added to it till 2020 and then projections were made, the projections have to be made by a so-called dependent approach. In other words, all the other variables projections had to be added with respect to time to plot the “total” graph and then deduce the trend. India as a test subject is heavily declining and need proper reforms to get better maybe a better leadership with good intent can do something better for this country.

Our 3rd research question “How oversampling technique supports the analysis of time series for better prediction?”, it was an idea to oversample the time series in order to get better prediction as there will be more data points but facebook’s Prophet library overcame on such things and it was not required all together. India as our test country, Facebook’s Prophet library gave us really good models that determined the fragility of India, this answers our 4th research question “How Facebook’s Prophet can forecast the fragility of a country fed with time series data?”.

6.3 Contributions

From this research new horizons are opening up for the use of machine learning in the field of socio-politico-economics. This research can help people of all the world how their particular country is fairing in the various indicators that affect their lives. People can take decisions on the basis forecasts and projections of individual indicators and can ask their governments to do better about them. This research in other words **strengthens the democracy worldwide**.

6.4 Future Works

There were few drawbacks in the research as pointed in the section discussion and conclusion. FSI can update the data according to the reality and those particular contrasts can be mitigated.

While working on this research, there were few indicators which could have been a great resource in making the clustering algorithms’ accuracy much better. Press Freedom Index and the Democratic Index could have been incorporated in this research and the countries which fall low in such Indices like China, North Korea etc. (rsf.org, 2021) can be discarded from the confusion matrix dataset. This is because any riot happening in these countries

will not be coming up in the world media as these are closed countries. There are many such countries which lay low in Press Freedom Index and the information doesn't come up in the list of riots (Wikipedia, 2020).

Technically there can be much more advancement in the machine learning models and can be advanced to the deep learning models too. LSTM can be used forecasting the individual indicator performance of a particular country. The accuracy of Facebook's Prophet and LSTM can also be compared in the future research. Individuals belonging to other country can take up this kind of research and can apply it on their own country and can test the FSI data as per the reality happening around them. Also, DEV-OPS / ML-OPS tools can be used to create such time-series projections for all the available countries with more resources. So, our first research question "Can Machine Learning techniques predict the future conflicts that might occur in any country across the world?" can be answered using this technique, maybe in near future someone will pick this research from here and apply that.

References

Wikipedia 2020 *List of riots*, Accessed on 4th Oct 2020. https://en.wikipedia.org/wiki/List_of_riots#2001%E2%80%932009

Fund for World Peace 2020. *Fragile States Index [Online]*, Accessed on 4th Oct 2020 http://ffp.statesindex.org

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

Wikipedia 2020 *K-means clustering*, Accessed on 4th Oct 2020. https://en.wikipedia.org/wiki/K-means_clustering

Wikipedia 2020 *Time Series*, Accessed on 4th Oct 2020. https://en.wikipedia.org/wiki/Time_series

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

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

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/

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/

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.

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.

Coyne, Christopher & Mathers, Rachel. (2010). Rituals: An economic interpretation. *Journal of Economic Behavior & Organization*. 78. 74-84. 10.1016/j.jebo.2010.12.009.

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.

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)

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)

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.

Bang, James & Shughart II, William & Basuchoudhary, Atin. (2019). *Predicting State Failure: Different Pathways into the Abyss.*

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

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

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

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)

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

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)

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

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)

Acemoglu, Daron, and James A. Robinson. 2001. "*A Theory of Political Transitions.*" American Economic Review 91 (4): 938–63.

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown Publishing House.

Akerlof, George. 1979. "*The Market for "Lemons*": *Quality Uncertainty and the Market Mechanism*." Quarterly Journal of Economics 84 (3): 488–500.

Anderton, Charles H., and John R. Carter. 2009. *Principles of Conflict Economics: A Primer for Social Scientists*. Cambridge: Cambridge University Press.

Bang, James T., Atin Basuchoudhary, John David, and Aniruddha Mitra. 2016. "*Predicting Aggregate Terror Risk: A Machine Learning Approach*." Working Paper.

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.

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.

Beard, Mary. 2010. "*State Capacity, Conflict and Development*." Econometrica 78 (1): 1–34. ———. 2015.

SPQR: *A History of Ancient Rome.* New York, London: Liveright Publishing Corporation, A Division of W.W. Norton and Co. Besley

Timothy J., and Torsten Persson. 2008. "*The Incidence of Civil War: Theory and Evidence*." National Bureau of Economic Research (NBER) Working Paper 14585 39.

Blomberg, S. Brock, and Gregory D. Hess. 2002. "*The Temporal Links Between Conflict and Economic Activity.*" Journal of Conflict Resolution 46 (1): 74–90.

Buchanan, James M. 1975. *The Limits of Liberty: Between Anarchy and Leviathan.* Chicago: University of Chicago Press.

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.

Center for Systemic Peace. 2016. *Global Conflict Trends: Assessing the Qualities of Systematic Peace*, July 13, http://www.systemicpeace.org/conflicttrends.html.

Charness, Gary, and Matthew Rabin. 2002. "*Understanding Social Preferences with Simple Tests.*" Quarterly Journal of Economics 117 (3): 817069.

Chassang, Sylvain, and Gerard Padri-i-Miquel. 2009. "*Economic Shocks and Civil War.*" Quarterly Journal of Political Science 4 (3): 211–28.

Choi, Jung-Kyoo, and Samuel Bowles. 2007. "*The Coevolution of Parochial Altruism and War.*" Science 318 (5850): 637–40.

Ciccone, Antonio. 2011. "*Economic Shocks and Civil Conflict: A Comment*." American Economic Journal: Applied Economics 3: 215–7.

Clark, Greg. 2007. A Farewell to Alms. Princeton, NJ: Princeton University Press.

Collier, Paul, and Anke Hoeffler. 1998. "*On Economic Causes of Civil War*." Oxford Economic Papers 50 (4): 563–73. ———. 2004.

"*Greed and Grievance in Civil War*." Oxford Economic Papers 56: 563–95.

Collier, Paul, and Nicholas Sambanis. 2002. "*Understanding Civil War." Journal of Conflict Resolution* 46 (1): 3–12.

Collier, Paul, Anke Hoeffler, and Dominic Roehner. 2009. "*Beyond Greed and Grievance: Feasibility and Civil War.*" Oxford Economic Papers 61: 1–27.

Congleton, Roger D. 1995. "*Ethnic Clubs, Ethnic Conflict, and the Rise of Ethnic Nationalism.*"

In Nationalism and Rationality, by Albert Breton, Gianluigi Galeotti, Pierre Salmon, and Ronald Wintrobe, 71–97. Cambridge: Cambridge University Press.

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.

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.

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.

Darley, John M., and Joel Cooper. 1998. *Attribution and Social Interaction: The Legacy of Edward E. Jones.* Washington, DC: American Psychological Association Press.

Desch, Michael C. 2008. *Power and Military Effectiveness: The Fallacy of Democratic Triumphalism.* Baltimore, MD: Johns Hopkins University Press.

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.

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.

McDermott, Rose. 2004. "*The Feeling of Rationality: The Meaning of Neuroscientific Advances for Political Science*." Perspectives on Politics 2 (4): 691– 706.

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.

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.

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.

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.

Indicators [online image] (n.d.) Available from: https://fragilestatesindex.org/indicators/ [Accessed: 30th November, 2020]

Weytjens, H., Lohmann, E. & Kleinsteuber, M. *Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet*. *Electron Commer Res* (2019). https://doi.org/10.1007/s10660-019-09362-7

"*Bitcoin Forecasting Using ARIMA and PROPHET*," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018, pp. 621-624, doi: 10.1109/UBMK.2018.8566476.

From Oxfam.org on "*Economic Inequality in India*", accessed on 15th February 2021. (https://www.oxfam.org/en/india-extreme-inequality-numbers)

From businessinsider.in on "*Rich People fleeing India*", accessed on 15th February 2021 by PRERNA SINDWANI (https://www.businessinsider.in/why-indias-millionaires-are-leaving-india/articleshow/69321959.cms)

A. Banerjee and R. N. Dave, "*Validating clusters using the Hopkins statistic*" 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542), Budapest, Hungary, 2004, pp. 149-153 vol.1, doi: 10.1109/FUZZY.2004.1375706.

Peter J. Rousseeuw, *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, *Journal of Computational and Applied Mathematics*, Volume 20, 1987, Pages 53-65, ISSN 0377-0427, https://doi.org/10.1016/0377-0427(87)90125-7. (https://www.sciencedirect.com/science/article/pii/0377042787901257)

From minorityrights.org on "*Maori: Minority rights group in New Zealand*", accessed on 15th February 2021. (https://minorityrights.org/minorities/maori/)

E. H. Ramirez, R. Brena, D. Magatti and F. Stella, "*Probabilistic Metrics for Soft-Clustering and Topic Model Validation,*" 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, Canada, 2010, pp. 406-412, doi: 10.1109/WI-IAT.2010.148.

From rsf.org on "*Press Freedom Index*", accessed on 15th February 2021. (https://rsf.org/en/ranking)

From carnegieendowment.org on "*Mounting Majoritarianism and Political Polarization in India*" by Niranjan Sahoo, accessed on 15th February 2021. (https://carnegieendowment.org/2020/08/18/mounting-majoritarianism-and-political-polarization-in-india-pub-82434)

From reuter.com on "*UK ever more polarized as Brexit Party storms to EU vote win*" by William James, accessed on 15th February 2021. (https://www.reuters.com/article/us-eu-election-britain/uk-ever-more-polarized-as-brexit-party-storms-to-eu-vote-win-idUSKCN1SW0Y8)

From pewresearch.org on "America is exceptional in the nature of its political divide" by Michael Dimock and Richard Wike, accessed on 15th February 2021. (https://www.pewresearch.org/fact-tank/2020/11/13/america-is-exceptional-in-the-nature-of-its-political-divide/)

I. Yenidoğan, A. Çayir, O. Kozan, T. Dağ and Ç. Arslan, "*Bitcoin Forecasting Using ARIMA and PROPHET,*" 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018, pp. 621-624, doi: 10.1109/UBMK.2018.8566476.

Appendix A

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

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

5. Can Machine Learning techniques predict the future conflicts that might occur in any country across the world?
6. How well will an unsupervised learning technique be able to segment the countries on the basis of conflict occurrences?
7. 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:

c. K-means
d. Time Series Analysis (if oversampling works)
e. Random Forest
f. 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:

a. Check the raw data for the number of variables that are numeric or category based.
b. Separate the numeric and category-based variables.
c. 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
d. Create Dummy variable for the categorical variables.
e. 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.

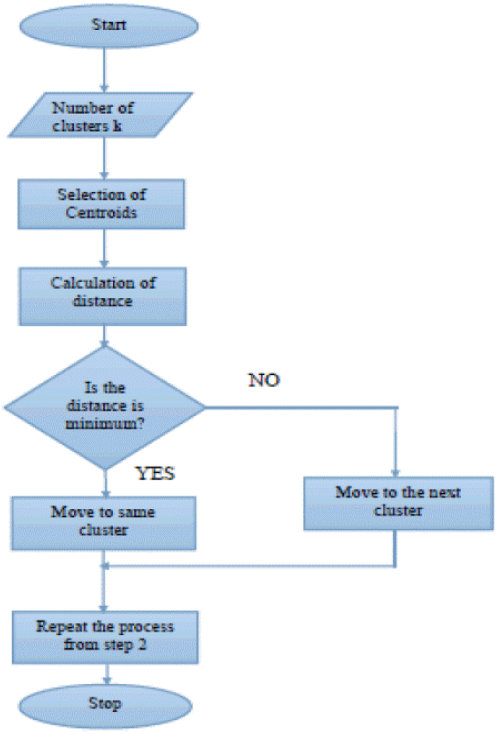

Fig 5: K-Means Algorithm on Flow Chart

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.

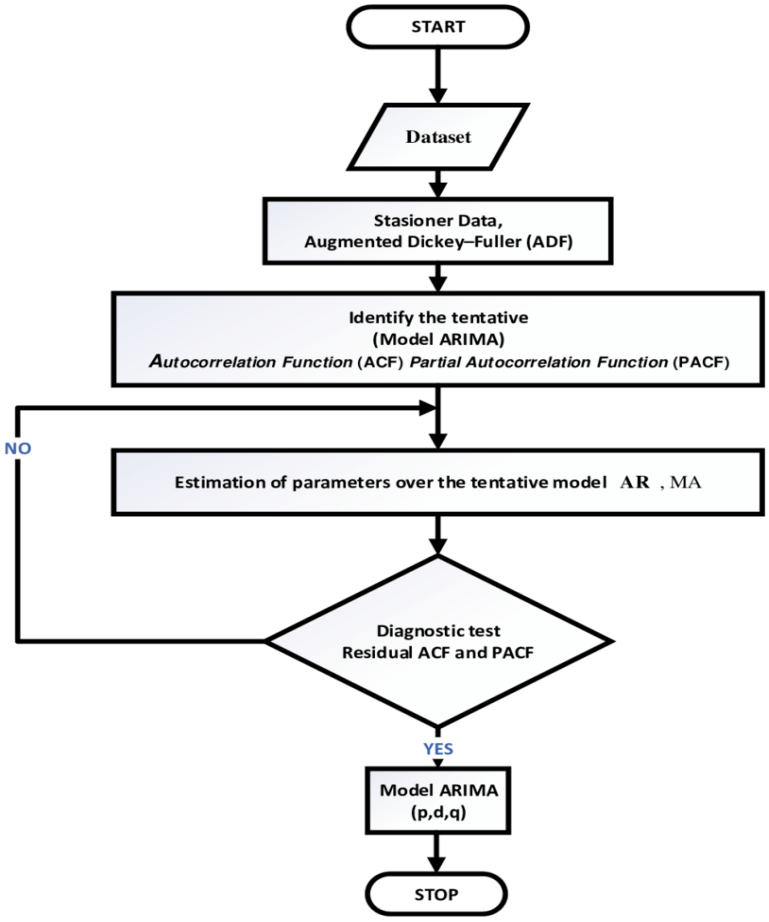

Fig 6: Time Series Forecasting Algorithm Flow Chart

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.

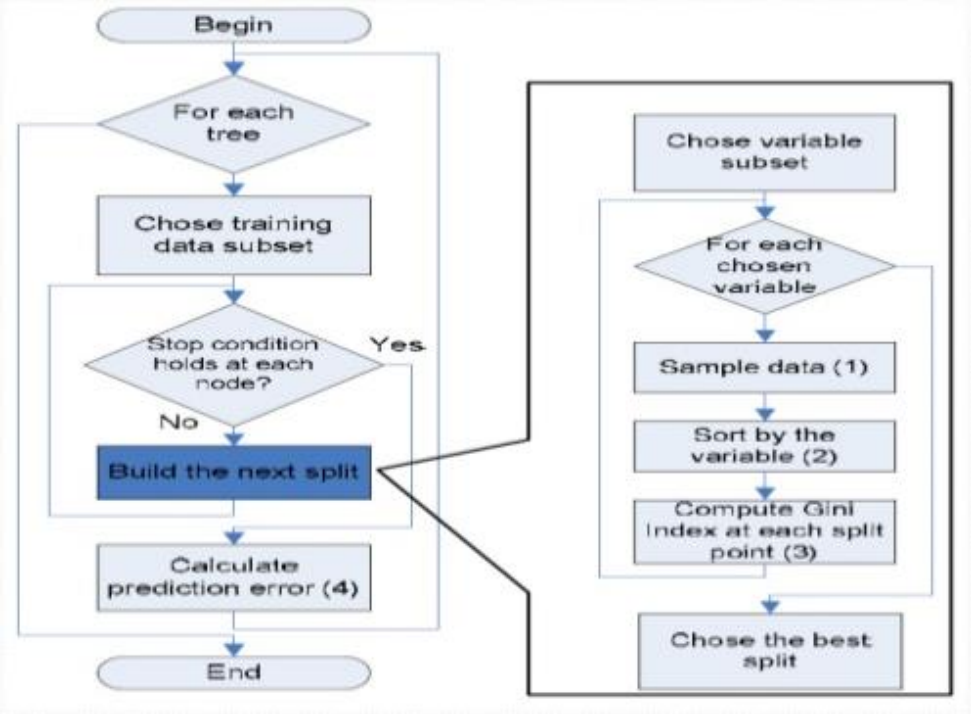

Fig 7: Random Forest Algorithm Flow Chart

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.

Appendix B

Code for Clustering the Countries

Predicting Political Instability across the world.html

The above-mentioned .html file can be opened up in any web browser for reference.

Appendix C

Code for Extracting Data from Wikipedia

WikipediaDataExtraction.html

The above-mentioned .html file can be opened up in any web browser for reference.

Appendix D

Code for Time Series Projections using Facebook's Prophet Library

Time Series with Facebook's Prophet Library.html

The above-mentioned .html file can be opened up in any web browser for reference.