Towards a Comprehensive CO2 Emission Model

1st Kavya Vedantham
Department of Information Science
University of North Texas
Texas, USA
kavyavedantham@my.unt.edu

2nd Hari Krishna Jammula
Department of Information Science
University of North Texas
Texas, USA
harikrishnajammula@my.unt.edu

3rd Naga Pavan Bonam
Department of Information Science
University of North Texas
Texas, USA
nagapavanbonam@my.unt.edu

4th Bhavesh Gujjula
Department of Information Science
University of North Texas
Texas, USA
bhaveshgujjula@my.unt.edu

5th Pavan Krishna Khanapuram
Department of Information Science
University of North Texas
Texas, USA
pavankrishnakhanapuram@my.unt.edu

***Abstract*—Global carbon dioxide (CO2) emissions from human activities continue to increase at an alarming rate, posing significant threats to the environment and existence on Earth. While past emission trends showed a moderate rise, recent years have witnessed a rapid increase of CO2 and other pollutants being released into the atmosphere. This concerning pattern underscores the urgent need for accurate modeling and forecasting of future emission levels to inform mitigation strategies and policies. This study aims to predict global CO2 emissions for future years by leveraging the power of machine learning algorithms. A rich data set consisting of data from countries worldwide spanning the past two decades (1998-2022) is collected, incorporating various factors that influence CO2 emission levels. By applying advanced techniques such as Long Short-Term Memory (LSTM), Linear Regression, and Logistic Regression, we aim to uncover intricate relationships between CO2 emissions and influential factors.**

***Index Terms*—CO2 emission, prediction, LSTM, random forest**

I. Introduction

Climate change is mainly driven by carbon dioxide (CO2) emissions. These are mostly caused by burning fossil fuels for energy, deforestation, and industrial procedures among others. Although there are natural causes of CO2 release into the atmosphere; however, since human beings started engaging in various activities that give off huge amounts of this gas into the atmosphere there has been a significant growth in its levels within our air cover. Greenhouse gasses like CO2 trap heat from escaping back out into space thus contributing to what is referred to as global warming through their ability to retain some of it within earth's atmosphere where we live. This leads to altered weather patterns worldwide such as increased frequency of severe storms or abnormal climate events and disturbances ecosystems face due to these changes occurring on planet earth. In face of the worldwide rise in global air pollution, we just have one goal to clear CO2-emission as fast as possible. The outcome of the heated increase in air temperature and subsequent weather change requires the development of a new system to stop the gas emissions and reduce their effects. The research agenda undertaken by our study, entitled Comprehensively Sequencing Carbon Dioxide Emission Model Using Machine Learning and Data Science, will focus on utilizing data-related sciences as well as machine learning methods to model as well as predict trends in CO2 emission.

After a meticulous data preparation that includes treatment of missing values, discovery of highly correlated features through correlation matrices, dimensionality reduction using PCA, and features scaling with MinMax scaler, we were ready for a deep analysis stage. This led to using more powerful computational tools, such as LSTM networks, linear regression, and random forest models. Through the implementation of powerful machine learning techniques, our targets are to unveil the complex network of CO2 emissions and their multifaceted factors. Through a systematic analysis, our goal is to study the effect of one variable on others, from short-term to long-term scenarios, thus being able to generate projections until the year 2038.

A. Statement of Problem

How can machine learning algorithms be applied to a comprehensive global dataset spanning 1998-2022 to accurately predict future carbon dioxide (CO2) emission levels up to the year 2038? This study aims to develop robust machine learning models, such as Long Short-Term Memory (LSTM), Linear Regression, and Logistic Regression, to investigate the vast data on emissions and influencing factors like population, economic activities, and energy usage patterns. In your opinion, which algorithm or combination of algorithms is most effective for reliably forecasting CO2 emission trajectories worldwide? The problem statement emphasizes creating predictive models that can enhance understanding of emission dynamics and inform strategies to mitigate environmental impact.

II. Review of Literature

The case study is on how CO2 emissions can be predicted by using machine learning techniques and assessing the potential climate change impact. It proves its fundamental meaning by being concerned with the problem of CO2 concentration monitoring to determine and combat climate change. The study shows the central role CO2 emissions play in terms of climate change and the need to bring down the emission level to a level that is needed to reduce the damage caused to the environment. Previous studies find a co-extracting relation between CO2 emissions and multifaceted socio-economic factors and demonstrate proper predicting models [4]. The main approaches being used are these ones, dimensionality reduction techniques like Principal Component Analysis (PCA) and the application of Decision Tree Regression algorithms, for the sake of better predictions of the CO2 emission level. The research focuses on feature engineering that should enable the selection of service parameters that are appropriate for the case of the world's population, GDP of states, and emissions of different types.

Applying machine learning algorithms such as the Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) for the prediction of CO2 emissions which are dependent on economic growth and energy consumption. This way AANN with its feature of linear relationship capturing, and ANFIS which is a synthesis of fuzzy logic and neural networks, have indicated that they can accommodate the multifaceted dynamics of CO2 emissions. Recent studies have shown that the effectiveness of prediction models (i.e. coming up with accurate models) largely depends on complete and clear source data that are obtained from sources such as the World Development Indicators (WDI) database. The employment of artificial intelligence tools overlapped with data dimensional reduction and clustering techniques improve the predictive power of the model with optimal decision-making in environmental management among other fields [8]. The research [7] used a few machine learning techniques to predict CO2 in India for twenty-ten (2020 - 2030). The types of models used are statistical models such as ARIMA and SARIMAX as well as machine learning models like linear regression, random forests, and long short-term memory neural nets. These models are trained and tested via past CO2 emission data of India, between 1980 and 2019, which is obtained from the CAIT database. The data reveals a steady rise in emissions over the years in the visualization. According to the ICOS (Integrated Carbon Observation System) [2]conducted a research on rapid raise on CO2 emission up to year 2017, where once a year carbon price range serves the dual cause of meeting the call for well-timed climate gadget facts and facilitating adaptation and mitigation efforts. With exceptional adjustments within the surroundings, frequent assessments and transparent information sets are important for constructing clinical information and addressing the weather mitigation undertaking.

From study [4] which is rapidly increasing carbon dioxide (CO2) emissions and the dramatic impact they're having on the environment and global economy. These historians are convinced that the concentration of 500 ppm of CO2 will be reached by the year 2047, which will come as an inevitable tragedy. To overcome the situation, they recommend that the emission rate should be reduced to 6.37% and reversed to 23.38%.

Often already-known methodologies have covered CO2 forecasting and mitigation previously. [4] use a non-equivalent equigap model to do emission forecasting in multiple countries and try to figure out energy consumption. The research paper [6] applied LSTMs a kind of machine learning model that they used in the prediction of greenhouse gas emissions quantity for the agricultural soil.

III. Objectives of the Study

In our Project we aim to use the machine learning techniques to predict the CO2 emissions values for the next 15 years (2024- 2038). Our approach is based on the below goals we want to accomplish in our research:

- Expanding our research beyond the United States or G20 countries and exploring global countries for Carbon dioxide (co2) emission patterns.
- Data extracted from the past two decades (1998 – 2022)to ensure the latest trends in global emissions particularly emissions related to CO2 are taken into consideration.
- Implementing machine learning models such as Linear Regression, Random Forest Regression, and Long Short-Term Memory (LSTM) to predict the future CO2 values for the next 15 years up to 2038.

IV. Research Design

The research design as shown in figure 1 our project was divided into Six steps in the research design. The first step is the WDI dataset where data is extracted from the World Development Indicators where the features are related to Co2 emissions which also includes year, country and country Code, the second step is data preprocessing where we have included two major things are one with handling missing values and another is with feature scaling, and the third step is Exploratory Data Analysis (EDA) and analysis steps. And the last step is where we applied all three machine learning algorithms Linear regression, Random Forest regression and Long Short-Term Memory (LSTM) for evaluation metrics and predicting the future Co2 values for next 15 years.

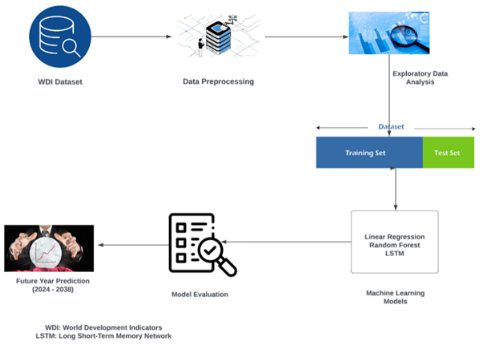

WDI Dataset
Data Preprocessing
Exploratory Data Analysis
Dataset
Training Set
Test Set
Linear Regression
Random Forest
LSTM
Machine Learning Models
Model Evaluation
Future Year Prediction (2024 - 2038)
WDI: World Development Indicators
LSTM: Long Short-Term Memory Network

Fig. 1. Research Design

V. Data Collection

Our data is collected from World Development Indicators (WDI) [3] and has been used for our research purposes and it is a data bank where all the data has been collected accurately includes data of separate groups such as poverty, people, GDP of global countries, Environment, emissions etc.

First, we have gathered specific targeting columns/features encompassing several factors activating influencing CO2 emissions. And the dataset contains 264 unique countries all over the world from the past years 1750 – 2022. And after some vast differences between the features, we have taken a total of 79 features including the features country and Year. These features are more likely related to CO2 emissions like methane gases, oil waste, cement waste, flaming co2 gases, nitrous oxide etc [9]. for the years 1998 – 2022. We have considered records from the last two decades (1998-2022) because most of the late 1700s and 1800s do not contain accurate records as included in the machine learning accuracy which results in more fluctuations in predicting future values.

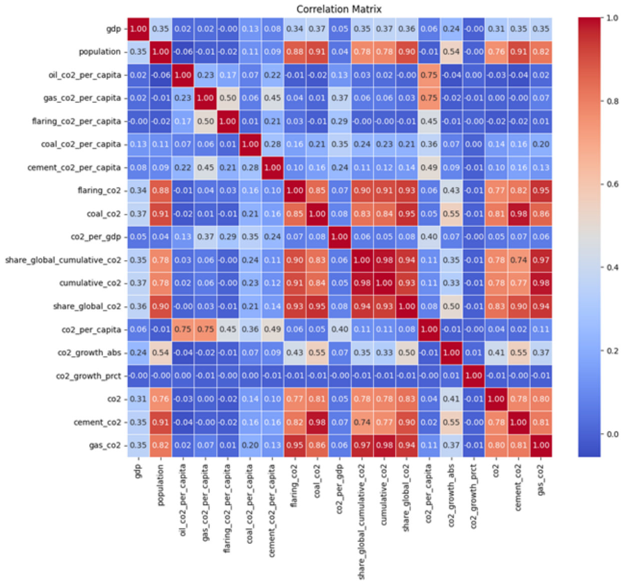


Fig. 2. Correlation Matrix

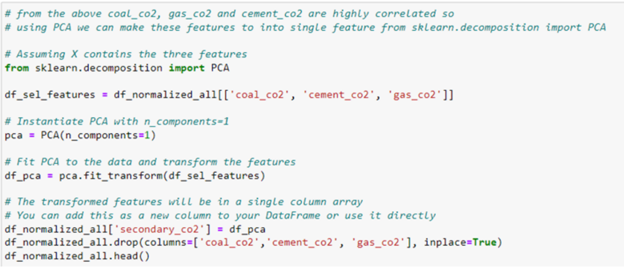
VI. EXPLORATORY DATA ANALYSIS

A. Correlation Matrix

Using Python we are detecting positive correlation among features of the data and matching the pairs above 0 coefficient to high correlation. First of all, the process performs elimination of all non-numeric columns from df_normalised_all that haven't been normalized and year from country that don't hold any useful information in correlation analysis. Subsequently, the corr() function calculates the correlation matrix for the latter columns. The current feature set contains plenty of such features, and correlation coefficients generally don't exceed zero or are less likely to do so at all. All correlation coefficients above and equal to 0.9 are thresholds and were put into high_correlation_pairs list. Then later we did feature pairs followed by their correlation values that might be scrutinized or skipped under their value, or stay as the basis for the dimensionality reduction, for example. It is imperative to grasp the relationship of factors in that joint. It thus is a crucial one within data analysis tasks and machine learning model building as it can solve the issue of multicollinearity and therefore improve the model's performance and interpretation.

B. Principal Component Analysis

This Python code snippet is grounded in the Principal Component Analysis (PCA) algorithm in the scikit-learn library and is dedicated to three main tasks, which are: (1) reducing dimensionality and (2) addressing multicollinearity among the highly correlated features of 'coal_co2', 'cement_co2', and 'gas_co2' within the read-in Data Frame In this case, the code choose these things as features and applies PCA with one principal component and maps them into this single feature space that retains the variability of the data. The process principal component transformation is then done and the resulting transformed principal component is now made the new column named 'secondary_co2' in the original Data Frame. The original correlated features that were earlier paired with this transformed principal component can now be excised. It is an approach that will improve computational efficiency and at the same time lead to enhanced model performance.



```
# from the above coal_co2, gas_co2 and cement_co2 are highly correlated so
# using PCA we can make these features to into single feature from sklearn.decomposition import PCA

# Assuming X contains the three features
from sklearn.decomposition import PCA

df_sel_features = df_normalized_all[['coal_co2', 'cement_co2', 'gas_co2']]

# Instantiate PCA with n_components=1
pca = PCA(n_components=1)

# Fit PCA to the data and transform the features
df_pca = pca.fit_transform(df_sel_features)

# The transformed features will be in a single column array
# You can add this as a new column to your DataFrame or use it directly
df_normalized_all['secondary_co2'] = df_pca
df_normalized_all.drop(columns=['coal_co2','cement_co2', 'gas_co2'], inplace=True)
df_normalized_all.head()
```

Fig. 3. Principal Component Analysis

C. Analysis

The graph has displayed a very unstable trend with more than one increase and troughs. The carbon dioxide emissions growth rate reached its highest peaks around 2005 and 2010, with some comparatively high readings during those periods.

On the other hand, the growth rate drops much more slowly in the other years that are not peak, exhibiting lower levels. In the part of the line chart that starts around 2015 and goes up to 2020, the volatility seems to have decreased a bit, when the growth rate varies through smaller fluctuations. we can see another point that is higher approximately in 2018, although it looks like the trend towards stabilization or even decrease starts towards the end of the years analyzed in the data.

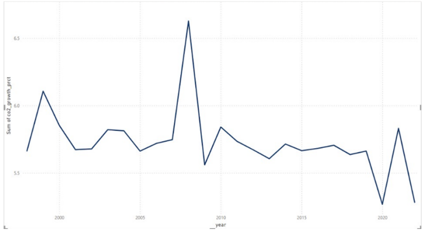


Fig. 4. Analysis of Trends in CO2 Emissions Growth Rate

The sharp drop of this variable from about 2010 to 2020 shows that there were many successes in emission cuts of the greenhouse gas “nitrous oxide” from such sources as agriculture, industrial processes, or transportation on a per capital basis over that period.

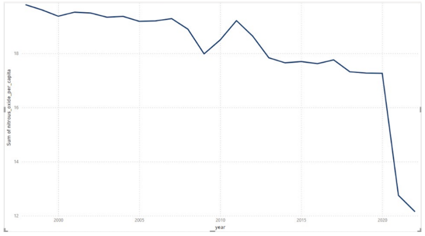


Fig. 5. Analysis of Emissions Reduction

The graph exhibits a line chart that shows an upward trend, with the CO2 emission increasing over time from around 2000 until the end of the covered period. Around 2010-2012, a remarkable inflection point or change in the rate of increase of CO2 emissions, which became sharper after that time, indicating a faster increase in emissions is evident. The overall ascent character indicates a smooth growth in such CO2-releasing activities that include but are not limited to fossil fuel combustion, industrial processes, deforestation, and other human activities that contribute to the pollution of the atmosphere with carbon dioxide.

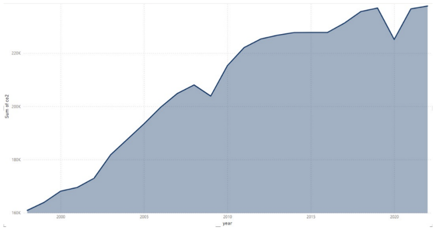


Fig. 6. CO2 Emissions: Analysis and Inflection Points

The bar height comparison makes it possible to contrast the carbon intensity of economies or their emissions efficiency for different countries. In countries with bars at the higher end, there is a higher amount of CO2 emissions per unit of GDP, which implies a more carbon-intensive. economy or a less efficient economy in terms of emissions relative to the economic output achieved. Aggregate CO2 emissions over time Sum of CO2 emissions per GDP (gross domestic product) across different countries or regions.

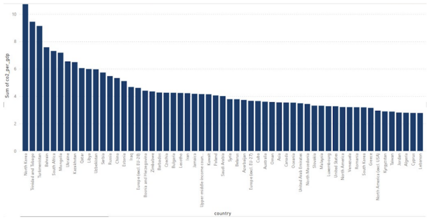


Fig. 7. A Bar Graph Comparison

VII. DATA ANALYTICS

Since our research is to predict future CO2 emission levels, the main task of the project after collecting the dataset is to ensure that the machine Learning algorithms we use can accurately and effectively analyze the dataset and future results. It is about transforming raw datasets into understandable details so that they can be understood. In the data preprocessing step we considered two main steps.

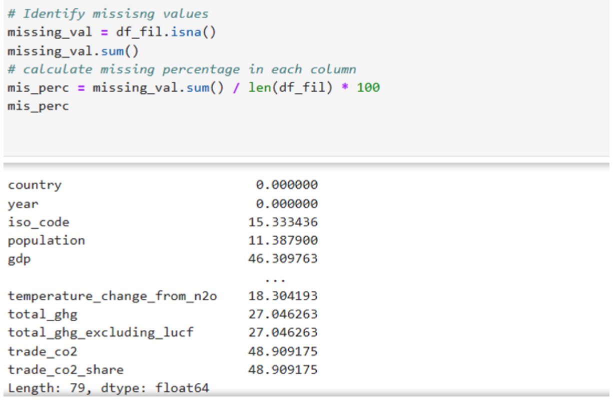
- Handling Missing Values
- Feature Scaling

A. Handling Missing Values

Handling missing values is a main aspect of data preprocessing to ensure the reliability and accuracy of the predictions. We have structured our approach into four steps while keeping the feature values unimpacted.

- Calculating the Missing Value Percentage: The step involves gathering the proportion of missing data for each of the features in our dataset and then calculating the percentage of the missing value of each feature, by this we have gained insights of each feature that are most affected by the missing data. This step describes the approach for handling missing data.
- Drop features more than 50% missing values: It often proved that more missing values could lead to problems in analysis and prediction purposes by affecting the performance of machine learning models. This ensures that only features with a perfect amount of available data can be used to perform analysis and prediction.
- Drop features more than 50% missing values: It often proved that more missing values could lead to problems in analysis and prediction purposes by affecting the performance of machine learning models. This ensures that only features with a perfect amount of available data can be used to perform analysis and prediction.
- Replace Missing values: The other percentages had the least percentage of missing values, and a reasonable percentage of missing values are handled differently as we need a different approach for each and every aspect.

- Replacing with Zero: For features where missing values are considered least percentage and not critical and do not impact the analysis or prediction, missing values are replaced with Zeros. A straightforward approach for handling missing values.
- Replacing with Median: And for features for a reasonable number of missing values which cannot be ignored, replacing them with the median of the same feature. Taking the median value for the features and replacing the missing values preserve the overall distribution and properties of features while reducing the impact of missing data.



```
# Identify missisng values
missing_val = df_fil.isna()
missing_val.sum()
# calculate missing percentage in each column
mis_perc = missing_val.sum() / len(df_fil) * 100
mis_perc
```

```
country                        0.000000
year                           0.000000
iso_code                      15.333436
population                    11.387900
gdp                           46.309763
                                 ...
temperature_change_from_n2o   18.304193
total_ghg                     27.046263
total_ghg_excluding_lucf      27.046263
trade_co2                     48.909175
trade_co2_share               48.909175
Length: 79, dtype: float64
```

Fig. 8. Handling Missing Values

The Figure 2 shows how we have calculated the missing values features so that according to highest percentage and lowest percentage has been handled using a specific techniques.

B. Feature Scaling

Scaling abilities of machine learning algorithms make it possible for the models to accurately consider our dataset and predict the future values of CO2, and thus we have conducted the necessary analysis [5]. We also made efforts to Change features values of all into one scale. The technical problem with differing feature scaling upon the dataset can make an enormous difference in the machine learning algorithms performance, especially the ones which are dependent on distance metrics and gradient descent optimization. A larger factor of variation in a specific feature compared with the other features is prone to overly influence the learning module, potentially causing the predictions to be biased and the model to act sub-optimally. Scale features imply factors that ensure each component makes a scaled up equal contribution to the model's learning output. In normalization, features are not given a disproportionate weight, which results in a more balanced distribution of influence. This allows the model to predict unbiased and rightful outcomes.

In our approach, we implemented a popularly used and widely known method that ensures the standardized scaling of the features in our dataset. One of the methodologies that we applied was scaling the range of feature values to lie between 0 and 1, thereby increasing the consistency across all features.

The formula for Min-Max scaling:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

Where:

- X is the original value of the feature,
- X_{min} is the minimum value of the feature,
- X_{max} is the maximum value of the feature, and
- X_{norm} is the normalized value of the feature.

Particularly, we utilize Min-max Scaling. In the process the minimum and maximum values of each feature are converted to the given range between 0 and 1 through this method, we intend to implement an imperative outline where every aspect is standardized, thus making it possible to have a fair and appropriate comparison between the entire dataset. The systematic normalization procedure is part of the implementation of the equalizing factor to maintain a balance among all the features, which makes the analysis more effective and the prediction better.

C. Splitting the data into training and testing models

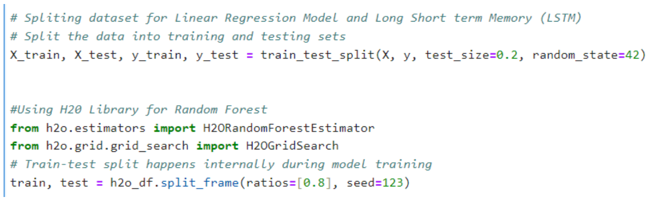
The code snippet splits the dataset into training and testing out sets for linear regression using scikit-research's train_test_split function. The enter features (X) and goal variable (y) are surpassed as parameters to the function, together with test_size=0.2, indicating that 20% of the facts will be reserved for trying out at the same time as the remaining 80% will be used for education. Additionally, random_state = 42 ensures implacability through solving the random seed. The feature returns four variables: X_train and y_train containing the training functions and target values, respectively, and X_test and y_test containing the testing functions and target values. This splitting manner allows the model to study on a subset of the facts and evaluated on unseen data, facilitating the evaluation of its generalization overall performance.

It also demonstrates the usage of H2O library for training a Random Forest model. First, the desired modules for Random Forest estimation and grid search are imported. The split_frame function from H2O library internally divides the records saved in h2o_df into training and checking out sets, with eighty% of the records allotted for schooling and the ultimate 20% for checking out.

D. Machine learning models

The Regression models which are applied for in our research are:

- Linear Regression



```
# Spliting dataset for Linear Regression Model and Long Short term Memory (LSTM)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


#Using H20 Library for Random Forest
from h2o.estimators import H2ORandomForestEstimator
from h2o.grid.grid_search import H2OGridSearch
# Train-test split happens internally during model training
train, test = h2o_df.split_frame(ratios=[0.8], seed=123)
```

Fig. 9. Splitting data into Test and Train

- Random Forest (H20)
- Long Short-Term Memory (LSTM) Networks

We have selected three models—Linear Regression, Random Forest, and Long Short-Term Memory (LSTM)—to analyze and predict CO2 emissions. After accurate data preprocessing to ensure accuracy and reliability. By using a distinct range of machine learning techniques, including traditional regression analysis, ensemble learning with Random Forest, and deep learning with LSTM networks, we aim to systematically assess and forecast CO2 emissions patterns. Through this complex approach, we reveal insights that contribute to an understanding of environmental dynamics and enable informed decision-making for sustainable development.

VIII. DATA VISUALIZATION AND RESULTS

A. Linear Regression

Analyzing the link between one or more independent variables—also referred to as predictor variables or contributing factors—and a dependent variable—in this case, CO2 emissions—linear regression is a basic statistical technique. Based on the supposition that the predictor variables and the target variable under investigation have a linear or straight-line relationship, the linear regression model works [10]. The equation represents this linear relationship analytically.

The simple linear regression:

$$y = a_0 + a_1 * x_1 + e$$

Where y = dependent variable, x = independent variable, a0 = y_intercept, a1 = slope, e = error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

The dependent variable, y, represents CO_2 emissions in this regression equation. The independent variables, $x_1, x_2, \ldots, x_n$, indicate contributing factors such as population, energy consumption, industrial output, and economic growth. The regression coefficients, $\beta_1, \beta_2, \ldots,$ and γ_n, represent the weighted values of the predictor variable, and the error term, ε, is shown.

The linear regression model is trained by identifying the values of regression coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_n)$ that minimize the error squared sum between what is predicted and the true CO_2 emission values. Such routine is often conducted with OLS, made for training data, or, more generally, with gradient descent optimization. Mean Squared Error (MSE) is frequently the goal function to be reduced, also referred to as the loss function or cost function: The Mean Squared Error (MSE) is frequently the goal function to be reduced, also referred to as the loss function or cost function.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The function will be n, which is how many data points we have, where y_i is the actual CO2 emission value for the i-th data point and ŷ_i is the predicted CO2 emission value for the i-th data point.

After the model is n trained, it can be used to predict outputs for new inputs by inputting the values of independent variables into the optimal equation that has been learned. Evaluation of the model's performance is based on MSE, RMSE, and R-squared values Metrics which are examples of the model's success in predicting events as well as economic movements. When linear regression is a basic and easy-to-sense model, it cannot include the non-linear relationships and complex interactions between the predictor and the target variables. However, in similar circumstances higher level statistical models such as Random Forest or LSTM would be more adequate.

B. Random Forest Regression

Random Forest Regression is consists of the ensemble approach, which occurs due to joined decision trees and leads to an increase in the accuracy of prediction and a low level of overfitting. Instead of H2O, we used the H2O library, which is designed to facilitate a variety of machine learning algorithms like Random Forest Regression which is a high-performance implementation of this algorithm.

The H2O library [1] has been developed to be highly integrable with both relational or non-relational databases and in the process making it easy for building and deploying machine learning models. It provides distributed parallel procession which allows to conduct of tasks among many nodes with only one execution. So, it is convenient for dealing with a CO2 emissions dataset that has an array of countries and attributes.

The Random Forest Regression algorithm in H2O works as follows: The Random Forest Regression algorithm in H2O works as follows:

- In the R sample mod, draw a bootstrap sample size of those N from the initial data.
- Grow a decision tree from the bootstrap sample, with the following modification: the experience can be random at every node instead of selecting the best split amongst all the predictors, randomly choosing m of the predictors (this is a parameter, m, that is set beforehand) and choosing the best split from those variables. It amplifies

the entropy and thus allows the decrease in correlation between individual tree stands.
- Restate steps 1 and 2 to make a tree forest of Bs (where B is again another parameter B which normally could be 50 or 100).
- Such as for a new data point x, we need to calculate the prediction by taking the average of the predicted values from all trees in the forest.

H2O provides a range of parameters and adjusting options for the Random Forest Regression model, through which we can ensure the model performs to the optimal level and preventively avoid overfitting. undefined

- ntrees: The number of trees to grow in the Random Forest (larger values usually are of help in getting better results with a risk of overfitting).
- max_depth: One of the parameters is the maximal depth of the trees in the Random Forest (the deeper the trees - the more complex patterns can be caught, but oversampling would be possible).
- sample_rate: The fraction of rows suggested for each tree (to leave enough room for randomness) is variation.
- col_sample_rate: Column fractions (features) to split (commonly a value between 0.3 and 0.7) is another important parameter for splitting to add an element of randomness.

Evaluation of the Random Forest Regression model consists of measuring such metrics as Mean Squared Error (MSE), Root Mean Squared Error(RMSE), and R-squared value. Moreover, tackling overfitting by employing approaches like cross-validation is another problem to be solved.

C. Long Short Term Memory (LSTM)

One kind of recurrent neural network (RNN) that works well with sequential input and provides a framework for long-term dependencies is the LSTM (long short-term memory) model. Particularly in CO2 emission prediction, LSTM has an evident merit in that it is able to identify time-ordered information and study the trend of the datasets.

The LSTM model is constructed of a string of memory cells interconnected and each containing gates that are used for the management of the flow information. These gates determine whether to process or destroy the inputs of the preceding time steps, let alone whether to keep minor or whatever information. This characteristic of LSTMs to ”selectively” remember or forget data makes them very useful in coping with long-term dependencies within the sequential data. The training algorithm, in turn, is based on a loss function minimization - e.g. mean squared error (MSE) - that compares the generated CO2 readings with the actual ones. The stochastic gradient descent, or its variations like Adam, are the optimization algorithms that are the most common in LSTM training.

In the forecasting stage, the LSTM model accepts the CO2 emission history together with other known relevant features (for example, population, matter consumption, and GDP) as input, and then predicts future time step values. The model’s design, consisting of memory cells along with the gating mechanisms, enables it to filter out the noise from the data and to store its past status in its memory, making it a suitable choice to be a long-term forecaster.

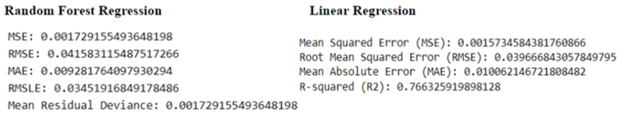
In this study, the LSTM model that can deal with long-term dependencies and trends observationally was implemented to predict CO2 emission value globally. We run the LSTM model, the historical CO2 emissions data as well as other features that are useful for forecasting, and at last use the trained model to predict future CO2 emission values.

D. Result Comparison Graph

The Figure 10 depicts two performance metrics, the linear regression appears to have a better fit for the dataset compared to random forest. Here’s a comparison of the two models: Here’s a comparison of the two models:

- Mean Squared Error (MSE): The Linear Regression has the lowest MSE in the computation for 6.9045 meaning that it is better than the Random Forest Regression (8.0017) which is higher. This statement shows that by and large, the simple regression forecasts could be considered more accurate than those coming from random forests.
- Root Mean Squared Error (RMSE): After the example of MSE, it is RMSE which comes lower of linear regression (2.6291) compared with random forest (2.8281). These findings thus affirm our conclusion that, over time, the linear regression model comparatively had a higher accuracy level.
- Mean Absolute Error (MAE): Yet the difference in a value is something minor, still MAE for linear regression (1.8911) is a bit better than random forest (1.9361). Therefore, this statistic stands for the average deviation of the error and it is the lowest value always for the model it means that the predicted values were good by the model from the actual.

Finally, in general, we can say that the Linear Regression model works quite well with this particular data set. The Tool accomplished a lower error on the entire metrics of (MSE, RMSE, and MAE) in comparison with Random Forest Regression. This is why the straightforward regression model for such tasks is likely to yield more precise predictions which means a better fit to the data.



Random Forest Regression

```
MSE: 0.001729155493648198
RMSE: 0.041583115487517266
MAE: 0.009281764097930294
RMSLE: 0.03451916849178486
Mean Residual Deviance: 0.001729155493648198
```

Linear Regression

```
Mean Squared Error (MSE): 0.0015734584381760866
Root Mean Squared Error (RMSE): 0.03966843057849795
Mean Absolute Error (MAE): 0.010062146721808482
R-squared (R2): 0.766325919898128
```

Fig. 10. Error Metrics

The ”CO2 Emission Forecasts of Different Countries” graph is a Bar graph that shows the anticipated CO2 emissions for several different nations between 2023 and 2033. Time is represented by the x-axis, while CO2 emission level is shown by the y-axis. Every country is depicted on one single line here giving the opportunity of a comparative assessment of their emission levels. Figure 11 shows that the bars are moving upwards which implies that the majority of the countries are predicted to bear an increase in CO2 emissions throughout the forecasted period. This allows countries with high or low forecast emission levels to be easily identified for subsequent emission reduction through the implementation of targeted strategies and policies.

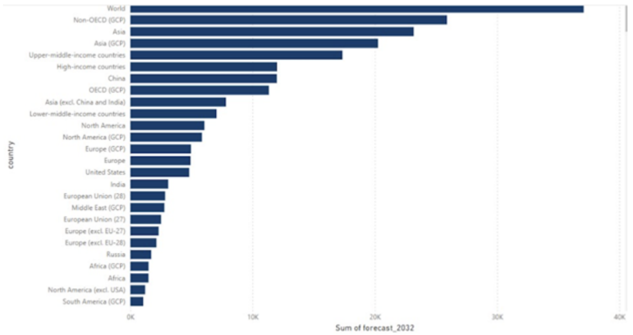


Fig. 11. CO2 emission forecasts of different countries

IX. CONCLUSION

As a result of the experimental research we conducted, we managed to employ machine learning methods that are linear Regression, Random Forest Regression (with the H2O library) and Long Short-Term Memory (LSTM) to predict future CO2 emissions on a global scale. Through such a study, carried out on a dataset that goes back to the last two decades and encompasses all the causes in many countries, we achieved core findings about the common trends, and interconnections governing global CO2 emissions. Since the Linear Regression model serves as the foundation for understanding the underlying relationships between predictors and CO2 emissions, we can get to a baseline level of knowledge This is quite uncomplicated and a machine will be able to understand the main points of the work. Random Forest Regression model which was an H2O-enabled library of impeccable speed and performance accuracy gave a superior forecasting result by using multiple decision trees to take care of the non-linear relationships while also addressing overfitting issues. LSTM model being able to characterize long-term dependencies and temporal patterns was a primary determinant of its capacity to predict future CO2 emission levels.

The linear regression model worked better than the random forest algorithm when this data set was studied. The model Linear Regression had a number MSE more lower - its - 6.9045 MSE than that of the model Random Forest Regression with MSE of 8.0017. On top of that, the Linear Regression model obtained smaller values for Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) that means the Linear Regression model was more accurate in predicting response as compared with the Univariate Linear Regression model. On the other hand, we should also bear in mind that linear regression is underpinned by the assumption of there being a linear connection between the variables. If the underlying data shows a linearity pattern, linear models may apply; however, if the data exhibits significant non-linearity, looking into other models will be necessary. For this specific duty, both layers can be applied effectively. However, while making long-term predictions on CO2 emissions, the LSTM network could be a suitable vocation since LSTM can handle complex time series data where in our reseach we have predicted CO2 values for next 15 years (2024 - 2038). Within the frame of the lower MSE and other indexes, the linear regression will be the most beneficial from the viewpoint of CO2 emissions prediction because of the given dataset and its environment.

Our research finally brings us a step closer to understanding global CO2 emissions and opens doors for further research and, hence, real-time solutions in the fight against global warming threats. Moving ahead, the deployment of extra complicated machine learning models and actual-time surroundings integration can similarly augment our predictive abilities and permit initiative-taking responses to emerging environmental demanding situations.

REFERENCES

[1] Distributed random forest (drf) - h2o 3.46.0.1 documentation. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html.

[2] Global carbon budget. 2018.

[3] The World Bank. World development indicators — databank, 2021.

[4] Harsh Bhatt, Manan Davawala, Tanmay Joshi, Manan Shah, and Ashish Unnarkat. Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques. *Cleaner Chemical Engineering*, 5:100095, 2023.

[5] Michail Fragkias, José Lobo, Deborah Strumsky, and Karen C Seto. Does size matter? scaling of co2 emissions and us urban areas. *Plos one*, 8(6):e64727, 2013.

[6] Abderrachid Hamrani, Arvind Agarwal, Amine Allouhi, and Dwayne McDaniel. Applying machine learning to wire arc additive manufacturing: A systematic data-driven literature review. *Journal of Intelligent Manufacturing*, pages 1–33, 2023.

[7] Surbhi Kumari and Sunil Kumar Singh. Machine learning-based time series models for effective co2 emission prediction in india. *Environmental Science and Pollution Research*, Jul 2022.

[8] Abbas Mardani, Huchang Liao, Mehrbakhsh Nilashi, Melfi Alrasheedi, and Fausto Cavallaro. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production*, 275:122942, 2020.

[9] Jos GJ Olivier, Jeroen AHW Peters, and Greet Janssens-Maenhout. Trends in global co2 emissions. 2012 report. 2012.

[10] Gülden Kaya Uyanık and Neşe Güler. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, 106:234–240, 2013.