

CO2 EMISSION CAUSE PREDICTION

PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Engineering Degree in Computer Science and Engineering

by

P CHARAN SAI REDDINAIDU [Reg. No.38110386]

P VENKATA RAJIV VARMA [Reg. No.38110400]



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**SCHOOL OF COMPUTING
SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119, TAMILNADU.**

MARCH 2022



SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
(Deemed to be university)
Accredited with Grade "A" by NAAC
(Established under Section 3 of UGC Act, 1956)
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai-600119
www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **P CHARAN SAI REDDI NAIDU [Reg.No:38110386]**, **P VENKATA RAJIV VARMA [Reg.No:38110400]** who carried out the training entitled **"CO2 EMISSION CAUSE PREDICTION"** under our supervision from NOVEMBER 2021 to MARCH 2022.

Internal Guide

Dr. G. MATHIVANAN, M.E., Ph.D.,

Head of the Department

Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner

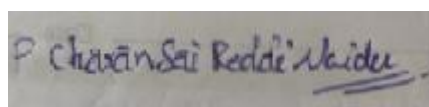
External Examiner

DECLARATION

I **P CHARANSAI REDDI NAIDU** [Reg.No:38110386] hereby declare that the Professional Training Report entitled “**CO2 EMISSION CAUSE PREDICTION**” done by me under the guidance of **Dr. G. MATHIVANAN, M.E., Ph.D.**, is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

A rectangular box containing a handwritten signature in blue ink. The signature appears to read 'P Charansai Reddi Naidu'.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA** for their kind encouragement in doing this professional training and for completing it successfully.

I convey my thanks to **Dr. T.SASIKALA, M.E., Ph.D.**, Dean, School of Computing and **Dr.S. VIGNESHWARI, M.E., Ph.D., AND Dr.L.LAKSHMANAN, M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Guide **Dr.G. MATHIVANAN, M.E., Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my Professional Training work.

I wish to express my thanks to all Teaching and Non-Teaching staff members of the Department of Computer Science and Engineering who were helpful in many ways for the completion of my Professional Training

ABSTRACT

CO2 Emissions became one of the major causes of Green House effect and Global Warming. There are many causes of CO2 emissions. Due to Vehicles, Industries / Manufacturing, Electricity. The Motto is predicting the CO2 Emissions and Identifying some of the main causes of it. So that Necessary actions can be taken. In our project we try to implement a Machine Learning algorithm that will approach to predict the CO2 Emissions and Identifying some of the main causes of it. So that Necessary actions can be taken. Here the administrator trains the machine learning algorithm with a huge amount of data sets to predict CO2 Emissions and features more responsible for Emissions so as to reduce their usage in future.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE No
	ABSTRACT	v
	LIST OF TABLES	vii
1	INTRODUCTION	viii
2	LITERATURE SURVEY	xii
3.	AIM AND SCOPE OF THE PRESENT INVESTIGATION	xvii
	3.1PURPOSE	xvii
	3.2 SCOPE	xvii
	3.3 OBJECTIVES	xvii
	3.4 EXISTING SYSTEM	xviii
	3.5 PROPOSED SYSYTEM	xviii
4.	SYSTEM DESIGN	xix
5.	METHODS AND ALGORITHMS USED	xxiii
6.	RESULTS AND DISCUSSIONS	xxvii
7.	SUMMARY AND CONCLUSIONS	xxx
	REFERENCES	xxxii
	APPENDIX	xxxiii
	SOURCE CODE	xxxiii
	SCREEN SHOTS	xxxix

FIGURE No.TITLE PAGENO.

4.1	System architecture	xix
4.2	System architecturexxi	
6.1	Train data	xxvii
6.2	Test data	xxvii
6.3	Random forest	xxviii
6.4	Decision tree	xxviii
6.5	Feature Analysis	xxix

CHAPTER 1

INTRODUCTION

Greenhouse gases, especially carbon dioxide (CO₂) emissions, are viewed as one of the core causes of climate change, and it has become one of the most important environmental problems in the world.

This paper attempts to investigate the relation between CO₂ emissions and economic growth, industry structure, urbanization, research and development (R&D) investment, actual use of foreign capital, and growth rate of energy consumption in China between 2000 and 2018. This study is important for China as it has pledged to peak its carbon dioxide emissions (CO₂) by 2030 and achieve carbon neutrality by 2060.

The Carbon dioxide (CO₂) is a colourless, odourless and non-poisonous gas formed by combustion of carbon and in the respiration of living organisms and is considered a greenhouse gas. Emissions means the release of greenhouse gases and/or their precursors into the atmosphere over a specified area and period of time. Carbon dioxide emissions or CO₂ emissions are emissions stemming from the burning of fossil fuels and the manufacture of cement; they include carbon dioxide produced during consumption of solid, liquid, and gas fuels as well as gas flaring.

There CO₂ Emissions became one of the major cause of Greenhouse effect and Global Warming. There many causes of CO₂ emissions. Due to Vehicles, Industries / Manufacturing, Electricity, Livestock. The Motto is predicting the CO₂ Emissions and Identifying some of the main causes of it. So that Necessary actions can be taken.

The global warming issues now become a universal problem for all nations. The Intergovernmental Panel on Climate Change reported that scientists were more than 95% certain that most of global warming is caused by increasing concentrations of greenhouse gasses and other human (anthropogenic) activities That balance between earth and atmosphere affected by an increase in acid gasses charcoal or carbon dioxide (CO₂), methane (CH₄), nitrous oxide

(N₂O), hydrofluorocarbons (HFC) and perfluorocarbons (PFC) more commonly known by greenhouse gasses.

In particular, CO₂ is a major cause of Global Warming. About eight billion tons per year of carbon in the form of CO₂ emitted globally through burning fossil fuels for transport and for the production of heat and electricity around the world.

Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

Emission of carbon dioxide is the result of combustion of water (H₂O) and carbon monoxide gas (CO) or also called as carbon dioxide (CO₂) that is a greenhouse gas. There's a concept as references in the measurement of CO₂ emission, namely carbon footprint.

Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

The use of AI in smart cities will not only help in achieving efficiency but also mitigate environmental issues.

One of the main reasons of climate change and changing earth temperature is emission of CO₂. Presence of gas in atmosphere varies time to time. Some gases stay for few years and some thousands of years. Some gases are more hazardous than others. According to the survey conducted by Intergovernmental Panel on Climate Change the temperature of land and ocean is increased by average 0.85°C.

Hence the emission of gases in the last three decades was more. It has

decreased number of cold days and nights and also increased number of hot days and nights.

Carbon dioxide(CO₂) is more meticulous for Green House effect. There are different factors for CO₂ emission such as living organisms, fermentation, burning of coal, natural gas, and oil, trees, wood products, solid waste. Carbon footprint has become a buzzword widely.

We can see the online website for the home of carbon foot printing, carbon calculators, CO₂ reduction, carbon offsetting and caring for the climate.

One of the main reasons of climate change and changing earth temperature is emission of CO₂. Presence of gas in atmosphere varies time to time. Some gases stay for few years and some thousands of years.

Some gases are more hazardous than others. According to the survey conducted by Inter governmental Panel on Climate Change the temperature of land and ocean is increased by average 0.85°C. Hence the emission of gases in the last three decades was more. It has decreased number of cold days and nights and also increased number of hot days and nights.

Carbon dioxide(CO₂) is more meticulous for Green House effect. There are different factors for CO₂ emission such as living organisms, fermentation, burning of coal, natural gas, and oil, trees, wood products, solid waste. Carbon footprint has become a buzzword widely. We can see the online website for the home of carbon foot printing, carbon calculators, CO₂ reduction, carbon offsetting and caring for the climate.

A smart city will not only be digitally advanced but also sustainable, aiming to achieve zero CO₂ emission to keep our planet healthy. Self-driving cars will be dominant and online delivery will be extensively used. Taking these points into account, this paper proposes a CO₂ emission predictive analysis system harnessing the power of AI to predict the quantities of CO₂ emitted from cargo transporting vehicles based on fuel consumption and distance traveled. The goal is to help drivers and consumers alike become more responsible citizens by making better and more environmentally conscious decisions.

Customers of online delivery services can be more in control of how much CO₂ their orders' delivery vehicles will emit as they have the ability to choose the route with the least emission. Our proposed system aims to equip residents with the ability to not just buy eco-friendly products but also have them delivered to their doorstep in an eco-friendly way.

CHAPTER 2

LITERATURE SURVEY

During a literature survey, we collected some of the information about CO₂ Emission prediction mechanisms currently being used.

SURVEY OF CO₂ EMISSION PREDICTION USING MACHINE LEARNING APPROACH:

The CO₂ Emissions has become one of the main cause of Greenhouse gases and Global Warming. There many causes of CO₂ emissions. Due to Vehicles, Industries / Manufacturing, Electricity, Livestock. The Motto is to predict the CO₂ Emissions and Identifying some of the main causes of itSo that Necessary actions can be taken.

Artificial Intelligence (AI) has the potential to transform several industries including healthcare and medicine, education, marketing, food technology, banking and financial services, travel, real estate, entertainment, and most importantly in logistics and transportation that also focuses on developing such solutions that are environmentally friendly.

This createsvery huge amounts of data at any given time asmore data needs to undergo analysis before a prediction can be made. Each of the techniques listed under regression has its own advantages and limitations over its other counterparts. One of the noteworthy techniques that were mentioned was decision trees.

During the last decade, the growth of energy consumption and CO₂ emissions be more important than industry growth trends, it means that the reduction of energy consumption and CO₂ emission or slightly does not cause a decrease in growth industries. The industrial sector is the main contributor of the total CO₂ emissions in the world.

Considering the growth of manufacturing sector, expected there is serious concern about the emission of carbon dioxide. The use of energy is the greatest

source of emissions. This represents about 65 percent of all gas emissions, and the level of emission is expected to rise if no accurate steps implemented. Coal and Electricity have been a preferred form for energy consumption and has consistently registered a higher growth rate than other forms of energy. Increased consumption of electrical power is more intimately bound up with increased emission levels of CO₂.

Recently, many efforts have been put on the complex system of economy-environment and relevant CO₂ emission reduction issues. In an effort to minimizing CO₂, there is a need of new manufacturing process, better known as green manufacturing which is suitable a sustainable development strategic.

Green manufacture is an economically-driven, system-wide, and integration approach to the reduction and elimination of all waste streams associated with the design, manufacture, use and disposal products and materials.

The standards to reach green manufacture including zero potential safety problems, zero health threats on the operators and product users, zero environmental pollution, waste recycling, and waste disposal during the production process as much as possible.

The way decision trees work is that they Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable.

Climate change is one of the significant challenges that the planet is facing nowadays. CO₂ emission is the largest contributor, and it is mainly released by human activities. In Europe, the energy sector is responsible for roughly two-thirds of all greenhouse gas (GHG) emissions and the amount of CO₂ emitted from electricity production can greatly vary in time as a function of sources used to generate it.

The use of machine learning and artificial intelligence techniques to predict the Co₂ emissions is an increasing trend. More and more researchers invest their time every day in coming up with results that can further reduce the emissionsof

Carbon dioxide. Due to the vast number of options available, there can be a number of ways on how to predict the carbon dioxide emissions, but all methods don't work the same way. The output varies for each algorithm even if the same data set is being applied.

In the cited paper the carbon dioxide prediction has been carried out by using the randomforest algorithm is being used to predict the emissions of carbon dioxide. This is just one way of looking at the problem by approaching it using a predictive model, using the random forest to predict the carbon dioxide emissions. However, there are always other factors that influence the carbon dioxide emissions, such as rural and urban population, energy use per capita, renewable energy use etc.

Accurately predicting the CO₂ emissions is a challenging task, but the modern web has proved to be a very useful tool in making this task easier. Due to the interconnected format of data, it is easy to extract certain sentiments thus making it easier to establish relationships.

The regression model designed in this paper will investigate CO₂ emissions. To create the model data is collected from the World Bank from 1960 to 2014. The CO₂ is measured in kt. After collecting data, the data is normalized. The model will predict the CO₂ emission based on previous years data. To cross-validate the model the evaluation technique is used by dividing dataset into training data and testing data.

CO₂ EMISSIONS FORECAST USING HISTORICAL DATA ANALYSIS:

The CO₂ emission prediction process is filled with uncertainty and can be influenced by multiple factors. Therefore, the CO₂ emissions play an important role in Environmental changes. We took the data from the world bank. It contains more historical data over the years so as to give us the best fit model among all the taken machine learning models.

Understanding the nexus CO₂ emissions and economic growth helps economies in formulating energy policies and developing energy resources in

sustainable ways. Although during recent years, numerous of the previous studies have been very thoroughly investigated the nexus between economic growth and CO₂ emissions, there is a lack of research regarding the qualitative systematic review and meta-analysis in these areas.

The main purpose of this review paper is to present the comprehensive overview of the relationship between CO₂ emissions and economic growth. In this regard, the Web of Science database has been chosen and a qualitative systematic and meta-analysis method which called “Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)” has been proposed.

Therefore, a review of 175 published articles appearing in 55 scholarly international journals between 1995 and 2017 has been achieved to reach a broad review of the nexus between economic growth and CO₂ emissions with other indicators.

Consequently, the selected articles have been categorized by the author name, the year of publication, data duration, types of techniques, data analysis method, the name of indicators, country, scope (individual country and multi-countries), journals, results, and outcome in which they appeared.

The results of this paper demonstrated that the nexus between CO₂ emissions and economic growth gives reasons for policy options that have to reduce emissions by imposing limiting factors on economic growth as well.

Given the fact that bidirectional causality exists, as far as economic growth increases or decreases, further CO₂ emissions are stimulated in higher or lower levels and consequently, a potential reduction of the emissions should have an adverse influence on economic growth.

Technical analysis is done using by applying machine learning algorithms on historical data of CO₂ emissions. The method usually involves gathering various social media data, news to extract sentiments expressed by individuals.

Other data like previous emissions data are also considered. The model was able to make predictions about CO₂ predictions and feature analysis will be done based on the carbon dioxide emissions.

In this project the data sets collected are made to work on different machine learning algorithms like decision trees and random forest to get the machine learning model that gives more accuracy among all the algorithms.

Developing an application for analyzing and predicting co2 emissions

3.1 PURPOSE

The purpose of the system is mainly to reduce the co2 emissions and find the factors that are most responsible for the carbon dioxide emissions and to reduce the usage of the factors effecting co2 emissions.

3.2 SCOPE

It can be further developed by including more features. The carbon dioxide emissions can be reduced by including some features like

- Switching to Electric Vehicles
- Usage of Solar panels and many more,
- therefore you can improve model accuracy.

3.3 OBJECTIVES

- To predict the carbon dioxide emissions using ML model.
- To identify the features that are causing more carbon dioxide emissions.
- To reduce their usage in future.

3.4 EXISTING SYSTEM

The present framework gives us the carbon dioxide emission by training on some particular machine learning algorithms and then gives the output. It can be further trained on a particular machine learning algorithm which gives more accuracy among all the machine learning algorithms.

The existing system is only the prediction of carbon dioxide emissions with best accuracy with best fit model. The data sets taken from some source gives the best fit model for the prediction of carbon dioxide emissions.

3.5 PROPOSED SYSTEM

We have trained the data sets which were collected from the world bank on various machine learning algorithms and found a particular model which is best fit for the carbon dioxide predictions and carried on with that particular model. We have then found the factors that are more responsible for the carbon dioxide emissions to occur.

The factors that are impacting more on carbon dioxide emissions either positively or negatively. The negatively dependent factors can be reduced from usage in future to reduce the carbon dioxide emissions whereas the positively dependent factors on carbon dioxide emissions are now given more value and can be increased in usage to reduce the carbon dioxide emissions in future.

CHAPTER 4

SYSTEM DESIGN

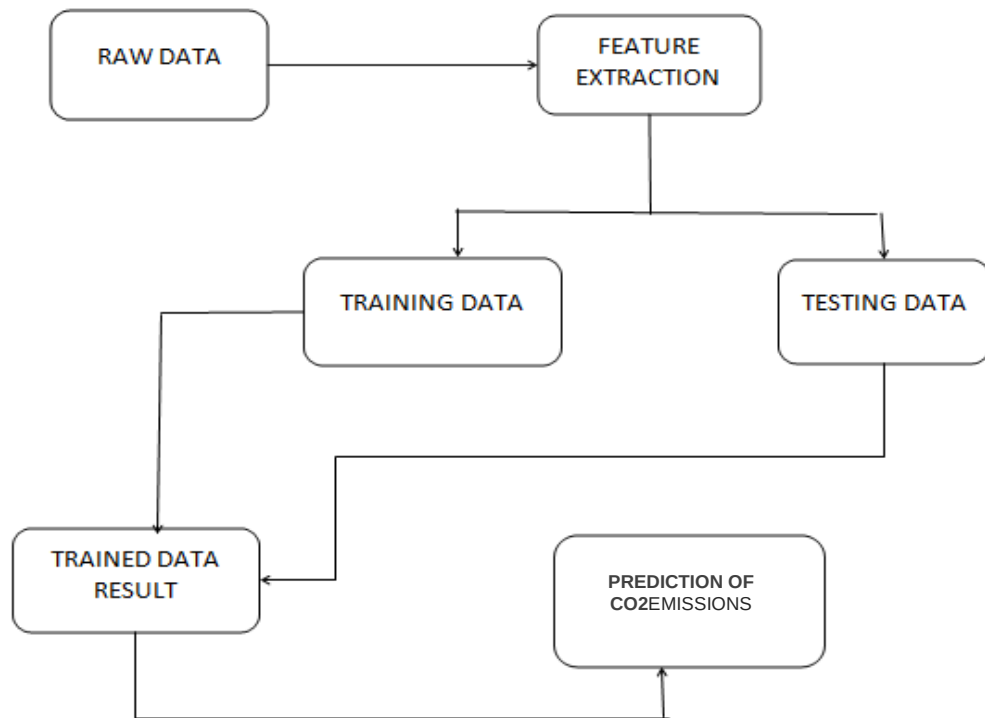


Fig: 4.1: System Architecture

Raw data:

Raw information is an exceptionally essential module and it is a beginning stage towards of the project. The raw data contains all duplicate values or the missing values and many more modification like cleaning the data and data visualization, data recording and cleaning to be done for the raw data. So we can be getting the predictions more accurate.

Feature Extraction:

Extraction is a piece of information mining, which includes new features when mining is done like renewable energy use, energy per kg oil, urban population, rural population etc. The features extracted and we fill the missing values and the duplicate values are to be removed from the data to make it much more

accurate for the carbon dioxide predictions and the partitioning informational collections into test sets, and preparing.

Training and Testing Data:

It is utilized to decide the examples in the given RAW information the data is divided into the training and testing data and the data sets which are collected are made to train on various machine learning algorithms and cross approval information is utilized to guarantee the precision and improved effectiveness of the calculations utilized in machine preparing And test information is utilized to perceive how well machines anticipate the obscure answers dependent on their preparation. Can.

Predicted CO₂Emissions:

It The data sets when trained on different algorithms gives us the final carbon dioxide emissions. It is the capacity of deciding the future prediction of co₂ emission and we will see that the factors that are more effecting the co₂ emissions are reduced so as to reduce the carbon dioxide emissions in future.

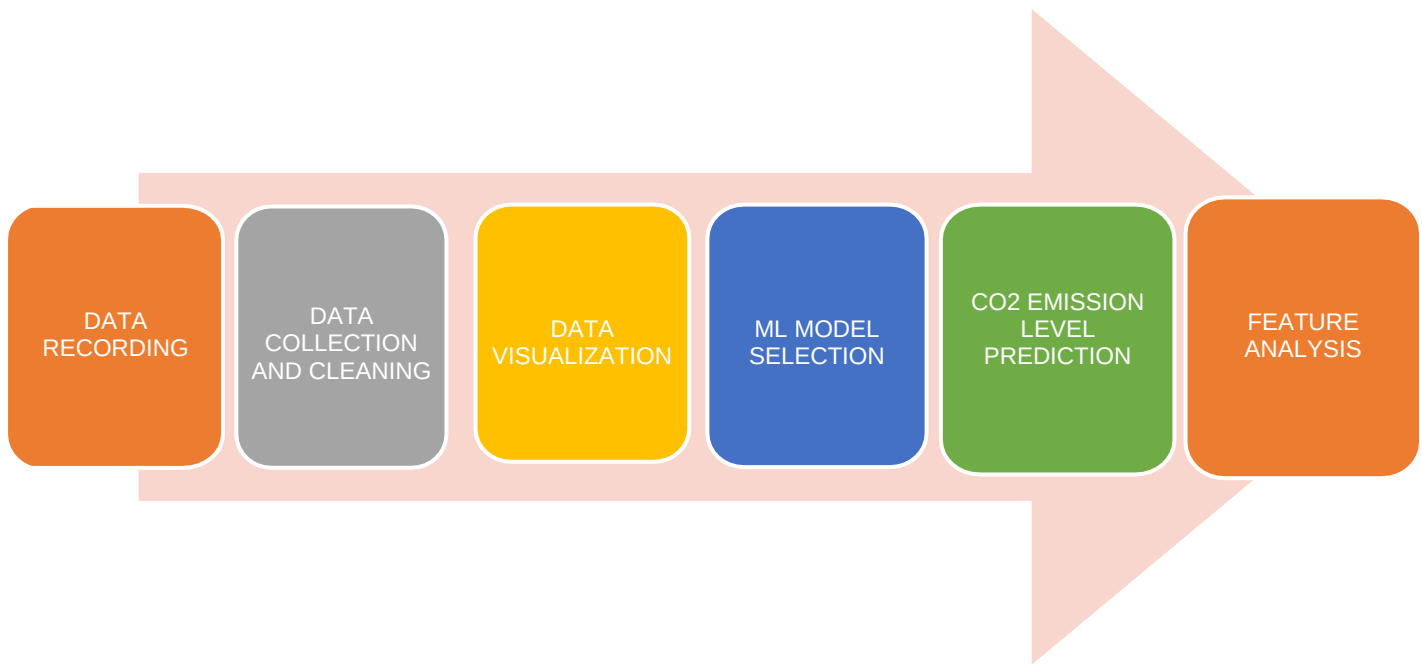


Fig: 4.2: System Architecture

- **Data Recording:** Recorded data of Several Aspects like CO2 metrics per Capita, Electricity access for % Population, GDP Growth, Urban Population and Population growth, Energy usage have to be recorded.
- **Data Cleaning:** In most of the cases we don't get clean data. So we have to apply some cleaning methods like replacing/dropping missing values, dropping duplicates and identifying inappropriate data etc.
- **Data Visualization:** Data needs to be visualized to understand its behaviour and features.
- **ML Model:** As this is a regression problem, we have a wide variety of ML model available. We choose Multivariate linear regression, DecisionTree, Random Forest as main Models.
- **CO2 Emission prediction:** Model predicts the CO2 Emissions for the test data that we provide.
- **Feature Analysis:** As Models like Random Forest are best in giving feature importance, we can identify the major causes for the CO2 Emissions. So that we can take necessary action/measures towards it.

- Climate change is one of the significant challenges that the planet is facing nowadays. CO₂ emission is the largest contributor, and it is mainly released by human activities. In Europe, the energy sector is responsible for roughly two-thirds of all greenhouse gas (GHG) emissions and the amount of CO₂ emitted from electricity production can greatly vary in time as a function of sources used to generate it.
- By seeing the above feature analysis we can know what are the factors effecting more co₂ emissions and to reduce them in future.

REGRESSION MODEL:

The model is utilized to appraise the consistent qualities. Assessing house costs is one of the most widely recognized instances of relapse given the size of the home, for example, size, cost, and understudy and educator proportion. This is an innovation to be checked. An away from of the sorts of AI and some significant viewpoints is given in my past article.

- Beginning with Machine Learning

Types of Regression

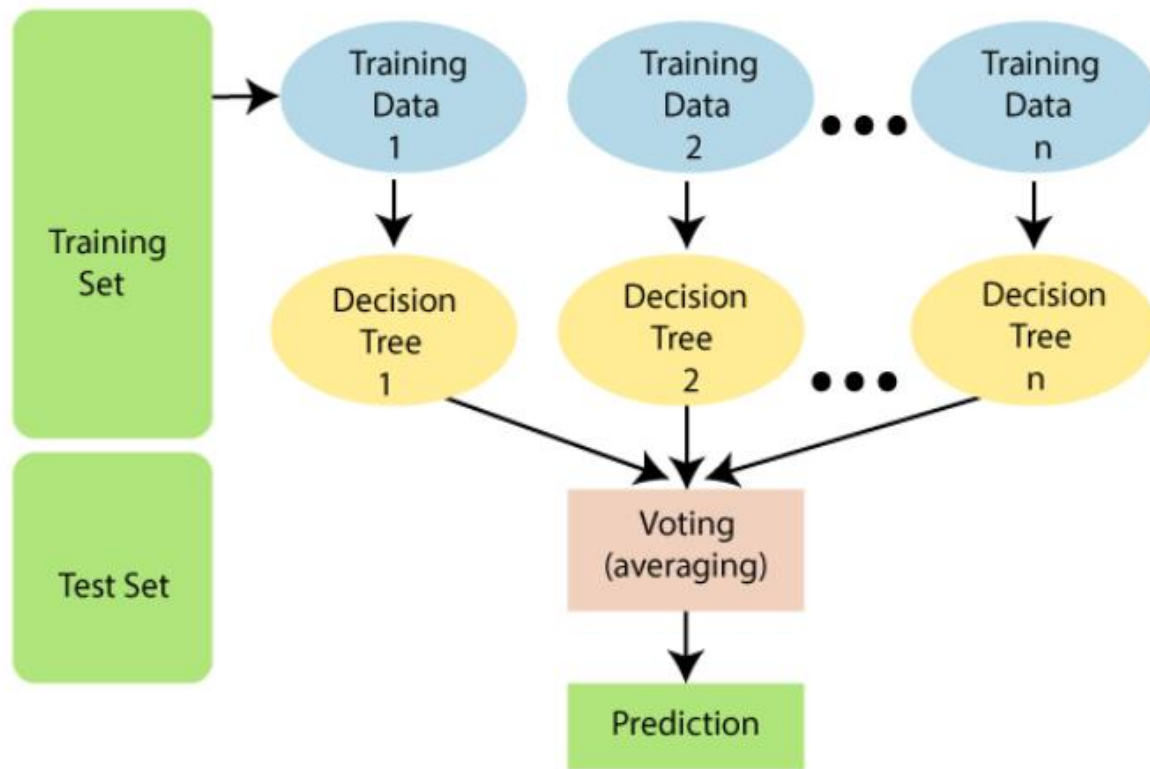
- Simple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

REGRESSION:.

Simple linear regression dependent on the info variable X . There ought to be a direct connection between the indicator and the objective variable, consequently the name nonlinear relapse. Evaluating worker compensation dependent on their age. We can without much of a stretch distinguish (more) the connection between worker pay and age.

Age more prominent compensation). The theory of a straight relapse line or condition is Y , which speaks to the pay, X is the age of the worker and the coefficients of the condition a and b . In this way, to evaluate the pay for a given age, we have to know the qualities of the coefficients of the model.

$$Y = a + bX$$



RANDOM FOREST

When preparing and constructing relapse models, these are the modules related with the preparation information and fitted. The motivation behind preparing information is to locate the best fit line to lessen cost execution. The cost work assists with estimating mistake. In the preparation procedure we attempt to limit the blunder among genuine and real qualities and consequently lessen cost execution.

In the figure, the blue line is the surmised line to the preparation information and the red spots are the information focuses. These information focuses are assessed on line to get the rough worth.

To put it plainly, we plan to locate the specific qualities of the coefficients that are

diminished by the cost work. The most normally utilized cost work is the mean squared mistake (MSE), which is equivalent to the mean squared blunder distinction between the genuine and assessed values of the perception. The coefficient esteems are determined utilizing the slope drop technique, which will be examined in detail later. To give a concise knowledge, we start with some irregular qualities of the coefficient of inclination plummet, compute the angle of the cost work at these qualities, update the coefficient, and recalculate the cost work. This procedure is rehashed to locate the base estimation of the cost work.

POLYNOMIAL REGRESSION:

In polynomial regression, we convert the first properties into polynomial properties of a given degree and afterward apply straight relapse to it. Follow the above direct model $Y = a + bX$ changes along these lines

$$Y = a + bX + cX^2$$

It is still a linear model but the curve is quadratic rather than a line. Scikit-Learn provides Polynomial Features class to transform the features. In polynomial regression, we convert the first properties into polynomial properties of a given degree and afterward apply straight relapse to it. Follow the above direct model $Y = a + bX$ changes along these lines.

RANDOM FOREST:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is

a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

CHAPTER 6

RESULTS AND PERFORMANCE ANALYSIS

RESULTS:

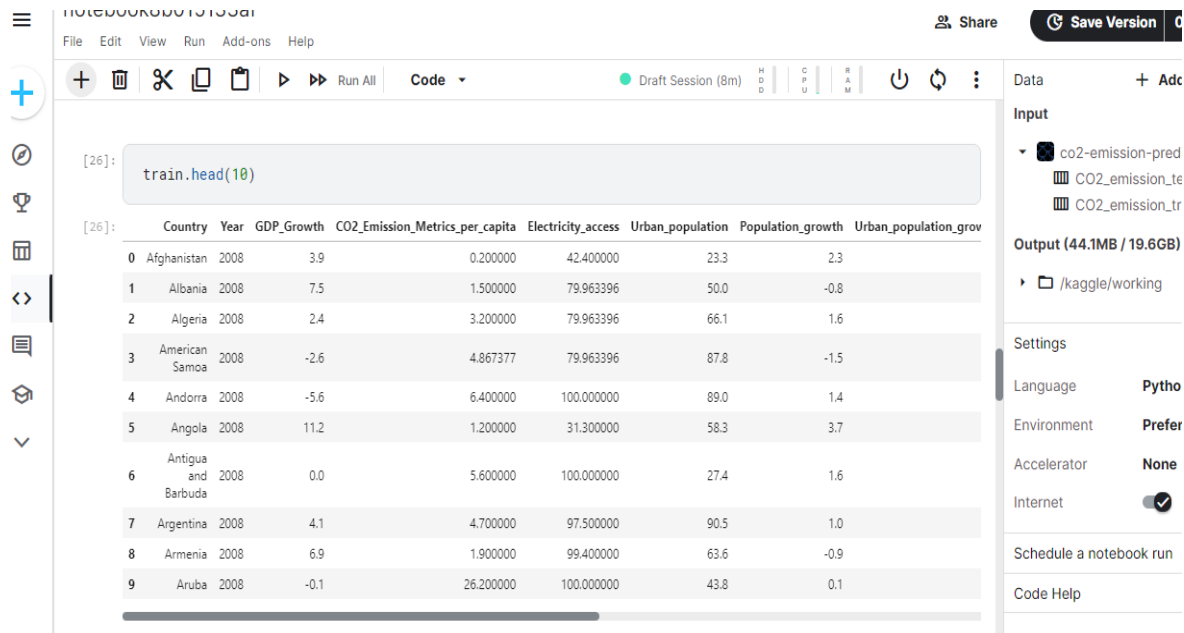


Fig. 6.1: Train Data

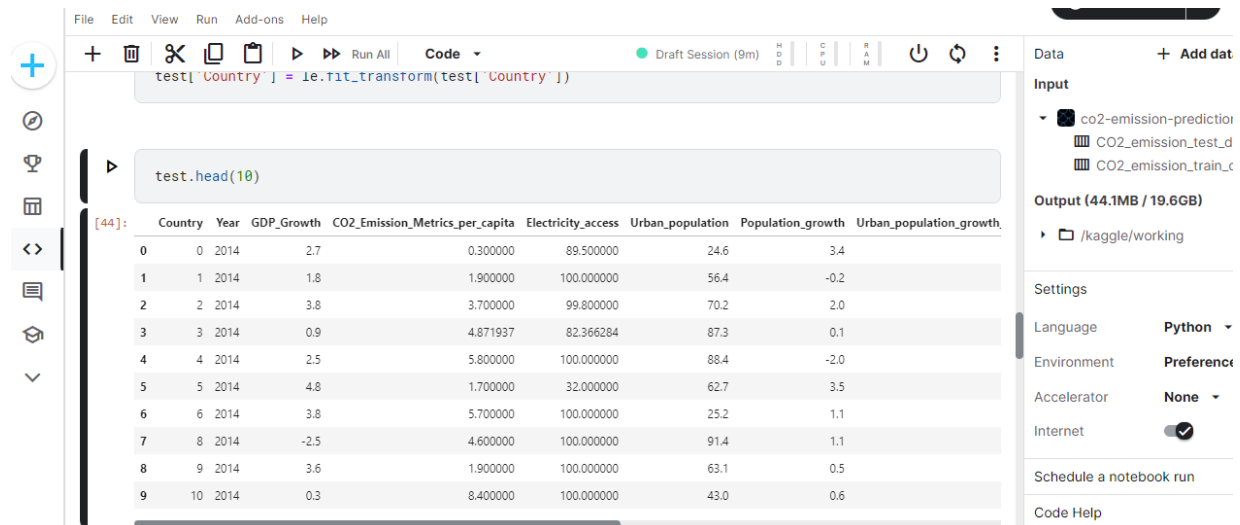


Fig. 6.2: Test Data

```

[156]: from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression

rf = RandomForestRegressor(min_samples_split=2,min_samples_leaf=2,random_state=0)
rf.fit(X_train, y_train)

[156_ RandomForestRegressor(min_samples_leaf=2, random_state=0)

[157]: predicted_test = rf.predict(X_test)
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predicted_test)))

Mean Absolute Error: 0.9006938176437221
Mean Squared Error: 2.4974077863444806
Root Mean Squared Error: 1.5803188875491176

```

Fig. 6.3: Random Forest

As we see that we got accuracy of 90 percentage if we train the data sets with the random forest regression machine learning algorithm

```

[160]: dt=DecisionTreeRegressor(min_samples_split=2,min_samples_leaf=1,random_state=0)
dt.fit(X_train,y_train)

[160_ DecisionTreeRegressor(random_state=0)

[161]: predicted_test_dt = dt.predict(X_test)
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test_dt))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test_dt))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predicted_test_dt)))

Mean Absolute Error: 1.1736218957627302
Mean Squared Error: 5.770587461832753
Root Mean Squared Error: 1.083338310853415

```

Fig. 6.4: Decision Tree

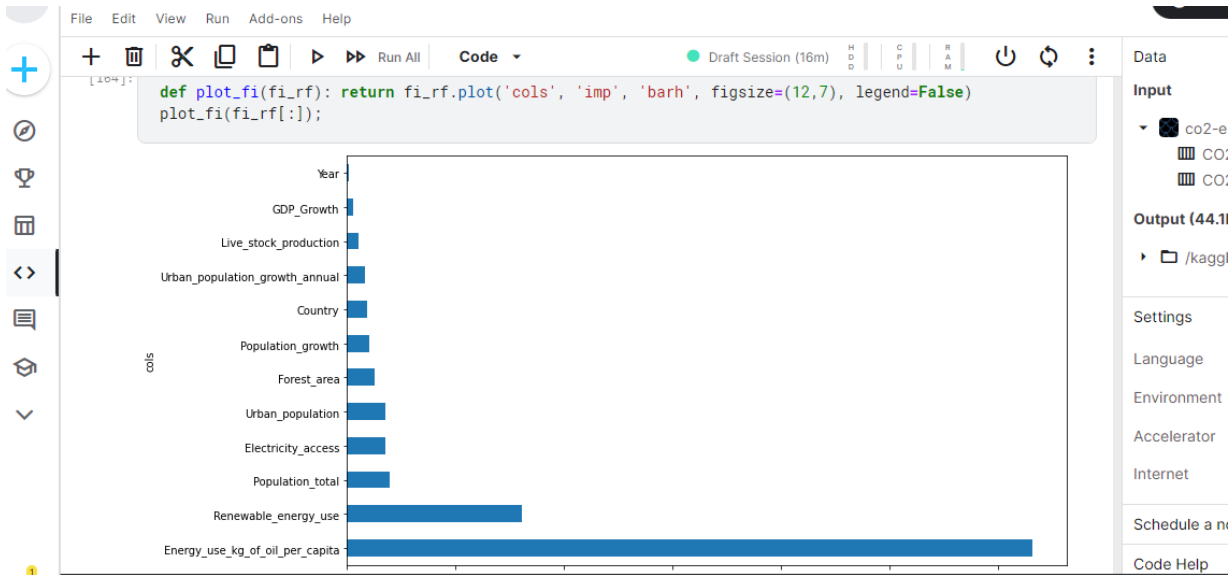


Fig: 6.5: Feature Extraction

SUMMARY

The project is used to demonstrate the machine learning models to predict the carbon dioxide emissions with high accuracy when compared with the previously implemented machine learning models. By using this project we can approximately predict the carbon dioxide emissions using various machine learning algorithms like random forest and decision trees. The feature extraction is also done so as to predict the factors that are mostly responsible for the carbon dioxide emissions. We can observe that there are some factors that gives negative impact on co2 emissions and some factors that are positive to co2 emissions. One should avoid factors that are related to co2 emissions in increase and increase the usage of factors that are positively related to co2 emissions so as to reduce the emissions in future.

CONCLUSION

As we have seen the feature importance: Cause Analysis,

- The Energy use per kg oil per capita
- Renewable energy usage (% Total)
- Urban Population(% total population)
- Population (total)
- Electricity access(% total population)

These are the features got most importance and therefore have more impact on the Target: 'CO2 Emission'.

In the above given features as we see that the energy use per kg oil percapita and renewable energy usage are the most dependent features on carbon dioxide emissions. But if we see the energy use per kg oil is the negative factor so we should stop using it and renewable energy use is the positive factor so we have to increase

its usage to reduce the carbon dioxide emissions in future. So we should increase the usage of renewable energy and decrease usage of energy per kg oil.

As we see after the feature extraction the above mentioned factors are more dependent on carbon dioxide emissions so we have to take necessary actions to reduce the negatively impacted factors so as to reduce the carbon dioxide emissions.

REFERENCES

- [1]. "Overview of Greenhouse Gases | US EPA", US EPA, 2020. [Online]. Available: <https://www.epa.gov>. [Accessed: 07- Sep- 2020].
- [2] S. Yeasmin, "Benefits of Artificial Intelligence in Medicine," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 2019, pp. 1-6, doi: 10.1109/CAIS.2019.8769557.
- [3] F. Kane, "Machine learning vs. deep learning: what's the difference? Udemy Blog, Udemy Blog, 2020. [Online]. Available: <https://blog.udemy.com>. [Accessed: 07- Sep- 2020].
- [4] D. Sullivan, "How Machine Learning Works, As Explained By Google", MarTech Today, 2020. [Online]. Available: <https://martechtoday.com/how-machine-learning-works-150366>. [Accessed: 07- Sep- 2020].

APPENDIX : SAMPLE CODE

```
import pandas as pd
import numpy as np
import sklearn
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
from sklearn.model_selection import train_test_split

train= pd.read_csv("../input/co2-emission-prediction-
data/CO2_emission_train_data_set.csv")
test= pd.read_csv("../input/co2-emission-prediction-
data/CO2_emission_test_data_2014.csv")

train.shape

test.shape

train.drop_duplicates()

test.drop_duplicates()

train.isnull().values.any()

test.isnull().values.any()

train.info()

train['Energy_use_kg_of_oil_per_capita']=
train['Energy_use_kg_of_oil_per_capita'].apply(lambda x: x.replace(',',''))
test['Energy_use_kg_of_oil_per_capita']=

test['Energy_use_kg_of_oil_per_capita'].apply(lambda x: x.replace(',',''))
train['Population_total']= train['Population_total'].apply(lambda x: x.replace(',',''))
test['Population_total']= test['Population_total'].apply(lambda x: x.replace(',',''))

train['Energy_use_kg_of_oil_per_capita'].tail()

train['Population_total'].head()

train['GDP_Growth']=train['GDP_Growth'].apply(lambda x:x.replace('..',''))

train['CO2_Emission_Metrics_per_capita']=train['CO2_Emission_Metrics_per_capita'].
```

```

apply(lambda x:x.replace('..',''))
train['Electricity_access']=train['Electricity_access'].apply(lambda x:x.replace('..',''))
train['Urban_population']=train['Urban_population'].apply(lambda x:x.replace('..',''))
train['Population_growth']=train['Population_growth'].apply(lambda x:x.replace('..',''))

train['Urban_population_growth_annual']=train['Urban_population_growth_annual'].ap
ply(lambda x:x.replace('..',''))
train['Live_stock_production']=train['Live_stock_production'].apply(lambda
x:x.replace('..',''))
train['Forest_area']=train['Forest_area'].apply(lambda x:x.replace('..',''))
train['Renewable_energy_use']=train['Renewable_energy_use'].apply(lambda
x:x.replace('..',''))

train['Energy_use_kg_of_oil_per_capita']=train['Energy_use_kg_of_oil_per_capita'].ap
ply(lambda x:x.replace('..',''))
train['Population_total']=train['Population_total'].apply(lambda x:x.replace('..',''))

test['GDP_Growth']=test['GDP_Growth'].apply(lambda x:x.replace('..',''))
test['CO2_Emission_Metrics_per_capita']=test['CO2_Emission_Metrics_per_capita'].a
pply(lambda x:x.replace('..',''))

test['Electricity_access']=test['Electricity_access'].apply(lambda x:x.replace('..',''))
test['Urban_population']=test['Urban_population'].apply(lambda x:x.replace('..',''))
test['Population_growth']=test['Population_growth'].apply(lambda x:x.replace('..',''))

test['Urban_population_growth_annual']=test['Urban_population_growth_annual'].appl
y(lambda x:x.replace('..',''))
test['Live_stock_production']=test['Live_stock_production'].apply(lambda
x:x.replace('..',''))
test['Forest_area']=test['Forest_area'].apply(lambda x:x.replace('..',''))
test['Renewable_energy_use']=test['Renewable_energy_use'].apply(lambda
x:x.replace('..',''))
test['Energy_use_kg_of_oil_per_capita']=test['Energy_use_kg_of_oil_per_capita'].appl
y(lambda x:x.replace('..',''))
test['Population_total']=test['Population_total'].apply(lambda x:x.replace('..',''))

train['GDP_Growth']=pd.to_numeric(train['GDP_Growth'])
train['CO2_Emission_Metrics_per_capita']=pd.to_numeric(train['CO2_Emission_Metri
cs_per_capita'])

train['Electricity_access']=pd.to_numeric(train['Electricity_access'])
train['Urban_population']=pd.to_numeric(train['Urban_population'])
train['Population_growth']=pd.to_numeric(train['Population_growth'])
train['Urban_population_growth_annual']=pd.to_numeric(train['Urban_population_gro
wth_annual'])

train['Live_stock_production']=pd.to_numeric(train['Live_stock_production'])

```

```

train['Forest_area']=pd.to_numeric(train['Forest_area'])

train['Renewable_energy_use']=pd.to_numeric(train['Renewable_energy_use'])
train['Energy_use_kg_of_oil_per_capita']=
pd.to_numeric(train['Energy_use_kg_of_oil_per_capita'])
train['Population_total']=pd.to_numeric(train['Population_total'])

test['GDP_Growth']=pd.to_numeric(test['GDP_Growth'])
test['CO2_Emission_Metrics_per_capita']=pd.to_numeric(test['CO2_Emission_Metrics_per_capita'])
test['Electricity_access']=pd.to_numeric(test['Electricity_access'])
test['Urban_population']=pd.to_numeric(test['Urban_population'])

test['Population_growth']=pd.to_numeric(test['Population_growth'])
test['Urban_population_growth_annual']=pd.to_numeric(test['Urban_population_growth_annual'])
test['Live_stock_production']=pd.to_numeric(test['Live_stock_production'])

test['Forest_area']=pd.to_numeric(test['Forest_area'])

test['Renewable_energy_use']=pd.to_numeric(test['Renewable_energy_use'])
test['Energy_use_kg_of_oil_per_capita']=
pd.to_numeric(test['Energy_use_kg_of_oil_per_capita'])
test['Population_total']=pd.to_numeric(test['Population_total'])

train['GDP_Growth']=train['GDP_Growth'].fillna(train['GDP_Growth'].mean())

train['CO2_Emission_Metrics_per_capita']=train['CO2_Emission_Metrics_per_capita'].fillna(train['CO2_Emission_Metrics_per_capita'].mean())
train['Electricity_access']=train['Electricity_access'].fillna(train['Electricity_access'].mean())

train['Urban_population']=train['Urban_population'].fillna(train['Urban_population'].mean())
train['Population_growth']=train['Population_growth'].fillna(train['Population_growth'].mean())

train['Urban_population_growth_annual']=train['Urban_population_growth_annual'].fillna(train['Urban_population_growth_annual'].mean())
train['Live_stock_production']=train['Live_stock_production'].fillna(train['Live_stock_production'].mean())

train['Forest_area']=train['Forest_area'].fillna(train['Forest_area'].mean())
train['Renewable_energy_use']=train['Renewable_energy_use'].fillna(train['Renewable_energy_use'].mean())

train['Energy_use_kg_of_oil_per_capita']=train['Energy_use_kg_of_oil_per_capita'].fill

```

```

na(train['Energy_use_kg_of_oil_per_capita'].median())
train['Population_total']=train['Population_total'].fillna(train['Population_total'].mean())

test['GDP_Growth']=test['GDP_Growth'].fillna(test['GDP_Growth'].mean())

test['CO2_Emission_Metrics_per_capita']=test['CO2_Emission_Metrics_per_capita'].fillna(test['CO2_Emission_Metrics_per_capita'].mean())
test['Electricity_access']=test['Electricity_access'].fillna(test['Electricity_access'].mean())
test['Urban_population']=test['Urban_population'].fillna(test['Urban_population'].mean())
test['Population_growth']=test['Population_growth'].fillna(test['Population_growth'].mean())
test['Urban_population_growth_annual']=test['Urban_population_growth_annual'].fillna(test['Urban_population_growth_annual'].mean())

test['Live_stock_production']=test['Live_stock_production'].fillna(test['Live_stock_production'].mean())

test['Forest_area']=test['Forest_area'].fillna(test['Forest_area'].mean())
test['Renewable_energy_use']=test['Renewable_energy_use'].fillna(test['Renewable_energy_use'].mean())

test['Energy_use_kg_of_oil_per_capita']=test['Energy_use_kg_of_oil_per_capita'].fillna(test['Energy_use_kg_of_oil_per_capita'].median())
test['Population_total']=test['Population_total'].fillna(test['Population_total'].mean())

train.isnull().sum()

test.isnull().sum()

train.describe()

test.describe()

train.head(10)

test.head(10)

le = preprocessing.LabelEncoder()

train['Country'] = le.fit_transform(train['Country'])
test['Country'] = le.fit_transform(test['Country'])

plt.figure(figsize=(10,7.5))
cor = train.corr(method='pearson')
sns.heatmap(cor, cmap="RdYlGn")

```

```

plt.show()

y_train= train['CO2_Emission_Metrics_per_capita']
X_train= train.drop(['CO2_Emission_Metrics_per_capita'], axis=1)

y_test= test['CO2_Emission_Metrics_per_capita']
X_test= test.drop(['CO2_Emission_Metrics_per_capita'], axis=1)

from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression

rf =
RandomForestRegressor(min_samples_split=2,min_samples_leaf=2,random_state=0
)
rf.fit(X_train, y_train)

predicted_test = rf.predict(X_test)
from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test,
predicted_test)))

print('R2 Score:', metrics.r2_score(predicted_test, y_test))

def rf_feat_importance(m, df):
    return pd.DataFrame({'cols':df.columns, 'imp':m.feature_importances_}
        ).sort_values('imp', ascending=False)
fi_rf = rf_feat_importance(rf, X_train);

dt=DecisionTreeRegressor(min_samples_split=2,min_samples_leaf=1,random_state=
0)
dt.fit(X_train,y_train)

predicted_test_dt = dt.predict(X_test)
from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test_dt))

print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test_dt))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test,
predicted_test_dt)))

print('R2 Score:', metrics.r2_score(predicted_test_dt, y_test))

```

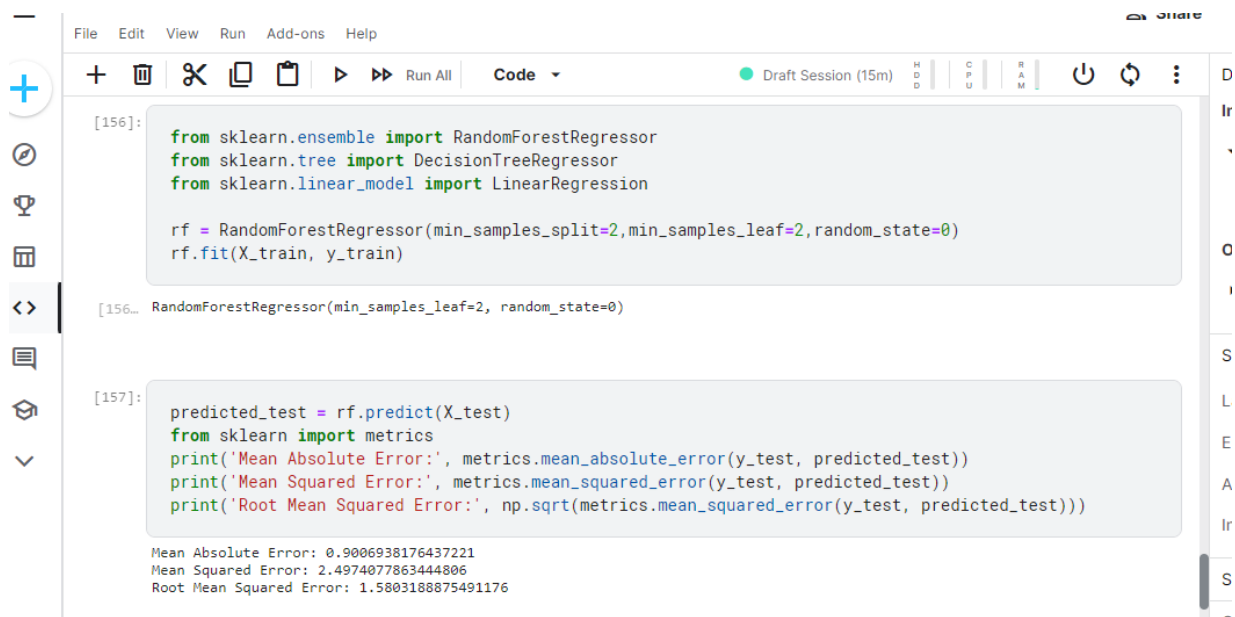
```

def dt_feat_importance(m, df):
    return pd.DataFrame({'cols':df.columns, 'imp':m.feature_importances_}
                        ).sort_values('imp', ascending=False)
fi_dt = dt_feat_importance(dt, X_train);

def plot_fi(fi_rf): return fi_rf.plot('cols', 'imp', 'barh', figsize=(12,7), legend=False)
plot_fi(fi_rf[:]);

```

SCREENSHOTS



The screenshot shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Run, Add-ons, Help) and a toolbar with icons for adding, deleting, copying, pasting, and running code. The notebook is titled "Draft Session (15m)". The code is written in a light blue box and includes imports for RandomForestRegressor, DecisionTreeRegressor, and LinearRegression from sklearn. The RandomForestRegressor is instantiated with min_samples_split=2, min_samples_leaf=2, and random_state=0. The fit method is called on X_train and y_train. The output shows the Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error.

```
[156]: from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression

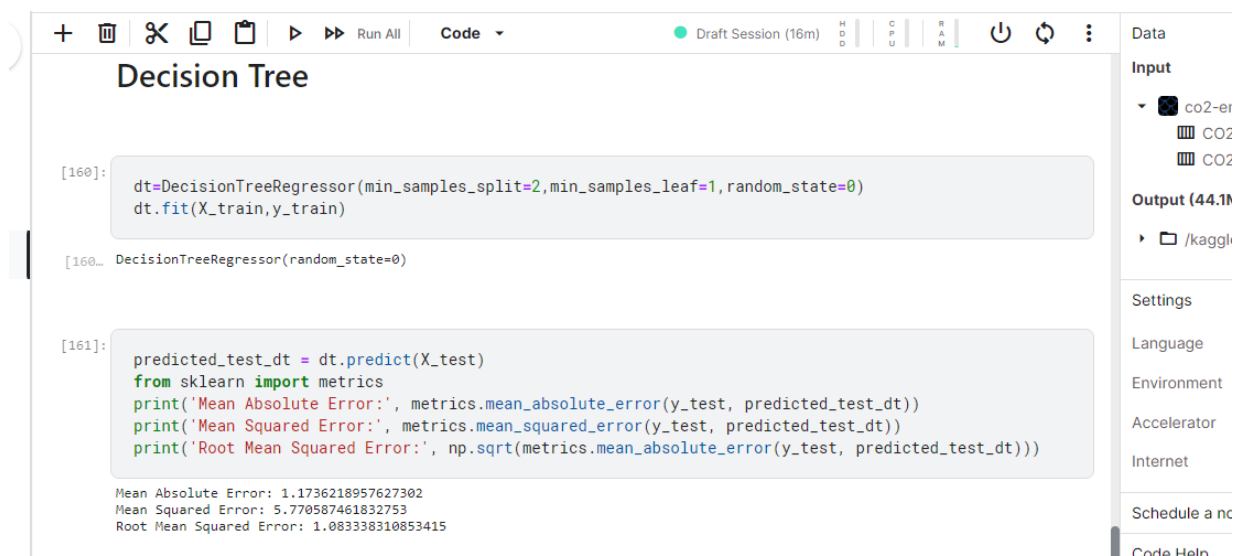
rf = RandomForestRegressor(min_samples_split=2,min_samples_leaf=2,random_state=0)
rf.fit(X_train, y_train)

[156... RandomForestRegressor(min_samples_leaf=2, random_state=0)

[157]: predicted_test = rf.predict(X_test)
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predicted_test)))

Mean Absolute Error: 0.9006938176437221
Mean Squared Error: 2.4974077863444806
Root Mean Squared Error: 1.5803188875491176
```

RANDOM FOREST



The screenshot shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Run, Add-ons, Help) and a toolbar with icons for adding, deleting, copying, pasting, and running code. The notebook is titled "Draft Session (16m)". The code is written in a light blue box and includes imports for DecisionTreeRegressor and metrics from sklearn. The DecisionTreeRegressor is instantiated with min_samples_split=2, min_samples_leaf=1, and random_state=0. The fit method is called on X_train and y_train. The output shows the Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error.

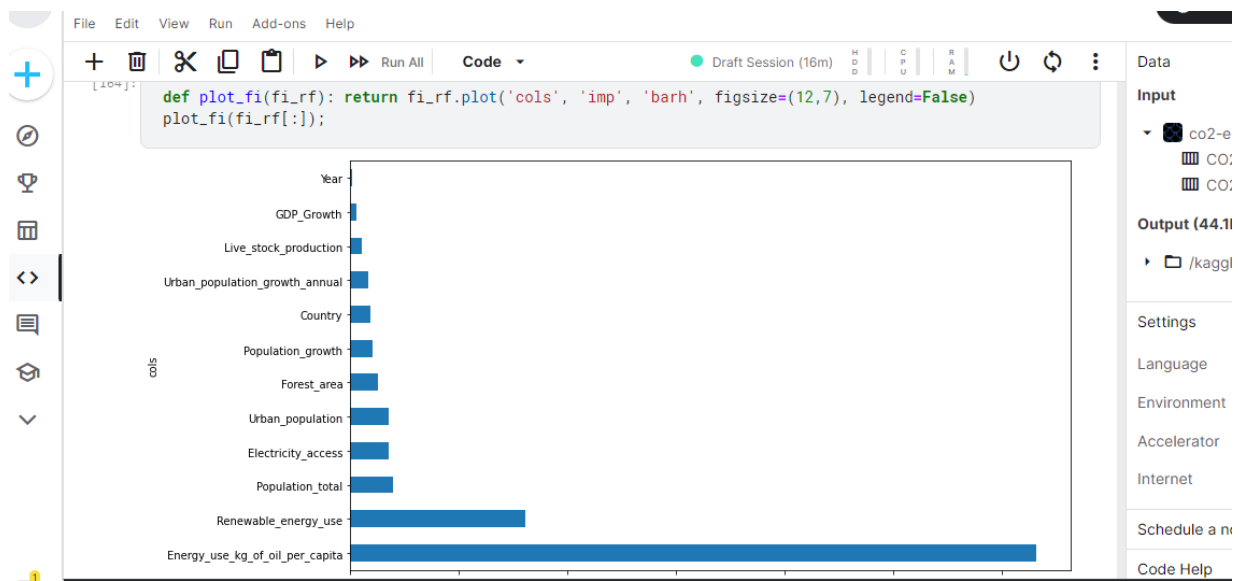
```
[160]: dt=DecisionTreeRegressor(min_samples_split=2,min_samples_leaf=1,random_state=0)
dt.fit(X_train,y_train)

[160... DecisionTreeRegressor(random_state=0)

[161]: predicted_test_dt = dt.predict(X_test)
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, predicted_test_dt))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, predicted_test_dt))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, predicted_test_dt)))

Mean Absolute Error: 1.1736218957627302
Mean Squared Error: 5.770587461832753
Root Mean Squared Error: 1.083338310853415
```

DECISION TREE



FEATURE ANALYSIS