Heriot-Watt University

Dissertation

Machine Learning based Analysis of Greenhouse Gases in Middle East

Author:
Usman Qureshi

Supervisor:
Smitha S Kumar

A dissertation submitted in fulfilment of the requirements for the degree of BSc.

in the

School of Mathematical and Computer Sciences

April 2021

Declaration of Authorship

I, Usman Qureshi, declare that this thesis titled, 'Machine Learning based Analysis of Greenhouse Gases in Middle East' and the work presented in it is my own. I confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed: _____

Date: _____

Abstract

Greenhouse gas emission in the world is increasing in a continuous way and has environmental and health effects with humans being the main factor contributing to the disruption of the scale of the Greenhouse gases. This project attempts to extract useful, hidden knowledge and to show meaningful data. There were several researches conducted on different datasets of the world on different countries in order to get deeper understanding of the Dataset. But no such research was conducted on the Middle Eastern Dataset.

The project experiments with multiple Machine Learning algorithms for analyzing the middle eastern GHG data from the year 1990-2016. Three machine learning algorithms are used namely Linear Regression, Support Vector Machines and K-means Clustering for evaluating the best algorithm through their accuracy. The paper looks at different papers which test algorithms in order to find the best machine learning algorithm.

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor :)

Contents

Chapter 1

Introduction

Data mining of Greenhouse Gases(GHG) is really important for the society as it helps to get a better understanding of the environment and to find solution for the problems caused by it, since, for decades, there has been a sharp increase in Greenhouse Gases(GHG) throughout the world and with it, the global average temperatures ha increased by more than 1° since pre-industrial times(Ritchie and Roser, 2017). Energy consumption, environmental degradation, and climate change are all closely related. To tackle such problems, this dissertation first understands the concentration of greenhouse gases in the different regions of the world and finds out which sector contributes most to the increase in GHG's.

For this project we are going to be implementing Data Mining and Machine Learning algorithms on the Middle Eastern dataset. For this dissertation, Middle Eastern dataset was chosen because there hasn't been much research conducted on this particular dataset.

1.1 Aim

The aim of the project is to use the Middle Eastern GGE's dataset and to extract useful, hidden knowledge and show meaningful data

1.2 Objective

The main objective of this dissertation is to get information from the dataset from the year 1990 till 2016 and accomplish the following tasks:

- Run multiple Machine Learning algorithms to find out which algorithm gives the best result.
- Data set analysis and preprocessing
- Creating a machine learning model to predict which sector contributes to the highest emission.
- Creating a machine learning model to predict which sector contributes to the highest emission.
- Design a web application to visualize and interact with the data.
- Design a web application to visualize and interact with the data.

1.3 Document Overview

The document is organized in such a way that the flow of the document is maintained, with that the **Literature Review** is first covering the background i.e., research into the data mining and machine learning algorithms on GGE datasets. Next, **Requirements Analysis**, which is going to cover requirements and outcomes of the project aim and objective. After that, **Design** which covers the design of the system. Following that is **Evaluation Strategy** to ensure the completion of specification requirement. Finally, **Project Management** discusses the plan of the system lifecycle and challenges.

Chapter 2

Literature Review

In this part of the chapter we are going to research the work done regarding the topic of the project. Each section covers different parts of the research by the authors.

2.1 Background

2.1.1 Greenhouse Gases (GHGs)

Human emissions of CO2 and other GHGs are a primary cause of climate change and present one of the world's most persistent challenges. This link between global temperatures and greenhouse gas concentrations, especially CO2, has been there throughout Earth's history [15]

The major greenhouse gases are Carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), hydrofluorocarbons (HFCs) and Sulphur hexafluoride (SF6). The most destructing gases of them all is Carbon Dioxide.

A changing climate has the potential to damage the ecological, physical and health, including weather events (such as floods, droughts, and heatwaves); sea-level rise; and altered crop growth. The most extensive source of analysis on the possible impacts of climatic change can be found in the 5th Intergovernmental Panel on Climate Change (IPCC) report [15].

CO2s and other GHGs go through a natural cycle as shown in Figure 1, large amounts of carbon travel back and forth first to the atmosphere and then towards the earth's surface. These processes have the tendency to keep the amount of CO2 moderately constant over time [16].

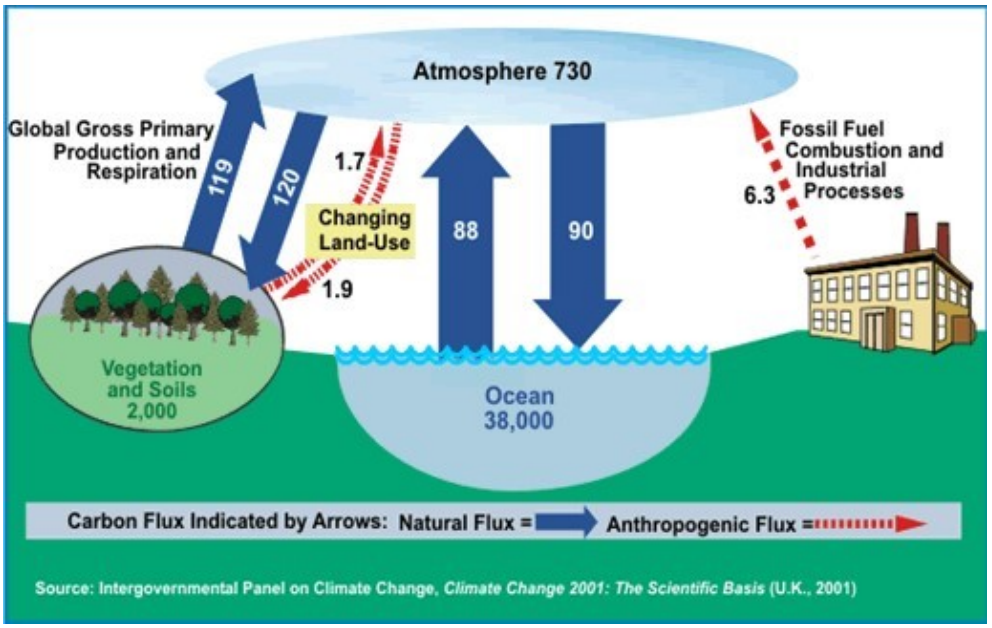

FIGURE 1: Carbon Cycle [17]

To get perspective as to where to even start reducing the GGE emission, the author is going to find out the sectors (such as industrial, agriculture, Fuel Combustion etc.) and countries responsible for this increase in GGEs.

2.1.2 Data Mining and Machine Learning

Data mining is the process of searching large stores of data to look for patterns that wouldn't be possible through simple analysis. Data mining uses mathematical algorithms to section the data and assess the chances of future events. Data mining is also known as Knowledge Discovery in Data (KDD)[2]

Data Mining combines tools from statistics and artificial intelligence (such as machine learning) to analyze large data sets. Data mining is broadly used in business, science research, and government security [7]

An array of tools and techniques are required in order to get the best result from the data. Some of the most commonly used functions include:

- Data cleansing and preparation- A step in which inaccurate or corrupt data is detected and removed. [6]
- Machine learning- It is a process where the machine acquires the skills and knowledge by identifying and using preexisting knowledge [19]
- Regression— A method used to predict a range of numeric values, for example sales, temperatures etc, using machine learning based on a data set. [3]

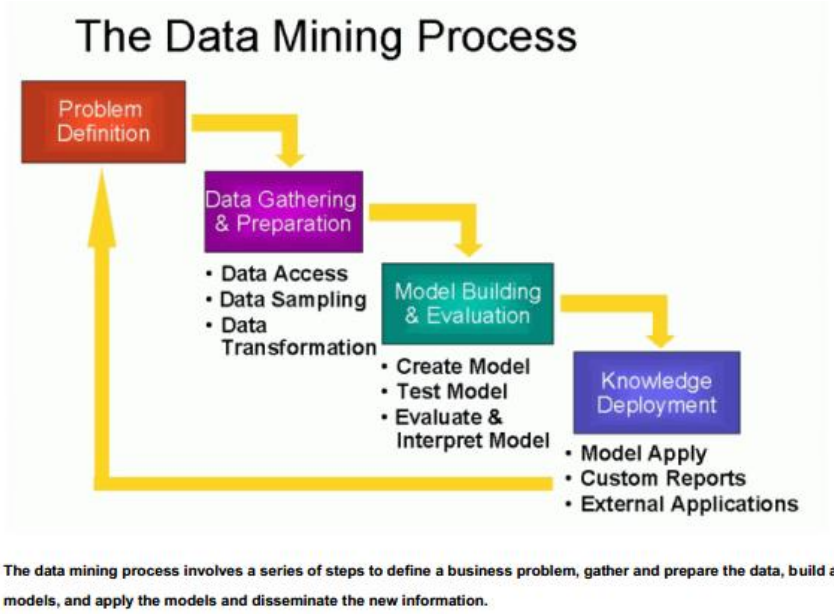

FIGURE 2: The Data Mining Process [14]

Machine learning is letting the computers to learn and consequently act the way humans do, improving its learning and knowledge over time autonomously. [4]

In Machine Learning there are three general categories, but for this project we are going to look into two:

Supervised machine learning can be used to make predictions about hidden or future data called predictive modeling. The algorithm develops functions that accurately predicts the output from input variables, such as predicting the market value of a home (output) from the square foot (input) and other inputs (type of construction, etc.). [4]

Two types of supervised learning are:

- **Classification**
- **Regression**

Supervised Learning Algorithm consists of k-Nearest Neighbor (kNN), Linear regression, Naive Bayes, support vector machine (SVM), logistic regression and gradient boosting.

In **unsupervised machine learning**, the algorithm is left on its own to find its input on its own. Discovering hidden patterns in data or a means to an end can be a goal in itself. This is also known as "feature learning [4]

For our project we will be using three Supervised machine learning algorithm:

- **Linear Regression**: Linear Regression is used to identify the relationship between the variables. Linear regression is either simple linear regression or multiple linear regression. The equation $y=x\beta+\epsilon$, is used to describe linear regression, where y is the independent variable and x is the dependent variable which is a continuous value. It is focused mainly upon the conditional probability distribution with analysis [11]

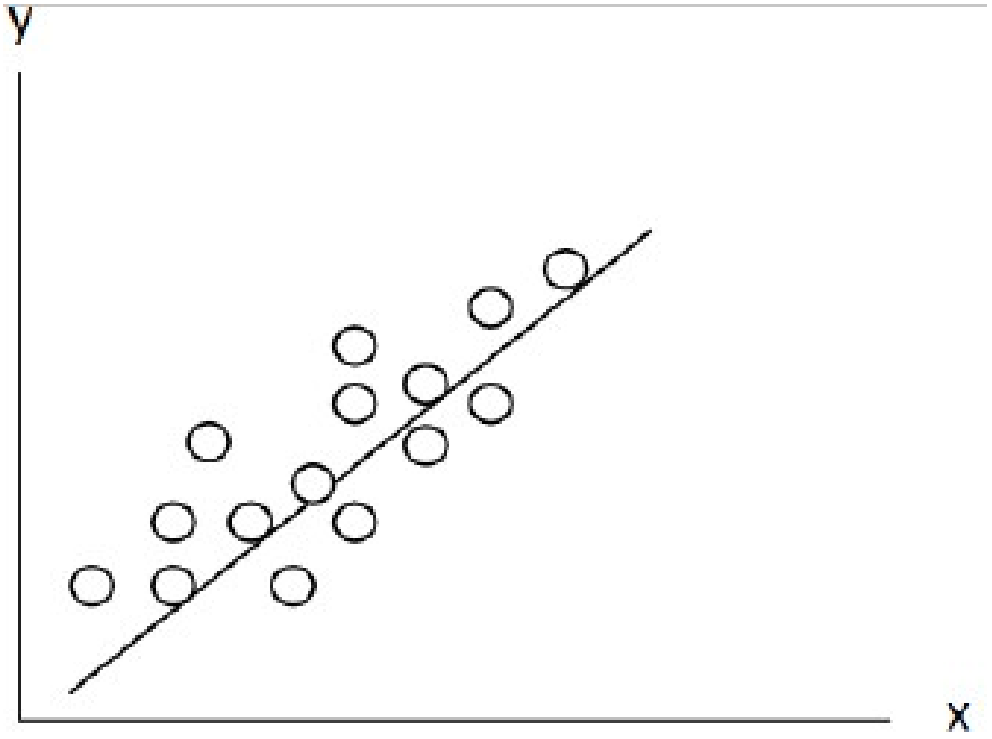

FIGURE 3: Linear Regression [11]

- **Support Vector Machine**: The objective of support vector machine is to find out the best hyperplane in more than two dimensions in order to find out how to separate the space into classes. The hyperplane is derived from the maximum margin i.e., the maximum distance between data points of both classes [13]

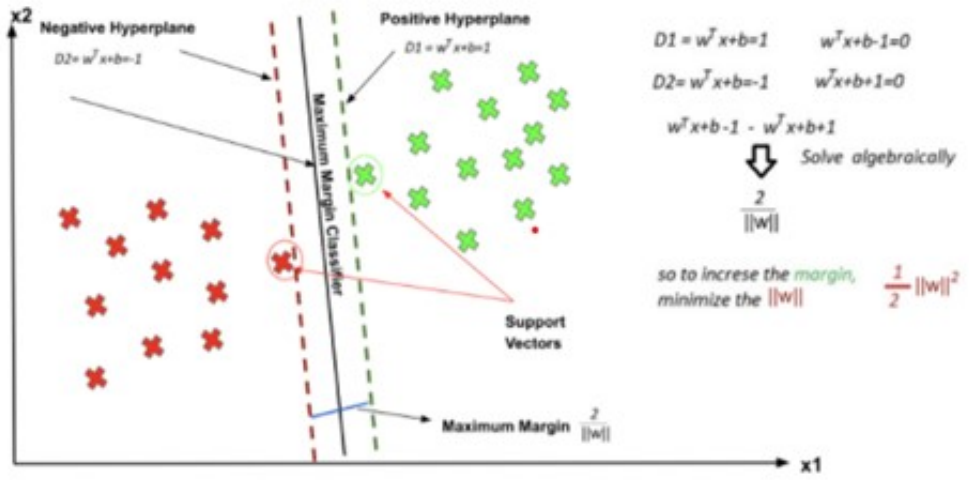

FIGURE 4: Support Vector Machine [13]

- **Decision Trees**: The Decision Tree is classified as a predictive model which is charting from observations form a dataset to conclude its target value. The leaves in a decision tree structure signify labels and the non-leaf parts are the features, and the branches signify conjunctions of the features that lead to the classification [18]

2.2 Related Works

2.2.1 Spatio-Temporal Analysis of Big Atmospheric Data

Cuzzocrea et al.[8] proposes an approach for supporting clustering-based analysis of big atmospheric data. In this paper the authors have researched the GHG's of the three European countries, namely United Kingdom, France and Italy. The authors had applied K-means clustering to the atmospheric big data. The aim of this paper was to produce a big data mining model which was used to analyze environmental data providing deeper understanding into high and low emitting sectors of GGE.

To achieve this, the author defined data workflow models which was used to analyze the environmental data. GGE's of several sectors were analyzed in order to find out where efforts and changes to reduce the emissions could be carried out. Then compared the emissions of the UK initially with the other countries. Also, over time, the GGE's of each country were investigated; in order to achieve that, a deeper insight was provided into the high and low emitting sectors [8].

In their paper, Cross-industry standard process for data mining (CRISP-DM) methodology, which is a standard methodology, which was in turn used to carry out data mining projects, and Weka Knowledge Flow, which is a graphical tool used to express the whole data mining process was used by the authors. In CRISP-DM, it consists of the six stages it is as shown in Figure 5. This methodology follows a cyclic pattern, and to fully meet the input requirements, the flexibility to return to previous stages from certain points ensures that completely[8].

Critical Review: The approach towards the issue was very well designed and its use of the CRISP-DM methodology is well balanced and quite flexible. The data used for this paper by the authors was not too large, as it consists of information of only 3 countries, therefore it's not clear how this will handle huge datasets. The author decided to only use K-Means Clustering which may or may not be the best algorithm for this dataset, more algorithms should have been used to get better accuracy.

FIGURE 5: CRISP-DM Process [8]

2.2.2 Prediction CO2 Emissions using Data Mining

Kunda et al. [12] analyzes the policy provision in Zambia regarding carbon emissions. This paper further provides a time series analysis of CO2 emission from 1964 to 2016 and makes a forecast for the year 2021 using Weka Knowledge Flow as the data mining tool. The dataset selected for this was for the country of Zambia and its transportation, manufacturing and construction information. The aim of this paper was to find out what has been the change in percentage contribution to carbon emission based on total fossil fuel combustion annually from different sectors and also to find out which sector had the highest contribution of emissions and to make predictions for the following five years on the basis of carbon emissions of each sector. And also, with this information, the authors hope that the policy makers introduce policies which will in turn regulate CO2 emissions.

Data mining was used to predict and analyze the data. To analyze this data, the authors used SMOreg algorithm for time series.

Results of SMOreg Algorithm	
Manufacturing and Construction	50.9%
Transport	31.7%
Electricity & Heat production	6.7%
Residential & Commercial	7.1%

After analyzing the data, the future predictions from the year 2017-2021 shows a continuous increase in CO2 emissions of transport and manufacturing. [12]

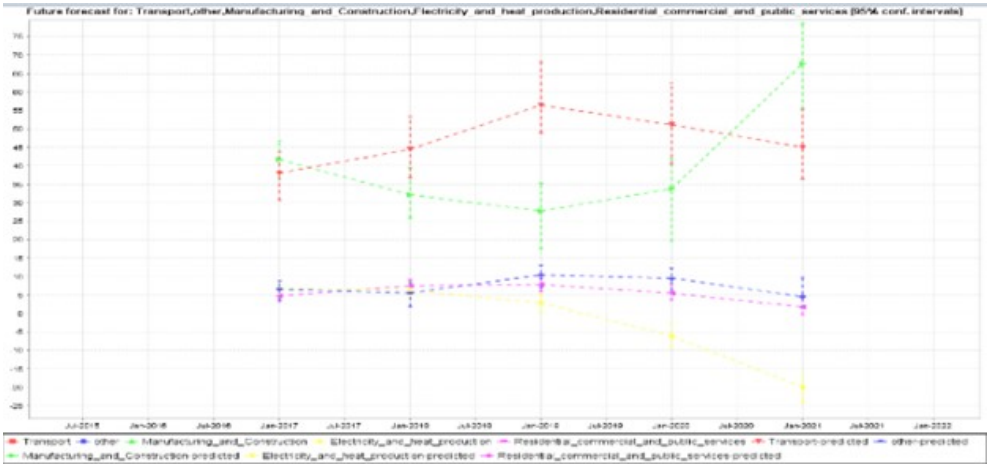

FIGURE 6: : Zambia CO2 forecast [12]

Critical Review: The implementation of the algorithm appears satisfactorily accurate and gives the insight of the future forecast for the next five years . This paper had some drawbacks, the dataset used was only for one specific country and hence like the previous paper it fails to elaborate information regarding huge datasets.

2.2.3 Global Warming Prediction in India using Machine Learning

Hema et al. [9] analyzed global warming through global temperature and GHG gases in India using the data set from 100-150 years timeline. Their objective was to forecast the temperature and the GHG gases for the next 10 years and to create a graphical interface based on the results for easy understanding. The data was first collected and then later preprocessed. Linear Regression algorithm was selected for the paper after running tests on multiple algorithms such as Multiple Regression and Support Vector Regression on the preprocessed data. Linear Regression gets more prediction accuracy in comparison to the other algorithms (Hema, Pal, Loyer and Gaurav, 2019). Pleasing accuracy results are obtained for the temperature as well as the GHG gases prediction after training of the temperature and the GHG gases data. The prediction was done by

dividing the data for training and testing, an object is created to predict the test value. The object is used to predict the data for the next 10 years. After prediction, forecasted data for the next 10 years and graphical representation of those predicted and forecasted data is presented [9].

Critical Review: The approach seems to be detailed and well-presented. The author was able to predict and forecast the temperature for the next few years. The author could have used some more algorithms for comparison and selection. The model predicts only the average of temperature and the GHG concentration from the data set of approximate 150 years, the prediction could be far more accurate if the author would use much larger data set.

2.2.4 Relation Between Global Temperature and Concentrations of Greenhouse Gases

Kalra et al. [10] analyses two different datasets, one for global temperature and another for GHG concentrations. The author finds and models the relationship between the two different datasets of equivalent to 65 years using Linear Regression, Decision trees, Random Forrest and Artificial Neural Network. These experiments were executed using the Keras Library for ANN and the Sklearn library for the rest of the algorithms. The independent variables were the GHG gases and the dependent variable was the global temperatures.

For the Linear Regression and Decision Trees the authors had used the default hyper-parameters. For the Random Forrest Algorithm, the authors observed that using thirty trees yielded the least MSE. For the Artificial Neural Network, three layers with three input nodes, one output node and two hidden nodes were used. The authors thought that using Mean Square Error loss function would be better as it is an excellent estimator of how well the network models the given data [10]

From the experiments conducted by the authors, they found out that the Artificial Neural Network performs in a way that is better than the other models performed on the same dataset.

Critical Review: The approach of using different types of algorithms on large datasets is well experimented upon which in turn helps to decide the final algorithm to run the

Comparison based on Mean Squared Error

Algorithm Used	MSE
Decision Tree Regression	0.0174
Linear Regression	0.0152
Random Forest Regression	0.0095
Artificial Neuron Network	0.0078

FIGURE 7: Comparison of Different Algorithms [10]

experiments on. Authors chose the best algorithm on the basis of mean square error which measures the squares of the errors. From the final algorithm chosen i.e. ANN, they were also able to calculate feature importance from which they were able to draw more conclusions.

2.2.5 Analysis of Global Warming Using Machine Learning

Zheng [20] compares the performance of several machine learning algorithms on the data. The algorithms used for the experiment included Linear Regression, Support Vector Regression and Random Forrest, these were used to build models which uses concentrations of different Greenhouse Gases to precisely predict the global atmosphere. The data chosen by the author was vast, it was from the past 800,000 years. This data was aligned using linear interpolation because the data was vast and there might have been missing values. For the three different machine learning algorithms, the parameters were altered to fit the data and produce accurate training results.

To measure the accuracy of the models, Mean Square Error (MSE) was used. The training and testing of the different algorithms show that the Random Forest creates the most accurate models as shown in Figure 8. Through machine learning the author confirms that the CO2 is the biggest contributor to temperature change [20]

Critical Review: The author uses a vast data source to predict the change in temperature which in turn gives far more accurate results. The implementation and proving of a point i.e. CO2 is the biggest contributor to temperature change was executed successfully. The author chose few machine learning algorithms for the experiment, however, more proven and accurate algorithms from the previous papers could be used such as XGBoost or Artificial neural Network which can produce better models.

Algorithm	Training MSE	Testing MSE
Random Forest	0.1289	0.9557
Lasso	1.8819	2.2088
SVR	1.3740	1.5267
Linear Regression	1.8796	2.2689

FIGURE 8: MSE of Each Algorithms [20]

2.3 Research Gaps

There hasn't been much research conducted on the Middle Eastern countries.

2.4 Challenges

The challenge of this project is to use the dataset and run the different machine learning algorithms. Each algorithms have different parameters, for instance, the dependent and the independent variable which makes it hard to switch around and keep track of different functionalities of each algorithms.

Chapter 3

Requirements Analysis

This chapter recognizes the Functional and the Non-functional requirements of the system with its priorities. These requirements are vital for the project as it ensures the development of the system.

The requirements are categorized on the basis of priority as:

- **Must Have**: These requirements are essential for the successful implementation of this project.
- **Should Have**: These requirements are essential and should be included in the system it possible.
- **Could Have**: These requirements are optional and have a small impact if left out.

3.1 Functional Requirements

Functional requirements are functions that developers must implement to enable users to complete their tasks. These requirements describe the system behavior.

FR No.	**Requirements Description**	**Priority**
1	The data should be preprocessed	Must
2	The system should be able to analyze the dataset	Must
3	The system should generate accurate results	Must
4	The system should be able to present Machine Learning models	Must

3.2 Non-Functional Requirements

The Non-functional requirements describes the constraints and the general characteristics of the system.

NFR No.	**Requirements Description**	**Priority**
1	The system should be scalable	Should
2	The system should be able to run on any OS	Could
3	The user should be able to read the presented data	Must
4	The GUI should be easy to use	Should

3.3 Evaluation Strategy

Evaluation strategy is used to find out whether the aims and objectives of the project have been satisfied. Hence, it is necessary to evaluate and test the usability, functionality of the system.

3.3.1 Functional Requirements analysis

- **The data should be preprocessed**
  To evaluate this, the author must preprocess the data to run machine learning algorithms and generate models.
- **The system should be able to analyze the Dataset**
  To evaluate this, the author needs to make sure that the system testing the dataset is correct and is in accordance to the topic.
- **The System should generate accurate results**
  This will be evaluated by making sure that the system generates accurate results for the dataset using different machine learning techniques.
- **The systems should be able to present Machine Learning Models** In order to accomplish this, the machine learning algorithms should be trained to identify certain type of patterns.

3.3.2 Non-Functional Requirements

- **The system should be scalable**
  The system should be designed in such a way that there's a scope for improvement and future work.
- **The System should be able to run on any OS**
  To ensure this, the author need to make sure that the system designed runs on any type of operating system i.e. Windows 10, Windows 7 or MacOS
- **The user should be able to read the presented data**
  The data presented by the system can be read by all the age groups.

- **The GUI should be easy to use** To ensure that it works the author needs to make sure that the GUI created can be used by all age groups.

3.4 Research Question

One of the main questions that need to be answered through this dissertation is, which Machine Learning algorithm achieves the best accuracy on the dataset. These various algorithms were selected on the basis of the research papers on Greenhouse gases and its results.

The next question that needs answering is, which country and/or sectors for the particular dataset is responsible for high GGE emissions.

Chapter 4

Project Implementation

4.1 Data Collection

The data is collected from an open source website (Climate Watch, 2020). The data set consists of Middle eastern countries with its sectors and their GHG emissions over the years from 1990-2016

Testing: The data set will be tested on the final model which will create machine learning model.

Evaluation: Upon testing, the results will be compared to find if the models are not overfitted. To evaluate this, different datasets of same type will be running in parallel to see if the results of the models are similar or not.

4.2 Data Preprocessing

For data preprocessing, non-numeric data and rows with no values will be removed.

Testing: The processing of the data should be tested to check that it doesn't cause issues when using it in the model.

Evaluation: A modelling technique will be used to evaluate that the data preprocessing is valid or not. Moreover, a deeper analysis will be carried out to check if the removal of non-numeric or rows with no data, affects the model.

4.3 GUI

For the GUI, a simple intuitive web application is designed for the user to interact with, it will hold the graphs and models generated from algorithms.

Testing: The GUI is tested before the deployment so that it does not crash whilst viewing the data.

Evaluation: For evaluating GUI, the graphs and models generated should be used on the GUI for rigorous testing.

Chapter 5

Project Management

This section covers the development phase of the dissertation with the help of Gantt chart and how the risks will be managed. Project management also covers the development methodology used.

5.1 Methodology

This project follows Waterfall Process which is perfect for the system. Waterfall process' management is sequential, linear process of project management and the requirements should be defined in the start. It comprises of several discrete phases. None of the phases begin until the previous phase is complete [5]

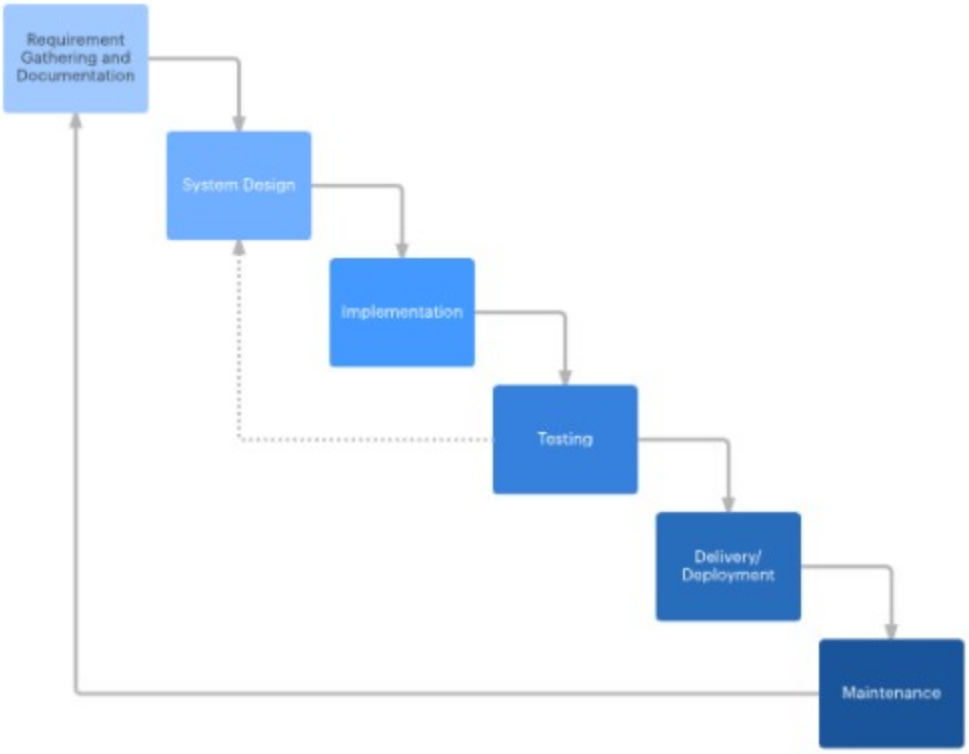

FIGURE 9: Waterfall Processe [5]

5.2 Risk Management

Risk management is a crucial part of the project management, it is important to understand all the risks in order to take actions upon those risks accordingly. Documenting the risks will safeguard the implementation of the project.

5.2.1 Risk Identification

This involves in identifying the risks that may be encountered during the development process. Metrics associated with risks are:

Probability

Very Low	Low	Moderate	High	Very High

Impact

Very Small	Small	Medium	Large	Severe

Priority

Medium	Large	Severe

Type

Technology	Organisation	Estimation	People	Tools	Requirements

ID	Risk	Type	Probability	Impact	Contingency
R1	No Dataset	Tools	Low	Severe	Find the dataset again from the source
R2	Machine Learning algorithms package is missing	Technology	Low	Severe	Reinstall the packages for the particular Machine Learning Algorithm
R3	Computational Feasibility	Tools	High	Moderate	If there are multiple tasks, make sure that other tasks are running on another device.
R4	Unable to meet the deliverable deadline	Organization	Low	Severe	Ensure that the task is planned and completed on time.
R5	Health is at risk	People	Moderate	Medium	Ensure to follow a proper healthy diet.

5.3 Project Plan

In this section the project plan for this dissertation is discussed, this includes the plan for the next semester as well. The tasks mentioned in the plan are completed iteratively. The project plan was created using the Gantt Charts.

5.3.1 Gantt Chart

This part covers the Gantt Chart diagram which shows the timeline of the project.

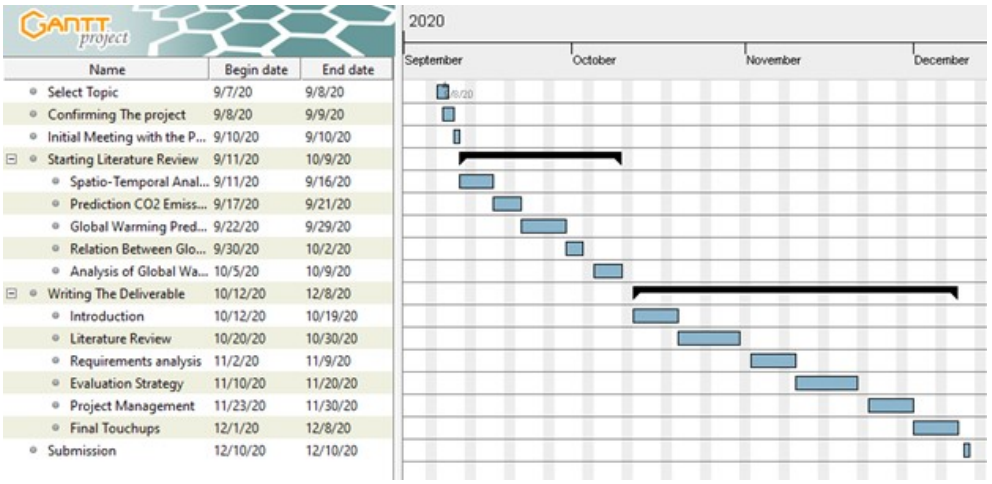

FIGURE 10: Gantt Chart for Deliverable 1

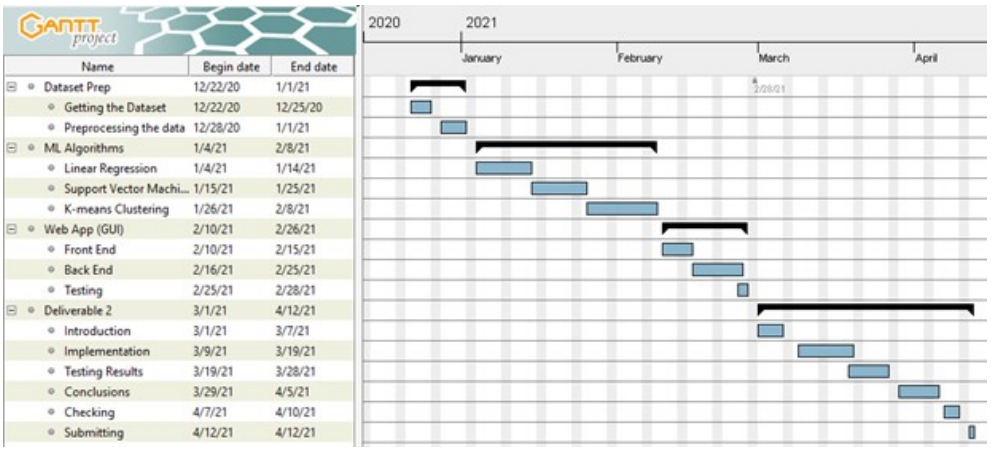

FIGURE 11: Gantt Chart for Deliverable 2

5.4 Professional, Legal, Ethical and Personal Issues

Professional and legal issues: Research papers that have been used in this dissertation will be properly referenced. The data sets and the software to be used is from the open source sites. *Climate Watch has an open data commitment and provides information free of constraints and restrictions on use* [1]

Ethical and Personal Issues: This dissertation does not involve testing with human beings, confidential or personal data. The project does not have any social issues.

Bibliography

[1] Data Explorer. *https://www.climatewatchdata.org/data-explorer/historical-emissions?historical-emissions-data-sources=caithistorical-emissions-gases=ch42021.*

[2] What Is Data Mining? https://docs.oracle.com/cd/B28359$_0$1/*datamine*.111/*b*28129/*process.htm*

[3] What is Data Mining? https://www.talend.com/resources/what-is-data-mining/: :text=Regression2021.

[4] What is Data Mining? *https://www.talend.com/resources/what-is-machine-learning/*, 2021.

[5] What the Waterfall Project Management Methodology Can (and Can't) Do for You. *https://www.lucidchart.com/blog/waterfall-project-management-methodology*, 2021.

[6] Z. S. Abdallah, L. Du, and G. I. Webb. Data Preparation. *https://doi.org/10.1007/978-1-4899-7687-1*$_6$2, 2021.

[7] C. Clifton. Data mining — computer science. *https://www.britannica.com/technology/data-mining*, 2021.

[8] A. Cuzzocrea, M. M. Gaber, S. Lattimer, and G. M. Grasso. Clustering-Based Spatio-Temporal Analysis of Big Atmospheric Data. *https://doi.org/10.1145/2896387.2900326*, 2016.

[9] D. D. Hema, A. Pal, V. Loyer, and Gaurav. Global Warming Prediction in India using Machine Learning. *https://doi.org/10.35940/ijeat.a1301.109119*, 2019.

[10] S. Kalra, R. Lamba, and M. Sharma. Machine learning based analysis for relation between global temperature and concentrations of greenhouse gases. *https://doi.org/10.1080/02522667.2020.1715559*, 2020.

[11] V. S. Kavitha S and R. R. A comparative analysis on linear regression and support vector regression. *https://doi.org/10.1109/GET.2016.7916627*, 2016.

[12] D. Kunda and H. Phiri. An Approach for Predicting CO2 Emissions using Data Mining Techniques. *https://doi.org/10.5120/ijca2017915098*, 2017.

[13] C. Liu. A Top Machine Learning Algorithm Explained: Support Vector Machines (SVM). *https://www.kdnuggets.com/2020/03/machine-learning-algorithm-svm-explained.html*, 2020.

[14] Michael.Walker. The Data Mining Process. *https://www.datascienceassn.org/content/data-mining-process*, 2016.

[15] H. Ritchie and M. Roser. CO and Greenhouse Gas Emissions. *https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions*, 2020.

[16] E. Takle and D. Hofstrand. Global warming – impact of greenhouse gases. *https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1291context=agdm*, 2020.

[17] E. Takle and D. Hofstrand. Fig.1. Simplified representation of the Global Carbon Cycle (Source:... *https://www.researchgate.net/figure/Simplified-representation-of-the-Global-Carbon-Cycle-Source-Intergovernmental-Panel-on*$_f ig 3_2 53152754, 2021$.

[18] L. Tan. The Art and Science of Analyzing Software Data. *https://www.sciencedirect.com/science/article/pii/B9780124115194000173*, 2015.

[19] X. Teng and Y. Gong. Research on Application of Machine Learning in Data Mining. *https://doi.org/10.1088/1757-899X/392/6/062202*, 2018.

[20] H. Zheng. Analysis of Global Warming Using Machine Learning. *https://www.scirp.org/journal/paperinformation.aspx?paperid=86337*, 2018.