Global Warming and Natural Disasters to Global Peace Index

Wakako O'Sullivan
Student ID: 17143951

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Wakako O'Sullivan
Student ID:	17143951
Programme:	Data Analytics **Year:** 2020
Module:	MSc Research Project (Top Up)
Supervisor:	Dr Catherine Mulwa
Submission Due Date:	28th September 2020
Project Title:	Global Warming and Natural Disasters to Global Peace Index
Word Count:	9901 **Page Count:** 31

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Wakako OSullivan

Date: 25th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Global Warming and Natural Disasters to Global Peace Index

Wakako O'Sullivan
x17143951

Abstract

Global Peace Index (GPI) is an index that ranks the safety level of 163 countries, including 23 factors in three categories which are safety & security, ongoing conflict, and militarisation. The impact of global warming and natural disasters on human life and the problems they cause are increasing year by year in various aspects. Considering this, the problem is that GPI does not contain any factors on global warming and natural disasters. These research objectives to confirm the adequacy of adding information on global warming and natural disasters to existing GPI, and to contribute to the fact that the future of the earth should be seriously considered at each country and individual level. As methodology, 28 factors which are related to global warming and natural disasters were collected from five data sources, and variables were selected from them and 12 machine learning prediction models were performed. The result was that eight models presented more than 70% of accuracy. These results demonstrate the validity of incorporating global warming and natural disaster data into GPI. Based on this result, this project will contribute to awareness of the safety rank of each country in the world and reduction of global warming.

1 Introduction

1.1 Project Background and Motivation

The problem caused by global warming has become more serious over time (Weart, 2008). Natural disasters have always occurred in the natural history of the earth (Council, 1999). Disasters, such as Covid-19 an outbreak of a pandemic disease (WHO, 2020), big forest fires (Filkov, et al., 2019), heavy rain and floods caused by hurricane (Zhang, et al., 2018) resulted from the effects of global warming and natural disasters had tremendous impacts on human lives. Meanwhile, GPI (IEP, 2019) as an index showing the degree of peace in the countries in the earth was created. It shows the safety ranking of 163 countries with Iceland ranking first and Afghanistan last in the GPI 2019 survey. The ranking factors were composed of three categories which were related to safety and security, ongoing conflict, and militarisation, in which 23 factors were defined. It turned out that the current GPI did not include factors from global warming and natural disasters. These three existing categories were important, however considering the recent impact of global warming and natural disasters on human life, it was clear that talking about safety rankings without information of global warming and natural disasters was not persuasive. Therefore, the adequacy of adding information on global warming and natural disasters to current GPI will be investigated and will discover out how the problem can be improved and can contribute to future human activities in this project. Global warming and natural disasters methodology based on CRoss-Industry Standard Process (CRISP-DM) (Chapman, 1998) was used with 12 machine learning prediction models.

1.2 Description of Factors That Affect the Outcome

The following factors were collected as necessary information. Regarding the global warming, the weather information on 10 factors from 1901 to 2014 and information on CO2 emissions from 1901 to 2017 were collected. Regarding natural disasters, 14 disasters factors from 1900 to 2000, and the number of people infected and killed by Covid-19 were collected. Table 1 presents the details of factors.

Table 1: Factors of global warming and natural disasters

Global Warming	**Natural Disasters**
Weather related	Disasters
• Cloud cover	• Animal accident
• Diurnal Temperature Range	• Drought
• Ground Frost Frequency	• Earthquake
• Maximum Temperature	• Epidemic
• Mean Temperature	• Extreme temperature
• Minimum Temperature	• Flood
• Potential Evapotranspiration	• Fog
• Precipitation	• Impact
• Rain Days	• Insect infestation
• Vapour Pressure	• Landslide
CO2 Emissions	• Mass movement (dry)
• Rate of 1970	• Storm
• Rate of 2017	• Volcanic activity
	• Wildfire
	Covid 19 pandemic
	• Infected case
	• Number of deaths

1.3 Research Question

This research question examined the current problems in the literature review and found the adequacy of adding factors of global warming and natural disasters to existing GPI. The scope of this project was to collect certain data from five different sources and select suitable variable for implementation which were conducted 12 machine learning prediction models and evaluate the results. This research question and sub-research question address the importance of adding information factors of global warming and natural disasters to existing GPI.

***RQ**: "Can prediction of factors (weather, CO2 emissions, death toll by natural disasters and Covid 19) that contribute to global warming and natural disasters provide insights into improving safety of countries and reduce the problem of global warming and natural disasters?"*

Sub-RQ: "Can identification of factors contributing to global warming and natural disasters be able to give significant impact in the ranking of safety country in the current GPI?"

1.4 Research Objectives and Contributions

Table 2 presents research objectives to obtain the right answer for a research query.

Table 2: Research objectives

Objective No.	Details	Evaluation Methods	Method Option
Objective 1	Review on related work on GPI, global warming, and natural disasters		
Objective 2	Data collection and preparation		
Objective 3	Select variables by Multivariate Linear Regression		
Objective 4.1	Implementation, evaluation, and results of Generalized Linear Model	**train** function in package **caret** in R: **metric** = Accuracy **control** = 10-fold cross validation **method** = see Method Option	glm
Objective 4.2	Implementation, evaluation, and results of Linear Discriminant Analysis		lda
Objective 4.3	Implementation, evaluation, and results of Classification and Regression Tree		rpart
Objective 4.4	Implementation, evaluation, and results of k-Nearest Neighbours		knn
Objective 4.5	Implementation, evaluation, and results of Support Vector Machines with Linear Kernel		svmLinear
Objective 4.6	Implementation, evaluation, and results of Support Vector Machines with Radial Basis Function Kernel		svmRadial
Objective 4.7	Implementation, evaluation, and results of Random Forest		rf
Objective 4.8	Implementation, evaluation, and results of Neural network		nnet
Objective 4.9	Implementation, evaluation, and results of Gradient Boosting with Component-wise Linear Models		glmboost
Objective 4.10	Implementation, evaluation, and results of Bagging (Bootstrap aggregation)		treebag
Objective 4.11	Implementation, evaluation, and results of Naive Bayes		nb
Objective 4.12	Implementation, evaluation, and results of Conditional Inference Tree		ctree
Objective 5	Compare and Contrast the results from objective 4.1 to 4.12		
Objective 6	Comparison of developed models (objective 5) verses existing models		

Contributions: The major contributions resulting from this project was 12 global warming and natural disasters prediction models. These models will help provide insights into impact on GPI and contribution to resolve the problems caused by global warming and natural disasters. The minor contributions include identified factors contributing to global warming and natural disasters that in future can be used to significantly influence in the ranking of safe countries in the current GPI.

The rest of this report presents as follows: chapter 2 literature review, chapter 3 methodology and design, chapter 4 data preparation, chapter 5 implementation and result evaluation, chapter 6 discussion and chapter 7 conclusion and future work.

2 Literature Review of Global Warming and Natural Disasters (1992-2020)

In this chapter, the literature was reviewed from related areas of this project and find the gap. The subsections are as follows: 2.1 global risk indicators from other research, 2.2 evidence of current GPI, 2.3 history and past research of global warming and natural disasters, 2.4 data analytic methodology from the other research and 2.5 identify the gap.

2.1 Research in a Similar Direction

2.1.1 Global Risks Reports 2020

Global risks reports 2020 defined uncertain events or situations that could have significant negative effect on several countries or different industries over the next decade (Marsh & Group, 2020). It consisted of five risks related categories which were economy, environment, geopolitics, social risk, and technology with 30 factors in them. As a sample the category of social risk had the factor -- rapid and large-scale spread of infectious diseases, and corresponds to Covid-19 scenario, which was the worst case of infectious disease in history (Liu, et al., 2020). It was helpful to see the risks in the world and how those were defined. As for the risk's forecasts were currently underway, however it was found there was no research on prevention or precautionary measures. Also, this report presented some categories of risks and their factors were similar to this project. Because this report focused on risk, therefore results showed no country-specific data was specified. It was recognised that approach is different from this project and requires specific data.

2.1.2 Sustainable Development Goals (SDG) Indicator

This indicator was provided by United Nations Development Program (UNDP). The purpose of SDG indicator was the action approach to end poverty, protect the earth and ensure peace and prosperity by 2030 and it had 17 goals and 169 targets with 170 countries (Hák, et al., 2016; Streich, et al., 2020). Figure 1 presents the details of the goals. SDG had a framework, but there are no specific conceptual indicators (Hák, et al., 2016). It is totally agreed with point of Hák, et al., (2016) as indicators is required that could be expressed in clear numbers and this indicator was a helpful sample for this project.

Figure 1: 17 goals of SDG indicator[1]

[1] https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

2.2 Evidence of Previous Study of GPI

2.2.1 Categories and Factors in GPI

GPI presented the rank of 163 countries peacefulness by using comprehensive data related on peace (IEP, 2019). In the analysis process there were three main categories with 23 factors in them and covered 99.7% of population in the world in the current GPI. The first category contains area of safety and security with 11 factors. The second category contains the area of ongoing conflict with six factors. The third category contains the area of militarisation with six factors. Table 3 presents the details of each category with factors. These information were from a trusted source valuable (IEP, 2019) and indispensable elements for measuring the overall security level of each country. As an evidence there were no factor from global warming and natural disasters included and justified to add them in this project.

Table 3: Current three categories and 23 factors of GPI

Safety & Security		**Ongoing conflict**	
1	Level of crime recognized in society	12	The severity of organized internal conflicts
2	Number of internal guards and police officers*	13	Death by organized internal conflict
3	Number of murders*	14	Number and duration of internal disputes
4	Number of prisoners*	15	Relationship with neighbouring countries
5	Availability of getting small and light weapons	16	Number, period, and role in external conflicts
6	Possibility of violent demonstration	17	Deaths from organized external conflicts
7	Violent crime level	**Militarisation**	
8	Political unrest	18	Transfers as recipients of major conventional weapons (imports)*
9	Terror scale of political	19	Military expenditure (% of GDP)
10	Impact of terrorism	20	Number of armed professionals*
11	Number of refugees (internally or external reason) **	21	Financial contribution to UN peacekeeping
		22	Nuclear and heavy weapon capabilities
	* per 100,000 people ** % of the population	23	Transfers as supplier of major conventional weapons (Exports)*

2.2.2 Concept and Method of Calculation in GPI

It was difficult to define the degree of peace, therefore the GPI measured negative factors, and defined with lower results indicating higher levels of peace as the concept (Clements, 2019). The calculation method of the score was banded or normalized in five steps for each of 23 factors. The qualitative attributes were grouped into five levels such as 1 to 5 (very low, low, moderate, high, and very high) and the quantitative attributes were separated by three decimal places and divided into five ranges. Because negative factors were used for scoring, therefore there was a lower score and the country can be considered more peaceful (IEP, 2019).

2.3 Research of Additional Categories

In this section, the literatures are highlighting about the knowledge of the history of global warming and the actions humans have taken, natural disasters, and the impact they gave on human livings.

2.3.1 History of Recognition of Global Warming and its Countermeasures

Global warming had two causes which were nature and humans. A typical example of nature were the originally present atmospheric carbon dioxide (CO_2) which was known greenhouse gas and volcanic eruptions. Humans had increased their carbon footprint and had accelerated global warming (Goel & Bhatt, 2012). The global warming was recognized as a greenhouse effect from around 1820s. Around 1890s, it was reported that halving CO_2 emissions would lower the Earth's surface temperature by four to five degrees. Later, it was discovered that human industrial activity was helping to increase temperatures, and in 1938 it was confirmed that levels of CO_2 emissions had steadily increased. However, it was only in the 1960s that scientists began reporting that CO_2 was causing global warming. It was reported that if there was no CO_2 on earth, today's temperature would be much lower (Letcher, 2019). As a global action, United Nations Framework Convention on Climate Change (UNFCC) was signed in 1992 as an environmental treaty (Sands, 1992) that set an international framework on global warming. The conference parties (COP) to the UNFCC had meetings to negotiate the details. The third COP was the first global countermeasure to address global warming as it was adopted as the 1997 Kyoto Protocol (Oberthür & Ott, 2013). It had set six greenhouse gas which are CO_2, methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulphur hexafluoride (SF_6) reduction targets, aiming to reduce the total emission of six greenhouse gases across developed countries from 2008 to 2012 by at least 5% compared to 1990 (Goel & Bhatt, 2012). Subsequently, the Paris agreement (NATIONS, 2015) was adopted at the 21st COP in 2015. This was the first framework of all 196 countries that will participate in the Framework Convention on Climate Change, prescribing global warming measures after 2020. The object was to keep the average temperature rise in the world below two degrees Celsius and to aim for the average temperature rise below 1.5 degrees Celsius by reducing CO_2 emissions (Rogelj, et al., 2016). The countermeasures and approaches varied from country to country. Although global warming was treated as a serious problem, the reason why it was not included in GPI cannot be explained. Therefore, the research of this project to incorporate the factors of the global warming were appropriate.

2.3.2 Natural Disasters

Natural disasters were categorized into four hazard families geophysical, meteorological, hydrological and climatological which are based on the data of natural catastrophe statistics online NatCatSERVIC[2] (Hoeppe, 2016). As the main events of geophysical family included earthquakes, volcanic eruptions, and mass movement (dry). The geophysical disasters included ground shaking, fire following, tsunami, volcanic eruptions, subsidence, rockfall and landslide. The meteorological disasters included a storm which were broken down to tropical cyclone, winter storm, tempest, hailstorm, lightning, tornado, local windstorm, sand/dust storm, blizzard, and snowstorm. The hydrological disasters included floods and mass movement (wet) which were broken down to general flood, flash flood, storm surge, subsidence, avalanche, and landslide. The climatological disasters included extreme temperatures, droughts and wildfire

[2] https://www.munichre.com/en/solutions/for-industry-clients/natcatservice.html

which were broken down to heat wave, cold/frost wave. This classification helped this project to classify the natural disasters dataset.

Earthquake was one of geophysical typical natural disasters. The mechanism of earthquake was related to plate tectonics on Earth. There were seven large plates (Eurasian, North American, Pacific, Australian, Nazca, South American and African) and many small plates on the earth (F.Luhr, 2013). Six of the large plates hold the continent. These were place of residence. The movement of the plate causes mountains to be built, volcanic activities, island formation, grand rifting at the points where the plates contact each other (Greig, 2017; Packham, 2017). It was understandable that the countries which had many disasters caused by earthquake were located near the joints of plates. Regarding earthquakes, location information, its scale, the number of occurrences, and the number of fatalities is important information.

This project considered Covid-19 as one of the natural disasters for the development of pandemic disease. Covid-19 began spreading from China in December 2019, and by the mid July 2020, nearly 12 million people had been infected and nearly 570,000 were killed worldwide (Liu, et al., 2020). Data on the number of infected people and the number of fatalities can be useful to check the virus countermeasures in each country and the degree of improvement of facilities such as hospitals, and to measure the safety of the country. Therefore, the research of this project to incorporate the factors of the natural disasters were appropriate.

2.4 Data Analysis Methods with Machine Learning

2.4.1 Sample Case from the Other Research

This literature (Fekete, 2018) presented the disaster resilience assessment at city level in Germany by providing indicators on resilience and vulnerability from the risk of disasters, and tested by using demographic (birth rate, care homes and number of hospital beds), infrastructure (water containers, wells and electricity) and socioeconomic (social aid, GDP and election participation) data to find if there was a significant difference between both indicators. The survey results were presented on a map and it tracks aspects of resilience over a period of 5 and 10 years. Of interest in this research was the procedure of generating indicators from natural disasters data. First, they prepared three attribute data in each of the three data sections and displayed the data from 2005 to 2015 on maps. The number of disasters, the amount of debt, and the number of fatalities were displayed in a table and compared. The data collection using the three attributes was like this project and it was helpful. However, unfortunately, detailed analysis of data such as machine learning was not provided, and details of how to handle the data was not described.

As a good reference example of machine learning from a banking system, an developing early warning system of bank distress by using machine learning models (Suss & Treitel, 2019) were reviewed. In detail, they were using linear statistical model, random forest, k-nearest neighbour (KNN), boosting, decision tree and support vector machine (SVM). And those results were compared to find the best model in the research. Although the field of research was different, the variety of uses of machine learning for non-parametric data and the design of reports including presentation of results were very understandable and helpful.

2.4.2 Parametric and Non-Parametric Test

This literature (Russell & Norvig, 2016) provided a clear explanation of parametric and non-parametric tests in machine learning. In the case of parametric, when it was known in advance that the distribution of the population was a certain probability distribution, if the parameters were known, it was possible to find what the distribution of the population looks like. And in the case of non-parametric, the population distribution was not known in advance and the population distribution could not be determined by some parameters. Common algorithms for parametric testing included logistic regression, Naive Bayes, and simple neural networks. The advantages were fast learning, the results were easy to understand, and could be analysed without the collection of large amounts of data. As the limitations, when functional form was selected, these methods were constrained to the selected form, suitable for only simple case study and the methods were poorly fit. The key algorithms for non-parametric testing include kNN, decision trees and SVM. The advantages were that could be adopted to many functional forms, no assumptions needed and useful performance for prediction models. As the limitation, it required large amount of data, slow run due to big number of parameters and high risk of overfitting. This project took the characteristics of these tests into consideration, and perform those tests using a model suitable for each case.

2.5 Identify the Gap

As for the gap, the current GPI only provided a world peace ranking with categories of safety and security, ongoing conflict, and militarisation, and did not include any factors from global warming and natural disasters. Due to the gap, it was believed that it was currently lacking the credibility of world peace rankings. It was understood the importance of the two missing categories in current GPI by the literature research. This project examines the adequacy of adding those new categories by using 12 machine learning prediction models along with global warming and natural disaster methodology to fill the gap.

In this chapter, four different areas of literature review were proceeded. It introduced similar indicator to GPI, introduction of GPI, history of global warming and human action on it, types of natural disasters and analytical methodologies. The next chapter 3 presents the methodology and design.

3 Methodology and Design

This section presents global warming and natural disaster methodology and design architecture.

3.1 Methodology

Global warming and natural disaster methodology based on CRISP-DM is a process model for this project, which consists of six stages. It is designed to cycle back and forth as needed to produce the correct result, as shown in Figure 2 (Wirth & Hipp, 2000). This section describes the process of this project in six stages.

3.1.1 Stage 1: Business Understanding

It was understood what GPI is and what is current issue. It was considered how global warming and natural disasters can give any impact to improve or resolve the issue of GPI and seek how this result can contribute to the problem on global warming and natural disasters.

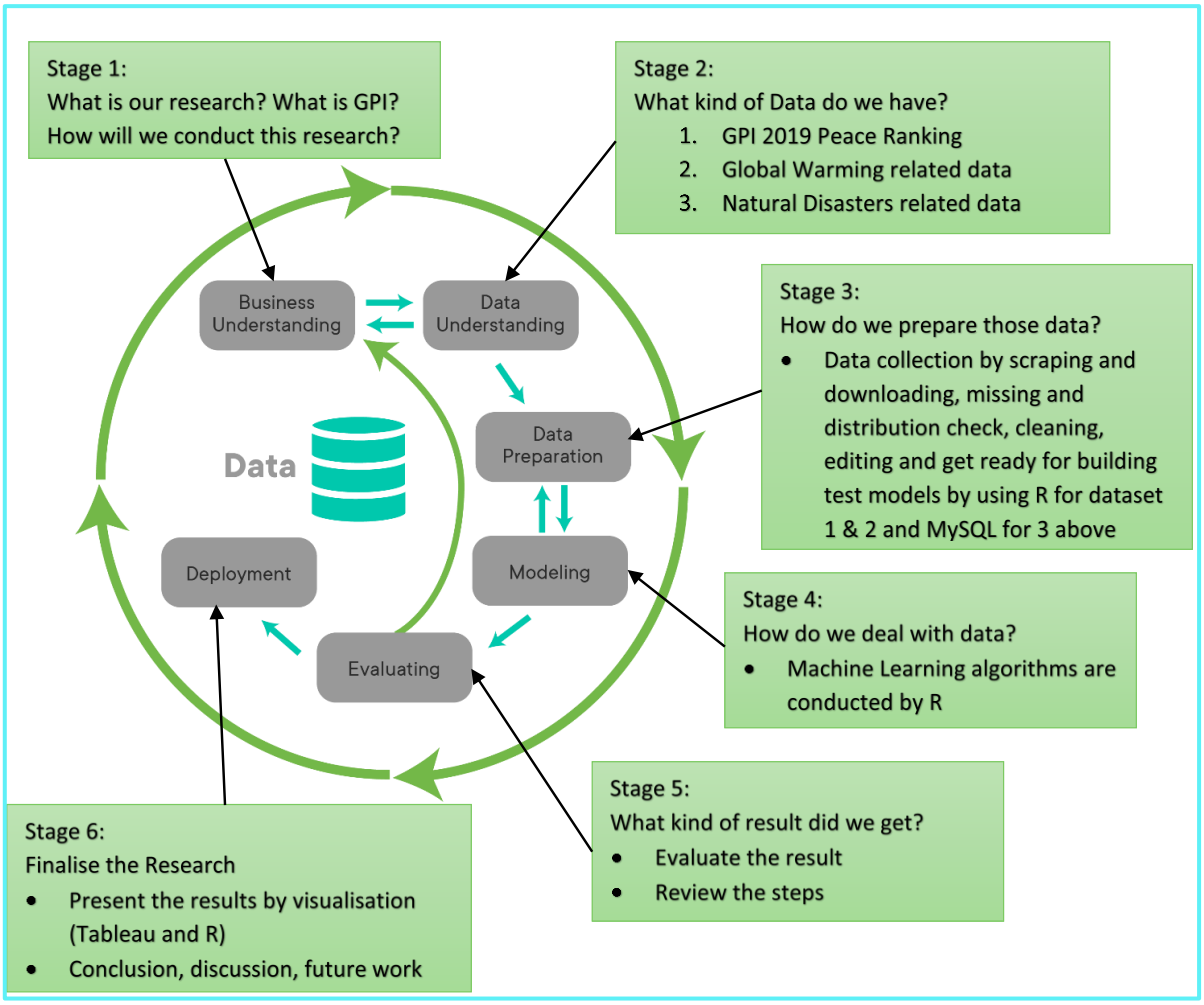

Figure 2: Global warming and natural disasters methodology

3.1.2 Stage 2: Data Understanding

This project required datasets from five different sources. Data 1 was the world peace score of 163 countries and its ranking and was the source from the GPI of 2019[3]. All five data sets were collected based on this country list. Data 2 is about weather-related data which was provided by Climatic Research Unit (CRU) of University of East Anglia data library[4]. This data includes 10 factors which were shown in the Table 1. This data helped to find the importance of weather information contributing to GPI. Data 3 was the time series of CO2 emissions rate from 1970 to 2017 which were provided by EU Open Data Portal[5] and it was considered as a part of global warming information. It is helpful to see the increase or the decrease of CO2 emission over 40

[3] http://visionofhumanity.org/app/uploads/2019/07/GPI-2019web.pdf
[4] https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_3.23/crucy.1506241137.v3.23/countries/
[5] https://data.europa.eu/euodp/data/dataset/a7fb0a23-2f71-4d03-a73f-3b41ab62febf

years period, and it is one of important measurement. Data 4 was natural disasters related and provided by School of Public Health Université catholique de Louvain [6], and contains natural disasters information included the 14 factors which were shown in Table 1. This data helped to find the importance of natural disasters information contributing to GPI. Data 5 was time series of Covid-19 related information such as number of infected cases, death count, number of beds, population etc. for each country[7] from December 2019 to June 2020. This data is helpful to use as pandemic of disease as part of natural disasters.

3.1.3 Stage 3: Data Preparation

The details of data preparation include data collection and data selection will be explain in the chapter 4: data preparation for global warming and natural disasters data.

3.1.4 Stage 4: Modelling

12 different machine learning prediction models which are, generalized linear model, linear discriminant analysis, classification and regression tree, k-nearest neighbours, support vector machines with linear kernel, support vector machines with radial basis function kernel, random forest, neural network, boosting, bagging, Naive Bayes and Ctree were conducted to find a solution to the research question. The details will be explained in the chapter 5: implementation.

3.1.5 Stage 5: Evaluating

The results from stage 4 and their comparative evaluation are presented in this section and all the processes are reviewed and the validity of the test models its methods are confirmed. If there are any corrections or better suggestions, it should be considered for improvement at this stage. Importantly, to find the solution to this research question, the adequacy of factors along with the results is discussed.

3.1.6 Stage 6: Deployment

This is the last section of this project. The project was discussed, concluded and future works were suggested, and created the final report. Visualization part which are generated by Tableau and R were the key method of the presentation in this project report.

3.2 Data Architecture Design

As Figure 3 presents, a three tier architecture design which was consistent with presentation tier, application tier and database tier (Bretl, et al., 1999) were applied to this project. Five different sources were selected, and those five data sets were collected and prepared at the database tier. The implementation was conducted by 12 machine learning prediction models and results were evaluated by using R and MySQL at application tier. At the presentation tier, the evaluated results were visualized by R and Tableau to create reports and presentation videos.

[6] https://emdat.be/database

[7] https://github.com/owid/covid-19-data/tree/master/public/data

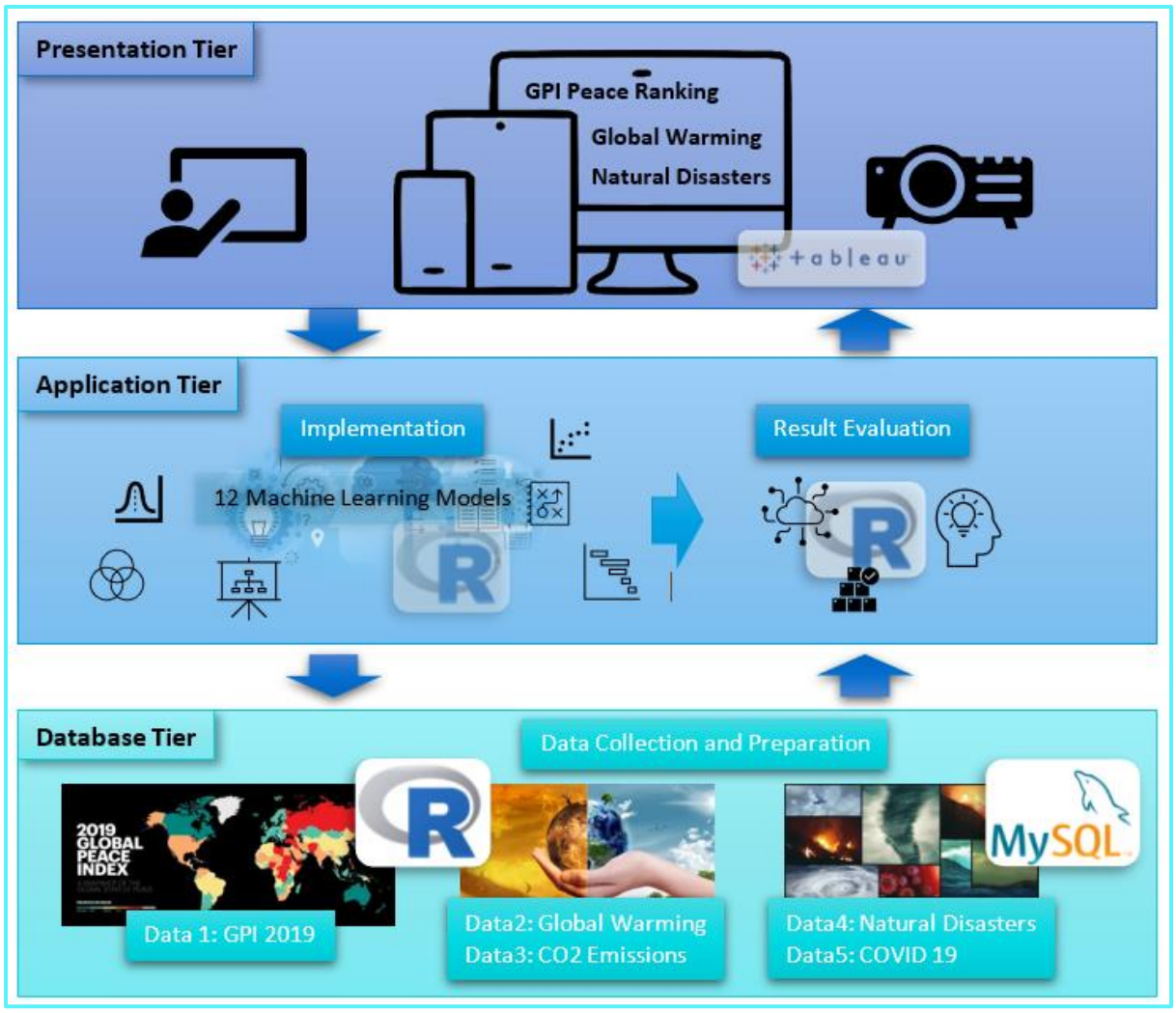

Figure 3: Design Specifications of Global Warming

This section introduced the process of the 6 stages of this project based on global warming and natural disaster methodology and each tier of the design by three tier architecture. The next chapter 4 will explain about data preparation.

4 Implementation of Data Preparation

This chapter describes the procedure and method of implementation which includes data collection and data preparation.

4.1 Introduction

Figure 4 presents the process flow diagram of the data preparation and the implementation. It starts from data collection from five different sources. The first data preparation stage dealt with data cleaning and missing data check by two different groups which are global warming and natural disasters. The second data preparation stage processed distribution check, correlation, normality test and outliers checking. After data sets were ready, tests were conducted by 12 machine learning prediction models followed by result evaluation. At the presentation stage, the technical report and presentation video were produced.

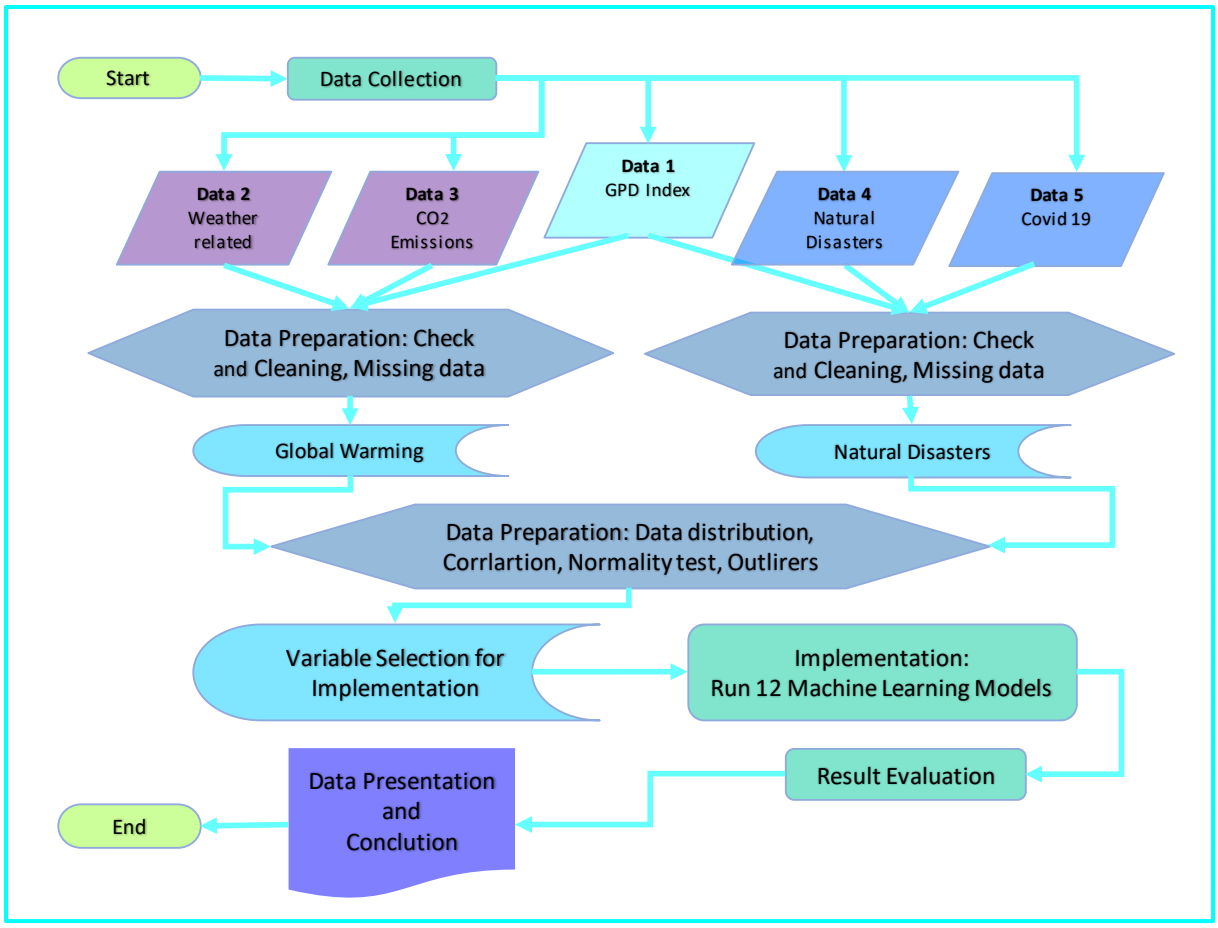

Figure 4: Process flow diagram of implementation

4.2 Data Collection and Pre-Processing

4.2.1 Data Collection

Data 1 is scraped from page 10 and 11 in a pdf file in the site. The image of the pdf is presented in configuration manual (abbreviated as CM) 4.1.1 Fig 30. The table in the pdf is displayed in six columns with the data of ranking, country name, GPI score, and national flag in one set. After scraping the data from the site, six sets of data excluding national flags were vertically concatenated, and the ranking, country name, and score of 163 countries were obtained as information and saved as GPI2019_country163. No missing data confirmed. The 163 countries included in data 1 apply to all data (data 2 to data 5). The programming by R is presented in CM 4.1.1 Fig 31.

Data 2 is scraped from data library. Before scraping the data, the country names of data 1 and data 2 were checked and modified to match (CM 4.1.1 Fig 32 & 33). Data 2 has ten factors (CM 4.1.2 Fig 34 and CM 4.1.2 Table 2) and each factor has independent site and has data for 289 countries independently (CM 4.1.2 Figure 35). Each country data contains 18 items which are YEAR (from 1901 to 2014), JAN to DEC (independent 12 columns, monthly means), MAM, JJA, SON, DJF (independent 4 columns, seasonal means) and ANN (annual mean) (CM 4.1.2 Figure 36). The main purpose was to scrape target 163 countries data for 10 factors from each site, combine them all, and extract ANN by country. The country name is the key to all data collection in the program therefore it should be an exact match. The following two rules

were noted in the country name modification. As rule No.1, if there is a space in the country name, it should be replaced with underscore, because data 2 used underscore in the URL. As rule No.2, if there is a special character such as special dots or some mark, it should be replaced with the relevant character, because R does not read them properly. Data 1 was also reviewed and applied the same rules as data 2 to the country names. Since there were big number of sites to be scraped (1630 sites = 163 countries for 10 factors), it was automated in the programming about scraping and combined them and saved as GW_ave_all. (CM 4.1.2 Fig 37 to 44). Since this is time-series data, the temperature difference and the temperature increase rate were obtained by calculating the mean temperature average of the oldest 10 years 1901 to 1910 and the average of the newest years 2005 to 2014 (CM 4.1.2 Fig 45 to 47). Data of 1901 to 2014 was averaged for each country and the temperature difference and the increase rate of temperature calculated above were jointed and saved as GW_agg_ave_all (CM 4.1.2 Fig 48).

Data 3 was downloaded from the website. First, the country name and CO2 emissions rate of 1970 and 2017 are extracted from the original source followed by correction of the country names as in data 2 (CM 4.1.3 Fig 49). Of the 163 countries, for those without data, 0 is set as a dummy and treated as missing data, and when two countries are treated as one country such as Serbia and Montenegro, they were made independent and finally unified into 163 countries (CM 4.1.3 Fig 50 to 53). From the CO2 emissions data of 1970 and 2017, the dataset was enriched by calculating[8] increase rate of CO2 emissions. Those four attributes were saved as CO2_Agg and joined with GPI information into GW_ALL (CM 4.1.3 Fig 54). It was decided to use the mean temperature, therefore minimum and maximum temperature were removed (CM 4.1.3 Fig 55). When all information was collected, the missing data were predicted and filled by using the random forest and saved as final version of data (CM 4.1.3 Fig 56 & 57).

Data 4 is downloaded from data library website. The programming was created with SQL by MySQL and programming was run by batch file from the Windows command prompt (CM 4.2 Fig 58) for data 4 and data 5. A table was created in a database in MySQL and data was loaded (CM 4.2.1 Fig 59). The country names were checked and modified according to the rules it was done in data 2 and made a note of any missing data countries (CM 4.2.1 Fig 60 to 66). From there, extract the number of deaths per disaster and total count of disaster per country.

Data 5 is downloaded from the website. This data was time series of Covid 19 information. A table was created in a database in MySQL and data was loaded. The latest data[9] of total case per million, total deaths per million and population for each country were extracted (CM 4.2.2 Fig 67). The country names were checked against GPI data and modified according to the rules it was done in data 2, and it was treated the same with data 3 for countries with missing data (CM 4.2.2 Fig 68 to 70). The death counts by disasters were converted in proportion to one million people and dummy data for missing country was inserted (CM 4.2.2 Fig 71). To enrich the dataset, the death rate was calculated and added as a variable by using the total cases and the total deaths of Covid 19 data. The data array was flat; therefore, it was transposed to a data frame type to obtain the data for each natural disaster by country and added the death rate[10].

[8] Increase rate = CO2 Emission 2017 /CO2 Emission1970

[9] Data as of 8th June 2020

[10] Covid 19 death rate = Covid 19 case/Covid 19 per million

(CM 4.2.2 Fig 72). The total number of disaster per country was added and joint the GPI score from data1.The header name was shortened for the later test and data was export from database in MySQL as ND_ALL with csv format (CM 4.2.2 Fig 73). Four attributes which are animal accident, fog, impact, and insect infestation were deleted because all those data value was 0. Regarding the missing data, the random forest was used to predict and fill in R (CM 4.2.2 Fig 74 & 75).

4.2.2 Collected Final Data

Table 4 presents the summary of the collection method, tools, data size and the final output (size). GW_ALL and ND_ALL were used in the stage of the implementation.

Table 4: Data summary of collected data

Data	File name	Collection method	Tool	Data size		Final output
				observation	attribute	
1	GPI2019_country163.csv	Scraped	R	163	3	
2	GW_ave_all.csv (GW_sgg_ave_all.csv)	Scraped	R	17670 (163)	12 (12)	*GW_ALL (163*15)
3	fossil_CO2_totals_by_country.csv	Downloaded	R	211	49	
4	emdat_public_2020_04_29_query_uid-0ZbMRD.csv	Downloaded	MySQL	15448	45	*ND_ALL (163*16)
5	owid-covid-data.csv	Downloaded	MySQL	22808	29	
						*GW: Global Warming, ND: Natural Disasters

Table 5 presents details of attributes, label, and data type. There are 15 in GW_ALL and 16 in ND_ALL.

Table 5: Details of final output of GW_ALL and ND_ALL

GW_ALL				ND_ALL			
#	**Attributes**	**Label**	**Type**	**#**	**Attributes**	**Label**	**Type**
1	Country name	Country	Char	1	Country name	Country	Char
2	Cloud cover	Cld_cv	Num	2	Drought	Drought	Num
3	Diurnal temperature range	Temp_day	Num	3	Earthquake	EarthQ	Num
4	Ground frost frequency	Gnd_Fr	Num	4	Epidemic	Epidemic	Num
5	Potential evapotranspiration	Pot_Eva	Num	5	Extreme temperature	Ex_temp	Num
6	Precipitation	Prcp	Num	6	Flood	Flood	Num
7	Mean temperature	Mean_Tmp	Num	7	Landslide	LandS	Num
8	Vapour pressure	Vap_prs	Num	8	Mass movement	Mass_move	Num
9	Rain days	Rn_day	Num	9	Storm	Storm	Num
10	Temperature difference	Tp_diff	Num	10	Volcanic activity	Volcanic	Num
11	Temperature increase rate	Tp_IncR	Num	11	Wildfire	Wildfire	Num
12	CO2 emission of 1970	CO2_1970	Num	12	Total disaster	Total_Disaster	Num
13	CO2 emission of 2017	CO2_2017	Num	13	Covid 19 infected case per million	COV_CASE	Num
14	Increase rate 2017/1970	Inc_Rate	Num	14	Covid 19 death per million	COV_DEATH	Num
15	GPI score	GPI_Score	Num	15	Death rate	COV_D_Rate	Num
	# is column number			16	GPI score	GPI_Score	Num

4.2.3 Data Pre-Processing -- Data Check Results and Reference

The initial check was conducted on data GW_ALL and ND_ALL by R. The details of checks and method, target variables, output style, programming code and result references are presented in Table 6.

Table 6: Details of data preparation

Purpose	Method	Target variables		Output	Reference to Configuration manual
		GW_ALL	ND_ALL		
Missing data check	sapply	2 to 15*	2 to 16*	Data	4.3 Fig 76
Data distribution check	Histogram	2 to 14*	2 to 15*	Chart	4.3.1 Fig 77 to 80
Outlier check	Q-Q plot	2 to 14*	2 to 15*	Chart	4.3.2 Fig 81 to 84
Normality test	Shapiro-Wilk	2 to 14*	2 to 15*	Data	4.3.3 Fig 87 to 89
Correlation check	ggpairs	2 to 14*	2 to 15*	Chart	4.3.4 Fig 90 to 92
	*presents column number of # in Table 5				

4.2.4 Data Pre-Processing – Variable Selection for Building Models

After outliers check, outliers were replaced with the value of median (CM 4.3.2 85 & 86). As a result of the correlation, ones with 60% or more relations were deleted, and those were applied in case 1 below. Table 7 presents cases of the test, description, and reference to configuration manual. All cases in Table 7 perform AIC[11] (Portet, 2020; Lukacs, et al., 2009) which is one of multivariate linear regression (K.Ardakani & Seyedaliakbar, 2019; Yanagihara, 2006) for variable selection, by using the GPI score as the response variable.

Table 7: Case of variable selection methods

Case #	Description	Reference to Configuration manual
Case 1	Remove highly correlated variables manually, from each data set	4.3.5 Fig 93 to 96
Case 2	Let AIC select variables on each data set	4.3.5 Fig 97 to 100
Case 3	Join both data sets and let AIC select variables from the data set	4.3.5 Fig 101 to 104
Case 4	Join both data sets, following transformation and let AIC select	
Case 4.1	Natural Log transformation	4.3.5 Fig 105 & 106
Case 4.2	Natural Log10 transformation	4.3.5 Fig 107 & 108
Case 4.3	Square root transformation	4.3.5 Fig 109 & 110
Case 4.4	Reciprocal transformation (1/x)	4.3.5 Fig 111 & 112
Case 4.5	Power of 3 transformation	4.3.5 Fig 113 to 116
Case 4.6	Exponential transformation	4.3.5 Fig 117 & 118
Case 4.7	Sine transformation	4.3.5 Fig 119 to 122
Case 4.8	Absolute transformation	4.3.5 Fig 123 to 126

4.2.5 Selected Variables Results

Table 8 shows the results of adjusted R-squared (R^2). Because of it was not high enough in case 1 as 0.3015, therefore several cases were conducted to obtain the higher value of R^2. Case 2 shows the highest value of R^2 as 0.3655 and the lowest value of AIC -285.5 therefore it was determined to use the variables from the case 2 which are 4 factors from global warming

[11] Akaike Information Criterion

(cloud cover, diurnal temperature range, vapour pressure, CO2 emissions increase rate 2017/1970) and four factors from natural disasters (volcanic activity, wildfire, Covid 19 infected case).

Table 8: Result summary of adjusted R-squared and validated models

Case	R^2	AIC	Model Provided
Case1	0.3015	271.53	GPI_Score ~ Temp_day + Mean_Tmp + Drought + Volcanic + Wildfire + COV_CASE
Case2	**0.3655**	**-285.5**	**GPI_Score ~ Cld_cv + Temp_day + Vap_prs + Inc_Rate + Drought + Volcanic + Wildfire + COV_CASE**
Case3	0.3196	-285.5	GPI_Score ~ Cld_cv + Temp_day + Vap_prs + Inc_Rate + Drought + Volcanic + Wildfire + COV_CASE
Case4.5	0.1577	677.26	GPI_Score ~ Cld_cv + Temp_day + Gnd_Fr + Drought + Wildfire
Case4.7	0.1318	-394.53	GPI_Score ~ Pot_Eva + Prcp + Tp_diff + LandS + Storm + Wildfire
Case4.8	0.3225	-285.5	GPI_Score ~ Cld_cv + Temp_day + Vap_prs + Inc_Rate + Drought + Volcanic + Wildfire + COV_CASE

4.2.6 Data Validation for Selected Variables

The selected variables from 4.2.5 were validated by using linear regression on the relationship between individual variables to GPI scores. (CM 4.3.6 Fig 127 to 140). Graphs in blue areas are from global warming and in purple areas are from natural disaster in Figure 5. Most of the data in blue can be relatively linearly regressed and most of the data in purple tend towards to x = 0, and many of them do not have a satisfied linear regression. Those biases in blue and purple can be inferred to reduces accuracy and one of the key causes. To verify the value obtained in case 2 of 4.2.5 (Cld_cv, Temp_day, Vap_prs, Inc_Rate, Drought, Volcanic, Wildfire, COV_CASE), selected variables are marked in red on the graph in Figure 5. It can be understood that it is relatively linear regression for Cld_cv, Temp_day, Vap_prs and Volcanic, but others are difficult to find. It should be kept in mind that even a small value of R^2 it has a significant difference from 0, and the regression model is statically significant (Neter, et al., 2004). Therefore, those selected variables are used for implementation. Next chapter 5 will explain data implementation and result evaluation.

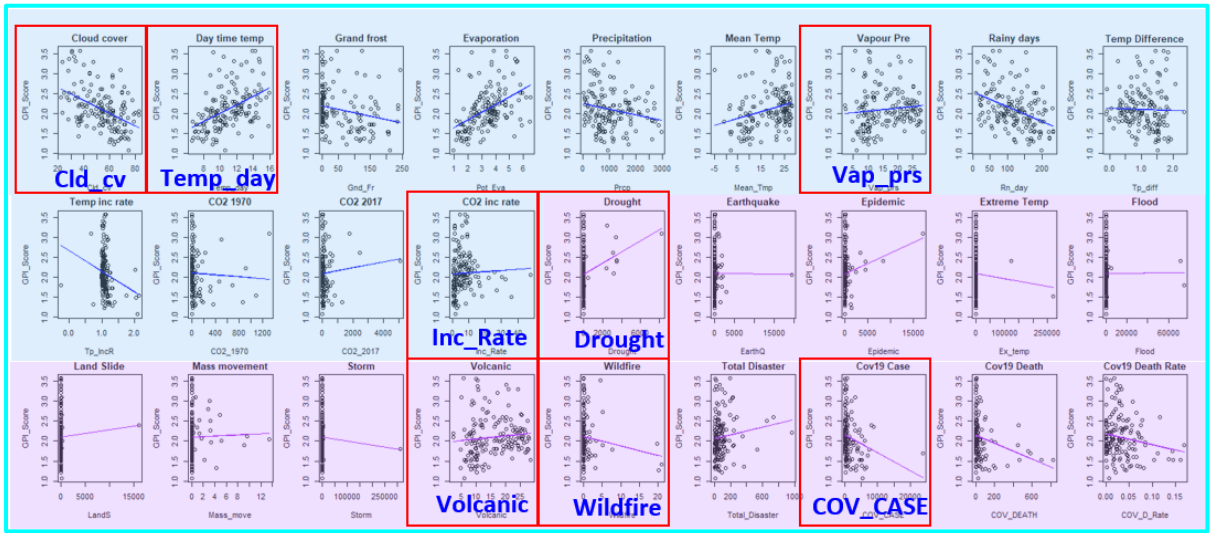

Figure 5: Linear regression for all variables on GPI score

5 Implementation, Evaluation and Results of Prediction Models

This chapter presents the implementation of 12 machine learning prediction models and the results evaluation.

Evaluation: To evaluate the developed models, the following evaluation methods were used: function train in R package caret and set (i) metric = Accuracy, (ii) control = 10-fold cross validation, (iii) method = depends on model. Experiments were conducted to test and validate each model.

5.1 Introduction

By using variables which were selected in 4.2.5, 12 following machine learning prediction models were conducted, and results were compared to find the most suitable model to understand the adequacy of adding global warming and natural disasters information to GPI. From literature review 2.4.2 and Figure 5, the distribution of data used in this project varies, therefore 12 machine learning prediction models were implemented assuming parametric tests and non-parametric tests.

To obtain binary response variable, the risk rank was set as 1 if the GPI score is above the average and 0 if it is below the average (CM5.1 Fig 141 to 143). The data was split into 70 percent and 30 percent for training and testing (CM 5.1 Fig 144). The tests were conducted by using the train function from the R package caret. The train function requires to set response variable, dataset, method, metric and trControl. Common options were set as an initial settings such as 10-fold cross-validation (Fushiki, 2011) as cv for trControl and accuracy was set as accuracy (CM 5.1 Fig 145). Risk rank was set as the response variable, train data was set as main data and method was specified by each model. As the results of running 12 machine leaning prediction models, those data were obtained as output: accuracy, kappa, sensitivity, specificity, positive predicted value, negative predicted value, prevalence, detection rate, detection prevalence and balanced accuracy (AUC). By using the confusion matrix, predicted data was used from the result of train function which was the cross-validation test was conducted between the test data and the predicted data for each model. Based on the results, AUC (Area Under the Curve) - ROC (Receiver Operation Characteristics) curve was created to visualize the model performance with confidence interval zone in blue in each result graphic which will be present in each experiment. Regarding the AUC-ROC curve and confusion matrix, literature reviews are in CM 5.2. From the literature review, following definition was adopted for the rate of AUR-ROC curve model evaluation: 0.9 – 1.0 excellent, 0.8 – 0.9 very good, 0.7 – 0.8 good, 0.6 – 0.7 satisfactory, 0.5 – 0.6 unsatisfactory, under 0.5 failed.

5.2 Experiment 1: Generalized Linear Model

5.2.1 Implementation

GLM is a linear model with residual distributions of arbitrary distribution. These models include linear regression, Poisson regression, logistic regression and it was advocated by (Nelder & Wedderburn, 1972). The implementation code is in CM 5.3 Fig 148 and 169.

5.2.2 Evaluation and Results

The key results are accuracy 0.7917, sensitivity 0.8571, specificity 0.70, precision (positive predictive value) 0.8 and AUC (balanced accuracy) 0.7786 in Figure 6. This is a good model.

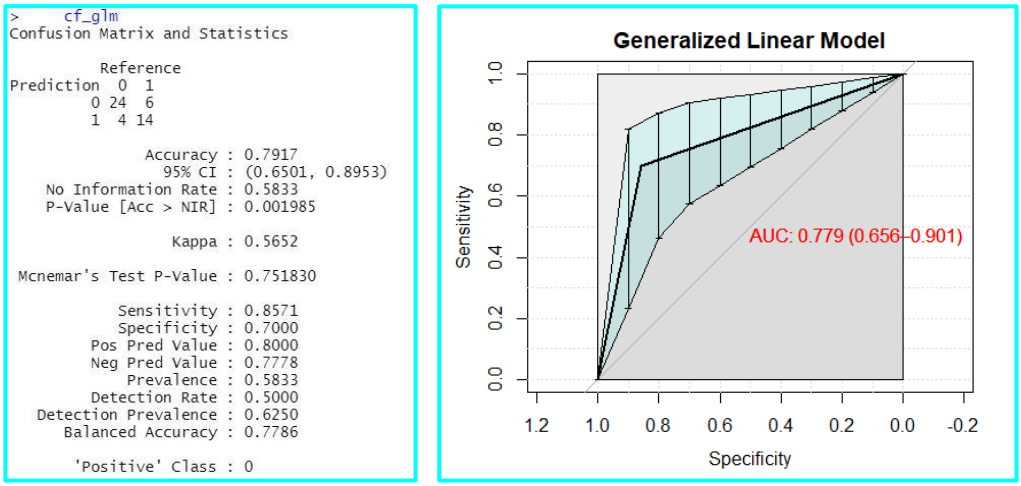

Figure 6: Result from GLM

5.3 Experiment 2: Linear Discriminant Analysis

5.3.1 Implementation

LDA is also called liner classifier and it is a statistical classifier that performs classification based on the value of linear combination of features. In machine learning, classification aims to classify items into groups based on feature values. (Yuan, et al., 2012). The implementation code is in CM 5.3 Fig 149 and 169.

5.3.2 Evaluation and Results

The key results are accuracy 0.75, sensitivity 0.8214, specificity 0.65, precision 0.7667 and AUC (balanced accuracy) 0.7357 in Figure 7. This is a good model.

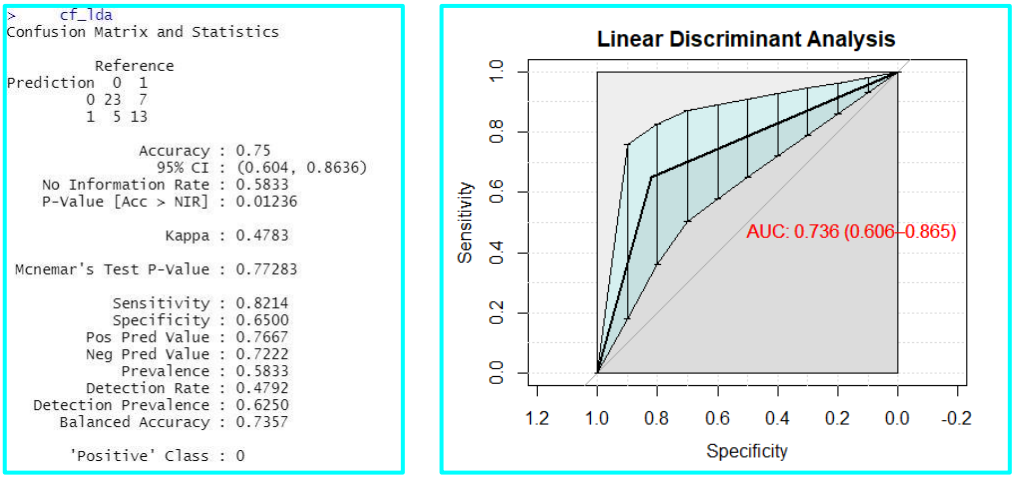

Figure 7: Result from LDA

5.4 Experiment 3: Classification and Regression Tree

5.4.1 Implementation

CART is a recursive partitioning machine learning algorithm for predicting continuous dependent variables (regression) and categorical predictors (classification). As the output, it builds regression tree and classification graphically. (Loh, 2011; Zhaoab, et al., 2016). The implementation code is in CM 5.3 Fig 150 and 170.

5.4.2 Evaluation and Results

The key results are accuracy 0.7292, sensitivity 0.6786, specificity 0.80, precision 0.8261 and AUC (balanced accuracy) 0.7393 in Figure 8. This is a good model.

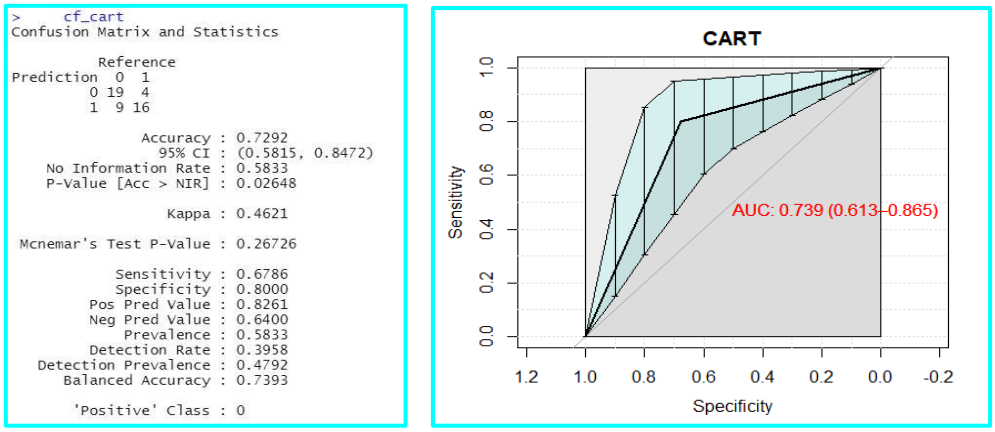

Figure 8: Result from CART

5.5 Experiment 4: k-Nearest Neighbours

5.5.1 Implementation

KNN is a classification and regression method based on the nearest training example in the feature space and is often used in pattern recognition. The classification of an object is determined by voting on its neighbours. (Ferreira, et al., 2015). The implementation code is in CM 5.3 Fig 151 and 170.

5.5.2 Evaluation and Results

The key results are accuracy 0.5417, sensitivity 0.6786, specificity 0.35, precision 0.5938 and AUC (balanced accuracy) 0.5143 in Figure 9. This is an unsatisfactory model.

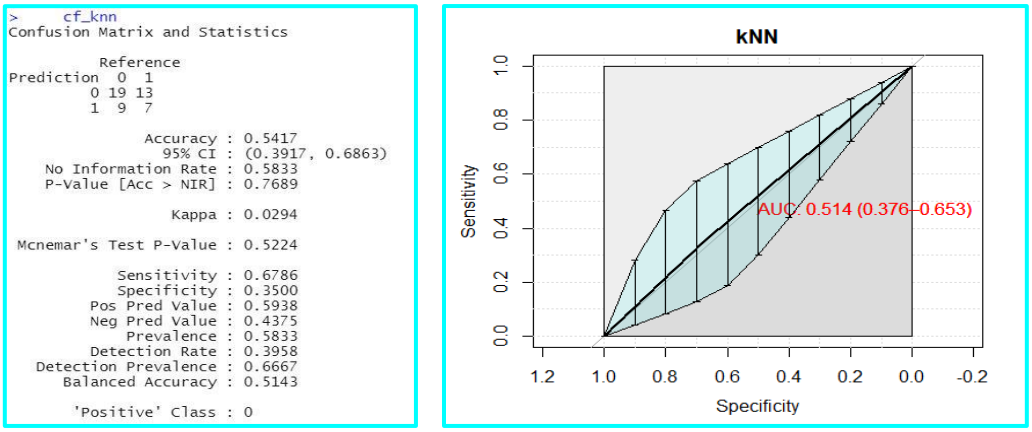

Figure 9: Result from kNN

5.6 Experiment 5: SVM with Linear Kernel

5.6.1 Implementation

SVM is one of the pattern recognitions models that uses supervised learning. It can be applied to classification and regression by drawing a boundary line that clearly separates classes. It draws a line by the method of maximizing the margin which is the distance from the boundary to the point that also has the closest feature. SVM was announced by (Vapnik & Lerner, 1963). The implementation code is in CM 5.3 Fig 152 and171.

5.6.2 Evaluation and Results

The key results are accuracy 0.75, sensitivity 0.8214, specificity 0.65, precision 0.7667 and AUC (balanced accuracy) 0.7357 in Figure 10. This is a good model.

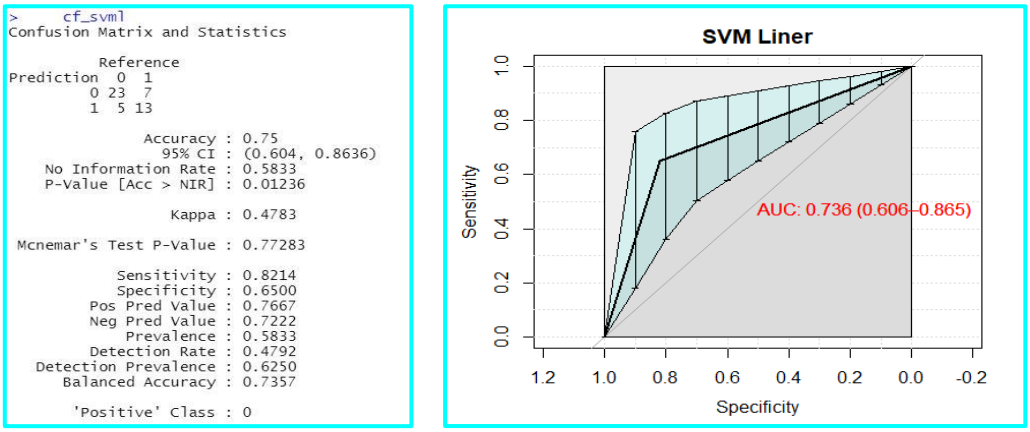

Figure 10: Result from SVM linear

5.7 Experiment 6: SVM with Radial Basis Function Kernel

5.7.1 Implementation

SVM with RBF kernel is recommended to use a non-linear data set, and the nonlinearity was extended by (Boser, et al., 1992). The implementation code is in CM 5.3 Fig 153 and 171.

5.7.2 Evaluation and Results

The key results are accuracy 0.7083, sensitivity 0.75, specificity 0.65, precision (positive predictive value) 0.75 and AUC (balanced accuracy) 0.70 in Figure 11. This is a good model.

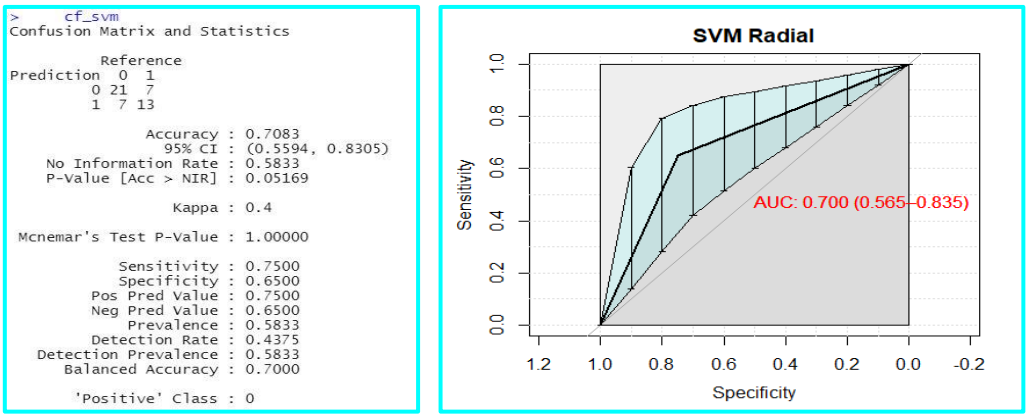

Figure 11: Result from SVM radial

5.8 Experiment 7: Random Forest

5.8.1 Implementation

Random forest is an algorithm used for classification, regression, and clustering. This is an ensemble learning algorithm that uses a decision tree as a learning device, and uses a large number of decision trees learned by randomly sampled training data (Zheng, et al., 2020). The implementation code is in CM 5.3 Fig 154 and 172.

5.8.2 Evaluation and Results

The key results are accuracy 0.6875, sensitivity 0.6786, specificity 0.7, precision 0.76 and AUC (balanced accuracy) 0.6893 in Figure 12. This is a satisfactory model.

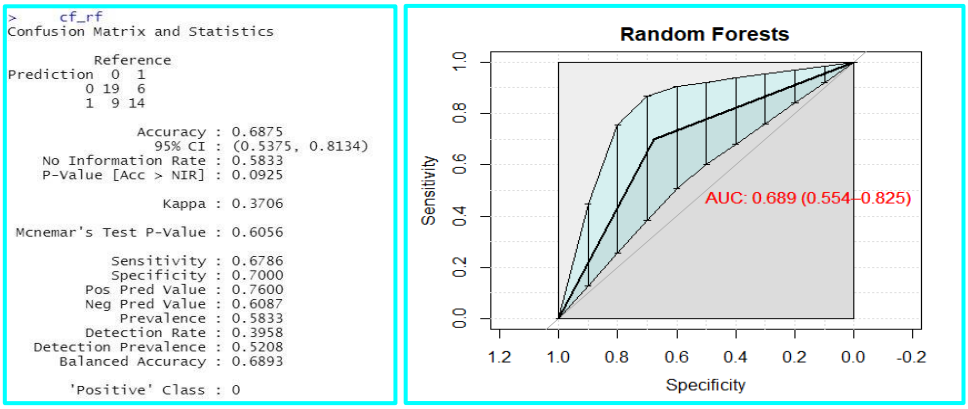

Figure 12: Result from random forest

5.9 Experiment 8: Neural Network

5.9.1 Implementation

A neural network is a mathematical model that imitates the neurons in the brain. It is used for pattern recognition, data classification, and future prediction because they can learn from data. A neural network consists of an input layer, one or more hidden layers, and an output layer (Abdullahi & Elkiran, 2017). The implementation code is in CM 5.3 Fig 155 and 172.

5.9.2 Evaluation and Results

The key results are accuracy 0.5833, sensitivity 0.7857, specificity 0.3, precision 0.6111 and AUC (balanced accuracy) 0.5429 in Figure 13. It is an unsatisfactory model.

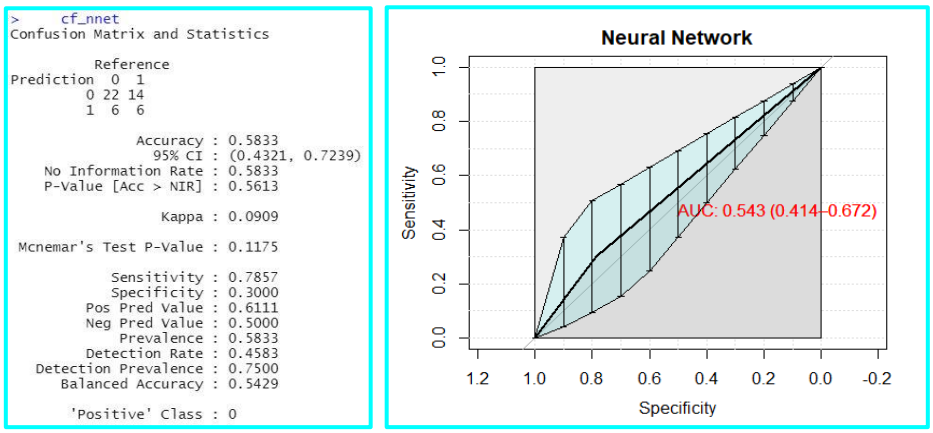

Figure 13: Result from neural network

5.10 Experiment 9: Gradient Boosting with Component-wise Linear Model

5.10.1 Implementation

Boosting is a meta-algorithm for performing supervised learning for regression and classification. It reduces bias and can convert weak learner to stronger (Bühlmann & Hothorn, 2007; Schonlau, 2005). The implementation code is in CM 5.3 Fig 156 and 173.

5.10.2 Evaluation and Results

The key results are accuracy 0.7292, sensitivity 0.8214, specificity 0.6, precision 0.7419 and AUC (balanced accuracy) 0.7107 in Figure 14. This is a good model.

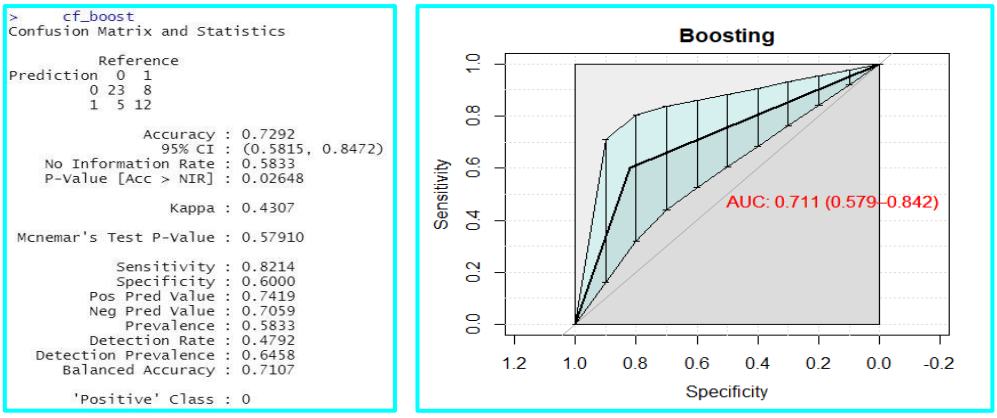

Figure 14: Result from boosting

5.11 Experiment 10: Bagging - Bootstrap Aggregation

5.11.1 Implementation

Bagging is a meta-algorithm that fits multiple models to different subsets of the training data set and combines the predicted values from all the models to make the prediction. It is also called an ensemble algorithm (Abdallah, et al., 2020). The implementation code is in CM 5.3 Fig 157 and 173.

5.11.2 Evaluation and Results

The key results are accuracy 0.6875, sensitivity 0.6786, specificity 0.70, precision 0.76 and AUC (balanced accuracy) 0.6893 in Figure 15. This is a satisfactory model.

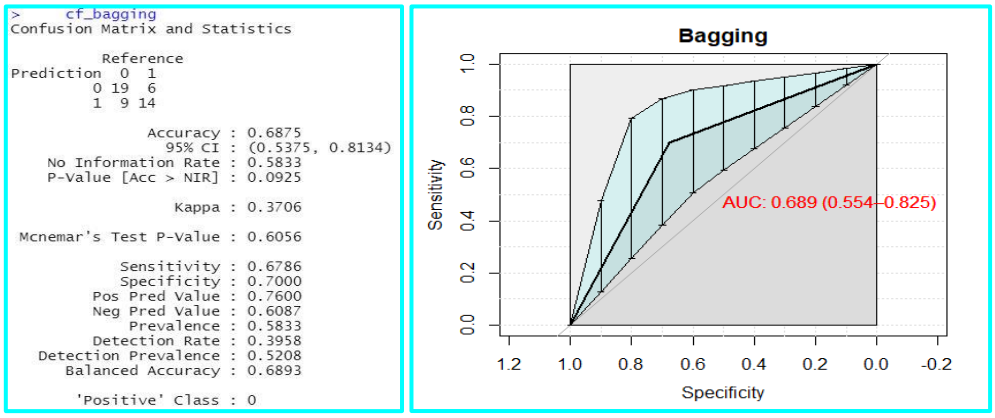

Figure 15: Result from bagging

5.12 Experiment 11: Naive Bayes

5.12.1 Implementation

Naive Bayes is a type of machine learning that probabilistically determines the category to which specific data belongs (Perez, et al., 2006). The implementation code is in CM 5.3 Fig 158 and 174.

5.12.2 Evaluation and Results

The key results are accuracy 0.7708, sensitivity 0.9643, specificity 0.50, precision 0.7297 and AUC (balanced accuracy) 0.7321 in Figure 16. This is a good model.

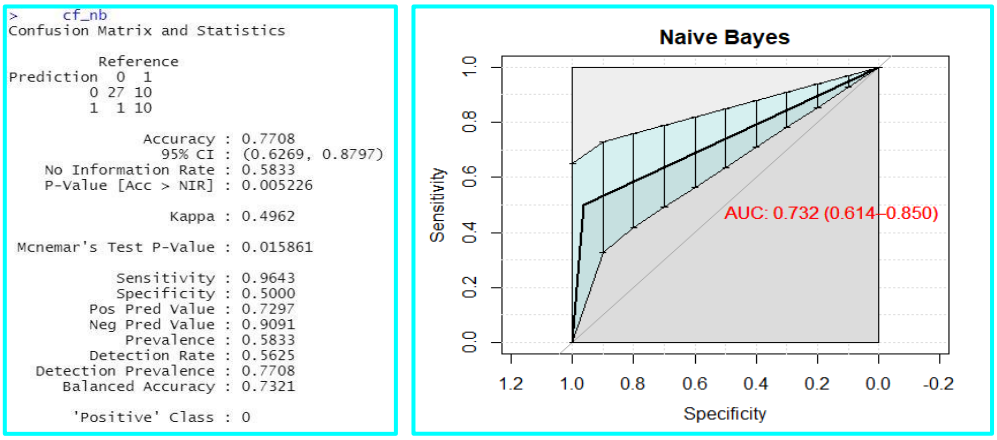

Figure 16: Result from Naive Bayes

5.13 Experiment 12: Conditional Inference Tree

5.13.1 Implementation

A conditional inference tree is a regression tree for a nonparametric type of data set and it embeds a tree-structured regression model with a well-defined theory (Hothorn, et al., 2006). The implementation code is in CM 5.3 Fig 159 and 174.

5.13.2 Evaluation and Results

The key results are accuracy 0.7083, sensitivity 0.5714, specificity 0.90, precision 0.8889 and AUC (balanced accuracy) 0.7357 in Figure 17. This is a good model.

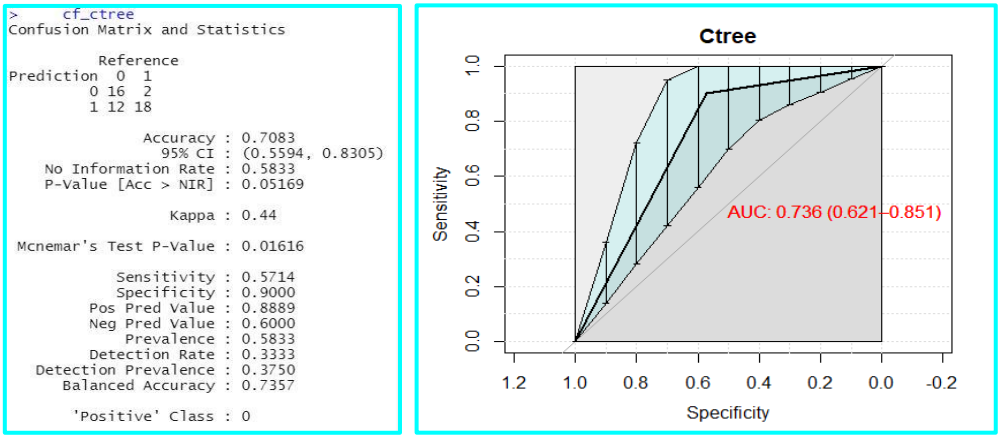

Figure 17: Result from Ctree

5.14 Results Comparison

5.14.1 Comparison in Data

Table 9 presents the summary of results of accuracy, AUC, sensitivity, specificity, and precision from 12 experiments (CM5.4). The highest values are indicated with blue in each category. The highest in each category is as follows: accuracy is GLM 0.7917. AUC is GLM 0.7786, sensitivity is Naïve Bayes 0.9643, specificity is Ctree 0.9, precision is Ctree 0.8899. Overall GLM has satisfied results as the highest result for accuracy and AUC, and high enough in sensitivity as 0.8571, specificity and precision are over 0.7. To make the comparison easier to see, in the next section, the results are visualized as a bar chart, a line chart, and an AUC-ROC curve.

Table 9: Result summary

	Model	Accuracy	AUC	Sensitivity	Specificity	Precision
1	**Generalized Linear Model (GLM)**	**0.7917**	**0.7786**	0.8571	0.7000	0.8000
2	Linear Discriminant Analysis (LDA)	0.7500	0.7357	0.8214	0.6500	0.7667
3	CART	0.7292	0.7393	0.6786	0.8000	0.8261
4	k-Nearest Neighbour (kNN)	0.5417	0.5143	0.6786	0.3500	0.5938
5	Support Vector Machine Linear (SVML)	0.7500	0.7357	0.8214	0.6500	0.7667
6	Support Vector Machine Radial (SVMR)	0.7083	0.7000	0.7500	0.6500	0.7500
7	Random Forest (RF)	0.6875	0.6893	0.6786	0.7000	0.7600
8	Neural network (Nnet)	0.5833	0.5429	0.7857	0.3000	0.6111
9	Boosting (Bst)	0.7292	0.7107	0.8214	0.6000	0.7419
10	Bagging (Bag)	0.6875	0.6893	0.6786	0.7000	0.7600
11	**Naive Bayes (NB)**	0.7708	0.7321	**0.9643**	0.5000	0.7297
12	**Ctree**	0.7083	0.7357	0.5714	**0.9000**	**0.8889**

5.14.2 Comparison by Bar Chart

Figure 18 presents the comparison of the results from table 9 in a bar chart (CM 5.4 Fig 160 to 166). The number in x-axis is experiment number. It compares the result among experiment on each category. There are dot lines at 0.7, as it is considered that over 0.7 is good level (Narkhede, 2018). The numbers that exceeded 0.7 are as follows: eight for accuracy, eight for AUC, seven for sensitivity, five for specificity, ten for precision.

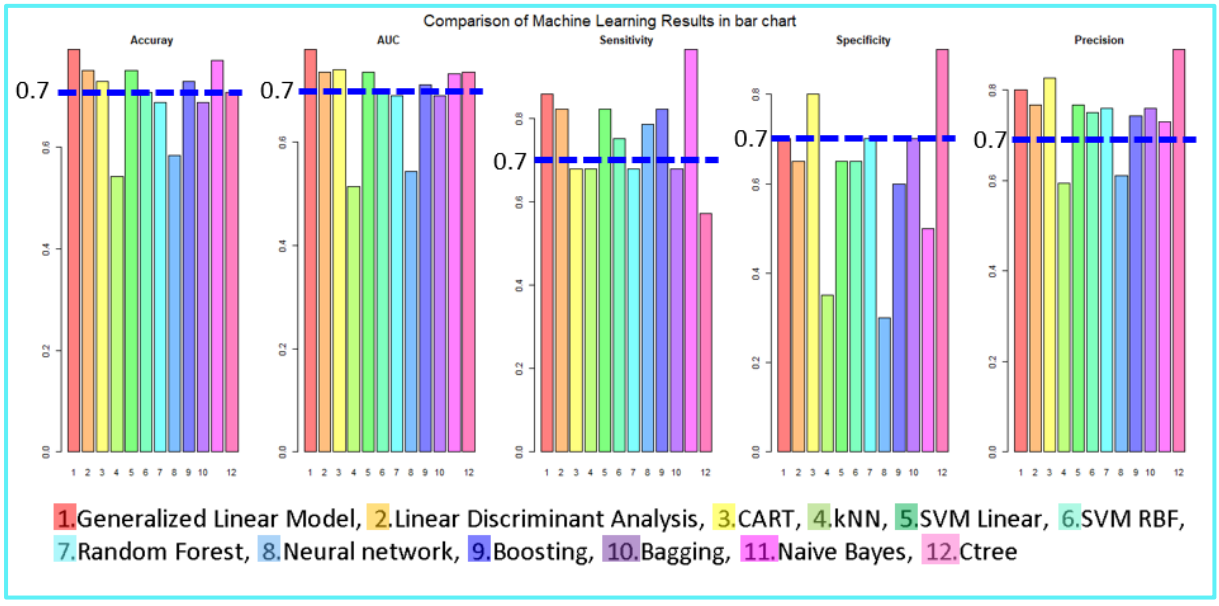

Figure 18: Comparison of the results in bar chart

5.14.3 Comparison by Line Chart

Figure 19 present the comparison of the results from table 9 in line chart (CM 5.4 Fig 167). The number in x-axis are experiments number. There is a dot line at 0.7. This graph presents the characteristics of each model and comparing the characteristics with the other models. Almost all the results of GLM are 70% or more. LDA, SVML, SVMR, Boosting, NB are showing the results of 70% or more excluding the specificity. kNN and Neural Network are all most all the results are less than 70%.

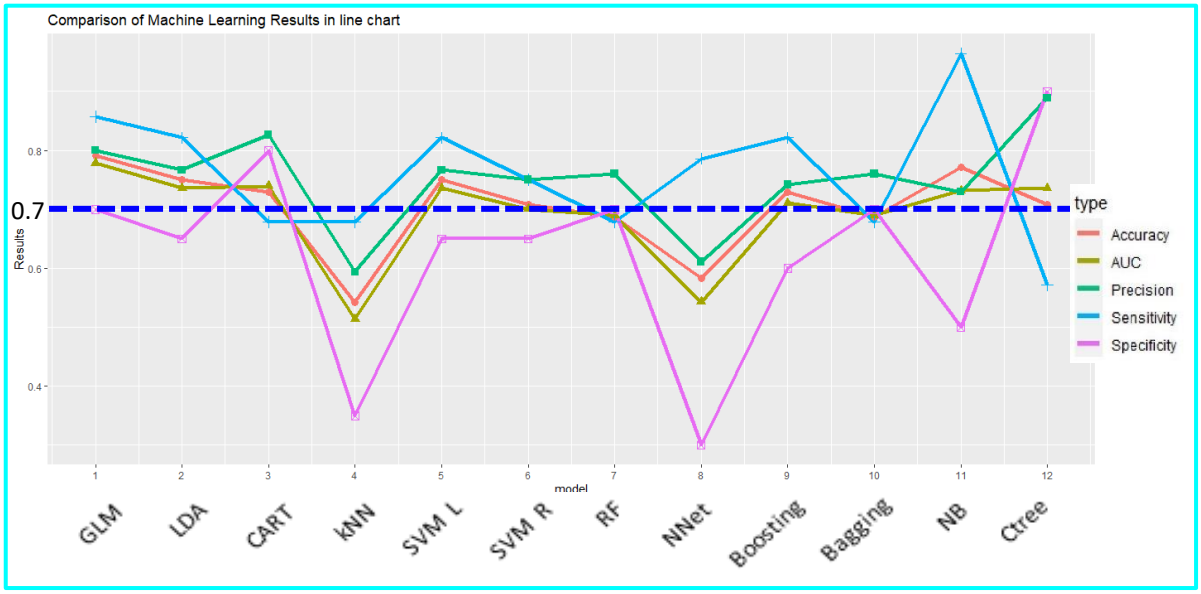

Figure 19: Comparison of the results in line charts

5.14.4 Comparison by AUC-ROC Curve graph

CM 5.4.1 Figure 175 and 176 present comparison of AUC-ROC Curve.

5.14.5 Comparison with Existed Model

Table 10 presents the details of proportion of the risk rank. Previous rank high (1) and low (0) indicate whether the GPI score is higher or lower than the average (CM5.1 Fig 143) from GPI 2019 existed model. Since the proportion of test data is 30%, the number of cases is 20.2 in high and 28.2 in low[12]. The results of 12 machine learning prediction models were collected from the confusion matrix (Fig 6 to Fig 17) with the true positive as high and the true negative as low. High pr, low pr, and outside[13] present the proportion in total. Outside indicates the misclassification ratio, which is the sum of the false positive and the false negative. Details of comparison will be presented in Figure 22 and 23.

[12] 69 * 0.3 = 20.7 (high), 94 * 0.3 = 28.2 (low)
[13] Value of outside = 1 – (high proportion + low proportion)

Table 10: Proportion of risk rank

Risk Rank	Previous Rank	12 Machine Learning prediction models											
		GLM	LDA	CART	kNN	SVML	SVMR	RF	Nnet	Bst	Bag	NB	Ctree
high 1	20.7	14	13	16	7	13	13	14	6	12	14	10	18
low 0	28.2	24	23	19	19	23	21	19	22	23	19	27	16
high pr	0.42	0.29	0.27	0.33	0.14	0.27	0.27	0.29	0.12	0.25	0.29	0.20	0.37
Low pr	0.58	0.49	0.47	0.39	0.39	0.47	0.43	0.39	0.45	0.47	0.39	0.55	0.33
outside	0.00	0.22	0.26	0.28	0.47	0.26	0.30	0.33	0.43	0.28	0.33	0.24	0.30

Figure 22 was created in Tableau and presents the proportion of rank high, low and misclassification by descending order by high, and comparing the results between previous rank which are previous data from GPI 2019 and the results of 12 machine learning prediction models. Previous rank shows 42.3% of high, 57.6% of low and no misclassification. Of the 12 machine learning prediction models, GLM had the lowest misclassification, 22.3%, followed by Naïve Bayes with 24.3%. The highest misclassification was kNN with 46.8%, followed by Neural Network with 42.7%.

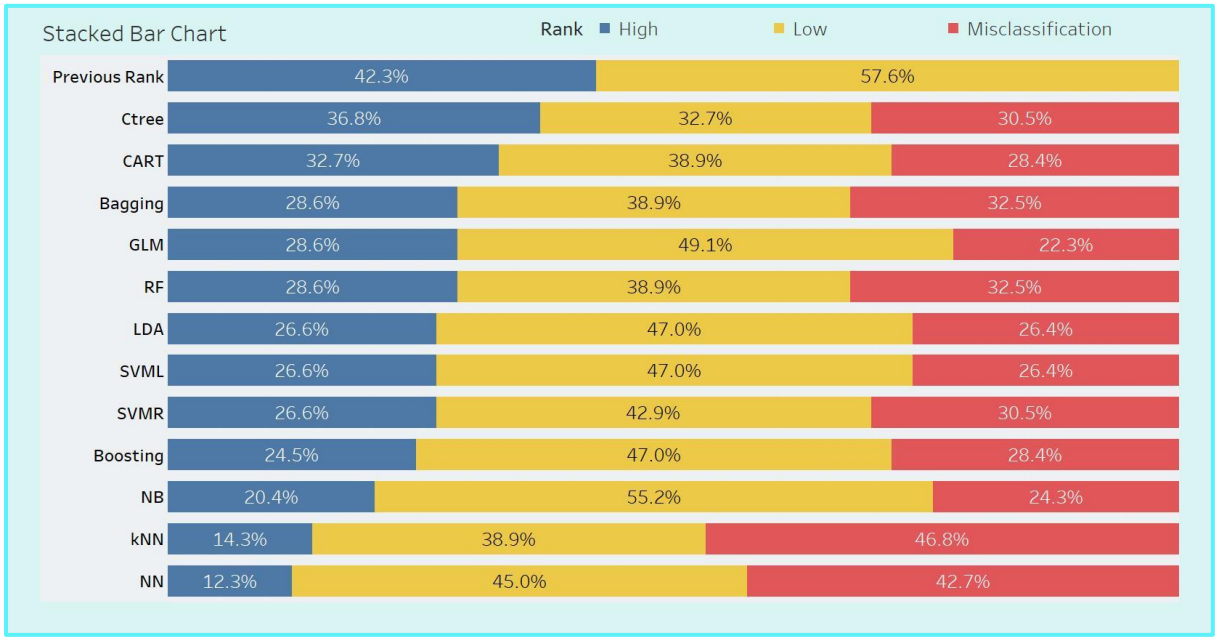

Figure 20: Proportion of risk rank -- previous reuslt vs results of 12 machine learning models

Figure 23 presents the comparison of the ranking of high, low and misclassification by Sankey charts (Riehmann, et al., 2005). It shows the rank flow from high to low and low to misclassification. As a sample, Ctree (indicated with light blue in Figure 23) is the second ranking in Risk-high in the left, and the last in the Risk-low in the middle, and 6th in the misclassification in the right. It is considered that lower in the misclassification is better in the accuracy. This point matches with the result in table 9 as GLM presents the highest accuracy followed by Naïve Bayes which are indicated in dark blue in Figure 23.

The creation detail of Tableau presentation is presented in CM 6.

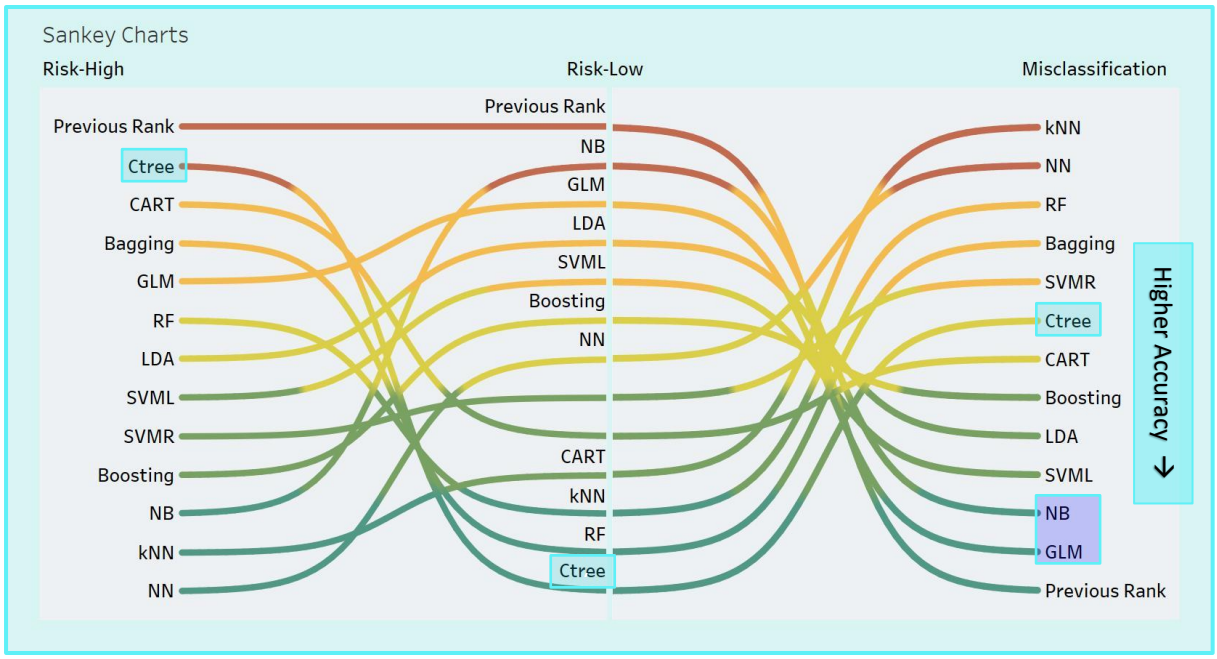

Figure 21: Ranking comparison of results by Sankey charts

5.15 Conclusion

In this chapter, the results of 12 machine learning prediction models were presented and evaluated. R was used for comparative evaluation using bar charts, line charts, and AUC-ROC curve graphs to search for answer to research queries based on the results of accuracy, sensitivity, specificity, precision, and AUC. Additionally, bar graphs and Sankey charts from the value of true positive and true negative were presented using Tableau.

6 Discussions

In this project, when the variables for implementation were selected by multivariate linear regression, unexpectedly result of high enough adjusted R-squared were not obtained. However, from the literature review in 4.2.6, it was validated to use selected variable for implementation. The reason why adjusted R-squared was not high enough was probably because there was a large bias in the distribution of natural disasters data. The reason for the poor distribution of natural disaster data is that some countries have few disasters or are in places where earthquakes occur frequently (near tectonic plate breaks or near volcanoes etc.). Therefore, it is difficult to compare countries. As for global warming information, more than half of the data are linearly regressing to the GPI Score, those were confirming the validity of using the data. Based on this evidence, it may be possible to improve the adjusted R-squared result by deleting information on natural disasters, increasing information on global warming, or grouping data by using PCA. As a result of searching the adequacy of using the data of global warming and natural disasters by 12 machine learning prediction models, the accuracy of GLM was close to 80%. Subsequently, the accuracy of the seven models which are LDA, CART, SVML, SVMR, boosting, Naïve Bayes and Ctree, were 70% or more. Regarding AUC, the results are similar to the results of accuracy. Regarding sensitivity, Naïve Bayes had the highest value of 96.4%,

followed by GML, LDA, SVML, and boosting of 80% or more, and SVMR and neural network of 70% or more. From these facts, it is possible to prove the adequacy of adding the results of using GLM, LDA, SVML or NB information on global warming and natural disasters to GPI. Furthermore, adding this information to GPI is an important method in understanding the safety of the country, and it can fully contribute to measures against global warming and natural disasters.

On the technical side, there were errors during the data transformation. Since data transformation was performed to improve the distribution of data, some methods[14] had errors and no result was obtained. Fortunately, there were several transform methods available, they could cover it. Regarding satisfied operations, weather information of data 2 could be automatically collected from 1630 websites by R, and data 4 and data 5 were done by batch file from Windows command prompt to run MySQL efficiently and it improved my programming skills. A Sankey chart was used in the presentation of the results, which may seem interesting, but it should be used more meaningful uses of rank comparison.

It is possible to continue this study by collecting data more regularly. The weather information has been available for over 100 years, and it is enough information for the research. Since there are only 10 types of factors in the current weather information in global warming data, it might be useful to incorporate pollution information such as UV index (Vanicek, et al., 1999) and PM2.5 (Zheng, et al., 2005), which can be considered to be rated to global warming.

7 Conclusion and Future Work

According to the literature review, existing GPI factors consisted of three factors: safety and security, ongoing conflict, and militarization. In this project, the adequacy of incorporating factors of global warming and natural disasters into GPI was investigated and it was researched whether this information could contribute to various problems caused by current global warming and natural disasters. As experiments 12 machine learning prediction models: GLM, LDA, CART, kNN, SVML, SVMR, random forest, neural network, boosting, bagging, Naïve Bayes and Ctree, were used as the objectives, and these verified how this information can contribute to the problems of global warming and natural disasters. From this project that have been conducted up to this point, the following research question "*Can prediction of factors (weather, CO2 emissions, death toll by natural disasters and Covid 19) that contribute to global warming and natural disasters provide insights into improving safety of countries and reduce the problem of global warming and natural disasters?*" and sub research question *"Can identification of factors contributing to global warming and natural disasters be able to give significant impact in the ranking of safety country in the current GPI?"* can be answered. In response to the research questions and sub research question, 28 variables were collected from the following five data sources: GPI, weather information, CO2 emissions, natural disasters and Covid 19. Then, by variables selection, following eight suitable variables were selected: cloud cover, diurnal temperature range, vapour pressure, CO2 emissions increase rate 2017/1970, drought, volcanic activity, wildfire and Covid 19 infected case. As a result of

[14] natural logarithm, natural logarithm 10, square root ,1/x transformation

performing 12 machine learning prediction models using the selected variables, it was found that GLM was the most appropriate model by comparing accuracy, AUC, sensitivity, specificity, and precision. It was found that 8 out of 12 machine learning prediction models were reliable models because of accuracy and AUC of 0.7 or higher. From this, it was possible to answer the research question that it is possible to make effective predictions using this information on global warming and natural disasters. From this project, it is recommended to add a new factor to GPI and review the global safety ranking. After understanding its importance, each country and individual level should focus more strongly on preventing global warming and protecting the future of the earth.

As a future task, it is recommended to try to limit the data of natural disasters to those related to global warming, and to use data that emphasizes global warming data in order to improve the adjusted R-squared value of model selection. Since the country name display differs for each data and it took time to adjust it, it would be more efficient to create a mapping table in the future. On the technical side, it is recommended to review outlier settings and AIC usage. In MySQL, it is recommended to use MySQL Benchmark as it makes it easier to see the code and process and tests can be improved.

Acknowledgements

I would like to thank to Dr. Catherine Mulwa, our supervisor for her useful lecture and detailed advice for this research project during this very hard pandemic circumstances.

References

Abdallah, I., Tatsis, K. & Chatzi, E., 2020. Unsupervised local cluster-weighted bootstrap aggregating the output from multiple stochastic simulators. *Reliability Engineering & System Safety*, Volume 199, p. 106876.

Abdullahi, J. & Elkiran, G., 2017. Prediction of the future impact of climate change on reference evapotranspiration in Cyprus using artificial neural network. *Procedia Computer Science,* Volume 120, pp. 276-283.

Boser, B. E., Guyon, I. M. & Vapnik, V. N., 1992. *A Training Algorithm for Optimal Margin Classifiers,* California: AT&T Bell Laboratories.

Bretl, B. et al., 1999. Persistent Java Objects in3 tier architectures. In: R. Morrison, M. Jordan & M. Atkinson, eds. *Advances in Persistent Object Systems: Proceedings of the Int'l Workshop on Persistent Object Systems (POS) and the Int'l Workshop on Persistence & Java (PJAVA).* California: Morgan Kaufmann, 1999, pp. 236-249.

Bühlmann, P. & Hothorn, T., 2007. Boosting Algorithms: Regularization,. *Statistical Science,* 22(4), p. 477–505.

Chapman, P., 1998. *The CRISP-DM User Guide,* Copenhagen: NCR Systems Engineering.

Clements, P. K. P., 2019. GPI Methdology. In: t. I. f. E. &. Peace, ed. *Global Peace Index 2019.* Otag, New Zealand: the Institute for Economics and Peace the Institute for Economics & Peace;, pp. 84-87.

Council, N. R., 1999. *The Impacts of Natural Disasters: A Framework for Loss Estimation.* illustrated ed. s.l.:National Academies Press.

F.Luhr, J., 2013. Tectonic Earth -- The Earth's Plates. In: *EARTH THE DEFINITIVE VISUAL GUIDE.* London, Great Britain: Dorling Kindersley Limited, pp. 488-489.

Fekete, A., 2018. Societal resilience indicator assessment using demographic and infrastructure data at the case of Germany in context to multiple disaster risks. *International Journal of Disaster Risk Reduction,* 31(1), pp. 203-211.

Ferreira, J. E. V., Costa, C. H. S. d., Miranda, R. M. d. & Figueiredo, A. F. d., 2015. The use of the k nearest neighbor method to classify the representative elements. *Educación Química,* 26(3), pp. 195-201.

Filkov, A. I., Duff, T. J. & Penman, T. D., 2019. Frequency of Dynamic Fire Behaviours in Australian Forest Environments. *Fire 2020 -- Technical Note,* 3(1), p. 10.3390/fire3010001.

Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing* , Volume 21, pp. 137-146.

Goel, A. & Bhatt, R., 2012. CAUSES AND CONSEQUENCES OF GLOBAL WARMING. *International Journal of Life Science Biotechnology and Pharma Research,* 1(1), pp. 27-31.

Greig, R., 2017. ROCKS, GEMS & FOSSILS. In: *HOW IT WORKS BOOK OF INCREDIBLE EARTH.* Sorset, Great Britain: Future publishing Ltd, pp. 112-123.

Hák, T., Janoušková, S. & Moldan, B., 2016. Sustainable Development Goals: A need for relevant indicators. *Ecological Indicators,* 60(1), pp. 565-573.

Hoeppe, P., 2016. Trends in weather related disasters – Consequences for insurers and society. *Weather and Climate Extremes,* 11(1), pp. 70-79.

Hothorn, T., Hornik, K. & Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework ctree: Conditional Inference Trees. *Journal of Computational and Graphical Statistics,* 15(3), p. 651–674. .

IEP, 2019. *Global Peace Index 2019: Measuring Peace in a Complex World.* 1 ed. Sydney: Institute for Economics & Peace.

IEP, 2019. GPI indicator sources, definitions & scoring criteria. In: T. I. f. E. &. Peace, ed. *Global Peace Index 2019.* Sydney: The Institute for Economics & Peace;, pp. 88-95.

K.Ardakani, M. & Seyedaliakbar, S. M., 2019. Impact of energy consumption and economic growth on CO2 emission using multivariate regression. In: M. K.Ardakani & S. M. Seyedaliakbar, eds. *Energy Strategy Reviews.* Florida: Florida Polytechnic University, p. 100428.

Leng, L., 2007. Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. *Journal of Physics: Conference Series,* 78(01).

Letcher, T. M., 2019. Why do we have global warming?. In: *Managing Global Warming -- An Interface of Technology and Human Issues.* Somerset, UK: Elsevier Inc, pp. 3-15.

Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine,* 27(2), p. https://doi.org/10.1093/jtm/taaa021.

Loh, W., 2011. Classification and regression trees. *WIREs DATA MINING AND KNOWLEDGE DISCOVERY,* 1(1), pp. 14-23.

Lukacs, P., Burnham, K. & Anderson, D., 2009. Model selection bias and Freedman's paradox.. *Annals of the Institute of Statistical Mathematics,* Volume 117, p. 62.

Marsh & McLennan and Zurich Insurance Group, 2020. *The Global Risks Report 2020 Insight Report 15th Edition,* Geneva: World Economic Forum.

Narkhede, S., 2018. *Understanding AUC - ROC Curve, towards data science.* [Online] Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Accessed 14 July 2020].

NATIONS, U., 2015. *PARIS AGREEMENT.* Paris, UNITED NATIONS.

Nelder, J. A. & Wedderburn, R. W. M., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General),* 135(3), pp. 370-384.

Neter, C. J. et al., 2004. *Applied Linear Regression Models.* 5th International edition ed. s.l.:McGraw-Hill Education; .

Oberthür, S. & Ott, H. E., 2013. *The Kyoto Protocol: International Climate Policy for the 21st Century International and European Environmental Policy Series.* illustrated ed. Berlin: Springer Science & Business Media, 2013.

Packham, C., 2017. Plate Tectonics. In: *NATURAL WONDERS OF THE WORLD.* London, Great Britain: Dorling Kindersley Limited, pp. 12-13.

Perez, A., Larranaga, P. & Inza, I., 2006. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning,* Volume 43, pp. 1-25.

Portet, S., 2020. A primer on model selection using the Akaike Information Criterion. *Infectious Disease Modelling,* Volume 5, pp. 111-128.

Rogelj, J., den Elzen, M. & H. N., 2016. Paris Agreement climate proposals need a boost to keep warming well below 2 °C. *Nature 534,* Volume https://doi.org/10.1038/nature18307, pp. 631-639.

Russell, S. J. & Norvig, P., 2016. Chapter 18: Learning from Examples. In: *Artificial Intelligence: A Modern Approach (3rd Edition).* Essex: Pearson Education Limited, pp. 693-767.

Sands, P., 1992. The United Nations Framework Convention on Climate Change. *RECIEL Review of European, Comparative & International Environmental Law,* 1(3), pp. 270-277.

Schonlau, M., 2005. Boosted regression (boosting): An introductory. *The Stata Journal,* 5(3), p. 330–354.

SIRSAT, M., 2019. *Data Science and Machine Learning.* [Online]
Available at: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html
[Accessed 14 July 2020].

Streich, J. et al., 2020. Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals?. *Current Opinion in Biotechnology,* Volume 61, pp. 217-225.

Suss, J. & Treitel, H., 2019. *Staff Working Paper No. 831 Predicting bank distress in the UK with machine learning,* London: Bnak of England.

Vapnik, V. & Lerner, A., 1963. Pattern recognition using generalized portrait method. *Automation and Remote Control,* p. 24.

Weart, S. R., 2008. *The Discovery of Global Warming.* revised ed. s.l.:Harvard University Press.

WHO, 2020. *Coronavirus disease 2019 (COVID-19) Situation Report – 67,* Geneva: WHO.

Wirth, R. & Hipp, J., 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* . London, UK: Springer-Verlag, pp. 29-39.

Yanagihara, H., 2006. Corrected version of AIC for selecting multivariate normal linear regression models in a general nonnormal case. *Journal of Multivariate Analysis,* 97(5), pp. 1070-1089.

Yuan, G.-X., Ho, C.-H. & Lin, C.-J., 2012. Recent Advances of Large-Scale Linear Classification. *Proceedings of the IEEE,* 100(9), pp. 2584 - 2603.

Zhang, W., Villarini, G., Vecchi, G. A. & Smith, J. A., 2018. Urbanization exacerbated the rainfall and flooding caused by hurricane Harvey in Houston. In: *Nature.* s.l.:Springer Nature, pp. 384-388.

Zhaoab, Y., Li, Y., Zhang, L. & Wang, Q., 2016. Groundwater level prediction of landslide based on classification and regression tree. *Geodesy and Geodynamics,* 7(5), pp. 348-355.

Zheng, C., Chen, C., Chen, Y. & Ong, S. P., 2020. Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure. *Patterns,* 1(2), p. 100013.