Applied Artificial Intelligence

SUMMATIVE ASSESSMENT BRIEF

Executive Summary

In this project, our primary objective was to investigate the impact of renewable energy consumption on carbon dioxide (CO2) emissions, a critical issue in the green energy sector. Our focus was to leverage artificial intelligence techniques, specifically supervised learning, to predict CO2 emissions based on renewable energy consumption data, and understand how increasing renewable energy usage can contribute to reducing greenhouse gas emissions.

Python, alongside its library Scikit-learn, was the primary tool utilized for developing and evaluating predictive models. The Scikit-learn library offers 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, improving accuracy, and reducing 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) exhibited better performance across all these metrics, suggesting that it fits the data better. Visualization of actual vs. predicted CO2 emissions further reinforced this conclusion.

In future studies, the exploration of other feature selection methods and machine learning algorithms could potentially yield improved results. Handling imbalanced data, if present, 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 that could influence 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 on 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 the integration and optimization of renewable energy sources to reduce CO2 emissions. This is not only crucial for meeting international climate targets but also for the sustainable development of global economies. Artificial Intelligence (AI) has emerged as a transformative tool in 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.

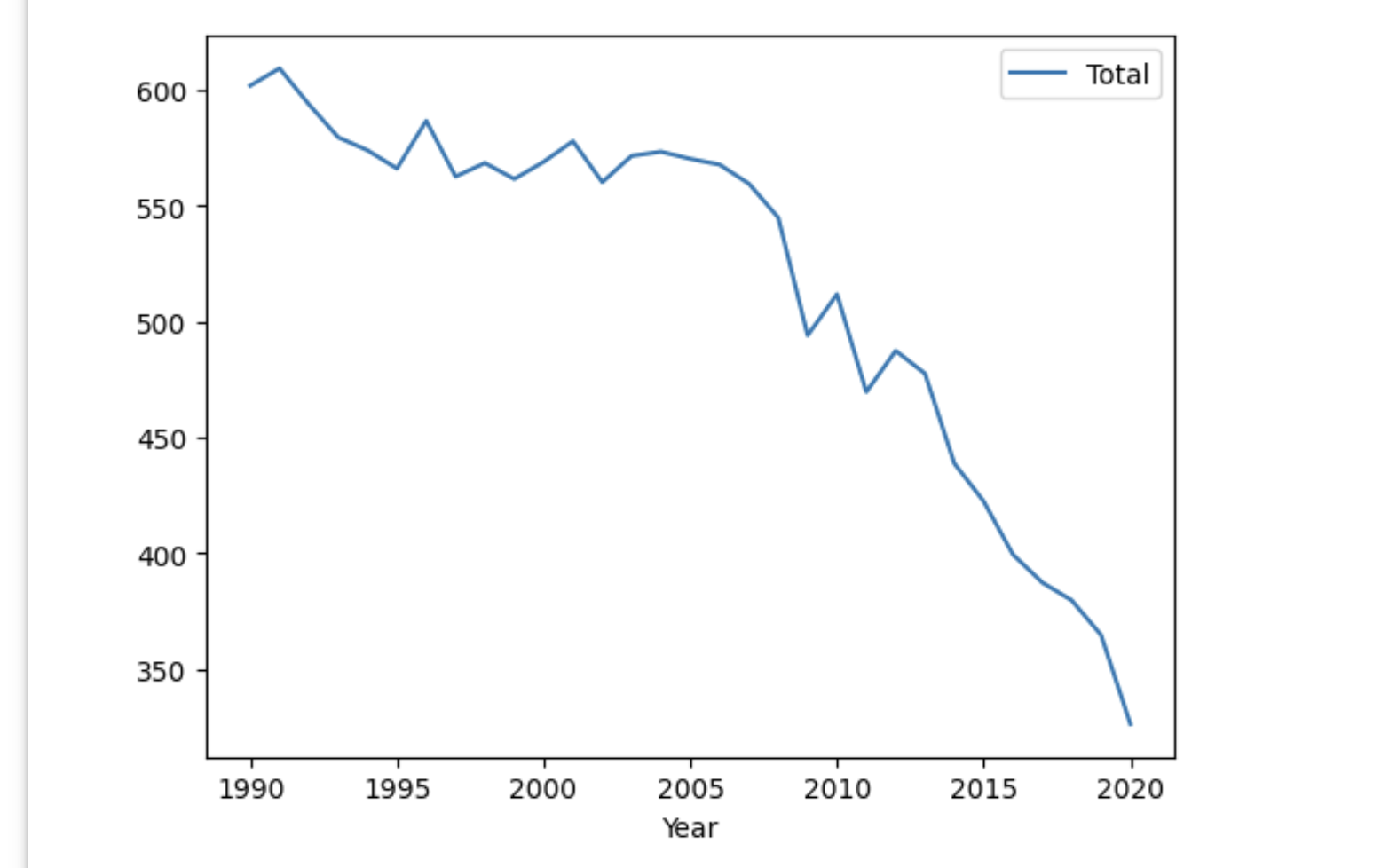
According from the data set, Coal, oil, and cement are strongly positively correlated with the total emissions, implying they are significant contributors to the UK's total emissions during these years.

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.

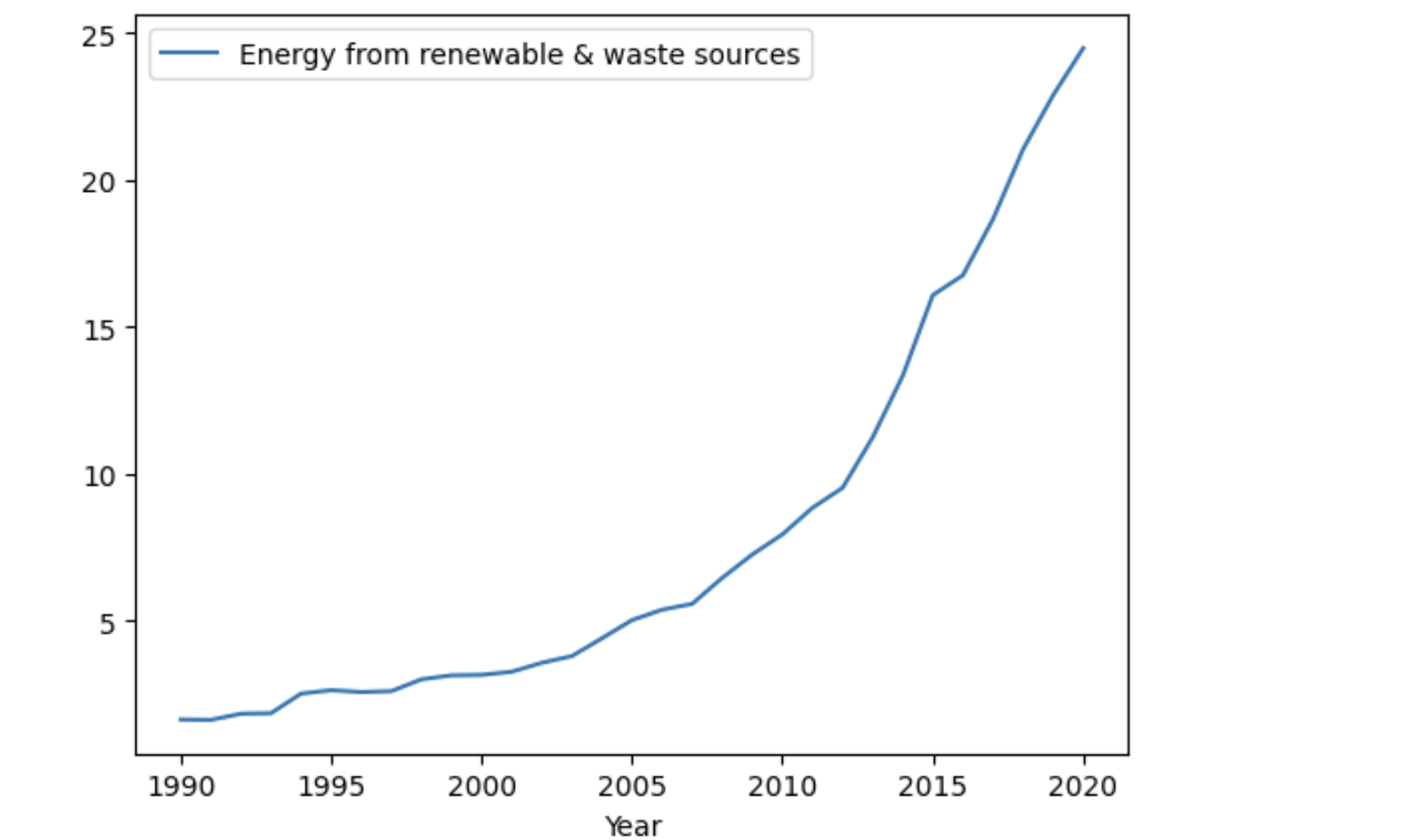
'Energy from renewable & waste sources', 'Fraction from renewable sources and waste', and various types of renewable energy sources (like wind, wave, tidal; solar photovoltaic; etc.) show a strong negative correlation with total emissions. This suggests that as the usage of these renewable sources increased, the total emissions likely decreased. However, the renewable energy sources seem to have less total energy than fossil fuels.

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

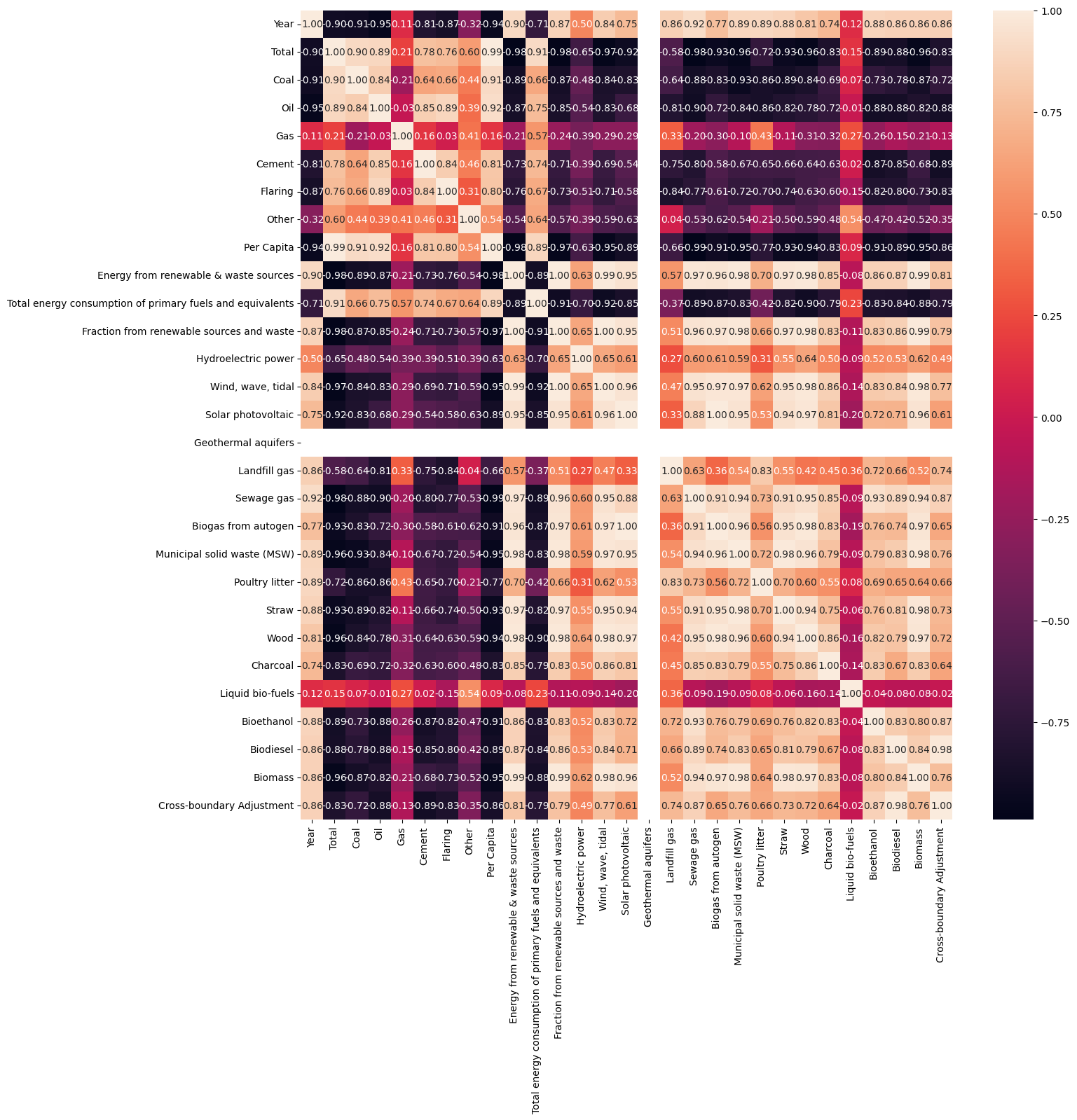
Over the years, the trend seems to be a reduction in total emissions, with the total amount decreasing from 601.945078 in 1990 to 326.263199 in 2020. Furthermore, there appears to be a significant shift in the source of emissions. In 1990, a large portion of the emissions was from Coal, but by 2020,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



Moreover, you can see an increase in biomass usage towards the end of the time series. It increased from 0.065 in 1990 to 5.478 in 2020, indicating a move towards more sustainable energy sources.



Coal, oil, and cement are strongly 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 types of renewable energy sources (like wind, wave, tidal; solar photovoltaic; etc.) show a strong negative correlation with total emissions. This suggests that as the usage of these renewable sources increased, the total emissions likely decreased. However, the renewable energy sources seem to have less total energy than fossil fuels. Liquid bio-fuels show a weak positive correlation with total emissions, implying that changes in its usage had less impact on total emissions.



A range of AI techniques has been developed to address these issues effectively. [2] The elements of statistical learning: data mining, inference, and prediction contains a wealth of 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 being used in smart 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 nature of the problem, the size of the dataset, and the computation resources available. Python, with its extensive libraries like Scikit-learn, TensorFlow, and PyTorch, 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 the use of performance metrics that depend on the specific problem at hand. 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 provide a quantitative measure of the model's ability to predict continuous values. In contrast, classification problems might use 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 their set of limitations and challenges. For example, the efficacy of supervised learning models 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 is largely dependent on the choice of hyperparameters, and tuning these hyperparameters can be computationally intensive. On the other hand, simple models like linear regression might not capture the complexity of real-world data well.

In this study, we propose to build two AI models using Python and Scikit-learn to evaluate the impact of renewable energy sources on CO2 emissions. The first model, a RandomForestRegressor, will be trained on the full dataset, while the second model, a DecisionTreeRegressor, will be trained on a subset of features deemed relevant through a Recursive Feature Elimination (RFE) method. The performance of both models will be evaluated using several metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared, and compared to determine the optimal approach.

Task 3 - Research Design:

Assumptions made 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's also assumed that the relationship between the features and the target variable is complex enough to benefit from a machine learning approach, but also consistent enough for such a model to generalize effectively.

Data preprocessing would initially involve checking for missing values, with suitable imputation strategies employed if necessary. In the absence of information about the distribution of data, mean or median values could be used for imputation. Furthermore, data normalization may be applied, 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) 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 the least important features one by one until the desired number of features is reached.

The supervised learning techniques selected are RandomForestRegressor and DecisionTreeRegressor. These techniques are suitable for regression tasks and can capture complex relationships in the data. 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 visualized and understood more easily.

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. Additionally, the Mean Absolute Error (MAE) and R2 score are 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.

Task 4 - Experimental Results and Analysis:

The experimental results show that both 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, a MAE of around 4.86, and an R2 Score of 0.993.

The RandomForestRegressor model (Model 1) thus performed considerably better than the DecisionTreeRegressor model (Model 2), as evidenced by the lower MSE and MAE and the higher R2 Score. This suggests that the RandomForestRegressor is better equipped to handle the complexity and the non-linear relationships in the data.

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 might not have been the optimal approach for this problem. Although RFE is a robust method for feature selection, in this case, it seems that it might have eliminated some features that were important for prediction.

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 that contribute 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, there are some limitations to this approach. One of them is the assumption of linear relationships between the features and the target variable by the models. 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 not have resulted in the optimal 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, it would be worthwhile to explore other feature selection methods, such as those based on optimization algorithms like genetic algorithms or simulated annealing. These methods can potentially uncover more complex relationships between features, improving the model's predictive power.

In conclusion, both the RandomForestRegressor and DecisionTreeRegressor models were able to predict CO2 emissions reasonably well, but there is potential for improvement by using more advanced feature selection methods and addressing the limitations noted above.

A screen shot of a graph

Description automatically generated with medium confidence

A graph with red and blue dots

Description automatically generated with low confidence

Task 5 - 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, MAE of around 2.66, and an R2 score of 0.998. In contrast, the DecisionTreeRegressor model achieved an MSE of approximately 49.79, 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 useful 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 significantly influence emissions.

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. By uncovering complex relationships between features, these techniques may enhance the model's predictive power.

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.