BIRATNAGAR
INTERNATIONAL COLLEGE

ing

UNIVERSITY OF
WOLVERHAMPTON

Report For Regression

Prediction of Primary Energy Consumption Using Regression Techniques

Module: Concepts and Technologies of AI (5CS037)

Assessment: Final Portfolio – Regression Task

Student Name: Bhawanath Sapkota

Student ID: 2513314

Module Leader: Siman Giri

Tutor: Ayush Regmi

Table of contents

Abstract

The objective of this project is to predict the primary energy consumption, which is a continuous variable, using machine learning regression models. For the purpose of this project, the World Energy Consumption data set, which has around 4,400 data points and more than 120 numerical features, has been employed (Ritchie, Roser, and Rosado, 2023). Moreover, the data set is consistent with the United Nations Sustainable Development Goal 7, which is "Affordable and Clean Energy" (United Nations, 2015).

In the proposed methodology, data preprocessing, exploratory data analysis, application of the Neural Network Regressor, development of two traditional machine learning models, i.e., Decision Tree Regressor and Random Forest Regressor, optimization of the model's hyperparameters, selection of the features, and comparison of the models are included. For the evaluation of the models, the metrics, i.e., Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared score, are employed. Based on the results, it is evident that the Random Forest model has the highest performance, while the Decision Tree model has also shown significant results in terms of accuracy.

List of Figures

1. Introduction

1.1 Problem Statement

Accurate prediction of energy consumption is a critical global challenge due to increasing energy demand, growing concerns over climate change, and the ongoing transition toward sustainable energy systems. Energy consumption patterns are shaped by multiple interacting factors, including fossil fuel usage, renewable energy adoption, and nuclear energy contribution. These relationships are often complex and non-linear in nature.

However, traditional statistical methods usually follow linear patterns, which may not be sufficient to describe the complexity of the problem. As a result, such models can produce limited predictive accuracy when applied to large and diverse datasets.

To overcome the limitations of traditional methods, the project employed machine learning regression models to predict the primary energy consumption. Regression models have the ability to learn from the historical data and can be used to make better predictions regarding the energy consumption patterns. These predictive insights are valuable for evidence-based energy planning, policy formulation, and sustainability assessment (Breiman, 2001).

1.2 Dataset

The World Energy Consumption Dataset was obtained from Our World in Data, a widely used and publicly accessible research repository that provides curated global energy statistics (Ritchie, Roser and Rosado, 2023). The dataset contains numerical indicators related to fossil fuel consumption, renewable energy usage, nuclear energy production, energy shares, and total primary energy consumption.

The target variable in this study is primary energy consumption, which represents the total amount of energy consumed across observations. This dataset is particularly suitable for regression analysis due to its numerical structure and comprehensive coverage of global energy indicators.

Furthermore, the dataset aligns with United Nations Sustainable Development Goal (SDG) 7 – Affordable and Clean Energy, as it supports predictive analysis of energy demand and contributes to understanding how energy systems can transition toward more sustainable and efficient usage (United Nations, 2015).

1.3 Objective

The objectives of this project are as follows:

- To explore and understand global energy consumption patterns using Exploratory Data Analysis (EDA)
- To build regression models capable of predicting primary energy consumption
- To evaluate and compare neural network and classical machine learning regression models
- To identify the most effective model based on quantitative performance metrics

---

2. Methodology

This section describes the systematic steps followed to preprocess the dataset, explore its characteristics, develop regression models, and evaluate their performance. Each stage was designed to ensure reliable, reproducible, and interpretable results.

2.1 Data Preprocessing

Data preprocessing was done to ensure the quality and consistency of the data before the actual model was developed. This was done by removing rows with missing data on the target variable, as this could lead to an invalid or biased model. Numerical features only were used, as the regression algorithm can only handle numeric features.

An 80-20 split was done on the data to test the models and ensure their ability to generalize the data. Feature scaling was done on the data, especially for the neural network model, to ensure the stability and convergence of the model. This was done to ensure the data was consistent and the model could be replicated, as required (Pedregosa et al., 2011).

2.2 Exploratory Data Analysis (EDA)

For this purpose, an Exploratory Data Analysis (EDA) was performed. EDA is a crucial step in understanding the features of the data, correlations, and possible non-linear relationships in the data.

Figure 1: Distribution of Primary Energy Consumption

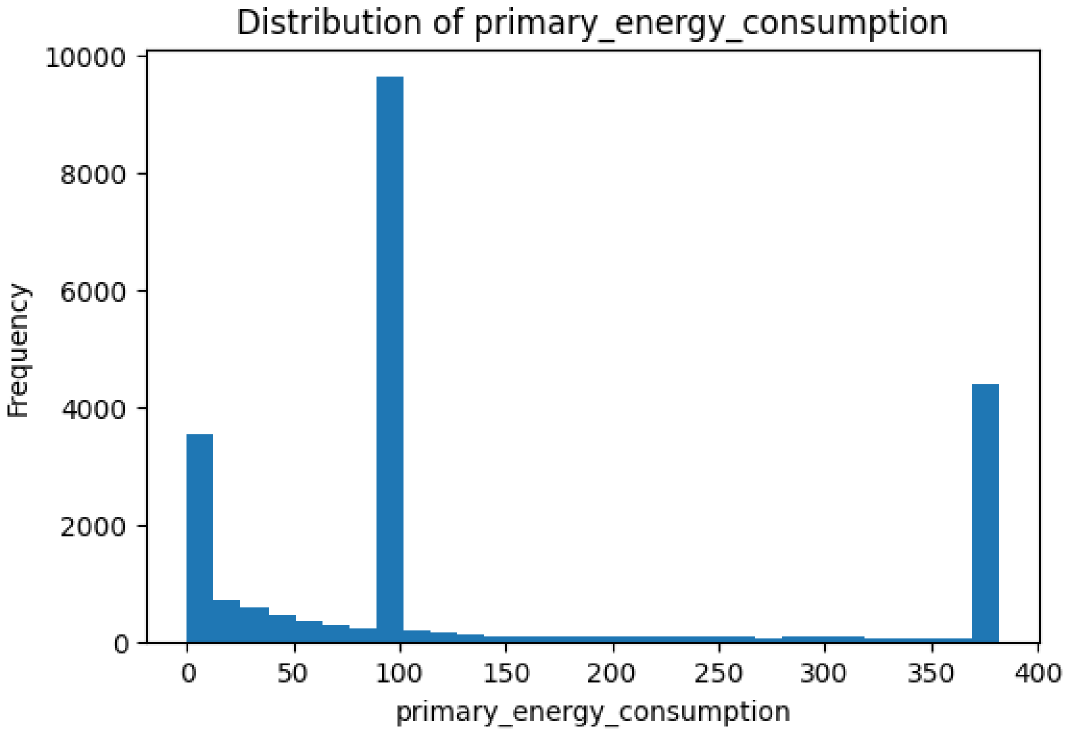


Figure 1 illustrates that primary energy consumption values are highly skewed. This indicates substantial variation in energy usage across observations, with some values significantly larger than others. Such skewness suggests that the assumptions of linear regression may not be satisfied.

Figure 2: Distribution of Energy-Related Features

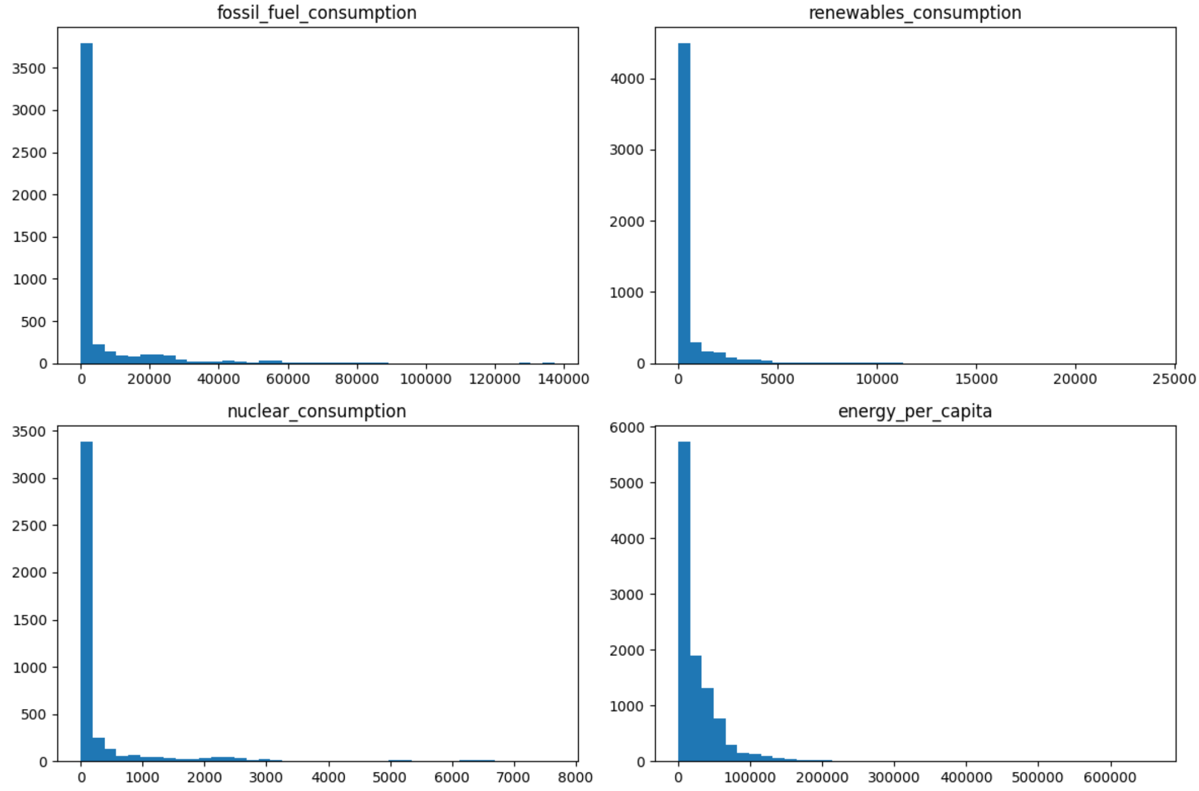


Figure 2 shows the distributions of selected energy-related features. Many features exhibit non-normal and skewed distributions rather than symmetric patterns. This observation supports the use of non-linear regression models that do not rely on strict normality assumptions.

Figure 3: Correlation Heatmap of Energy Features

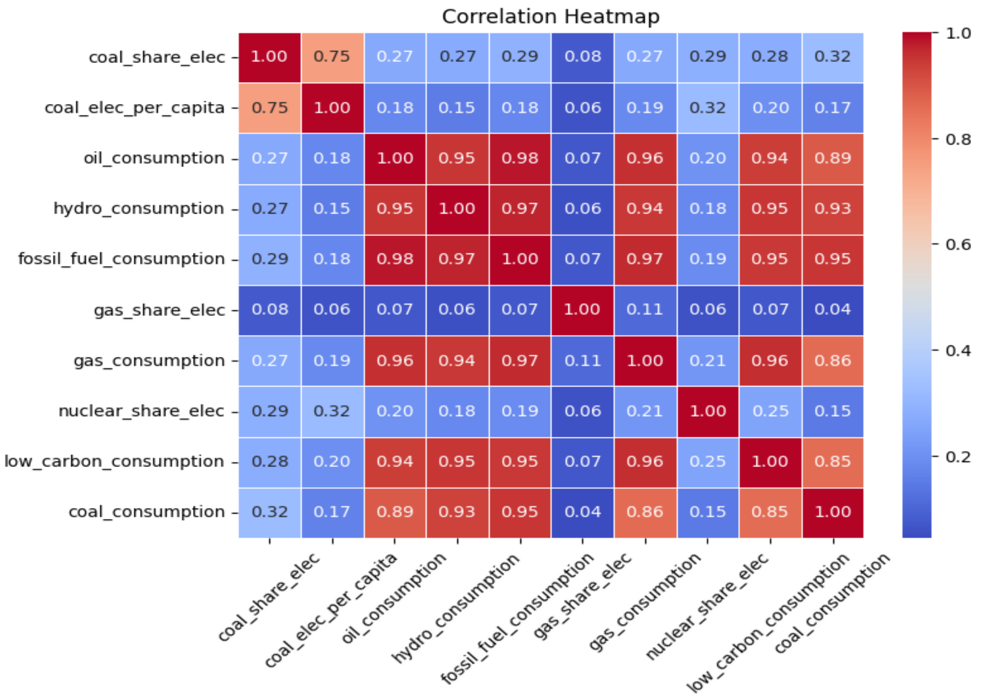


Figure 3 presents the correlation heatmap among energy-related variables. Strong correlations are observed among fossil fuel consumption and energy share indicators, indicating multicollinearity. This multicollinearity can negatively impact linear models and explains their limited performance. Consequently, tree-based regression techniques were considered more appropriate .

Figure 4: Scatter Plot Showing Non-Linear Relationships

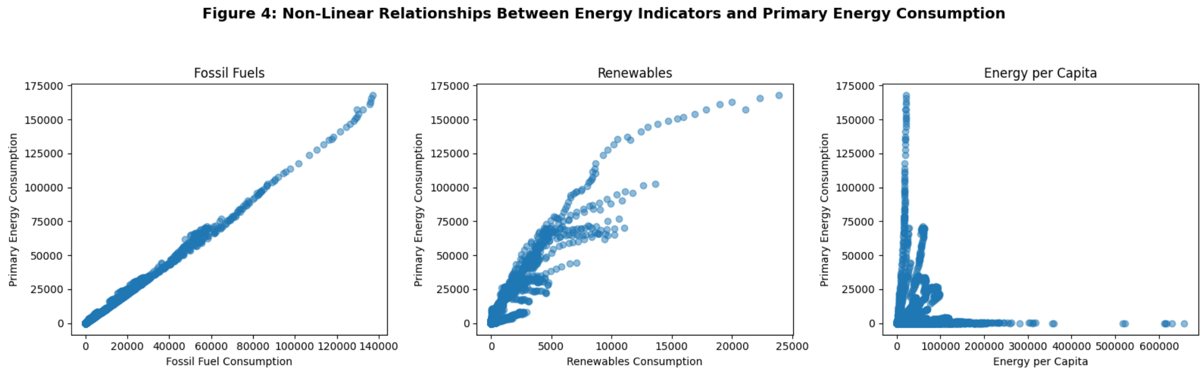


Figure 4 demonstrates non-linear relationships between selected predictors and primary energy consumption. These patterns confirm that energy consumption is influenced by complex interactions among features rather than simple linear trends.

2.3 Model Building

After completing data preprocessing and Exploratory Data Analysis (EDA), multiple regression models were developed to predict primary energy consumption. The selection of models was guided by insights obtained from EDA, particularly the presence of non-linear relationships, skewed feature distributions, and multicollinearity among energy-related variables.

To ensure a comprehensive comparison, both a Neural Network–based regression model and classical machine learning regression models were implemented. The Neural Network model has been used to capture complex non-linear patterns in the data. Tree-based models have also been chosen due to their robustness to multicollinearity, non-linear relationships, and good performance on structured numerical data sets. This combination allowed for a balanced evaluation of model complexity, interpretability, and predictive performance.

2.3.1 Neural Network Regressor

For the Neural Network requirement, a Neural Network Regressor has been implemented. This will help in fulfilling the requirement of the Neural Network model. It has been implemented by including several hidden layers with ReLU activation functions. This will help the model learn complex non-linear relationships in the data. MSE has been used as the loss function, while the Adam optimizer has been used for efficient parameter updates during the training process (Goodfellow, Bengio, and Courville, 2016).

2.3.2 Classical Machine Learning Models

Two classical regression models were developed for comparison:

Decision Tree Regressor

The Decision Tree Regressor captures non-linear relationships by recursively splitting the data based on feature thresholds. It does not require feature scaling and provides interpretable decision rules, making it useful for understanding feature influence.

Random Forest Regressor

Random Forest is an ensemble-based regression model. It is a combination of several decision trees used for regression predictions. It has been chosen due to its good performance on structured numerical data sets. It also performs well in complex non-linear relationships.

2.4 Model Evaluation

The performance of the models was assessed using various regression evaluation metrics, as follows:

- Mean Absolute Error (MAE)
- Measured Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

These metrics are useful in obtaining a thorough evaluation of the models' precision, as they are capable of comparing the models' performances comprehensively (Pedregosa et al., 2011).

2.5 Hyperparameter Optimization

For the Decision Tree and Random Forest models, hyperparameter tuning was carried out using GridSearchCV with 5-fold cross-validation. Parameters such as maximum depth, minimum samples per split, and number of estimators were considered for tuning. Cross-validation was used for robust parameter tuning, avoiding the risk of overfitting..

2.6 Feature Selection

Feature selection was conducted using **Recursive Feature Elimination (RFE)**, a wrapper-based feature selection technique that evaluates feature importance based on model performance.

Figure 5: Feature Importance after Wrapper-Based Feature Selection

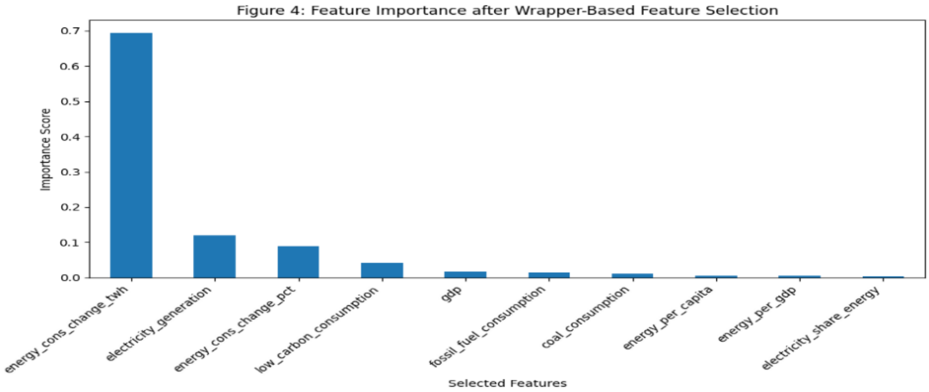


Figure 5 highlights the most influential energy indicators contributing to primary energy consumption. Feature selection reduced model complexity while maintaining strong predictive performance, thereby improving interpretability and generalization.

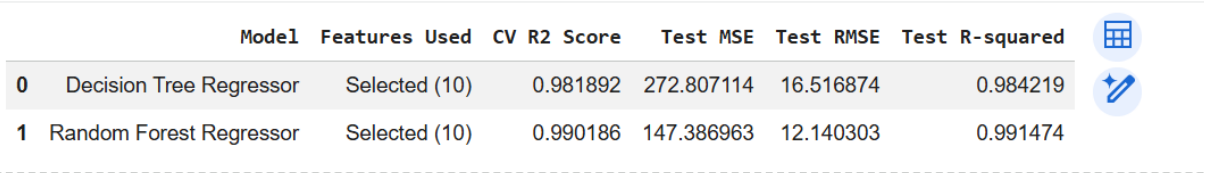
3. Results and Conclusion

This segment of the report outlines the final experimental results of the regression models. It also includes a summary of the important findings of the models' evaluation and comparison.

3.1 Final Model Comparison

The final version of the Decision Tree Regressor model and the Random Forest Regressor model were trained utilizing the optimized parameters of the models, which were found during the cross-validation step, as well as the selected features subset that was found during the feature selection step.

Figure 6: Comparison of Final Regression Models



	Model	Features Used	CV R2 Score	Test MSE	Test RMSE	Test R-squared
0	Decision Tree Regressor	Selected (10)	0.981892	272.807114	16.516874	0.984219
1	Random Forest Regressor	Selected (10)	0.990186	147.386963	12.140303	0.991474

Figure 6: Comparison of final regression model performance by evaluation metrics. The Random Forest Regressor model had the lowest error values for MAE and RMSE and the highest R² value among all models evaluated.

The Decision Tree Regressor model also showed promising results with low error values for MAE and RMSE and a high R² value. However, its performance is slightly low compared to the Random Forest Regressor model because of its ensemble nature that reduces variance and overfitting..

3.2 Key Findings

The experimental results show that tree-based regression models perform much better than linear regression models for predicting primary energy consumption. This is consistent with the findings of Exploratory Data Analysis, which showed non-linear relationships and multicollinearity among features related to energy.

Among all regression models evaluated for this study, the Random Forest Regressor model showed promising results with a good balance between accuracy and generalization. The Decision Tree Regressor model also showed promising results with a focus on model interpretability.

3.3 Final Model

On the basis of evaluation criteria and cross-validation scores, the final model was chosen as Random Forest Regressor. The model has a high R² value and low prediction error, which shows that the model is capable of handling complex patterns in the data and making accurate predictions for unseen data points. Therefore, Random Forest is the best model for primary energy consumption prediction among all models used in this study.

3.4 Challenges

Several challenges were encountered during this project. The dataset contained a large number of features, leading to high dimensionality and potential redundancy. Additionally, multicollinearity among energy indicators complicated the use of linear regression models. The presence of non-linear relationships further required careful model selection and tuning to achieve optimal performance.

The above-mentioned issues were resolved using feature selection, hyperparameter tuning, and the application of robust tree-based regression models.

3.5 Future Work

Future work could focus on improving predictive performance by applying advanced ensemble techniques such as gradient boosting models (e.g., XGBoost or LightGBM). Incorporating time-series forecasting methods could further enhance energy consumption prediction by capturing temporal trends. Additionally, feature engineering and the inclusion of external socio-economic indicators may provide further improvements in model accuracy and interpretability.

4. Discussion

In this section, the performance of the proposed regression models will be discussed, and the results of the optimization techniques' impact will be highlighted, along with the limitations of the proposed models based on the characteristics of the dataset used.

4.1 Model Performance

As shown by the results, the Random Forest Regressor performed better in terms of generalization ability compared to the Decision Tree Regressor model. The better performance of the Random Forest Regressor is attributed to the ensemble property of the model, where a combination of decision trees is used, thus avoiding overfitting.

In contrast, the **Decision Tree Regressor**, while slightly less accurate, provided greater interpretability through its hierarchical decision structure. This highlights a common trade-off in machine learning between predictive performance and model interpretability. Overall, both models performed well, but Random Forest achieved more consistent and robust predictions across the test dataset.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning was instrumental in improving the stability and performance of the models significantly. By selecting optimal parameter values through cross-validation, both regression models achieved better generalization and reduced sensitivity to training data variations.

Feature selection was used to improve the performance of the models by removing noise and irrelevant features, as well as redundant features, from the models. Importantly, reducing the feature set did not lead to a decrease in predictive accuracy, indicating that a smaller subset of energy indicators was sufficient to capture the underlying patterns influencing primary energy consumption.

4.3 Interpretation of Results

The results confirm that primary energy consumption is governed by complex and non-linear interactions among multiple energy indicators, including fossil fuel usage, renewable energy contribution, and energy shares. These interactions are difficult for linear regression models to capture effectively.

Ensemble methods such as Random Forest are more suitable for this problem, as they can handle non-linear relationships and interactions without making any specific assumption about the data distribution. The good results of tree-based methods in this study are consistent with the characteristics of structured numerical data with correlations.

4.4 Limitations

Although good results have been obtained, it is important to highlight a series of limitations of this study. In particular, it only considers historical numerical data, which may not allow us to capture future changes in energy systems. Additionally, temporal dynamics were not explicitly modeled, and external socio-economic factors such as population growth, policy changes, or economic conditions were not included.

These limitations suggest that while the models perform well within the scope of the available data, their predictive capability could be further enhanced with additional contextual information.

4.5 Suggestions for Future Research

As a suggestion for future research, it may be interesting to consider more powerful methods of deep learning and ensemble methods, such as gradient boosting methods. Incorporating time-series forecasting approaches may allow the models to capture temporal trends in energy consumption. Furthermore, integrating real-time data sources and additional socio-economic indicators could enhance both accuracy and practical applicability of energy consumption prediction models.

References

Breiman, L. (2001) ‘Random Forests’, *Machine Learning*, 45(1), pp. 5–32. Available at: https://link.springer.com/article/10.1023/A:1010933404324 (Accessed: 1 February 2026).

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. MIT Press. Available at: https://www.deeplearningbook.org/ (Accessed: 1 February 2026).

Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12, pp. 2825–2830. Available at: https://scikit-learn.org/stable/ (Accessed: 1February 2026).

Ritchie, H., Roser, M. and Rosado, P. (2023) *Energy*. Our World in Data. Available at: https://ourworldindata.org/energy (Accessed: 1February 2026).

United Nations (2015) *Sustainable Development Goal 7: Affordable and Clean Energy*. Available at: https://sdgs.un.org/goals/goal7 (Accessed: 1 February 2026).