Comprehensive Analysis of Global Energy Sustainability

Laxmi priya Kotha
Likitha Reddy Kotla
Laxmikanth Reddy Jitta
Siddardh Reddy Garlapati
CS652 Data Science Architecture
Sacred Heart University
Fairfield, CT USA

Abstract

This paper presents a comprehensive study on global energy sustainability trends, correlations, and predictive modeling. The research utilizes a diverse dataset encompassing economic indicators, energy consumption metrics, and environmental factors across multiple countries. Initial analyses include exploratory data visualizations such as choropleth maps and time series plots to highlight regional disparities and trends in energy access, consumption, and CO2 emissions. Correlation matrices and regression analyses further examine relationships between key variables, revealing insights into factors influencing energy usage and environmental impact. Subsequently, machine learning models—Gradient Boosting, Ada Boosting, Support Vector, and Random Forest—are applied to predict primary energy consumption per capita, demonstrating the efficacy of different approaches in modeling and forecasting energy demands. Evaluation metrics such as R2 score, mean absolute error (MAE), and root mean squared error (RMSE) assess model performance, identifying Random Forest as the top-performing model. This study contributes valuable insights into sustainable energy planning and policy formulation, emphasizing data-driven approaches for addressing global energy challenges.

Keywords

Global energy sustainability, economic indicators, energy consumption metrics, environmental factors, choropleth maps, time series plots, correlation matrices, regression analysis, machine learning models, Gradient Boosting, Ada Boosting, Support Vector Regression, Random Forest Regression, R2 score, mean absolute error (MAE), root mean squared error (RMSE), sustainable energy planning, policy formulation

I. INTRODUCTION

The pursuit of sustainable energy practices is paramount in addressing global challenges such as climate change and socio-economic development. This paper investigates the complex interplay between economic indicators, energy consumption patterns, and environmental impacts across diverse global contexts. The study utilizes a rich dataset encompassing variables ranging from GDP per capita and primary energy consumption to access to clean fuels and CO2 emissions, providing a comprehensive overview of global energy trends.

The motivation behind this research stems from the critical need to understand how various factors influence energy consumption and environmental sustainability on a global scale. By leveraging advanced data analytics techniques, including exploratory data analysis (EDA), correlation analysis, and machine learning modeling, this study aims to uncover significant relationships and predictive insights.

The methodologies employed include the visualization of spatial and temporal trends using interactive maps and time series plots, highlighting disparities in energy access and consumption across different regions. Correlation matrices and regression analyses further delve into the relationships between economic growth, energy consumption behaviors, and environmental impacts, shedding light on critical factors influencing energy policy and planning.

Moreover, the application of machine learning algorithms—such as Gradient Boosting, Ada Boosting, Support Vector, and Random Forest—aims to predict primary energy consumption per capita. Evaluation metrics such as R2 score, mean absolute error (MAE), and root mean squared error (RMSE) assess the efficacy of these models, providing insights into their suitability for forecasting future energy demands.

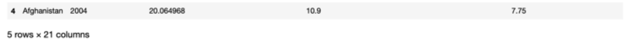
Ultimately, this study contributes valuable insights into sustainable energy strategies and informs policymakers, researchers, and practitioners about data-driven approaches to address global energy challenges effectively.

II. METHODOLOGY

1. Dataset Description

The dataset used in this study consists of 3649 entries spread across 21 columns, capturing a wide array of global socio-economic and energy-related metrics. It encompasses essential indicators such as access to electricity and clean cooking fuels, renewable energy capacity and consumption shares, and electricity generation from fossil fuels, nuclear sources, and renewables. Moreover, it includes critical environmental metrics like CO2 emissions and the percentage of renewable energy in the total energy mix. Economic dimensions such as GDP growth, GDP per capita, and demographic data such as population density and geographical coordinates (latitude and longitude) are also incorporated, offering a comprehensive view of global energy trends and economic characteristics.

This dataset is instrumental for conducting in-depth analyses and developing predictive models that explore the relationships between energy accessibility, consumption patterns, economic development, and environmental sustainability across countries and over time. By examining the interplay between energy policies, economic factors, and environmental impacts on a global scale, this dataset facilitates a nuanced understanding of the challenges and opportunities in achieving sustainable development goals.



	Entity	Year	Access_to_electricity_(% of population)	Access_to_clean_fuels_for_cooking	Renewable_electricity_generating_capacity_per_capita	Financial_flows_to_devel
0	Afghanistan	2000	1.613591	6.2	9.22	
1	Afghanistan	2001	4.074574	7.2	8.86	
2	Afghanistan	2002	9.409158	8.2	8.47	
3	Afghanistan	2003	14.738506	9.5	8.09	
4	Afghanistan	2004	20.064968	10.9	7.75	

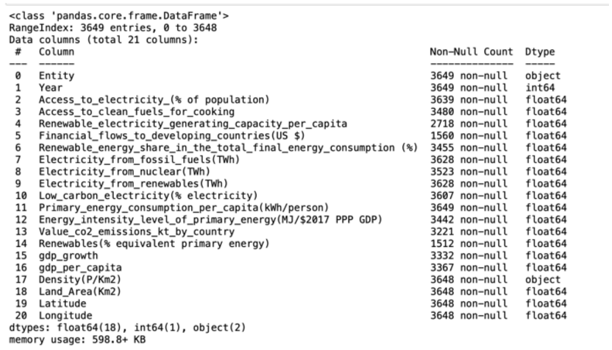
5 rows × 21 columns

Figure - 1 Sample Dataset

The above fig 1 shows the sample data. The dataset's extensive coverage and diverse metrics make it a valuable resource for interdisciplinary research spanning energy, economics, and environmental studies. Its inclusion of geographical and socio-economic dimensions enables researchers to explore regional variations in energy consumption, renewable energy adoption rates, and economic impacts. Moreover, the dataset supports comparative analyses between countries and regions, providing a foundation for evidence-based policy recommendations and strategic interventions to enhance energy security, mitigate environmental degradation, and foster inclusive economic growth.

2. *Dataset Preprocessing*

Data Changing Data Types: To ensure consistent data types for analysis, the column 'Density(P/Km2)' was converted from a string format (with commas) to a numeric format. Initially, the commas were removed, and then the column was converted to an integer type. This step helped standardize the data for further processing.
Figure-2 shows the column description.



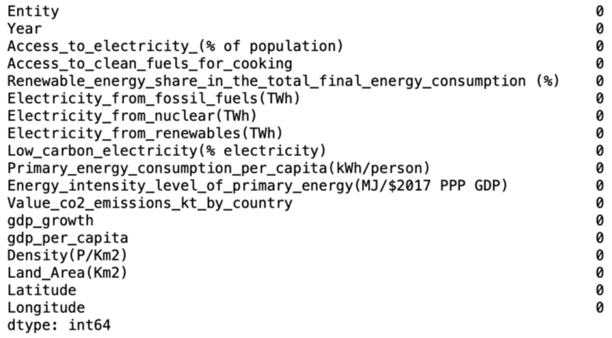
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3649 entries, 0 to 3648
Data columns (total 21 columns):
 #   Column                                                        Non-Null Count  Dtype  
---  ------                                                        --------------  -----  
 0   Entity                                                        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(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 - 2 Info about Dataset

The above fig 2 shows the information about dataset and its values with its datatypes.

Handling Missing Values: The dataset was examined for missing values. Several columns had null values, which were addressed as follows:

- Columns with a significant number of missing values were dropped from the dataset. These included 'Renewables', 'Financial flows to developing countries(US $)' etc.
- Columns with minimal missing values were imputed with the mean of the respective columns. This included columns like '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', and 'gdp_per_capita'.



```
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(P/Km2)                                                      0
Land_Area(Km2)                                                      0
Latitude                                                            0
Longitude                                                           0
dtype: int64
```

Figure - 3 Zero Null Values

In figure 3, Dropping Rows with Null Values: After handling columns with many missing values and imputing others, any remaining rows with null values were dropped to ensure a complete dataset for analysis.

Duplicate Row Verification: The dataset was checked for duplicate rows to ensure data integrity. It was confirmed that there were no duplicate entries, maintaining the uniqueness of the data.

Final Dataset Verification: After all preprocessing steps, the dataset was verified to ensure it was free from missing values and duplicates. The final dataset consisted of 18 columns with 3597 entries, ready for analysis. This comprehensive preprocessing ensured that the data was clean, standardized, and suitable for subsequent analytical methods.

3. *Analytical Methods*

Hub The analytical methods employed in this study encompass a variety of techniques to explore, analyze, and model the relationships between different energy and economic indicators. These methods include exploratory data analysis (EDA), machine learning modeling, and various visualization techniques. Below is a detailed description of each method used:

Exploratory Data Analysis (EDA)

1. Descriptive Statistics:
   - o Descriptive statistics such as mean, median, standard deviation, and range were calculated for key variables to summarize the central tendency and dispersion of the data.
   - o Distribution Analysis: The distribution of key variables was analyzed using histograms and box plots to understand the underlying data distribution and identify any skewness or kurtosis.

2. Correlation Matrix:
   - o Fig 4 is correlation matrix was computed to understand the relationships between different numerical variables in the dataset. This matrix helped identify both positive and negative correlations between variables such as CO2 emissions, electricity consumption from various sources
   - o Visualization: A heatmap was created to visualize the correlation matrix, highlighting the strength and direction of relationships between variables. This provided a visual summary of how different factors are interrelated.
   - o Significance Testing: Statistical significance of the correlations was tested to ensure that the observed relationships were not due to random chance.figure-4 below is a pictorial presentation of heat map.

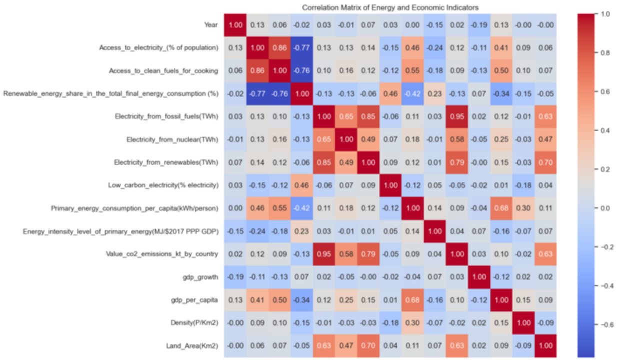


Figure - 4 Correlation Matrix

3. Scatter Plots and Regression Analysis:

   - Scatter plots with regression lines were used to visualize and quantify the relationships between pairs of variables. For example, plots were created to examine the relationship between:

     - Electricity from fossil fuels and CO2 emissions
     - Access to electricity and access to clean fuels for cooking
     - GDP per capita and primary energy consumption per capita
     - Renewable energy share and access to clean fuels for cooking.

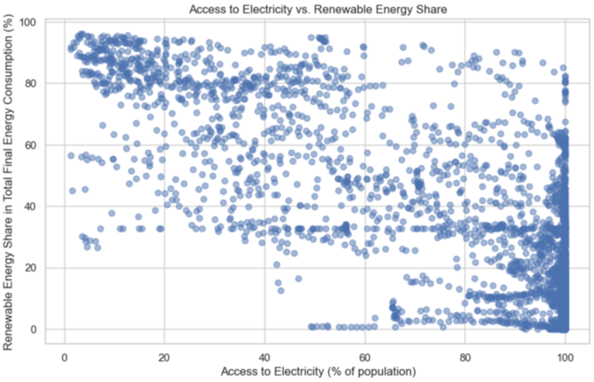


Figure-5 Plot for electricity access

   - Regression Lines: These plots in Figure-5 included regression lines to show the trend and strength of relationships, allowing for the identification of significant trends and patterns.
   - Residual Analysis: Residual plots were used to assess the fit of the regression models and to check for any patterns in the residuals that might indicate model inadequacies.

4. Geospatial Analysis:

   - Latitude and longitude data were utilized to perform geospatial analysis, examining how geographic location influences access to energy and other variables. This added a spatial dimension to the analysis, helping to identify regional disparities.
   - Spatial Clustering: Spatial clustering techniques were used to identify regions with similar energy and economic characteristics, providing insights into regional patterns and clusters.

Machine Learning Modeling

1. Regression Models:

   - Multiple regression models were implemented to predict primary energy consumption per capita based on various features. The models used included:
     - Random Forest Regressor: An ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction.
     - Gradient Boosting Regressor: An ensemble technique that builds models sequentially, each one correcting the errors of its predecessor.
     - AdaBoost Regressor: An ensemble method that combines multiple weak learners to create a strong learner by focusing on errors of previous models.
     - Support Vector Regressor (SVR): A machine learning model that uses support vector machines for regression tasks.

   - Model Training and Testing:

     - The dataset was split into training and testing sets using an 80-20 split to ensure that the models could be evaluated on unseen data. This split helped in assessing the generalization capability of the models.

   - Hyperparameter Tuning: Hyperparameters for each model were tuned using techniques such as grid search and cross-validation to optimize model performance.

2. Model Evaluation:

   - The performance of these models was evaluated using metrics such as R-squared (R2), adjusted R-squared, mean absolute error (MAE), and root mean squared error (RMSE). These metrics provided insights into the accuracy and reliability of the predictions made by each model.

   - Comparison of Models:

     - The performance of each model was compared based on the evaluation metrics. The Random Forest Regressor emerged as the best-performing model with the highest R2 score and the lowest RMSE, indicating its superior predictive capability for this dataset.

   - Cross-Validation: K-fold cross-validation was employed to ensure that the evaluation metrics were robust and not dependent on a particular train-test split.

Model	R2Score	Adjusted R2 Score	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Gradient Boosting	0.964	0.964	3708.472	6898.599
Ada Boosting	0.865	0.863	9091.661	13417.416
Support Vector	-0.143	-0.167	22181.347	39094.552
Random Forest	0.979	0.978	2354.981	5323.621

Table 1. Shows scores of models

3. Feature Importance Analysis:

   - Feature importance analysis was conducted to identify which variables had the most significant impact on the target variable (primary energy consumption per capita). This helped in understanding the key drivers of energy consumption.
   - Partial Dependence Plots: Partial dependence plots were created to visualize the effect of each feature on the predicted outcome, providing deeper insights into the model's behavior.

4. Ensemble Methods:

   - Ensemble methods such as stacking and voting were explored to combine the predictions of multiple models, aiming to improve overall prediction accuracy by leveraging the strengths of different models. Various scores of models are shown in **table 1**.

III. Visualization Techniques

1. Choropleth Maps:

   - Choropleth maps in Figure-6 were created to visualize the geographical distribution of variables such as CO2 emissions and access to electricity. These maps provided a spatial perspective on the data, highlighting regional differences and trends.
   - Interactive Features: The maps included interactive features such as hover information and zoom capabilities, enhancing the user's ability to explore the data.

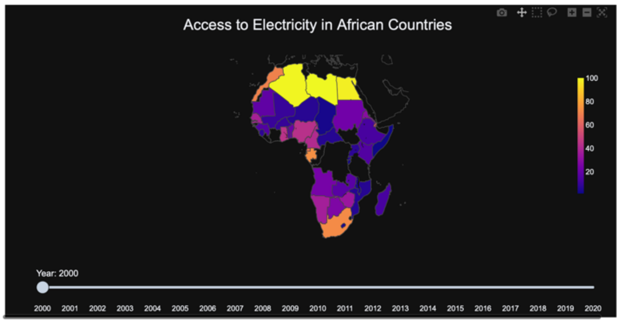


Figure-6 African countries

2. Time Series Plots:

   - Interactive Dashboards: An interactive dashboard in Figure-7 and figure 6 was created using Dash, allowing users to select different countries and view corresponding time series plots and choropleth maps. This made the analysis more dynamic and user-friendly.

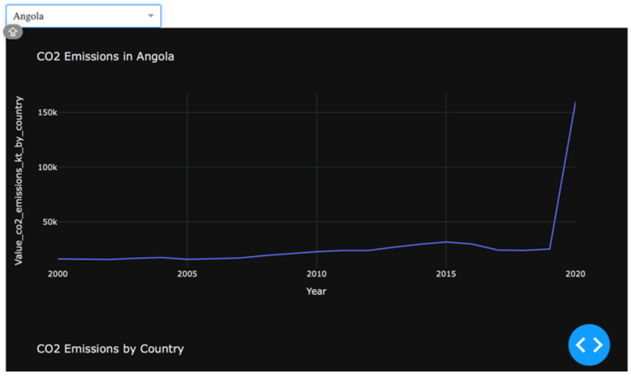


Figure-7 CO2 Emissions

   - Seasonal Decomposition: Seasonal decomposition of time series data was performed to separate the trend, seasonal, and residual components, providing a clearer view of underlying patterns.

3. Pairwise Plot Analysis:

   - Pairwise plots were used to visualize relationships between multiple variables simultaneously, providing a comprehensive view of interactions between different indicators.
   - Multivariate Analysis: Multivariate analysis techniques such as principal component analysis (PCA) was employed to reduce the dimensionality of the data and visualize the relationships between multiple variables in lower-dimensional spaces.

4. Regression and Residual Plots:

   - Detailed regression and residual plots were used to assess the fit of the models and to check for patterns that might indicate model inadequacies or the presence of outliers.

These analytical methods collectively enabled a comprehensive exploration and analysis of the dataset, providing valuable insights into the relationships between energy consumption, economic factors, and environmental impacts. The combination of EDA, machine learning modeling, and visualization techniques facilitated a robust and thorough investigation of sustainable energy trends and their economic implications.

IV. RESULTS

CORRELATION MATRIX

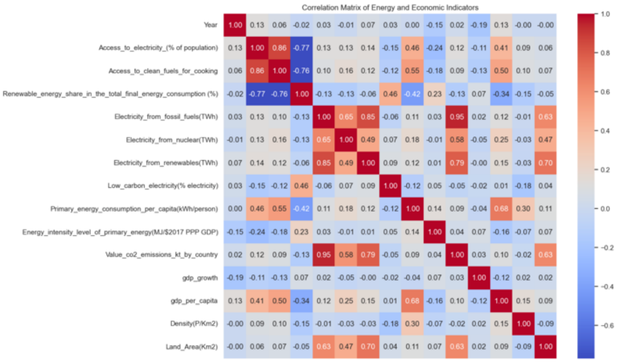


Figure-8 Correlation matrix

Figure-8 represents correlation of various elements/columns in our dataset.

CORRELATION RELATIONSHIPS

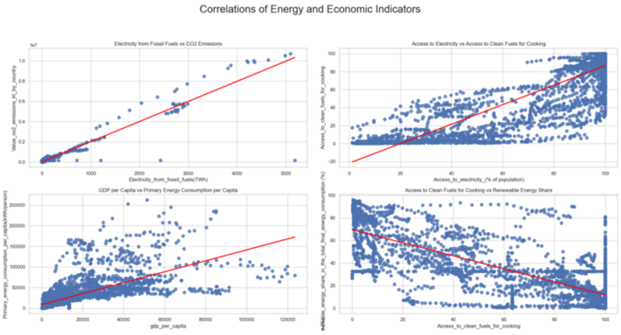


Figure-9 Correlational relationships

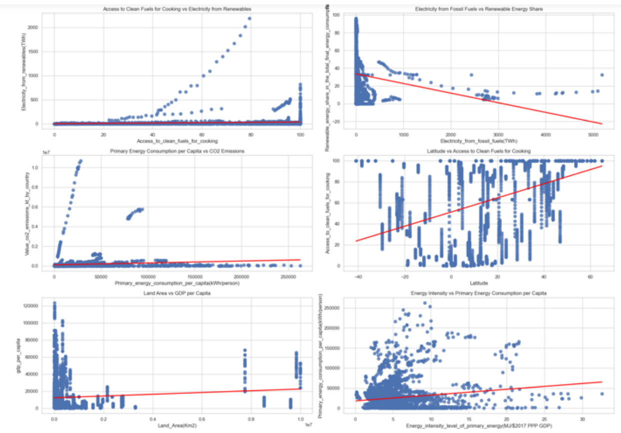


Figure-10 Correlation relationships

Figure 9 and Figure-10 are graphical representations of the Correlation of the columns.

MODEL PERFORMANCE

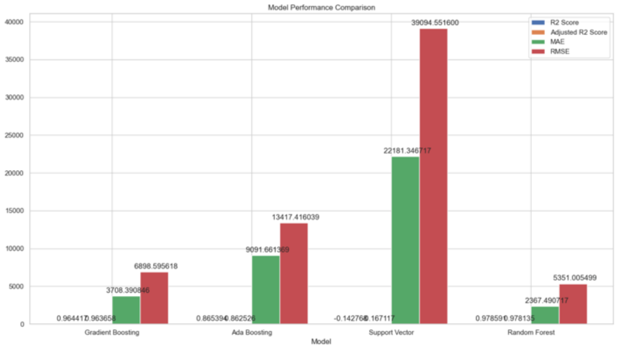


Figure-11 model performance

Figure-11 Shows the overall model performance of the each model.

V.ABBREVIATIONS AND ACRONYMS

1. GDP: Gross Domestic Product
2. CO2: Carbon dioxide
3. TWh: Terawatt-hour (unit of energy)
4. MJ/$2017 PPP GDP: Megajoules per 2017 purchasing power parity GDP dollar
5. R2: R-squared (coefficient of determination)
6. MAE: Mean Absolute Error
7. RMSE: Root Mean Squared Error
8. SVR: Support Vector Regressor
9. PCA: Principal Component Analysis

VI.CONCLUSION

To sum up, this research provides a comprehensive examination of global energy sustainability by utilizing a variety of economic, energy-related, and environmental data. Advanced analytics methods such as correlation analysis, exploratory data analysis, and machine learning models (such as Random Forest, Support Vector Regression, Ada Boosting, and Gradient Boosting) have provided important new insights into the intricate relationships influencing energy consumption and environmental impact globally. While correlation matrices and regression studies have deepened understanding of how economic growth effects energy use and sustainability, spatial and temporal visualizations have highlighted important regional trends.

These results are essential for creating data-driven plans and policies that will improve environmental protection, promote economic expansion, and increase global energy security. Through its focus on clean energy transitions and flexible regulatory frameworks, this study advances the objectives of sustainable development in the changing worldwide issues. All things considered, including multidimensional data analysis has given thorough insights into the dynamics of the world's energy supply, guiding plans for a robust and sustainable energy future.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Professor Geoffrey Thomas for his invaluable guidance, support, and feedback throughout the development of this research paper.Professor Thomas's expertise and mentorship have enhanced in shaping the direction of this work and quality.

REFERENCES

[1] Introducing Sustainable Energy & Fuels' newest Editorial Board member, Tharamani C. Nagaiah
08 Dec 2023By Lily Newton, Development Editor.

[2] Wu, Lin and Weng, "Probability estimates for multi-class classification by pairwise coupling", JMLR 5:975-1005, 2004.

[3]"A Tutorial on Support Vector Regression" Alex J. Smola, Bernhard Schölkopf - Statistics and Computing archive Volume 14 Issue 3, August 2004, p. 199-222.