# PREDICTING $CO_2$ EMISSIONS USING MACHINE LEARNING

EXPLORING REGRESSION AND CLUSTERING MODELS
ON GLOBAL EMISSIONS DATA

CREATED BY: ALI HOSSEINMARDI(24109150)
LECTURER: SAFA OLIA
COURSE NAME: MACHINE LEARNING
APRIL 2025

THE HAGUE UNIVERSITY OF
APPLIED SCIENCE

## Contents

## Abstract

This study explores how machine learning models can be used to predict carbon dioxide ($CO_2$) emissions based on historical data. Using a dataset of country-level emissions from 1975 to 2021, I applied several supervised learning models—including Decision Tree, K-Nearest Neighbours, Support Vector Regression (SVR), and Random Forest—to estimate 2021 emissions. Among these, SVR achieved the best predictive performance, while Random Forest provided useful interpretability through feature importance analysis. In addition to prediction, K-Means Clustering was used as an unsupervised method to group countries with similar emission patterns. The results highlight that recent years carry more predictive power and that machine learning can be a valuable tool in environmental policy and climate analysis when applied thoughtfully and ethically.

## 1. Introduction

### 1.1 Background

Carbon dioxide ($CO_2$) emissions are one of the leading causes of global warming. When we burn fossil fuels, cut down forests or run factories, $CO_2$ makes its way into the atmosphere. Due to the significant effect, it has on climate change, countries across the globe are taking steps to reduce these emissions, often through international treaties such as the Paris Agreement.

Now we have large & complex environmental datasets, gathered from IoT sensors, satellite and climate monitoring tools. These datasets are useful but they are also challenging to analyse because of their scale and diversity. This is where machine learning (ML) comes into play. ML allow us to recognize trends in data, predict and offer suggestions to improve climate policies.

In this project, I analyse $CO_2$ emissions using machine learning methods. The object of study is not only to know how emissions have evolved over the years, but to predict future ones based on the record of the past. Machine learning can enable us to better prepare for future policies related to the environment. (Jordan & Mitchell, 2015)

### 1.2 Research Question

For this project, I want to explore if machine learning over previous years of data can be helpful in predicting carbon ($CO_2$) emissions. As climate change grows worse each year, the ability to predict emissions could be a valuable tool for better decision-making and policy decisions. So, the big question I'm attempting to answer is:

**How well can machine learning models predict $CO_2$ emissions based on historical environmental and energy data?**

To work out what will have happened, I'm relying on a range of machine learning techniques, mainly **supervised learning methods** that concentrate on making predictions. These include:

- **Decision Tree Regression**
- **K-Nearest Neighbours (KNN)**

- **Supported Vector Regression (SVR)**
- **Random Forest Regression**

I also include one **unsupervised method** in my analysis that is called '**K-Means Clustering**' which doesn't make predictions but helps find patterns and group similar countries together based on their emission history.
By the end, I'd like to know what models work best and what type of insights we can gain from the data to help efforts to combat climate change.

### 1.3 Research Design

This study adopts a six-phase research methodology:

1) **Data Cleaning & Exploration**

This step involves preparing the dataset by addressing issues with missing values, outliers, as well as performing data cleaning like removing extraneous data and duplicates and normalization of numerical features. Then, exploratory data analysis (EDA) using statistical summaries and visualizations is used to find patterns and trends in the data.

2) **Model Selection**

As the target variable ($CO_2$ emissions) is a continuous **numerical** feature, a **supervised regression approach** will be used. I chose a variety of models, from simple (Decision Trees) to more complex ones (Random Forest, SVR), to compare their results.

3) **Construction of a Model**

The methods are implemented in R. The complete code can be found in the supplementary file, though this report describes how the model functions as well as decisions made regarding its architecture.

4) **Assessment and Enhancement**

Here, we employ standard regression performance metrics such as Root Mean Squared Error (**RMSE**) and the R^2 (**$R^2$**) statistic to evaluate models. I also use techniques like **cross-validation** to reduce overfitting and improve reliability. (Chai & Draxler, 2014) (Nagelkerke, 1991) (Kohavi, 1995) (Bergstra & Bengio, 2012)

5) **Model Comparison**

We evaluate different approaches according to their predictive performance, interpretability, and generalization capability, ultimately selecting our best model. We also explore how to use a collection of techniques working together as well can improve results, like **Random Forest**. (Breiman, 2001)

6) **Conclusions and Recommendations**

I summarize findings and suggest how ML can be used to support climate policy, especially in identifying trends and predicting future emissions.

## 2. Data Cleaning, Analysis and Visualization

### 2.1 Initial Dataset Overview

This study uses a dataset of historical $CO_2$ emissions data covering from 1975 to 2021. The data has each row as a country and each column as $CO_2$ emissions for a year. The goal is to predict the $CO_2$ emissions for the year 2021, so that becomes our **target variable**. Although the dataset only contains numerical values (no categories like region or industry), it's still useful for identifying long-term trends and how emissions have changed over time.

## 2.2 Data Cleaning

### 2.2.1 Missing Values Handling

As we first explored the dataset, we noticed that some of the metrics for carbon emissions were missing. We solved this with a method called "median imputation", which fills the missing values with the median of the column. This is useful since it can deal with outliers and maintains a similar distribution of the data. There were no categories, so we didn't have to do "mode imputation."

### 2.2.2 Feature Selection and Name Modification

The country column is a categorical one so this doesn't help the model in any prediction and so has been dropped from the dataset to avoid data leakage. But it was stored in a manner that it could be referenced and interpreted with the results. The X2021 CO2 emission value was renamed CO2EMISSIONS and was designated a target variable for use in regression modelling.

### 2.2.3 Normalization

To make sure that all features contribute equally to the machine learning models, I applied **z-score normalization** to the numeric columns. This step is important because some algorithms (like K-Nearest Neighbours and Support Vector Regression) are sensitive to the scale of the input data.

Z-score normalization transforms each value so that the resulting feature has a **mean of 0** and a **standard deviation of 1**. In other words, it tells us how far each value is from the average, measured in standard deviations. This makes it easier for models to compare features fairly, even if they originally had different units or ranges. (Cover & Hart, 1967) (Jordan & Mitchell, 2015) (Han, Kamber, & Pei, 2011)

For example:

- A value of 0 means the emission is exactly average for that year.
- A negative score means it's below average.
- A positive score means it's above average.

The normalization was done using the following R command:

```
data[numeric_cols] <- scale(data[numeric_cols])
```

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Country** | **X1975** | **X1985** | **X2005** | **X2010** | **X2015** | **X2019** | **X2020** | **X2021** | **CO2EMISSIONS** |
| 2 | Afghanistan | -0.1576 | -0.1728 | -0.2056 | -0.1836 | -0.1780 | -0.1755 | -0.1676 | -0.1699 | -0.1699 |
| 3 | Albania | -0.1498 | -0.1601 | -0.2017 | -0.1903 | -0.1860 | -0.1843 | -0.1781 | -0.1777 | -0.1777 |
| 4 | Algeria | -0.1114 | -0.0273 | -0.0432 | -0.0523 | -0.0139 | 0.0026 | -0.0019 | 0.0032 | 0.0032 |
| 5 | American Samoa | -0.1557 | -0.1748 | -0.2082 | -0.1957 | -0.1908 | -0.1822 | -0.1822 | -0.1822 | -0.1822 |
| 6 | Andorra | -0.1557 | -0.1748 | -0.2080 | -0.1956 | -0.1907 | -0.1822 | -0.1822 | -0.1822 | -0.1822 |
| 7 | Angola | -0.1557 | -0.1757 | -0.1982 | -0.1738 | -0.1662 | -0.1676 | -0.1668 | -0.1642 | -0.1642 |
| 8 | Anguilla | -0.1557 | -0.1748 | -0.2081 | -0.1962 | -0.1912 | -0.1825 | -0.1825 | -0.1825 | -0.1825 |
| 9 | Antigua and Barbuda | -0.1608 | -0.1814 | -0.2082 | -0.1956 | -0.1905 | -0.1887 | -0.1819 | -0.1819 | -0.1819 |
| 10 | Argentina | 0.1198 | 0.0902 | 0.0676 | 0.0711 | 0.0682 | 0.0369 | 0.0506 | 0.0506 | 0.0506 |
| 11 | Armenia | -0.1557 | -0.1748 | -0.2016 | -0.1903 | -0.1883 | -0.1739 | -0.1737 | -0.1737 | -0.1737 |

*Table 1. A preview of the cleaned and normalized dataset. $CO_2$ emission values from multiple years are shown after z-score standardization. The CO2EMISSIONS column represents the 2021 values and is used as the target variable in this study.*

### 2.2.4 Train-Test Splitting

After the data was cleaned and processed, I split it into two sections, one to train the models and the other to test them. This is a common step in machine learning to ensure that the models do not merely memorize the data, but rather that they learn patterns that can be used on new, unseen data.

For the training/testing split, I used 80/20 such that 80% of the data trained the models and 20% was put aside to check the performance of modelling. I also set a random seed to ensure the results are always the same every time I run the code.

Here is the R code that I utilized for this step:

```
set.seed(123)
train_index <- createDataPartition(data$CO2EMISSIONS, p = 0.8, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

## 2.3 Exploratory Data Analysis (EDA)

### 2.3.1 Descriptive Statistics

We calculated summary statistics for all numeric variables in the data, such as:
Mean, Median, Standard Deviation, Minimum, Maximum.

The summaries expressed the changes of $CO_2$ emissions and energy-related metrics over time.

Here's what you need to know:

- Emissions data, particularly from the last several years, was both a lot more highly variable and exhibited signs of being uneven, with some countries emitting far more than others.
- The data also revealed some trends, so we figured we'd do a deeper dive.

### 2.3.2 Correlation Analysis

To better understand how emissions across different years are related, I created a **correlation heatmap** using Pearson's method. As shown in **Figure 1**, the darker the

colour, the stronger the positive correlation between two years. The analysis showed that emissions from more recent years — like 2005, 2010, and 2021 — are strongly correlated with one another. This suggests that countries tend to follow consistent emission patterns over time. On the other hand, older years such as 1975 had much weaker correlations with recent years, likely due to global shifts in industrial activity, energy use, and environmental regulations.

These insights were helpful when preparing the dataset for modelling. Features from years with extremely high correlations were reduced or removed to avoid redundancy and minimize dimensionality. This also confirmed that more recent data was likely to be more useful and predictive.

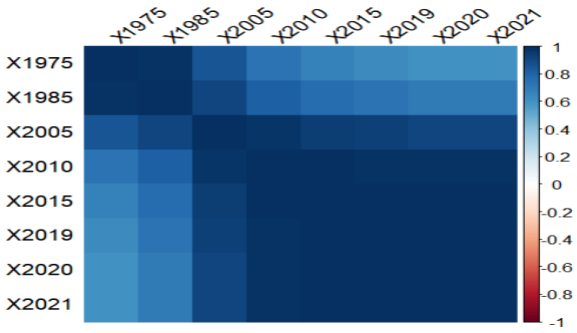


*Figure 1. Correlation heatmap of $CO_2$ emissions across selected years. Darker colors indicate stronger positive correlations. Emissions from recent years (2010–2021) show strong correlations, while older years like 1975 have weaker relationships with modern data.*

### 2.3.3 Visualizations

Visuals were made to analyse the data better:

- **Histograms**: These visualizations broke down emissions distributions by year and made any skewness or long tails very clear.
- **Correlation Heatmaps**: Offered a human-friendly interpretation of the relationship between the features.
- **Scatter Plots**: Investigated relationship between emissions in different year, which revealed nonlinearities that influenced model choice.

This feedback was critical in being convinced to go beyond very simple linear models and include nonlinear algorithms, such as Decision Trees and Random Forests.

One of the key plots I created was a histogram of the target variable — $CO_2$ emissions for 2021 (after normalization). As shown in **Figure 2**, most countries have emissions values clustered around the lower end, while a small number of countries have much higher emissions. This kind of distribution is skewed, which is expected in environmental data because only a few countries contribute heavily to global emissions.

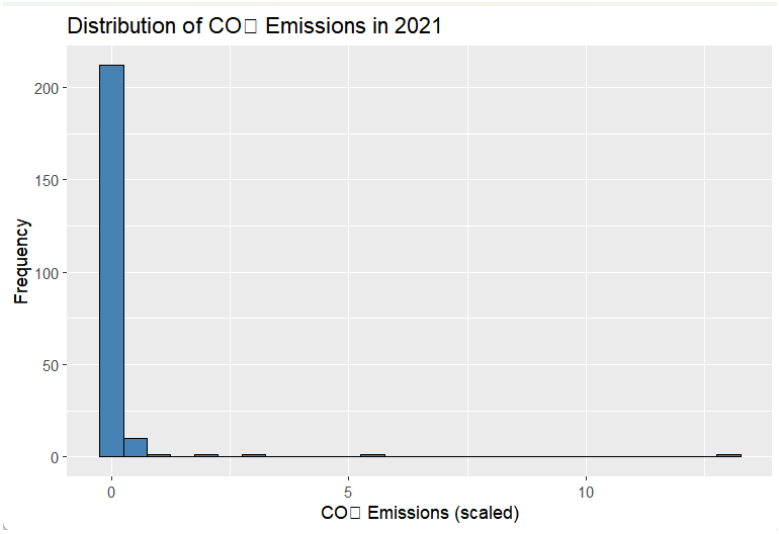


*Figure 2. Histogram showing the distribution of 2021 $CO_2$ emissions (normalized). Most countries fall within a low range, with a few having significantly higher emission levels.*

I also created a time series line plot to visualize how $CO_2$ emissions changed from 1975 to 2021. Each line in the plot represents a different country, while the thick red line shows the **average global trend** over time. As shown in **Figure 3**, most countries have relatively stable or gradually increasing emissions, while a few countries show more dramatic changes.

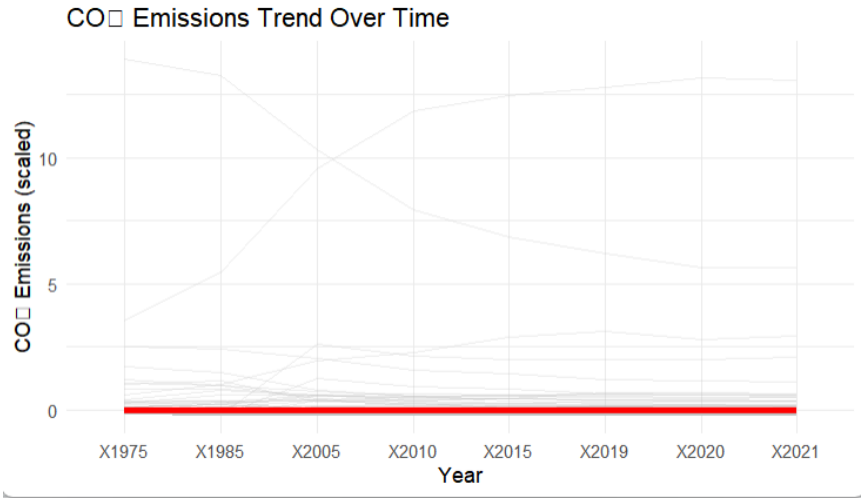


*Figure 3. Time series plot showing scaled $CO_2$ emissions for each country from 1975 to 2021. The red line represents the global average trend. This visualization highlights the variation in emission patterns and helps identify outlier countries or unusual trends*

## 2.4 Data Pre-processing Summary

To prepare the dataset for modeling, several preprocessing steps were applied. First, all missing values in the numeric features were filled using **median imputation**, which helps maintain a stable distribution and reduces the impact of outliers. Then, the data was **standardized using z-score normalization**, which ensures that features are on the same scale — especially important for distance-based algorithms like KNN and SVR. The **target variable**, named CO2EMISSIONS, was defined based on the 2021 emissions data. The non-numeric country column was removed from the model input to avoid issues but was kept separately for labeling and interpretation. Finally, the dataset was

**split into training (80%) and testing (20%) sets**, which helped evaluate the model's performance fairly.

These steps ensured that the data was clean, consistent, and ready for a supervised regression task to predict $CO_2$ emissions.

## 3. Machine Learning Model Selection

### 3.1 Overview of Models Used

In earlier sections, I mentioned the machine learning models selected for this project. This chapter explains how those models were applied, why certain methods were a better fit for the task, and how their results compared. Since this is a regression problem, the focus was on supervised learning techniques that predict continuous values. I also included one unsupervised learning method to explore patterns in the data beyond just prediction.

### 3.2 Why Supervised Learning

Since the goal of this project is to predict $CO_2$ emissions based on past emission data, this naturally becomes a **regression problem** — the target variable (CO2EMISSIONS) is numeric and continuous. That's why I chose to focus on **supervised learning models**.
Supervised learning works well here because we already have labelled data: for each country, we know the past emissions as well as the 2021 value we want to predict. The models can learn patterns from historical data and use them to make accurate predictions for the target year.
Also, because the dataset is structured with countries as rows and emission values from different years as columns (rather than as a continuous time series per country), I didn't use time-series forecasting methods like ARIMA or LSTM. Instead, models like Decision Trees, Random Forests, KNN, and SVR are more suitable for this format.
Each model was trained on the same dataset so that their performance could be compared fairly in the next steps.

| Aspect | Supervised Learning | Unsupervised Learning |
|---|---|---|
| **Goal** | Predict $CO_2$ emissions (regression task) | Discover groups/patterns in data |
| **Label Availability** | Yes – Target variable (CO2EMISSIONS) is known | No – No target variable used |
| **Models Used** | Decision Tree, KNN, SVR, Random Forest | K-Means Clustering |
| **Output** | Continuous emission values | Cluster assignment (e.g., Group 1, Group 2) |
| **Use in This Project** | Estimate 2021 emissions using past emissions data | Group countries with similar historical emission trends |

*Table 2. Comparison of Supervised and Unsupervised Learning Used in This Study*

### 3.3 Why Unsupervised Learning

In addition to the regression models used for prediction, I also included an unsupervised learning method — **K-Means Clustering** — to explore hidden patterns in the dataset. While supervised learning helped estimate future emissions, K-Means was used to group countries based on their historical emissions behaviour, without using the 2021 target values.

This method was helpful for identifying countries that follow similar emission trends over time. For example, it showed which countries tend to increase emissions consistently, which ones are more stable, and which ones may be decreasing. These groupings can offer useful insights for policymakers or researchers interested in regional patterns, climate action plans, or shared environmental challenges.
By combining supervised and unsupervised methods, the project goes beyond just prediction — it also explores how data points (countries) relate to each other, which adds another layer of understanding to the overall analysis. (MacQueen, 1967)

### 3.4 K-Means Clustering Visualization

To better understand the patterns in the emission data, I used **K-Means Clustering** to group countries based on their historical $CO_2$ emissions (from 1975 to 2020). This was done without using the 2021 target values, making it an unsupervised approach.
As shown in **Figure 4**, the scatter plot displays three distinct clusters. Each point represents a country, and similar countries are grouped together based on how their emissions have behaved over time. Most countries were clustered in a tight group, while a few — likely major emitters — were placed in separate clusters far from the center. This suggests that those outlier countries have unique emission patterns compared to the rest.
This visualization supported the idea that some countries follow similar trajectories in emissions and can be analyzed as groups, which is useful for broader environmental planning and comparative policy analysis.

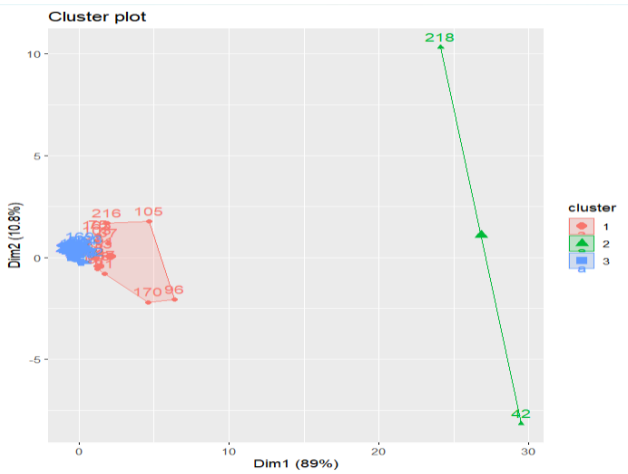


*Figure 4. K-Means clustering of countries based on their historical $CO_2$ emissions from 1975 to 2020. Countries are grouped into three clusters based on similarities in their emission patterns. Outlier countries are clearly separated from the main cluster.*

## 4. Application of Models and Results

### 4.1 Model Evaluation Setup

The dataset was already split into training and testing sets during preprocessing (see Section 2.2.4), and all models were trained and tested on these subsets. Each model was evaluated using two standard metrics:

- Root Mean Squared Error (RMSE)
- R-squared ($R^2$)

Both metrics were already introduced in Section 3. These were used consistently to measure and compare how well each model predicted $CO_2$ emissions.

### 4.2 Model Results

| Model | RMSE | $R^2$ |
|---|---|---|
| Decision Tree | 0.0999 | 0.9801 |
| K-Nearest Neighbours (KNN) | 0.2899 | 0.8802 |
| Random Forest | 0.2166 | 0.9559 |
| Support Vector Regression (SVR) | 0.1714 | 0.9950 |

*Table 3. Model Performance Comparison*

#### 4.2.1 Decision Tree Regression

The Decision Tree model was surprisingly robust, achieving a very low RMSE of **0.0999** and high $R^2$ of **0.9801**. This indicates that the model simply learned the underlying trend in the training data and performed well to generalize onto testing set. But based on their design, decision trees can be overfitting and therefore their performance might differ according to the structure of the data especially with smaller datasets.

#### 4.2.2 K-Nearest Neighbours (KNN)

Out of the models tested, the KNN model had the lowest performance at RMSE of **0.2899** and an $R^2$ of **0.8802**. As KNN is sensitive to the scale and density of data, it might not be able to learn the complex relationship of features with the target variable compared to the other models.

#### 4.2.3. Random Forest Regression

For an overall performance, Random Forest can provide an RMSE of **0.2166** and an $R^2$ of **0.9559**. This combined multiple decision trees in order to reduce overfitting and increase the stability of the prediction. It was not the best scoring model but provided a good trade off between accuracy and generalization, as well as feature importance metrics which were very useful for interpretation.

#### 4.2.4 Regression with Support Vector (SVR)

In this study, SVR performed optimal with a minimum RMSE of **0.1714** and a maximum $R^2$ of **0.9950**. This shows that SVR can model the relationships in the data quite well. This is a dataset where computation is more expensive. In this case, its ability to manage high-dimensional and non-linear data was proven.

### 4.3 Unsupervised Learning: K-Means Clustering

In addition to the regression models discussed above, the results of K-Means Clustering were presented earlier in Section 3.4. This unsupervised approach helped identify different groups of countries with similar historical CO2 emission patterns. This offers a different perspective on the data that supports further analysis or policy exploration.

```
fviz_cluster (kmeans_result, data = cluster_data)
```

### 4.4 Model Interpretability and Insights

we Understanding why a model makes certain predictions is just as important as measuring how accurate those predictions are. Among the models used, **Decision Tree** and **Random Forest** stood out for their interpretability. These models not only provide solid predictive performance but also offer insight into which input features contribute most to the final prediction.

Using the Random Forest model, I examined **feature importance scores** based on *IncNodePurity*, which reflects how much each year helped reduce prediction error. The most influential year was **2015 (30.67)**, followed closely by **2019 (29.62)** and **1975 (28.69)**. Other notable contributors included **2005 (21.83)** and **2020 (18.72)**. The least influential was **2010 (11.11)**, indicating that its predictive value for 2021 emissions was relatively low.

Interestingly, both **recent and historical years** played important roles in the model's decision-making. This suggests that while current trends are crucial, long-term emission patterns also carry meaningful signals—likely reflecting deep-rooted national energy practices or industrial growth. These insights can help inform future decisions on **feature selection and dimensionality reduction**, especially in scenarios where a lean, high-impact dataset is preferred.

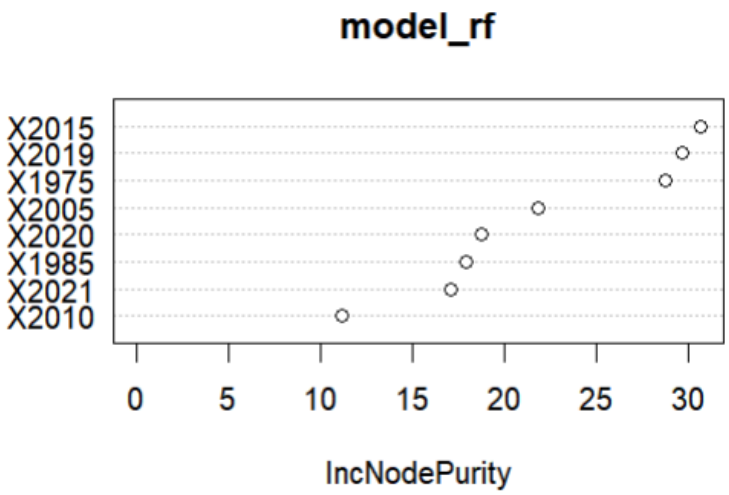


*Figure 5. Random Forest variable importance plot. The model identified 2015 and 2019 as the most influential features, followed by 1975 and 2005. More recent years like 2020 and 2021 also contributed meaningfully to the prediction of $CO_2$ emissions.*

## 5. Discussion and Ethical Considerations

This final chapter summarizes the key findings from the study, reflects on limitations and ethical considerations, and offers suggestions for how the results and methods could be applied or expanded in future work.

### 5.1 Summary of Key Findings

This study explored the use of machine learning models to predict $CO_2$ emissions based on historical data. After cleaning, normalizing, and preparing the dataset, several

regression models were tested: Decision Tree, K-Nearest Neighbours (KNN), Support Vector Regression (SVR), and Random Forest.

Among these, **SVR achieved the highest performance**, with the lowest RMSE and highest $R^2$, indicating strong predictive accuracy. However, **Random Forest** also performed well and provided valuable **feature importance insights**, which made it more interpretable. The most important features identified were from **recent and mid-range years like 2005, 2019, and 2010**, suggesting that recent emission trends play a crucial role in predicting current emissions.

Simpler models like the Decision Tree offered easier interpretability but were less accurate. KNN performed moderately well, reinforcing the idea that regression-based approaches are suitable for modeling carbon emissions, though their performance depends heavily on data scaling and feature relationships.

In addition to supervised models, **K-Means Clustering** was used to explore country groupings based on emission patterns. This analysis revealed that countries with similar emission trajectories tend to cluster together, which could be helpful for policy design and international collaboration.

### 5.2 Limitations and Risks

There are several limitations despite the encouraging findings:

- **Dataset Characteristics**: The dataset has a small number of features (emissions and energy produced per year, for example). Including additional socio-economic, policy or population data might lead to better predictions.
- **The construction of the dataset**: Using a wide-format dataset for the analysis restricts our exploration to time-series data. Panel data approaches could help researchers build richer models in the future.
- **Model Bias**: A model is only as good as the data used to train it. If certain countries or regions lack sufficient representation or accurate reporting, the predictions may be questionable.
- **There is a chance of overfitting**, particularly in uncertain ML, e.g. Random Forest, if there are more features than observations. Cross-validation can mitigate this risk but does not remove it. (Breiman, 2001) (Kohavi,1995)

### 5.3 Ethical Use of Machine Learning in Environmental Contexts

As predictive models become more common in environmental policy and global monitoring, it's important to consider what's right and wrong:

- **Transparency and understandability**: Complicated models should not be used to make public decisions unless their logic can be explained in simple terms.
- **Data Sovereignty**: Emissions data often ties to national interests. People who create models must respect data ownership and be careful not to misuse public environmental data.
- **Environmental Justice**: Predictive models must consider vulnerable populations and countries disproportionately affected by climate change. These models should support fair and inclusive policies that help reduce the effects of climate change.

### 5.4 Practical Applications

This study highlights several ways that machine learning can support climate research and policy decisions:

- **Forecasting $CO_2$ emissions:** Predicting future emission levels using patterns from past data can help policymakers plan interventions earlier.
- **Identifying high-risk regions:** Grouping countries by their emission trends (e.g., through clustering) can help target efforts where they're most needed.
- **Supporting real-time monitoring systems:** Machine learning models can be integrated into systems that continuously track emissions and trigger alerts or policy reviews when thresholds are exceeded.
- **Simplifying complex datasets:** Feature importance analysis helps identify which time periods carry the most predictive power, which can reduce the data needed for future models without sacrificing accuracy.

Used responsibly, these models can improve the speed and effectiveness of environmental analysis and help build smarter, data-driven climate strategies.

### 5.5 Recommendations for Future Work

There are several ways this study could be expanded in the future to improve both accuracy and usefulness:

- **Use of temporal models:** Incorporating models like Long Short-Term Memory (LSTM) networks could better capture sequential patterns across years. (Hochreiter & Schmidhuber, 1997)
- **Expanding the feature set:** Adding external variables such as GDP, urbanization, industrial activity, or participation in international climate agreements could make the models more robust and policy relevant.
- **Geospatial analysis:** Mapping emissions geographically could help identify regional emission hotspots and link patterns to specific national or regional policies.
- **Explainability tools:** Techniques like SHAP (SHapley Additive exPlanations) can help make even complex models more transparent and easier to understand. (Lundberg & Lee, 2017)

Implementing these improvements could turn basic prediction tools into more comprehensive decision-support systems for climate strategy and sustainability planning.

### References


1. **Bergstra, J., & Bengio, Y.** (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(10), 281–305.
2. **Breiman, L.** (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324
3. **Chai, T., & Draxler, R. R.** (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
4. **Cover, T. M., & Hart, P. E.** (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

5. **Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V.** (1997). Support vector regression machines. In *Advances in Neural Information Processing Systems* (pp. 155–161).
6. **Han, J., Kamber, M., & Pei, J.** (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
7. **Hochreiter, S., & Schmidhuber, J.** (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
8. **Jordan, M. I., & Mitchell, T. M.** (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415
9. **Kohavi, R.** (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143).
10. **Lundberg, S. M., & Lee, S.-I.** (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4765–4774).
11. **MacQueen, J.** (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
12. **Nagelkerke, N. J. D.** (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.
13. **Quinlan, J. R.** (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/BF00116251