Applied Artificial Intelligence
Summative Assessment

Word count: 2512

Executive Summary

In this study, our primary focus was to explore the effect of renewable energy utilization on carbon dioxide (CO2) emissions - a paramount concern as we track climate change. We utilized artificial intelligence methodologies, mainly supervised learning, to predict CO2 emissions relying on data about renewable energy consumption and to comprehend how an increase in renewable energy can mitigate greenhouse gas emissions. [1] Supervised learning involves predicting an outcome measure based on multiple input measures.

Python, alongside its library Scikit-learn, was the primary tool used for developing and evaluating predictive models. The Scikit-learn library offered a diverse range of supervised learning algorithms and was selected due to its simplicity and efficiency.

The data was first processed, and a correlation matrix was used to understand the relationships between the variables. Two models were then developed: a RandomForestRegressor and a DecisionTreeRegressor. The RandomForestRegressor utilized the entire dataset, while the DecisionTreeRegressor applied a subset of features determined through a feature selection process. The feature selection method used was Recursive Feature Elimination (RFE), aiming to improve the model's performance by reducing overfitting, accuracy, and training time.

Results from both models were evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and the R-Squared score. Model 1 (RandomForestRegressor) performed better across all these metrics. Visualization of actual vs. predicted CO2 emissions further reinforced this conclusion.

Future studies could improve results by exploring other feature selection methods and machine learning algorithms. Handling imbalanced data could also be an area of focus, as imbalanced data can bias the model towards the majority class. Moreover, expanding the dataset to include other variables influencing CO2 emissions, such as industrial output or population, could help create a more holistic model. This project serves as a foundational step towards understanding and quantifying the benefits of renewable energy sources in reducing CO2 emissions, providing valuable insights for policymakers, researchers, and businesses in the green energy sector.

Introduction

In the rapidly evolving landscape of the energy sector, a primary area of focus is integrating and optimizing renewable energy sources to reduce CO2 emissions. This is crucial for meeting international climate targets and the sustainable development of global economies. Artificial Intelligence (AI) has emerged as a transformative tool for addressing these problems. Its capacity to process and analyze large amounts of data enables the formulation of predictive models and actionable insights, thereby contributing significantly to decision-making processes and policy formulation in the energy sector.

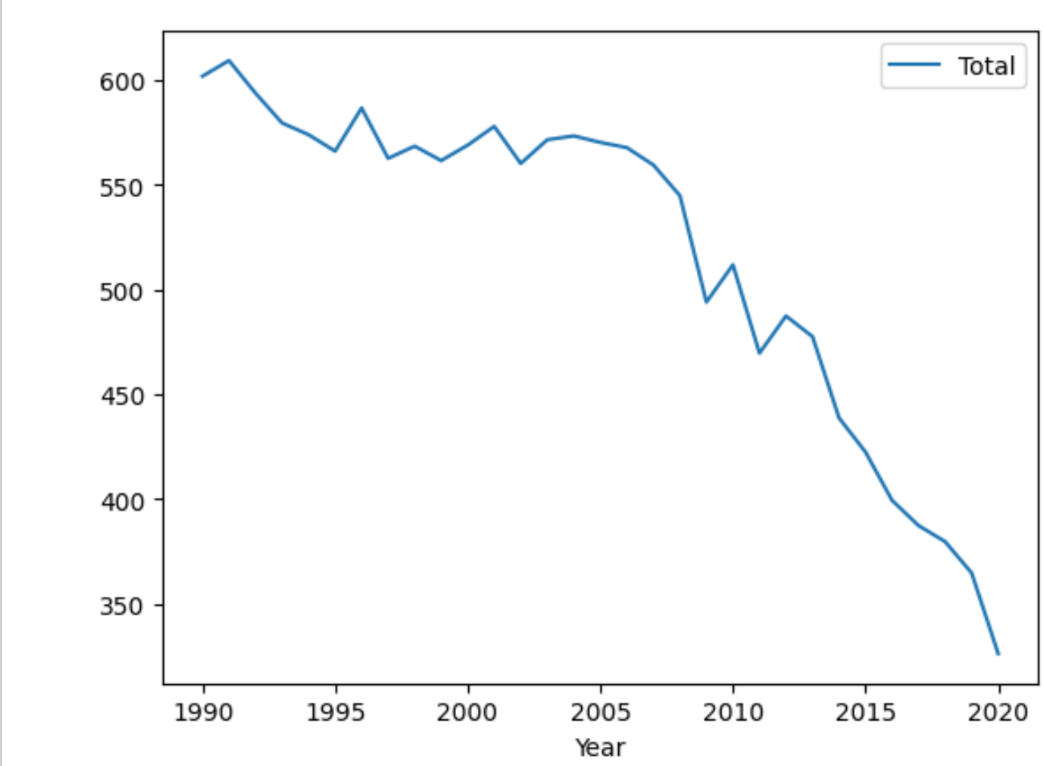
According to the data set, coal, oil, and cement are positively correlated with the total emissions, implying they are significant contributors to the UK's total emissions during these years.

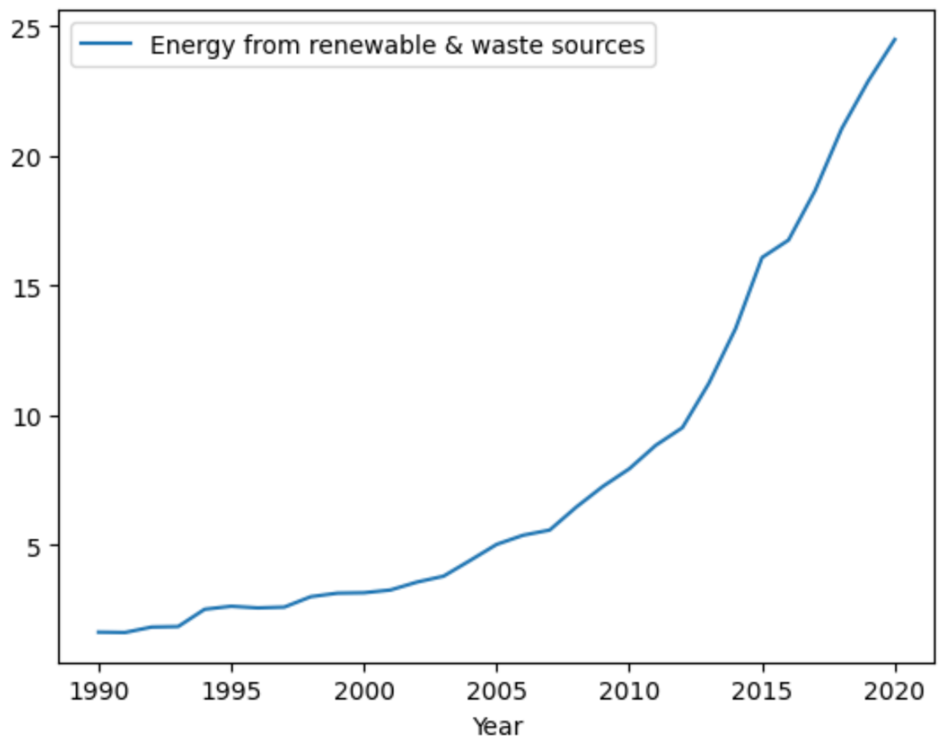
Gas shows a lower correlation of 0.21 with total emissions, which indicates that changes in gas usage might not have had as much of an impact on total emissions as changes in the usage of coal or oil.

'Energy from renewable & waste sources,' 'Fraction from renewable sources and waste,' and various renewable energy sources (like wind, wave, tidal, and solar photovoltaic) show a negative correlation with total emissions. This suggests that as the usage of these renewable sources increase, the total emissions likely decrease. However, renewable energy sources have less total energy compared to fossil fuels.

Liquid bio-fuels show a weak positive correlation with total emissions, implying that changes in their usage had less impact on total emissions.

Over the years, the trend is a reduction in total emissions, with the total amount decreasing from 601.945078 in 1990 to 326.263199 in 2020. Furthermore, there is a significant shift in the source of emissions. In 1990, a large portion of the emissions was from coal, but by 2020, the emissions from coal had significantly decreased, and Oil and Gas made up a larger share. This suggests efforts to shift away from coal as a source of energy.

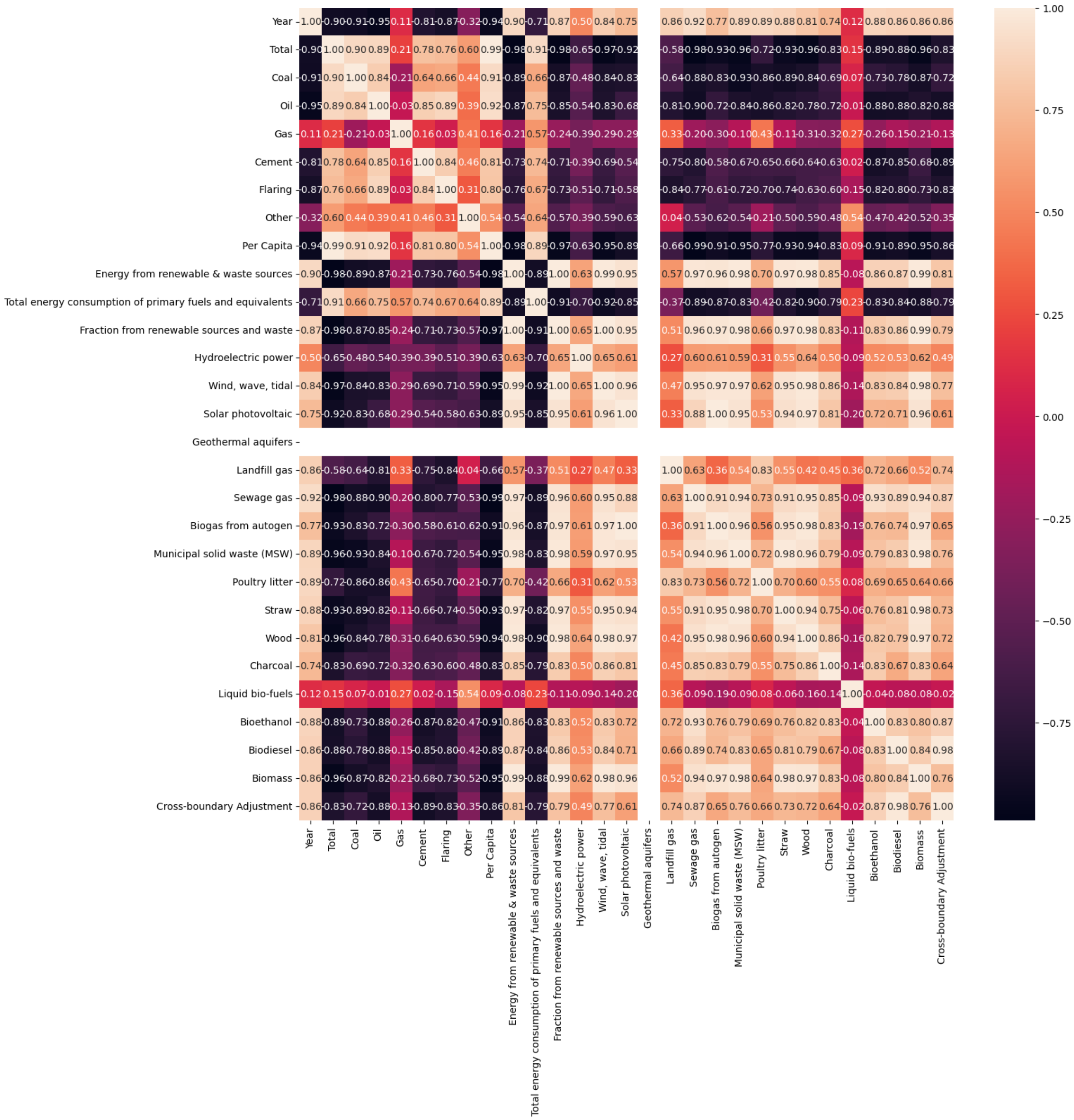




Based on the correlation heatmap derived from the 'merged_cleaned_data.csv' dataset, it is evident that coal, oil, and cement have a strong positive correlation with total CO2 emissions. This signifies their substantial contribution to the UK's total emissions over these years. Conversely, gas demonstrates a comparatively low correlation of 0.21 with total emissions, suggesting that fluctuations in gas usage may not have been as influential on the total emissions as those observed in coal or oil usage.

Interestingly, 'Energy from renewable & waste sources,' 'Fraction from renewable sources and waste,' and various renewable energy sources such as wind, wave, tidal, and solar photovoltaic exhibit a robust negative correlation with total emissions. This suggests that as the utilization of these renewable sources amplified, the total emissions correspondingly diminished. Renewable energy sources appear to possess less total energy than fossil fuels, indicating a possible trade-off between sustainability and energy yield.

Moreover, liquid biofuels display a weak positive correlation with total emissions, implying that variations in their usage have a smaller impact on total emissions. This information, visualized in the heatmap, provides a clearer understanding of the relationship between different energy sources and CO2 emissions in the UK.



A range of AI techniques has been developed to address these issues effectively. [2] The elements of statistical learning: data mining, inference, and prediction contain information on various data mining and machine learning methods. Supervised learning algorithms like linear regression, decision trees, random forests, and neural networks have been extensively used for predictions and pattern recognition. In contrast, unsupervised learning methods like clustering have been employed to uncover hidden patterns and relationships in data. Techniques like reinforcement learning are used in intelligent grid management, where the algorithm learns to make decisions (like load balancing and energy distribution) through a reward system.

The choice of tools to implement these techniques often depends on the problem's nature, the dataset's size, and the computation resources available. With its extensive libraries like Scikit-learn, TensorFlow, and PyTorch, Python is a widely adopted language in the AI domain. These libraries offer a range of pre-built functions for data preprocessing, machine learning, and evaluation. They also provide flexibility, ease of use, and compatibility with different data formats, making them suitable for various applications.

Literature Review

Evaluation of these AI techniques often involves using performance metrics that depend on the specific problem. For regression problems like predicting CO2 emissions, metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared are commonly used. These metrics quantitatively measure the model's ability to predict continuous values. In contrast, classification problems might use [3] accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic curve (AUROC) as evaluation metrics.

While the approaches discussed have significantly contributed to addressing the challenges in the energy sector, they come with limitations and challenges. For example, the efficacy of supervised learning models [4] heavily depends on the quality and quantity of the available data. They are also prone to overfitting, where the model performs exceptionally well on the training data but poorly on unseen data. Unsupervised learning models, while capable of discovering unknown patterns, might present interpretability challenges.

Deep learning models, such as neural networks, can capture highly complex relationships in data but are often criticized as "black-box" models due to their lack of interpretability. Also, their performance largely depends on the choice of hyperparameters, and tuning them can be computationally intensive. [5]On the other hand, simple models like linear regression might not capture the complexity of real-world data well.

In this research, we aim to construct two AI models using Python and Scikit-learn to assess the effect of renewable energy sources on CO2 emissions. The underlying concept of [2] random forests is to enhance the variance reduction of bagging by minimizing the correlation between the trees while simultaneously avoiding a significant increase in variance. This is accomplished during tree construction via the random selection of input variables.

Using the complete dataset, we will train our first model, the RandomForestRegressor. On the other hand, we will train our second model, the DecisionTreeRegressor, using a select set of pertinent features identified through the Recursive Feature Elimination (RFE) method. To evaluate the efficacy of both models, we will utilize a variety of metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared. A comparative analysis of these metrics from both models will help us establish the most effective approach.

Research Design

Assumptions about the given scenario include that the available data is reliable and comprehensive and that the features selected (i.e., 'Energy from renewable & waste sources,' 'Total,' 'Year') are the most relevant for predicting CO2 emissions. It has also assumed that the relationship between the features and the target variable is complex enough to benefit from a machine-learning approach and consistent enough for such a model to generalize effectively.

In the absence of information about data distribution, mean or median values could be used for data pre-processing imputation. Furthermore, data normalization may be used, especially considering the use of algorithms like Decision Trees and Random Forests, which could potentially be sensitive to the range of the data.

Recursive Feature Elimination (RFE) [2] was chosen as the feature selection technique due to its effectiveness in removing irrelevant features, improving the model's interpretability, and potentially its performance by reducing overfitting. RFE uses a model (in this case, RandomForestRegressor) to rank features by importance, eliminating important minor features one by one until the desired number of features is reached.

The supervised learning techniques selected were RandomForestRegressor and DecisionTreeRegressor. These techniques are good for regression tasks and can capture complex relationships in the data. [6]RandomForestRegressor, an ensemble method, is generally more robust and accurate than a single decision tree (DecisionTreeRegressor), as it averages the predictions of multiple decision trees trained on different subsets of the data. However, a single Decision Tree can offer better interpretability, as it can be easily visualized and understood.

The primary evaluation metric is the Mean Squared Error (MSE), which measures the average squared difference between the actual and predicted values, with lower values indicating better performance. The Mean Absolute Error (MAE) and R2 score are also calculated to provide a more comprehensive assessment. MAE gives a direct average measure of prediction error magnitudes, while R2 indicates the proportion of variance in the target variable that is predictable from the input features.

Experimental Results and Analysis

The experimental results show that the RandomForestRegressor (Model 1) and DecisionTreeRegressor (Model 2) models performed well, with high R2 scores indicating that the models could explain a large proportion of the variance in CO2 emissions. However, there was a clear difference in performance between the two models, with Model 1 outperforming Model 2.

Model 1 - The RandomForestRegressor model achieved a Mean Squared Error (MSE) of approximately 14.96, a Mean Absolute Error (MAE) of around 2.66, and an R2 Score of 0.998. This model used all available features in the dataset.

Model 2 - The DecisionTreeRegressor model, which used a subset of features selected using Recursive Feature Elimination (RFE), achieved an MSE of approximately 49.79, an MAE of around 4.86, and an R2 Score of 0.993.

Random Forests is a highly efficient predictive tool [7] offering robust performance without succumbing to overfitting, courtesy of the Law of Large Numbers. The method's unique approach, which injects a specific type of randomness, allows them to be accurate classifiers and regressors. Furthermore, the conceptual framework offers critical insights into the predictive capabilities of the Random Forest tool. The concrete manifestation of strength and correlation values through an out-of-bag estimation further reinforces these insights.

The RandomForestRegressor model (Model 1) outperformed the DecisionTreeRegressor model (Model 2) in predicting CO2 emissions. This performance is indicated by a lower MSE and MAE and a higher R2 Score for Model 1, suggesting that the RandomForestRegressor is more adept at dealing with the intricate and non-linear relationships.

The improvement in the MSE score (the difference between the MSE of Model 1 and Model 2) was negative, indicating that Model 1 performed better than Model 2. This outcome also suggests that the RFE feature selection method could have been better for this problem. Although RFE is a robust method for feature selection, it may have eliminated some essential features for prediction in this case.

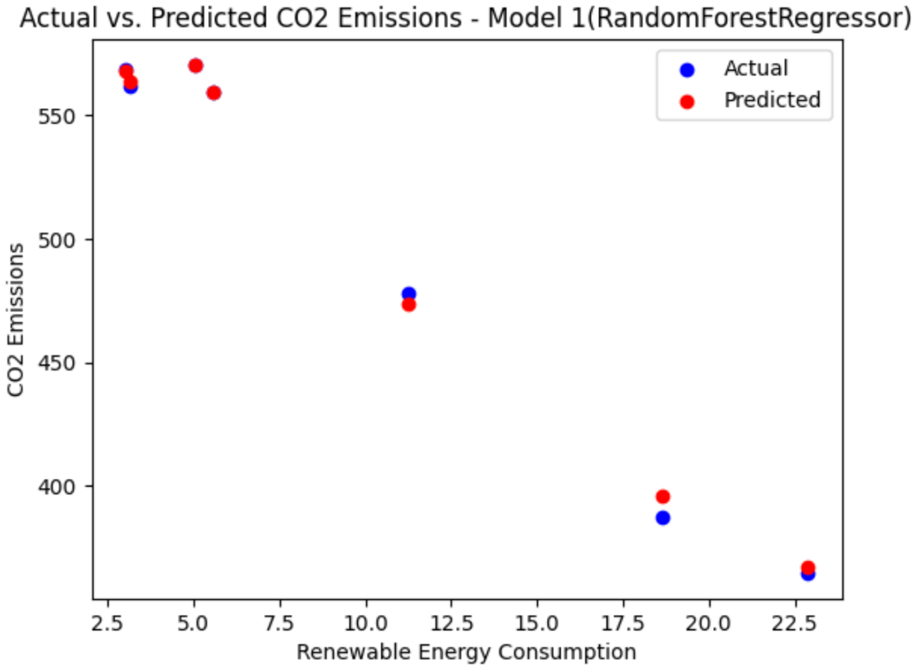
In terms of business implications, these results can provide valuable insights into the impact of green energy on CO2 emissions. As businesses look to reduce their carbon footprint, understanding the variables contributing to CO2 emissions can guide decisions regarding energy use. For instance, if energy from renewable and waste sources is found to be a significant variable, businesses might consider investing more in these types of energy sources.

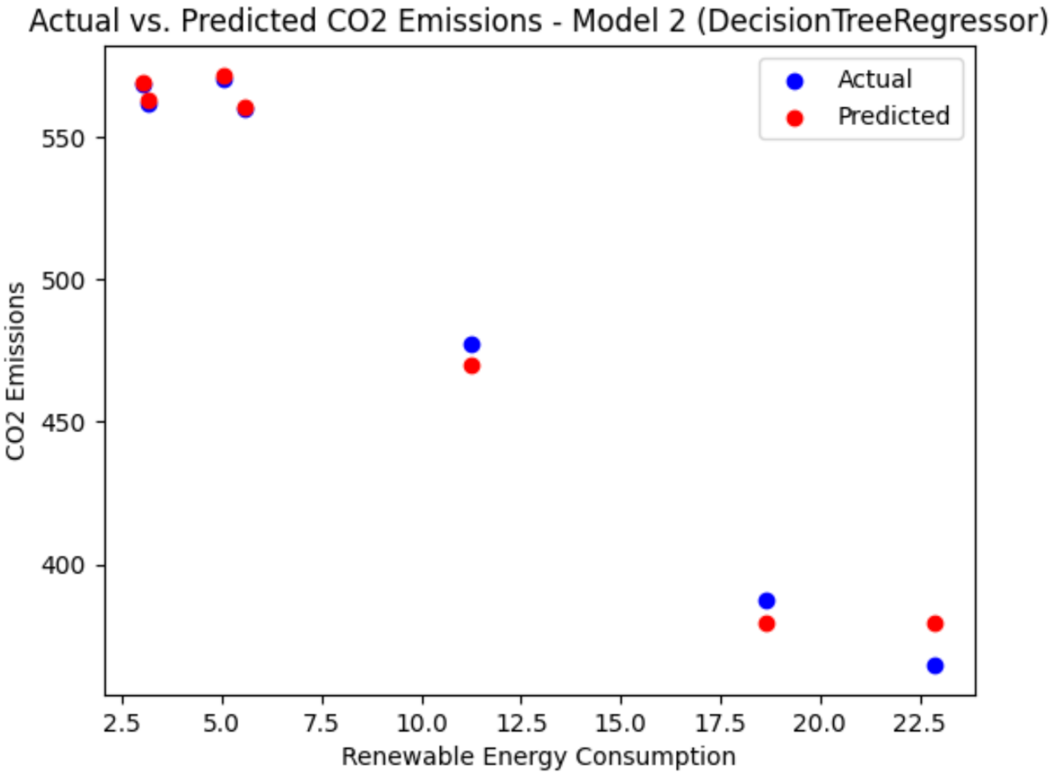
However, some things could be improved with this approach. One is the assumption of linear relationships between the models' features and target variables. In reality, these relationships may be non-linear or influenced by other factors not included in the model. Moreover, the use of RFE for feature selection, while convenient, may have resulted in a different set of features for this problem, as evidenced by the performance of Model 2. Further research and more advanced feature selection methods might yield improved results.

In future work, explore other feature selection methods, such as those based on optimization algorithms like genetic algorithms or simulated annealing. These methods can uncover complex relationships between features, improving the model's predictive power.

In conclusion, the RandomForestRegressor and DecisionTreeRegressor models could predict CO2 emissions well. However, there is potential for improvement by using more advanced feature selection methods and addressing the abovementioned limitations.

```
Model 1 - MSE: 14.961531562949702 MAE: 2.657506968571543 R2 Score: 0.9978387708051486
Model 2 - MSE: 49.785838927153364 MAE: 4.863924857142861 R2 Score: 0.9928083159049046
MSE Improvement: -34.82430736420366
```





Conclusion

In this study, we utilized RandomForestRegressor and DecisionTreeRegressor models to investigate the impact of green energy on CO2 emissions. These supervised learning techniques enabled us to generate predictive models, and through evaluation, we discerned that the RandomForestRegressor model significantly outperformed the DecisionTreeRegressor model.

The RandomForestRegressor model yielded an MSE of approximately 14.96, an MAE of around 2.66, and an R2 score of 0.998. In contrast, the DecisionTreeRegressor model achieved an MSE of approximately 49.79, an MAE of around 4.86, and an R2 score of 0.993. This disparity in performance signifies that the RandomForestRegressor model was more adept at handling the complexity and non-linear relationships in the data.

These findings have significant implications, especially for organizations aiming to mitigate their carbon footprint. Our results can provide valuable insights for these entities, helping them understand the variables that substantially influence CO2 emissions. Specifically, they may realize the importance of investing in renewable and waste energy sources if these sources are found to influence emissions significantly.

However, we should address a few limitations for future studies. The assumption of linearity between features and the target variable is a prime concern that should be evaluated further, as real-world relationships may be more complex. Additionally, the RFE feature selection technique might not be optimal for this problem, as demonstrated by the poorer performance of Model 2.

As a recommendation, future investigations could benefit from adopting advanced feature selection techniques such as genetic algorithms or simulated annealing. These techniques enhance the model's predictive power by uncovering complex relationships between features.

In conclusion, the study demonstrated that while both models could predict CO2 emissions to some extent, there is room for improvement. Addressing the mentioned limitations and recommendations could lead to more robust and precise predictions in future studies.

REFERENCES

[1] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In Emerging artificial intelligence applications in computer engineering (pp. 3-24). IOS Press.
[2]Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
[3]Lee, J. (2021). Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. PLoS One, 16(12), e0261737.
[4]Which Factors Determines Economic Growth ? – Aim Institute of Economics. https://ecoaim.in/2020/07/07/which-factors-determines-economic-growth/
[5]Molnar, Christoph. 2022-02-06, Interpretable machine learning, Second edition, https://christophm.github.io/interpretable-ml-book/
[6]Techniques to Boost True Positive Rates using Independent Combinatorics - Hyperspec AI. https://hyperspec.ai/techniques-to-boost-true-positive-rates-using-independent-combinatorics/
[7]Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.