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 types of research conducted on different Datasets of the world in different countries in order to get a 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 Decision Tree 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:

- Data set analysis and preprocessing
- Run multiple Machine Learning algorithms to find out which algorithm gives the best result.
- Find out which Sector of the Country is responsible for its high emission.
- Design a web application to visualize 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 life cycle 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 [21]

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 [21].

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 [22].

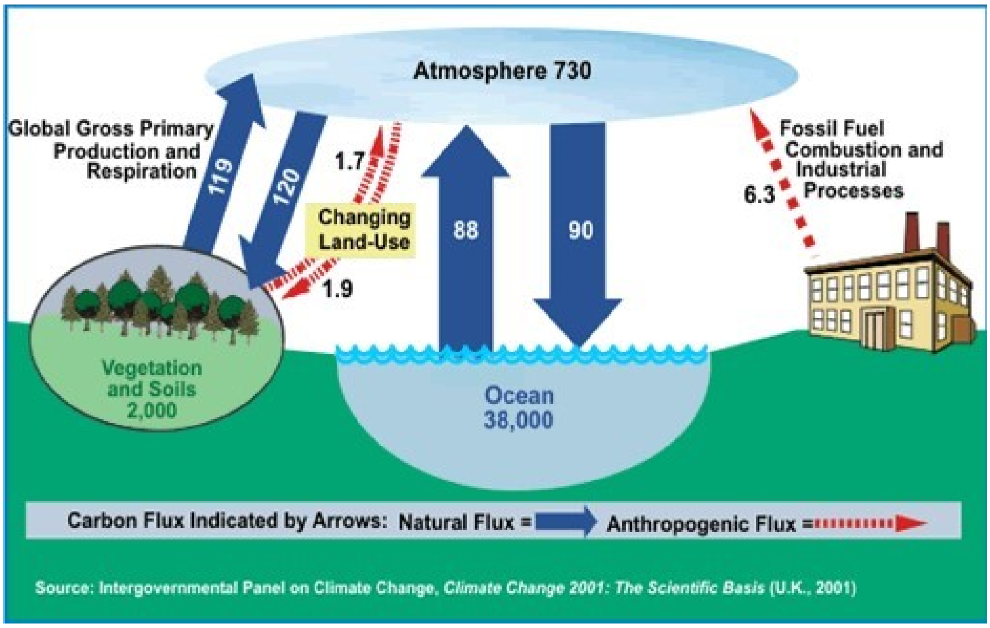

Atmosphere 730
Global Gross Primary Production and Respiration
119
120
1.7
Changing Land-Use
1.9
88
90
Fossil Fuel Combustion and Industrial Processes
6.3
Vegetation and Soils 2,000
Ocean 38,000
Carbon Flux Indicated by Arrows: Natural Flux = Anthropogenic Flux =
Source: Intergovernmental Panel on Climate Change, *Climate Change 2001: The Scientific Basis* (U.K., 2001)

FIGURE 1: Carbon Cycle [23]

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

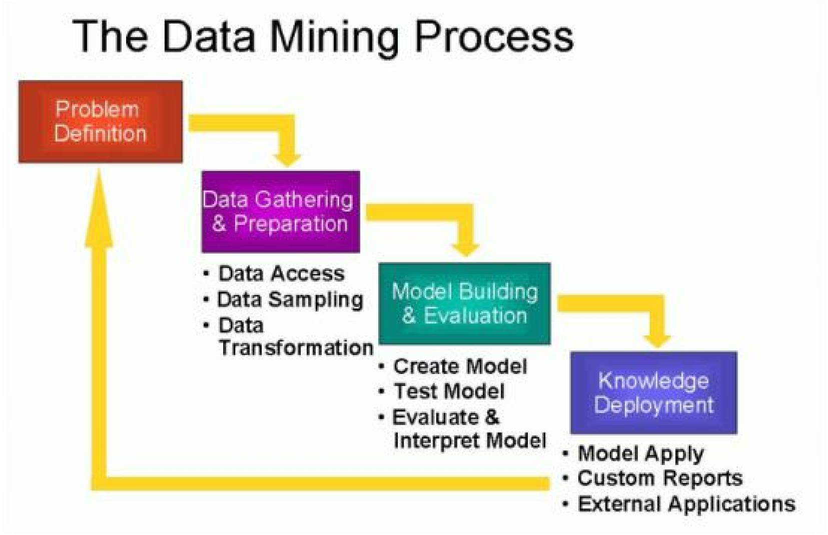
Data mining is the process of searching large stores of data to look for patterns that wouldn't be possible through simple analysis. Data mining uses mathematical algorithms to section the data and assess the chances of future events. Data mining is also known as Knowledge Discovery in Data (KDD)[8]

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 [12]

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. [11]

- Machine learning- It is a process where the machine acquires the skills and knowledge by identifying and using preexisting knowledge [25]
- Regression— A method used to predict a range of numeric values, for example sales, temperatures etc, using machine learning based on a data set. [9]



The data mining process involves a series of steps to define a business problem, gather and prepare the data, build and evaluate mining models, and apply the models and disseminate the new information.

FIGURE 2: The Data Mining Process [20]

Machine learning is letting the computers to learn and consequently act the way humans do, improving its learning and knowledge over time autonomously. [10]

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.). [10]

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 [10]

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=xβ+ϵ, 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 [17]

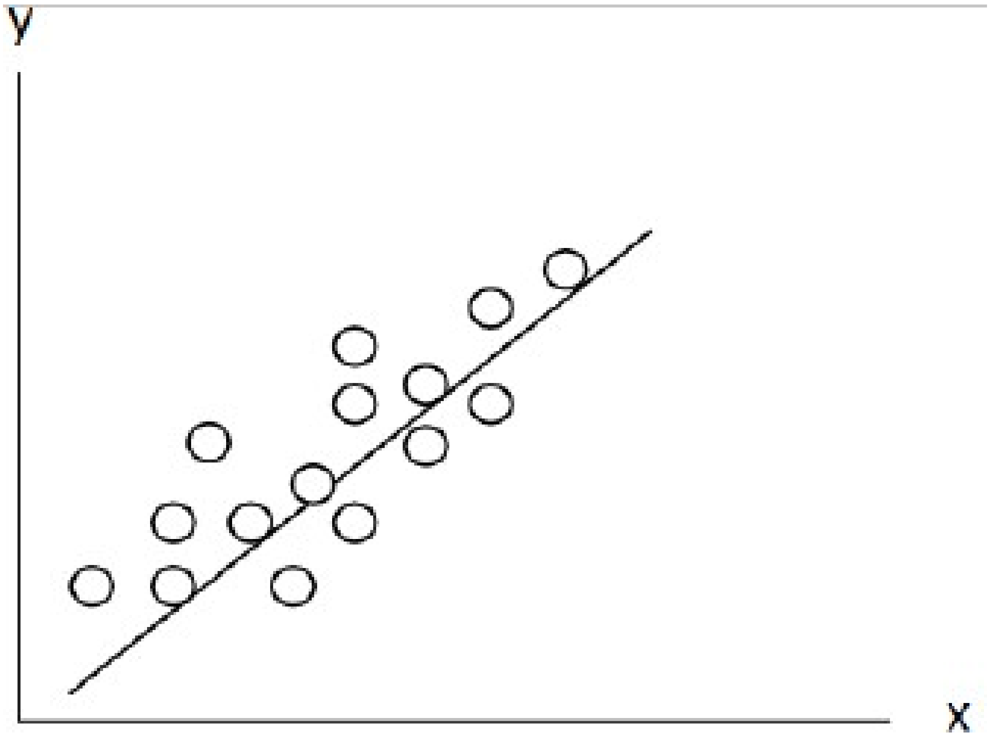


FIGURE 3: Linear Regression [17]

- **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 [19]

- **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 [24]

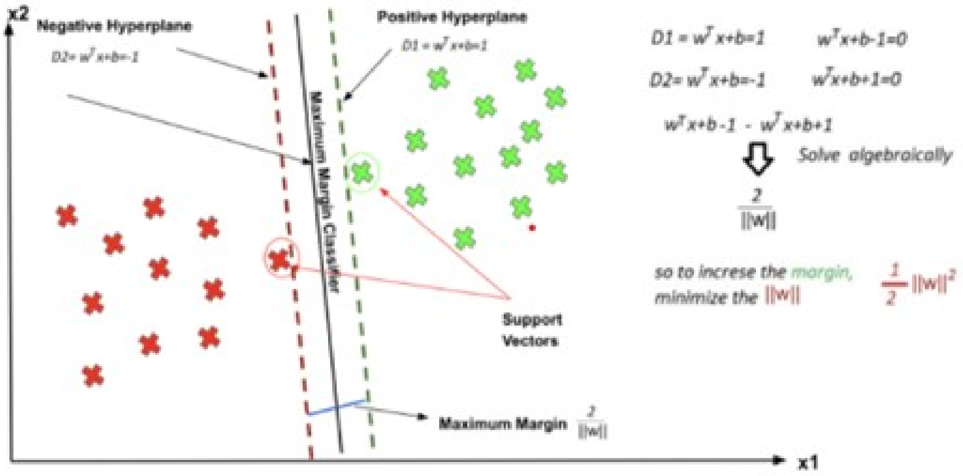


Figure 4: Support Vector Machine [19]

2.2 Related Works

2.2.1 Spatio-Temporal Analysis of Big Atmospheric Data

Cuzzocrea et al.[13] 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 [13].

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[13].

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.

2.2.2 Prediction CO2 Emissions using Data Mining

Kunda et al. [18] 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

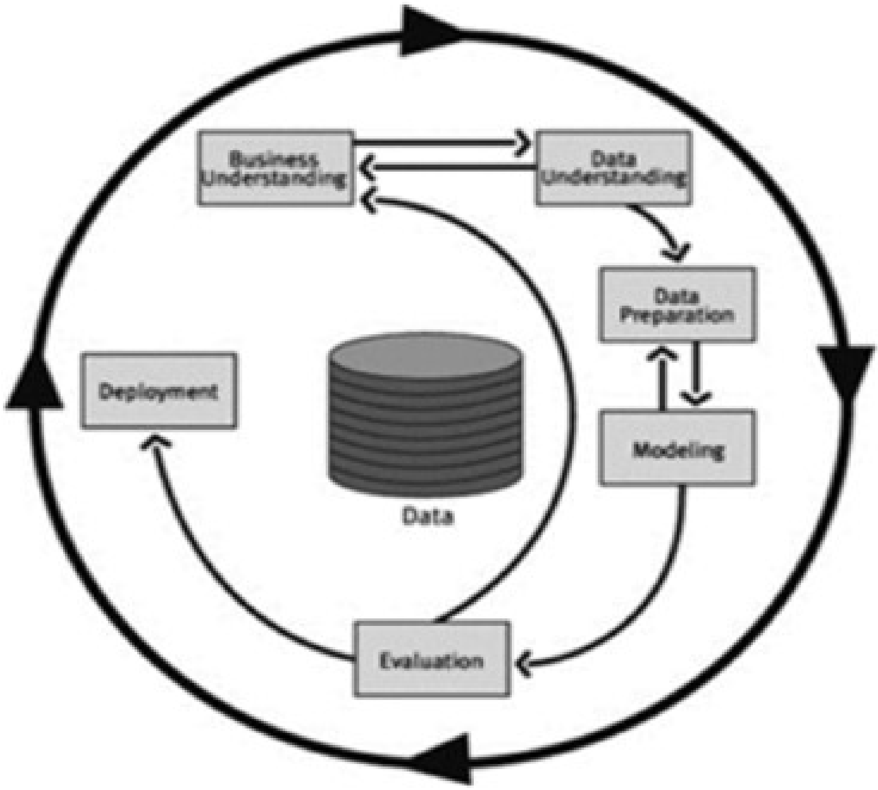
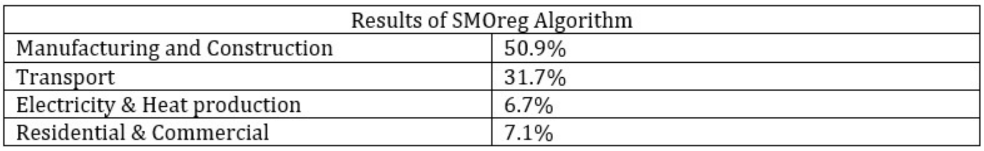


FIGURE 5: CRISP-DM Process [13]

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. [18]

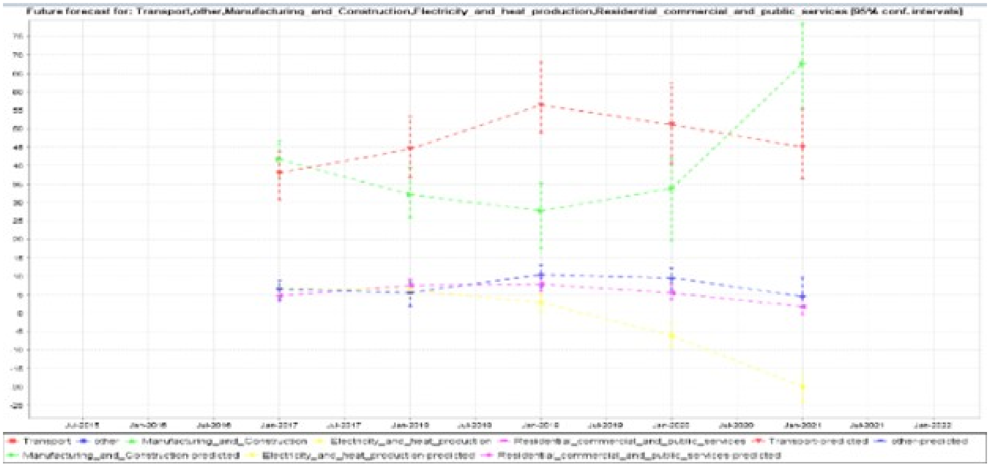


Figure 6: : Zambia CO2 forecast [18]

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. [14] 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 [14].

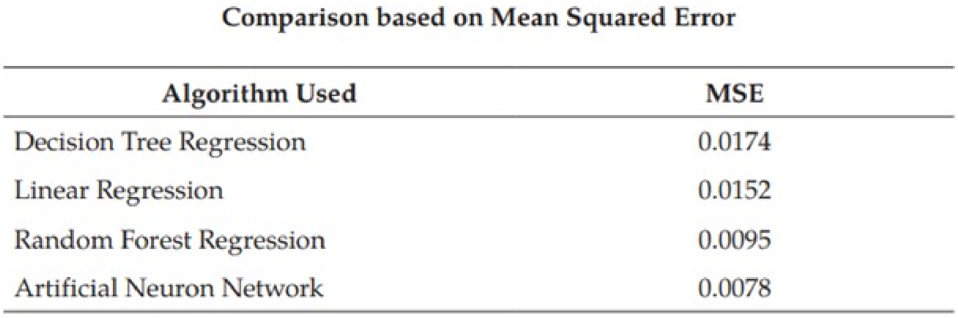
Critical Review: The approach seems to be detailed and well-presented. The author was able to predict and forecast the temperature for the next few years. The author could have used some more algorithms for comparison and selection. The model predicts only the average of temperature and the GHG concentration 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. [16] 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 hyperparameters. 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 [16]

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.



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

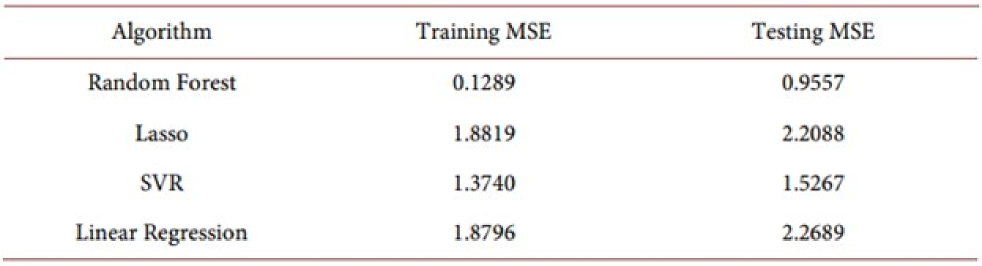
FIGURE 7: Comparison of Different Algorithms [16]

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 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 [26] 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 [26]



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 [26]

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.

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.

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 Functional and Non-Functional Analysis

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.

Chapter 4

Evaluation Strategy

After the creation of the models and predicting the testing values with the actual, we need to interpret the models reliability. It is possible in many ways, but for this experiment we will look into the following:

- Coefficient of Determination (R^2)
- Mean Square Error (MSE)
- Mean Absolute Error (MAE)
- Explained Variance

4.1 Coefficient of Determination (R^2)

R^2 is a statistical measure that represents the proportion of the variance explained by an independent variable or variables in a regression model for a dependent variable [15]. R^2 value lies between 0 and 1, and can be easily interpreted, for example, R^2: 0.75 means that 75% of the dependent variable is predictable.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$SS_{Regression}$: Residual sum of squares.

SS_{Total}: Total sum of squares.

4.2 Mean Squared Error (MSE)

Mean Squared Error (MSE) also known as Mean Squared Deviation (MSD) measures the average of the squares of the error in other words it measures the average squared difference between the actual value and the expected value. Most regression algorithms commonly use Mean Squared Error for evaluating the results.

$$\textbf{MSE} = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - \hat{y}_i)^2$$

y_i: Actual Value

$\hat{y}_i$: Expected Value

4.3 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) in statistics is a measure of the difference in errors between paired observations describing the same phenomenon [15]. It is simply known as the average of all absolute errors. The values of MAE lie between 0 and ∞ and small value of MAE indicates a better model.

$$\textbf{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

y_i: Actual Value

x_i: Expected Value

4.4 Explained Variance

Explained Variance also known as Explained Variation is a measurement of the difference between actual data and a model. Explained variance is similar to Coefficient of Determination, the main difference between the two is that Mean Error is being subtracted in Explained Variance. If the Mean Error is equal to 0, then R^2 is equal to Explained Variance.

$$\textbf{Explained Variance Score} = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

$Var(y - \hat{y})$: Variance of predicted error

$Var(y)$: Variance of actual values

Chapter 5

Implementation

5.1 Tools Used

5.1.1 Programming Language

For the project, Python version 3.8 [6], was used as the programming language of choice. Python is recognized as a high-level and general purpose programming language. Python is very easy to write, read and understand. There are various libraries that are supported in python. Although its usability and readability are a huge factor in the decision for choosing the programming language, but it is not memory efficient as it uses a lot of memory to run a program and hence slowing down the process. As the Dataset being used for the project is not that huge, it shouldn't be much of a problem.

5.1.2 Libraries

Python has vast libraries support, so for the project, ***Numpy*** Library [4] was used for the preprocessing stage to get rid of zeros in the dataset. ***Pandas*** Library [5] was used to read the Comma-separated values and convert it into a ***DataFrame*** [2]. To visualize the results ***Matplotlib*** Library [3] was used to show different types of graphs like Scatter Plot and Line Graph. To use the different algorithms, create its models and get metrics for assessing the errors, ***Scikit-Learn*** Library [7] was used.

5.1.3 Hardware

A good hardware is required to create models because it requires a lot of memory and CPU power. The hardware used for the implementation of all the objectives in order for it to work perfectly are as follows:

- Processor: Intel Core i7-9750H CPU @ 2.60 GHz
- RAM: 16 GB
- Operating System: Windows 10 Home 20H2

5.2 Dataset

5.2.1 Dataset Collection

The data was acquired from a Open Source website [1] consisting of different emissions recorded from the year 1990 to 2016 from the Middle Eastern countries.

The table 5.1 shows the list of attributes in the Dataset giving a description of the attributes and whether the attribute is required or not for the evaluation.

Attribute Name	**Description**	**Required**
Country	Country Names	Yes
Data Source	The source of the data	No
Sector	The various sectors of the country	Yes
Gas	The different types of Gases	No
Unit	The unit measurement of the gases	No
Years	Recorded GHG emissions of the years from 1990 to 2016	Yes

Table 5.1: Attribute Description

The removed attributes carry no weight in regards for the evaluation on the Dataset and hence they were removed.

5.2.2 Dataset Preparation

Prior to the preprocessing of the dataset, the dataset consisted of 930 rows and 31 columns, out of which, 26 columns consists of the years from 1990-2016 with the values of Greenhouse Gases.

To prepare the data, all the rows with 0 were first converted to ’NaN’ (Not a Number) type values to easily remove them and then convert back the ’NaN’ values to 0, if any, using the Numpy Library in Python.

```
data = data.replace(0, np.nan)
data = data.dropna(how=’any’, axis=0)
data = data.replace(np.nan, 0)
```

Originally, the Year attributes in the Dataset started from the year 2016 all the way till 1990. To make it readable and also to make a model systematically, I decided to reverse the columns in order.

5.3 Machine Learning Algorithms

After the Dataset has been preprocessed according to the requirements for the Machine Learning Algorithms, we are going to explore the Three Algorithms for the experiment.

5.3.1 Linear Regression

To create a Linear Regression model, the Dataset was first preprocessed to meet the requirements for the model i.e., removing the zeroes from the rows. As a part of preprocessing the data, I also removed the columns ”Country” and ”Sector” for the exact reason that it hinders the creation of the model, and to create one, we just need the numerical values.

The model learns from the Dataset by dividing it into two variables i.e., Independent variable denoted by ’X’ and Dependent variable denoted by ’y’. The independent variable consists of the GHG data from the year 1990 to the year 2015 whereas the dependent variable consists of GHG data from the year 2016.

```
y = test['2016']
X = test.drop(['2016'], axis=1)
```

These divided datasets are split into two i.e., 80% training data and 20% testing data using the ***train_test_split*** method from the Scikit-Learn library [7]. The model is then trained using both the independent and dependent training datasets. The model created can now be used to make predictions.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
model.fit(X_train, y_train)
```

Results	
Evaluations	Accuracy
Mean Square Error (MSE)	3.49
Mean Absolute Error (MAE)	0.7965
Variance Score	0.9994
R^2	99.94%

Table 5.2: Linear Regression Results Table

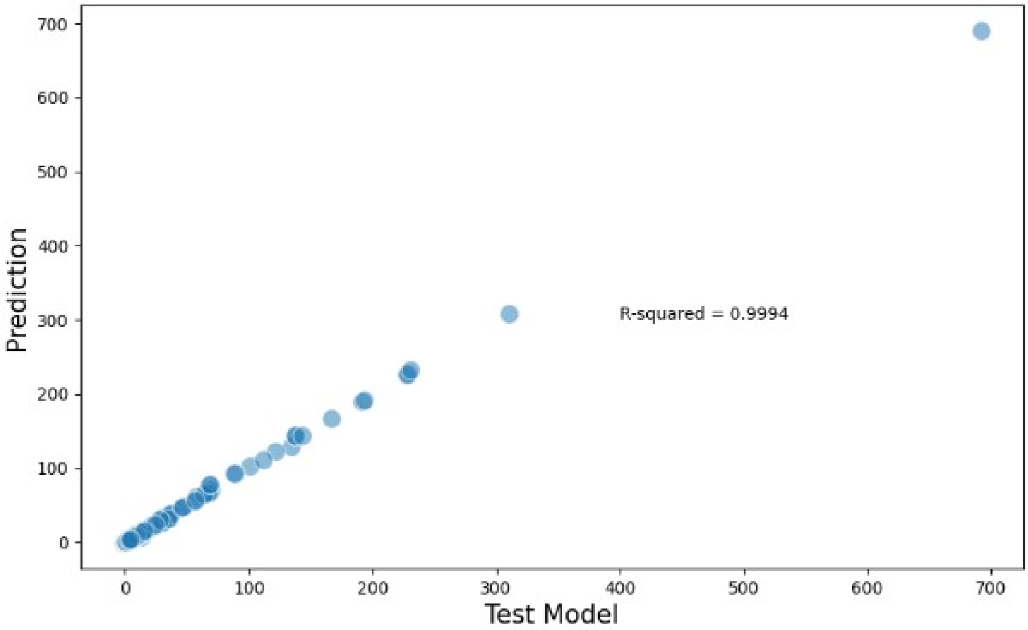


Figure 9: Test Model vs Prediction

Above is the graph for Test Model vs Prediction where it shows how accurate the created model is. The Test Model contains the GHG values of the year 2016 and the prediction is what the model predicts values on the basis of the Testing Dataset. The accuracy depicted from the graph is **99.94%**. More description of the results is shown in table 5.2.

5.3.2 Support Vector Machine

To create Support Vector Machine model, the Dataset was similarly preprocesssed like the one in Linear Regression to meet the requirements for the model.

SVM model learns from the Dataset divided into Independent variable and Dependent variable denoted by ’X’ and ’y’. The independent variable consits of the GHG data from the year 1990 to the year 2015 while the dependent variable consists of GHG data from the year 2016.

```
X = data[
    [’1990’, ’1991’, ’1992’, ’1993’, ’1994’, ’1995’, ’1996’, ’1997’, ’1998’, ’1999’,
     ’2000’,’2001’, ’2002’, ’2003’, ’2004’, ’2005’, ’2006’, ’2007’, ’2008’, ’2009’,
     ’2010’, ’2011’,’2012’, ’2013’, ’2014’, ’2015’]]
y = data[’2016’]
```

These divided datasets are split into two i.e., 70% training data and 30% testing data using the Scikit-Learn package ***train_test_split*** method. A ***Random State*** of 50 is selected after changing comparing it with other states as it gives us the most optimal result. The random state helps decide the splitting of the dataset.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=50)
```

To get a good model, I decided to use the ***GridSearchCV*** method provided from the Scikit-Learn package, it helped me run different hyperparameters options, without it, I would have to create a manual loop to check each hyperparameter to get an optimum accuracy. In SVM there are two parameters **’C’** which denotes misclassification, and **’Gamma’** which specifies how far the measurement of a plausible line of separation is influenced, excluding ’Kernel’ which I decided for it to be default i.e., ’rbf’. A model is created after an optimum parameter is selected.

```
grid_parameter = {’C’: [0.1, 1, 10, 100], ’gamma’: [1, 0.1, 0.01, 0.001]}
grid = GridSearchCV(SVR(kernel=’rbf’), grid_parameter, verbose=2)

grid.fit(X_train, y_train)
```

Results	
Evaluations	Accuracy
Mean Square Error (MSE)	2108.67
Mean Absolute Error (MAE)	8.9998
Variance Score	0.5991
R^2	59.56%

TABLE 5.3: SVM Results Table

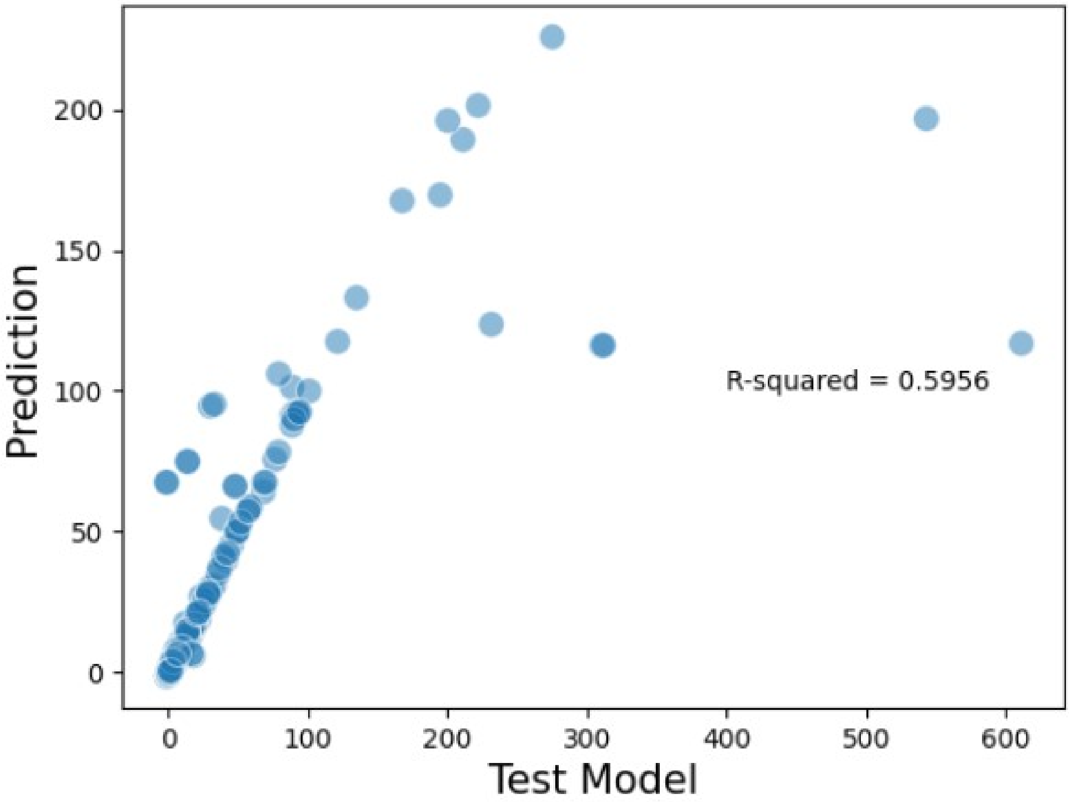


FIGURE 10: Test Model vs Prediction

Above is the graph for Test Model vs Prediction where it shows how accurate the created model is. The Test Model contains the GHG values of the year 2016 and the prediction is what the model predicts values on the basis of the Testing Dataset. The accuracy depicted from the graph is **59.56%**. A much more detailed information is depicted in table 5.3

5.3.3 Decision Tree

The Decision Tree model was constructed by preprocessing the data as stated in section 5.2.2. Further preparation was done to create the model, the columns named "Countries" and "Sector" were removed because it is not either of the types of categorical or numerical and hence it would hinder the creation of the model.

The Dataset was divided into Independent and Dependent variable denoted by 'X' and 'y' comprising of the GHG data from the year 1990 to the year 2015 and the GHG data from the year 2016 respectively.

```
X = data[
    ['1990', '1991', '1992', '1993', '1994', '1995', '1996', '1997', '1998', '1999',
     '2000','2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009',
     '2010', '2011','2012', '2013', '2014', '2015']]

y = data['2016']
```

These variables are then split into two, 70% training data and 30% testing data using the ***train_test_split*** method from the Scikit-Learn library. A Random state of 50 is chosen as the parameter as it helps the splitting of the dataset. This was chosen after creating and running the model multiple times.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=50)
```

To build the perfect model, I experimented with the parameters of the ***DecisionTreeRegressor*** by changing its 'max_depth' and creating multiple models to check the accuracy on each of the 'max_depth'. After experimenting, I came to a conclusion that after a max_depth of 5, there seems to be little to no change in the overall accuracy of the model. It is depicted clearly in Figure 11.

For the accuracy of the model, I chose Depth being equal to 20 as it gives the highest accuracy from the rest. The result of the model is depicted in the table below. Also, the visualization of the tree is shown in.....

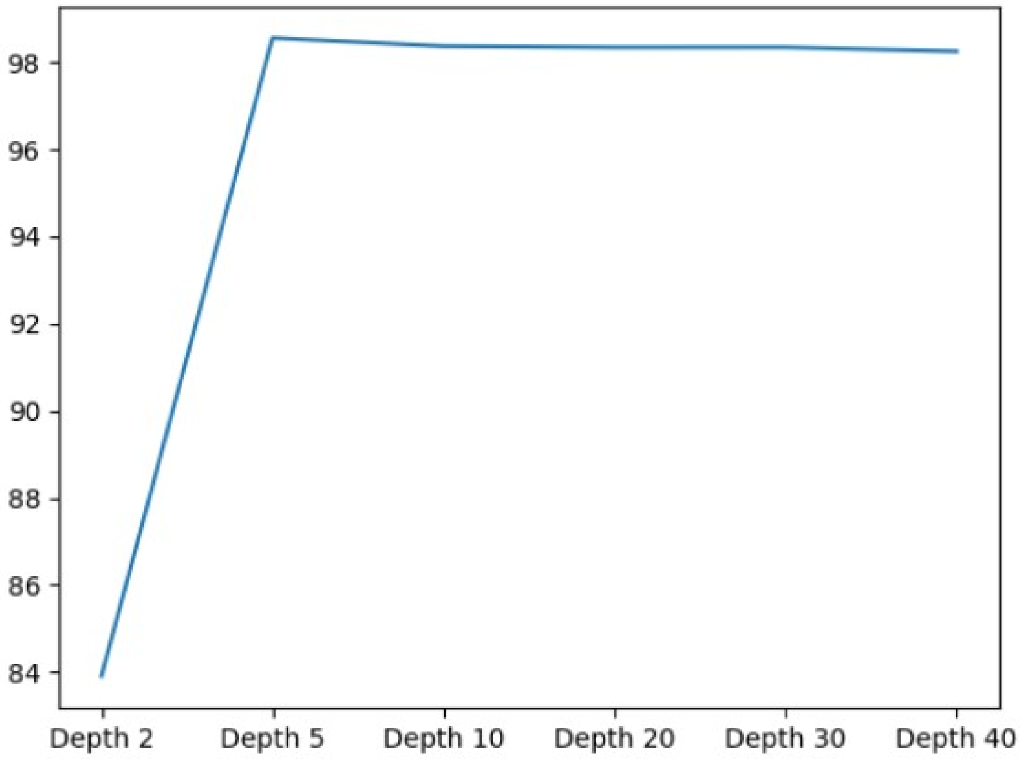


FIGURE 11: Change in accuracy vs the Depth

Results	
Evaluations	Accuracy
Mean Square Error (MSE)	84.36
Mean Absolute Error (MAE)	1.6486
Variance Score	0.9838
R^2	98.38%

TABLE 5.4: Decision Tree Results Table

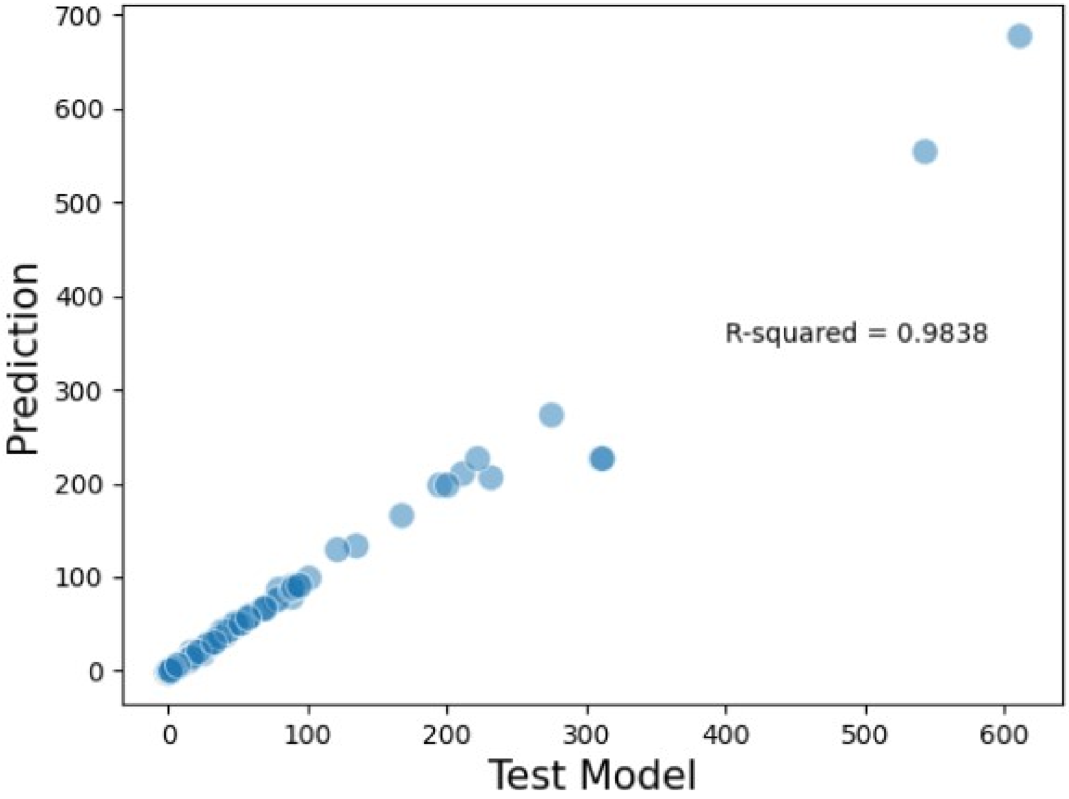


FIGURE 12: Change in accuracy vs the Depth

The Figure 12 shows the graph for Test Model vs Prediction for the Depth of **5** where it shows the accuracy of the model. The Test Model contains the GHG values of the year 2016 and the prediction is what the model predicts values on the basis of the Testing Dataset. The accuracy depicted from the graph is **98.38%**. A much more detailed information is depicted in table 5.4.

5.4 Sector

Finding out sectors within the country which are responsible for high emission can help the people who would like to take action for the same in order to reduce the overall emission.

To find out which sector in a country is responsible for its high emission, I firstly processed the data i,e., I removed the columns ’Data source’ and ’Unit’ for the reason that they were redundant and held no real value.

```
data.drop(’Data source’, inplace=True, axis=1)
data.drop(’Unit’, inplace=True, axis=1
```

I then moved on to selecting only those rows which have ’All GHG’ because the data contained mixed gases constituting of ’N20’, ’CO2’, ’CH4’ and ’F-Gas’. ’All GHG’ consisted of all theses gases and hence opting to choose only this was a good option. ’Total including LUCF’ and ’Total excluding LUCF’ was removed from the dataset because it consisted of the total GHG values of all the sectors including and excluding ’Land-Use Change and Forestry’ which is not a good estimate to find out the exact sector in a country responsible for high emission.

```
data = data.loc[data[’Gas’] == ’All GHG’]
data = data.loc[data[’Sector’] != ’Total including LUCF’]
data = data.loc[data[’Sector’] != ’Total excluding LUCF’]
```

Now that we have filtered data with only the right columns and rows, we can proceed to sum each rows with sector’s GHG values and put the values in a new column called ’GHG’.

```
data["GHG"] = data.sum(axis=1)
```

I then proceed to group the countries with GHG values to convert it into a list, because in a DataFrame, we cannot group by multiple columns and so therefore I decided to put the calculated GHG values in a list and then locate the rows which have the same values as in the list.

```
countriesGHG = data.groupby(’Country’, as_index=False)[’GHG’].max()
listOfGHG = countriesGHG[’GHG’].tolist()
filtered = data.loc[data[’GHG’].isin(listOfGHG)]
```

The data we have right now still needs to be filtered. This new data has 20 rows with unique countries and its sectors which was selected on the basis of the highest total emission over the years from 1990 to 2016 along with the individual GHG values of each year. I decided to remove the individual GHG values to get a much cleaner data.

```
filtered = filtered.drop(['1990', '1991', '1992', '1993', '1994', '1995', '1996', '1997',
                          '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005',
                          '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
                          '2014', '2015', '2016', 'Gas'], axis=1)
```

Dataset with Sectors		
Country	Sector	GHG
Algeria	Energy	3121.63
Bahrain	Energy	570.35
Djibouti	Agriculture	18.21
Egypt	Energy	3975.08
Iran	Energy	12604.98
Irqa	Energy	2709.41
Israel	Energy	1522.25
Kuwait	Energy	1615.02
Lebanon	Energy	411.46
Libya	Energy	2927.94
Malta	Bunker Fuels	66.10
Morocco	Energy	1097.53
Oman	Energy	938.02
Qatar	Energy	1140.15
Saudi Arabia	Energy	8685.04
Syria	Energy	1645.66
Tunisia	Energy	571.58
United Arab Emirates	Energy	3694.03
Yemen	Energy	435.37

Table 5.5: Country and Sectors Table

The table 5.5 shows the filtered data with the unique countries, its sector and the total GHG values from the years.

These values from the table 5.5 is better depicted by a graph. This graph was created using the **Matplotlib** Library.

```
filtered.set_index(['Country', 'Sector']).plot.bar(color='royalblue')
plt.xlabel('Country-Sector', size=15)
plt.ylabel('GHG Value', size=15)
plt.title('Countries and its Sector responsible for high GHG', size=15)
plt.grid(color='#95a5a6', linestyle='--', linewidth=2, axis='y', alpha=0.7)
figure(figsize=(12, 5), dpi=2060)
plt.rcParams["figure.figsize"] = (12,5)
```

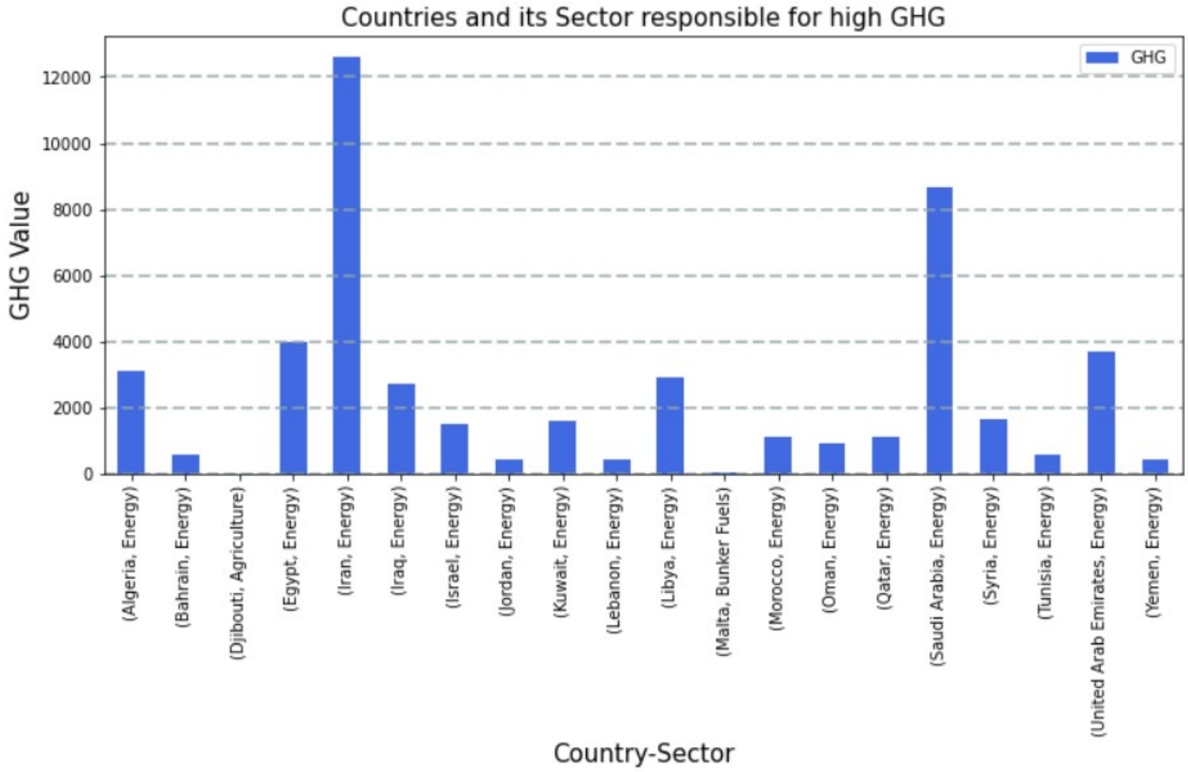


FIGURE 13: Country and its Sector Graph

Figure 13 shows us that the Country **Iran** is responsible for the highest emission in the Middle East. And the sector contributing to the most of the emission in Iran is the **Energy** sector, followed by **Saudi Arabia** and its sector being **Energy**. The country with the least amount of emission is **Djibouti**.

5.5 GUI

Bibliography

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

[2] *Data Frame.* https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html*, 2021.*

[3] *Matplotlib.* https://matplotlib.org/*, 2021.*

[4] *Numpy.* https://numpy.org/*, 2021.*

[5] *Pandas.* https://pandas.pydata.org/*, 2021.*

[6] *Python 3.8.* https://www.python.org/dev/peps/pep-0569/*, 2021.*

[7] *scikit-learn.* https://scikit-learn.org/stable/*, 2021.*

[8] *What Is Data Mining?* https://docs.oracle.com/cd/B28359$_0$1/$datamine.111/b28129/process.htmCHDFC$

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

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

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

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

[13] 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.

[14] 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.

[15] S. Hiregoudar. Ways to Evaluate Regression Models. *https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70*, 2021.

[16] 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.

[17] 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.

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

[19] 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.

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

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

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

[23] 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}ig3_{2}$53152754, 2021.

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

[25] 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.

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