6-14-2023

Risk Management in Design of Distributed Supply Chains Using a Bi-Objective Optimization and Machine-Learning Approach

Shirin Shahsavary
University of Windsor

Risk Management in Design of Distributed Supply Chains Using a Bi-Objective Optimization and Machine-Learning Approach

by

Shirin Shahsavary

A Thesis

Submitted to the Faculty of Graduate Studies

through the Industrial Engineering Graduate Program

in Partial Fulfillment of the Requirements for

the Degree of Master of Applied Science at the

University of Windsor

Windsor, Ontario, Canada

Risk Management in Design of Distributed Supply Chains Using a Bi-Objective Optimization and Machine-Learning Approach

by

Shirin Shahsavary

APPROVED BY:

_____

T. Najem
Department of Political Science

_____

M. Wang
Department of Mechanical, Automotive & Materials Engineering

_____

A. Azab, Advisor
Department of Mechanical, Automotive & Materials Engineering

May 23, 2023

DECLARATION OF ORIGINALITY

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

In this day and age, distributed manufacturing is a fact of life. While basic facility location models are typically static and deterministic, the environment in which facilities operate is quite volatile due to factors such as the various encountered risks, population dynamics, market trends, distribution costs, and demand patterns. As a result, in the design of modern supply chains, adopted strategies may need to be revised and redistribution of the supply chain relocating a company's manufacturing facilities may need to take place accordingly. To address this challenge, a Mixed-Integer Linear Programming (MILP) employing a bi-objective cost/risk function is developed. A lexicographic preemptive optimization approach is followed to solve the model using the CPLEX commercial solver. To predict risk associated with a geographic location, Machine Learning (ML) is utilized. The dataset used in this study covers the period from 2006 to 2021 and represents 139 countries in an unbalanced panel. Several ML models are considered to predict the degree of fragility of a geographic location based on political, economic, and social factors. Random Forests have demonstrated superior performance with an accuracy of 96% and an F1-score of 95%. It is believed that this combined ML and optimization approach is deemed to be of benefit to decision makers in design and improvement of their supply chains.

Keywords: Machine Learning, Mixed-Integer Linear Programming, Facility Location Problem, Bi-Objective Optimization

DEDICATION

To my wonderful deeply missed mom,

forever you remain in my soul

To my husband, father, daughter, and sisters

For their continuous love and support

To my supervisor

for his continuous guidance and effort

ACKNOWLEDGMENT

Words cannot express my gratitude to my dear professor Dr. Ahmed Azab with his ongoing support, guidance, and feedback throughout this thesis. I also would like to thank my committee members, Dr. Michael Wang and Dr. Tom Najem, for their invaluable and constructive feedbacks leading to major improvement in my thesis.

Special thanks to my awesome husband for all his guidance, help and encouragement and my beautiful daughter for her love and understanding.

Finally, I would like to thank my father and my sisters for their endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
FLP	Facility Location Problem
FP	False Positive
FN	False Negative
FSI	Fragile States Index
GDP	Growth Domestic Product
KNN	K-Nearest Neighbors Algorithm
GBDT	Gradient-Boosted Decision Trees
ML	Machine Learning
MIPL	Mixed-Integer Linear Programming
PCA	Principle Component Analysis
RF	Random Forests
SMOTE	Over Sampling Method
TN	True Negative
TP	True Positive
VIF	Variance Inflation Factor
XGBoost	Extreme Gradient Boosting

NOMENCLATURE

n	Set of locations
m	Set of customers
$d_{i,t}$	Demand of customer i at time t
$u_{j,t}$	Production cost per unit of location j at time t
$C_{j,t}$	Capacity of location j at time t
$f_{j,i,t}$	Freight cost per unit to ship from location j to customer i at time t
$R_{j,t}$	Risk (fragile state) associated with location j at time t
B_i	Minimum units shipped from a location to customer i (batch size for shipment)
i	Customers, $i \in m$
j	Locations, $j \in n$
t	Time period, $t \in T$
$y_{j,i,t}$	Number of units shipped from location j to customer i in period t
$x_{j,i,t}$	1 If location j is selected to manufacture and ship units to customer i at t 0 Otherwise

CHAPTER 1: INTRODUCTION

1.1 Overview

The Facility Location Problem (FLP) refers to the problem of determining the optimal location for a facility, such as a factory, warehouse, or distribution center. The objective of the most facility location problem is to minimize the total cost of serving the customers, which includes the fixed costs associated with opening and removing the facility, as well as the variable costs associated with serving each customer, such as transportation costs, inventory carrying costs, and labor costs. While basic facility location models are typically static and deterministic, the environment in which facilities operate can change over time due to population dynamics, market trends, distribution costs, and demand patterns. As a result, supply chain designers may need to revise their strategies and relocate their facilities accordingly. To address this challenge, dynamic models have been developed to determine the best facility locations over all time periods (Arabani and Farahani, 2012). In dynamic facility location models, the decision-maker considers the impact of removing or adding facilities over time, considering factors such as changes in demand, transportation costs, and availability of resources. Typically, these models consist of planning across multiple time periods, wherein the decision-maker uses the present circumstances and predictions of future conditions to make decisions at each interval. (Thanh et al.2008).

Optimizing facility location involves selecting the best location for a business to establish its operations. The process of facility location optimization can be complex and involves

using mathematical models to determine the optimal location. When optimizing facility location models, there are several factors to consider, such as:

- Demand: The amount of demand for a product or service in a specific area is an important consideration for facility location. If demand is high, locating a facility closer to the customers may be beneficial to reduce transportation costs.
- Transportation costs: The cost of transporting goods and materials to and from a facility is a key factor in determining the optimal location of a facility. If transportation costs are high, locating the facility closer to suppliers or customers may be more cost-effective.
- Labor costs: The cost of labor can vary significantly between regions and countries. It is important to consider the availability of skilled labor and the cost of hiring and training employees when selecting a location for a facility.
- Country stability: The stability of a country or region can impact the security of a facility's operations, as well as the ability to conduct business in the region and the risk of political, social, and economic instability can impact the viability of a facility in a specific location.

Nowadays, many manufacturing companies invest in various countries worldwide to manufacture their products. There are multiple reasons behind such a strategy, such as benefiting from low wage rates, relaxed regulations, and accessing the local markets and regions (Kalantari,2013). Choosing the location for manufacturing products is a key aspect of strategic and logistical decision-making. Location may contribute to a company's success and give it a competitive advantage (McCarthy, 2003).

One factor that should be considered when optimizing facility location is "country risk". Acknowledging the term “country risk” in facility location is important. Country risk refers to the potential risks that a business may face when operating in a particular country. These potential risks in facility location can arise from various sources, such as changes in government regulations, political instability, financial risks, or social unrest (Howell, 2007). In other words, the location may no longer be able to support the facility's operations, leading to potential disruptions and losses. The instability of a country refers to the possibility that the government may become unstable due to changes in the environment, political or market conditions. Countries can be at different stages of “stability,” and based on the various pressures, causes instability (Ades and Chua, 1997). The instability of a country is a subject of active research, and previous studies have used several techniques to identify optimal methods for dealing with political events. To the best of the authors’ knowledge, most literature on this topic focuses only on political factors. Still, this research focuses not only on political but also social and economic factors when it comes to the country's stage. Political, social, and economic risks are important considerations regarding facility location decisions. According to Erb et al. (1996), political risk can be defined when political events, such as elections, wars, or policy changes, will negatively impact investments. Economic risk is defined as the risk that economic factors, such as inflation, interest rates, or economic growth, will negatively impact investments, and financial risk is defined as the risk that the financial performance of a company or investment will be worse than expected. It is essential to carefully evaluate each location's potential risks and benefits before making a final decision (Prakash et al. 2015). Therefore, considering country risk in FLP is vital to mitigate potential negative impacts on the

business and to ensure long-term sustainability and profitability. It is important to conduct thorough research and predict and analyze the country's risks before deciding on a facility location.

Facility location and country risk are two complex and interdependent factors that require careful consideration in any optimization model. When defining the objectives of the optimization model, it is important to consider both the short-term and long-term goals of the organization. For example, a company may prioritize minimizing costs in the short term but may also want to ensure that the chosen facility location is sustainable and will not create instability over time. To address this problem, a bi-objective optimization approach is employed. Bi-objective optimization is an optimization problem that involves finding the best possible solutions to problems with two conflicting objectives or criteria.

To optimize facility location while considering country risk, businesses need to thoroughly analyze the risks associated with operating in a particular country. This analysis should include an assessment of the political and economic stability of the country, as well as the potential for social unrest and other risks that may affect the business's operations. In this research, several ML models are proposed to predict the fragile state of a country using social, political, and economic factors.

ML has gained significant attention in recent years due to its ability to analyze complex data and accurately predict outcomes. ML can be used to analyze a wide range of data sources, including demographic data, customer data, and location data, to identify patterns and make predictions about the most optimal facility locations. ML has been applied in various fields, including economics, social, and political, to analyze the causes and consequences of instability and identify potential solutions.

States that are considered fragile are at a greater risk of facing domestic and international conflicts and unexpected disruptions. These states are also more likely to experience crises in one or more of their subsystems (CSRC, 2006). The Fragile states index (Fragile State Index, 2022), which is developed a long time ago and is widely used by policymakers, refers to measuring the collapse vulnerability of a country in a given time. It considers various indicators when scoring the countries. The countries are labeled into four categories in the developed ML models. The categories are based on the aggregate score in Fragile States Index (FSI), as being in the alert zone (90–120 aggregate score), warning zone (60–89.9), stable zone (30–59.9), and sustainable zone (29.9 or less) (Sekhear, 2010).

1.2 Problem Background

In this study, an MILP model is developed. Two objectives addressed by the model are minimizing the total production/distribution cost and minimizing the risk to the country of the operating facility.

In order to calculate the risk of a country, various factors such as political, social, and economic factors are considered that impact a country's stability at a given time. One hundred thirty-nine countries are studied from 2006 to 2021. Various ML models, namely Random Forest (RF) classifier, Artificial Neural Network (ANN) classifier, Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbor (KNN), are implemented and compared to predict the state of a county in a given year.

1.3 Research Objectives

- Develop an MILP with bi-objective optimization model minimizing production/distribution cost and risk factor of the supplier is developed.

- To determine the risk factor of suppliers' countries of origin, several ML models are developed to predict how fragile a geographic region is based on the potential impact of political, economic, and social factors.
- The combine ML and optimization approach can be deemed to be of benefit to decision makers in design and improvement of their supply chains.

1.4 Research Gaps and Novelty

The main contribution of this research is the combination of both ML approach and optimization model. The output of the ML model is used as one of the inputs (risk factor) to the optimization model. In addition, to the best of our knowledge, there is a dearth of work in literature that considers all political, social, and economic features used in this research to predict the fragile state of a country.

CHAPTER 2: LITERATURE REVIEW

This chapter covers two topics in literature namely facility location, and important political, social, and economic factors in determining the instability of a country including the use of ML approach.

2.1 Facility Location Problem (FLP)

The Facility Location Problem (FLP) is a well-known optimization problem in operations research and management science. The primary objective of the FLP is to minimize costs or maximize efficiency by strategically locating facilities in a way that minimizes transportation costs, satisfies customer demand, and takes into account other relevant factors. The problem considers both the geographic location and capacity location decisions. Thanha et al. (2008) present a facility location model for complex supply chains, which is dynamic in nature. The study addresses the multi-period, and multi FLP with deterministic customer demands. This article aims to assist with strategic and tactical decision-making processes, such as deciding whether to open or close facilities, selecting suppliers, and determining the flow of materials in the supply chain.

According to Hinojosa et al. (2000), the Multiple FLP falls under the NP-hard category. and involves determining the optimal locations for facilities, deciding which facilities should serve which customers, and calculating the optimal quantity of goods to be transported to minimize transportation costs. The problem also considers the possibility of setting up new facilities during the planning horizon and assigning facilities to locations. It is a capacitated, single commodity, multiple facilities, and multiple time-period location and allocation problem that considers changing customer demand over the planning

horizon. Gebennini et al. (2009) have researched optimization models for the dynamic facility location and allocation problem. Their study aim to create and utilize advanced MILP models to handle multi-stage, multi-commodity, and dynamic Location-Allocation Problems (LAP) in design and management. They suggest that a complete definition for the class of problem involving capacitated, single commodity, multiple facilities, and multiple time period (dynamic) facility location-allocation would be appropriate.

Musonera et al. (2009) propose an optimization model that considers various risk factors influencing FDI attraction, such as market size, labor costs, infrastructure, political stability, and economic policies. They use a linear programming technique to estimate the optimal combination of these factors that maximize FDI attraction. Their model is based on the idea that FDI tends to attract countries with low political, financial, and economic risk levels.

Cui et al. (2010) discuss the risk of disruptions in the context of facility location design. Disturbances such as natural disasters, transportation strikes, or power outages can significantly impact the ability of facilities to serve customers, leading to service disruptions, delays, or even complete shutdowns. These disruptions can result in financial losses, damage to the firm's reputation, and even loss of life in extreme cases. The authors argue that facility location design needs to consider the risk of disruptions and develop strategies to mitigate the impact of such disruptions. They propose a multi-objective model that considers the probability of disruption for each main facility and designs a system of backup facilities that can be activated when disruptions occur. The algorithm generates a set of solutions representing the trade-off between cost and reliability, and decision-makers can choose the solution that best fits their preferences.

The paper by Prakash et al. (2015) focuses on assessing risk associated with facility location in a global supply chain. The authors argue that supply chain risk management has become increasingly important due to the growing complexity and global nature of supply chains, making them more vulnerable to various risks such as natural disasters, political instability, economic downturns, and supply disruptions. The authors propose a grey-based approach for risk assessment to address this issue. Grey Relational Analysis (GRA) is a mathematical framework for dealing with systems that have limited or uncertain information. The authors apply the proposed approach to a hypothetical global supply chain scenario case study. The case study considers several factors that affect the risk associated with facility location, such as transportation costs, labor costs, infrastructure availability, political stability, and environmental risks. The authors use GRA to evaluate the relationship between each factor and the overall risk level associated with facility location. The authors suggest that the approach can be used by supply chain managers to make informed decisions about facility location and to develop risk mitigation strategies.

Akgün et al. (2015) propose an optimization model that minimizes the risk that a demand point may be subjected to because the located facilities do not establish it. To address this issue, the authors propose a risk-based approach to facility location that uses Fault Tree Analysis (FTA) to identify the factors that contribute to a disaster's occurrence and evaluate the effectiveness of potential emergency response facilities in mitigating those risks. The authors demonstrate the effectiveness of their approach through a case study of Istanbul, Turkey, where they use FTA to model the risks associated with earthquakes, floods, and fires, and then use an optimization model to identify the optimal locations for emergency response facilities. The authors note that their approach can be applied to other types of

disasters and can be used by decision-makers in the public sector and private companies involved in disaster management. They also suggest extending their approach to include other factors such as social and environmental impacts to provide a more comprehensive evaluation of emergency response facilities.

Hedaoo (2016) proposes a mathematical model for optimizing the supply chain network and uses a meta-heuristic algorithm to solve the model. The model considers factors such as facility capacity, transportation costs, and demand variability. The meta-heuristic algorithm finds an optimal solution that balances the trade-offs between these factors. He also presents a case study to demonstrate the effectiveness of the proposed model and algorithm in a real-world setting.

Habibi et al. (2017) formulates a multi-objective robust optimization model for the site-selection and capacity allocation of municipal solid waste (MSW) facilities. The proposed model aims to minimize the MSW management system's total cost and negative environmental impacts while considering uncertainties in the input parameters. The proposed model includes two main stages: (1) site selection and (2) capacity allocation. In the first stage, a multi-criteria decision-making approach is used to identify potential sites for MSW facilities based on technical, environmental, and social criteria. In the second stage, an optimization model is developed to allocate the MSW to the selected sites and determine the capacity of each facility to minimize the total cost and the negative environmental impacts of the MSW management system. The model is formulated as an MILP and solve using the robust counterpart approach to consider uncertainties in the input parameters.

Chen et al. (2021) propose a multi-objective optimization that considers both financial and non-financial factors, such as political risk, environmental impact, and social responsibility, to make the selection process more comprehensive. The authors first identify the key decision-making criteria based on the literature review and expert opinions. Then, they use the Analytic Hierarchy Process (AHP) to calculate the weights of each criterion and the International Country Risk Guide (ICRG) index to associate risk for the countries and a multi-objective genetic algorithm (MOGA) to find the optimal solution. The proposed method is applied to a case study involving four potential oil projects in different countries, and the results demonstrate that the proposed approach can effectively balance financial and non-financial factors to select the best overseas oil project. The study provides a useful tool for decision-makers in the oil industry to make more informed and comprehensive decisions when selecting overseas oil projects.

Chen and Lai (2022) propose a multi-objective optimization approach for determining the location and allocation of emergency medical service (EMS) facilities in rural areas. The approach considers two main objectives: minimizing the EMS system's total cost and maximizing the EMS facilities' population coverage. The authors use a genetic algorithm to solve the multi-objective optimization problem and conduct a case study in a rural region in China to demonstrate the approach's effectiveness. The results show that the proposed approach can generate a set of optimal solutions that provide a trade-off between the two objectives. The authors also perform a sensitivity analysis to investigate the impact of different parameters on the optimization results.

Maliki et al. (2022) develop a multi-objective optimization model that aims to minimize the costs of deploying and operating mobile facilities, maximize the population coverage,

and minimize the environmental impacts associated with facility locations. The model considers a set of decision variables that represent the locations and schedules of mobile facilities in each period, as well as a set of constraints that ensure the coverage of the population and the compliance with environmental regulations. To solve the proposed model, the paper presents a genetic algorithm-based approach that generates a set of non-dominated solutions. The algorithm uses a set of heuristics and operators to generate the initial population and evolve the population over multiple generations.

Forecasting country risk and FLP involves dynamic decision-making processes. Country risk is not a static phenomenon, and political, social, and economic conditions can change over time. Therefore, businesses need to continuously monitor and update their risk assessments, adapt their strategies, and adjust their facility locations accordingly. Similarly, facility location decisions may require a long-term perspective, considering potential changes in political stability, economic conditions, or regulatory frameworks that can impact the suitability of a particular location.

2.2 Risks Associated with Geographic Locations

Risk is an inherent and critical factor in decision-making processes across various domains, including FLP, and represents the uncertainty and potential negative consequences associated with different choices and outcomes. In the case of country risk, businesses or investors need to assess the potential risk or policy changes that may affect their operations or investments in a particular country or region. Similarly, in FLP models, companies need to make decisions on where to locate their facilities, considering various factors such as market conditions, infrastructure, and the potential risks associate with political, social and

economic instability or regulatory changes. In order to assess the risk of a country, the literature review examines various factors including political, social, and economic aspects. These factors are analyzed through different methods to forecast the level of risk.

Political factors cover elements such as the stability of the government, the level of corruption, political institutions, policy predictability, and legal frameworks. Alesina et al. (1996) define political instability as the propensity for a government to collapse. They estimate a model that consider both political instability and economic growth together. According to their study, growth is significantly lower in countries and time periods with a high likelihood of government collapse. Growth is affected by political instability because it increases policy uncertainty which negatively affects investment and savings decisions. According to Siermann (1998), political instability occurs when the government changes. This definition may result in a measure of political instability that is either underestimate or overestimate, depending on how it is applied. For instance, in the current state of Iran, most political scientists agree that the country is not politically stable, although according to this definition, Iran is considered politically stable since its government has not changed. However, Jong-A-Pin (2009) explain that instability of the political system has four dimensions: politically motivate violence, mass civil protest, instability within the political regime, and instability of the political regime. He use the GMM (Generalized Method of Moments) method and find that economic growth is affect differently by the four dimensions of political instability. As a result, only the instability of the political regime has a significant negative effect on economic growth, and instability within the political regime appears to increase economic growth.

Other studies, such as Aisen and Veiga (2013), define political instability by using cabinet changes, when a new premier is named several times a year and/or 50% or new ministers fill more of cabinet positions. They show the GDP per capita growth rate is lower when political instability is high. As a result, they find out that political instability significantly reduces economic growth by lowering productivity since human capital and economic growth are positively correlate.

Political instability can have significant effects on both social and economic factors within a country. Collier and Hoeffler (2004) research the civil war examining various factors such as primary commodity exports, GDP per capita, secondary schooling, per-capita income, economic growth, population, democracy score, time since the previous conflict, dominance of one ethnic group, social fractionalization, geographic dispersion. They use the data from 1960 to 1999. They develop a logit regression in their work, concluding that political and social variables have little explanatory power while economic variables have the highest explanatory power. The paper suggests that civil wars are caused by a combination of economic and political factors and that policies aimed at reducing poverty and promoting economic growth can help prevent conflicts in developing countries. However, the use of logit regression in their research limits the ability to establish causal relationships between variables. Additionally, the study's time frame (1960 to 1999) may not capture the contemporary dynamics of political risk, warranting an update with more recent data. According to Mustapha (2014), corruption harms GDP per capita. In this study, the main finding of their research is that corruption significantly negatively impacts GDP per capita. Their data cover 20 countries from 2003 to 2011. The author concludes that reducing corruption through policy interventions such as increase transparency and

accountability can increase economic prosperity. While the study suggests that only corruption negatively impacts GDP per capita, it is also possible that a higher GDP per capita leads to reduced corruption.

Murad and Alshyab (2019) study the effect of political instability on economic growth using Jordan as their case study. They evaluate various internal and external factors affecting the political instability from 1980 to 2015 and apply a ML technique called "fully modified ordinary least squares." They find that internal factors such as increased crime and cabinet changes would negatively impact economic growth. Controversially, they also find out that the degree of freedom negatively impacts GDP growth, meaning that as the level of freedom increases, economic growth significantly decreases.

Perles-Ribes et al. (2019) study the effect of political instability on the country's tourism. They consider Catalonia in Spain for their case study. They show that although Catalonia is one of the most attractive regions for visitors in Spain, due to the political instability in that region in late 2017, there is a significant decline in the number of visitors and their spending there. They conclude that political instability could negatively impact tourism even in a prosperous region like Catalonia. For their analysis, ARIMA univariate model is deployed. Their study focusing on only political instability may not capture the full range of factors that can influence tourism. Other variables, such as economic conditions, global travel trends, or natural disasters, can also impact tourism independently of political instability. It is important to consider and control these factors to isolate the specific impact of political instability.

Other study by Karnane and Quinn (2019) investigate the effect of corruption and ethnic fractionalization on economic growth by empirically evaluating them in two ways: 1) their

direct effects on economic growth and 2) their indirect effects through political instability. They study 157 countries from 1996 to 2014. They find that both corruption and ethnic fractionalization can adversely impact economic growth indirectly through increasing political instability.

Morrissey et al. (2020) consider household economic instability by focusing on family income and employment in the U.S. over the period between 2008 to 2013. Their results show that household economic instability is exceptionally high in families with low education backgrounds, families with little children, and those families who did not possess a home. They also find that public assistance income has the highest instability during economic recession, and they recommend that safety net programs consider the economic instability when planning the benefits and services. However, their study focus on the period between 2008 and 2013, which primarily encompasses the aftermath of the 2008 global financial crisis. While this period is relevant for understanding the impact of economic instability, it may not capture the full range of economic fluctuations and their effects on households over a longer time frame. Other recent study by Khafaga and Albagoury (2022) analyze the impact of political instability on economic growth in Ethiopia between 2005 and 2019. They use a method called Auto Regression Distributed lag by analyzing both direct and indirect effects. They find that political instability has an important influence on economic growth in both the short and long term. The study specifically focuses on Ethiopia, and the findings may not be easily applicable or generalizable to other countries or regions. Political instability can manifest differently in various contexts, and its impact on economic growth can vary significantly depending on

the specific circumstances and factors at play. Therefore, caution should be exercised when extrapolating the results to other settings.

Forecasting country risk involves predicting the potential risks and uncertainties associated with a particular country's economic, political, and social environment. ML is often used for risk assessment and involves training models on historical data and making predictions based on new data inputs. There have been numerous studies conducted on forecasting country instability using various approaches and methodologies. De La and Neckar (1988) discuss the importance of forecasting political risks for companies engaged in international operations. The authors propose a methodology for assessing and forecasting political risks using qualitative and quantitative methods and suggest using a set of country-specific and industry-specific factors to evaluate political risks. The country-specific factors include political stability, economic conditions, social conditions, and the legal environment. They use different features such as income inequality, social unrest, human rights issues, inflation rates, GDP growth, balance of trade, property rights, contract enforcement, and intellectual property protection. Goldstone et al. (2010) develop a model with a two-year lead time to separate unstable countries from stable countries based on the data gathered from 1995 to 2003. They find out that regime type has the most impact on the political instability of a country' through other variables such as infant mortality, discrimination, and lousy neighborhood effects are still important factors. They argue that once regime type is considered in the model, other social, political, and economic features would not impact their model. For the regime type, they use a 21-point Polity scale with five non-linear categorical features. The developed model has an accuracy level of above 80% in distinguishing countries concerning political stability and instability.

Basuchoudhary et al. (2015) proposed a ML approach to predicting state failure. The authors focus on predicting the onset of state failure, defined as the point at which a state experiences a significant and sustained decline in its ability to govern effectively. The authors develop a predictive model using a dataset of 162 variables derived from the World Bank, International Monetary Fund, and other sources. They use various ML algorithms, including RF, Gradient Boosting, and Support Vector Machines, to predict state failure in 37 countries. They find that ML approaches can significantly improve the accuracy of predicting state failure compared to traditional statistical methods. The authors report that their model can predict the onset of state failure up to two years in advance with high accuracy. Another notable finding of this paper is the identification of the most important predictors of state failure.

Li and Yao (2018) apply two ML methods, namely "Support Vector Machine" and "Gradient Boosted Regression Trees," to classify the countries based on their states for a period between 2007 and 2015. They argue that since both models have some weaknesses, they combine them together using an ensemble approach. However, combining multiple ML models through an ensemble approach is commendable, the study could benefit from a more comprehensive evaluation of the strengths and weaknesses of each model. Providing more insights into the decision-making process behind the ensemble approach would be valuable.

Sofuoğlu and Ay (2020) conduct a study to find the correlation between climate change and political instability. Their research consider 18 countries from the Middle East and North Africa from 1985 to 2016. They apply the "panel causality test" developed by Dumitrescu and Hurlin (2012). They find out that there is a cause-and-effect relationship

between climate change and political instability in 16 countries, and there is a conflicting relationship in 15 of the studied countries. They also find out the more stable and richer countries such as Saudi Arabia, UAE, and Kuwait have opposite relationships compare to other studied countries, which can be due to the spillover effect. They explain that the agriculture sector has an important role in the GDP of the studied countries, and climate change will negatively impact agriculture and, consequently, their GDP.

Baillie et al. (2021) propose a minimal forecasting model of political instability. Three predictors are used in their model: polity code (a measure of government type), infant mortality, and years of stability (i.e., years since the last instability event). They use a logistic regression model for the data from 1976 to 2017. They find that the most important feature that impacts political instability is "polity code" and predict that infant mortality and years of stability increase the likelihood of instability.

Matta et al. (2022) examine 38 regime crises between 1970 to 2011 to check the effect of non-violent political instability on the economy. They divide those regime crises into two categories: those that accompany mass civil protest and those in which civil protest is absent. They apply a synthetic control approach and find that those regime crises accompanied with mass civil protests has an adverse impact on the economy of the country which not only last for more than five years but tend to increase over time. In contrast, those regime crises, in which civil protest is absent, have no significant impact on economic growth.

Bittencourt et al. (2022) develop a model call "overlapping generations (OLG)" with socio-political instability with features including strikes, riots, and protests. They show that socio-political instability has a direct positive relationship with inflation. They apply their

model to data from 156 countries from 1980 to 2012 from the World Bank. Other features use in their model include GDP per Capita, Cabinet Change, and the Gini coefficient of income inequality. They find out that the more affluent countries have lower socio-political instability.

Khudari et al. (2023) examine the impact of political stability and macroeconomic factors on FDI. They consider Turkey for their case study from 1974 to 2017. They apply the Auto-Regressive Distributed Lag (ARDL) model and show a positive correlation between GDP per capita, political stability, and direct foreign investment. They also discover that pollution, energy consumption, and inflation can negatively affect direct foreign investment.

In summary, these studies have made valuable contributions to understanding political, social and economic risks, but there is room for improvement in terms of methodology, data sources, inclusion of additional factors, and broader analysis of political risks beyond economic growth.

In this research, selected features are used to predict the risk of a country including, growth GDP rate, employment to population ratio, GDP per capita income, levels of education, population, individuals using the internet, life expectancy, the level of democracy, voice and accountability and regime type. The label "Fragile State Index" (Fragile State Index, 2022) is used, which categorizes the countries into four groups, namely "Alert," "Warning," "Stable," and "Sustainable." To the best of our knowledge, no work in literature considers all the addressed features together. Also, no work in literature has studied the impact of the addressed features on the label of a country belonging to one of the four mentioned categories. Furthermore, various ML models and approaches are provided in

this research to find the essential features in categorizing a country. Finally, the model's output is used as the risk factor input to the developed MILP model, which has a bi-objective function of minimizing production/distribution cost and risk of the facility location.

Researchers can advance our knowledge and provide more comprehensive insights into the complex nature of a country's risks. Such advancements can ultimately support policymakers, businesses, and society at large in effectively managing and mitigating the potential challenges posed by potential risks.

CHAPTER 3: DEVELOP MACHINE LEARNING APPROACHES

In this chapter, multiple supervised ML algorithms are developed to predict the fragile state. The objective is to increase the model accuracy, precision, and F1-score. Steps are involved in using supervised ML include:

1. Data Collection: The first step is collecting relevant data about the country's political, economic, and social situations. The data is from open source (World Bank Group Archives), and we borrowed the labels from Fragile State Index, 2022. Based on the literature reviews and the availability of the data, we selected 14 features related to political, social, and economic factors.
2. Data Preparation and Feature Selection: Once the data is collected, it must be preprocessed and cleaned to remove noise, missing values, and irrelevant features. In this step, we used Hot-Deck imputation to impute the missing values, variance inflation factor (VIF), to remove the multicollinearity among the independent feature, balancing the data and feature scaling.
3. Model Training: After selecting the features, the next step is to choose multiple appropriate supervised learning algorithms that could effectively predict the country's state and train each model using the prepared data. This step involves splitting data into the train and test (80:20), feeding the algorithm with the train data, and using Grid-Search to optimize the hyperparameters. In this research, we used algorithms for multiclass classification problems that include K-nearest neighbors (kNN), Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN).

4. Model Evaluation: After training the model, it must be evaluated using various performance metrics such as accuracy, precision, recall, F1-score.
5. Model Prediction: After choosing the best model based on its performance, it is used for future predictions. This utilization of the selected model for future predictions is described in Chapter 4 within the test case in section 4.2.

At the end of this chapter, an unsupervised ML approach is also presented.

3.1 Data Collection

In this study, country-year data is utilized due to its availability and potential for replication by others. Fourteen indicators are selected for analysis, with data obtained from the official website of the World Bank that offers comprehensive information on indicators for various countries and regions globally. Independent variables used in different studies typically belong to economic, social-demographic, and political categories. The dataset used in this study comprises 2224 country-year observations, covering the period from 2006 to 2021 and representing 139 countries in an unbalanced panel. To facilitate predictions, the dataset is randomly divided into a testing sample consisting of 1779 observations (around 80% of the entire dataset) and a separate test sample of 445 observations.

3.1.1 Independent Variables

Economic Variable:

- **Growth GDP rate:** Annual percentage growth rate of gross domestic product. In order to determine how healthy the economy is, we measure GDP growth. Positive

numbers indicate that the economy is growing. Negative numbers indicate a contracting economy.

- **GDP per Capita (US dollar):** GDP per capita is the gross domestic product divided by midyear population in dollars. The GDP per capita measures output per person, which is an indirect measure of income per capita. It is expected to impact the stability of the countries positively.
- **Employment to Population Ratio, 15+, Total (%):** A country's employment rate measures how many citizens are employed. High employment rates indicate that many people are employed. The working-age population is generally considered to be those 15 and older. In some cases, however, a lower employment-to-population ratio can be considered a positive trend, especially for young people, if it results from a higher education level.
- **Import:** The value of goods and services a country receives from other nations is represented by its imports and includes the value of various items such as merchandise, transport, insurance, royalties, license fees, and other market services like communication, construction, financial, personal, and government services. However, the calculation does not consider employee compensation, investment income, and transfer payments. All the data are calculated in current U.S. dollars.
- **Export:** The value of goods and services a country provides to other nations is represented by its exports and includes the value of various items such as merchandise, transport, insurance, royalties, license fees, and other market services like communication, construction, financial, personal, and government services. However, the calculation does not consider the compensation of employees,

investment income, and transfer payments. All the data are calculated in current U.S. dollars.

- **Foreign Direct Investment (FDI):** FDI refers to the net inflows of investment made by an entity, where the intention is to acquire a lasting management interest of 10 percent or more of the voting stock in an enterprise that operates in an economy other than that of the investor. Inflow of investment is calculated as the sum of equity capital, reinvestment of earnings, additional long-term capital, and short-term capital and is reflected in the balance of payments. The FDI data series displays the total net inflows of foreign direct investment. All the data are expressed in current U.S. dollars.

Social Variable:

- **Life Expectancy at Birth:** An estimated lifetime at birth indicates how long an infant is expected to live if the mortality patterns at birth remain the same as they are now. Life expectancy at birth reflects a population's happiness based on factors other than economics. It can be argued that the longer the life expectancy, the better the population's living conditions will be.
- **Secondary School Enrolment:** The value is calculated by dividing the number of students at official school age enrolled in secondary education by the population of the age group multiplied by 100. Individuals using the Internet (% of the population): Individuals who use it (regardless of where they are) are considered Internet users. In addition to computers, mobile phones, PDAs, gaming systems, and digital TVs, the Internet can also be accessed via various devices.
- **Population:** Indicate the total population.

Political Variable:

- **Regime Type:** Scores are assigned to closed autocracies (score 0), electoral autocracies (score 1), electoral democracies (score 2), and liberal democracies (score 3). Larger numbers indicate regimes are more democratic. In closed autocracies, citizens cannot elect the government's chief executive or the legislature by way of multi-party elections. In electoral autocracies, citizens have the right to choose the chief executive and legislature. However, they lack some of the freedoms that make elections meaningful, free, and fair, such as freedom of association. In electoral democracies, the right to meaningful, free, and fair elections is guaranteed. In liberal democracies, citizens have additional rights as individuals and as minorities, and legislative and judicial checks limit the executive's actions.
- **Democracy Score:** This variable indicates that freedoms of association and expression are guaranteed while political leaders are elected through free and fair elections. Indexes range from 0 to 1 (most democratic). Countries with more democratic get a higher democracy score.
- **Control of Corruption:** A corruption offense is when a public official uses their position for private gain, including bribery and theft of public funds. Corruption is more prevalent in countries where institutions are weak and often affected by fragility and conflict. According to the estimate, a country's aggregate indicator score will range from approximately -2.5 to 2.5 in units of a standard normal distribution. Higher values indicate that government can control corruption.

- **Voice and Accountability:** Measures how much citizens can participate in selecting their government, as well as their freedom of expression, association, and access to information. An aggregate indicator's percentile rank represents a country's rank among all countries covered, ranging from 0 to 100.

3.1.2 Dependent Variables (Class Label)

The dependent variable in this study, known as the "Class Label," is obtained from the Fragile States Index (FSI) produced by Foreign Policy magazine and the Fund for Peace since 2006 (Fragile State Index 2022). The FSI assigns scores to countries based on their overall performance. Sekhar (2010) classified countries into different zones based on their aggregate FSI scores: Alert zone (90–120), Warning zone (60–89.9), Stable zone (30–59.9), and Sustainable zone (29.9 or less). Figure 3.1 illustrates the distribution of these classes among the 139 countries from 2006 to 2021.

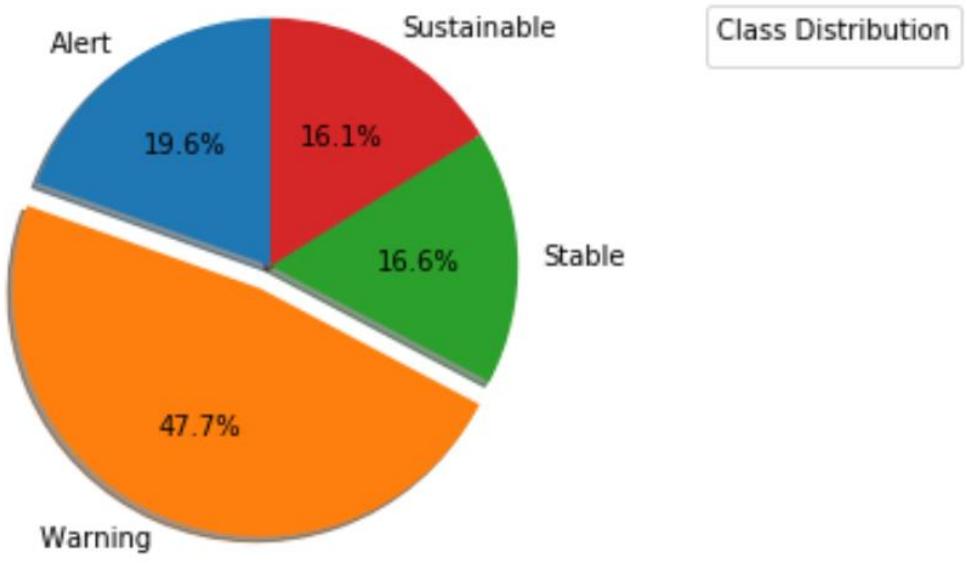

Figure 3.1 Class Distribution of Dataset

3.2 Data Preprocessing

An important aspect of ML involves data preprocessing and feature selection, which can significantly impact the model's performance. Data preprocessing is important as it ensures the accuracy, completeness, and meaningfulness of the data used for analysis. The selection of appropriate data features plays a crucial role in determining the success of ML models. This section focuses on identifying and imputing missing data, addressing multicollinearity using the variance inflation factor, balancing the imbalanced dataset, and applying feature scaling to normalize the dataset. These steps are vital for improving the quality and effectiveness of the ML process.

3.2.1 Imputation of Missing Data

Dealing with missing values posed a significant challenge when working with the data. Numerous ML algorithms are unable to handle missing values, requiring data scientists to determine the most suitable strategy for addressing missing values. Table 3.1 provides the percentage of missing values for each feature, while Figure 3.2 visually represents the distribution of missing data among the countries under study. The features with the most missing values are school enrollment, followed by the employment.

Table 3.1 The Percentage of Missing Values for Each Feature in the Dataset

Foreign Direct Investment	7.46	GDP Per Capita	0.58
Export	5.03	Population	0
Import	5.02	Individuals Using the Internet	6.42
GDP Growth	0.62	Secondary School Enrolment	30.66
Employment	10.79	Life Expectancy at Birth	5.35
Democracy Score	0	Voice and Accountability	0
Control of Corruption	0	Regime Types	0

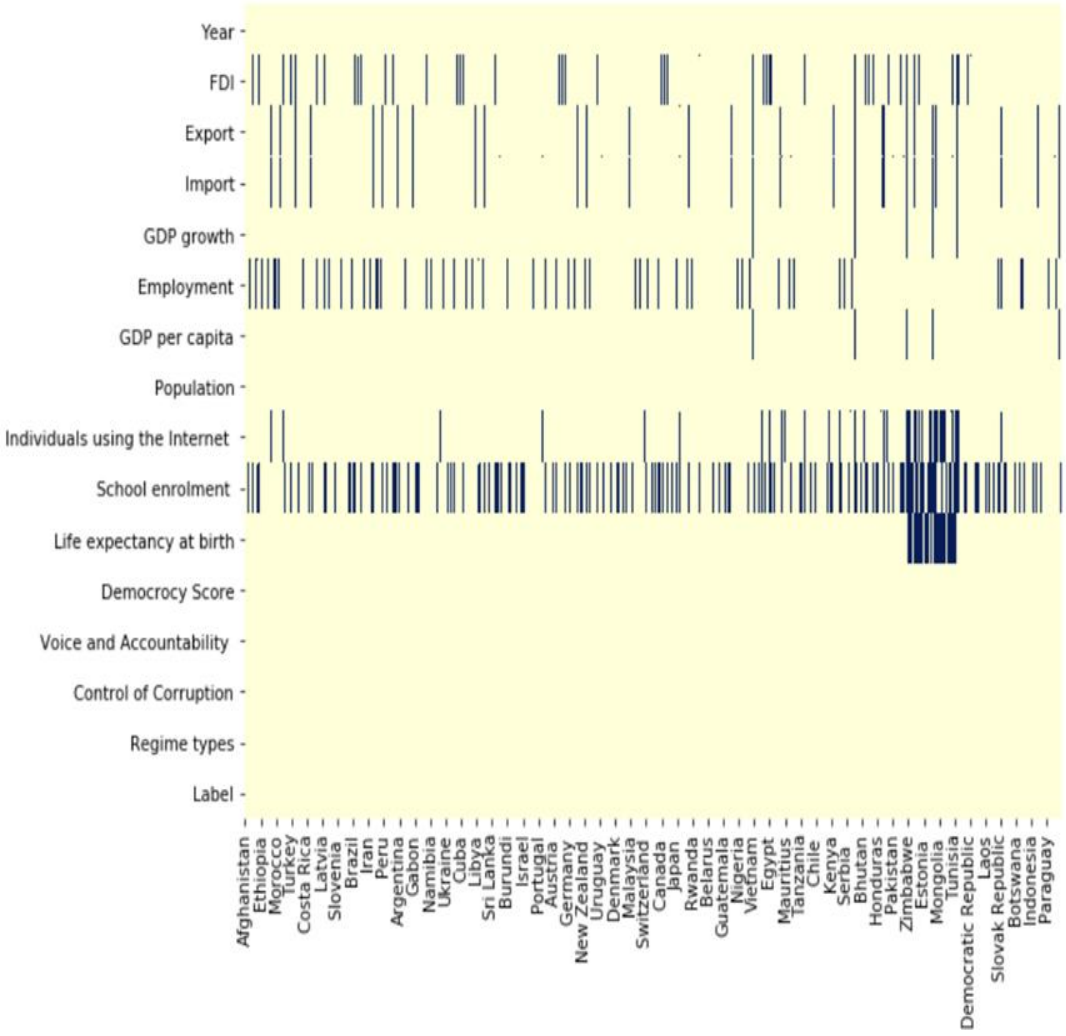

Figure 3.2 Distribution of Missing Values in Dataset

For the data imputation process, the Hot-Deck imputation method is employed. Hot-Deck imputation involves replacing missing data with values obtained from similar or identical cases within the dataset. Hot-Deck imputation utilizes the value of the most comparable observed data point to impute missing values, as described by Strike et al. (2001). The provided Pseudo-Code outlines the sequential steps involved in the imputation process.

Pseudo-Code of the proposed Hot-Deck imputation

1: Set k as the total number of potential donors used for imputing each missing value. In our case, we used k=1.

2: Read the raw data and convert it to a Data Frame (D).

3: Divide the raw data into two groups: complete data and missing data:

• M: Set of data with missing values-at least one column of each data point in this set is missing ($M \subseteq \boldsymbol{D}$)

4: Determine the type of our original data (D): (Numerical, Categorical, or Mix)

5: For each missing data point x in set M ($\boldsymbol{x} \in \boldsymbol{M}$) store its index (row number) and its missing columns

6: For each missing data point x in set M from Step 5:

- 6.1: Based on the data type from Step 4:
  - if the data type is Numerical only:
    - Calculate the Nan-Euclidian distance between each missing data point x in set M and every data point in D and go to step 6.2.
  - else if the data type is Categorical only:
    - Calculate the Hamming distance between x and every data point in D and go to step 6.2.
  - else (the data type is a Mix of categorical and numerical):
    - Calculate the Gower distance between x and every data point in D, excluding missing columns, and go to step 6.2.
- 6.2: Sort the calculated distances in ascending order and put them in list L.
- 6.3: Select the closest index in list L (call it y) to x (i.e., highest similarity to x) and replace missing columns of x with their corresponding in data point y.

7: Repeat Step 6 until all missing data points are imputed.

3.2.2 Variance Inflation Factor (VIF)

In this section, the presence of multicollinearity among the features is examined. Modifying one variable in a model will inevitably affect another when independent variables exhibit a high correlation. Therefore, it is crucial to ensure the independence of the independent variables before proceeding with data training. To achieve this, the Variance Inflation Factor (VIF) is utilized to assess the correlation of a single independent variable with a group of other variables. VIF helps to detect and address the presence of multicollinearity features. (Rawlings et al. 1998). Equation (3.1) shows the formula for calculating the VIF.

$$VIF_i = \frac{1}{1 - R_i^2} \qquad (3.1)$$

In Equation (3.1), R_i^2 is the unadjusted coefficient of determination for regressing the *i*th independent variable on the rest of the variables. Generally, any VIF greater than ten is considered high multicollinearity (Kutner et al. 2004). Some researchers consider a VIF of 5 as the cut-off point for high multicollinearity (Faraway, 2002). However, this research considers a VIF of 10 as the cut-off point. Table 3.2 shows the VIF number for each feature. At each step, the feature with the highest VIF greater than ten is eliminated from the list of features.

Table 3.2 VIF Calculation of the Initial Features

Features	VIF
GDP Growth	1.1074
Employment	0.9087
GDP Per Capita	3.9567
Population	1.9350
Individuals Using the Internet	3.6846
Secondary School Enrolment	3.3435
Life Expectancy at Birth	0.8075
Democracy Score	7.4332
Voice and Accountability	12.3166
Control of Corruption	5.0070
Regime Types	6.0253
Import	29.8353
Export	**33.0350**
Foreign Direct Investment	1.1525

According to Table 3.2, “Export” has the highest VIF among others. Table 3.3 shows the newly calculated VIFs after removing the “Export” feature:

Table 3.3 Second Run VIF Calculation of the Features

Features	VIF
GDP Growth	1.1056
Employment	0.9084
GDP Per Capita	3.8869
Population	1.7841
Individuals Using the Internet	3.6711
Secondary School Enrolment	3.3434
Life Expectancy at Birth	0.8043
Democracy Score	7.4298
Voice and Accountability	**12.2884**
Control of Corruption	4.9682
Regime Types	6.0163
Import	2.3853
Foreign Direct Investment	1.1520

With each feature removal, the VIF value for each variable decreases. The feature with the next highest VIF value, "Voice and Accountability," is excluded. Table 3.4 shows the third run of the VIF calculation.

Table 3.4 Third Run VIF Calculation of the Features

Features	VIF
GDP Growth	1.1055
Employment	0.9066
GDP Per Capita	3.8605
Population	1.7658
Individuals Using the Internet	3.6333
Secondary School Enrolment	3.3391
Life Expectancy at Birth	0.8042
Democracy Score	4.9421
Control of Corruption	4.0101
Regime Types	5.2747
Import	2.3538
Foreign Direct Investment	1.1504

After excluding each feature with a VIF > 10, Table 3.4 indicates the features with the final VIF; since there is no VIF of more than ten, there is no multicollinearity between all independent variables.

3.2.3 Class Imbalance

A dataset with an unequal distribution of classes is considered to have an imbalanced classification. An imbalanced classification dataset has skewed class proportions. Oversampling and undersampling are ways to address imbalanced datasets. (Shelke et al. 2017). When datasets are almost balanced, most ML algorithms perform better. In order to address the imbalanced class problem, various techniques of undersampling and oversampling can be used. This research employed the SMOTH method for oversampling,

while the NearMiss technique is utilized for undersampling. Figure 3.3 indicates the oversampling and undersampling method.

3.2.3.1 Undersampling Method (NearMiss)

Using undersampling techniques, examples belonging to the majority class are removed from the dataset to achieve a more balanced distribution of classes. Due to the smaller data size, undersampling reduces the time it takes to learn. NearMiss method (Bao et al. 2016) is used for undersampling. In NearMiss sampling, examples are selected based on distance from minority class examples. The data is reduced to 1432 samples after undersampling.

3.2.3.2 Oversampling Method (SMOTH)

Oversampling of minority groups is one way to address imbalanced datasets. SMOTE technique involves generating synthetic samples for the minority class (He et al. 2008). When SMOTE is applied, examples close together in the feature space are selected, a line is drawn between them in the feature space, and a new sample is drawn at the point along that line. SMOTE function generates 4244 samples.

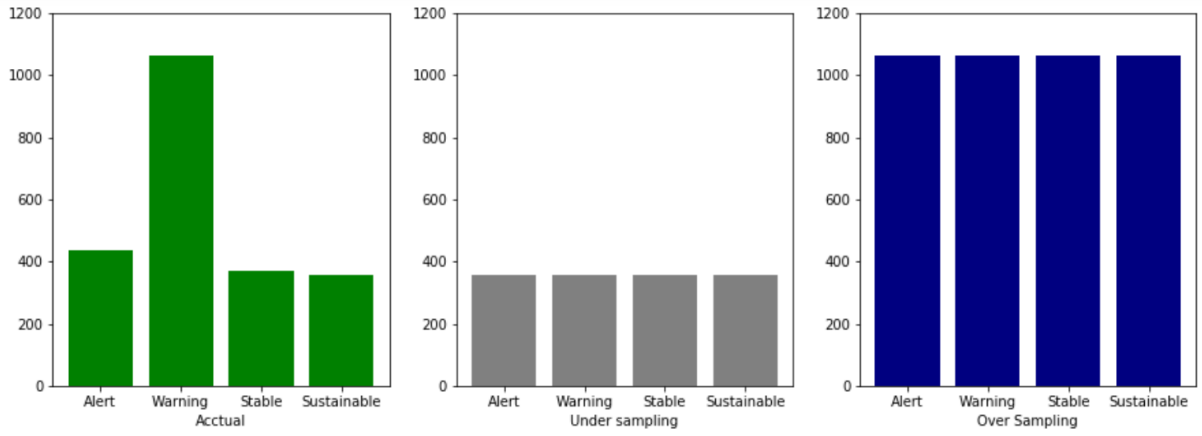

Figure 3.3 Over and UnderSampling

3.2.4 Feature Scaling

The final step before training the model is to scale the numerical variables to the standard range. Scaling refers to normalizing the values of different features within a dataset so that they are on a similar scale. Feature scaling is important because many ML algorithms perform better when the input features are normalized, allowing them to more accurately compare the relative importance of different features when making predictions.

In this research, the StandardScaler method is employed to scale the features. StandardScaler standardizes the features by subtracting the mean and scaling to achieve unit variance. Consequently, the mean of each feature is transformed to 0, while the standard deviation is transformed to 1. The normalization process results in features with a similar scale, facilitating effective comparison by ML algorithms. Additionally, StandardScaler has the advantage of transforming the feature distribution to approximate a normal distribution, which can be beneficial in certain scenarios.

3.3 Model Training

In ML, training is the most important step. ML models are trained by inputting prepared data, allowing them to discover patterns and make future predictions. The model learns from the data through this iterative process and improves its ability to perform the specified task. Over time, the trained model becomes more accurate in its predictions.

This research employed the "train-test split" procedure to fit and evaluate ML learning models. The typical split ratio is 80:20, indicating that 80% of the dataset is allocated to the training set, while the remaining 20% is assigned to the testing set. To ensure the

reproducibility of results, a random state is set, ensuring consistent outcomes each time the code is executed. For the train-test split, a random state value of one is employed.

Hyperparameters play a critical role in determining the performance of a model. The candidate models considered for the multi-class classification in this study encompassed KNN, RF, XGBoost, and ANN. To identify the optimal hyperparameters for each model, the GridSearchCV method is utilized.

3.3.1 K-Nearest-Neighbor (KNN) Classifier

K-Nearest-Neighbor (KNN) classifier is based on learning by analogy and is supervised learning. When given an unknown sample, a KNN classifier searches the pattern space for the KNN that is closest to the unknown sample. The unknown sample is assigned to the most common class among its K (Guo et al. 2003). Closeness is defined in terms of distance. Finding the best 'k' for a K-NN algorithm, which determines the class by majority vote, is essential. In the KNN model, the number of neighbors to inspect is a hyperparameter.

The variance of an algorithm is controlled by K in KNN, where small values, such as K=1, result in high variance (overfitting), and large values result in low variance. Figure 3.4 indicates the training and testing accuracy with different numbers of K.

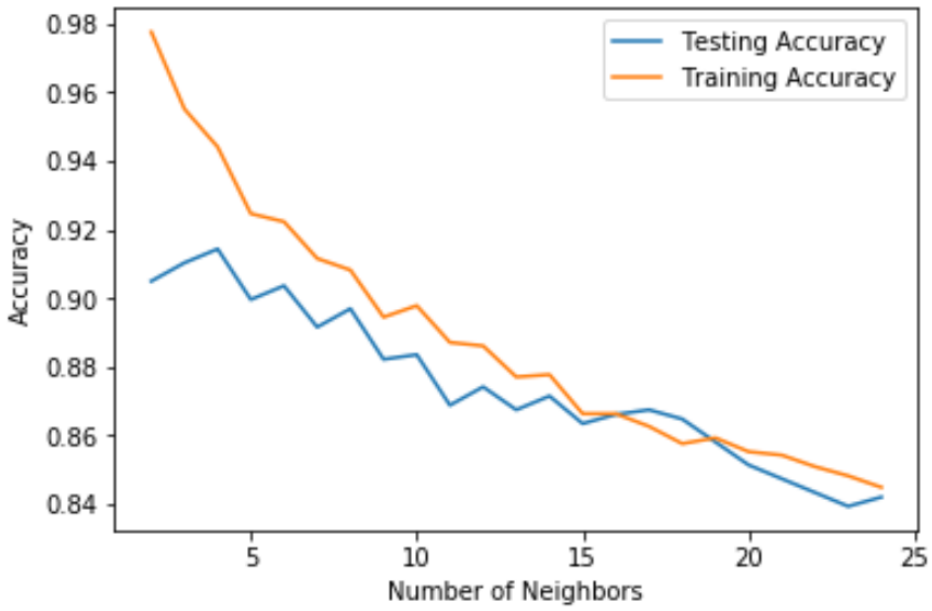

Figure 3.4 Result of Different K Values in KNN

3.3.2 Random Forests (RF) Classifier

RF is a supervised learning algorithm and combines several trees where each tree classifier uses a random vector (Breiman, 1996).

RF, also known as a bagging model, creates decision trees and estimates their predictions in parallel. The impurity present in each group is calculated by entropy and information gain. We can also use the RF algorithm to find important features in our dataset.

The Entropy of a group of observations is an information-theoretic metric that measures the impurity present in that group. Decision trees choose how to split data based on entropy factor. Equation (3.2) demonstrates how the entropy function is calculated:

$$Entropy = -\sum_{i=1}^{N} p_i \log_2 p_i \qquad (3.2)$$

where p_i is the probability of randomly selecting an element of class i. Information gain can be defined as a measure of how much information we can gain from a feature about a class. GridSearchCV utilized hyperparameters for RF, including a n_estimator value of 200 (indicating the number of trees) and the entropy as a criterion.

By reducing the max_depth parameter, one of the decision trees from the RF model is visualized. The top main nodes and branches of the RF model are shown in Figure 3.5.

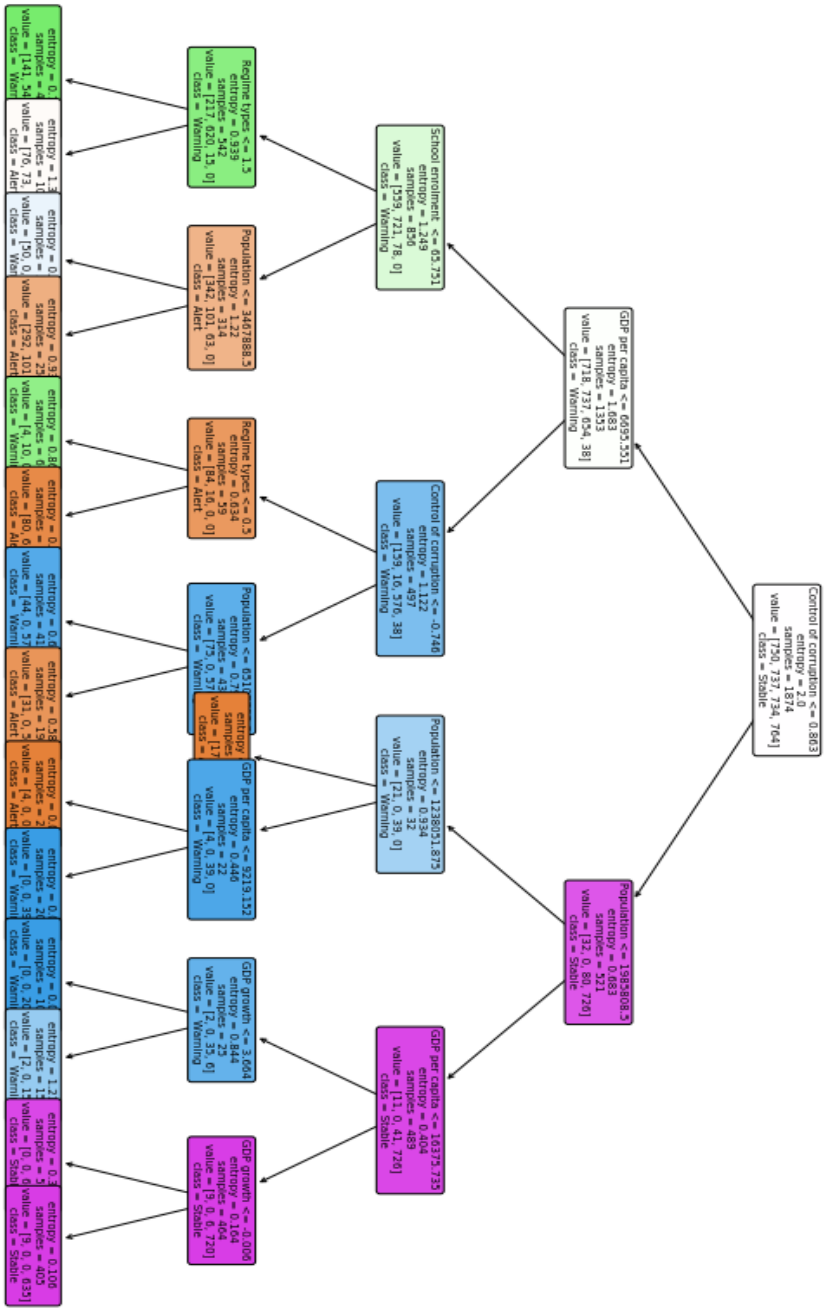

Figure 3.5 Top Main Nodes and Branches of the RF.

RF provides a measure of feature importance, indicating the contribution of each feature in the model's predictions. Feature importance can be valuable for feature selection and understanding the underlying patterns in the data.

RF provides different methods for calculating feature importance, including the mean decrease impurity, mean decrease accuracy, and permutation importance. In this research, we used the Mean Decrease Accuracy method to find the impotence of each feature. RF feature importance (Figure 3.6) suggested that “Control of Corruption” and “GDP per Capita” are the most impacting factors in predicting the labels.

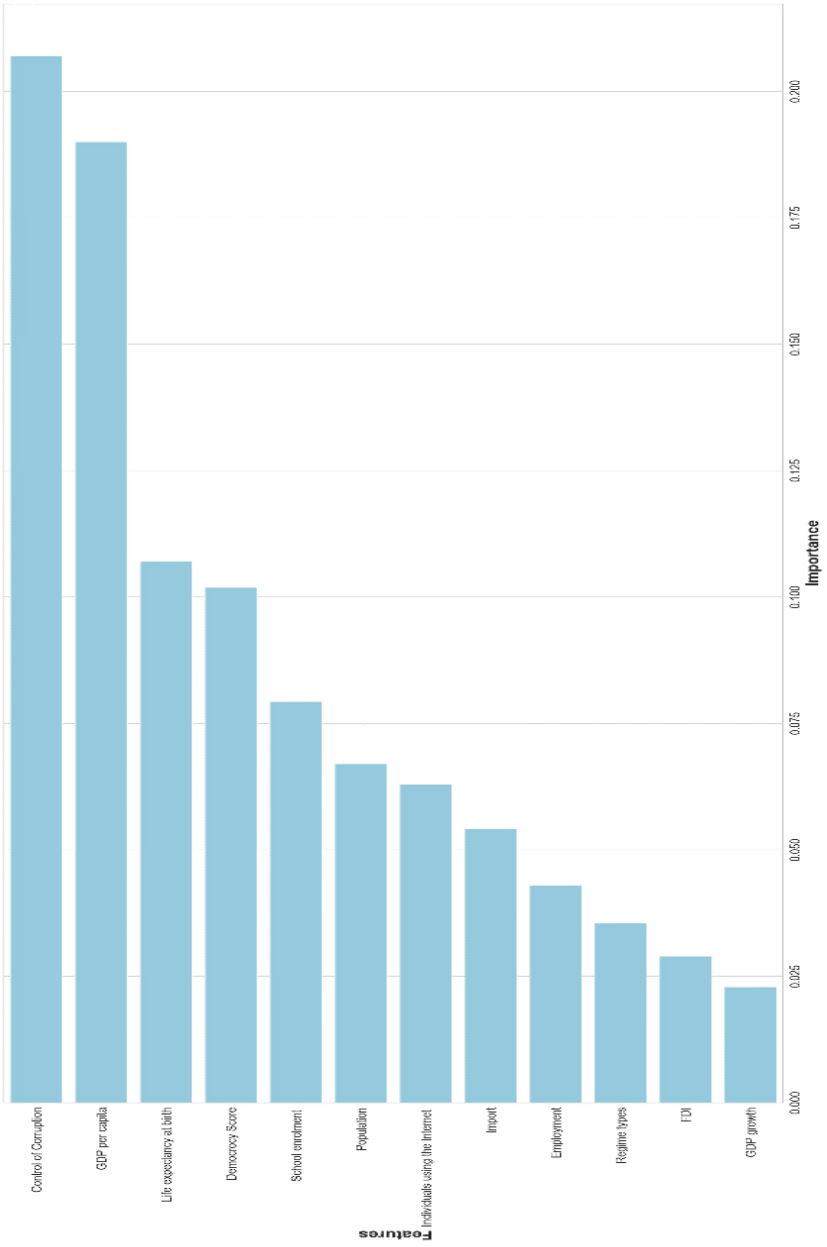

Figure 3.6 RF Feature Importance

3.3.3 Extreme Gradient Boosting (XGBoost) Classifier

XGBoost can be used to solve regression, classification, and ranking problems. The algorithm is a sequential model, which means that it starts by training a decision tree where all observations are given equal weights, and the tree is dependent on the outcome of the last tree (Pan, 2018). The drawback of XGBoost algorithm is that it takes longer time to generate classifier outcomes. RF and XGBoost both construct models based on multiple decision trees, however, there is a difference in the way trees are constructed and combined.

The main hyperparameters that affect the complexity of XGBoost are the number of trees (i.e., iterations), the maximum depth of each tree, the learning rate, and the regularization parameters. Choosing the optimal combination of these hyperparameters can require significant experimentation and tuning, particularly for large and complex datasets. Figure 3.7 shows the hyperparameters find by GridSearchCV for XGBoost model.

```
xgb_model= xgb.XGBClassifier(eta=0.2,gamma=.01,max_depth= 12,
                min_child_weight=1,subsample=0.5,
                objective='multi:softmax',n_estimators=200)
```

Figure 3.7 Hyperparameter for XGBoost

3.3.4 Artificial Neural Networks (ANN)

A neural network for multi-class classification is an artificial neural network designed to classify input data into multiple categories or classes. In contrast to binary classification, which involves only two possible outcomes, multi-class classification involves three or more possible outcomes. In ANN for multi-class classification, the input data is typically represented as a vector of numeric values fed into a series of interconnected layers of artificial neurons. Each neuron in the network receives input from the neurons in the

previous layer and uses information to calculate its own output (Anderson,1995). The output is then passed on to the neurons in the next layer, and so on until the final output layer is reached.

Neural networks are defined as comprising neurons, which are basic components that receive a real value, multiply it by weight, and apply a nonlinear activation function, as represented by Equation (3.3) (Kröse et al. 1993).

$$y = f\left(\sum_{i=1}^{n} x_i\, w_i\ + b\right) \qquad (3.3)$$

One commonly used ANN architecture for multi-class classification is the softmax function. The softmax function is applied to the network's output layer and converts each node's output into a probability value between 0 and 1. The probabilities for all the output nodes sum up to 1, which ensures that the network's output is a valid probability distribution over all possible classes (Bishop, 1995).

The final output layer of a neural network for multi-class classification typically consists of one neuron for each possible class. Each output neuron generates a probability score for its corresponding class based on the input data and the network's learned weights and biases. The class with the highest probability score is then considered to be the predicted class for the input data.

In ANN, hyperparameters are the parameters that are set before training the model, which cannot be learned during training. These parameters control the learning process and model architecture, and they can significantly impact the model's performance. Choosing the right hyperparameters can be challenging, and it often involves a combination of trial and error

and expertise. It's important to perform hyperparameter tuning to find the best combination of hyperparameters that result in the highest performance on a test set. In this research, we used Grid search to find the hyperparameter for ANN (Figure 3.8).

```
model=Sequential()
model.add(Dense(250, activation='relu',kernel_initializer='he_normal',input_shape=(x_train.shape[1],)))
model.add(Dropout(0.3))
model.add(Dense(100, activation='relu',kernel_initializer='he_normal'))
model.add(Dropout(0.3))
model.add(Dense(50, activation='relu',kernel_initializer='he_normal'))
model.add(Dropout(0.3))
model.add(Dense(25, activation='relu',kernel_initializer='he_normal'))
model.add(Dropout(0.3))
model.add(Dense(10, activation='relu',kernel_initializer='he_normal'))
model.add(Dense(7, activation='relu',kernel_initializer='he_normal'))
model.add(Dense(4, activation='softmax'))
```

```
opt=Adam(learning_rate=0.0001)
model.compile(optimizer=opt, loss=SparseCategoricalCrossentropy(from_logits=True), metrics=['accuracy'])
history=model.fit(x_train,y_train,epochs=250, verbose=1, validation_data=(x_test,y_test),batch_size=4)
```

Figure 3.8 Hyperparameter for ANN

3.4 Models Evaluation

In order to determine how well a trained model will perform on new, unseen data, it is common practice to split the available data into separate training and testing sets. The model is trained on the training set, and its performance is evaluated on the testing set. Figure 3.9 shwas the accuracy performance for KNN, RF and XGBoost in actual data size, undersampling and oversampling methods. Based on the results, oversampling of RF classifier model achieved the highest accuracy of 96%, indicating that it is the most effective model for this particular problem and dataset.

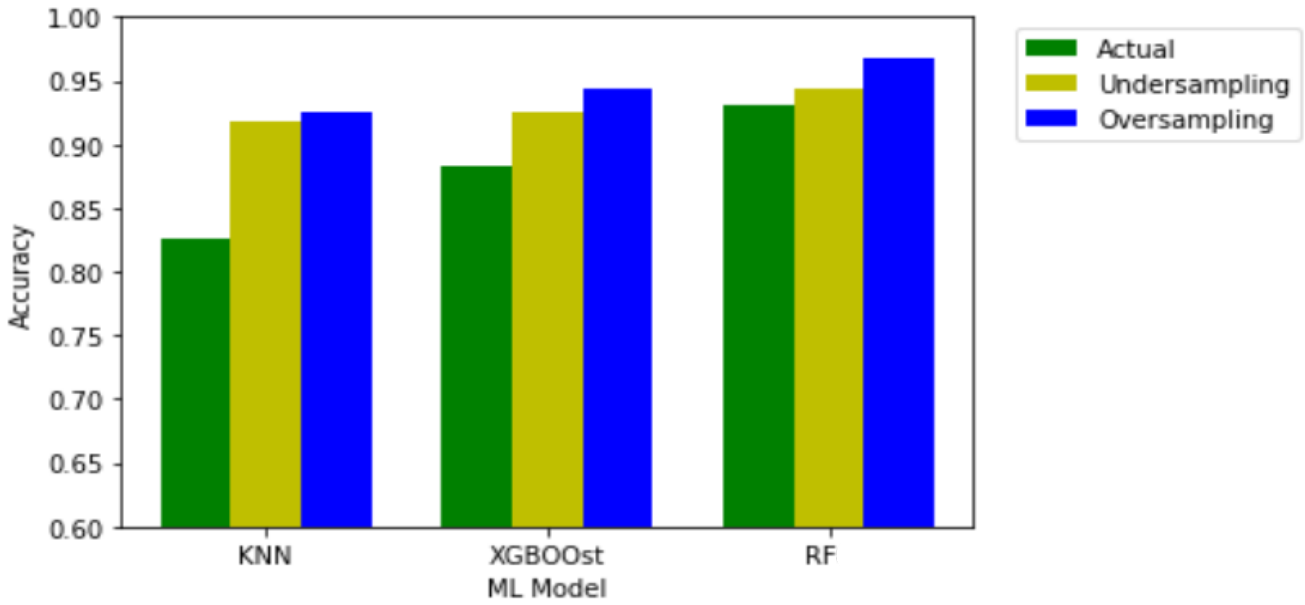

Figure 3.9 Performance for KNN, RF, GBDT

3.4.1 Confusion Matrix

Defining the performance of a classification algorithm is done by creating a confusion matrix (Provost and Fawcett 2001). Table 3.5 shows the confusion matrix for RF classifier.

Data labeled as belonging to an actual class is considered True Positives (TP). For example, the TP for the Alert label is 203, meaning the model correctly classified 203 Alert labels in the test set. The green boxes in Table 3.5 show the TP for each label.

Table 3.5 Confusion Matrix, TP

	Prediction Label			
Actual Labels	Alert	Warning	Stable	Sustainable
Alert	203	12	0	0
Warning	9	202	4	1
Stable	0	4	203	0
Sustainable	0	0	4	207

When the model predicts the negative class correctly, it is called a True Negative (TN). False Positive (FP) refers to positive outcomes the model predicted incorrectly. Type I error is also referred to as FP. For example, the FP for the Alert label is 9+0+0=9.

False Negative (FN) is an outcome that the model incorrectly predicted or the number of misclassified negatives. For example, FN for Alert label is 12+0+0=12. FN is also referred to as a type II error.

The previous tables show the confusion matrix for RF, and we highlight the TP, TN, FP, and FN. Additionally, the normalized confusion matrix for XGBoost, is illustrated in Figure 3.10. The next section explains how to calculate the Accuracy, Precision, Recall and F1-score values for these classes to check the performance of RF model.

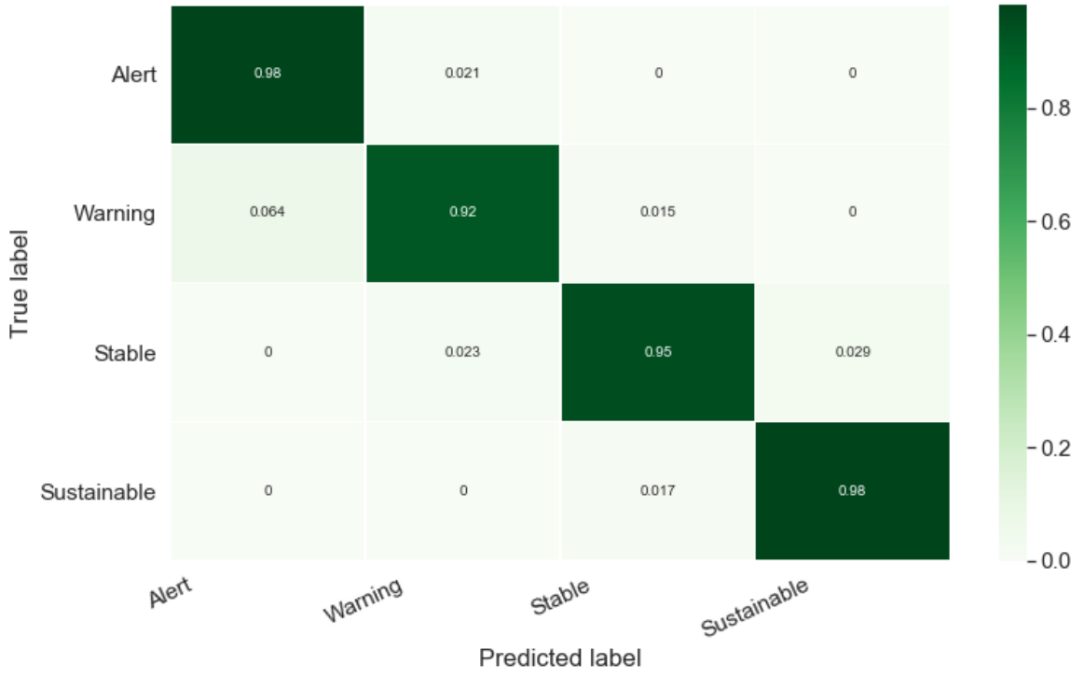

Figure 3.10 Normalized Confusion Matrix for XGBoost

3.4.2 Classification Report

The previous section calculates TP, TN, FP, and FN values for each class label for RF. The classification report for RF can be obtained since the highest accuracy on the test set has been achieved. The classification report serves as a performance evaluation metric for a classifier ML model. By employing classification report method, a better understanding of the trained model's performance can be obtained. The formulas below illustrate the calculation of Accuracy (3.4), Precision (3.5), Recall (3.6), and F1-score (3.7).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3.4)$$

$$Precision = \frac{TP}{TP + FP} \qquad (3.5)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3.6)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (3.7)$$

Table 3.6 indicates the classification report for RF classifier. The accuracy achieved by RF is more than 96% on the test data.

Table 3.6 Classification Report for RF

Classification report:				
	Precision	Recall	F1-score	Support
Alert	0.96	0.94	0.95	215
Warning	0.93	0.94	0.93	216
Stable	0.96	0.98	0.97	207
Sustainable	1.00	0.98	0.98	211
Accuracy			0.96	849
Macro avg	0.96	0.96	0.96	849
Weighted avg	0.96	0.96	0.96	849

3.4.3 Artificial Neural Network Results

This section indicates the neural network accuracy and loss results. Figure 3.11 indicates the accuracy plot for ANN. It appears that the model could benefit from additional training, as the accuracy trend for both datasets has continued to improve over the last few epochs. Furthermore, the model's performance on the test dataset is similar to that of the training dataset, indicating that it has not yet overfit the training data. The X-axis is the number of epochs, and the Y-axis is the accuracy score over time (epoch).

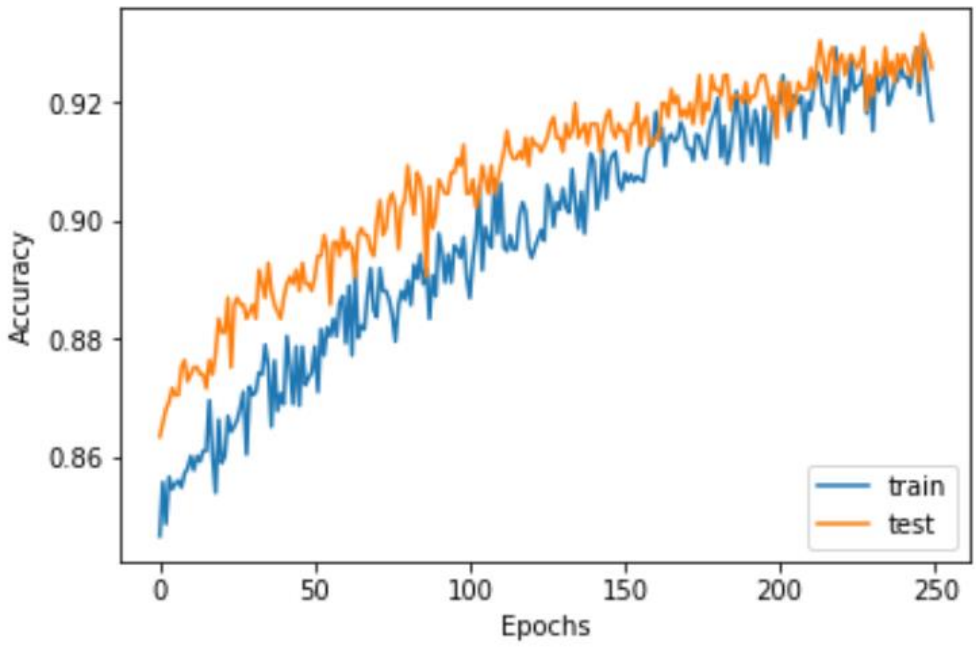

Figure 3.11 Accuracy Plot for ANN

A loss function is a mathematical function that quantifies the difference between the predicted output values and the actual or target output values. The loss function serves as a measure of how well a neural network is able to model the training data. During the training process, the goal is to minimize the value of the loss function, which means minimizing the difference between predicted and target outputs.

Figure 3.12 indicates that the model performs similarly on the training and test datasets. If the loss plots begin to diverge consistently, it may indicate that training should be halted at an earlier epoch. The X-axis is the number of epochs, and the Y-axis is the loss score over time (epoch).

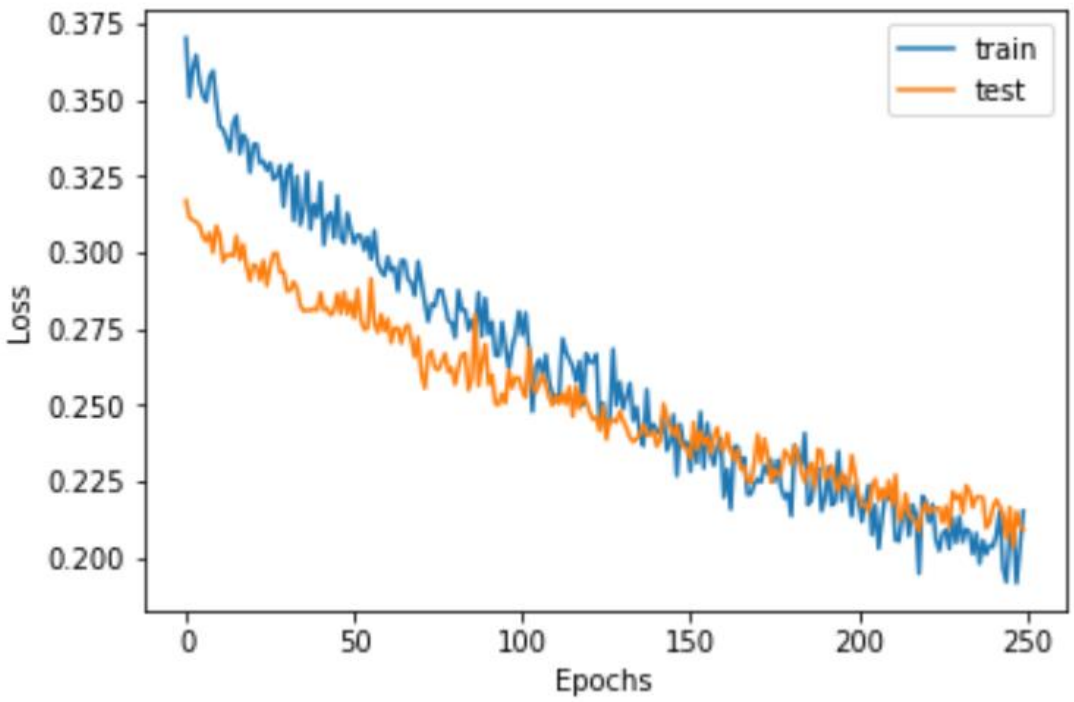

Figure 3.12 Loss Plot for ANN

3.5 Model Selection

In this chapter, various ML models are examined, and RF classifier is chosen as the best-performing model based on its performance. Compared to XGBoost, which tends to create more complex models and may be overfitted with small datasets, RF is generally less prone to overfitting. Neural networks typically require a large amount of training data to generalize well, and small datasets may not provide enough diverse examples for effective learning. In such cases, the RF classifier model may outperform the neural network since it can handle smaller datasets more efficiently. RF is used for the prediction of the risk in the case study of Chapter 4.

3.6 Unsupervised ML

This section aims to categorize countries based on similar variables into clusters while separating dissimilar countries into different clusters. The unsupervised ML method allows the model to uncover patterns and gather information autonomously. Unlike supervised learning algorithms, unsupervised learning algorithms empower users to automate more intricate tasks (Gentleman and Carey, 2008). In contrast to other natural learning methods, unsupervised learning tends to be more unpredictable.

Examples of unsupervised learning algorithms encompass clustering, anomaly detection, and various other techniques. The variables examined in this section are derived from the same dataset used in the previous section. However, considering the feature importance identified by the RF classifier in section 3.3.2, only the significant features are utilized here. The features employed in this section include control of corruption, GDP per capita, life expectancy at birth, democracy score, secondary school enrolment, and population.

Clustering is one of the important approaches when dealing with unsupervised learning. Clustering is a method used in data mining that categorizes unlabeled data based on their similarities. There are different types of clustering algorithms, such as K-Means, Hierarchical, DBSCAN, etc. In this research, K-Means clustering is used to cluster data since the size of the dataset is small (Bradley and Fayyad, 1998)

Before applying K-Means clustering, principal component analysis (PCA) is recommended. Yeung and Ruzzo (2001) proposed a clustering algorithm that first applies PCA to the data to obtain the principal components and then clusters the samples based on the principal components.

Combining PCA and K-Means clustering involves using PCA to reduce the dimensionality of the data and then applying K-Means clustering to the transformed data to identify clusters. The steps involved in combining PCA and K-Means clustering are:

1. Before applying PCA, the data should be standardized to ensure that all features are on the same scale.
2. Use PCA to transform the data into a lower-dimensional space. The number of principal components to retain can be determined using various methods such as the scree plot, percentage of variance explained, or cross-validation.
3. Examine the PCA to identify which features contribute most to each principal component that can help in interpreting the clusters identified by K-Means clustering.
4. Apply K-Means clustering to the transformed data to identify clusters. The optimal number of clusters can be determined using methods such as the elbow method, silhouette score, or gap statistic.
5. Examine the clusters to understand their characteristics and how they relate to the original features.

3.6.1 Principal Component Analysis (PCA)

PCA is a widely used technique in the field of data science for dimensionality reduction and can be useful for visualizing high-dimensional data, reducing noise in the data, and improving the efficiency of ML algorithms by reducing the number of features. PCA can also be used as a preprocessing step before applying clustering algorithms. By reducing the

dimensionality of the data, PCA can help to simplify the clustering problem and make it more computationally tractable (Abdi and Williams, 2010).

Normalizing the data before applying PCA is recommended to prevent the scenario where one component contains all the variance due to the unequal scaling of variables (Ringnér, 2008). Figure 3.13 shows the quantity of variance that is captured (depicted on the y-axis) based on the number of components included (shown on the x-axis). Typically, it is recommended to maintain approximately 80% of the variance (Linacre, 2006). Therefore, it has been determined to retain 2 principal components.

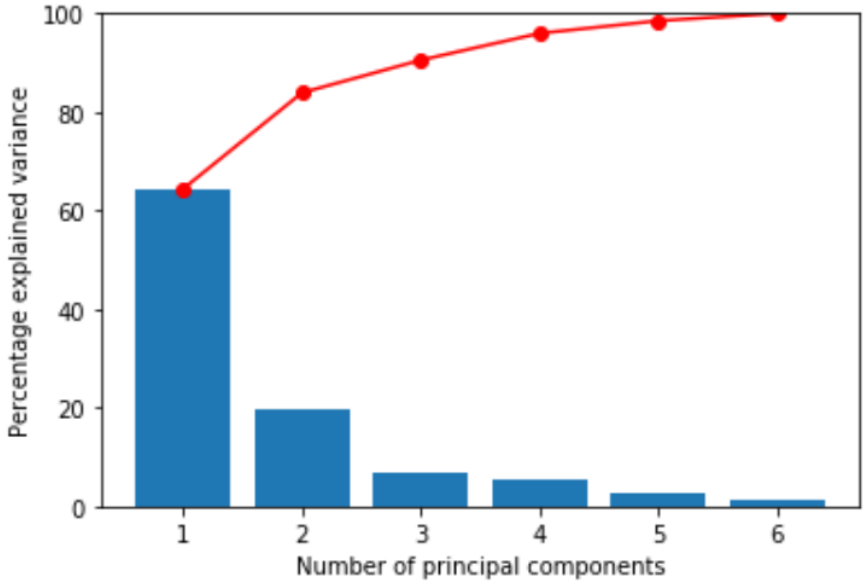

Figure 3.13 Explained Variance Based on Number of Principal Components

Table 3.7 indicates the loading vectors of the principal components. Each loading vector is a row vector showing each feature's contribution to the corresponding principal component.

The sign of each coefficient indicates the direction of the relationship between the feature and PC1. A positive coefficient indicates that the feature has a positive relationship with

PC1, meaning that high values of the feature are associated with high values of PC1. On the other hand, a negative coefficient indicates a negative relationship between the feature and PC1.

Table 3.7 Loading Vector of PC1 and PC2

Variables	PC1	PC2
GDP per capita	0.458825	0.053466
Population	-0.068727	0.817743
School enrolment	0.468649	0.259422
Life expectancy at birth	0.472234	0.282430
Democracy score	0.358825	-0.383288
Control of corruption	0.461914	-0.185637

3.6.2 Combining PCA and K-Means Clustering

The main idea of combining PCA and K-Means clustering approach is to use PCA to transform the data into a lower-dimensional space that captures most of the variation in the data and then apply K-Means clustering to transformed data to identify clusters. The most important hyperparameter in K-Means clustering is the number of clusters in the algorithm. The optimal value of K depends on the data's characteristics and the analysis's goals.

The Elbow method is a popular heuristic used to determine the optimal number of clusters in K-Means clustering. In the Elbow method, the number of clusters (K) is varied from 1

to 10, and for each value of K, the Within-Cluster Sum of Square (WCSS) is calculated. WCSS represents the sum of the squared distance between each point and the centroid in a cluster. The plot exhibits an elbow shape when the WCSS is plotted against the K value (Clayman et al., 2020). As the number of clusters increases, the WCSS value decreases, and it is the highest when K = 1. The Elbow method and the silhouette score determine the optimal number of clusters. Figure 3.14 shows the results of the Elbow method.

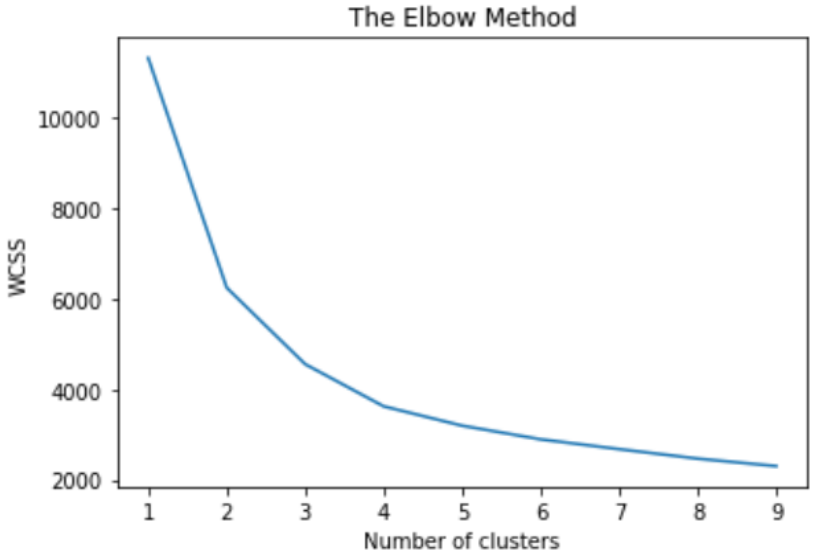

Figure 3.14 WCSS Values of Different Number of Clusters Using the Elbow Method

Based on the above Figure 3.14, the Elbow method suggests that the optimal number of clusters lies between 3 and 4.

The silhouette score is a metric used to evaluate the quality of clustering results in K-Means clustering. It provides a measure of how well each data point fits into its assigned cluster and how well separated each cluster is from the other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better clustering results.

Figure 3.15 shows the silhouette score of 0.388 when the number of clusters is three; however, when K=4, the silhouette score is 0.424 which is higher (Figure 3.16). Considering the higher silhouette score, K=4 is the optimal number of clusters.

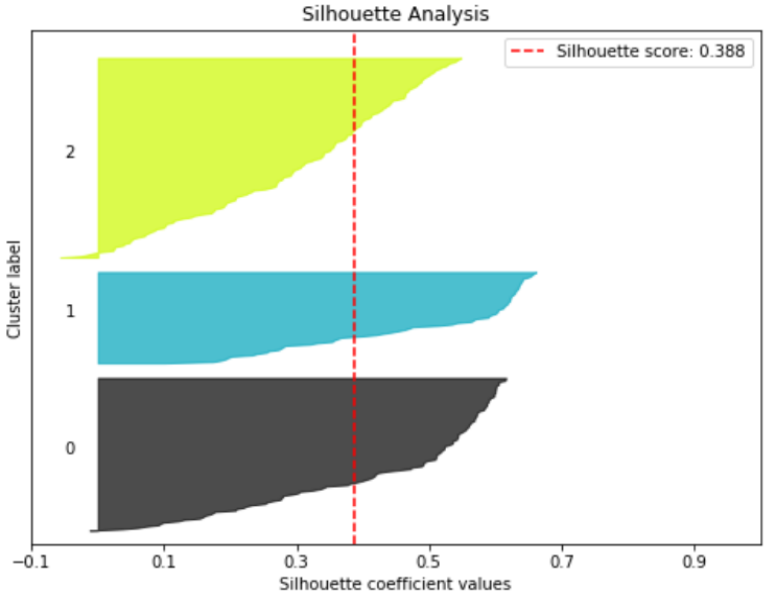

Figure 3.15 Silhouette Score for K=3

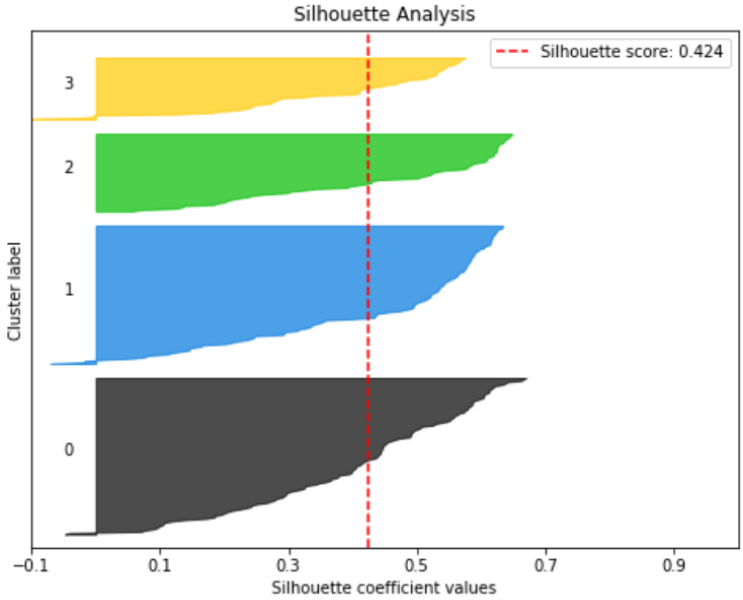

Figure 3.16 Silhouette Score for K=4

One of the benefits of PCA besides data preparation is visualization. PCA can be used for data visualization. Figure 3.17 demonstrates how clustering can be performed based on PC 1 and PC 2. Each cluster is identified by the countries that belong to it, and its label can be modified through supervised learning, such as, Cluster 0: Warning, 1: Stable, 2: Sustainable, and 3: Alert. PC 1 is plotted along the X-axis, while PC 2 is plotted along the Y-axis.

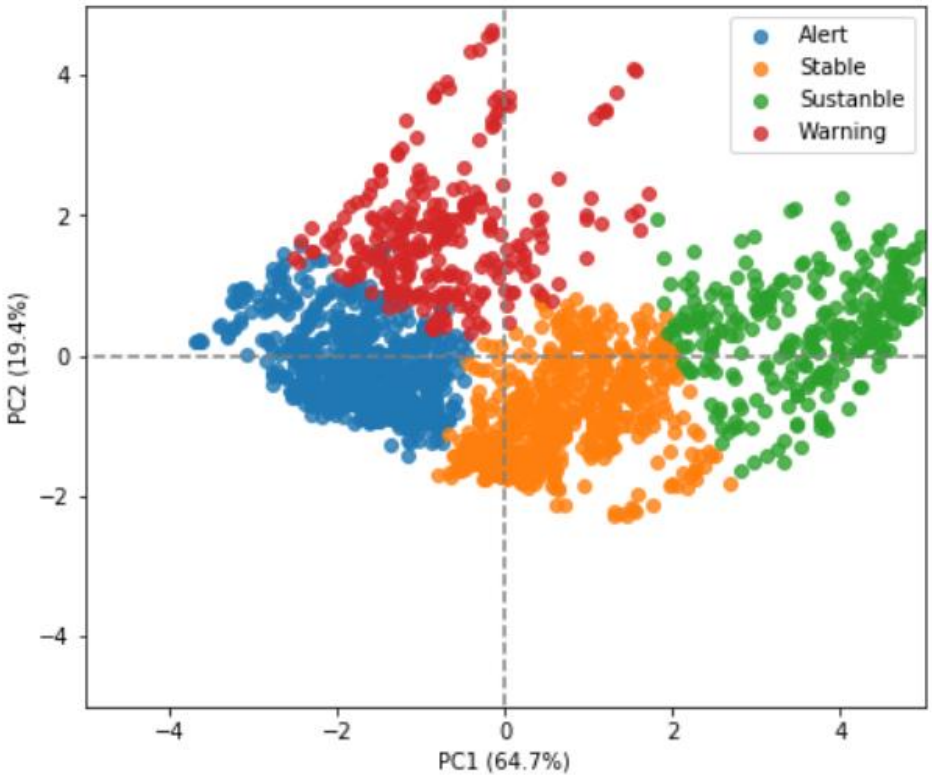

Figure 3.17 Data Visualization Using Two Principal Components

Overall, combining K-Means clustering and PCA effectively categorizes countries into meaningful clusters, and is particularly useful if there is no label available for the data. A clustering approach can be valuable for gaining insights into different groups of countries, aiding decision-making processes, and providing a foundation for further research in the field.

CHAPTER 4: FACILITY LOCATION MODEL

Facility location models are used to determine the optimal location of facilities (such as factories, warehouses, or distribution centers) and the distribution of resources (such as supplies or personnel) to those facilities. In this section, we develop a bi-objective MILP model. The first objective is to minimize transportation costs by locating facilities closer to customers and reduce inventory costs by optimizing the allocation of resources to different facilities. The second objective function is to minimize the associated with the fragile state of the country that can be predicted by the chosen ML model (RF) in chapter 3. Table 4.1 shows the parameters and indices used in the model.

Table 4.1 Model Parameters, Indices and Variables

Parameters	Definitions
n	Set of locations
m	Set of customers
$d_{i,t}$	Demand of customer i at time t
$u_{j,t}$	Production cost per unit of location j at time t
$C_{j,t}$	Capacity of location j at time t
$f_{j,i,t}$	Freight cost per unit to ship from location j to customer i at time t
$R_{j,t}$	Risk (fragile state) associated with location j at time t
B_i	Minimum units shipped from a location to customer i (batch size for shipment)
i	Customers, $i \in m$
j	Locations, $j \in n$
t	Time period, $t \in T$
Decision Variables	
$y_{j,i,t}$	Number of units shipped from location j to customer i in period t
$x_{j,i,t}$	1 If location j is selected to manufacture units of customer i at time t

There are two objective functions: The first objective of the model is to minimize the total production and freight cost (production/distribution cost) and is defined by Equation (4.1):

$$Min\ OF1 = \sum_{j\in n}\sum_{i\in m}\sum_{t\in T} y_{j,i,t} \times f_{j,i,t} + \sum_{j\in n}\sum_{i\in m}\sum_{t\in T} y_{j,i,t} \times u_{j,t} \qquad (4.1)$$

The second objective function is to minimize the risk associated with the fragile state of the country where the manufacturing plant is located (Equation (4.2)). The riskier the manufacturing location, the fewer units should be allocated to that location.

$$Min\ OF2 = \sum_{j\in n}\sum_{i\in m}\sum_{t\in T} y_{j,i,t} \times R_{j,t} \qquad (4.2)$$

The Equations (4.3) to (4.8) represents the sets of the constraints of the model:

1. Total units assigned to location *j* cannot exceed its capacity at time *t*:

$$\sum_{i\in m} y_{j,i,t} \leq C_{j,t} \qquad \forall\, j \in n, t \in T \qquad (4.3)$$

2. The forecasted demand of each customer should be met at each time period t:

$$\sum_{j\in n} y_{j,i,t} \geq d_{i,t} \qquad \forall\, i \in m, t \in T \qquad (4.4)$$

3. The units to be shipped to customer *i* from each assigned location cannot be less than the minimum batch size:

$$y_{j,i,t} \geq B_i \times x_{j,i,t} \quad \forall\, j \in n, i \in m, t \in T \qquad (4.5)$$

4. If location *j* is not assigned to customer *i*, $x_{j,i,t}$ will be zero:

$$x_{j,i,t} \leq y_{j,i,t} \qquad \forall j \in n, i \in m, t \in T \qquad (4.6)$$

5. If there is a shipment from location j to customer i, $x_{j,i,t}$ will be one:

$$x_{j,i,t} \geq \frac{y_{j,i,t}}{C_{j,t}} \qquad \forall j \in n, i \in m, t \in T \tag{4.7}$$

6. $x_{j,i,t}$ is a binary variable and $y_{j,i,t}$ is a non-negative integer:

$$x_{j,i,t} \in \{0,1\}, y_{j,i,t} \in \mathbb{Z}^+ \tag{4.8}$$

4.1 Solving MILP Model

FLP models are important tools for improving efficiency and reducing costs in a variety of industries. These models can be solved using a variety of approaches. We used a lexicographic preemptive optimization approach to solve the developed bi-objective MILP model.

Lexicographic preemptive optimization approach is a technique used in decision-making problems where multiple objectives must be optimized simultaneously. So, the objectives are prioritized in a hierarchical manner, and the optimization is performed one objective at a time, starting with the most important one. (Rentmeesters et al. 1996). In lexicographic preemptive optimization, the objectives are ranked in order of importance. In our case, let's say the company minimizing cost is over minimizing the risk of the country. So, the optimization process would start with minimizing cost while keeping the risk of the country within certain limits. Once the cost is minimized, the optimization process moves to the next objective, minimizing the country's risk while keeping the cost within certain limits. The lexicographic preemptive optimization approach ensures that the most important objective is achieved first, and the other objectives are optimized while maintaining the

priority of the higher-ranked objectives. The lexicographic approach can be useful when the objectives are difficult to optimize simultaneously or when there are trade-offs between objectives.

As illustrated in Figure 4.1, the flowchart below shows the overall approach in this research.

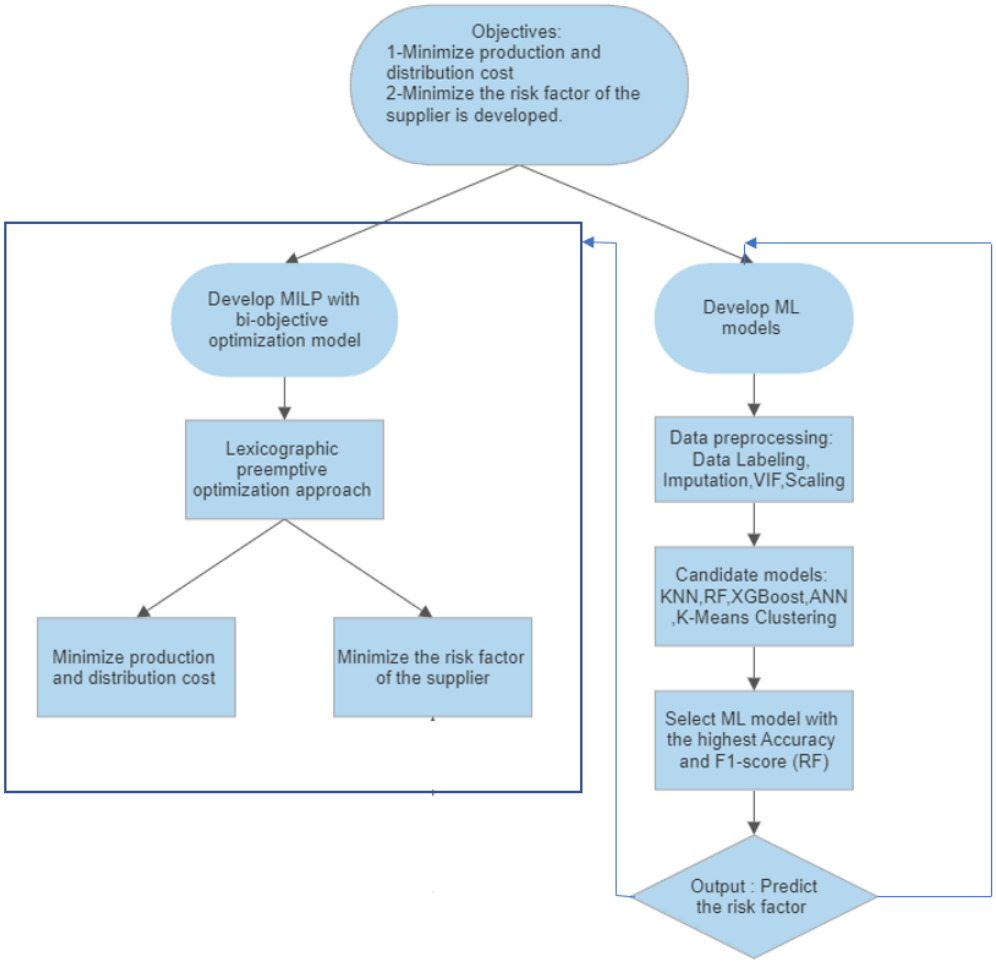

Figure 4.1 Block Diagram Summarizing the Combined ML-Mathematical Approach

4.2 Test Case

Company ABC is an international manufacturer that has manufacturing facilities in five countries: Mexico (MX), Argentina (AR), Romania (RO), USA(US), and Vietnam (VN). It also has three distribution centers in the USA, Spain (SP), and China (CH), acting as the regional hub covering demands in America, Europe, and Asia.

The study period is five years covering the years of 2021 to 2025. Table 4.2 shows the demand of each region.

Table 4.2 Demand of Each Region

Year	Distrub1-US	Distrub2-SP	Distrub3-CH
2021	9000	5000	6000
2022	10000	7000	5500
2023	11000	7500	6000
2024	10000	7500	5500
2025	9000	8000	5000

Table 4.3 shows the available capacity at each manufacturing facility in different years (in units) and Table 4.4 indicates the product cost per unit (in dollars).

Table 4.3 Capacity of Each Location

Year	MX	AR	RO	US	VN
2021	5000	4500	4000	12000	3000
2022	5000	4500	4000	12000	3000
2023	5000	4500	4000	12000	3000
2024	5000	4500	4000	12000	3000
2025	5000	4500	4000	12000	3000

Table 4.4 Cost Per Unit in Each Location

Year	MX	AR	RO	US	VN
2021	8.0	18.0	6.0	28.0	8.0
2022	8.3	22.5	6.6	28.8	8.1
2023	8.7	28.1	7.3	29.7	8.2
2024	9.0	35.2	8.0	30.6	8.4
2025	9.4	43.9	8.8	31.5	8.5

Table 4.5 indicates the freight cost per unit for each distribution center.

Table 4.5 Freight Cost Per Unit for Each Distribution Center.

Distribution Center	MX	AR	RO	US	VN
Distrub1-US	3.5	8	9	0.5	14
Distrub2-SP	11	14	5	8	10
Distrub3-CH	13	19	6	12	3

In order to calculate the risk, the best performed ML model in Chapter 3, RF classifier is used for prediction and the results are obtained for each country and each year in Table 4.6.

Table 4.6 Predicted State of Each Facility's Country

Country	2021	2022	2023	2024	2025
VN	Warning	Warning	Warning	Warning	Warning
US	Stable	Sustainable	Sustainable	Sustainable	Sustainable
AR	Stable	Stable	Stable	Warning	Warning
MX	Warning	Warning	Warning	Warning	Warning
RO	Stable	Warning	Stable	Stable	Stable

Based on Sekhar's article, the labels indicated above can be converted to numerical values; in this test case, we have used the average value of each label, and we have used the value as the indicator of the risk level in that country. The higher value indicates a higher risk in that country.

Table 4.7 Numerical Value of the Risk for Each Category

Label	Alert	Warning	Stable	Sustainable
Range	90–120	60–89.9	30–59.9	29.9 or less
Average	105	75	45	15

Since we have considered two objective functions in our model, depending on the priority we select for each objective, the results of the model will change. In this test case, we have used the CPLEX solver to optimize the model using a lexicographic preemptive approach. It is also assumed that the minimum batch size to ship from a manufacturing facility to a distribution center is 1000 units.

Scenario One:

In scenario one, the cost is assumed to have more priority over the risk. Therefore, the model will first try to minimize the total cost, and under equal conditions, it then chooses the locations with lower risks.

The results obtained after running the model under scenario one are presented in Table 4.8

Table 4.8 Result of the First Scenario

Year	2021	2022	2023	2024	2025
Location	Country, units	Country, units	Country, units	Country, units	Country, units
Distrub1-US	MX, 5000 US, 4000	MX, 4000 US, 6000	MX, 1000 US, 10000	US, 10000	US,9000
Distrub2-SP	AR,4000 RO,1000	AR,4500 MX,1000 RO,1500	MX,4000 RO,1000 US,1500	MX,5000 RO,1500 US,1000	MX,5000 RO,2000 US,1000
Distrub3-CH	RO,3000 VN,3000	RO,2500 VN,3000	RO,3000 VN,3000	RO,2500 VN,3000	RO,2000 VN,3000

Scenario Two:

In scenario two, it is assumed that selecting a high-risk country should be avoided. Therefore, the model first minimizes the objective function based on the risk level of the distribution centers and under equal conditions; it then chooses the location with the lowest production and freight costs. The solution after running the model under scenario two indicates in Table 4.9.

Table 4.9 Result of the Second Scenario

Year	2021	2022	2023	2024	2025
Location	Country, units	Country, units	Country, units	Country, units	Country, units
Distrub1-US	US, 9000	US, 10000	US, 11000	US, 6000 MX,4000	US,6000 MX,3000
Distrub2-SP	AR,3500 US,1500	AR,4500 US,1000 RO,1500	AR,4500 RO,1000 US,1000	RO,1500 US,6000	RO,2000 US,6000
Distrub3-CH	RO,4000 US,1000 AR,1000	RO,1500 US,1000 VN,3000	RO,3000 VN,3000	RO,2500 VN,3000	RO,2000 VN,3000

In scenario one, it is clear that MX is one of the most favorable manufacturing facilities in terms of low production/distribution costs. In all five years, the model suggests that MX runs at its maximum capacity (5000) to fulfill the demands of different regions.

In the second scenario, where the risk factor is more important than the cost, the model recommends using US plants to operate with full capacity in all years except 2021, where countries US, RO, and AR have the same level of risk. Since RO and AR have less cost, they are chosen to operate at full capacity. In addition, in 2021, 2022 and 2023 where AR is predicted to be a low-risk country, the model recommends that AR operate at its full capacity in those years. However, since the model predicts that AR will be a high-risk country in 2024 and 2025, the model suggests not using this plant. This can be considered as one of the factors in deciding whether a certain manufacturing facility should be shut down.

4.3 Discussion

The primary objective of the MILP model is to minimize production and distribution costs while also considering the risk factor associated with the supplier's country of origin. Comparative analysis of ML models in this research revealed that the RF classifier emerges as the best-performing model based on its performance characteristics. While XGBoost and ANN have their strengths, they may not be well-suited for scenarios with limited datasets. RF classifier offers a balance between complexity and generalization, making it less prone to overfitting and capable of handling small datasets efficiently (Breiman, 1996). RF is known to be less sensitive to hyperparameter tuning compared to XGBoost. While XGBoost offers more flexibility in terms of hyperparameter customization, it requires

careful tuning to optimize its performance. On the other hand, RF's default hyperparameters often provide reasonable results, and it is less sensitive to small changes in hyperparameter values. When faced with limited or small datasets, ANN may struggle to learn diverse representations due to insufficient examples. Consequently, the RF classifier often outperforms neural networks in such cases by efficiently leveraging the available data and producing reliable predictions. We also categorized countries based on similar variables into clusters while separating dissimilar countries into different clusters. A clustering approach can be valuable for gaining insights into different groups of countries, aiding decision-making processes, and providing a foundation for further research in the field.

In the specific test case conducted, the problem is solved by utilizing the RF predictions for the years 2021 to 2025 (Table 4.6). The output predictions are integrated into the optimization model, enabling a comprehensive analysis of production and distribution costs while factoring in the risk associated with the supplier's country of origin. This holistic approach allows decision-makers to make informed choices that optimize costs and account for potential risks in the supply chain. By leveraging the power of the RF model and incorporating its predictions into the optimization framework, this novel approach offers a valuable tool for decision-makers in forecasting a country's fragility and making strategic decisions. It empowers organizations to proactively manage risks, optimize resource allocation, and enhance operational efficiency and resilience in an increasingly complex and uncertain global landscape.

CHAPTER 5: CONCLUSION

5.1 Research Contribution

The primary achievement of this study is integrating a ML technique with an optimization model. The ML model's results are utilized as a risk factor input in the optimization model. The best-performing ML model (RF Classifier) achieves an accuracy rate of 96% on the test set. Moreover, to our knowledge, no existing literature has considered all the political, social, and economic characteristics utilized in this research to forecast a country's fragility. The MILP model with bi-objective optimization aims to minimize production and distribution costs while also considering the risk factor of the supplier's country of origin. ML approach is utilized to determine the supplier's country of origin risk factor and is considered an input for a risk factor in the MILP model.

The contribution of this research can be addressed to:

1. Better allocation of resources: By combining a ML approach with an optimization model and considering a comprehensive set of political, social, and economic features, this research may provide a more accurate and reliable forecast of a country's fragility, which could help decision-makers take appropriate action to prevent or mitigate the risk of instability and conflict. In other words, the results of this research could inform policymakers and aid organizations about which countries are at the highest risk of fragility, enabling them to allocate resources and prioritize interventions more effectively.

2. Contribution to academic literature: This research's innovative approach in integrating ML techniques and optimization approach and use of a comprehensive set of features to predict a country's fragility could contribute to the academic literature on this topic and potentially inform future research and policymaking.

5.2 Limitation Study

One of the limitations of this study can be addressed as the limitation of data. ML models can perform better with larger amounts of data. This is because ML models use statistical algorithms to identify patterns and relationships within the data. With more data, the model has a larger pool of examples to learn from, which can improve its ability to predict outcomes or classify data points accurately; on the other side collecting and processing large amounts of data can be expensive and time-consuming. The other limitation that can be mentioned is the complexity of political, social, and economic factors. While using a comprehensive set of political, social, and economic features to predict a country's fragility is a strength of this research, it also poses a challenge. These factors are complex and multifaceted, and it may be challenging to determine which variables are most important and how they should be weighted in the models.

5.3 Conclusions

In this research, a novel approach combining mathematical optimization and ML is proposed to enhance supplier selection and mitigate supply chain risks.

Here are some important notes from this research:

- The proposed MILP model provides a comprehensive approach to decision-making, considering both cost minimization and risk assessment in the supplier selection process.
- The utilization of various ML models and techniques demonstrates the researchers' efforts to identify the most accurate and reliable method for predicting the risk factor of a supplier's country of origin.
- The dataset used in this study covers the period from 2006 to 2021 and represents 139 countries in an unbalanced panel.
- The high accuracy rate of 96% achieved by the RF Classifier model suggests its effectiveness in providing dependable predictions for future events and enhances the robustness of the overall decision-making process.

The test case solved by the CPLEX solver demonstrates the practical applicability of the proposed approach by effectively incorporating the risk factors predicted by the RF Classifier for the period spanning from 2021 to 2025. This showcases the ability of the approach to handle real-world scenarios and provides evidence of its effectiveness in addressing risk factors in the specified time frame.

By considering the risk factor in the decision-making process, companies can make more informed choices, reducing potential disruptions and vulnerabilities in the supply chain while achieving cost savings. The research contributes to the field of supply chain management by offering a comprehensive framework that integrates mathematical

optimization, ML, and risk assessment to support strategic supplier selection and mitigate supply chain risks.

5.4 Future Work

Further improvements can be made to enhance the performance and efficiency of the MILP model and could involve exploring alternative optimization algorithms, incorporating additional decision variables or constraints. In the context of capacity planning and optimization, future research can focus on the development and implementation of a reconfigurable capacity Facility Location Problem (FLP) model.

Future research can explore the application of System Dynamics methodology to model the dynamics involved in current conflictual global political systems.

The ML models used for predicting the risk factor of the supplier's country of origin can be further developed and expanded. Including a wider range of features, such as environmental factors or industry-specific indicators, may improve the accuracy and robustness of the predictions.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.

Ades, A., & Chua, H. B. (1997). They neighbor's curse: regional instability and economic growth. *Journal of Economic Growth*, 2(3), 279-304.

Aisen, A., & Veiga, F. J. (2013). How does political instability affect economic growth? *European Journal of Political Economy*, 29, 151-167.

Akgün, İ., Gümüşbuğa, F., & Tansel, B. (2015). Risk based facility location by using fault tree analysis in disaster management. *Omega*, 52, 168-179.

Alesina, A., Özler, S., Roubini, N., & Swagel, P. (1996). Political instability and economic growth. *Journal of Economic growth*, 1(2), 189-211.

Ali, A., & Bibi, C. (2017). Determinants of social progress and its scenarios under the role of macroeconomic instability. *Pakistan Economic and Social Review*, 55(2), 533-568.

ANDERSON, J. A. 1995. An introduction to neural networks, MIT press. Applications. *Computers & Industrial Engineering*, 62(1), 408–420.

Arabani, A. B., & Farahani, R. Z. (2012). Facility location dynamics: An overview of classifications and

Baillie, E., Howe, P. D., Perfors, A., Miller, T., Kashima, Y., & Beger, A. (2021). Explainable models for forecasting the emergence of political instability. *Plos one*, 16(7), e0254350.

Bao, L., Juan, C., Li, J., & Zhang, Y. (2016). Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. Neurocomputing, 172, 198-206.

Basuchoudhary, A., Bang, J., & Shughart, W. F. (2015). A Machine Learning Approach to Predicting State Failure. *Working Paper.*

BISHOP, C. M. 1995. Neural networks for pattern recognition, Oxford university press.

Bittencourt, M., Gupta, R., Makena, P., & Stander, L. (2022). Socio-political instability and growth dynamics. *Economic Systems*, 46(4), 101005.

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining initial points for k-means clustering. In *ICML* (Vol. 98, pp. 91-99).

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining initial points for k-means clustering. In *ICML* (Vol. 98, pp. 91-99).

Breiman, L. (1996). Bagging predictors Machine Learning 24 (2), 123-140 (1996) 10.1023. A: 1018054314350.

Chen, H., Li, X. Y., Lu, X. R., Sheng, N., Zhou, W., Geng, H. P., & Yu, S. (2021). A multi-objective optimization approach for the selection of overseas oil projects. *Computers & Industrial Engineering*, 151, 106977.

Chen, Y., & Lai, Z. (2022). A Multi-Objective Optimization Approach for Emergency Medical Service Facilities Location-Allocation in Rural Areas. *Risk Management and Healthcare Policy,* 473-490.

Clayman, C. L., Srinivasan, S. M., & Sangwan, R. S. (2020). K-means clustering and principal components analysis of microarray data of L1000 landmark genes. Procedia Computer Science, 168, 97-104.

Collier, P., & Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford economic papers*, 56(4), 563-595.

CSRC (2006). War, state collapse and reconstruction: Phase 2 of the Crisis States Program. *Working Paper No. 1, Series No. 2, London School of Economics and Political Science.*

Cui, T., Ouyang, Y., & Shen, Z. J. M. (2010). Reliable facility location design under the risk of disruptions. *Operations research*, 58(4-part-1), 998-1011.

De La Torre, J., & Neckar, D. H. (1988). Forecasting political risks for international operations. *International journal of forecasting,* 4(2), 221-241.

Dumitrescu, E. I., & Hurlin, C. (2012). Testing for Granger non-causality in heterogeneous panels. *Economic modelling*, 29(4), 1450-1460.

Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1996). Political risk, economic risk, and financial risk. *Financial Analysts Journal,* 52(6), 29-46.

Faraway, J. J. (2002). Practical regression and ANOVA using R (Vol. 168). Bath: University of Bath.

Gebennini, E., Gamberini, R., & Manzini, R. (2009). An integrated production–distribution model for the dynamic location and allocation problem with safety stock optimization. *International Journal of Production Economics,* 122(1), 286-304.

Gentleman, R., & Carey, V. J. (2008). Unsupervised machine learning. In Bioconductor case studies (pp. 137-157). *Springer, New York, NY*.

Goldstone, J. A., Bates, R. H., Epstein, D. L., Gurr, T. R., Lustik, M. B., Marshall, M. G., ... & Woodward, M. (2010). A global model for forecasting political instability. *American Journal of Political Science*, 54(1), 190-208.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 986-996). *Springer, Berlin*, Heidelberg.

Habibi, F., Asadi, E., Sadjadi, S. J., & Barzinpour, F. (2017). A multi-objective robust optimization model for site-selection and capacity allocation of municipal solid waste facilities: A case study in Tehran. *Journal of cleaner production*, 166, 816-834.

He, H., Bai, Y., Garcia, E. A., & Li, S. A. (2008). adaptive synthetic sampling approach for imbalanced learning. *IEEE international joint conference on neural networks*. In 2008 (IEEE World Congress On Computational Intelligence).

Hedaoo, S. M. (2016). Mathematical modelling and a meta-heuristic for cross border supply chain network of re-configurable facilities (Doctoral dissertation, University of Windsor (Canada)).

Hinojosa, Y., Puerto, J., & Fernández, F. R. (2000). A multiperiod two-echelon multicommodity capacitated plant location problem. *European Journal of Operational Research*, 123(2), 271-291.

Howell, Llewellyn D., (2007) Country and Political Risk Assessment for Managers, in Howell, Llewellyn D. (ed.) The Handbook of Country and Political Risk Analysis (4th Edition) The PRS Group, Inc.

Jong-A-Pin, R. (2009). On the measurement of political instability and its impact on economic growth. *European Journal of Political Economy*, 25(1), 15-29.

Kalantari, Amir Hossein, "Facility Location Selection for Global Manufacturing" (2013). *Theses and Dissertations*. 233.

Karnane, P., & Quinn, M. A. (2019). Political instability, ethnic fractionalization and economic growth. *International Economics and Economic Policy*, 16(2), 435-461.

Khafaga, R. H., & Albagoury, S. H. (2022). Political Instability and Economic Growth in Ethiopia: An Empirical Analysis. *Journal of Social and Political Sciences*, 5(1).

Khudari, M., Sapuan, N. M., & Fadhil, M. A. (2023). The impact of political stability and macroeconomic variables on foreign direct investment in Turkey. In International Conference on Business and Technology (pp. 485-497). *Springer, Cham*.

KRÖSE, B., KROSE, B., VAN DER SMAGT, P., and SMAGT, P. 1993. An introduction to neural networks.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4, pp. 563-568). New York: McGraw-Hill/Irwin.

Li, Y., & Yao, H. (2018). Classification of fragile states based on machine learning. In MATEC Web of Conferences (Vol. 173, p. 02044). EDP Sciences.

Linacre, J. M. (2006). Data variance explained by Rasch measures. Rasch Measurement Transactions, 20(1), 1045.

Maliki, F., Souier, M., Dahane, M., & Ben Abdelaziz, F. (2022). A multi-objective optimization model for a multi-period mobile facility location problem with environmental and disruption considerations. *Annals of Operations Research*, 1-26.

Matta, S., Bleaney, M., & Appleton, S. (2022). The economic impact of political instability and mass civil protest. Economics & Politics, 34(1), 253-270.

McCarthy, John J., "Optimality Theory: An overview" (2003). Oxford International Encyclopedia of Linguistics. 56.

Messner, J. J. (2022). Fragile States Index... Fund for Peace.

Morrissey, T. W., Cha, Y., Wolf, S., & Khan, M. (2020). Household economic instability: Constructs, measurement, and implications. *Children and youth services review*, 118, 105502.

Murad, M. S. A., & Alshyab, N. (2019). Political instability and its impact on economic growth: the case of Jordan. *International Journal of Development Issues*.

Murad, M. S. A., & Alshyab, N. (2019). Political instability and its impact on economic growth: the case of Jordan. *International Journal of Development Issues*.

Musonera, E., Yaprak, A., & Monplaisir, L. (2009). Modeling FDI Attraction-An Optimization Model. *Journal of International Business Research and Practice*, 3, 15.

Mustapha, N. (2014). The impact of corruption on GDP per capita. Journal of Eastern European and Central Asian research, 1(4), 1-5.

Pan, B. (2018, February). Application of XGBoost algorithm in hourly PM2. 5 concentration prediction. In IOP conference series: earth and environmental science (Vol. 113, p. 012127). *IOP publishing*.

Perles-Ribes, J. F., Ramon-Rodriguez, A. B., Such-Devesa, M. J., & Moreno-Izquierdo, L. (2019). Effects of political instability in consolidated destinations: The case of Catalonia (Spain). *Tourism Managemen*t, 70, 134-139.

Prakash, S., Soni, G., & Rathore, A. P. S. (2015). A grey based approach for assessment of risk associated with facility location in global supply chain. Grey Systems: Theory and Application.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine learning, 42(3), 203-231.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (Eds.). (1998). Applied regression analysis: a research tool. New York, NY: *Springer New York*.

Rentmeesters, M. J., Tsai, W. K., & Lin, K. J. (1996). A theory of lexicographic multi-criteria optimization. In Proceedings of ICECCS'96: *2nd IEEE International Conference on Engineering of Complex Computer Systems* (held jointly with 6th CSESAW and 4th IEEE RTAW) (pp. 76-79). IEEE.

Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, *26*(3), 303-304.

Salman, D., & Bassim, M. A. (2023). Political stability, austerity measures, external imbalance, and debt impact on the Egyptian economy. *In Research Anthology on Macroeconomics and the Achievement of Global Stability* (pp. 1635-1656). IGI Global.

Sekhar, C. S. C. (2010). Fragile states: The role of social, political, and economic factors. *Journal of Developing Societies*, 26(3), 263-293.

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A review on imbalanced data handling using undersampling and oversampling techniques. *Int. J. Recent Trends Eng*. Res, 3(4), 444-449.

Siermann, C. L. (1998). Politics, institutions and the economic performance of nations. Books.

Sofuoğlu, E., & Ay, A. (2020). The relationship between climate change and political instability: the case of MENA countries (1985: 01–2016: 12). Environmental Science and Pollution Research, 27(12), 14033-14043.

Strike, K., El Emam, K., and Madhavji, N. (2001). Software cost estimation with incomplete data. *IEEE Transactions on Software Engineering*, 27(10), 890-908.

Thanh, P. N., Bostel, N., & Péton, O. (2008). A dynamic model for facility location in the design of complex supply chains. *International Journal of Production Economics*, 113(2), 678-693.

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics, 17(9), 763-774.

APPENDICES

Appendix A: Scale /unit of each feature and source it

Economic Features	**Scale /Unit**	**Source**
Growth GDP rate	Number	World Bank
Employment to population ratio	Precent	World Bank
GDP per capita income	US Dollars	World Bank
Import	US Dollars	World Bank
Export	US Dollars	World Bank
Foreign direct investment	US Dollars	World Bank
Social Features		
Secondary school enrolment	Number	World Bank
Population	Number	World Bank
Individuals using the Internet	Precent	World Bank
Life expectancy at birth	Precent	World Bank
Political Features		
The level of democracy	0 to 1	Our world in data
Voice and Accountability	Precent	World Bank
Regime type	0,1,2,3	Our world in data
Control of Corruption	-2.5 to 2.5	World Bank

Appendix B: Statistical result for each feature

Features	count	mean	std	min	25%	50%	75%	max
FDI	2.2E+03	-1.3E+09	2.7E+10	-3.5E+11	-2.7E+09	-7.5E+08	-8.7E+07	2.2E+11
Import	2.2E+03	1.4E+11	3.4E+11	8.0E+07	7.4E+09	2.2E+10	1.0E+11	3.4E+12
GDP growth	2.2E+03	3.3E+00	5.3E+00	-5.0E+01	1.5E+00	3.6E+00	6.0E+00	8.7E+01
Employment	2.2E+03	5.6E+01	1.3E+01	2.4E+01	4.7E+01	5.7E+01	6.3E+01	8.8E+01
GDP per capita	2.2E+03	1.3E+04	1.8E+04	1.7E+02	1.5E+03	5.0E+03	1.5E+04	1.0E+05
Population	2.2E+03	5.0E+07	1.6E+08	1.4E+05	5.5E+06	1.2E+07	3.7E+07	1.4E+09
Individuals using the Internet	2.2E+03	4.1E+01	3.0E+01	1.8E-01	1.2E+01	3.6E+01	7.0E+01	1.0E+02
School enrolment	2.2E+03	8.0E+01	3.1E+01	1.1E+01	5.2E+01	8.9E+01	1.0E+02	1.6E+02
Life expectancy at birth	2.2E+03	7.1E+01	8.6E+00	4.4E+01	6.4E+01	7.3E+01	7.7E+01	8.5E+01
Democracy Score	2.2E+03	5.4E-01	2.6E-01	1.5E-02	3.1E-01	5.4E-01	7.9E-01	9.3E-01
Control of Corruption	2.2E+03	-1.5E-01	1.0E+00	-1.8E+00	-8.9E-01	-4.3E-01	3.6E-01	2.5E+00
Regime types	2.2E+03	1.7E+00	9.4E-01	0.0E+00	1.0E+00	2.0E+00	2.0E+00	3.0E+00

VITA AUCTORIS

NAME:	Shirin Shahsavary
PLACE OF BIRTH:	Tehran, Iran
YEAR OF BIRTH:	1990
EDUCATION:	Alborz University, B.Sc, Qazvin, Iran, 2009-2013
	St. Clair College, Diploma, Windsor, ON, 2019-2021
	University of Windsor, M.Sc, Windsor, ON, 2021-2023