DATA-DRIVEN SOLUTIONS FOR SUSTAINABLE DEVELOPMENT

Neema Martin Manadan, Meenu Hani, Resmi K R

Keywords-sustainable development goals (SDG); data-driven; deep learning (DL); machine learning (ML), cloud computing; green computing

Abstract

United Nations have come up with 17 sustainable development goals and 169 related targets that each member state should follow to reach the goals by 2030. However, as a result of the emergence of a variety of environmental issues over time, such as climate change and air pollution, sustainable development, which aims to improve human lives on Earth, has become widely accepted. Sustainability is a long-term economic and environmental benefit. In today's world of information and communication technology, businesses and organizations are much more concerned with developing long-term operational strategies. The primary objective is to lower operating costs while simultaneously reducing their environmental and Carbon footprint. To achieve sustainability objectives, businesses can use automated data collection and analysis to make strategic, real-time decisions. In addition, cutting-edge technology is able to extract profound insights from data, opening up a vast array of novel strategies for businesses to encourage sustainable behavior.

1. Introduction

Globally, sustainability has been a significant focus due to society's rapid growth. According to Hollaway et al. (2020), one viewpoint from the previous century holds that the development framework only needs to satisfy current requirements without taking into account whether or not future generations will be able to meet their requirements. However, as a result of the emergence of a variety of environmental issues over time, such as climate change and air pollution, sustainable development, which aims to improve human lives on Earth, has become widely accepted. According to Ahmad T. et al. (2022), sustainable development is the growth of people, the economy, and society at the expense of nature, life support, and community. We must preserve the environment because it may take hundreds of years for it to recover from damage, manage the use of resources because they are limited, and protect the Earth's ecosystems. In the meantime, it's essential that everyone has equal access to resources like healthcare, education, and housing so they can live their everyday lives.

The United Nations (U.N.) n.d.) proposed eight Millennium Development Goals (MDGs) to be met by 2015, including extreme poverty and hunger, education, equality, well-being, and the environment. According to "The Millennium Development Goals Report 2015" (United Nations, 2015a), the establishment and achievement of the MDGs have saved and improved the lives of millions of people, particularly in impoverished nations. The MDGs' achievement is made possible by international cooperation and support. The United Nations adopted a new Sustainable Development Agenda in 2015 that included 17 Sustainable Development Goals (SDGs) and 169 related targets that each member state must accomplish by 2030 (United Nations, 2015b). According to United Nations (2015a), despite significant progress and improved living conditions, the poorest people were still being left behind. With the goal of "leaving no one behind," the Sustainable Development Goals (SDGs) help people achieve a more prosperous and sustainable future. (Fernando, C., and Koswatte) .

Because (Big) data-driven analytics, artificial intelligence (A.I.), the Internet of Things (IoT), deep learning (DL), and machine learning (ML) have profoundly altered processes in recent years and have impacted our lives in every way.

2. Literature Review

Defining industry-specific goals to achieve the most significant and pertinent results is the first step toward a data-driven, sustainable organization. The next step is to construct a model for treating data, refine that model, and then use that model to examine how current procedures relate to sustainability objectives. Combining human and artificial intelligence to confidently make decisions for a business based on data is the final step.(Bachmann, et al., 2022).

Google's carbon-intelligent computing platform maximizes the energy efficiency of its data centers by utilizing machine learning methods. Their most recent sustainability innovation aims to shift the timing of many computer tasks so that they occur at times when renewable energy sources like wind and solar are most abundant. Flexible computing workloads can also be moved between data centers, allowing for more work to be done when and where it is best for the environment.

Tomorrow's electricity Map, which enables businesses to evaluate the carbon footprint of their electricity usage, is another excellent example of data-driven sustainable technology. By utilizing data-driven solutions from Google Cloud to accelerate the process of designing solar panels, Sun Power is making solar energy significantly more accessible to consumers. Sun Power is a tool that, to increase the use of solar panels, end users must be empowered to design rooftop solar configurations. This task is made more accessible by machine learning.

The Stella McCartney Foundation and Current Global are utilizing data analytics and machine learning to assist fashion brands in becoming more sustainable. The foundation is working on a tool to give fashion companies a complete picture of their supply chain, where brands have little to no visibility and account for most environmental impact.

In addition to meeting legal requirements, these technologically driven sustainable solutions provide a competitive advantage through green technology innovation.

Cloud computing is offered as a long-term solution to address and achieve these objectives. The ability to access computing resources like applications, storage, services, video games, movies, and music on demand so that Cloud clients do not need to know how or where they are receiving these contents makes Cloud computing an emerging technology rapidly gaining attention. They only required broadband Internet access to the Cloud.

The purpose of this research is to find out how cloud-based data-driven solutions support the U.N.'s SDGs. As a result, the main focus was figuring out the opportunities various kinds of Big Data provided for cloud-based data-driven solutions.

3. Methodology

The research was conducted through the quantitative research design using a comparative approach. The main objective of the study was to analyze the regional Carbon free emission and various carbon emissions from multiple countries and analyze the trends or differences between them.

The primary data sources for this research were the Regional Carbon free emissions dataset obtained from the Google Cloud Public dataset (BigQuery Public Datasets Program,2021) available on https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/ and the carbon emission dataset from the EDGAR (Emissions Database for Global Atmospheric Research) available on https://edgar.jrc.ec.europa.eu/report_2022#data_download for the various CO2 emissions of all world countries. The regional Carbon free emission contained the different year's carbon-free emissions obtained with the regions and locations where the area is present, along with the CFE metrics. The EDGAR dataset was further divided to find the analysis for only the Asian countries, including the type of emission, EDGAR code, and the various Asian countries.

The datasets were then analyzed using various machine-learning algorithms. The regional Carbon free emission dataset was analyzed by multiple supervised and unsupervised learning due to incomplete records for analysis. The EDGAR dataset was only used for EDA(Exploratory Data Analysis) as it was used only to find the trends in carbon emissions within various Asian countries.

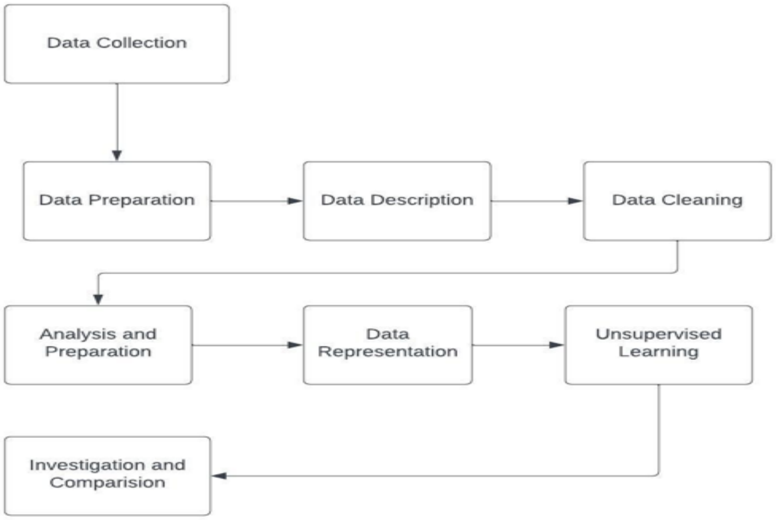

Fig 3.1 Flow Diagram for Regional Carbon Free Dataset

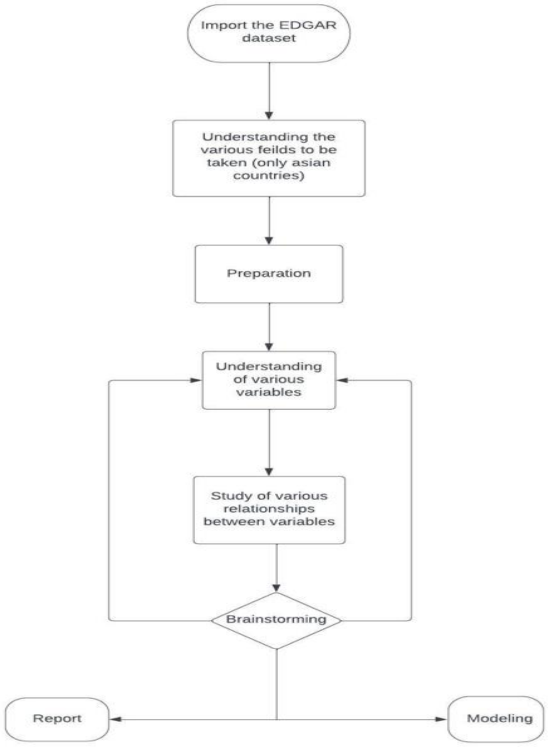
Import the EDGAR dataset
Understanding the various feilds to be taken (only asian countries)
Preparation
Understanding of various variables
Study of various relationships between variables
Brainstorming
Report
Modeling

Fig 3.2 Flow Diagram for EDGAR Dataset

4. Experimental Results

4.1 Datasets

This research was done on two datasets, primarily the Regional Carbon Free Emission and the EDGAR dataset, for the various CO2 emissions of all world countries.

Regional Carbon Free data was acquired by taking the various regions of google cloud datacenter and their respective locations, as each of these regions consists of grids where electricity is generated. These grids have high and low carbon emissions based on the type of electricity consumed. Google Cloud has set up a goal by 2030 that all their data centers would match their energy consumption with CFE(%) metric.CFE metric measures the percentage of carbon-free energy produced every hour and in every region. In addition to this, they use an A.I. model called carbon-intelligent computing, which allows shifting all compute tasks between different data centers based on the CFE(%). This is Google's first step towards green computing as it provides various data centers around their regions to shift their computing periods to an hour where more renewable resources like solar and wind are available. It also gives a day-ahead prediction of how much energy a grid will produce that Carbon free emission which allows computing activities to be shifted globally. As Google uses this method, it can move most of its computing tasks to energy consumption that utilizes less carbon-intensive energy consumption.

The EDGAR dataset consists of the various findings of the countries that use GHG emissions. We can see that the leading factor of global warming is the CO2 emission which is still increasing worldwide. Therby, EDGAR gives an independent estimate of greenhouse gases for each world country based on the IPCC guidelines.

4.2 Preprocessing

The regional Carbon free emission dataset included 96 rows and six columns. The columns were year, CFE region, zone I.D., cloud region, location, and Google CFE. The detailed data description includes

- **Year:** Year when the Google CFE metric has been aggregated.
- **CFE Region:** Regional boundary of the electric grid when calculating the Google CFE score.
- **Zone ID:** I.D. associated with the CFE Region on Tomorrow's ElectricityMap API.
- **Cloud Region:** Google Cloud Region that is mapped to the CFE region.
- **Location:** The various Cloud Region.
- **Google_CFE:** This metric is calculated for every hour in every region, which tells the percentage of Carbon Free Emission.

The dataset was imported to Google Colab for various machine learning algorithms. The following was then used with the isnull() function to check for the different null values present in the dataset. Accordingly, the values were dropped, and the fillna() function replaced the missing values. After that, all the categorical data is encoded using the Label Encoder. The Label Encoder() function converts all the categorical attributes into ranges of numeric values that are easier for the machine to interpret. Various scaling features like the StandardScaler() and MinMaxScaler were used to scale the various columns to a range that is better for interpretation. Then the dataset is split into testing and training sets. All the columns except for the Google CFE are taken as the independent variable, and the remaining column is, Google CFE is taken as the dependent variable. The test size was 20% of the data.

Various supervised learning methods, including KNN, Linear Regression, and Decision Tree, were used. When Knn was used with five neighbors, it had an accuracy of 0.25, which means the model was only correct 25% of the time, which was very low. Thereby this algorithm was not taken. Linear Regression was carried, giving a Mean squared error of 1.61, indicating that predictions were off by about 1.61 from the actual values. The root means a squared error of 1.27 was an interpretable measure displaying an average forecast of 1.27 units. Then R-squared score of 0.26 told the model could only predict 26% of the variance in the dataset, which meant that this dataset also could not be taken. Then finally Decision tree algorithm was used, but it also gave a satisfactory result. Thereby, supervised learning methods could not be used for this dataset.

Unsupervised learning methods included K-means, a clustering algorithm that allows users to cluster the dataset according to the user. It randomly assigns K points in the dataset as initial cluster centers. Then gradually changes the cluster heads(centroids) according to the mean of all the data points in that particular cluster. By doing this algorithm, we can see that within each cluster, data points with similar characters are grouped together, making it easier for analysis. The elbow method finds the minimal number of clusters that a dataset can have. Following the region Carbon free dataset, the elbow method was used, and it showed a minimum of 4 clusters could be grouped from the dataset. Accordingly, Google CFE% also was grouped into 4 clusters shown in Table 4.1.

Table 4.1 Various Clusters within the Region Carbon-Free Emission Data

Google CFE(%)	**Cluster**
0.0-0.25	Low Carbon-Free Emission
0.25-0.50	Medium Carbon-Free Emission
0.50-0.75	High Carbon-Free Emission
0.75-1	Very Carbon Free Emission

Accordingly, the regions and locations were mapped to the various clusters. This analysis was better than the others because it gave good insight into how Google Cloud is working towards the goal they have set.

The EDGAR dataset includes 55 rows and 11,715 records. The columns had substance, EDGAR country code, Country, and the various CO2 emission from 1970-2021. The dataset was further divided by only taking the multiple countries in the Asia continent.

The dataset was imported to Google Colab to do the exploratory data analysis. No null values were present when the isnull() function was used. First, a histogram was made using the matplot library. The hist() function was used further to plot the data point. By using this, we can represent the Histogram of CO2 Emissions from 1970 to 2021. The changes in the total carbon emission in various countries were described using a line graph. Finally, to get an overview of the Total Carbon Emissions by year, a bar graph was used, and all the years(1970-2021) as the x-axis and all total carbon emissions were represented in the y-axis.

4.3 Results

Table 4.2 shows the various cloud regions' locations, google CFE, and the multiple clusters. This was obtained by using the K-means algorithm.

Table 4.2 Regional Carbon-Free Emission Data using K-Means

Cloud Region	**Location**	**Google CFE(%)**	**Cluster**
us-east1	South Carolina	0.19,0.25,0.27	Low Carbon-Free Emission
us-east4	Northern Virginia	0.48,0.58,0.64	High Carbon-Free Emission
us-east5	Columbus	0.64	High Carbon-Free Emission
us-west1	Oregon	0.89,0.9,0.88	Very High Carbon Free Emission
us-west2	Los Angeles	0.55,0.54,0.53	High Carbon-Free Emission
us-west3	Salt Lake City	0.25,0.28,0.31	Low Carbon-Free Emission
us-west4	Las Vegas	0.13,0.19,0.21	Low Carbon-Free Emission
asia-east2	Hong Kong	0.28	Low Carbon-Free Emission
asia-south1	Mumbai	0.12,0.1	Low Carbon-Free Emission
asia-south2	Delhi	0.12,0.1,0.08	Low Carbon-Free Emission
us-central1	Iowa	0.78,0.93,0.97	Very High Carbon Free Emission
europe-west1	Belgium	0.68,0.79,0.82	High Carbon-Free Emission
europe-west2	London	0.54,0.59,0.57	High Carbon-Free Emission
europe-west3	Frankfurt	0.63,0.6,0.61	High Carbon-Free Emission
europe-west4	Netherlands	0.61,0.6,0.53	High Carbon-Free Emission
europe-north1	Finland	0.77,0.91,0.94	Very High Carbon Free Emission
asia-notheast1	Tokyo	0.12,0.16	Low Carbon-Free Emission
asia-northeast3	Seoul	0.31,0.31	Low Carbon-Free Emission
asia-southeast1	Singapore	0.03,0.04,0.03	Low Carbon-Free Emission
asia-southeast2	Jakarta	0.13	Low Carbon-Free Emission
europe-central2	Warsaw	0.2	Low Carbon-Free Emission
southamerica-east1	Sao Paulo	0.87,0.88,0.78	Very High Carbon Free Emission
southamerica-west1	Santiago	0.69	High Carbon-Free Emission
australia-southeast1	Sydney	0.11,0.11,0.21	Low Carbon-Free Emission
australia-southeast2	Melbourne	0.31	Low Carbon-Free Emission
non-cloud-data-center	Chile	0.63,0.65	High Carbon-Free Emission
northamerica-northeast1	Montreal	0.1	Low Carbon-Free Emission

northamerica-northeast2	Toronto	0.92	Very High Carbon Free Emission

From Table 4.2, these are the findings that are found. There are no data points in the Medium Carbon Free Emissions. There are 14 regions with Low Carbon Free Emissions, nine regions with High Carbon Free Emissions, and five with Very High Carbon Free Emissions out of the total 28 areas taken.

When you take region-wise analysis, these are the observations:

1. us-east has 3 locations within it; among them, 1 is a Low Carbon Free Emission, and two are in High Carbon Free Emission. Thereby, the majority is High Carbon Free emissions, so we can say that this region can produce at least 75% CFE emissions.
2. us-west has 4 locations, and two come under Low Carbon Free Emission,1 High Carbon Free Emission, and 1 Very High Carbon Free Emission. Thereby the majority is Low Carbon Free emissions, so we can say that this region can only produce at least 25% CFE emissions.
3. asia-east,europe-central has only one location: a Low Carbon Free Emission. Thereby it can only grow at least 5% CFE emissions
4. asia-south, asia-northeast ,asia-southeast and austrialia-southeast have two regions, and both the locations produce Low Carbon Free Emission so only 10% of the emissions are CFE.
5. us-central, europe-north has only one place, and it produces Very High Carbon Free Emission so this region can produce atleast 90% CFE emissions.
6. europe-west has 4 locations, in which all four come under High Carbon Free Emission; thereby, it can produce 85% CFE emissions.
7. southamerica-west and non-cloud datacenter has only 1 location, and it produces High Carbon-Free emissions, which means that 50% of the emissions are CFE.
8. northamerica-northeast has 2 locations within it. One makes a Low Carbon Free Emission, and the other produces a Very High Carbon Free Emission, producing 60% CFE.

The EDGAR datasets consist of various graphs. These were the graphs obtained by using the matplot library in Python.

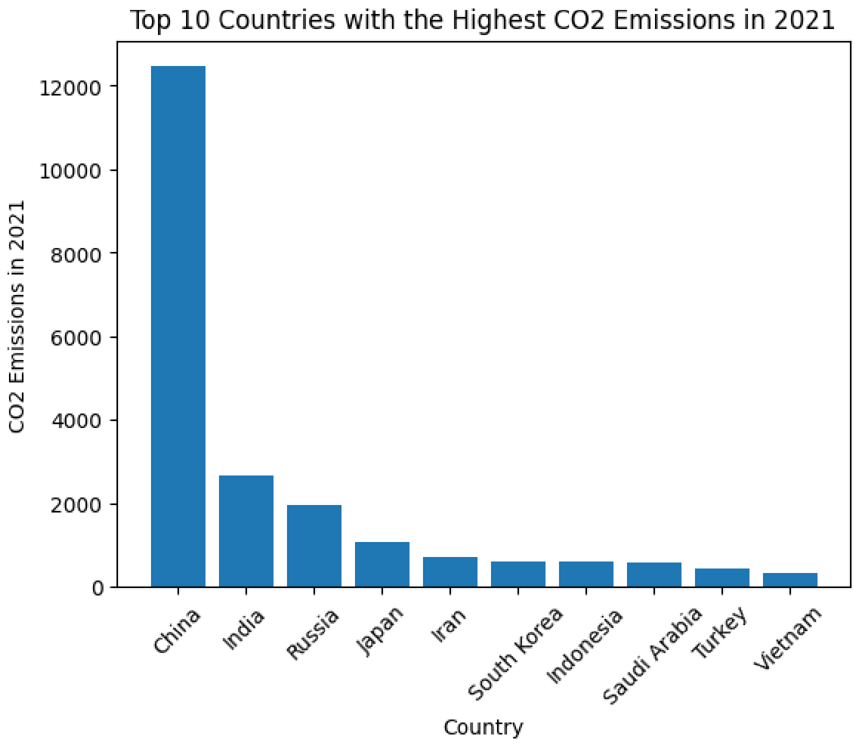

Fig 4.1 Stacked Bar Chart for EDGAR Dataset

In Fig 4.1, we can see the top 10 countries with the highest CO2 emissions. When analyzing the graph, we can see that China is the highest Country that produces carbon emissions due to its high population, rapid industrialization, and economic growth. It is the highest producer of manufactured goods for export. Whereas Vietnam is the lowest Country because it has a low population and more dependency on renewable resources like wind and solar, most of the economy is earned through agriculture.

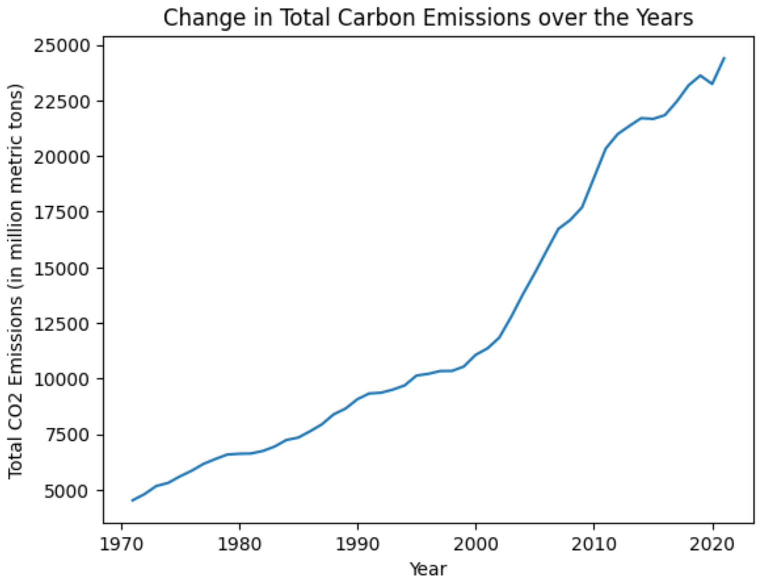

Fig 4.2 Line Graph for EDGAR Dataset

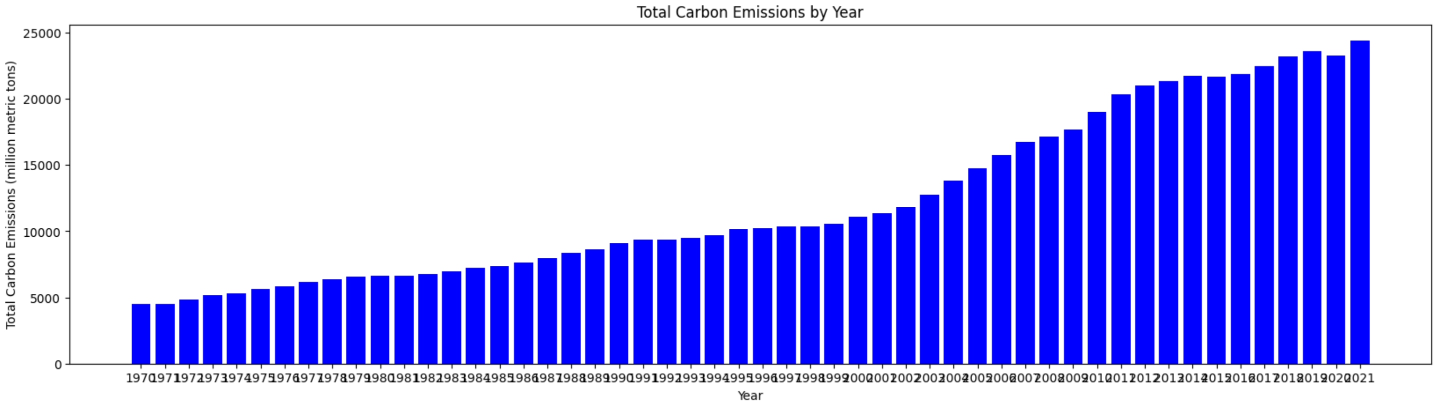

Fig 4.3 Bar Chart for EDGAR Dataset

When analyzing Fig 4.2 and Fig 4.3, we can see that a drastic change occurred from 1970-2021 due to the enormous population growth from 3.7 billion in 1970 to 7.8 billion in 2021. There has been a sudden increase in industrialization, and high usage of fossil fuels over the past 50 years, contributing to the changes. Deforestation and urbanization have led to changes in land usage which changed the balance of CO2 in the atmosphere.

5. Conclusion

To help Cloud computing save more energy and have a more minor impact on the environment, some technologies and methods in data center infrastructures are being investigated. A Green Cloud architecture is emerging from these trends, leading to energy efficiency and a concept that is carbon emission aware. We explain how Cloud computing can reduce energy use and carbon footprint, create a greener environment, and reduce costs to fully satisfy the meaning of sustainability that can be built upon the energy-efficient data center infrastructure. The idea of cloud computing involves a variety of concerns, technologies, and issues.

6. References

[1] Hollaway, M.J., Dean, G., Blair, G.S., Brown, M., Henrys, P.A. and Watkins, J., 2020. Tackling the challenges of 21st-century open science and beyond A data science lab approach. *Patterns*, *1*(7), p.100103.b

[2] Ahmad, T., Madonski, R., Zhang, D., Huang, C. and Mujeeb, A., 2022. Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renewable and Sustainable Energy Reviews*, *160*, p.112128.

[3] Bachmann, N., Tripathi, S., Brunner, M. and Jodlbauer, H., 2022. The contribution of data-driven technologies in achieving the sustainable development goals. *Sustainability*, *14*(5), p.2497.

[4] Fernando, C. and Koswatte, I., Data Science for Economic Development: A Sustainable Strategy. *REVITALIZING THE ECONOMY THROUGH SUSTAINABLE STRATEGIES*, p.247.

[5] BigQuery Public Datasets Program,2021, Regional Carbon-free Energy Data, https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/regional-cfe

[6] EDGAR - Emissions Database for Global Atmospheric Research, CO2 emissions of all world countries, https://edgar.jrc.ec.europa.eu/report_2022#data_download