ANALYSIS AND PREDICTION OF CARBON FOOTPRINTS

Introduction

The objective of this report is to document the entire process of data sourcing, analysis, and model development for predicting carbon footprints for different countries. The case study aims to explore the relationships between various country-specific features and CO2 emissions to create an accurate predictive model.

Data Sourcing

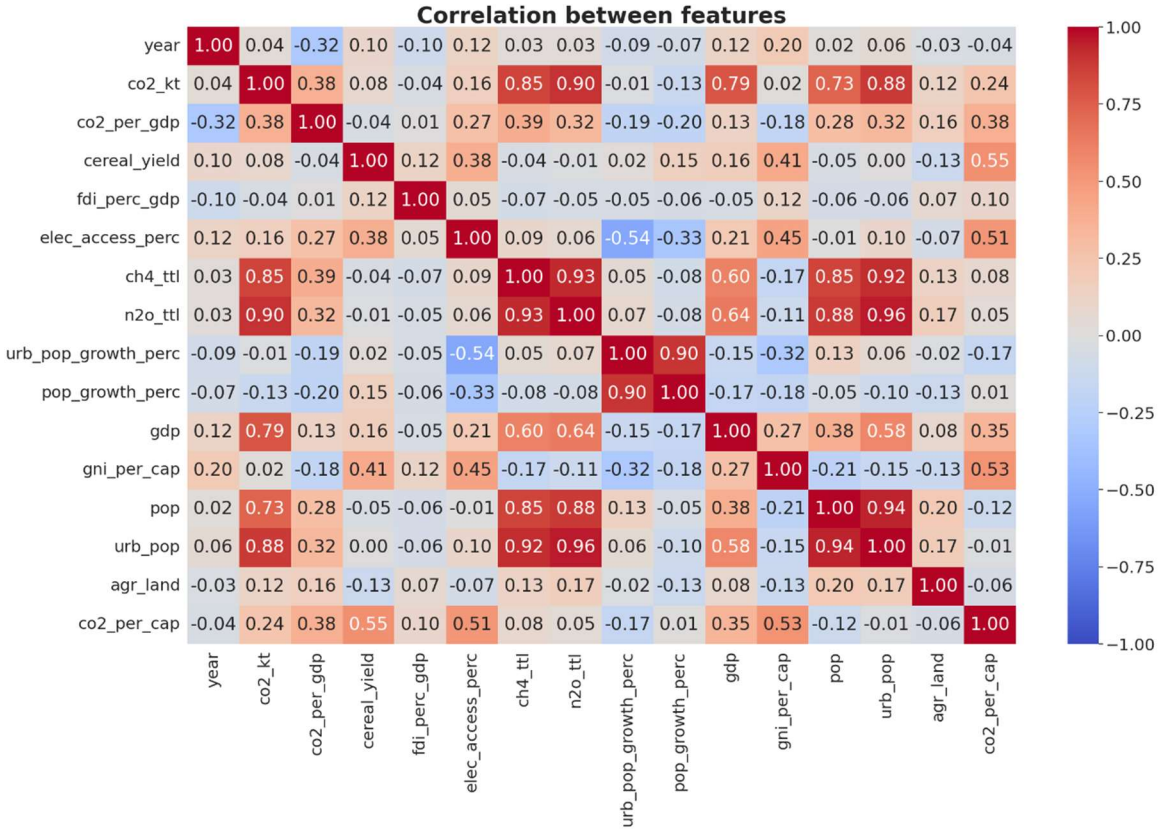
The data used in this analysis was sourced from the publicly available dataset "Climate Change Data" provided by the World Bank Group. The dataset contains comprehensive country-specific information over the range of 2000-2019, encompassing essential variables such as 'CO2 emissions (kt)', 'CO2 emissions (kg per PPP $ of GDP)', 'Cereal yield (kg per hectare)', 'Foreign direct investment, net inflows (% of GDP)', 'Access to electricity (% of population)', 'Energy use (kg of oil equivalent) per $1,000 GDP (constant 2017 PPP)', 'Other greenhouse gas emissions, HFC, PFC and SF6 (thousand metric tons of CO2 equivalent)', 'Methane emissions (kt of CO2 equivalent)', 'Nitrous oxide emissions (thousand metric tons of CO2 equivalent)', 'Urban population growth (annual %)', 'Population in urban agglomerations of more than 1 million', 'Population growth (annual %)', 'Terrestrial protected areas (% of total land area)', 'GDP (current US$)', 'GNI per capita, Atlas method (current US$)', 'Population, total', 'Urban population', 'Agricultural land (% of land area)', 'Fossil fuel energy consumption (% of total)', 'CO2 emissions (metric tons per capita)', 'CO2 emissions from transport (% of total fuel combustion)'.

Data Preprocessing

Before conducting the analysis, the data underwent rigorous preprocessing steps. Missing values were handled using imputation techniques, and redundant columns were removed to ensure data integrity. Additionally, outliers were identified and dealt with to prevent any undue influence on the analysis.

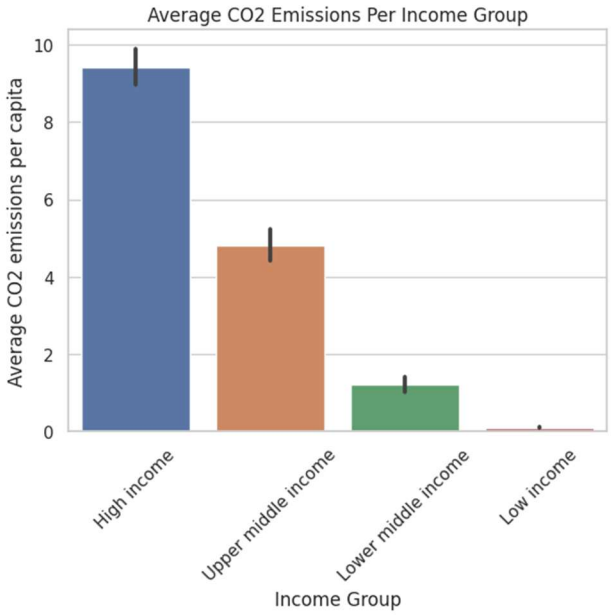
Data Analysis

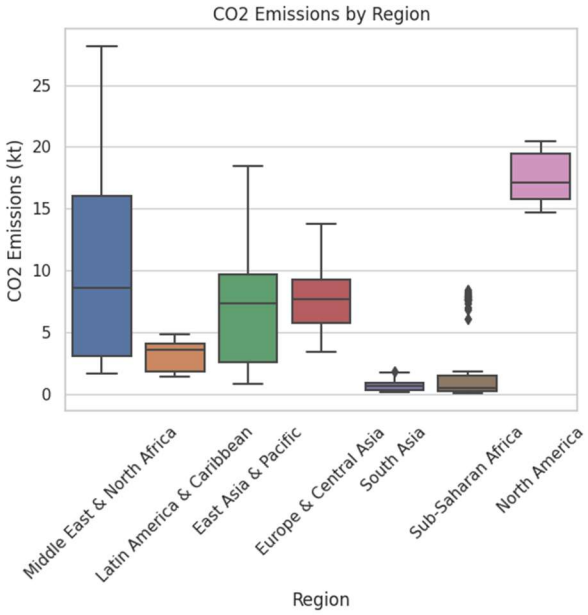
Exploratory Data Analysis (EDA) was carried out to gain insights into the distribution and relationships between different variables. Visualizations such as scatter plots, histograms, and correlation matrices were utilized to identify patterns, trends, and potential dependencies. The analysis highlighted the nonlinear characteristics of most dependencies and the presence of clustered data points in specific countries.

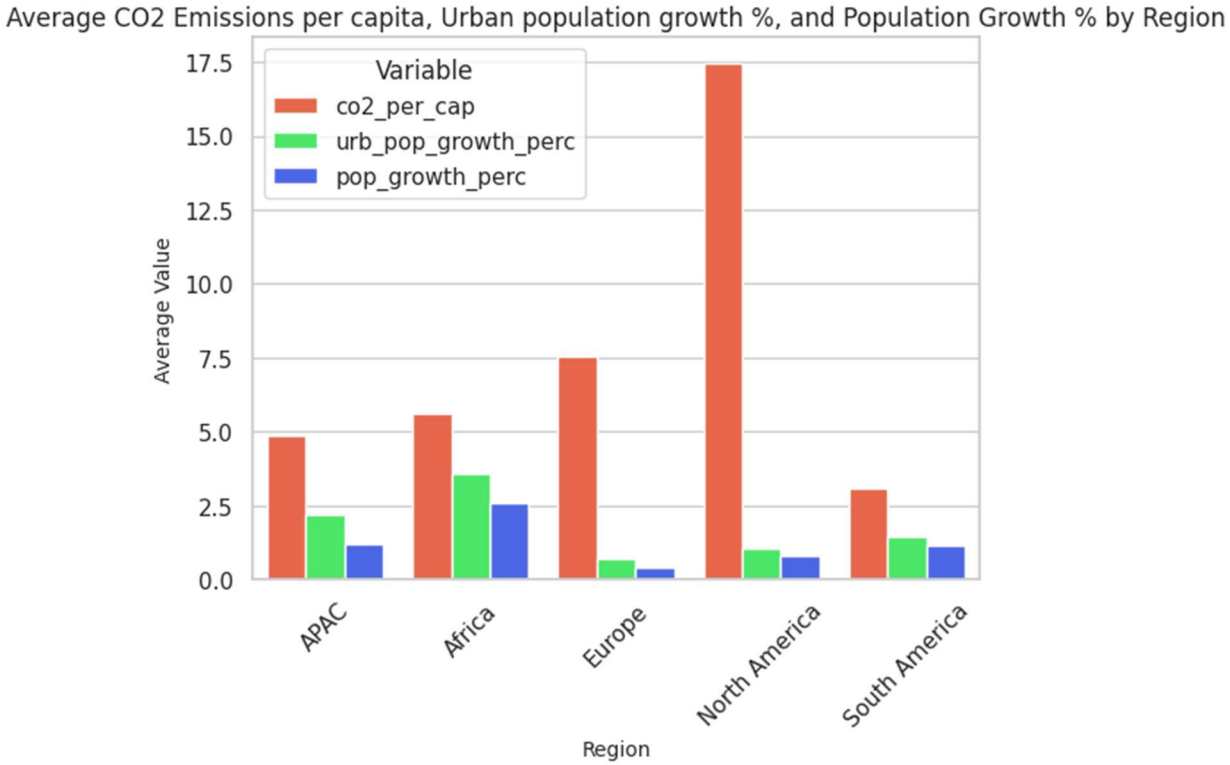


From the correlation matrix, only the variables that are more significant with co2_kt are considered. Most of the variables include 'fdi_perc_gdp', 'elec_access_perc', 'ch4_ttl', 'n2o_ttl', 'urb_pop_growth_perc', 'pop_growth_perc', 'gdp', 'gni_per_cap', 'urb_pop', 'agr_land'.

Also, the pair plots in the notebook suggested that the country ‘CHN’ have the outlier variables thus removed from further analysis. Plots have been made against different income groups as well as plots against different geographical regions.



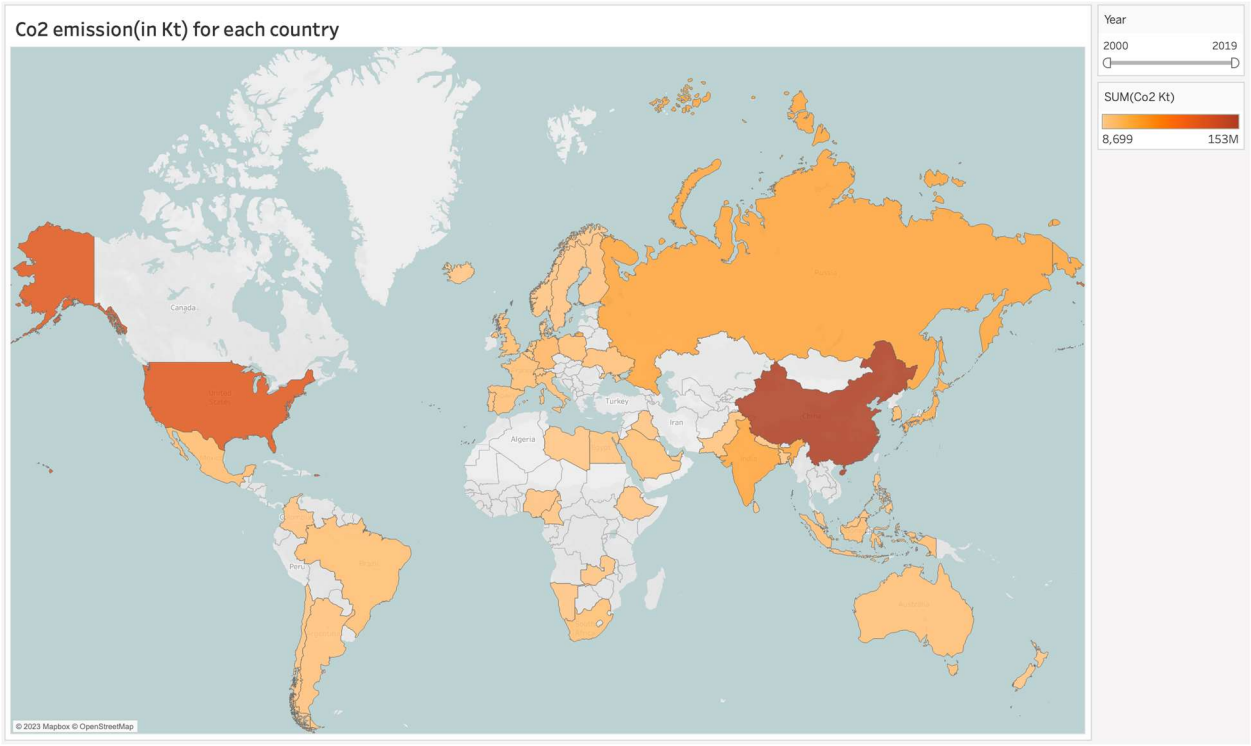




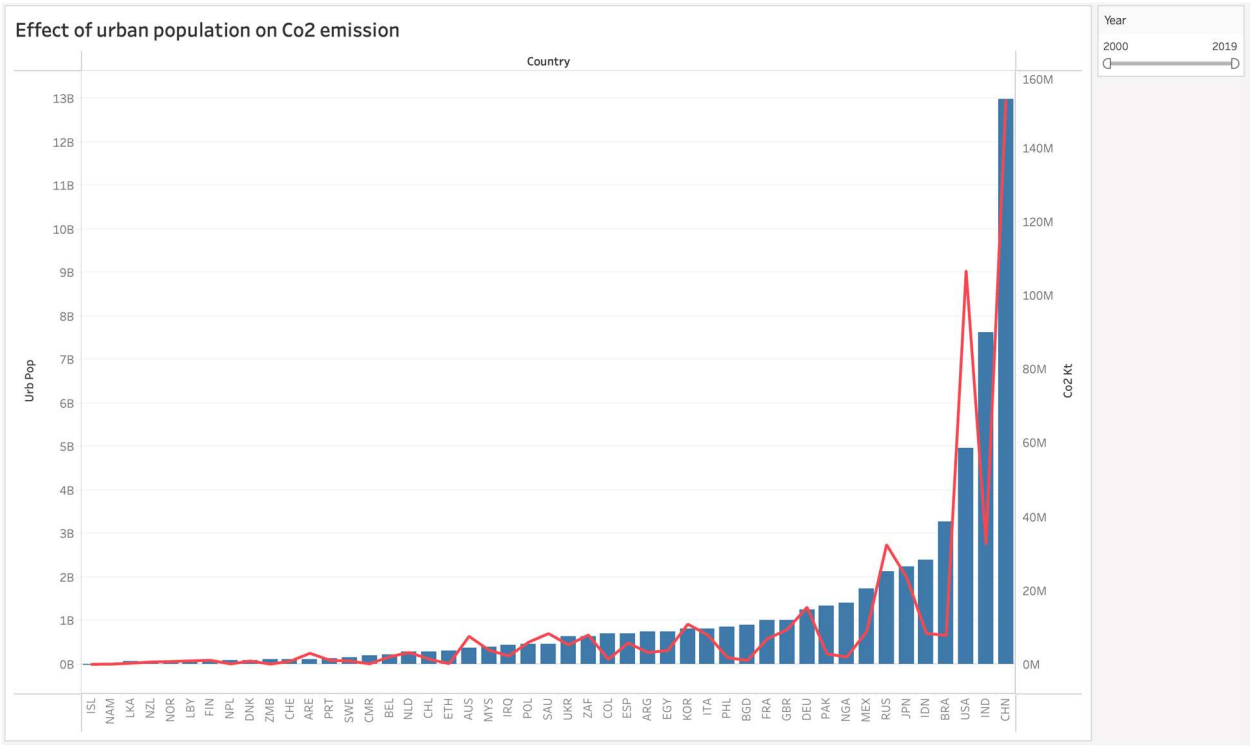
Additional visualizations and insights from Tableau

For getting a better understanding of the trends in the time series data, I have leveraged the knowledge of Tableau software and made some visualizations.

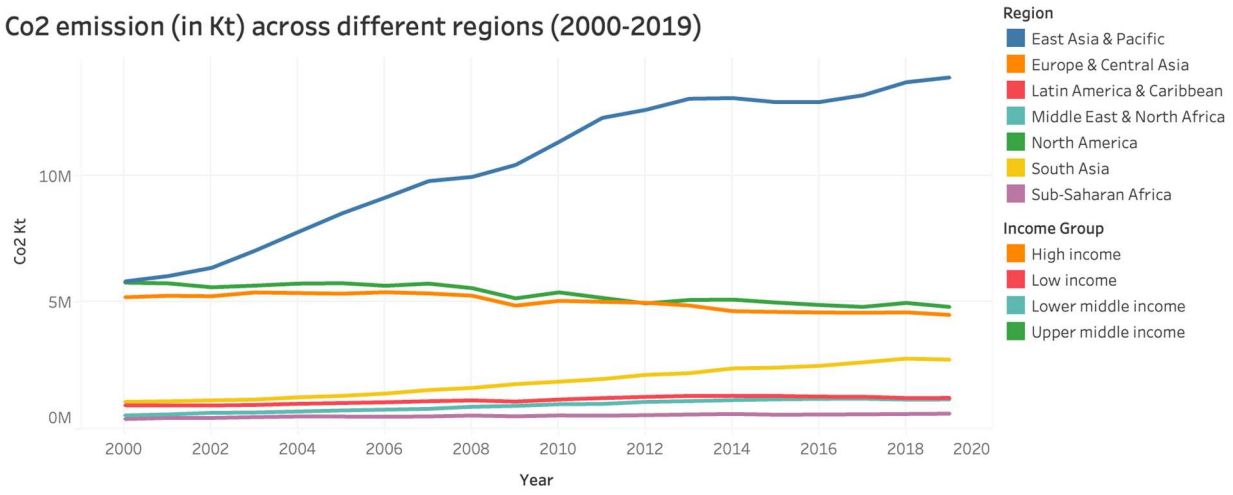
1. Geographical view of CO2 emissions across the different countries. It shows that countries like China, USA has more carbon footprint emissions among all the countries.

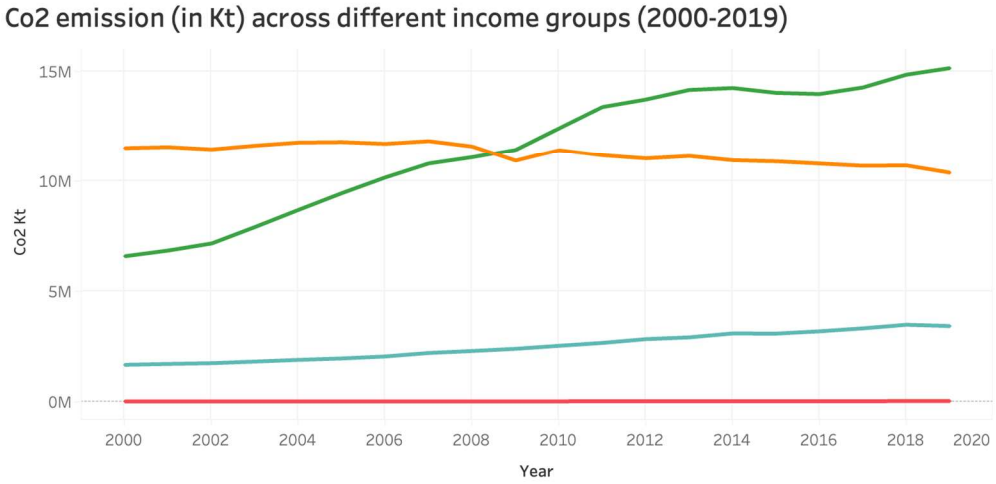


2. The below bar chart shows that urban population has a significant impact on the co2 emissions.



3. The below line chart shows the co2 emissions across different regions and income groups.





Model Development

Given the nonlinear nature of the data and the presence of clustered points, the use of machine learning algorithms capable of handling such complexities was deemed appropriate. A Random Forest Regression model was selected due to its ability to capture nonlinear relationships and effectively handle groups of data points.

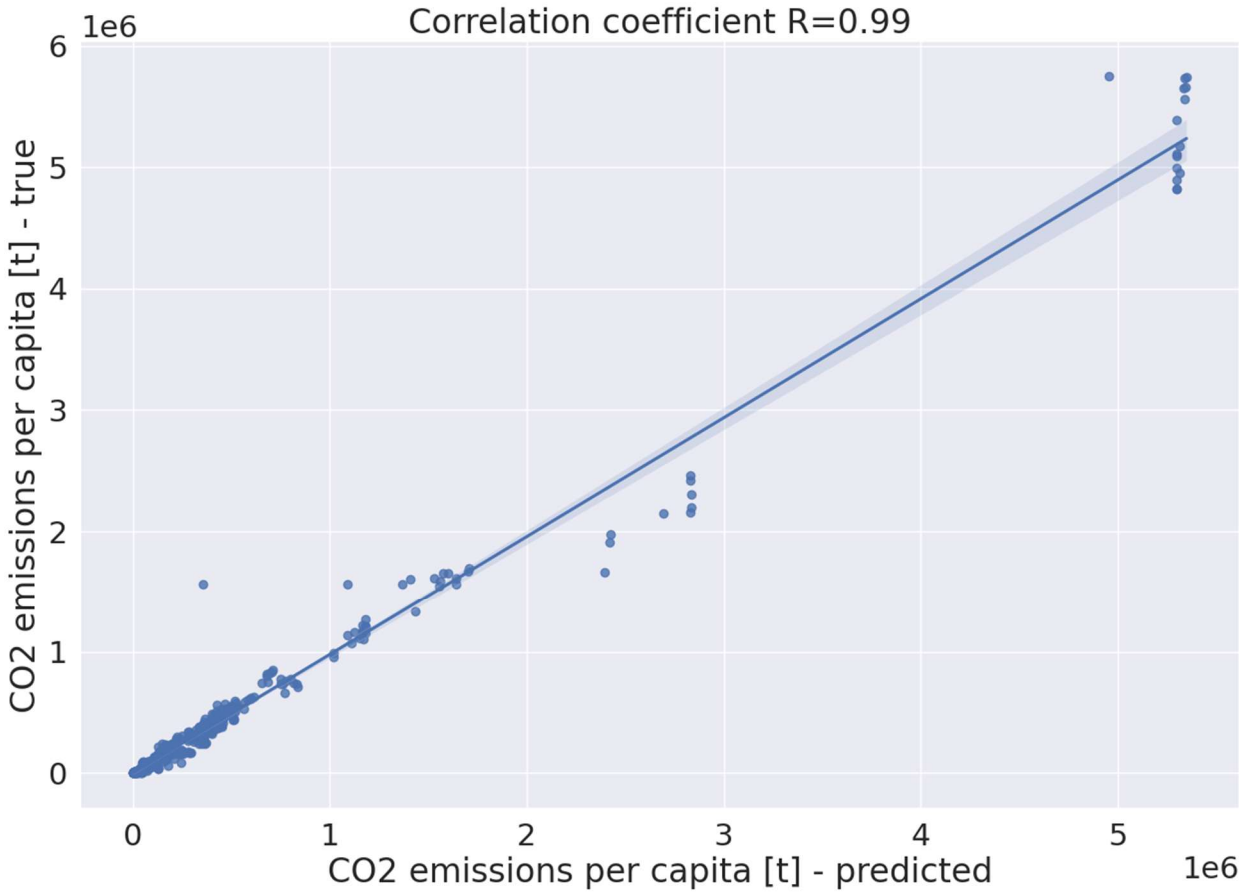
Feature Selection and Hyperparameter Tuning

Feature selection techniques were employed to identify the most relevant country-specific features for predicting CO2 emissions. 'RandomizedSearchCV' from the sklearn.model_selection module was utilized to perform hyperparameter tuning. This technique randomly samples a defined number of hyperparameter combinations from the specified ranges, allowing for efficient exploration of the hyperparameter space.

Model Evaluation

The performance of the final model was evaluated using various metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score. The model

demonstrated a high precision, achieving an R2 score of 98.3%, indicating its excellent ability to explain the variance in CO2 emissions.



Conclusion

In conclusion, this report presents a comprehensive analysis of data sourcing, data analysis, and model development for predicting CO2 emissions. The use of Random Forest Regression, coupled with appropriate feature selection and hyperparameter tuning, resulted in a predictive model with remarkable precision. The insights gained from this study can contribute to a deeper understanding of the factors influencing carbon footprints and assist in making informed decisions towards a sustainable future.