**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

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

## Details of the research project

### Aims and Objectives

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

**Data Explanation**

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


Fig-1: Cohesion Indicators

**COHESION INDICATORS**

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

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

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

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


Fig 2: Economic Indicators

**ECONOMIC INDICATORS**

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

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

preparing with the end goal that regardless of whether there is financial imbalance present, how much it is public oriented and strengthening.

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


Fig 3: Political Indicators

**POLITICAL INDICATORS**

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

capacity of a state to practice essential capacities that inference a populace's trust in its administration and organizations, for example, through the capacity to gather duties.

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

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

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

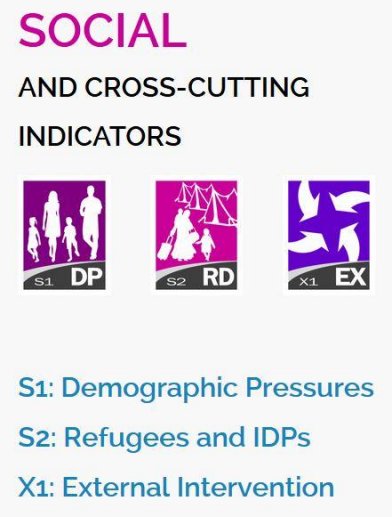


Fig 4: Social and Cross-Cutting Indicators

**SOCIAL INDICATORS**

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

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

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

**CROSS-CUTTING INDICATORS**

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

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

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

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

## Research Methodology

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

a. K-means
b. Time Series Analysis (if oversampling works)
c. Random Forest
d. XGBoost

### Data Preparation

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

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

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

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

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

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

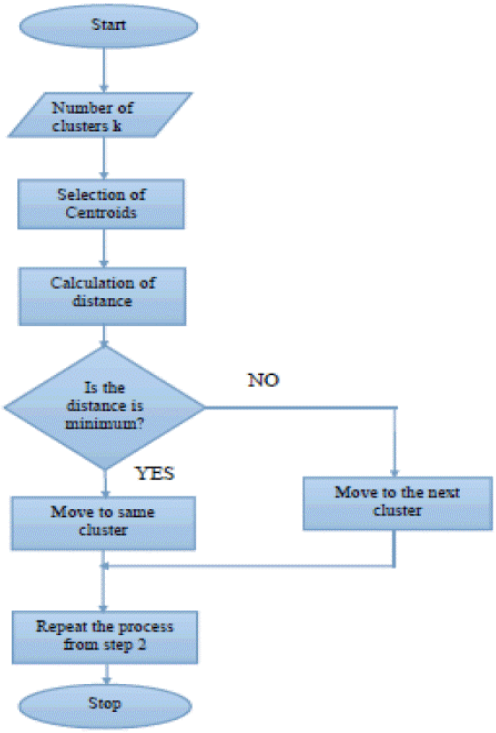



Fig 5: K-Means Algorithm on Flow Chart

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

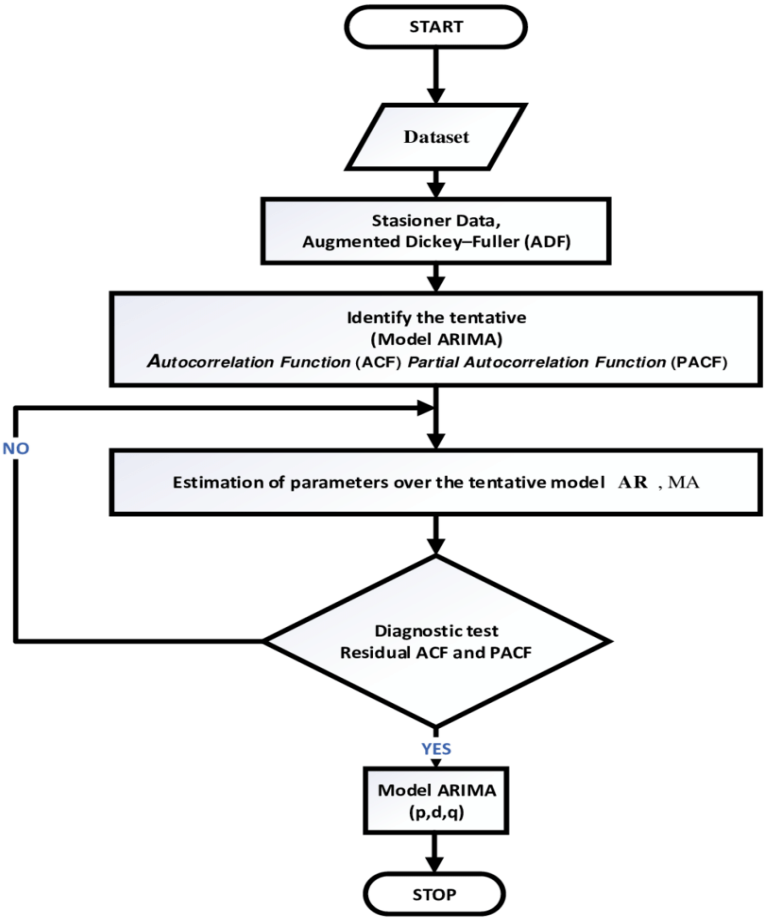



Fig 6: Time Series Forecasting Algorithm Flow Chart

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

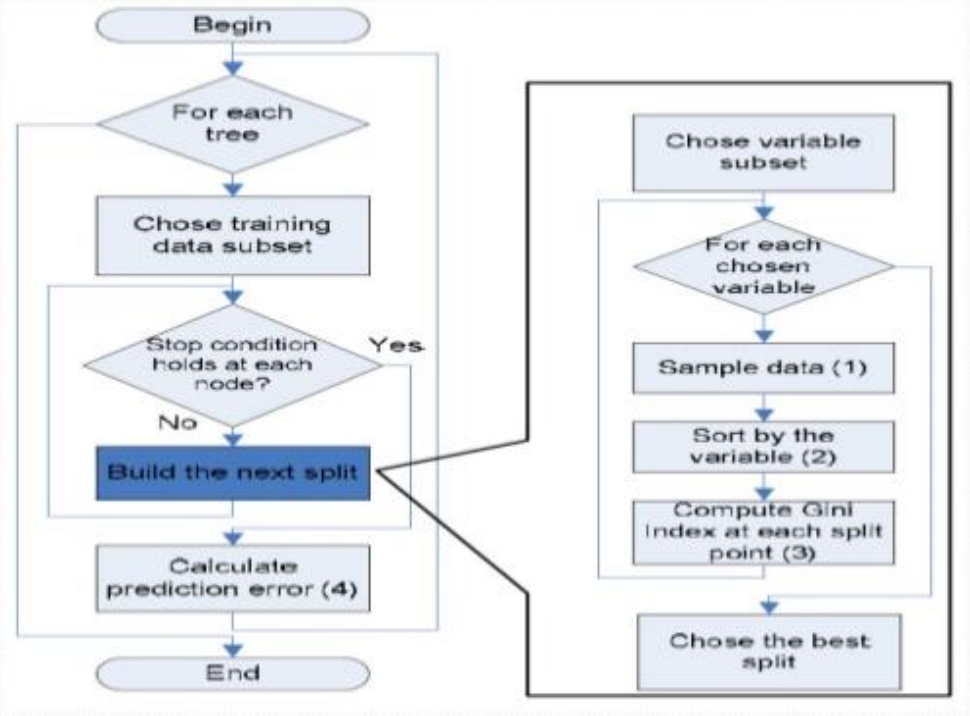



Fig 7: Random Forest Algorithm Flow Chart

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

## Outcome

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

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

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

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

## Risks or contingency plan

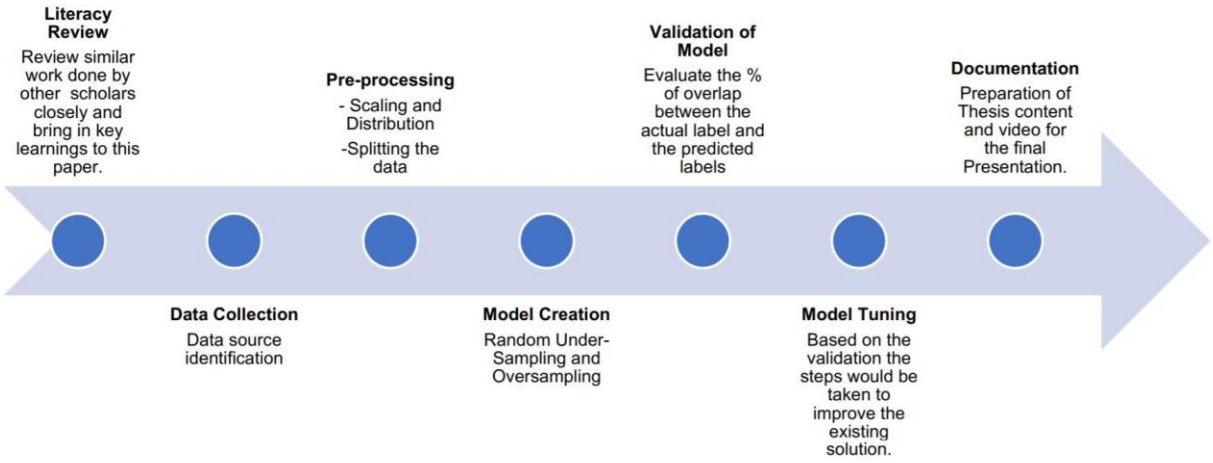
I might skip one approach in the methodology if the time is limited. Time Series is going to be a cumbersome and highly time-consuming exercise, if there is less time, I might have to skip that part.

## Resource Requirements

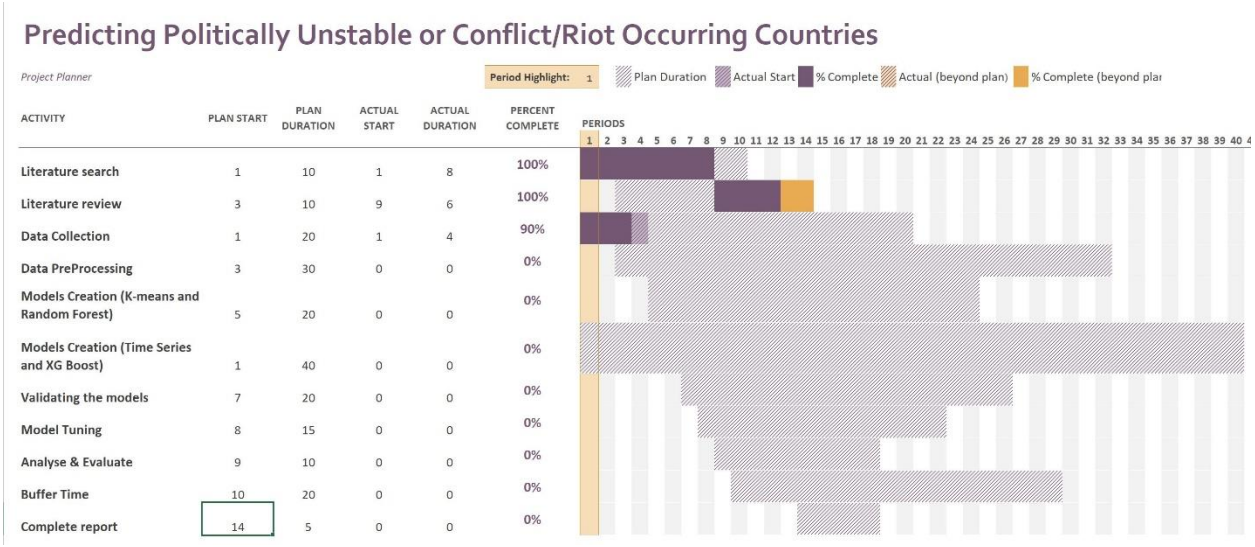
A machine with i5/i7 processor, 16 GB RAM and 4 GB graphics card is sufficient for this type of machine learning research. The data size is not that huge for a requirement of an array of graphics card. The resources mentioned above would be sufficient. In software and utilities part the requirement is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

# Timetables and Milestones using Gantt chart

## Analysis Plan for components




## Gantt Chart



## References:


1. From Wikipedia, on “List of riots”, 4th Oct 2020. https://en.wikipedia.org/wiki/List_of_riots#2001%E2%80%932009
2. From Fund for World Peace. Fragile States Index [Online], 4th Oct 2020 http://ffp.statesindex.org
3. Predicting High-Risk Countries for Political Instability and Conflict by Blair Huffman, Emma Marriott, April Yu Stanford University. (2016) http://cs229.stanford.edu/proj2014/Blair%20Huffman,%20Emma%20Marriott,%20April%20Yu,%20Predicting%20high-risk%20countries%20for%20political%20instability%20and%20conflict.pdf
4. From Wikipedia on “K-means clustering”, 4th Oct 2020. https://en.wikipedia.org/wiki/K-means_clustering
5. From Wikipedia, on “Time Series”, 4th Oct 2020. https://en.wikipedia.org/wiki/Time_series
6. Iwana, Brian & Uchida, Seiichi. (2020). Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher. https://www.researchgate.net/publication/340805406_Time_Series_Data_Augmentation_for_Neural_Networks_by_Time_Warping_with_a_Discriminative_Teacher/citation/download
7. From Wikipedia on “Random Forest”, 4th of Oct 2020. https://en.wikipedia.org/wiki/Random_forest
8. From Wikipedia on “List of wars and anthropogenic disasters by death toll”, 4th of Oct 2020. https://en.wikipedia.org/wiki/List_of_wars_and_anthropogenic_disasters_by_death_toll
9. From Vision of Humanity, 4th of Oct 2020. http://visionofhumanity.org/economists-on-peace/predicting-civil-conflict-can-machine-learning-tell-us/