"Modeling CO_2 Emissions Trends: A Machine Learning Framework for Climate Mitigation, A Data-Driven Approach for Sustainable Policy Development"

A Project by Manaswini Deshpande

INDEX:

SUSTAINABLE
DEVELOPMENT
GOALS
4 QUALITY EDUCATION
5 GENDER EQUALITY

Acknowledgment

I would like to express my heartfelt gratitude for the opportunity to participate in the **IBM** Skill Build Internship in AI and ML, sponsored by **CSRBOX**. This experience has been invaluable in enhancing my knowledge and skills in the field.

A special thanks to Amit Tiwari for his exceptional teaching and guidance throughout the program. His insights and expertise have been instrumental in my learning journey. I am also deeply appreciative of the support provided by my mentors—Kurian, Raju, and Meenakshi. Their encouragement and feedback have played a significant role in my development during this internship.

Thank you all for your dedication and support. This experience has truly inspired me, and I look forward to applying what I have learned in my future endeavors.

1. Introduction to Sustainable Development Goals

What are SDG goals: The United Nations Sustainable Development Goals (SDGs) are a universal call to action to end poverty, protect the planet, and ensure that all people enjoy peace and prosperity by 2030. Adopted by all United Nations Member States in 2015, these 17 interlinked goals serve as a blueprint for a more sustainable future for both the planet and its inhabitants. The SDGs address the global challenges we face, including those related to poverty, inequality, climate change, environmental degradation, peace, and justice. The goals are designed to be broad in scope yet specific in objectives, providing a comprehensive framework for governments, private sector, civil society, and individuals to contribute towards achieving them. The United Nations established these goals to mobilize global efforts around a common set of targets and indicators that promote human well-being, environmental sustainability, and economic prosperity, ensuring no one is left behind.

Actions to Be Implemented: As part of our initiative to integrate innovation with global development, each student is tasked with selecting one United Nations Sustainable Development Goal (SDG) to focus on. Our challenge is to deeply understand the intricacies of the chosen goal and craft a precise problem statement addressing a specific issue within that goal. Following this, we will leverage our technical skills to design an Artificial Intelligence (AI) and Machine Learning (ML) solution aimed at resolving this issue. This project provides an opportunity to apply AI and ML to real-world problems, promoting social impact while enhancing our skills. It is about harnessing technology for positive change and sustainability.

2. Goals Of Sustainable Development

- No Poverty - End poverty in all its forms everywhere.
- Zero Hunger - End hunger, achieve food security and improved nutrition, and promote sustainable agriculture.
- Good Health and Well-being - Ensure healthy lives and promote well-being for all at all ages.
- Quality Education - Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.
- Gender Equality - Achieve gender equality and empower all women and girls.
- Clean Water and Sanitation - Ensure availability and sustainable management of water and sanitation for all.
- Affordable and Clean Energy - Ensure access to affordable, reliable, sustainable, and modern energy for all.
- Decent Work and Economic Growth - Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all.
- Industry, Innovation, and Infrastructure - Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.
- Reduced Inequality - Reduce inequality within and among countries.
- Sustainable Cities and Communities - Make cities and human settlements inclusive, safe, resilient, and sustainable.
- Responsible Consumption and Production - Ensure sustainable consumption and production patterns.
- Climate Action - Take urgent action to combat climate change and its impacts.
- Life Below Water - Conserve and sustainably use the oceans, seas, and marine resources for sustainable development.

- Life on Land - Protect, restore, and promote sustainable use of terrestrial ecosystems, manage forests sustainably, combat desertification, and halt and reverse land degradation and halt biodiversity loss.
- Peace, Justice, and Strong Institutions - Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels.
- Partnerships for the Goals - Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development

3. Causes of Climate Pollution

Air pollution is the main cause of climate change. Human activities such as burning fossil fuels and mass deforestation lead to the increase of **carbon dioxide** in the atmosphere, which traps heat inside the atmosphere through a process called the greenhouse effect. This impacts climate patterns and sea levels around the world.

Four Major Gases That Contribute to the Greenhouse Effect

- **Carbon Dioxide**

A vital component of the atmosphere, carbon dioxide (CO_2) is released through natural processes (like volcanic eruptions) and through human activities, such as burning fossil fuels and deforestation.

- **Methane**

Like many atmospheric gases, methane comes from both natural and human-caused sources. Methane comes from plant-matter breakdown in wetlands and is also released from landfills and rice farming. Livestock animals emit methane from their digestion and manure. Leaks from fossil fuel production and transportation are another major source of methane, and natural gas is 70% to 90% methane.

- **Nitrous Oxide**

A potent greenhouse gas produced by farming practices, nitrous oxide is released during commercial and organic fertilizer production and use. Nitrous oxide also comes from burning fossil fuels and burning vegetation and has increased by 18% in the last 100 years.

- **Chlorofluorocarbons (CFCs)**

These chemical compounds do not exist in nature – they are entirely of industrial origin. They were used as refrigerants, solvents (a substance that dissolves others), and spray can propellants.

Another Gas That Contributes to the Greenhouse Effect:

- **Water Vapor**

Water vapor is the most abundant greenhouse gas, but because the warming ocean increases the amount of it in our atmosphere, it is not a direct cause of climate change.

4. Measures Taken to Improve Climate Conditions

Climate actions should integrate both mitigation and adaptation strategies, particularly for vulnerable communities facing climate impacts. Here are five solutions that serve both purposes:

1. **Protect Coastal Wetlands**: Coastal ecosystems like mangroves store carbon and shield against storms. Protecting and restoring these areas can mitigate emissions while supporting local livelihoods.
2. **Promote Sustainable Agroforestry**: Combining trees with crops and livestock enhances carbon sequestration and diversifies farmers' income, making them more resilient to climate changes.
3. **Decentralize Energy Distribution**: Smaller, community-based renewable energy systems are more resilient to climate disruptions and provide reliable power, especially in remote areas.
4. **Secure Indigenous Land Rights**: Empowering Indigenous communities to manage their lands helps protect forests, which store significant carbon, while preserving traditional knowledge and sustainable practices.
5. **Improve Mass Transit**: Expanding and retrofitting public transportation reduces emissions and enhances resilience to climate impacts, benefiting urban populations and reducing congestion.
6. **Integrating both mitigation and adaptation strategies is essential for effectively lowering CO2 levels** while supporting vulnerable communities. Actions like protecting coastal wetlands and promoting sustainable agroforestry not only sequester carbon but also enhance resilience against climate impacts. Additionally, decentralizing energy systems and improving mass transit can reduce emissions and strengthen community infrastructure, creating a comprehensive approach to the climate crisis.

By prioritizing these multifaceted solutions, policymakers can effectively tackle the climate crisis while supporting vulnerable communities.

5. Impact of CO_2 Emissions on Climate Change.

CO_2 emissions are a key driver of climate change, exceeding 35 billion tonnes annually, up from 6 billion in 1950.

Sources and Trends:

- Fossil Fuels are the main contributors, while emissions from land use have slightly decreased.
- Regional Shifts: Historically led by Europe and the US, Asia now contributes about half, with China alone responsible for over a quarter of global emissions.

Per Capita Emissions:

- High in oil-rich countries like Qatar and in the US, Canada, and Australia.
- Many Sub-Saharan African countries have minimal emissions, averaging around 0.1 tonnes per person.

Cumulative Emissions:

- Over 1.5 trillion tonnes have been emitted since 1751, with the US being the largest contributor historically.

Challenges: Disparities in emissions complicate international agreements. Major emitters (China, the US, and EU) must take the lead in reducing emissions.

Data Estimation: Estimating emissions involves analysing energy data and trade, with uncertainties impacting global figures, especially in regions like China.

Effective action requires global cooperation on both current and historical emissions.

6 Developing a Machine Learning Model to Predict CO_2 Emissions based on year

6.1 Project Overview

This project aims to predict CO2 emissions over time using a linear regression model. The dataset consists of historical CO2 emissions data from the year 1750 to 2023. We will preprocess the data, handle outliers, train a linear regression model, and visualize the results.

We begin by importing the dataset using Pandas:

Importing Necessary Libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

6.2 We begin by importing the dataset using Pandas:

Data Extraction

```
]: df = pd.read_csv(r"C:\Users\deshp\Downloads\Global Co2 Emission.csv")
```

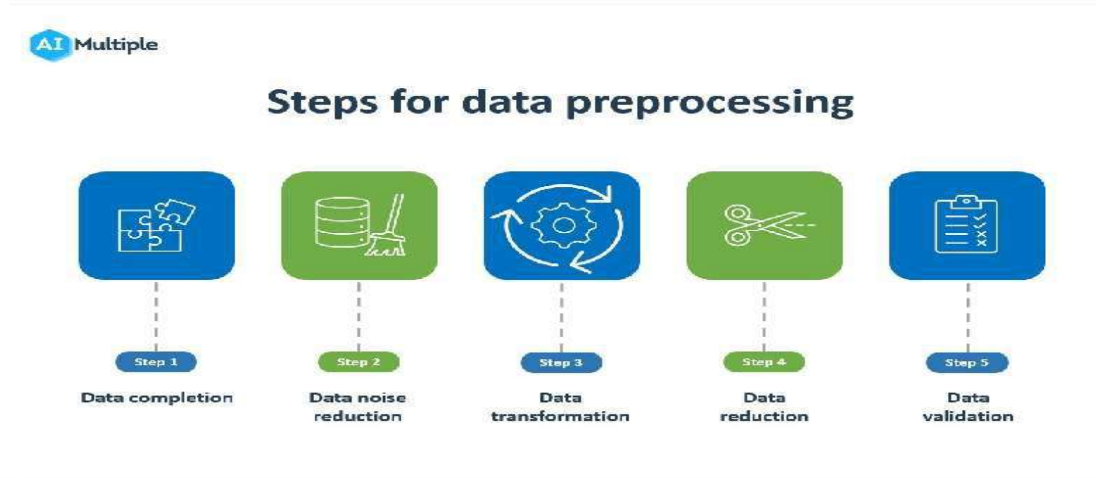
```
]: df
```

6.3 Data Preprocessing

Preprocessing refers to the transformations applied to our data before feeding it to the algorithm. Data preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format, which cannot be used for analysis hence we do pre-processing data which makes data easy to pass through algorithms and thus is helpful in designs good models

- Preprocessing refers to a set of techniques and steps used to clean, transform, and prepare raw data for analysis or modelling.
- In data preprocessing, we also perform specific tasks such as dropping unwanted columns and renaming columns to ensure our dataset is streamlined and structured for further analysis and modelling.

The below pic represents the data preprocessing and data validation after preprocessing steps



Here are required data preprocessing steps that I have followed in order to achieve a clean data for data validation

A. Dropping Unwanted Columns and Indexing

We drop the unnecessary columns and rows, then reset the index:

```
[4]: df= df.drop(columns=['Unnamed: 0', 'Unnamed: 1'])
```

```
[5]: df= df.iloc[1:]
```

```
[6]: df.reset_index(drop=True, inplace=True)
```

B. Renaming Columns

Next, we rename the columns for clarity:

```
df.rename(columns={'Unnamed: 2': 'Emission_Year', 'Unnamed: 3': 'CO2_Emissions'}, inplace=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 2 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Emission_Year  29 non-null     object 
 1   CO2_Emissions  29 non-null     float64
dtypes: float64(1), object(1)
memory usage: 592.0+ bytes
```

C. Outlier Detection and Treatment

We check for outliers using the Interquartile Range (IQR) method and replace them with the mean of non-outlier values:

```
Q1 = np.percentile(df['CO2_Emissions'], 25)
Q3 = np.percentile(df['CO2_Emissions'], 75)

# Calculate the Interquartile Range (IQR)
IQR = Q3 - Q1

# Define the lower and upper bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

```
outliers = [value for value in df['CO2_Emissions'] if value < lower_bound or value > upper_bound]
```

```
outliers
```

```
[20.64, 28.65, 37.15, 40.9]
```

```
#Calculate the median of non-outlier emissions
non_outlier_values = [value for value in df['CO2_Emissions'] if value >= lower_bound and value <= upper_bound]
median_value = np.median(non_outlier_values)
```

```
# Replace outliers in the 'Emissions' column with the median
df['CO2_Emissions'] = np.where(df['CO2_Emissions'] < lower_bound, median_value, df['CO2_Emissions'])
df['CO2_Emissions'] = np.where(df['CO2_Emissions'] > upper_bound, median_value, df['CO2_Emissions'])
```

D. Feature Selection

Define the input (X) and output (y) variables:

```
x=df.iloc[:,0].values
y=df.iloc[:,1].values
```

```
x
```

```
array(['1750', '1760', '1770', '1780', '1790', '1800', '1810', '1820',
       '1830', '1840', '1850', '1860', '1870', '1880', '1890', '1900',
       '1910', '1920', '1930', '1940', '1950', '1960', '1970', '1980',
       '1990', '2000', '2010', '2022', '2023'], dtype=object)
```

```
y
```

```
array([ 1.03,  1.03,  3.03,  3.03,  4.04,  6.04,  4.05,  5.06,  6.07,
        7.09,  8.11,  7.14,  9.19,  9.28,  9.43, 10.6 ,  8.81, 10.07,
       10.38, 10.71, 11.34,  9.74,  6.73, 10.4 , 14.45,  8.11,  8.11,
        8.11,  8.11])
```

```
print("Type of x:", type(x))
print("Shape of x before encoding:", x.shape)
print("Type of y:", type(y))
print("Shape of y:", y.shape)
```

```
Type of x: <class 'numpy.ndarray'>
Shape of x before encoding: (29,)
Type of y: <class 'numpy.ndarray'>
Shape of y: (29,)
```

6.3 Data Splitting

Data Splitting Purpose,**：** We split the data into training and testing sets,Data splitting is a critical step in machine learning that involves dividing the dataset into distinct subsets for training and testing the model. Here are the key purposes:

1. Model Training:
   - The training set is used to fit the model, allowing it to learn the underlying patterns in the data.
2. Model Evaluation:
   - The test set is reserved for evaluating the model's performance. It helps assess how well the model generalizes to unseen data.
3. Avoid Overfitting:
   - By using a separate test set, you can detect if the model is overfitting (performing well on training data but poorly on new data).
4. Performance Metrics:
   - Splitting the data allows for the calculation of performance metrics (e.g., accuracy, RMSE) on the test set, providing insights into the model's predictive capabilities.
5. Hyperparameter Tuning (if applicable):
   - If further splitting into training, validation, and test sets is done, the validation set can be used for tuning model hyperparameters without biasing the test set.

Splitting data into Train and Test

```
37]: import sklearn
     from sklearn.model_selection import train_test_split
     x_train,x_test, y_train,y_test = train_test_split(x,y, test_size = 0.2, random_state = 0)
```

```
38]: x_test
38]: array(['1770', '1950', '1980', '1880', '1860', '2000'], dtype=object)
```

```
90]: x_train = x_train.reshape(-1, 1)
     x_test = x_test.reshape(-1,1)
```

6.4 Model Training: Model training involves using a training dataset to teach the machine learning algorithm the relationships between input features and target outputs. During this process, the algorithm adjusts its internal parameters to minimize the error between its predictions and the actual target values. Effective training results in a model that can accurately predict outcomes on new, unseen data, while avoiding overfitting to the training set.

Linear Regression: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane) that minimizes the sum of squared differences between predicted and actual values. This technique is widely used for predictive analytics due to its simplicity and interpretability.

Training the Simple Linear Regression model on the Training set

```
[91]: from sklearn.linear_model import LinearRegression
      regressor = LinearRegression()
      regressor.fit(x_train, y_train)
[91]: LinearRegression()
```

6.9 Making Predictions

Prediction in machine learning refers to the process of using a trained model to estimate outcomes for new, unseen data based on learned patterns from the training dataset. After training, the model generates predictions by applying the learned relationships to input features, producing an output value. Effective predictions enable decision-making and insights across various applications, such as finance, healthcare, and environmental analysis.

Predicting the Test set results

```
n [98]: y_pred = regressor.predict(x_test)
```

```
 [100]: y_pred
t[100]: array([ 4.0871616 ,  9.23870622, 10.09729699,  7.23532776,  6.66293391,
               10.66969084])
```

6.10 Visualization

Visualize the actual vs predicted values: Visualization in data analysis involves creating graphical representations of data to uncover patterns, trends, and insights that may not be immediately apparent in raw data. Effective visualizations, such as scatter plots, line graphs, and histograms, enhance understanding by providing a clear and intuitive way to interpret complex information. In machine learning, visualizations are essential for evaluating model performance, comparing predictions to actual values, and communicating results effectively to stakeholders.

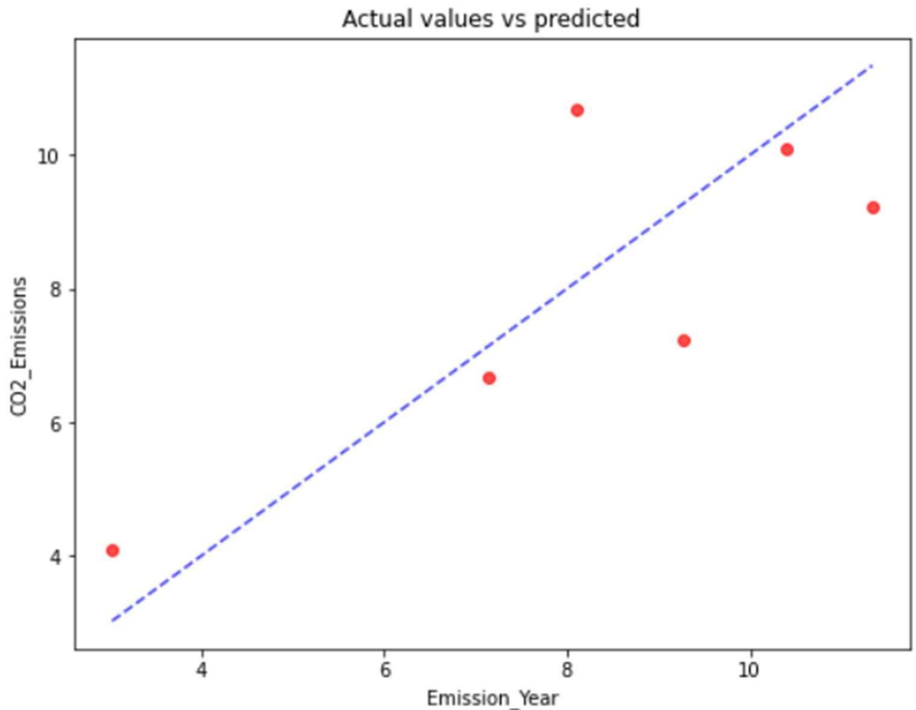
Types of Visualizations:

- **Scatter Plots**: Used to visualize the relationship between two continuous variables, often revealing correlations.
- **Line Graphs**: Effective for showing trends over time, illustrating how a variable changes sequentially.
- **Bar Charts**: Useful for comparing categorical data, displaying the frequency or value of different categories.
- **Histograms**: Help visualize the distribution of a single continuous variable by grouping data into bins.
- **Box Plots**: Useful for summarizing the distribution of data based on five-number summaries (minimum, first quartile, median, third quartile, and maximum), highlighting outliers

```python
# Plotting
plt.figure(figsize =(8,6))
plt.scatter(y_test, y_pred, color='red',alpha = 0.7)  # Scatter plot of the data points
plt.title("Actual values vs predicted")
plt.xlabel('Emission_Year')
plt.ylabel('CO2_Emissions')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='blue', alpha=0.7, linestyle='--')

plt.show()

```



6.11 Making Future Predictions

Finally, we can predict future CO2 emissions based on user input:

Let's predict Co2 emission with respective to year

```
Emission_Year= input('Climate pollution with respective to Emission_Year : ')
Emission_Year = float(Emission_Year)
prediction = regressor.predict([[Emission_Year]])  # Use lowercase 'prediction'
print("The Climate pollution with respective to CO2_Emissionsis:", prediction[0])
```

```
Climate pollution with respective to Emission_Year : 2024
The Climate pollution with respective to CO2_Emissionsis: 11.35656345567186
```

Conclusion:

This project successfully demonstrates how to preprocess historical emissions data, detect and treat outliers, train a linear regression model, and make predictions for future emissions. Further work could involve exploring more complex models or incorporating additional features for improved accuracy.

References:

causes of climate pollution - Search

Causes - NASA Science

Climate Mitigation + Adaptation Strategies

- Kaggle Dataset

Scikit-Learn Documentation