|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Structure of the research proposal** | **Word limit** |
| 1 | Your Name(Mention whether you are in MSc Data Science (DS) or Machine Learning (ML) programme | 10-20 words |
| 2. | Research title | max 10 words |
| 3. | Summary of the proposed research | max 200 words |
| 4. | Research question | max 350 words |
| 5. | Details of the research project (Includes aims, objectives, references, data and methods used for analysis) | max 2000 words |
| 6. | Outcome | max 200 words |
| 7. | Risks or contingency plan | - |
| 8. | Timetables and milestones using Gantt chart | - |
| 9. | References | - |

Predicting Politically Unstable or Conflict Occurring Countries

Rachit Dev

Master of Data Science Research Proposal

Summary

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. By being able to detect trends within a nation and respond with the right political, economic, and developmental sanctions, the use of military intervention can be avoided and prevent conflict or total governmental collapse. Our motivation is to be able to capture and interpret these trends on a grand scale and build a model that can indicate the fragility of a nation, as well as identify the crucial indicators that attribute to its instability. Currently, there are organizations that rank countries based on their political instability, but these systems use targeted quantitative and qualitative data, and are annual reviews rather than instantaneous predictions. We hope that by applying machine learning techniques we can predict the conflicts which might occur in the countries which are prone to it.

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?

Suggested Questions:

1. Can a research using Machine Learning predict conflict in any country using relevant data across the world?
2. Can a 15 years data (with 1 row per year) be used for time series analysis for prediction using oversampling techniques, where we can introduce more entries (for example, 1 entry per quarter rather than 1 entry per year)? Which oversampling technique would be suitable?
3. We currently cannot use supervised leaning according to the data (although we can use the ranking / total points). How we will be proceeding with the unsupervised learning to supervised learning remains a question, maybe we might be incorporating more data or just by ranking we will be supervising the things?
4. We will be trying the below mentioned algorithms:
5. K-means
6. Time Series Analysis (if oversampling works)
7. Random Forest
8. ANN

For predicting the future conflicts that might occur in any country across the world. Which algorithm will predict the best result?

1. How well will K-means be able to segregate the countries on the basis of conflict occurrences?
2. If oversampling works well, will time series be able to predict in a good way as there will be synthetic data incorporated in our research?
3. Time Series analysis will be a cumbersome process as each and every country will be dealt individually by our model, so repeating it for all the countries will be a tedious task, we will have to find out a good programming concept to deal with this problem.

Details of the research project

**Aim** =  what you hope to achieve.  
**Objective** = the action(s) you will take in order to achieve the aim.  
  
**Aims** are statements of intent. They are usually written in broad terms. They set out what you hope to achieve at the end of the project.  
  
**Objectives**, on the other hand, should be specific statements that define measurable outcomes, e.g. what steps will be taken to achieve the desired outcome.   
When writing your objectives try to use strong positive statements.   
  
**Strong verbs -** collect, construct, classify, develop, devise, measure, produce, revise, select, synthesise   
  
**Weak verbs** - appreciate, consider, enquire, learn, know, understand, be aware of, appreciate, listen, perceive

Aims

Our aim in this research is to create a model which will be able to predict politically unstable or civil war prone countries using the FSI dataset.

Objectives

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 Artificial Neural Networks as the main algorithmic approach towards our problem. (Add some bullet points about the usage of the ML alogs you are going to use.)

Data

The FSI data 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:



**COHESION INDICATORS**

**C1: Security Apparatus:** The Security Apparatus indicator considers the security threats to a state, such as bombings, attacks and battle-related deaths, rebel movements, mutinies, coups, or terrorism. The Security Apparatus also takes into account serious criminal factors, such as organized crime and homicides, and perceived trust of citizens in domestic security. In some instances, the security apparatus may extend beyond traditional military or police forces to include state-sponsored or state-supported private militias that terrorize political opponents, suspected “enemies,” or civilians seen to be sympathetic to the opposition. In other instances, the security apparatus of a state can includes a “deep state”, that may consist of secret intelligence units, or other irregular security forces, that serve the interests of a political leader or clique. As a counter example, the indicator will also take into account armed resistance to a governing authority, particularly the manifestation of violent uprisings and insurgencies, proliferation of independent militias, vigilantes, or mercenary groups that challenge the state’s monopoly of the use of force.

**C2: Factionalized Elites:** The Factionalized Elites indicator considers the fragmentation of state institutions along ethnic, class, clan, racial or religious lines, as well as and brinksmanship and gridlock between ruling elites. It also factors the use of nationalistic political rhetoric by ruling elites, often in terms of nationalism, xenophobia, communal irredentism (e.g., a “greater Serbia”) or of communal solidarity (e.g., “ethnic cleansing” or “defending the faith”). In extreme cases, it can be representative of the absence of legitimate leadership widely accepted as representing the entire citizenry. The Factionalized Elites indicator measures power struggles, political competition, political transitions, and where elections occur will factor in the credibility of electoral processes (or in their absence, the perceived legitimacy of the ruling class).

## C3: Group Grievance: The Group Grievance Indicator focuses on divisions and schisms between different groups in society – particularly divisions based on social or political characteristics – and their role in access to services or resources, and inclusion in the political process. Group Grievance may also have a historical component, where aggrieved communal groups cite injustices of the past, sometimes going back centuries, that influence and shape that group’s role in society and relationships with other groups. This history may in turn be shaped by patterns of real or perceived atrocities or “crimes” committed with apparent impunity against communal groups. Groups may also feel aggrieved because they are denied autonomy, self-determination or political independence to which they believe they are entitled. The Indicator also considers where specific groups are singled out by state authorities, or by dominant groups, for persecution or repression, or where there is public scapegoating of groups believed to have acquired wealth, status or power “illegitimately”, which may manifest itself in the emergence of fiery rhetoric, such as through “hate” radio, pamphleteering, and stereotypical or nationalistic political speech.

## 

**ECONOMIC INDICATORS**

## E1: Economic Decline and Poverty: The Economic Decline Indicator considers factors related to economic decline within a country. For example, the Indicator looks at patterns of progressive economic decline of the society as a whole as measured by per capita income, Gross National Product, unemployment rates, inflation, productivity, debt, poverty levels, or business failures. It also takes into account sudden drops in commodity prices, trade revenue, or foreign investment, and any collapse or devaluation of the national currency. The Economic Decline Indicator further considers the responses to economic conditions and their consequences, such as extreme social hardship imposed by economic austerity programs, or perceived increasing group inequalities. The Economic Decline Indicator is focused on the formal economy – as well as illicit trade, including the drug and human trafficking, and capital flight, or levels of corruption and illicit transactions such as money laundering or embezzlement.

## E2: Uneven Economic Development: The Uneven Economic Development Indicator considers inequality within the economy, irrespective of the actual performance of an economy. For example, the Indicator looks at structural inequality that is based on group (such as racial, ethnic, religious, or other identity group) or based on education, economic status, or region (such as urban-rural divide). The Indicator considers not only actual inequality, but also perceptions of inequality, recognizing that perceptions of economic inequality can fuel grievance as much as real inequality, and can reinforce communal tensions or nationalistic rhetoric. Further to measuring economic inequality, the Indicator also takes into account the opportunities for groups to improve their economic status, such as through access to employment, education, or job training such that even if there is economic inequality present, to what degree it is structural and reinforcing.

## E3: Human Flight and Brain Drain: The Human Flight and Brain Drain Indicator considers the economic impact of human displacement (for economic or political reasons) and the consequences this may have on a country’s development. On the one hand, this may involve the voluntary emigration of the middle class – particularly economically productive segments of the population, such as entrepreneurs, or skilled workers such as physicians – due to economic deterioration in their home country and the hope of better opportunities farther afield. On the other hand, it may involve the forced displacement of professionals or intellectuals who are fleeing their country due to actual or feared persecution or repression, and specifically the economic impact that displacement may wreak on an economy through the loss of productive, skilled professional labor.

## 

## POLITICAL INDICATORS

## P1: State Legitimacy: The State Legitimacy Indicator considers the representativeness and openness of government and its relationship with its citizenry. The Indicator looks at the population’s level of confidence in state institutions and processes, and assesses the effects where that confidence is absent, manifested through mass public demonstrations, sustained civil disobedience, or the rise of armed insurgencies. Though the State Legitimacy indicator does not necessarily make a judgment on democratic governance, it does consider the integrity of elections where they take place (such as flawed or boycotted elections), the nature of political transitions, and where there is an absence of democratic elections, the degree to which the government is representative of the population of which it governs. The Indicator takes into account openness of government, specifically the openness of ruling elites to transparency, accountability and political representation, or conversely the levels of corruption, profiteering, and marginalizing, persecuting, or otherwise excluding opposition groups. The Indicator also considers the ability of a state to exercise basic functions that infer a population’s confidence in its government and institutions, such as through the ability to collect taxes.

## P2: Public Services: The Public Services Indicator refers to the presence of basic state functions that serve the people. On the one hand, this may include the provision of essential services, such as health, education, water and sanitation, transport infrastructure, electricity and power, and internet and connectivity. On the other hand, it may include the state’s ability to protect its citizens, such as from terrorism and violence, through perceived effective policing. Further, even where basic state functions and services are provided, the Indicator further considers to whom – whether the state narrowly serves the ruling elites, such as security agencies, presidential staff, the central bank, or the diplomatic service, while failing to provide comparable levels of service to the general populace – such as rural versus urban populations. The Indicator also considers the level and maintenance of general infrastructure to the extent that its absence would negatively affect the country’s actual or potential development.

## P3: Human Rights and Rule of Law: The Human Rights and Rule of Law Indicator considers the relationship between the state and its population insofar as fundamental human rights are protected and freedoms are observed and respected. The Indicator looks at whether there is widespread abuse of legal, political and social rights, including those of individuals, groups and institutions (e.g. harassment of the press, politicization of the judiciary, internal use of military for political ends, repression of political opponents). The Indicator also considers outbreaks of politically inspired (as opposed to criminal) violence perpetrated against civilians. It also looks at factors such as denial of due process consistent with international norms and practices for political prisoners or dissidents, and whether there is current or emerging authoritarian, dictatorial or military rule in which constitutional and democratic institutions and processes are suspended or manipulated.

## 

## SOCIAL INDICATORS

## S1: Demographic Pressures: The Demographic Pressures Indicator considers pressures upon the state deriving from the population itself or the environment around it. For example, the Indicator measures population pressures related to food supply, access to safe water, and other life-sustaining resources, or health, such as prevalence of disease and epidemics. The Indicator considers demographic characteristics, such as pressures from high population growth rates or skewed population distributions, such as a “youth or age bulge,” or sharply divergent rates of population growth among competing communal groups, recognizing that such effects can have profound social, economic, and political effects. Beyond the population, the Indicator also takes into account pressures stemming from natural disasters (hurricanes, earthquakes, floods or drought), and pressures upon the population from environmental hazards.

## S2: Refugees and IDPs: The Refugees and Internally Displaced Persons Indicator measures the pressure upon states caused by the forced displacement of large communities as a result of social, political, environmental or other causes, measuring displacement within countries, as well as refugee flows into others. The indicator measures refugees by country of Asylum, recognizing that population inflows can put additional pressure on public services, and can sometimes create broader humanitarian and security challenges for the receiving state, if that state does not have the absorption capacity and adequate resources. The Indicator also measures the Internally Displaced Persons (IDP) and Refugees by country of origin, which signifies internal state pressures as a result of violence, environmental or other factors such as health epidemics. These measures are considered within the context of the state’s population (per capita) and human development trajectory, and over time (year on year spikes), recognizing that some IDPs or refugees for example, may have been displaced for long periods of time.

## CROSS-CUTTING INDICATORS

## X1: External Intervention: The External Intervention Indicator considers the influence and impact of external actors in the functioning – particularly security and economic – of a state. On the one hand, External Intervention focuses on security aspects of engagement from external actors, both covert and overt, in the internal affairs of a state at risk by governments, armies, intelligence services, identity groups, or other entities that may affect the balance of power (or resolution of a conflict) within a state. On the other hand, External Intervention also focuses on economic engagement by outside actors, including multilateral organizations, through large-scale loans, development projects, or foreign aid, such as ongoing budget support, control of finances, or management of the state’s economic policy, creating economic dependency. External Intervention also takes into account humanitarian intervention, such as the deployment of an international peacekeeping mission.

## Our Dataset has 178 countries before that there maybe lesser countries as some countries got split for example Sudan got split into Sudan and South Sudan.

Methods used for analysis (elaborate as much as possible also including data preparation part, summary of the algorithms, justification for the usage of algorithms)

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

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

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
  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.

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

K-Means: Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build 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.

Time Series: **Time series *analysis*** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series *forecasting*** is the use of model to predict future values based on previously observed values. 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 state of a country exercise rather than predicting the conflict.

Random Forest:

Outcome

Four five lines

Risks or contingency plan

Define the risks you have in data, backup plan if something fails.

Timetables and milestones using Gantt chart

References:

* + 1. List of riots <https://en.wikipedia.org/wiki/List_of_riots#2001%E2%80%932009>
    2. Fund for World Peace. Fragile States Index [Online]. Available: <http://ffp.statesindex.org>
    3. Predicting High-Risk Countries for Political Instability and Conflict by Blair Huffman, Emma Marriott, April Yu Stanford University<http://cs229.stanford.edu/proj2014/Blair%20Huffman,%20Emma%20Marriott,%20April%20Yu,%20Predicting%20high-risk%20countries%20for%20political%20instability%20and%20conflict.pdf>
    4. K-means clustering. <https://en.wikipedia.org/wiki/K-means_clustering>
    5. Time Series. <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