

LD6053

UG Computing Project

**Analysis and Predicting Global Energy Sustainability Using
Machine Learning**

by

23028446 Yingxuan Zhang

Supervisor Name: Dr Syed Raza

Research Coordinator & Module Leader Name: Dr Rose Fong

A Dissertation

Submitted to Department of Computer and Information Sciences

Northumbria University

In Partial Fulfilment of the Requirements

For Bachelor of Science (Hons) in Computing

<submission date: 19/5/2024>

Declaration

I declare the following:

1. that the material contained in this project is the end result of my own work and that due acknowledgement has been given in the bibliography and references to ALL sources be they printed, electronic or personal.
2. the Word Count of this report is 6521.
3. that unless this dissertation has been confirmed as confidential, I agree to an entire electronic copy or sections of the dissertation to being placed on Blackboard, if deemed appropriate, to allow future students the opportunity to see examples of past reports. I understand that if displayed on Blackboard it would be made available for no longer than five years and that students would be able to print off copies or download. The authorship would remain anonymous.
4. I agree to my report being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service.
In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second marker, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.
5. I have read the UNN/CEIS Policy Statement on Ethics in Research and Consultancy and I confirm that ethical issues have been considered, evaluated and appropriately addressed in this research.

Signature: Yingxuan Zhang

Date: 18/5/2024

Acknowledgements

I would like to express my gratitude to Dr Rose Fong and Dr Syed Raza at Northumbria University (London) for their assistance. Furthermore, I would like to extend my gratitude to the professors from other academic disciplines. In addition to imparting knowledge, they also monitored my academic experience, providing me with a great deal of warmth. Finally, I would like to express my gratitude to my family and friends for their unwavering support and encouragement, which has provided me with the foundation to successfully complete my undergraduate studies. It is my hope that I will be able to make further contributions to the fields of data analysis and artificial intelligence in the future, and that I will be able to propose further solutions to achieve the world's sustainable development goals.

Abstract

This article deeply analyzes and predicts the sustainability of global energy through data analysis and machine learning techniques. The article first outlines the current global environmental crisis and the achievement of Sustainable Development Goals (SDGs), emphasizing the key role of SDG 7 in promoting other goals. The main objective of this study is to use data analysis and artificial intelligence technology (machine learning) to create more solutions for achieving sustainable energy consumption. The article provides a detailed introduction to research methods, including data preprocessing, feature engineering, machine learning model construction and evaluation, as well as the design and implementation of visual dashboards. Research has found that data exploration engineering can extract effective information from data, but the selected machine learning model has certain limitations in predicting CO₂ emissions. Finally, the article summarizes the research findings and proposes directions for future work. At the same time, the author also reflects on their own growth and learning in the project.

Contents

Declaration.....	1
Acknowledgements	2
Abstract.....	3
1 Introduction.....	5
1.1 Background	5
1.2 Objectives.....	6
1.3 Structure of the Report.....	6
2 Literature Review	7
2.1 Sustainable Energy.....	7
2.2 Machine Learning.....	9
2.3 Current Studies.....	10
2.4 Conclusions.....	11
3 Design of Practical Work	12
3.1 Applying Case	12
3.2 Data Structure	12
3.3 Approach and Methodology.....	14
4 Implementation and Testing.....	17
4.1 Data Pre-processing	17
4.2 Exploratory Data Analysis (EDA)	20
4.3 Building a Machine Learning Model for CO2 Emissions Prediction	26
4.4 Visualization Dashboard	29
5 Implementation and Testing.....	31
5.1 Discussion.....	31
5.2 Critical Evaluation.....	31
6 Conclusions and Self-reflection.....	33
6.1 Conclusion	33
6.2 Self-reflection.....	33
References.....	34
Appendix A – Extra EDA charts.....	37
Appendix B – Data Column Names and Definition.....	40
Appendix C – Ethical Approval	41

1 Introduction

The global situation is becoming increasingly complex today. All humanity is facing an unprecedented environmental crisis. Although the United Nations proposed a systematic framework for sustainable development goals in 2015-SDG (The Sustainable Development Goals). However, the progress in achieving sustainable goals is still worrying, and it is urgent to use innovative technologies to promote sustainable development.

1.1 Background

The Sustainable Development Goals are a set of seventeen interrelated global goals planned to be achieved by 2030. Approved by the United Nations General Assembly in 2015 (United Nations, 2016). The year 2030 is fast approaching, and in the past decade, climate change has become increasingly evident. Indeed, nine out of the ten years have been the hottest on record. Based on the current situation, it is not difficult to predict that the Earth is heading towards a drastic climate change, some of which will occur in the near future (Yadav, 2022). Not only the environment, but also the international situation has become increasingly tense, which has increased the resistance to global progress towards sustainable development and exacerbated the inequality in world development. This complex landscape is further complicated by the power struggles between major world players, including Russia, America, and China, which have been intensified by recent events such as the Ukraine crisis and the COVID-19 pandemic (Aslakhanova et al., 2021). This deviates from the global human expectations for the direction of future development. In this situation, it is necessary to take more actions and efforts to put sustainable development back on track. Nevertheless, in the course of implementing sustainable objectives, numerous countries have identified the necessity to elucidate the interrelationships between objectives, given that different objectives exert varying degrees of mutual influence. It is particularly important to identify and prioritize the most important of the seventeen sustainable development goals, so as to promote the achievement of other sustainable development goals by achieving a single goal (Allen et al., 2018). Many studies have shown that improving energy issues is more conducive to achieving a great blueprint for sustainable development (Yadav, 2022). In research on the interrelationships between SDGs, it has been found that SDG 7 “Ensures access to affordable, reliable, sustainable, and modern energy for all”, which is one of the most synergistic goals for other objectives. In its current form, Sustainable Development Goal 7 is specifically stated through five objectives, namely (7.1) ensuring universal access to affordable, reliable, and modern energy services, (7.2) increase global percentage of renewable energy, (7.3) doubling global energy efficiency, (7.4) strengthening international cooperation in clean energy research and technology, and (7.5) expanding and upgrading energy services in developing countries (United Nations, 2016). SDG 7 shows positive associations with ten or more goals (Griggs et al., 2017, Anderson et al., 2021). Dave Griggs and his colleagues (2017) found that Goal 7 has varying degrees of promotion with all 17 goals, especially with SDG-1 (“no poor”), SDG-2 (“zero hunger”), and SDG-3 (“good health and wellbeing”), SDG-6 (“clean water and sanitation”), SDG-8 (“decent work and economic growth”), and SDG-13 (“climate action”) have key interactions. Of course, relying on data

technology and artificial intelligence technology can more effectively achieve sustainable energy layout, and at the same time, the implementation of sustainable energy will greatly help the development of technology. Based on this background, this study will utilize data analysis and artificial intelligence technology (machine learning) to create more solutions for achieving sustainable energy consumption.

1.2 Objectives

For the research on data analysis technology and artificial intelligence technology in the direction of sustainable energy, this study mainly has three objectives.

Task 1 Through data exploration and analysis, provide an overview of energy consumption and explore the relationship between data columns.

Task 2 Build and train multiple machine models, compare model performance, and predict carbon emissions data in 2020.

Task 3 Building an interactive sustainable energy data visualization dashboard.

1.3 Structure of the Report

In order to achieve these stated objectives, the following chapter structure has been adopted for this report:

Chapter 1 - Introduction

Chapter 2 - Literature Review

Chapter 3 - Design of Practical Work

Chapter 4 - Implementation and Testing

Chapter 5 - Discussion and Evaluation

Chapter 6 - Conclusions, Recommendation and Self-reflection.

2 Literature Review

The literature review chapter will interpret concepts of sustainable energy, data analysis and machine learning, and provide a review and critical analysis of existing related research.

2.1 Sustainable Energy

2.1.1 Energy Consumption

The current energy demand is usually met by the sustained basic load of coal or nuclear power plants and the demand for other renewable energy sources such as wind and solar energy (Vincent et al., 2021). At the beginning of the 21st century, the world was depleting the available energy of fossil fuels (oil, natural gas, coal, natural gas, and nuclear energy), while renewable energy (wind and solar) had not yet developed enough to provide comprehensive and flexible alternatives. Although the transition to renewable energy has received a lot of attention in the sustainable development agenda and technologies already exist and are being rapidly deployed (Bórawski et al., 2019), the fact that fossil fuels largely dominate global energy production has not changed.

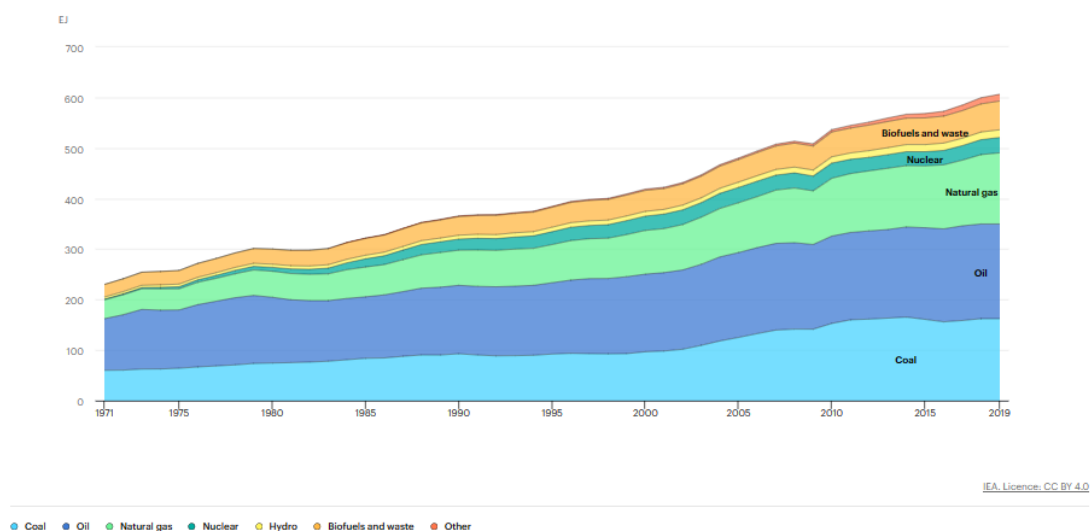


Figure 1. World total energy supply by source, 1971-2019 (IEA, 2021a).

As shown in Figure 1, although the global energy consumption growth rate has slowed down, the overall trend of sustained growth has not changed.

2.1.2 Renewable Energy

Renewable energy, including solar energy, wind energy, hydropower, biofuels, etc., is the core of the transition to a more carbon intensive and sustainable energy system. Although the overall global energy structure remains fragile, the development momentum of clean energy has been increasing year by year. According to the official data of SDG7.2 (increase global percentage of renewable energy), hosted by the International Energy Agency in Figure 2, in 2020, the share of modern renewable energy in the total final energy consumption decreased to

12.5% (International Energy Agency, 2021). However, with the popularization of sustainable development awareness and the efforts of all parties, the proportion of renewable energy has unprecedentedly increased and still maintains a good upward trend. The International Energy Agency (IEA) has forecast that, within the next five years, it will be able to achieve certain milestones in the field of renewable energy, including: by 2028, renewable energy accounts for more than 42% of global electricity generation, and the share of wind and solar photovoltaics will double to 25%.

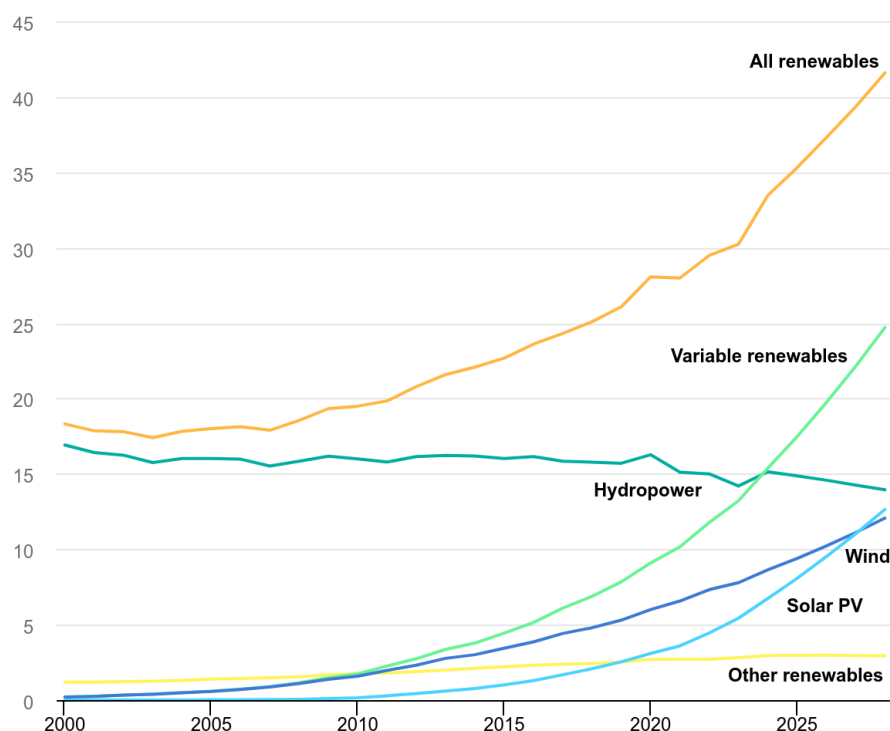


Figure 2. Share of renewable electricity generation by technology (IEA, 2021).

2.1.3 Energy Crisis

After the pandemic, the energy market began to tighten in 2021 due to various factors, but after the Ukraine war in 2022, the world is facing a comprehensive energy crisis. High energy costs have increased poverty, led to a decline in energy supply, and generated enormous economic, social, and political pressures. From the perspective of regional situation analysis, different regions of the world are also facing different energy crises. Due to population growth and pandemics, the number of people without electricity globally increased to 760 million in 2022. Currently, less than 20% of African countries have set the goal of achieving universal electricity supply by 2030. In 2022, approximately 2.3 billion people worldwide are unable to access clean cooking facilities and rely on traditional solid biomass, kerosene, or coal as their primary cooking fuels. This seriously damages health and productivity, with women and children being the most severely affected. The inefficient combustion of fuelwood and charcoal leads to the significant release of methane and other greenhouse gases, and unsustainable logging also contributes to deforestation, further exacerbating climate impacts (IEA,2024).

2.1.4 Energy & IT technology

It can be seen that the world is still facing a relatively serious energy problem at present. However, with the vigorous development of IT technology, in recent years, there have been new technological innovation deployments in the upstream and downstream of the energy industry, greatly improving energy efficiency. At the same time, it is the key for renewable energy to replace fossil energy sources (Kupzog et al., 2020).

The integration of artificial intelligence (AI) and big data in the energy sector is a growing area of interest, with potential to revolutionize decision-making and operations (Wang, 2021; Ahmad, 2021). AI models, including deep learning and machine learning, are increasingly being applied to energy control and decision-making processes (Jeon, 2022). These technologies are particularly beneficial in controllability, big data handling, cyberattack prevention, smart grid, IoT, robotics, energy efficiency optimization, predictive maintenance control, and computational efficiency. The utilisation of AI in the energy sector is anticipated to enhance operational performance and efficiency, thereby serving as a pivotal enabler within the industry (Ahmad, 2021). Energy analytics, powered by big data and machine learning, has seen significant growth, particularly in the use of energy computing techniques (Dhanalakshmi, 2021).

2.2 Machine Learning

Machine learning is a key component of artificial intelligence, and its technological core lies in using algorithms to enable computers to learn and evolve behaviours based on data. It brings together technologies from multiple disciplines, such as statistics and neuroscience, to gain insights through data and computation (Schneider&Guo, 2018). This field is closely related to pattern recognition, computational statistics, and artificial intelligence, and is used in various daily life applications such as image recognition and predictive analysis. Once they understand how to process specific data, they can automatically work (Kumar et al., 2020).

Machine learning algorithms are mainly divided into four categories :

- Supervised learning is the most widely used machine learning algorithm for learning functions that map inputs to outputs based on sample input-output pairs. It uses a set of labeled training data and training examples to infer functions. When certain goals are determined to be achieved from a specific set of inputs, supervised learning, also known as task driven methods, is performed. The most common supervisory task is to separate the "classification" of data and the "regression" of fitted data.
- Unsupervised learning is a data-driven process that does not require human intervention when analyzing unlabeled datasets. This is widely used for extracting generated features, identifying meaningful trends and structures, grouping results, and exploring purposes.
- Semi supervised learning falls between unsupervised learning and supervised learning. Semi supervised learning is useful for analyzing unlabeled data types. The

ultimate goal of semi supervised learning models is to provide better prediction results than using only labeled data in the model.

- Reinforcement learning enables software agents and machines to automatically evaluate the best behavior in a specific context or environment to improve their efficiency, which is an environment driven approach. It is a powerful tool for training artificial intelligence models, which can help improve the operational efficiency of automation or optimization of complex systems, but it is not suitable for solving basic or direct problems (Sarker, 2021).

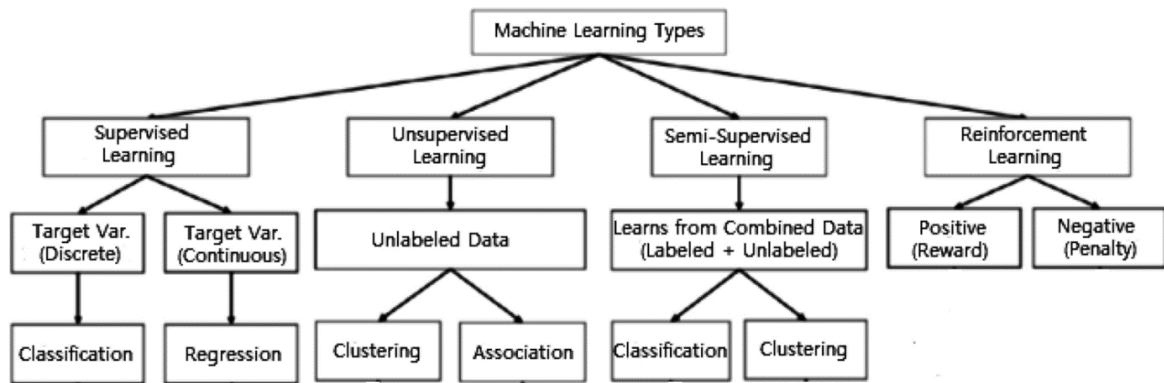


Figure 3. Various types of machine learning technologies (Kumar et al., 2020).

Common machine learning algorithms and application areas include support vector machines (SVM), decision trees, random forests, K-nearest neighbors (KNN), naive Bayes, and multi-layer perceptron for prediction, classification, regression, and various complex tasks (Mathur&Badone, 2019). The general structure of a machine learning based prediction model is shown in Figure 4, where the model is trained from historical data in the first stage and generates results for new test data in the second stage.

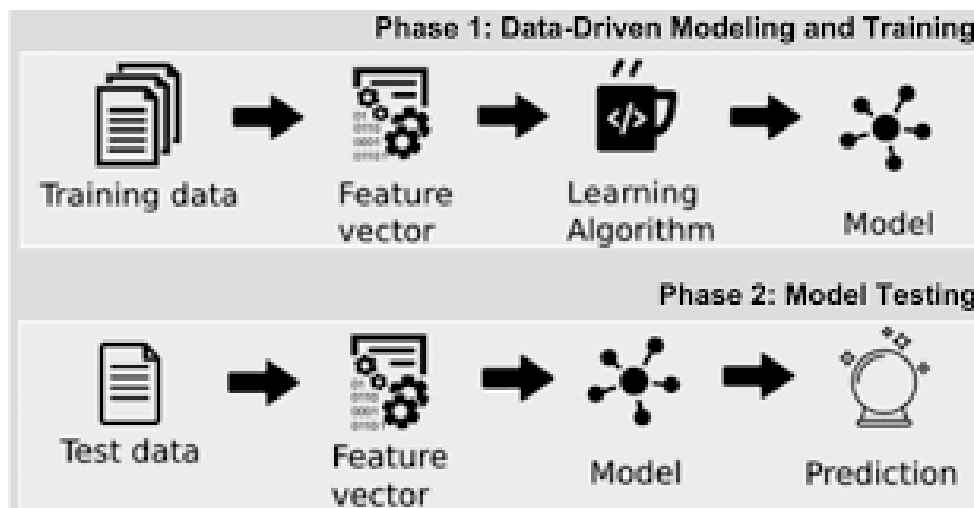


Figure 4. A general structure of a machine learning model (Sarker, 2021).

2.3 Current Studies

From material design and equipment management to system deployment, machine learning has achieved remarkable achievements in multiple fields of energy

technology. Machine learning is particularly suitable for discovering new materials, and researchers in this field predict that it will bring new materials that may completely change the energy industry (Yao et al., 2022). Bhatt et al (2023a) used multiple machine learning models (linear regression, K-nearest neighbor regression, decision tree regression, etc.) to predict global environmental carbon dioxide emissions, proposed key CO₂ concentration thresholds and emission reduction requirements, and concluded that a threshold of 500 ppm and emission reduction rate will be achieved by 2047. In the field of renewable energy, classic machine learning models such as linear regression, random forests, and support vector machines have been widely applied, and have the characteristics of simplicity, ease of interpretation, and requiring less computational resources (Benti et al., 2023b). A study has also proposed a method for predicting the total ecological footprint (EF) of primary energy consumption based on population and various sources. And four hybrid machine learning models based on Bayesian parameter estimation were proposed: (i) KNNReg, (ii) RFR, (iii) ANN ReLU, and (iv) ANN SPOCU. Finally, after horizontal comparison, the KNNReg model with the best performance was selected to develop a graphical user interface for automatically calculating EF based on user input data (Janković et al., 2020). At the same time, there is research focused on SDG 7 and machine learning. Matenga (2022) used unsupervised learning methods (ordinal K-means clustering) to form eight health level states. And judge the proximity of energy markets in different countries and regions to achieving Sustainable Development Goal 7 based on their health levels. Through machine learning models, it is necessary to improve the electricity supply in sub-Saharan Africa through or through policy changes, as the region has the largest number of low-level electricity supply countries. Developed countries such as China and the United States face sustainability challenges.

2.4 Conclusions

In summary, the development of sustainable energy has become particularly important in today's world, and with the development of artificial intelligence technology, the application of machine learning technology in the industry is becoming increasingly widespread, gradually shifting from cost reduction and efficiency improvement to a more sustainable direction. However, there is limited application of machine learning models for overall indicators of sustainable energy, and more research reviews have summarized the development and application of machine learning in different scenarios in the energy industry. And for the data analysis of indicators, unsupervised machine learning algorithms are used. Sustainable energy development, as an important indicator of SDG goals, should explore more machine learning applications.

3 Design of Practical Work

3.1 Applying Case

- Government and Decision-makers

This project will assist the government in formulating energy related policies for better achieving SDG 7. At the same time, the government can better understand the energy use situation in different regions of the world through data analysis, in order to formulate more data-driven energy policies and plans. They can formulate sustainable energy policies based on predictions from machine learning models to promote economic growth and reduce environmental impacts.

- Energy Companies and Suppliers

For those employed in the energy industry, the utilisation of data-driven methodologies for strategic deployment is of paramount importance. The incorporation of artificial intelligence technology represents a significant advancement in the transformation of the industry. The use of data analysis methods can enrich industry research more scientifically and uncover potential market demands. Energy companies can use machine learning to predict energy demand in different regions.

- Related Institutions and Organizations

Related institutions and organization can use this technology to monitor and evaluate the development of sustainable energy in various countries, and continuously promote the progress of national or relevant organizations in improving energy structures. Concurrently, they can monitor the effect of energy consumption on the environment and propose improvements to achieve more sustainable production and consumption in the energy industry and its related industries.

- Investors

Investors can choose to invest in the most attractive countries or regions based on factors such as energy demand, production capacity, and policy environment. Investors can use machine learning technology to predict the development trends, competitive landscape, and prospects of energy markets in different countries, in order to make wiser investment decisions. In the contemporary era, an increasing number of investors are giving consideration to environmental, social, and governance factors (ESG) when making investment decisions. Energy data analysis can help them evaluate the environmental friendliness, social responsibility, and governance level of energy industries in different countries to support their ESG investment strategies.

3.2 Data Structure

3.2.1 Data Source

This study uses a dataset from Kaggle, Global Data on Sustainable Energy (2000-2020), to explore the predictive performance of different machine learning models

on energy indicators. The appendix summarises the definitions of all indicators used in this article.

Link to the dataset: [Global Data on Sustainable Energy \(2000-2020\) \(kaggle.com\)](https://www.kaggle.com/datasets/tanwar12345/global-data-on-sustainable-energy-2000-2020) (Tanwar, 2023).

3.2.2 Data Information

Using the 'pandas' data analysis library, import the original CSV format file into Python, as shown in Figure 5. And using the 'info()' method to read the basic information of the data, a total of 21 columns and 3649 rows of data can be obtained in the dataset. The majority of data columns are of the float data type (see Figure 6). The output results reveal the presence of missing data points in the data.

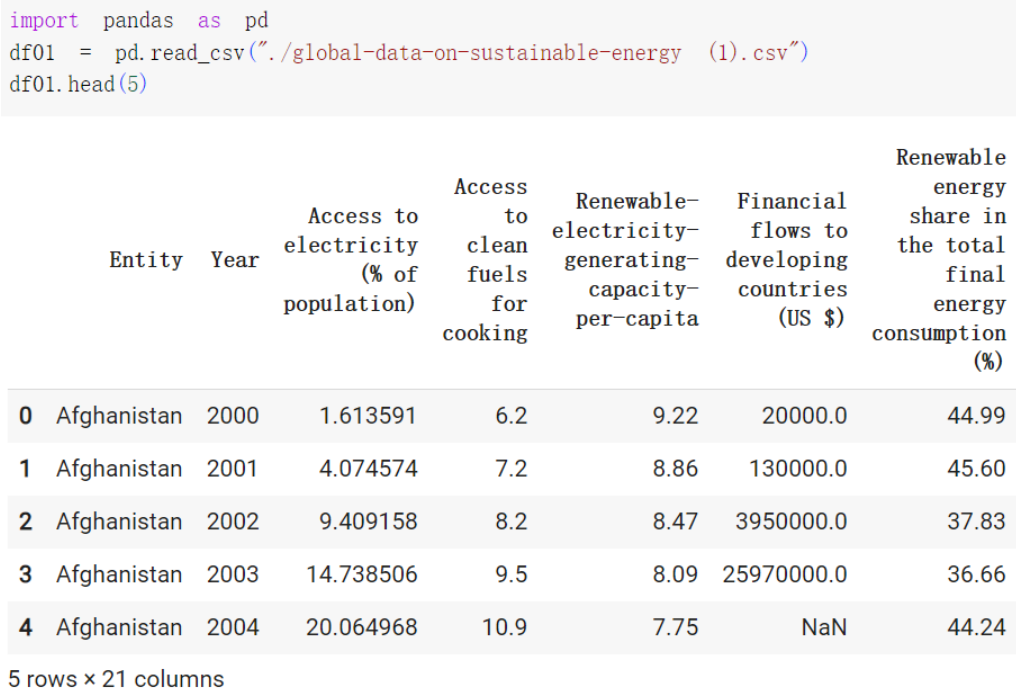


Figure 5. Import Data.

```
df01.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3649 entries, 0 to 3648
Data columns (total 21 columns):
#   Column                                                                 Non-Null Count  Dtype
---  -
0   Entity                                                                3649 non-null   object
1   Year                                                                  3649 non-null   int64
2   Access to electricity (% of population)                             3639 non-null   float64
3   Access to clean fuels for cooking                                   3480 non-null   float64
4   Renewable-electricity-generating-capacity-per-capita               2718 non-null   float64
5   Financial flows to developing countries (US $)                     1560 non-null   float64
6   Renewable energy share in the total final energy consumption (%)    3455 non-null   float64
7   Electricity from fossil fuels (TWh)                                 3628 non-null   float64
8   Electricity from nuclear (TWh)                                      3523 non-null   float64
9   Electricity from renewables (TWh)                                   3628 non-null   float64
10  Low-carbon electricity (% electricity)                             3607 non-null   float64
11  Primary energy consumption per capita (kWh/person)                 3649 non-null   float64
12  Energy intensity level of primary energy (MJ/$2017 PPP GDP)        3442 non-null   float64
13  Value_co2_emissions_kt_by_country                                  3221 non-null   float64
14  Renewables (% equivalent primary energy)                            1512 non-null   float64
15  gdp_growth                                                            3332 non-null   float64
16  gdp_per_capita                                                        3367 non-null   float64
17  Density\n(P/Km2)                                                     3648 non-null   object
18  Land Area(Km2)                                                       3648 non-null   float64
19  Latitude                                                             3648 non-null   float64
20  Longitude                                                            3648 non-null   float64
dtypes: float64(18), int64(1), object(2)
memory usage: 598.8+ KB
```

Figure 6. Data Information.

3.3 Approach and Methodology

3.3.1 Machine Learning Model Construction

The process of machine learning modelling and predicting data includes the following steps: first, collect relevant data, and then perform data preprocessing, such as cleaning, transforming, and processing missing values. Next, proceed with feature engineering, including feature selection and extraction, and then split the data into training and testing sets. Select the appropriate algorithm based on the type of problem and train the model using the training set. Evaluate the performance of the final model on the test set to ensure its accuracy and effectiveness (see Figure 7).

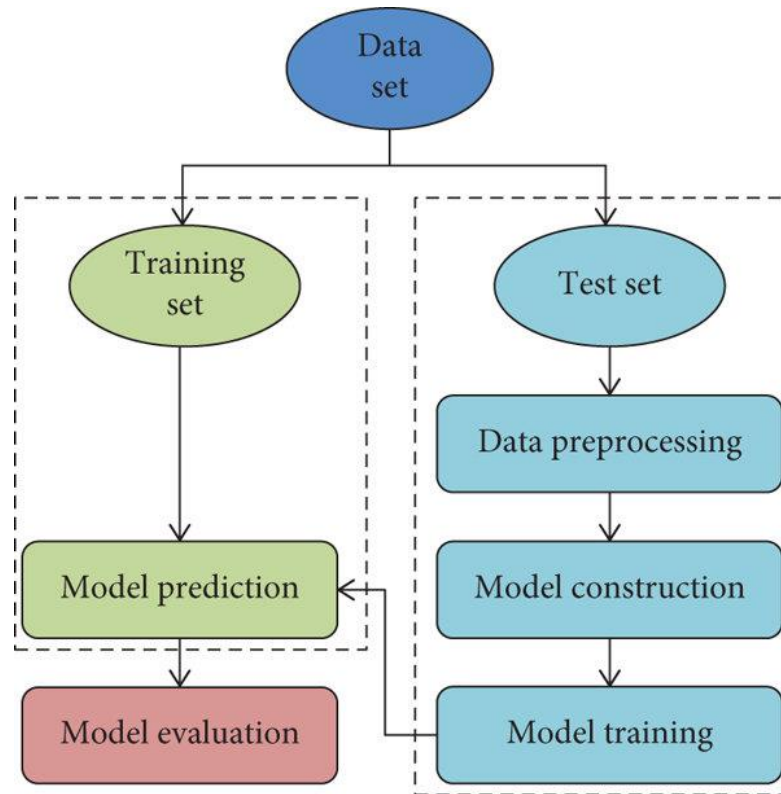


Figure 7. Machine learning model construction process (Zhang et al., 2022).

3.3.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an important step in the process of data science, which involves analyzing and drawing charts to extract information from data. The purpose of EDA is to understand the basic characteristics of data, identify patterns and anomalies in the data, and provide a basis for further modelling and analysis. In the EDA process, commonly used methods include statistical description (such as mean, median, standard deviation), data distribution analysis (such as histograms, box plots), and visualization of relationships between variables (such as scatter plots, correlation matrix heatmaps). Through these technologies, EDA can extract valuable information from raw data, guiding subsequent feature engineering and model selection.

3.3.3 Visualization Dashboard Design

This study will also construct a visual dashboard based on the information from the dataset. The purpose of building a dashboard is to provide a comprehensive and easily understandable visual tool to help stakeholders understand the status and trends of global energy sustainability. The visualization instrument will visualize energy data based on the various dimensions contained in the dataset, allowing users to analyze energy data from different perspectives and dimensions, including energy types, geographic locations, time trends, etc. The visualization dashboard will include interactive charts, making the data more vivid and intuitive.

This project plans to use the ECharts chart library, which is an open-source JavaScript data visualization library developed by Baidu. It provides a variety of chart types, including line charts, bar charts, pie charts, etc. It has flexible configuration options, supports interactive and dynamic effects, is compatible

across platforms, and has rich map functions. Its ease of use and scalability enable developers to quickly build various charts and are suitable for various data analysis and presentation scenarios. Figure 8 shows the initial draft of a visualization instrument built using Power BI.

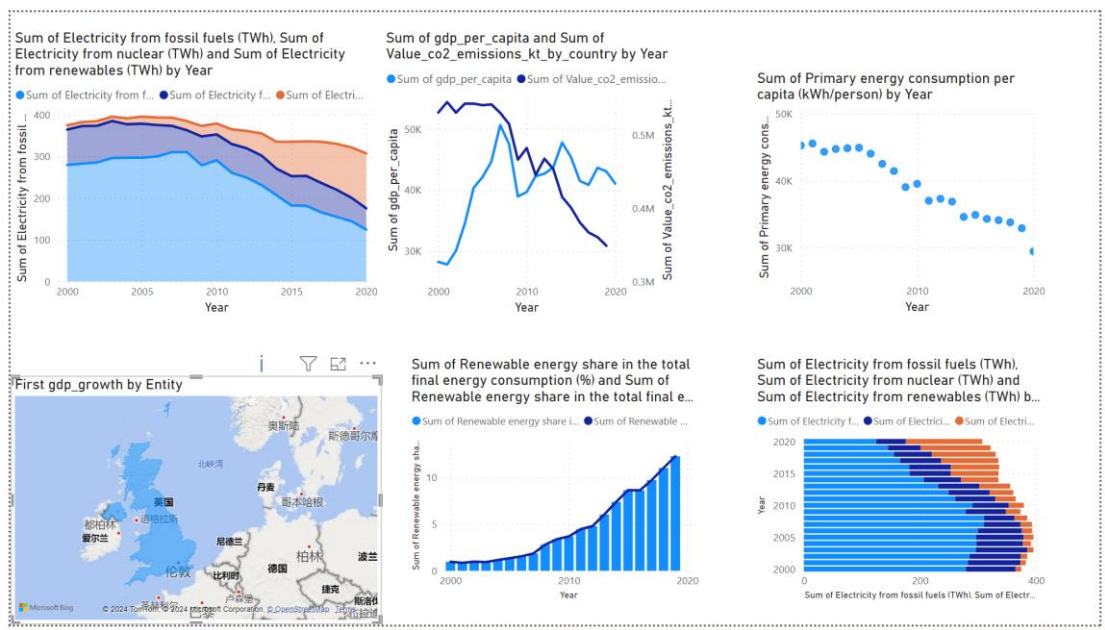


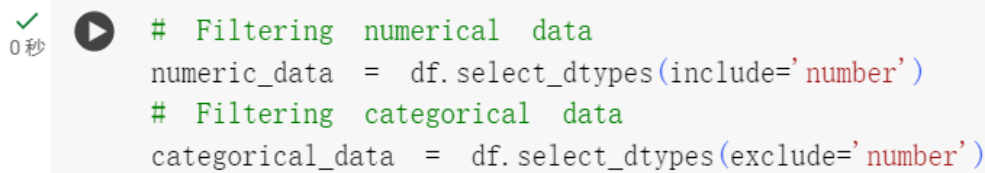
Figure 8. Visualization Dashboard Draft.

4 Implementation and Testing

4.1 Data Pre-processing

(1) Splitting Data

Because the data contains different categories of data, it is necessary to distinguish between numerical and categorical data before conducting data analysis. Using the 'select_dtypes()' method to set parameters for data classification(see Figure 9).

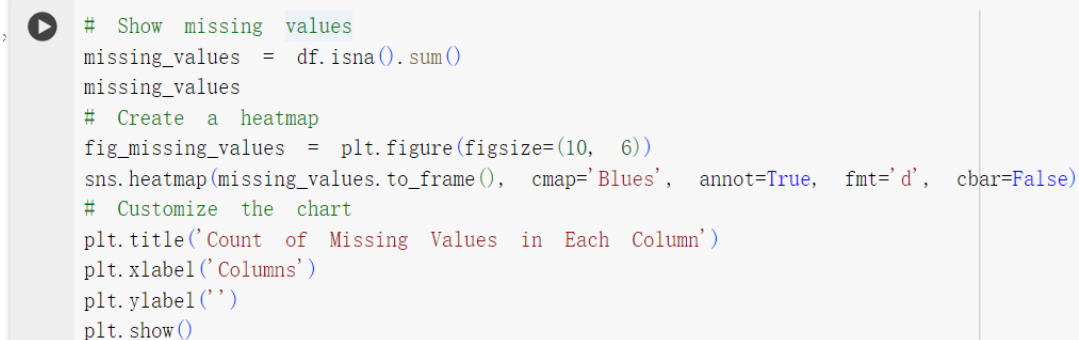
A code snippet in a Jupyter Notebook cell. It starts with a green checkmark and a play button icon, followed by '0 秒'. The code contains two comments: '# Filtering numerical data' and '# Filtering categorical data'. The first line of code is 'numeric_data = df.select_dtypes(include='number')'. The second line of code is 'categorical_data = df.select_dtypes(exclude='number')'.

```
# Filtering numerical data
numeric_data = df.select_dtypes(include='number')
# Filtering categorical data
categorical_data = df.select_dtypes(exclude='number')
```

Figure 9. Splitting Different Data Types.

(2) Managing Missing Values

This step involves identifying and correcting missing values in the data. As shown in Figure 10, the program first uses the 'isna()' method to identify missing values in each column of data, and calculates the total sum of missing values using the 'sum()' method. In order to express the missing values more intuitively, the program constructed a heatmap that reflects the overall situation of the missing values.

A code snippet in a Jupyter Notebook cell. It starts with a play button icon. The code contains several lines: a comment '# Show missing values', 'missing_values = df.isna().sum()', 'missing_values', a comment '# Create a heatmap', 'fig_missing_values = plt.figure(figsize=(10, 6))', 'sns.heatmap(missing_values.to_frame(), cmap='Blues', annot=True, fmt='d', cbar=False)', a comment '# Customize the chart', 'plt.title('Count of Missing Values in Each Column')', 'plt.xlabel('Columns')', 'plt.ylabel('')', and 'plt.show()'.

```
# Show missing values
missing_values = df.isna().sum()
missing_values
# Create a heatmap
fig_missing_values = plt.figure(figsize=(10, 6))
sns.heatmap(missing_values.to_frame(), cmap='Blues', annot=True, fmt='d', cbar=False)
# Customize the chart
plt.title('Count of Missing Values in Each Column')
plt.xlabel('Columns')
plt.ylabel('')
plt.show()
```

Figure 10. Splitting Different Data Types.

In Figure 11, it is evident that 'Financial flows to developing countries (US \$)', 'Renewables (% equivalent primary energy)' and 'Renewable-electricity-generating-capacity-per-capita' have too many missing values, reaching 2089, 2137 and 931 respectively. The presence of an excessive number of missing values renders the data unsuitable for filling and consequently affects the subsequent data analysis process.

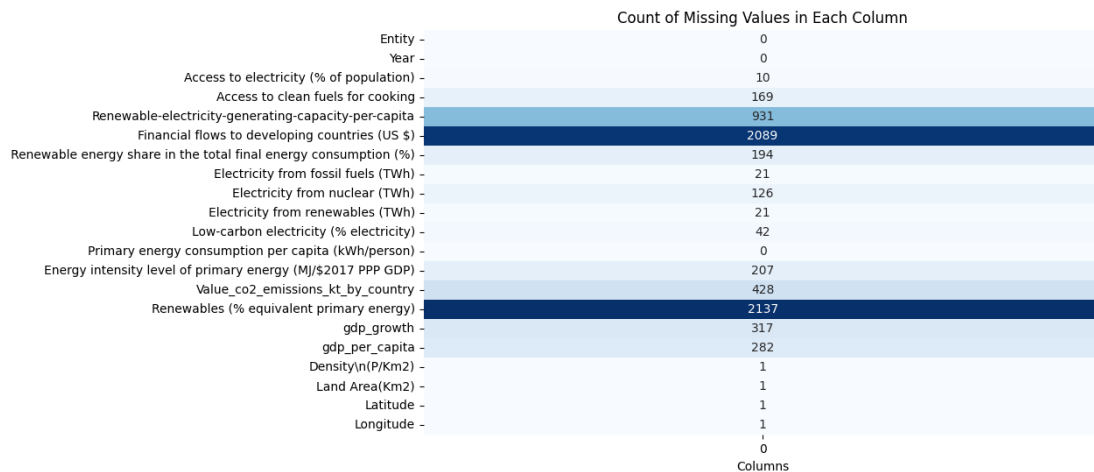


Figure 11. Heatmap of Missing Values.

In Figure 12, the program removes columns with too many missing values and uses the 'drop' method to delete specific columns. For other columns with fewer missing values, the program uses mean padding to minimize the impact on the results of data analysis.

```
[22] # Drop columns with a high number of missing values
df.drop(columns=['Financial flows to developing countries (US $)',
                 'Renewables (% equivalent primary energy)',
                 'Renewable-electricity-generating-capacity-per-capita'],
        inplace=True)

[23] # Fill missing values with mean
columns_to_fill_mean = ['Access to clean fuels for cooking',
                        'Renewable energy share in the total final energy consumption (%)',
                        'Electricity from nuclear (TWh)',
                        'Energy intensity level of primary energy (MJ/$2017 PPP GDP)',
                        'Value_co2_emissions_kt_by_country', 'gdp_growth', 'gdp_per_capita']
df[columns_to_fill_mean] = df[columns_to_fill_mean].apply(lambda x: x.fillna(x.mean()))

[24] # Drop remaining rows with missing values
df = df.dropna()
```

Figure 12. Using Deletion and Mean Padding to Handle Missing Values.

After processing the missing values, check the missing values again and it can be seen in Figure 13 that the missing values after processing are null.



	<code>df.isnull().sum()</code>	
	Entity	0
	Year	0
	Access to electricity (% of population)	0
	Access to clean fuels for cooking	0
	Renewable energy share in the total final energy consumption (%)	0
	Electricity from fossil fuels (TWh)	0
	Electricity from nuclear (TWh)	0
	Electricity from renewables (TWh)	0
	Low-carbon electricity (% electricity)	0
	Primary energy consumption per capita (kWh/person)	0
	Energy intensity level of primary energy (MJ/\$2017 PPP GDP)	0
	Value_co2_emissions_kt_by_country	0
	gdp_growth	0
	gdp_per_capita	0
	Density\n(P/Km2)	0
	Land Area(Km2)	0
	Latitude	0
	Longitude	0
	dtype: int64	

Figure 13. Check Missing Values.

(3) Checking Duplicate Rows

Checking for duplicate rows is a crucial step in data preprocessing, which helps improve data quality, avoid overfitting, improve computational efficiency, and provide a more reliable foundation for data analysis and modelling. As shown in Figure 14, there are no duplicate rows in this dataset.

```
# Check for duplicate rows
num_duplicates = df.duplicated().sum()
print("Number of Duplicate Rows:", num_duplicates)
```

Number of Duplicate Rows: 0

Figure 14. Check Missing Values.

(4) Managing Special Categories

In order to facilitate the next step of data visualization, the program converted some column names, extracted special names and areas of countries from the 'Entity' column, and processed the 'Land Area (Km2)' column with integers (see Figure 15).

```

] # Create a copy of the DataFrame to avoid SettingWithCopyWarning
df_copy = df.copy()
# Rename columns in the copied DataFrame
df_copy.rename(columns={"Value_co2_emissions_kt_by_country": "CO2", "Land Area(Km2)": "Land"}, inplace=True)
# Rename the 'Density' column
df.rename(columns={"Density\\n(P/Km2)": "Density"}, inplace=True)
# Convert 'Density' to string and then replace commas and convert to integer using .loc
df.loc[:, 'Density'] = df['Density'].astype(str).str.replace(',', '').astype(int)

] print(df.columns)

Index(['Entity', 'Year', 'Access to electricity (% of population)',
      'Access to clean fuels for cooking',
      'Renewable energy share in the total final energy consumption (%)',
      'Electricity from fossil fuels (TWh)', 'Electricity from nuclear (TWh)',
      'Electricity from renewables (TWh)',
      'Low-carbon electricity (% electricity)',
      'Primary energy consumption per capita (kWh/person)',
      'Energy intensity level of primary energy (MJ/$2017 PPP GDP)',
      'Value_co2_emissions_kt_by_country', 'gdp_growth', 'gdp_per_capita',
      'Density', 'Land Area(Km2)', 'Latitude', 'Longitude'],
      dtype='object')

```

Figure 15. Check Missing Values.

4.2 Exploratory Data Analysis (EDA)

This dataset contains multiple numerical data columns. Before starting to explore the data of individual columns, use the Matplotlib Pylot library to draw histograms of all data columns for an overview. Figure 16 shows the code implementation process.

```

[34] # Visualize histograms for each numerical column
df.hist(figsize=(40, 20))
plt.show()

```

Figure 16. Code Implementation of Histograms of Numerical Columns.

Figure 17 illustrates that the majority of countries have achieved near-universal electricity penetration, with the proportion of clean fuel use and the capacity for renewable energy generation varying considerably across the globe. Fossil fuel and nuclear power generation also show significant differences between countries, with the proportion of low-carbon electricity mostly concentrated between 0-50%. The per capita primary energy consumption and energy intensity level are mainly concentrated in a lower range, but a few countries have higher levels. The proportion of renewable energy in total terminal energy consumption and primary energy is relatively low, while the CO2 emissions per unit of GDP are higher in certain countries.

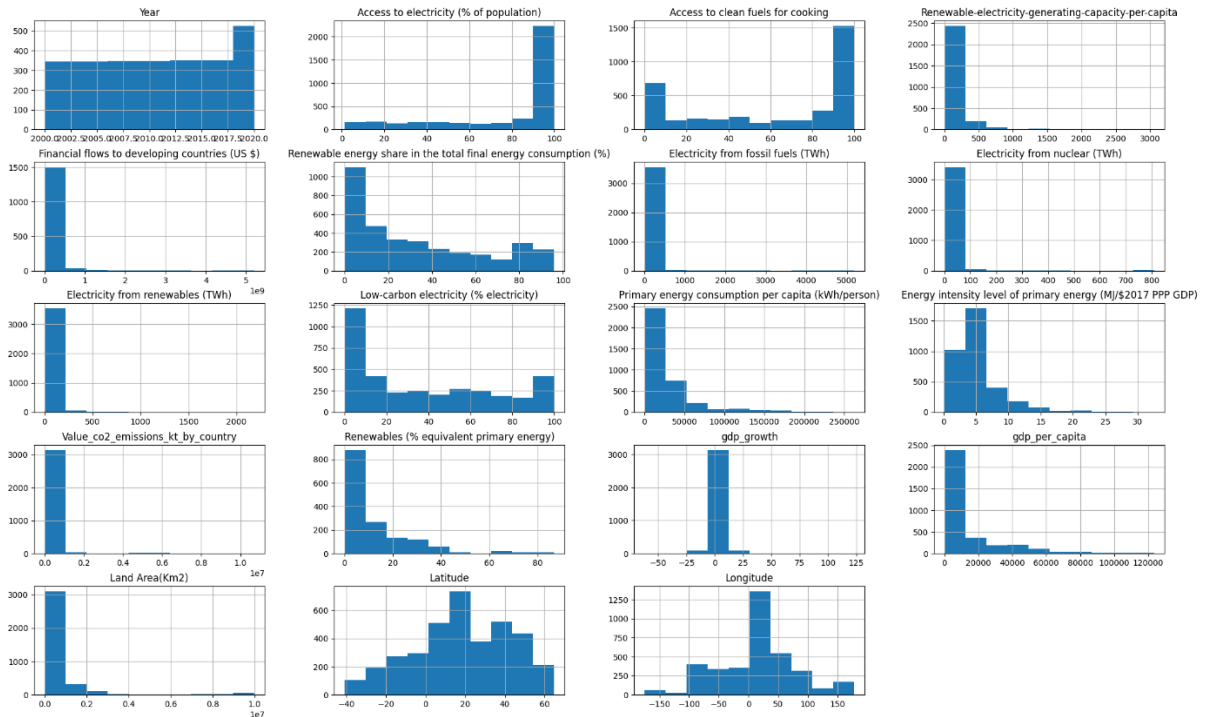


Figure 17. Histograms of Numerical Columns.

4.2.1 CO2 Emissions

(1) Top 10 Countries with Highest Average CO2 Emissions

Carbon emissions are an important indicator for measuring the progress of sustainable energy. Based on data distribution, this study calculated the average total carbon emissions of each country from 2000 to 2019 and used code to construct a bar chart of top ten countries with carbon emissions (see Figure18).

```
[ ] average_CO2_by_country = df.groupby('Entity')['CO2'].mean()
top_10_countries01 = average_CO2_by_country.nlargest(10)
plt.figure(figsize = (10, 6))
sns.barplot(x = top_10_countries01.index, y = top_10_countries01.values)
plt.xlabel('Country')
plt.ylabel('Average CO2 consumption per capita (kWh/person)')
plt.title('Top 10 Countries with Highest Average CO2')

plt.xticks(rotation = 90, ha = 'center')

plt.tight_layout()
```

Figure 18. Code Implementation of Bar chart of CO2 Consumption.

As shown in Figure 19, the bar chart shows the top ten countries with the highest average CO2 consumption. China and the United States are ranking first and second, significantly higher than third place India. The following countries include Japan, Germany, and Canada. The figure also shows the UK, Mexico, Indonesia, and Saudi Arabia, which have similar CO2 consumption levels. Overall, China and the United States are far ahead of other countries in CO2 consumption.

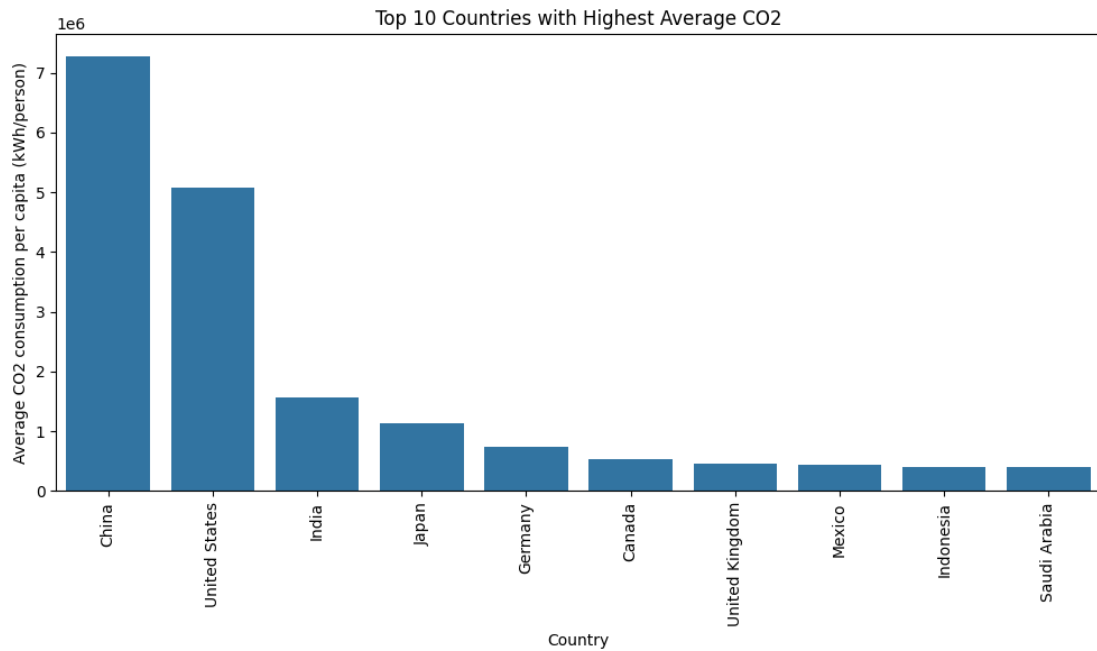


Figure 19. Bar chart of Top 10 Countries in CO2 Consumption.

(2) Top 10 Countries with Highest Average Primary Energy Consumption

Analyzing only based on the total carbon emissions is not comprehensive. The population, geographical environment, and economic development of each country are different. Therefore, as shown in Figure 20, the program drew a bar chart based on the per capita carbon emissions of each country to see if there is a difference in the ranking of the top 10 international and total carbon emissions.

```

average_primary_energy_by_country = df.groupby('Entity')['Primary energy consumption per capita (kWh/person)'].mean()
top_10_countries02 = average_primary_energy_by_country.nlargest(10)
plt.figure(figsize = (10, 6))
sns.barplot(x = top_10_countries02.index, y = top_10_countries02.values)

plt.ylabel('Average Primary energy consumption per capita (kWh/person)')
plt.title('Top 10 Countries with Highest Average Primary energy consumption')

plt.xticks(rotation = 90, ha = 'center')

plt.tight_layout()
plt.show()

```

Figure 20. Code Implementation of Bar chart of Top 10 Countries in Primary CO2 Emissions.

As shown in Figure 21, the bar chart shows the top ten countries with the highest per capita carbon dioxide consumption. Qatar ranks first at 215565.21 kWh per person, possibly due to its abundant oil and gas resources and small population base. Iceland and Bahrain both achieved over 150000 kWh/person, ranking second and third, followed closely by Singapore, United Arab Emirates, Trinidad and Tobago, Kuwait, Canada, Norway, and Luxembourg. Overall, the high per capita energy consumption of these countries reflects multiple factors such as their abundant resources, level of economic development, and climate conditions.

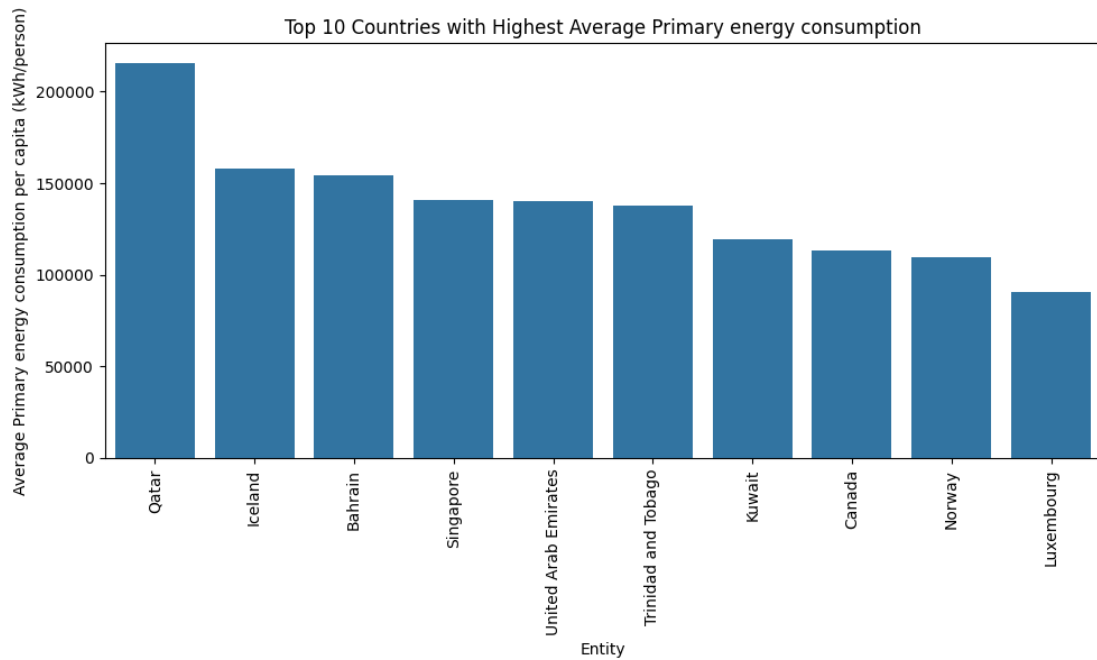


Figure 21. Bar chart of Top 10 Countries in Primary CO2 Emissions.

4.2.2 Correlation Matrix Heatmap

Correlation Matrix Heatmap is a data visualization tool used to display the correlation between multiple variables. As shown in Figure 22, a correlation matrix data frame based on dataset was constructed through coding. Using cool and warm colour tones to represent the positive and negative correlations of the data. Find the relevant relationships in the data column.

```
[36] numeric_df = df.select_dtypes(include=['float64', 'int64']) # Select only numeric columns
# Calculate the correlation matrix
correlation_matrix = numeric_df.corr()
```

```
[37] # Increase the figure size for a clearer heatmap
plt.figure(figsize=(14, 12))
# Visualize the correlation matrix using a heatmap
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5,
            cbar_kws={"shrink": .75}, xticklabels=correlation_matrix.columns,
            yticklabels=correlation_matrix.columns,
            annot_kws={"size": 8}, square=True)
plt.xticks(rotation=45, ha='right')
plt.title('Correlation Matrix Heatmap')
plt.show()
```

Figure 22. Histograms of Numerical Columns.

This correlation heatmap displays the correlation coefficients between multiple variables. The correlation coefficient ranges from -1 to 1, with positive values indicating positive correlation and negative values indicating negative correlation. The larger the absolute value, the stronger the correlation. In Figure 23, there is a strong positive correlation (0.86) between Access to Electricity and clean fuel usage, indicating that regions with high electricity access rates often have higher clean fuel usage rates. Fossil fuel power generation is positively correlated with nuclear and renewable energy power (0.65 and 0.85), and highly positively correlated with CO2 emissions (0.95), indicating that although multiple power

generation methods coexist, fossil fuel power generation remains the main source of CO2 emissions. It is interesting that Renewable energy in energy consumption% is negatively correlated with electricity access rate (-0.77) and clean fuel usage rate (-0.76), indicating that in areas with high electricity access rate and clean fuel usage rate, the proportion of renewable energy usage is actually lower.

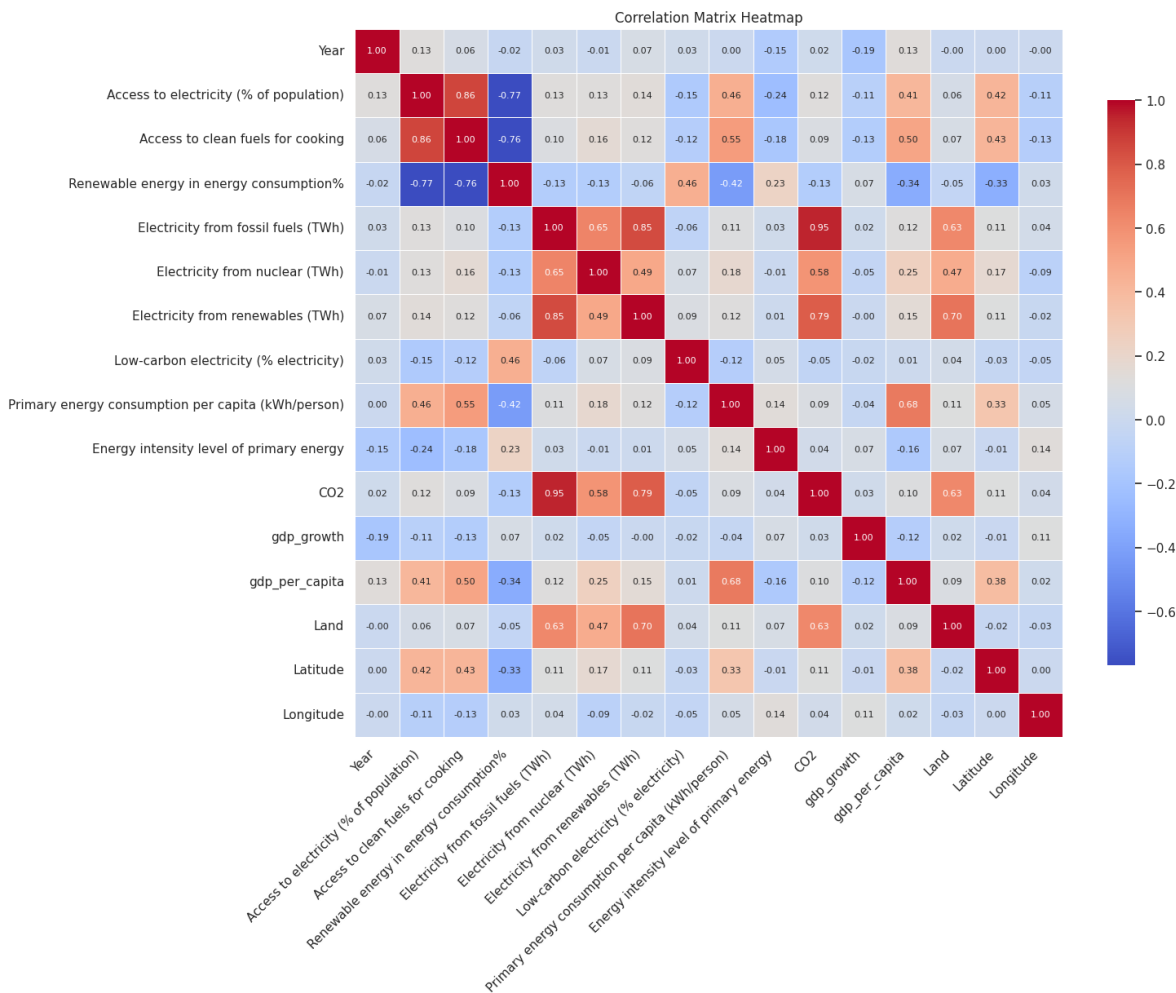


Figure 23. Histograms of Numerical Columns.

4.2.3 Trends In Electricity from Different Sources

The dataset contains data on three energy acquisition methods, including fossil fuel power generation, renewable energy power generation, and nuclear power generation. By summarizing data from all countries every year, the trend of changes in power generation by different methods and construct a multi-line chart for comparison can obtained (see Figure 24).

```

average_EFFF_by_year = df_copy01.groupby('Year')['Electricity from fossil fuels (TWh)'].mean()
average_EFN_by_year = df_copy01.groupby('Year')['Electricity from nuclear (TWh)'].mean()
average_EFR_by_year = df_copy01.groupby('Year')['Electricity from renewables (TWh)'].mean()

average_EFFF_by_year = average_EFFF_by_year.reset_index()
average_EFN_by_year = average_EFN_by_year.reset_index()
average_EFR_by_year = average_EFR_by_year.reset_index()

plt.figure(figsize = (10, 6))
sns.lineplot(data = average_EFFF_by_year, x = 'Year',
              y = 'Electricity from fossil fuels (TWh)',
              label = 'Electricity from Fossil Fuels')
sns.lineplot(data = average_EFN_by_year, x = 'Year',
              y = 'Electricity from nuclear (TWh)',
              label = 'Electricity from Nuclear', color = 'aqua')
sns.lineplot(data = average_EFR_by_year, x = 'Year',
              y = 'Electricity from renewables (TWh)',
              label = 'Electricity from Renewables', color = 'green')
plt.title('Average Growth of Energy from Different Sources Over the Years')
plt.xlabel('Year')
plt.ylabel('Energy from Different Sources (TWh)')

plt.xticks(average_EFFF_by_year['Year'], rotation = 0, ha = 'center')
plt.xlim(2000, 2019) #2020 doesn't contain data and will be predicted later

plt.tight_layout()
plt.show()

```

Figure 24. Building Line Chart.

From the line chart in Figure 25, since 2000, fossil fuel power generation has been dominant and has been on an upward trend, but the growth rate has slowed down. Although renewable energy generation lags behind fossil fuels in terms of quantity, it maintains a good upward trend and its growth rate is increasing year by year. Nuclear power generation began to decline around 2006 and showed a relatively weak upward trend starting in 2015.

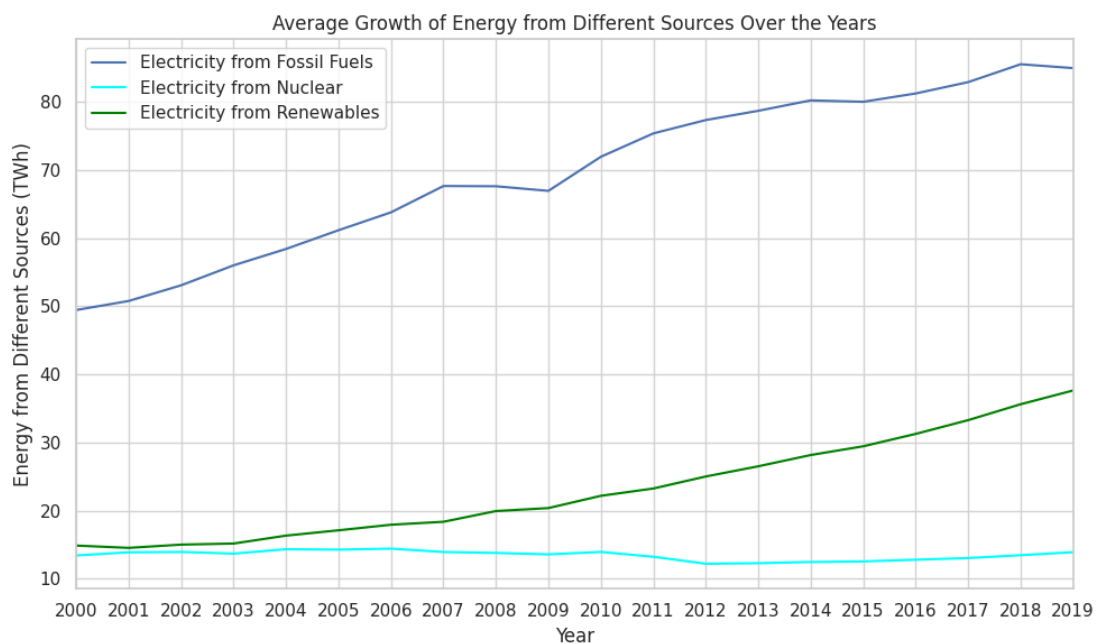


Figure 25. Average Growth of Energy from Different Sources Over the Years.

4.3 Building a Machine Learning Model for CO2 Emissions Prediction

(1) Feature Engineering

Select existing features to improve the performance of machine learning models. As shown in Figure 26, the program used a data feature selection tool from the Sklearn data analysis library, selecting twelve features from the dataset to construct predictions for per capita carbon dioxide emissions in various countries.

```
from sklearn.feature_selection import SelectKBest, f_regression
# Divide the dataset into features and labels
X = numeric_df.drop(['CO2'], axis=1)
y = numeric_df['CO2']

# Feature Selection
selector = SelectKBest(f_regression, k=12) #f_regression
X_new = selector.fit_transform(X, y)
selected_features = X.columns[selector.get_support()]
selected_features

Index(['Year', 'Access to electricity (% of population)',
      'Access to clean fuels for cooking',
      'Renewable energy in energy consumption%',
      'Electricity from fossil fuels (TWh)', 'Electricity from nuclear (TWh)',
      'Electricity from renewables (TWh)',
      'Low-carbon electricity (% electricity)',
      'Primary energy consumption per capita (kWh/person)',
      'Energy intensity level of primary energy', 'gdp_per_capita', 'Land'],
      dtype='object')
```

Figure 26. Feature Selection.

(2) Model Selection

In this study, it is necessary to purchase models based on numerical data to solve prediction problems and select appropriate machine learning algorithms. The following are the selected machine learning models:

- Random Forest

Random forest is an ensemble learning method that constructs multiple decision trees and combines their prediction results for classification or regression. Random forest has good robustness and high accuracy, suitable for processing high-dimensional data and large-scale datasets.

- Linear Regression

Linear regression is a statistical method used to establish and analyse relationships between continuous variables. Linear regression describes the relationship between the independent variable (feature) and the dependent variable (target variable) by fitting a linear model. Linear regression is suitable for situations where

the target variable is a continuous variable, such as house price prediction, sales volume prediction, etc.

- Gradient Boosting

Gradient Boosting regression is an ensemble learning method that constructs a powerful regression model by sequentially training multiple weak learners (usually decision trees). Gradient Boosting gradually improves the predictive performance of the model by iteratively fitting residuals, usually with high predictive accuracy.

- KNeighborsRegressor

K-nearest neighbor regression is an instance-based learning method that uses the average of nearest neighbors to predict new data. Suitable for situations where there is no clear pattern in data distribution, complex nonlinear relationships, or small sample sizes of data.

(3) Model Building

Splitting the data into training and testing sets is required before building the model. The training set is used to train the model, enabling it to learn patterns and relationships in the data, in order to fit the data as closely as possible. The test set is used to evaluate the performance of the model on unseen data and check if the model overfits the training data as shown in Figure 27.

```
[10] from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split (
          X_new, numeric_df['CO2'], test_size =0.2, random_state = 42)

[11] scaler = StandardScaler()
      X_train_scaled = scaler.fit_transform(X_train)
      X_test_scaled = scaler.transform(X_test)
```

Figure 27 Splitting Data.

After selecting the model required for this prediction task, import it into the relevant model library and establish a corresponding method to save the model. When importing the KNeighborsRegressor model, set the k value to 5 (see Figure 28).

```
✓ 0秒 [12] from sklearn.model_selection import train_test_split, GridSearchCV
      from sklearn.preprocessing import StandardScaler
      from sklearn.impute import SimpleImputer
      from sklearn.metrics import mean_squared_error, r2_score
      from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
      from sklearn.linear_model import LinearRegression
      from sklearn.neighbors import KNeighborsRegressor

✓ 0秒 [13] random_forest_model = RandomForestRegressor()
      gradient_boosting_model = GradientBoostingRegressor()
      linear_regression_model = LinearRegression()
      knn_regressor_model = KNeighborsRegressor(n_neighbors=5)
```

Figure 28. Feature Selection.

After importing the model, each model is trained using the previously segmented training dataset. Figures 29-32 indicate the successful training of the four models.

```
[14] random_forest_model.fit(X_train, y_train)
```



```
RandomForestRegressor  
RandomForestRegressor()
```

Figure 29. Random Forest Model.

```
[15] gradient_boosting_model.fit(X_train, y_train)
```



```
GradientBoostingRegressor  
GradientBoostingRegressor()
```

Figure 30. Gradient Boosting Model.

```
[16] linear_regression_model.fit(X_train, y_train)
```



```
LinearRegression  
LinearRegression()
```

Figure 31. Linear Regression Model.

```
[17] knn_regressor_model.fit(X_train, y_train)
```



```
KNeighborsRegressor  
KNeighborsRegressor()
```

Figure 32. KNeighborsRegressor Model.

(4) Model Evaluation

Evaluate trained models on separate validation datasets or through cross validation to evaluate generalization performance and identify potential overfitting or underfitting issues. Mean Squared Error (MSE) and R-squared (R2) are common metrics used to evaluate the performance of regression models. The MSE of the model should be as small as possible, and R2 should be as close as possible to 1, indicating better model performance. Figure 33 shows the process of statistical MSE and R2 values for each model.

```
[18] rforest_predictions = random_forest_model.predict(X_test_scaled)  
linreg_predictions = linear_regression_model.predict(X_test_scaled)  
gradboost_predictions = gradient_boosting_model.predict(X_test_scaled)  
KNeighbors_prediction = knn_regressor_model.predict(X_test_scaled)
```

```
[19] rf_mse = mean_squared_error(y_test, rforest_predictions)  
lr_mse = mean_squared_error(y_test, linreg_predictions)  
gb_mse = mean_squared_error(y_test, gradboost_predictions)  
knn_mse = mean_squared_error(y_test, KNeighbors_prediction)
```

```
[20] rf_r2 = r2_score(y_test, rforest_predictions)  
lr_r2 = r2_score(y_test, linreg_predictions)  
gb_r2 = r2_score(y_test, gradboost_predictions)  
knn_r2 = r2_score(y_test, KNeighbors_prediction)
```

Figure 33. Calculate the MSE and R2 of Four Models.

By evaluating four different regression models, their performance metrics on a certain dataset were obtained. As shown in Figure 34, the KNeighborsRegressor

model performs the best in terms of MSE, followed by Linear regression model and Gradient boosting model, while the random forest model performs the worst. However, in terms of coefficient of determination, the KNeighborsRegressor model also performed the best, reaching a high level of 0.971175. At the same time, the performance of the linear regression model and the gradient boosting model were also very close. Based on the results, it can be seen that the KNeighborsRegressor Model performs the best, and using this model for prediction is the best choice.

	Model	MSE	R-squared
0	Random Forest	1.795360e+10	0.964400
1	Linear Regression	2.279022e+10	0.954810
2	Gradient Boosting	3.054656e+10	0.939430
3	KNeighborsRegressor	1.453703e+10	0.971175

Figure 34. Evaluation Results of Four Models in MSE and R2.

4.4 Visualization Dashboard

Visualization dashboard is a form of applying data visualization technology to display on a large screen. By displaying data in the form of charts, maps, etc. on a large screen, it is possible to visually present information such as the trend and distribution of data, thereby improving the readability and comprehension of the data. Visualization dashboards are commonly used in monitoring, data display, decision analysis, and other fields. In the process of creating visualisation panels, ECharts is a commonly used tool for creating various interactive and customizable data visualization, thereby achieving intuitive display and analysis of data. ECharts is an open-source JavaScript-based visualization library developed by Baidu. It supports multiple types of charts, including line charts, bar charts, pie charts, scatter charts, and other special types, which can meet different data display needs. Figure 35 shows the data structure of the visualization dashboard constructed in this project, which includes files in HTML, JSON, and CSS formats.

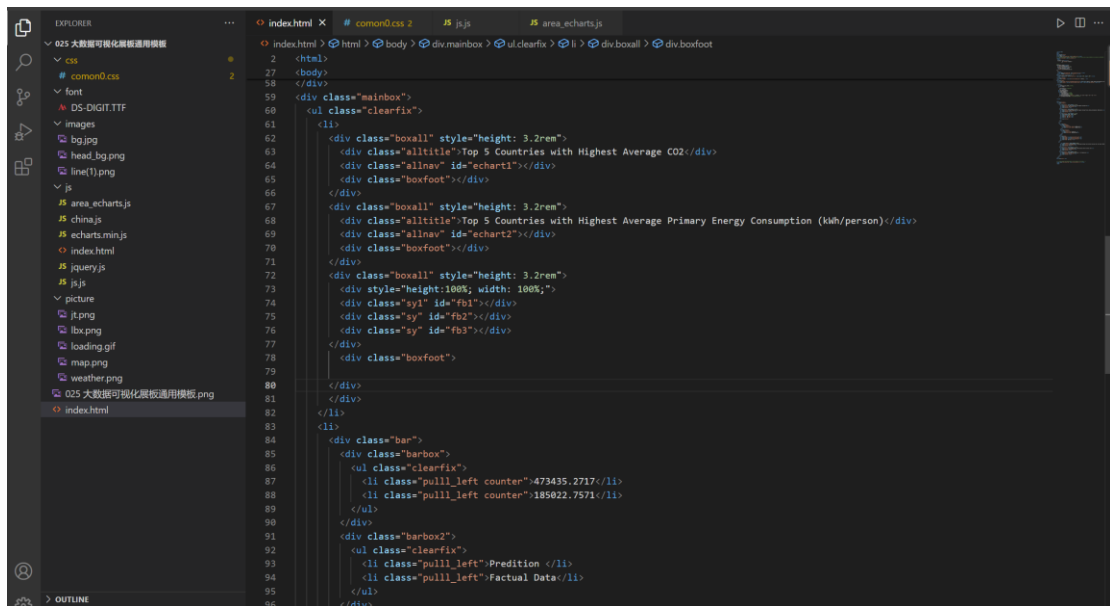


Figure 35. Structure of Visualization Dashboard.

Figure 36 shows the final presentation effect of the visualization dashboard on the web. The visual dashboard displays some of the conclusions in data analysis, including the top 5 countries with average carbon dioxide emissions, the top 5 countries with per capita carbon dioxide emissions, the trend of sustainable power generation year by year, and the proportion of population supplied by different countries. Charts have interactive functions, and as the mouse moves, they can display more specific information about the data.

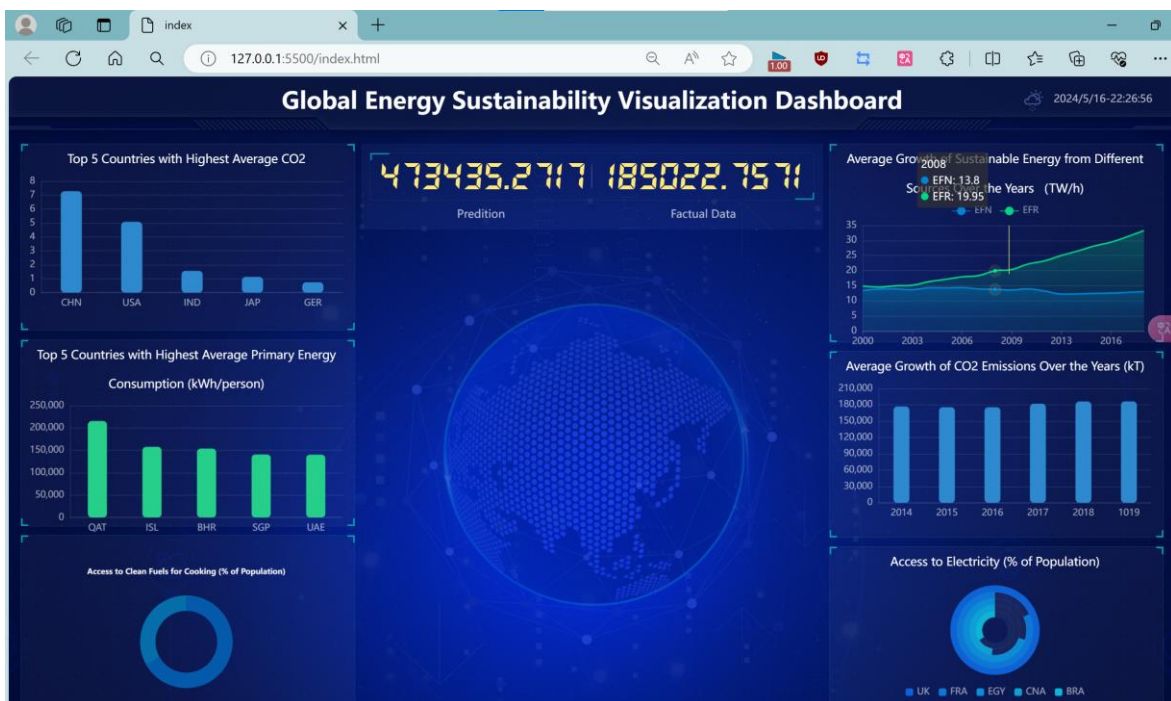


Figure 36. Visualization Dashboard Display.

5 Implementation and Testing

5.1 Discussion

(1) Profession Issues: This experiment was conducted on a 64-bit Windows 10 system running on an Intel (R) Core (TM) i7-10710U CPU and 16 GB RAM. The programming software used is Google Colab (a cloud-based Python programming environment provided by Google), with a Python version of 3.10.12(see Figure 37).

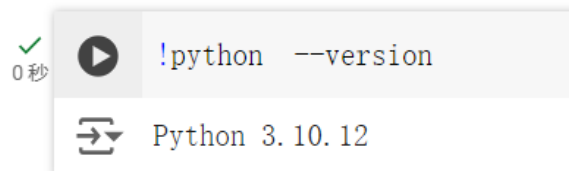


Figure 37. Python Version of Google Colab.

(2) Legal: The dataset used in this study is from a publicly available dataset on the Kaggle website and has the Attribute 4.0 International (CC BY 4.0) License. The dataset and project report are stored in a personal laptop with password and facial recognition protection. During the project research period, only the author used this laptop. The running code is saved in the Google Colab cloud and needs to be logged in with author Google account and password and has not been accessed by anyone else.

(3) Social: In the process of data analysis, the report found strong correlations between some key data. For stakeholders in related industries, it is recommended to focus on observing these data to provide data support for policy formulation or investment contributions. Through the display of visual dashboards, it is possible to have a more intuitive understanding of the current status of data related to sustainable energy, strengthen the understanding of sustainable energy among all sectors of society, and increase public attention to sustainable energy.

(4) Ethical: During the data processing, except for the necessary steps of preprocessing, no changes were made to the content of the data, indicating that the data content is authentic and trustworthy. In the process of training machine learning models, only the content from the dataset was used. The data of the visualization dashboard is also based on the data exploration and analysis content of this study.

5.2 Critical Evaluation

(1) Limitation

When this study attempt to use the KNeighborsRegressor model to predict carbon emissions data for 2020, while extracting actual carbon emissions data for 2019 for comparison. As shown in Figures 38-39, the predicted values are generally larger and more average, while there are significant data differences between each country in the actual data for 2019. Finally, by comparing the average predicted value for 2020 (473435.27 kT) with the average actual data for 2019 (185022.756 kT), it was found that there is a significant difference between the two values.


```
predicted_df = pd.DataFrame({'Predicted_CO2': predicted_df})
```

	Predicted_CO2
0	487139.308177
1	564174.011240
2	562458.007820
3	484407.301337
4	68552.000428
...	...
168	562458.007820
169	484407.301337
170	559836.010760
171	486531.300357
172	486531.300357

173 rows × 1 columns

后续步骤: [查看推荐的图表](#)

```
[27] predicted_df = pd.DataFrame({'Predicted_CO2': predicted_df})
predicted_mean = predicted_df['Predicted_CO2'].mean()
# print Average
print("Mean of Predicted_CO2:", predicted_mean)
```

Mean of Predicted_CO2: 473435.27166773105

Figure 38. Predicting Outcomes of 2020 CO2 Emissions.

```
selected_columns = ['CO2']
df_2019_CO2 = data_2019[selected_columns]
```

	CO2
19	6079.999924
40	4829.999924
61	171250.000000
82	25209.999080
103	519.999981
...	...
3563	116709.999100
3584	209.999993
3605	159866.462686
3626	6800.000191
3647	11760.000230

173 rows × 1 columns

后续步骤: [查看推荐的图表](#)

```
[31] data_2019mean = df_2019_CO2['CO2'].mean()
data_2019mean
```

185022.7570805759

Figure 39. Real Data on CO2 Emissions in 2019.

The significant difference between the predicted results and the actual data may be caused by multiple factors. Firstly, the model may be overfit on the training set, resulting in poor performance on new data. Secondly, there may be quality issues with the data itself. Due to the significant differences between countries and the fact that the dataset only records data from each country for 20 years. In addition, the KNeighborsRegressor model is sensitive to feature selection and data distribution. If inappropriate features are selected or data distribution is uneven, it may also lead to inaccurate prediction results.

(2) Future Development

Next, for the optimization of the model, the model parameters can be further adjusted. Research has also found that the model may require a higher quality and larger dataset, which can increase the span of the dataset in terms of years or make more accurate predictions for a single country. To improve the predictive performance of the model and reduce differences from actual data. For visual dashboards, it is possible to consider adding more interactive features, such as users entering country names and selecting the content they want to predict, the dashboard can call machine models to make predictions and output results.

6 Conclusions and Self-reflection

6.1 Conclusion

This project ultimately achieved the analysis of sustainable data and identified some trends and key information from the data. The operation of the visual dashboard is also very smooth, and it can effectively display all the content. For the machine learning module, building the model was successful and comparisons were made with different types of models, but the results were not satisfactory, and more solutions need to be explored. In the exploratory data analysis section, this study also made some extra visual charts. However, due to word limit, those charts are not possible to provide more explanation in the main text. The relevant charts will be placed in the attachment.

This project is very interesting and important, but unfortunately, the number of similar datasets is limited. Hopefully, future work will be able to access more relevant datasets and continue to try and deploy machine learning models to contribute more technical strength to the sustainable industry.

6.2 Self-reflection

I am pleased to report that the project was implemented without incident. The initial goal of the project has been met, with all content covered. However, it should be noted that machine learning prediction still requires a significant investment of time and effort. Over the past seven months, I have successfully completed five papers for the BSc (Hons) Computing with Data Science and Big Data Techn project. Additionally, I have completed my graduation project on deep learning image recognition at Beijing City University. This is an achievement I have never achieved before. In this process, I believe the biggest gain was learning self-management. In the first semester, I will complete my paper overnight and submit it only a few hours in advance. Later, I realized that this was not a sustainable way of working, and there was not much time to optimize and adjust the paper after completion. In the second semester, I have achieved the goal of completing the paper ahead of schedule and discussing it with the professor. I highly value everything this project has brought me, and it has also strengthened my determination to develop in the direction of sustainable development and artificial intelligence.

References

Lee, B. X., Kjaerulf, F., Turner, S., Cohen, L., Donnelly, P. D., Muggah, R., ... & Gilligan, J. (2016). Transforming our world: implementing the 2030 agenda through sustainable development goal indicators. *Journal of public health policy*, 37, 13-31.

Yadav, P. (2022) Global and Local Environmental Issues, *Ijrasnet Journal for Research in Applied Science and Engineering Technology*. Available at: <https://www.ijrasnet.com/research-paper/global-and-local-environmental-issues> (Accessed: 20 April 2024).

Aslakhanova, S.A., Amadayev, A.A. and Bakashev, E.D. (2021) 'World Power Relations in the modern context', *European Proceedings of Social and Behavioural Sciences [Preprint]*. doi:10.15405/epsbs.2021.11.18.

Allen, C., Metternicht, G. and Wiedmann, T. (2018) 'Initial progress in implementing the sustainable development goals (sdgs): A review of evidence from countries', *Sustainability Science*, 13(5), pp. 1453–1467. doi:10.1007/s11625-018-0572-3.

Griggs, D. et al. (2017) A guide to SDG interactions: From science to implementation. Paris: *International Council for Science*.

Anderson, C.C. et al. (2021) 'A systems model of SDG target influence on the 2030 Agenda for Sustainable Development', *Sustainability Science*, 17(4), pp. 1459–1472. doi:10.1007/s11625-021-01040-8.

Vincent, I. et al. (2021) 'The WASP model on the symbiotic strategy of renewable and nuclear power for the future of "renewable energy 3020" policy in South Korea', *Renewable Energy*, 172, pp. 929–940. doi:10.1016/j.renene.2021.03.094.

Bórawski, P. et al. (2019) Development of renewable energy market in the EU with particular regard to Solar Energy, *Conference Proceedings Determinants Of Regional Development*. Available at: <http://web.pwz.pila.pl/~pes/index.php/proceedings/article/view/162> (Accessed: 21 April 2024).

IEA (2021) World Total Energy Supply by source, 1971-2019 – charts – Data & Statistics, *IEA*. Available at: <https://www.iea.org/data-and-statistics/charts/world-total-energy-supply-by-source-1971-2019> (Accessed: 26 April 2024).

IEA (2024) Executive summary – renewables 2023 – analysis, *IEA*. Available at: <https://www.iea.org/reports/renewables-2023/executive-summary#abstract> (Accessed: 10 May 2024). IEA (2023) SDG7: Data and projections – analysis, *IEA*. Available at: <https://www.iea.org/reports/sdg7-data-and-projections> (Accessed: 26 April 2024).

Kupzog, F., King, R. and Stefan, M. (2020) 'The role of it in Energy Systems: The Digital Revolution as part of the problem or part of the solution', e & i

Elektrotechnik und Informationstechnik, 137(7), pp. 341–345. doi:10.1007/s00502-020-00818-5.

Wang, Y. and Huang, S. (2021) 'A decision analysis platform for energy big data based on Artificial Intelligence', *IOP Conference Series: Earth and Environmental Science*, 781(4), p. 042044. doi:10.1088/1755-1315/781/4/042044.

Ahmad, T. et al. (2021) 'Artificial Intelligence in sustainable energy industry: Status quo, challenges and opportunities', *Journal of Cleaner Production*, 289, p. 125834. doi:10.1016/j.jclepro.2021.125834.

Jeon, G. (2022) 'Artificial Intelligence Approaches for energies', *Energies*, 15(18), p. 6651. doi:10.3390/en15186651.

Dhanalakshmi, J. and Ayyanathan, N. (2021) 'A systematic review of big data in Energy Analytics using energy computing techniques', *Concurrency and Computation: Practice and Experience*, 34(4). doi:10.1002/cpe.6647.

Schneider, W.F. and Guo, H. (2018) 'Machine learning', *The Journal of Physical Chemistry B*, 122(4), pp. 1347–1347. doi:10.1021/acs.jpcb.8b00035.

Kumar, A., Upadhyay, P. and Kumar, A.S. (2020) 'Machine learning', *Fuzzy Machine Learning Algorithms for Remote Sensing Image Classification*, pp. 1–8. doi:10.1201/9780429340369-1.

Sarker, I.H. (2021) 'Machine learning: Algorithms, real-world applications and Research Directions', *SN Computer Science*, 2(3). doi:10.1007/s42979-021-00592-x.

Mathur, S. and Badone, A. (2019) 'A methodological study and analysis of machine learning algorithms', *International Journal of Advanced Technology and Engineering Exploration*, 6(51), pp. 45–49. doi:10.19101/ijatee.2019.650020.

Yao, Z. et al. (2022) 'Machine learning for a sustainable energy future', *Nature Reviews Materials*, 8(3), pp. 202–215. doi:10.1038/s41578-022-00490-5.

Bhatt, H. et al. (2023a) 'Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques', *Cleaner Chemical Engineering*, 5, p. 100095. doi:10.1016/j.clce.2023.100095.2017

Benti, N.E., Chaka, M.D. and Semie, A.G. (2023b) 'Forecasting renewable energy generation with machine learning and Deep learning: Current advances and future prospects', *Sustainability*, 15(9), p. 7087. doi:10.3390/su15097087.

Janković, R. et al. (2020) 'Machine learning models for ecological footprint prediction based on energy parameters', *Neural Computing and Applications*, 33(12), pp. 7073–7087. doi:10.1007/s00521-020-05476-4.

Matenga, Z. (2022) 'Assessment of Energy Market's progress towards Achieving Sustainable Development goal 7: A clustering approach', *Sustainable Energy Technologies and Assessments*, 52, p. 102224. doi:10.1016/j.seta.2022.102224.

Tanwar, A. (2023) Global Data on Sustainable Energy (2000-2020), *Kaggle*. Available at: <https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy/data> (Accessed: 12 May 2024).

Zhang, H., Yang, G. and Dong, A. (2022) 'Prediction model between serum vitamin D and neurological deficit in cerebral infarction patients based on machine learning', *Computational and Mathematical Methods in Medicine*, 2022, pp. 1–6. doi:10.1155/2022/2914484.

.

Appendix A – Extra EDA charts

(1) Dynamic World Map

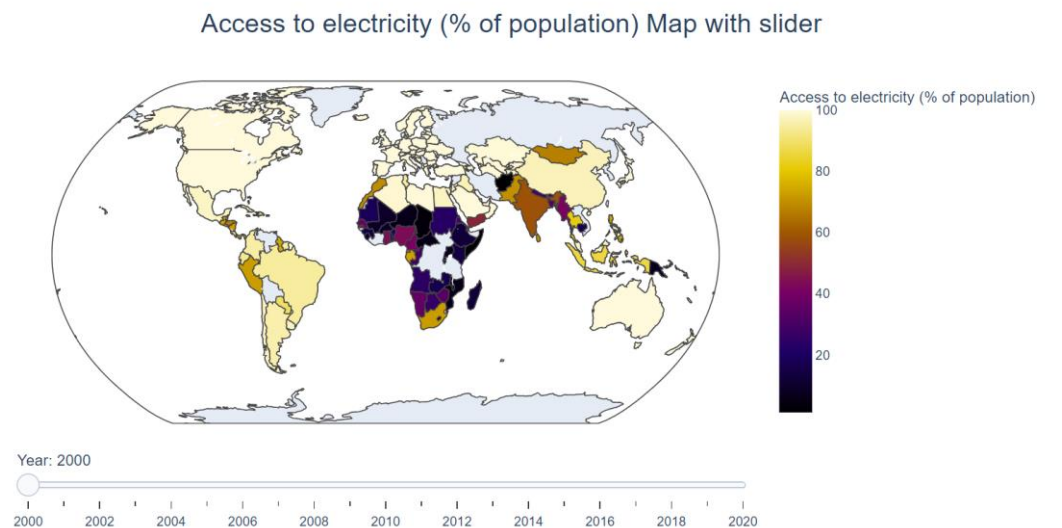


Figure 40. Access to electricity (% of population) in 2020.

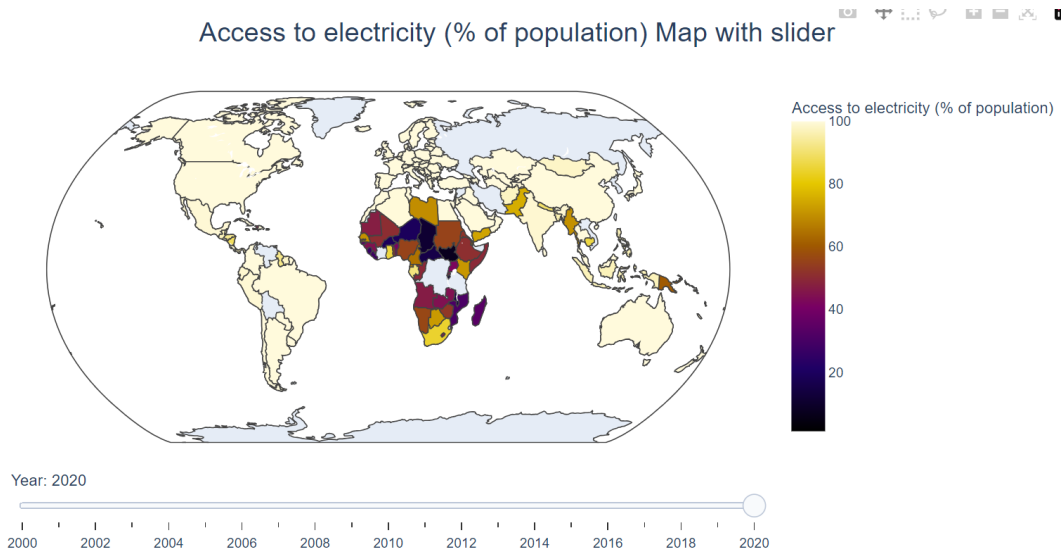


Figure 41. Access to electricity (% of population) in 2019.

Access to clean fuels for cooking Map with slider

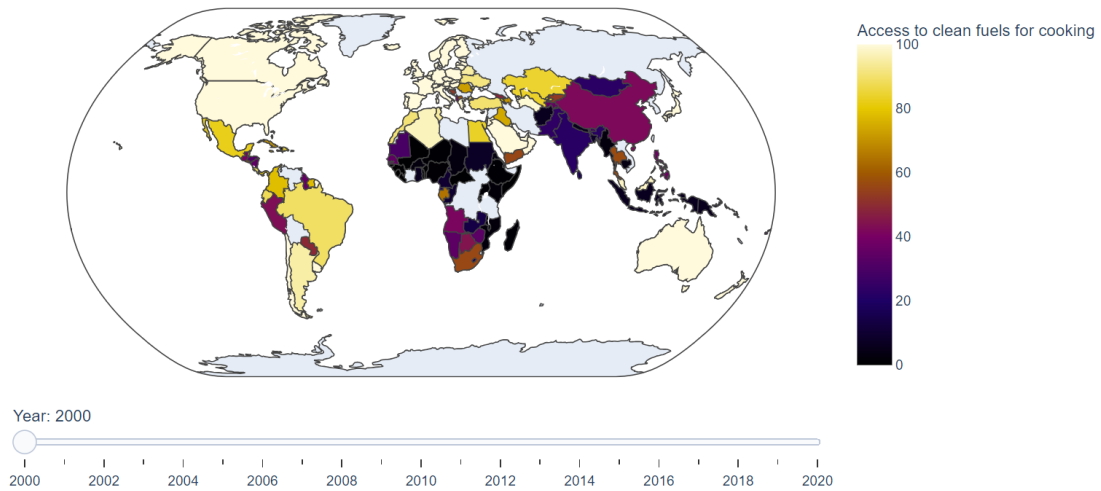


Figure 42. Access to clean fuels for cooking (% of population) in 2000.

Access to clean fuels for cooking Map with slider

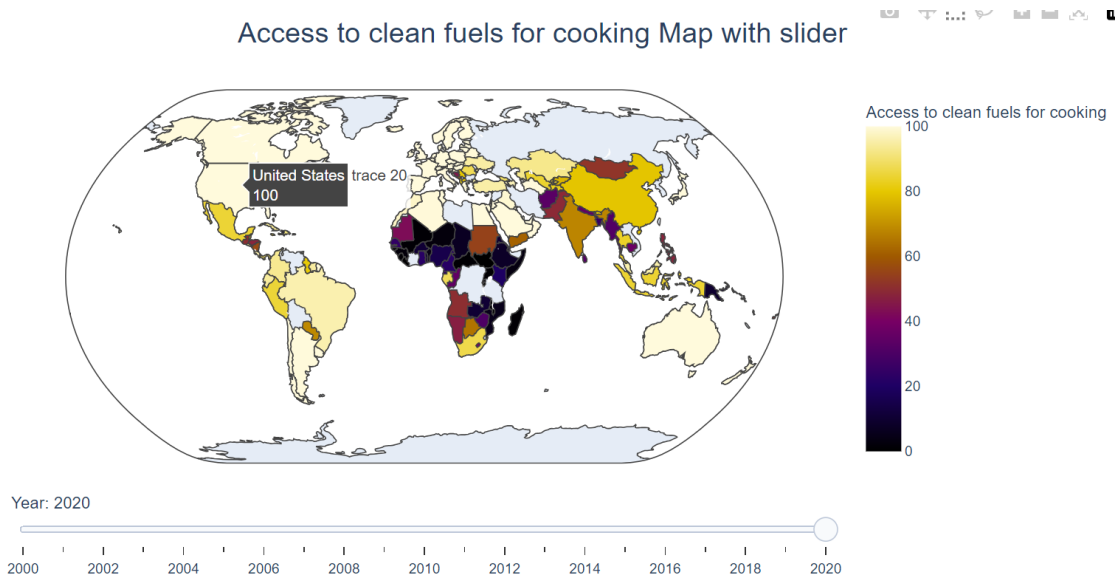


Figure 43. Access to clean fuels for cooking (% of population) in 2019.

(2) Scatter Diagram

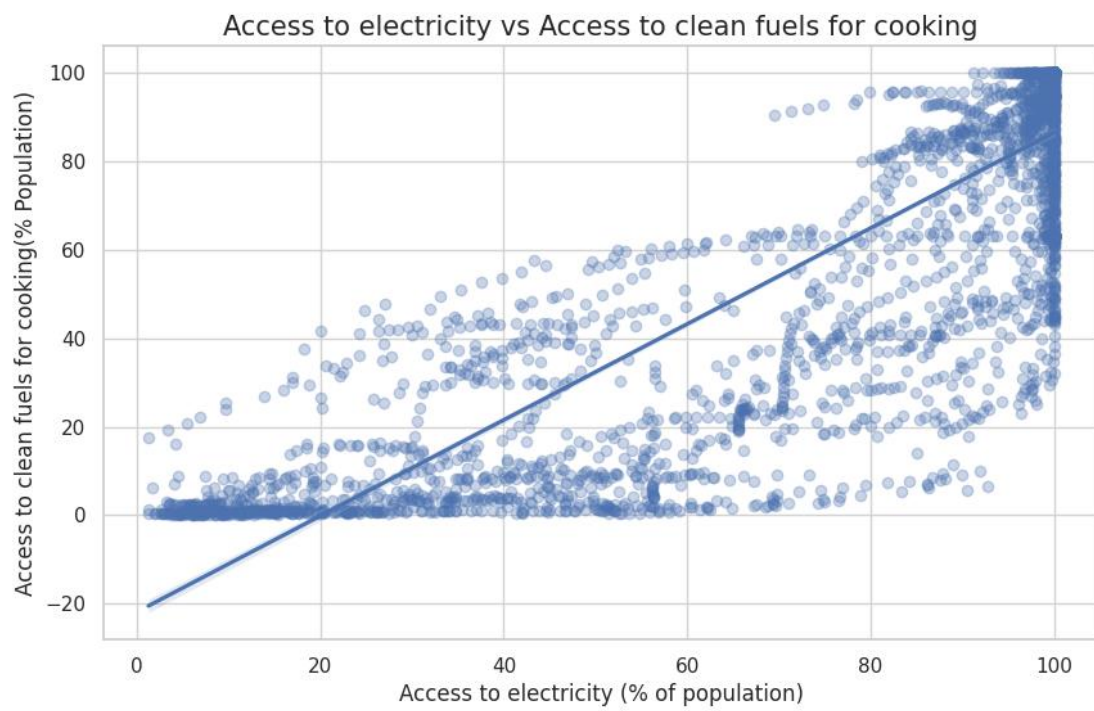


Figure 44. Access to electricity vs Access to clean fuels for cooking.

Appendix B – Data Column Names and Definition

Indicator	Definition
Entity	The name of the country or region for which the data is reported
Year	The year for which the data is reported, ranging from 2000 to 2020
Access to electricity	The percentage of population with access to electricity
Access to clean fuels for cooking	The percentage of the population with primary reliance on clean fuels
Renewable-electricity-generating-capacity-per-capita	Installed Renewable energy capacity per-person
Financial flows to developing countries	Aid and assistance from developed countries for clean energy projects (US \$)
Renewable energy share in total final energy consumption	Percentage of renewable energy in final energy consumption
Electricity from fossil fuels	Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours (TWh)
Electricity from nuclear	Electricity generated from nuclear power in terawatt-hours (TWh)
Electricity from renewables	Electricity generated from renewable sources (hydro, solar, wind, etc) in terawatt-hours (TWh)
Low-carbon electricity	Percentage of electricity from low-carbon sources (nuclear and renewables)
Primary energy consumption per capita	Energy consumption per person in kilowatt-hours (kWh/person)
Energy intensity level of primary energy	Energy use per unit of GDP at purchasing power parity (MJ/\$2011 PPP GDP)
Value_co2_emissions	Carbon dioxide emissions per person in metric tons (metric tons per capita)
Renewables	Equivalent primary energy that is derived from renewable sources (% equivalent primary energy)
GDP growth	Annual GDP growth rate based on constant local currency
GDP per capita	Gross domestic product per person
Density	Population density in persons per square kilometer (P/Km ²)
Land Area	Total land area in square kilometers (Km ²)
Latitude	Latitude of the country's centroid in decimal degrees
Longitude	Longitude of the country's centroid in decimal degrees

Appendix C – Ethical Approval

Student Project Approval Form

LD6053/ UG Computing Project: Student Project Approval Form

You should use this document if you intend to use one of the existing module level approval ethics applications. Please complete this document and discuss your study with your supervisor before you collect any data. *Failure to complete this document and have all aspects signed off and approved by your supervisor risks a notable deduction in your grade and may risk a case of Academic misconduct. Please see the module Bb site for more details.*

Supervisor sign off	
Ethics form complete	<input checked="" type="checkbox"/>
Ethical concerns acknowledged	<input checked="" type="checkbox"/>
Research tool(s) checked	<input checked="" type="checkbox"/>
All relevant forms included (consent etc.)	<input checked="" type="checkbox"/>
Is not high risk	<input checked="" type="checkbox"/>

Please ensure that your project meets the conditions of the existing ethics application (available on Module Bb site). ***If it does not, then you will need to submit a full ethics application instead.***

Student Name:	Yingxuan Zhang
Project Title:	Analysis and Predicting Global Energy Sustainability Using Machine Learning
Supervisor Name:	Syed Raza
Ethics application you are amending (check box):	<input type="checkbox"/> Low-risk Lab-based research <input checked="" type="checkbox"/> Low Risk Secondary Data Science project <input type="checkbox"/> Medium Risk Secondary Data Science project from the private domain required membership <input type="checkbox"/> Questionnaire/ survey Study <input type="checkbox"/> Interview Study or other Usability Study

Introduction to the project: *Treat like an introduction to the study. Why is your proposed study important? What has already been done on the topic? How does your proposed study ‘fit’ with the current literature and what does it add? What is the aim of the proposed study? Make reference to appropriate studies.*

Sustainable development is a hot topic today and research on ESG (Environment, Society, Governance) is still booming. This project aims to analyse and forecast publicly available datasets on ESG issues for each country from a country perspective by constructing data analysis methods such as regression analysis models and machine learning models. The ESG progress of each category of countries will be analysed from the distilled key indicators. This research helps to visualize the baseline status of countries in terms of ESG, and also explores further ways to address ESG issues using advanced techniques in computer technology.

Methodology: Please complete the table below, using the following info to guide you. Write this as a future tense method. Describe the **participants** that you will recruit, how many you are going to recruit, and indicate if you have any additional exclusion criteria. Include the **research design** (e.g. randomised/repeated measures/quantitative/qualitative/case study etc) and detail of your proposed **procedures** (i.e., how are you collecting the data?). Include information on all of the equipment you plan to use. If this is a low-risk study, outline how you will extract data and list the criteria you will use to do this. Somebody should be able to read this and replicate it. Describe all planned **data analysis** for both quantitative (e.g. t-tests, ANOVA, correlation etc.) and qualitative (content analysis, thematic analysis etc.) data. If doing a low-risk study explain how you intend to analyse the data you have collected. Use literature to justify your method.

1. Is this a low-risk secondary data or lab-based study? If Yes please go to questions 6 and 7.	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO
2. Who are your participants and what is the inclusion criteria?	This is a personal assignment and does not involve anyone else.
3. How many will you recruit and from where?	This is a personal assignment and does not involve anyone else.
4. Are there any exclusion criteria (reasons why people should not participate)?	This is a personal assignment and does not involve anyone else.
5. Research design:	<p>(1) Data Cleaning and Preprocessing: Clean the data to address issues like missing values, duplicates, outliers, and inconsistencies. Preprocess the data by standardizing variables, encoding categorical variables, and scaling numerical features as necessary.</p> <p>(2) Exploratory Data Analysis (EDA): Explore the data through descriptive statistics, data visualization, and graphical analysis. Identify patterns, trends, correlations, and insights that can inform subsequent analysis. Creating an overview of the dataset. Use multiple icons to mine data and try to interpret some data trends and phenomena.</p> <p>(3) Feature Engineering: Create new features or transform existing features to improve the performance of machine learning models. Using correlation and confusion matrix to find the most</p>

	<p>important value. Explain the concept of correlation and confusion matrix.</p> <p>(4) Model Selection: Choose appropriate machine learning algorithms or statistical techniques based on the nature of the problem, the available data, and the project objectives. Consider factors like interpretability, accuracy, scalability, and computational efficiency.</p> <p>(5) Model Training: Train the selected models on the training dataset using appropriate algorithms and techniques. Optimize model hyperparameters and evaluate model performance using suitable metrics like accuracy, precision, recall, F1-score, or ROC AUC. Explain the concept of suitable metrics.</p> <p>(6) Model Evaluation: Evaluate the trained models on a separate validation dataset or through cross-validation to assess generalization performance and identify potential overfitting or underfitting issues.</p> <p>(7) Visualization: Present the findings and insights from the data analysis in a clear, concise, and visually appealing manner. Use tables, charts, graphs, dashboards, or interactive visualizations to communicate complex information effectively to stakeholders.</p>
6. Procedures (describe what you will do to collect data, include all equipment/methods you plan to use).	The data for this project can be download from the Kaggle website. They are all published dataset.
7. Data analysis methods:	Correlation matrix, Decision tree regression model.
8. Additional information:	<p>here is the link:</p> <p>https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy/code</p>

Health and Safety: *Relevant risk assessments are listed in the ethics application. If your project needs additional risk assessments, then you will need to submit a new ethics application. Please identify the elements of the listed risk assessment that are relevant for your study and the risk assessment(s) you are working with.*

Please check the relevant boxes*:

- ☐ HL_RISK_173 Testing in an external environment
- ☐ HL_RISK_722 face to face interview
- ☐ HL_RISK_727 Group interview

Areas of potential risk <i>Please indicate how you will eliminate, or as a minimum ameliorate, the following areas of potential risks throughout the processes of research design, data generation, data analysis and dissemination</i>		
Area of risk	Questions relating to this risk	How will you mitigate against this risk?
Avoiding harm to all involved in or potentially affected by the research	How will you ensure that your participants/ respondents come to no harm (psychological; emotional; physical). e.g. not subjecting them to questioning about sensitive issues without advance agreement?	In this study, the data utilized is sourced from the Kaggle dataset titled "Global Data on Sustainable Energy (2000-2020)." This dataset involves aggregated and anonymized information related to sustainable energy, and no direct interaction with human participants is involved in the collection process.
	How will you ensure your own safety (beyond just physical) in undertaking the Enquiry?	As my research solely involves utilizing the "Global Data on Sustainable Energy (2000-2020)" dataset for data analysis and constructing a decision tree model, I do not conduct on-site investigations or engage in direct interactions with human participants. Therefore, personal safety considerations, beyond those related to data privacy and information security, are not applicable to this study.
Ensuring the anonymity of all participants/respondents	How will you ensure anonymity in collecting/generating data	Not involving other participants
	How will you ensure anonymity in reporting the data?	Not involving other participants
Gaining informed consent from all participants / respondents	How will you ensure respondent/participant consent in advance? You should provide a copy of the	Not involving other participants



	necessary consent form/s with this document	
	(How) might participants/respondents be able to withdraw their data?	Not involving other participants
Avoiding deception	How will you how you promote accuracy in recording, analysis, reporting of the data/findings?	<p>Prioritize Data Cleaning: Thoroughly clean and integrate the dataset to address inconsistencies and ensure a reliable foundation.</p> <p>Transparent Data Virtualization: Document and ensure transparency in the data virtualization process to accurately represent the original dataset.</p> <p>Rigorous Model Validation: Employ cross-validation and holdout datasets to rigorously validate the accuracy and transparency and address any discrepancies promptly.</p>
Data storage and destruction	How will you transport and store your data securely (e.g. password protected; cloud storage)	Save with a personal computer password
	How will you destroy the data and when?	There is no need to delete data
Secondary data sets	<i>Is your data set(s) from a domain requires membership?</i>	No
	<i>Does this data set can be used for educational or academic research purpose?</i>	Yes

Please check this box after you have read and understood ethics and health and safety information.

☒ I confirm I have read the University's health and safety policy and ethics policy. I have read and understood the requirement for the mandatory completion of risk assessments and that my study does not deviate from the module level approval ethics forms on Blackboard.

Further information (add below, if applicable)

- Consent forms
- Participant information sheet
- Debrief form
- Recruitment materials
- Permission letters
- Data collection tools

Student's Name and sign 	Date 11/3/2024
<hr/> (Name)	
Supervisor's name and sign 	Date 06/05/2024
<hr/> (Name) Syed Aqeel Raza	