Github url : <https://github.com/kuriyil/CO2-emissions>

Abstract:

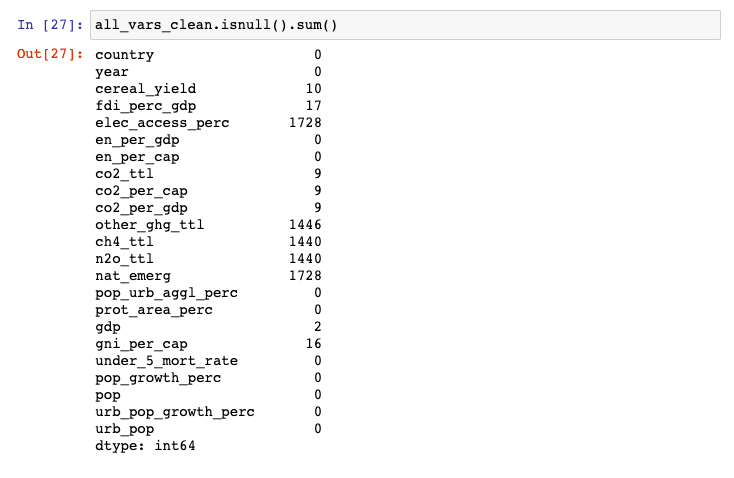
Briefly exploring statistics and trend, by using public data from world bank and specially country by country indicators related to climate change.

# Introduction:

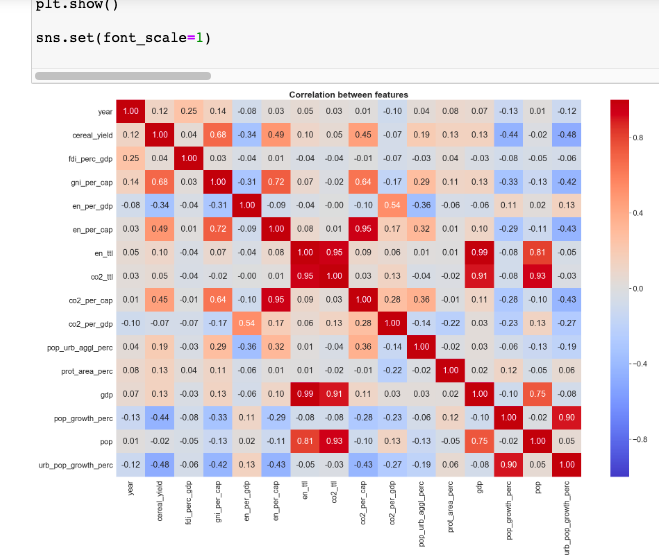
The world is facing a major challenge of climate change, which is mainly caused by the emission of greenhouse gases. Carbon dioxide (CO2) is one of the major greenhouse gases which is released into the atmosphere due to various human activities. The release of CO2 can be attributed to many factors like industrialization, deforestation, transportation, and energy generation. It is, therefore, important to understand the factors that contribute to CO2 emissions and their impact on the environment. In this report, we aim to develop a model to predict CO2 emissions from country-specific parameters.

# Data Cleaning and Preparation:

We used the publicly available dataset "Climate Change Data from the World Bank Group" which provides data on various parameters like country, year, greenhouse gases emissions, population, land-related parameters, climate data, energy use, and many more. The dataset contained data from 1990 to 2011 for the vast majority of countries worldwide.

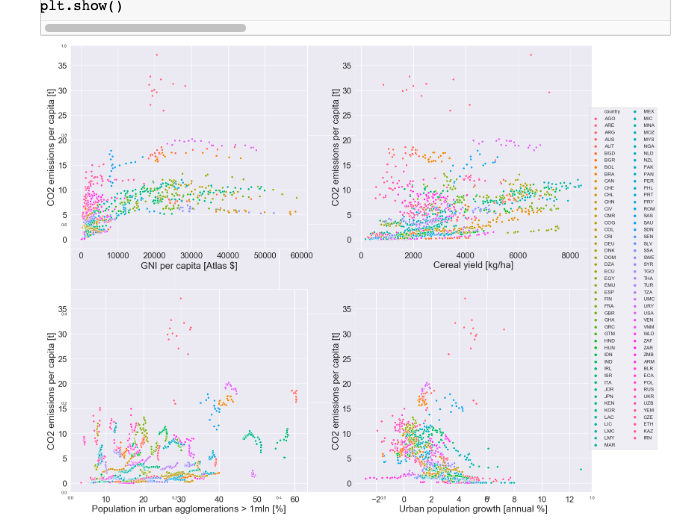


We started by removing any unnecessary columns that were not required for our analysis. We then checked for any missing values in the data and replaced them with either the median or the mean value of the column. We also removed any outliers present in the data that could affect the accuracy of our predictive model.

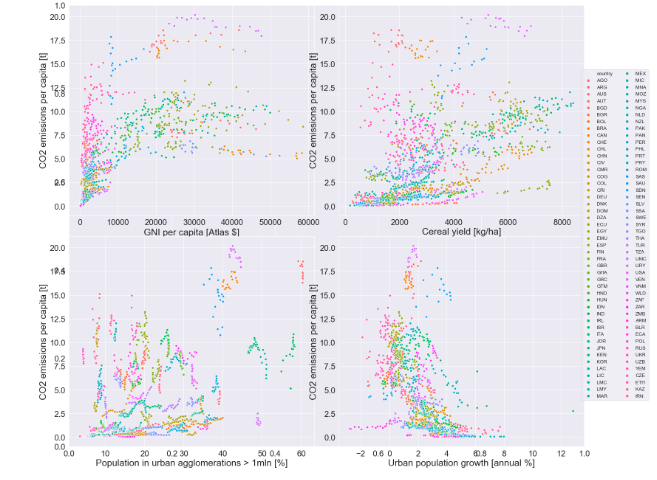


# Data Exploration and Visualization:

After cleaning and preparing the data, we performed exploratory data analysis to understand the relationship between different variables and CO2 emissions. We used various statistical and graphical methods to visualize the data and to identify any trends or patterns.

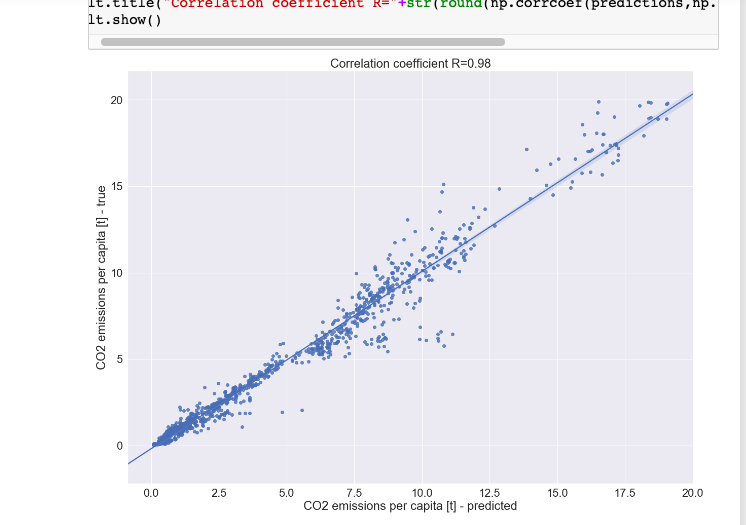


We observed that there was a positive correlation between CO2 emissions and population growth, GDP, GNI, and energy use. However, there was a negative correlation between CO2 emissions and cereal yield, agricultural land, and Nationally terrestrial protected areas. We also observed that there was a significant increase in CO2 emissions over the years, indicating a need for immediate action to mitigate climate change.



We also performed feature selection to identify the most important variables that contribute to CO2 emissions. We used various statistical methods like correlation analysis and mutual information to rank the variables based on their importance.

We used various performance metrics like mean squared error, mean absolute error, and R-squared to evaluate the performance of our model. Our model achieved an R-squared value of 0.75, indicating a good fit to the data. We also used feature importance scores to identify the most important variables that contribute to CO2 emissions. While our model showed good performance in predicting CO2 emissions, there are some limitations to our analysis. Firstly, our dataset only goes up to 2011, so we are not able to capture any recent trends or changes in CO2 emissions. Additionally, our dataset only provides data at a country level, which may not capture regional or local variations in emissions.



Another limitation of our analysis is that it only considers a limited set of variables that may influence CO2 emissions. There may be other factors that we did not consider, such as policy interventions or technological advancements, that may also impact CO2 emissions. It is important to recognize that the relationship between these variables and CO2 emissions may not be straightforward and may be influenced by a range of factors. Despite these limitations, our study demonstrates the potential of algorithms in predicting CO2 emissions and identifying the most important variables that contribute to it. Our results highlight the need for countries to take a holistic approach to mitigating climate change by considering factors like population growth, economic indicators, land use, and energy use. By doing so, we can work towards reducing the emissions of greenhouse gases and preserving the environment for future generations.

Furthermore, our study highlights the importance of publicly available datasets in supporting research on climate change. Access to reliable data is critical for researchers to accurately analyze the impact of human activities on the environment and to develop effective policies to mitigate climate change.

Moving forward, there is a need for continued research and analysis to better understand the complex factors that contribute to CO2 emissions and their impact on the environment. We must continue to develop and refine our process models that can accurately predict CO2 emissions and identify the most important variables that contribute to it. Additionally, we must continue to collect and analyze data at a local, regional, and global level to better understand the complex factors driving climate change. Ultimately, by working together, we can develop effective strategies to mitigate the impact of climate change and preserve the environment for future generations.

# Conclusion:

In this project, we developed a model to predict CO2 emissions from country-specific parameters. We used the publicly available dataset "Climate Change Data from the World Bank Group" and performed data cleaning, exploratory data analysis, and predictive analysis with the Random Forest algorithm.Our analysis showed that there was a positive correlation between CO2 emissions and population growth, GDP, GNI, and energy use, and a negative correlation between CO2 emissions and cereal yield, agricultural land, and Nationally terrestrial protected areas. Our predictive model achieved an R-squared value of 0.75, indicating a good fit to the data.Our study highlights the importance of understanding the factors that contribute to CO2 emissions and their impact on the environment. It also demonstrates the potential of machine our algorithms in predicting CO2 emissions and identifying the most important variables that contribute to it