**Predicting the Impact of Renewable Energy Transition on UK's CO2 Emissions: A Support Vector Regression Approach**

John Philip

University of York

Applied Artificial Intelligence

24.06.2023





### **Table of Contents**

Executive Summary

1. Introduction
2. Literature Review
3. Research Design
4. Experimental Results and Analysis
5. Conclusion
6. References
7. Appendix

### **Executive Summary**

This research report details the development of a supervised learning model with feature optimisation to predict how changes in renewable energy usage influence CO2 emissions in the UK. Firstly, an exploratory data analysis was conducted to understand the properties and characteristics of two datasets from the Office of National Statistics and the Global Carbon Project. The data preprocessing phase involved handling different units of measurement and addressing data collection challenges such as nominal attributes and missing data. Additionally, the report examines the assumptions underlying the regression analysis, including the established relationship between renewable energy and carbon emissions.

The research employed a support vector regression (SVR) model with a linear kernel parameter and feature selection optimisation techniques such as tabu search and greedy search. The techniques were chosen for their ability to avoid local optima, reduce overfitting, and provide effective models for energy predictions using limited training data.

The results of the study demonstrated the effectiveness of supervised learning models in assessing the impact of renewable energy on CO2 emissions. By carefully analysing and selecting influential features, such as 'Energy from renewable & waste sources' and 'Total Energy', more accurate predictions of carbon emissions were achieved. The model's efficiency was evaluated by examining its R-squared and Mean Squared Error (MSE) values, utilising K-fold cross-validation, and employing feature importance and optimisation techniques. However, the research also identified limitations and areas for future reconsideration. The study acknowledged that the predictive models did not account for all features influencing CO2 emissions, such as economic growth, energy type, energy efficiency, urbanisation, and government policies.

In conclusion, this research provides valuable insights for policymakers and businesses in the renewable energy sector. The model demonstrates the impact that increased renewables have on lowering carbon emissions. While the study succeeded, it also emphasises the need for continued research and refinement to create more robust predictive models encompassing a broader range of influencing factors.

### **Introduction**

Human-induced climate change is one of the 21st century's most pressing existential issues. Extensive evidence identifies greenhouse gas emissions as a main driver of global warming affecting all regions worldwide[1]. Climate change consequences, including ozone depletion, ocean warming, ice loss, sea level rises, temperature extremes, water and air pollution, and droughts, pose threats to food and water resources, human health, economies, and infrastructure [1].

The energy sector, responsible for over two-thirds of global greenhouse gas emissions [2], plays a significant role in climate change. The Intergovernmental Panel on Climate Change (IPCC) identifies emission reduction as crucial to limiting warming to 1.5°C or less than 2°C by the century's end [1, p. 32]. Understanding factors driving CO2 emissions is key to designing effective reduction policies. Accurate energy forecasting is vital for planning energy logistics, environmental conservation, and economic growth [3, p. 5].

This report seeks to build a predictive model for the UK's energy profile by linking renewable energy and CO2 emissions for potential business clients. It utilises two datasets: one from the Global Carbon Project detailing 250 years of global CO2 emissions and another from the Office of National Statistics (ONS) documenting 30 years of UK energy consumption, focusing on renewables[22][23]. The challenge lies in synthesising these datasets to create a model that shows the CO2 emission impact of transitioning from non-renewable to renewable energy. AI techniques, specifically supervised learning and feature selection optimisation, help manage the datasets' complexity, uncover patterns, handle nonlinearity, and manage noise and errors [4]. This report outlines developing a predictive model using Support Vector Regression and tabu search optimisation and evaluating its performance via R-squared, Mean Squared Error, and K-fold validation metrics.

### **Literature Review**

Energy prediction research has led to various sophisticated models, with Artificial Intelligence (AI) models highly utilised due to their specific strengths. AI models effectively handle large datasets and nonlinear problems without needing explicit relationships [4]. They learn from data without predefined rules, offer flexibility in data types, and adapt to changing scenarios. AI models, such as artificial neural networks, decision trees, naïve Bayes, and support vector machines, are frequently employed in energy prediction [4]. While these models are used for classification and regression, our review will focus on regression, particularly exploring the link between renewable energy and carbon emissions.

Artificial neural networks (ANNs) emulate the human brain's learning process, responding to external stimuli through weighted inputs from a training dataset, which adapt based on the ANN's input-output accuracy. They serve as a structure for diverse machine learning algorithms to process complex data inputs [4, p.5][8, p.2]. Various ANNs are utilised in energy prediction, including FFNN, BPNN, ANFIS, WNN, ESN, and DL models [4]. ANNs excel by leveraging increased data availability and modern computational power, outpacing traditional ML algorithms' limits. However, for smaller, limited datasets, other ML models may still perform better [8, p.4].

Decision trees predict the label associated with an instance of x by travelling from a root node to a leaf based on a binary classification system. At each node on the root-leaf path, a child is chosen according to a predefined set of rules to arrive at a final prediction. These can be used for classification or regression. With regression, the splitting decision is based on the values of the features and the predicted values. Decision trees are often gradually constructed using optimisation techniques such as the greedy approach. They are generally less accurate than other methods in returning the global optimum decision but tend to work reasonably well in practice and are relatively easy to interpret and implement [9, pp. 212-214].

Naive Bayes are a generative approach to machine learning that aims to model the distribution over the data. It is generally a classification model that assumes that the probability of each feature of a data point is independent of the probability of the other features [9]. For this reason, it would not be best suited for regression forecasting in a situation where we are looking to examine the relationship between variables such as carbon emissions and renewable energy use.

Support Vector Machines are another widely used model in energy prediction that can better avoid being trapped in local minima and overfitting compared to ANN models making it well suited in energy forecasting [4]. They are also very good at solving nonlinear problems while using small quantities of training data [10]. This can be useful for understanding the nonlinear relationships between points in our datasets, such as renewable energy consumption, carbon-based energy consumption, and per capita income.

In the context of the energy prediction scenario, feature selection is a critical task in machine learning models. Several optimisation techniques, such as hill climbing, simulated annealing, tabu search, and genetic algorithms, have been developed for this purpose. Hill climbing iteratively improves a solution by making incremental changes to feature subsets. Simulated annealing employs a probabilistic approach to explore new regions while gradually reducing acceptance of worse solutions. Tabu search uses a memory-based approach to avoid revisiting recently explored solutions. Genetic algorithms mimic biological evolution to create new generations of feature subsets. These techniques offer diverse strategies for feature selection, allowing the identification of feature combinations that maximise model performance based on the problem, dataset characteristics, and trade-offs between exploration and exploitation in the search process [17].

### **Research Design**

This research design aims to forecast the impact of renewable energy usage on the UK's carbon emissions. A predictive model was built using Support Vector Regression (SVR) with a linear kernel, supplemented by tabu search optimisation for feature selection. The model is based on supervised learning and has been trained on preprocessed input data. Its performance is assessed through the coefficient of determination (R-squared), Mean Squared Error, and K-fold validation. The model development process utilised Python in a Datalore Jupyter notebook environment, incorporating the following libraries: pandas, numpy, matplotlib, sklearn, and seaborn.

**Assumptions:**

This regression analysis has two important assumptions: the established relationship between renewable energy and carbon emissions, and the quality and reliability of the data used. The literature supports the notion that replacing carbon-based emissions with renewables can contribute to over 90% of the required reductions in energy-related carbon emissions [12]. Both datasets are taken from reliable governmental sources, however, it is crucial to recognise the limitations of all data, particularly with respect to missing features that could influence our target outcome, such as economic growth, population growth, urbanisation, income, technology, lifestyle and culture, climate, energy prices, and government policy [5].

**Data Exploration and Processing:**

Data exploration and preprocessing presented challenges, including differing measurement units—energy in million tonnes of oil equivalent (Mtoe) in the ONS dataset and metric tons of carbon dioxide equivalent (MtCO2) in the GCB dataset [6][13][22][23]. Additionally, there were disparities in dataset years and limitations in data verification, with unverifiable emissions estimates stemming from survey data, average values, and large data compilations [6, p. 2].

Addressing these issues required focusing exclusively on UK data, aligning years between the Global Carbon Project and UK Office of National Statistics datasets (1990-2020), and handling nominal data. Further preprocessing steps involved differentiating 'Total Energy' from 'Total Emissions', replacing missing values, identifying outliers, and compiling the refined datasets into a unified CSV file [22][23].

**Supervised Learning Model:**

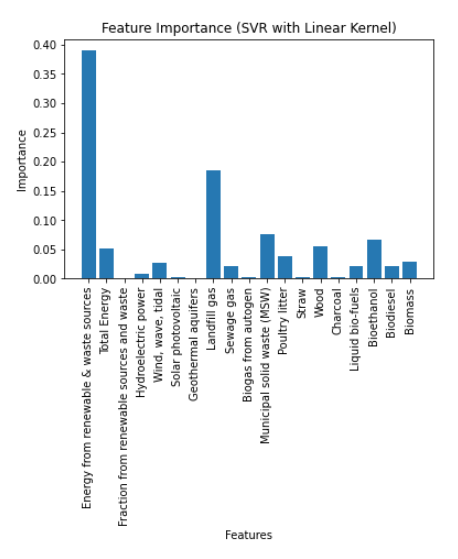
Support Vector Machines are widely used in regression and are well suited for energy predictions as they can better avoid being trapped in local minima and overfitting compared to ANN models [4]. Additionally, they are capable of solving problems while using small quantities of training data [10]. The development of the SVM model used a Python environment with the necessary libraries including, pandas, numpy, matplotlib and sklearn. The processed dataset, 'FINAL\_UK.csv', was loaded into a pandas DataFrame. The dataset was divided into input features 'x' and the target variable 'y', ' Total Emissions'. The dataset was split into an 80-20 training and testing set split using sklearn's 'train\_test\_split' function. The data was scaled using the 'StandardScaler' function from sklearn to avoid any bias due to the different scales of input features. An SVR model with a linear kernel parameter was then initialised and trained on our scaled training data. Once trained, the model was used to make predictions on the test data. Initially, an SVR model with a nonlinear kernel was attempted, however, the predictive capabilities were unsatisfactory in comparison to the linear kernel.

**Feature Selection & Search Optimisation**

The model refinement process incorporated Tabu Search and Greedy Search optimisation methods. Tabu Search, known for its superior performance and faster solution time, uses a hill climbing strategy to optimise the model, specifically adept at evading local optima compared to Greedy Search [17., p. 327]. This technique started with a random feature set, evaluated the SVR model performance via Mean Squared Error and R squared metrics, and made iterative modifications, embracing those which boosted the score. A 'Tabu list' tracked past alterations, averting their reversal and thereby avoiding revisitation of previous solutions [17].

An SVR feature importance analysis, executed using Python, sklearn, and numpy, found a prominent correlation between 'Total Emissions' and variables like 'Energy from renewable & waste sources', 'Total Energy', and 'Landfill Gas'. Intriguingly, despite initial assumptions, 'Fraction from renewable sources and waste' emerged as an insignificant feature, a finding reinforced by the optimisation process, as shown below in Figure 1:

Figure 1:



**Validation**

The model's initial performance was measured using two metrics: Mean Squared Error (MSE) and the Coefficient of Determination (R-squared). The MSE calculates the average squared differences between predicted and actual values, providing insight into the model's overall prediction error rate [11]. Smaller MSE values indicate more accurate predictions. Additionally, R-squared quantifies the percentage of the dependent variable's variance that the model can predict from the independent variables, with values closer to 1 indicating a more robust prediction capability [11]. The performance of the model was further validated using K-fold Cross Validation, as discussed in the results.

### **Experimental Results and Analysis**

The results from our initial model underscore the effectiveness of the supervised learning SVR in evaluating the influence of renewable energy on CO2 emissions (Table 1). Interestingly, while Tabu Search optimisation enhanced model efficiency, the Simple Greedy Search method delivered excellent performance. This can be ascribed to Greedy Search's approach of making locally optimal decisions at each stage, enabling a quick convergence towards the global optimum [17]. Notably, the incorporation of optimised search features only yielded slight improvements over the base SVR model with a linear kernel and all features (Table 4.1). This implies that the initial feature set was already comprehensive and efficient for predictive purposes. Intuitively, i.e. randomly, selected features like 'Energy from renewable & waste sources', 'Total Energy', and 'Fraction from renewable sources and waste' resulted in a less effective model for predicting Carbon Emissions, highlighting the value of thorough feature analysis.

**Table 1 Model Results**

| **Model Name** | **Mean Squared Error (MSE)** | **R-squared** | **K-Fold Cross-validated MSE** | **K-Fold Cross-validated R^2** |
| --- | --- | --- | --- | --- |
| **Initial SVR w/ all features** | 61.66 | 0.991 | 177.74 | -6.73 |
| **SVR with randomly chosen features** | 1482.17 | 0.78 | 2249.19 | -15.88 |
| **SVR w/ Tabu** | 76.18 | .988 | 286.41 | -0.23 |
| **SVR w/ Greedy Approach** | 18.34 | 0.997 | 132.76 | -3.83 |

**K-Fold Validation:**

A K-Fold Cross-Validation method was employed to further assess the performance of the Support Vector Regression (SVR) model. This approach divided the dataset into 'K' equally sized folds. Following this, the model is trained K times, each time using (K-1) folds for training and the remaining fold for validation. This technique ensures that every observation in the dataset is used for both training and validation exactly once. The ultimate performance metrics are then derived by averaging the individual results from all K iterations. The advantage of this method is that it mitigates any potential bias in the model and provides a more reliable estimate of how the model is expected to perform on unseen data [11]. For this validation, the K value was chosen at 10, as empirical evidence shows this to be a good balance between bias and variance [11, p. 177].

**Business Challenge**

*“The challenge is to make a meaningful connection between the two sets of data in order to build a predictive model to demonstrate the impact on CO2 emissions of replacing non-renewable energy sources with renewable energy sources”*

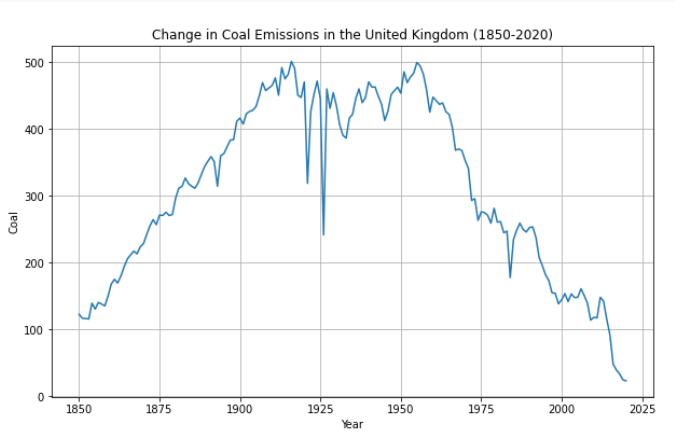
The predictive model developed in this study serves as a valuable asset for clients, enabling them to understand the potential impacts of different renewable energy scenarios. The following examples offer a practical application of the model, exploring hypothetical questions related to the interplay between renewable energy and CO2 emissions:

| What happens to total emissions if you double the amount of renewable energy? |
| --- |
|  |
| What happens if we increase renewables by 1000% |
|  |
| What might be the impact of doubling a particular type of renewable energy on CO2 emissions? |
|  |
| What might total CO2 emissions look like as we reach 100% Renewable Energy? |
|  |

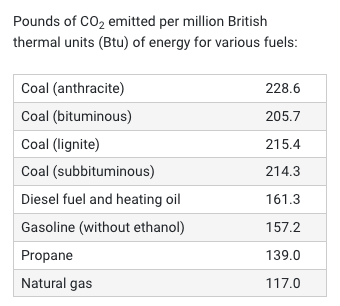
**Limitations**

The question, 'What might total CO2 emissions look like as we reach 100% Renewable Energy?', reveals limitations of the model. While certain renewable energy sources generate CO2, particularly biomass fuels like 'wood', 'charcoal', 'straw', and 'bioethanol', 100% renewable use should result in an enormous decrease in overall emissions [18]. However, the model fails to show this due to insufficient information in our datasets. A comprehensive analysis of global CO2 emissions would require examining numerous features like economic growth, energy type, industry, urbanisation, and growth rate of energy consumption [21]. Accurate forecasting also requires an extensive consideration of forecast scope, data, model selection, and forecast accuracy [5, p.108].

Our datasets demonstrate inconsistency between energy and emissions. The GCP CO2 emissions data is not tied to the amount of energy used; unlike the ONS dataset, GCP data only reflects total emissions [6]. For instance, the observed reduction in CO2 could primarily be attributed to the UK's decline in coal energy over the past 70 years, not necessarily an uptick in renewables [19]. There is evidence of this in the GCP dataset, illustrated below, which shows a decrease in coal emissions without shedding light on how those changes relate to coal energy usage or types of coal.



Coal generates more CO2 per million Btu than other fossil fuels, and this model overlooks the impact of reduced coal usage or cleaner fossil fuel production on CO2 emissions.

*Source: The American Geosciences Institute [20]*

### **Conclusion**

This report presented an in-depth analysis, employing a robust Support Vector Regression (SVR) model with a linear kernel, aiming to examine the impact of renewable energy on CO2 emissions. The findings emphasise the importance of optimised feature selection in making predictive models as opposed to random or intuitive selection approaches.

While the report does provide a valuable tool, it is important to note the limitations of the model. The complexity of factors influencing CO2 emissions is vast and intricate, exceeding the scope of this study with a focus on renewable energy and a handful of features. CO2 emissions are affected by an intricate matrix of variables such as economic growth, industry structure, urbanisation, research and development (R&D) investment, actual use of foreign capital, and the growth rate of energy consumption, to name a few [3][4][7][21]. A more holistic understanding of CO2 emissions would require a multidimensional analysis that accounts for a wider array of influencing factors and careful handling of complex and often messy data.

Despite these limitations, the report holds significance in helping to gain context in how renewable energy influences climate change and sustainable development. The transition to renewable energy, as emphasised in the analysis, is an essential strategy for reducing greenhouse gas emissions. As our planet grapples with the escalating threats of climate change, the transition to renewable energy sources stands as one of the most effective strategies for reducing greenhouse gas emissions [1][2][12].

The model developed is a valuable asset for simulating various scenarios and exploring critical questions in our global energy landscape. Future steps in this research could explore additional machine learning models for comparative analysis and performance enhancement. Expanding the dataset, and incorporating more predictive factors, as described, could also improve the model's accuracy. Another important consideration is that Artificial intelligence alone will not be able to solve climate change, but rather it will require us to think about where and how we employ these tools.

*“You cannot get through a single day without having an impact on the world around you. What you do makes a difference, and you have to decide what kind of difference you want to make.”*

*— Jane Goodall*

### **References**

[1] Lee, H., Calvin, K., Dasgupta, D., et al. (2023). Synthesis Report of the IPCC Sixth Assessment Report (AR6): Summary for Policymakers. Intergovernmental Panel on Climate Change. Retrieved from <https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_SPM.pdf>

[2] International Energy Agency. (2020). If the energy sector is to tackle climate change, it must also think about water. Retrieved from <https://www.iea.org/commentaries/if-the-energy-sector-is-to-tackle-climate-change-it-must-also-think-about-water>

[3] Z. Zuo, H. Guo, and J. Cheng, ‘An LSTM-STRIPAT model analysis of China’s 2030 CO 2 emissions peak’, *Carbon Management*, vol. 11, no. 6, pp. 577–592, Nov. 2020, doi: [10.1080/17583004.2020.1840869](https://doi.org/10.1080/17583004.2020.1840869).

[4] N. Wei, C. Li, X. Peng, F. Zeng, and X. Lu, ‘Conventional models and artificial intelligence-based models for energy consumption forecasting: A review’, *Journal of Petroleum Science and Engineering*, vol. 181, p. 106187, Oct. 2019, doi: [10.1016/j.petrol.2019.106187](https://doi.org/10.1016/j.petrol.2019.106187).

[5] M. Hasanuzzaman and N. A. Rahim, Eds., *Energy for sustainable development: demand, supply, conversion and management*. London, United Kingdom: Academic Press, an imprint of Elsevier, 2020.

[6] R. Andrew and G. P. Peters, ‘The Global Carbon Project’s fossil CO2 emissions dataset: 2022 release’.

[7] Hong, T., Pinson, P., Fan, S., & Zareipour, H. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. International Journal of Forecasting, 32(3), 896-913.

[8] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Cham: Springer International Publishing, 2018. doi: [10.1007/978-3-319-94463-0](https://doi.org/10.1007/978-3-319-94463-0).

[9] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.

[10] M. A. Mat Daut, M. Y. Hassan, H. Abdullah, H. A. Rahman, M. P. Abdullah, and F. Hussin, ‘Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review’, *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 1108–1118, Apr. 2017, doi: [10.1016/j.rser.2016.12.015](https://doi.org/10.1016/j.rser.2016.12.015).

[11] S. Raschka, Y. Liu, V. Mirjalili, and D. Dzhulgakov, *Machine learning with PyTorch and Scikit-Learn: develop machine learning and deep learning models with Python*. in Expert insight. Birmingham Mumbai: Packt, 2022.

[12] *International Renewable Energy Agency. (2019). Global energy transformation: A roadmap to 2050. Retrieved from* [*https://www.irena.org/publications/2019/Apr/Global-energy-transformation-A-roadmap-to-2050-2019Edition*](https://www.irena.org/publications/2019/Apr/Global-energy-transformation-A-roadmap-to-2050-2019Edition)

[13] Office for National Statistics. (2023, June 5). Energy use: renewable and waste sources. Retrieved from <https://www.ons.gov.uk/economy/environmentalaccounts/datasets/ukenvironmentalaccountsenergyconsumptionfromrenewableandwastesources>

[14] BP. (2022). Statistical Review of World Energy 2022. Retrieved from <https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2022-full-report.pdf>

[15] S. Voß, ‘J. Dreo, A. Petrowski, P. Siarry, E. Taillard: Metaheuristics for Hard Optimization: Springer, Berlin, ISBN 10 354023022X, 2006, 369 pages’, *Math Meth Oper Res*, vol. 66, no. 3, pp. 557–558, Nov. 2007, doi: [10.1007/s00186-007-0180-y](https://doi.org/10.1007/s00186-007-0180-y).

[16] Brownlee, J. (2012). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Packt Publishing.

[17] K.-L. Du and M. N. S. Swamy, *Search and Optimization by Metaheuristics*. Cham: Springer International Publishing, 2016. doi: [10.1007/978-3-319-41192-7](https://doi.org/10.1007/978-3-319-41192-7).

[18] Forest Research. (2023). Carbon emissions of different fuels. Retrieved from <https://www.forestresearch.gov.uk/tools-and-resources/fthr/biomass-energy-resources/reference-biomass/facts-figures/carbon-emissions-of-different-fuels/> Accessed: June 17, 2023.

[19] International Energy Agency. (2023). Achieving a swift reduction in global coal emissions is the central challenge for reaching international climate targets. Retrieved from <https://www.iea.org/news/achieving-a-swift-reduction-in-global-coal-emissions-is-the-central-challenge-for-reaching-international-climate-targets> Accessed: June 17, 2023

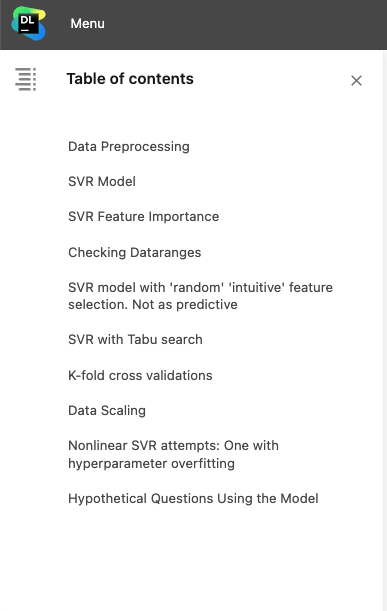
[20] American Geosciences Institute. (2023). How much carbon dioxide is produced when different fuels are burned? Retrieved from <https://www.americangeosciences.org/critical-issues/faq/how-much-carbon-dioxide-produced-when-different-fuels-are-burned> Accessed: June 17, 2023

[21] S. Li, Y. W. Siu, and G. Zhao, ‘Driving Factors of CO2 Emissions: Further Study Based on Machine Learning’, *Front. Environ. Sci.*, vol. 9, p. 721517, Aug. 2021, doi: [10.3389/fenvs.2021.721517](https://doi.org/10.3389/fenvs.2021.721517).

[22] Office for National Statistics. (2023). UK Environmental Accounts: Energy consumption from renewable and waste sources. Retrieved from <https://www.ons.gov.uk/economy/environmentalaccounts/datasets/ukenvironmentalaccountsenergyconsumptionfromrenewableandwastesources>

[23] GCP. Dataset "Emissions by Country" by Anon (2023), accessed on January 2, 2023, available at <https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022>.

### **Appendix: Jupyter Notebook Overview** [LINK](https://datalore.jetbrains.com/view/notebook/B7Aj8RvyYGyBwNuKTjTFif)

The code for this project is also included as a zipfile, however, I am also providing this accompanying [Jupyter Notebook](https://datalore.jetbrains.com/view/notebook/B7Aj8RvyYGyBwNuKTjTFif), prepared in DataLore, as this is where I completed the majority of the coding tasks for this project. A Table of Contents can be accessed on the left hand side as noted here:

The notebook is systematically structured as follows:

1. **Data Preprocessing:** This section provides the code and description of the initial steps in preparing the dataset for modeling
2. **SVR Model**: This part details the implementation of the Support Vector Regression (SVR) model for this study.
3. **SVR Feature Importance**: This section focuses on the different features in the SVR model.
4. **Checking Dataranges:** Exploring the ranges and distributions of the dataset's features.
5. **SVR Model with 'Random' 'Intuitive' Feature Selection**: This portion documents an attempt at manually selecting features for the SVR model based on intuition or random choice, noting that the resulting model was not as predictive as the optimised model.
6. **SVR with Tabu Search:** This section details the application of the Tabu Search method for feature selection in the SVR model.
7. **K-Fold Cross Validations**: Here, the process and results of using K-Fold Cross-Validation to assess the model's robustness and predictive performance
8. **Data Scaling**: This part visualises the scaling of the dataset's features for machine learning modeling
9. **Nonlinear SVR Attempts** - This section provides insights into attempts at using a Nonlinear SVR model
10. **Hypothetical Questions Using the Model:** The final section uses the developed model to answer hypothetical questions about the relationship between renewable energy use and carbon emissions.