# CO2 EMISSION PREDICTION USING MACHINE LEARNING

P Jeevanasai, Shreedhara N Hegde.

*Student, Master of Computer Application, Dayananda Sagar Academy of Technology and Management, Karnataka, India.*
*Assistant Professor, Department of Master of Computer Application, Dayananda Sagar Academy of Technology and Management, Karnataka, India.*

## ABSTRACT

*The use of fossil fuels is increasing at an alarming rate as a result of the ongoing progress of global industrialization and the development of human society. A rate, which causes severe environmental problems, including the greenhouse effect. Carbon dioxide (CO2) is one of the primary gases that cause the greenhouse effect. There is evidence that, during the last 200 years, the average yearly temperature of the Earth's surface has been rising globally. This essay examines and projects carbon dioxide (CO2) emissions. use the Indian dataset from the years 1995 to 2018. The purpose of this study is also to inform readers of the gravity of the present environmental problems. Multiple linear regression is a statistical method used to predict and analyze data. The year, India's population, and its power consumption are regarded as independent factors, whereas CO2 emissions are the dependent variables. The test result produced by the multiple linear regression model was 96.40%.*

**Keyword***: - Machine learning, multiple linear regression.*

## 1. INTRODUCTION

The topic of environmental change concentrates on how climatic examples change over the course of several years or more. It is widely known that people are to blame for environmental change. The Intergovernmental Panel on Climate Change (IPCC) has stated in a few reports that the change in environmental conditions is largely the result of human activities and caused by excessive discharge of substances that harm the ozone layer, such as greenhouse gases (GHGs), after industrialization. [1] Natural and man-made elements both contribute to environmental change. Some examples of human-caused factors for global warming include using non-renewable energy sources (such as oil, coal, and diesel), removing trees from forests, and clearing land for ranches, cities, and roads because all of these activities release greenhouse gases into the atmosphere. Historically, industrialised countries have made up around 75% of the total amount of anthropogenic ozone-depleting emissions due to their industrialization processes and associated production and utilisation patterns. At the start of the Industrial Revolution (from 1850 to the present), substances were introduced into the air. Even though they have large populations today, emerging countries contribute far less to such anthropogenic discharges because of their lower levels of industry. With only 15% of the global population as of now, industrialised countries account for 45% of CO2 emissions [2].

India is emitting carbon dioxide (CO2) at a faster pace than any other big energy consumer nation. According to research by the International Energy Agency, which is located in Paris, CO2 emissions in the country increased 4.8% in 2018 over the previous year. In 2018, India's emissions made up 7% of the global CO2 problem, compared to 14% for the United States [3]. This study uses multiple linear regression to analyse data on carbon dioxide emissions in India from 1995 to 2018 and then forecast those emissions for the years to come based on factors like population and power use.

## 2. LITERATURE SURVEY

The investigation into carbon emissions originally started in 1981, with the major focus being on unstable natural carbon discharges from cooling tower water. From that point on, a few professionals explored the area without experiencing any serious consequences up until 1997, when the convention known as Kyoto was marked. However, following 2007, the study of carbon emissions became a rising issue, leading to a significant increase in research output.

The study by Allen et al. [4] on "warming caused by total carbon emissions towards the trillionth tonne" was identified as the research paper that receives the most citations in the Web of Science research database. The unit, which researchers referred to 632 times, revealed that manmade carbon emissions of one trillion metric tonnes are likely to result in a two-degree Celsius warming of the global climate.

For the City of Beijing from more than 2005 to 2030, Feng et al. [5] developed the Beijing-STELLA Model, a system dynamics-based simulation model, specifically for estimating and forecasting urban vitality consumption patterns and CO2 emissions.

Direct relapse models and traditional least squares methods were used by Murad et al. [6]. The main goal of the essay was to identify and investigate the relationship between environmental change and rural development in Malaysia.

Relapse investigations were used by Kone and Buke [7] to determine energy-related CO2 fluxes. The method of pattern examination was used to demonstrate it. First, trends in CO2 outflows for the top 25 countries and global absolute CO2 outflows from 1971 to 2007 were identified.

Hajji and Lewis [8] developed a tool that may be used to assess the production rate, action span, full fuel usage, and total toxins emanating from earthwork activities by fusing the several direct relapse approaches with the EPA's NONROAD model.

Say and Yucel [9] developed a model using different direct relapse investigations in Turkey that analysed total energy consumption (TEC) in relation to gross national product (GNP) and country population (CP). Sleuth was expected to depend on the 6.7% annual growth in GNP that was the centrepiece of the previous development plan.

Greta Thunberg [10], a 16-year-old Swedish environmental activist, and 15 other youth from around the world documented a human rights protest against five countries in September 2019 for continuing to promote petroleum derivatives and failing to reduce carbon emissions

## 3. METHODOLOGY

The alteration in the environment has had a significant impact on Earth's many regions. This has had a significant impact on the global climate. Environmental change affects the atmosphere. By increasing the amount of greenhouse gases, which cause droughts, we are increasing the temperature. As a result of higher temperatures, hurricanes have also become, in a sense, stronger throughout time. More sweltering temperatures result in hotter water in the oceans. Ocean water gets warmer as the temperature rises. Typhoons are becoming more rare as a result of warmer waters. The shrinking of ice sheets is significantly influenced by environmental changes. The melting of ice raises sea levels, putting the disappearance of several islands in danger. Water quality, quantity, and drinking water treatment are all negatively impacted by environmental change [11].

As the world's top producers of carbon, nations like Russia, India, China, the United States, and Japan have been under tremendous stress as a result of rising carbon outflow. Research on carbon emissions spans more than a few fields, including the common sciences, architecture, financial views, needs, etc. It is obvious that the majority of research on carbon discharge aligns with ideas relating to the atmosphere [12].

## 4. MULTIPLE LINEAR REGRESSION

In this study, multiple linear regression is used to examine how India's population and power usage relate to the nation's CO2 emissions. It is a type of predictive model system that investigates the link between a single dependent variable or target variable and several independent variables or predictors. This method may be used to determine the causal effect link between variables and the time series estimate. The year, India's population, and its energy consumption are regarded as the independent factors (predictors) in this study, whereas CO2 emission is the dependent variable (goal) [13].

To produce predictions, a variety of regression techniques are available. Three metrics (the number of autonomous factors, the kind of ward factors, and the condition of the regression line) largely determine these tactics. In order to determine a direct correlation between the target and at least one indicator, linear regression [14] is used. Multiple. To explain the relationship between one dependent continuous variable and several independent factors—which may be both continuous and categorical—many regression models are used.

The link between one categorical dependent variable and numerous independent factors is explained by multiple linear regression [15]. The independent variables may be continuous or categorical. A line's equation is given by Y = mX+c, where m is the line's slope and c is its y intercept. Y is the dependent variable, and X is the independent variable.

The equation of the line is changed to y = b0 + b1*x1 + b2*x2 +... + bn*xn in the case of multiple linear regression. Here, y stands for the goal or dependent variable, x1, x2,..., xn for the corresponding independent variables or predictors, b0 for the model's intercept, and b1, b2,..., bn for the regression coefficients.

## 5. APPROACH

The International Energy Agency (IEA), an autonomous international organisation with headquarters in Paris, provided statistics for CO2 emissions, population, and power use from 1995 to 2018 [16] to start the process. The information gathered is time series data that was easily accessible and reveals intriguing patterns or shifting trends in the development of the items under examination. The obtained raw data was afterwards transformed into comma-separated values (csv) format. The population is measured in millions, the amount of CO2 emitted is measured in metric tonnes (Mt), and the amount of power consumed is measured in terawatt-hours (TWh).

Jupyter, a Python IDE (integrated development environment), which comes with Anaconda, served as the coding environment. The multiple regression model was constructed using the scikit-learn machine learning toolkit for Python. First, the necessary libraries were imported: Pandas, NumPy, Seaborn, and Matplotlib. The existence of any null values in the data was then verified. There were no null values in the dataset.

The pair plot function from Seaborn (a Python module) was used to plot the data in the next step to determine whether any outliers existed and to determine the distribution of the various relationships between the various variables. The Seaborn Heatmap was then produced, from which we obtained the correlation value, which ranges from -1 to +1. Then the features and labels were extracted; in this case, CO2 emission is the goal variable and is labelled, while year, population, and India's power consumption are the independent/predictor variables and are considered features.

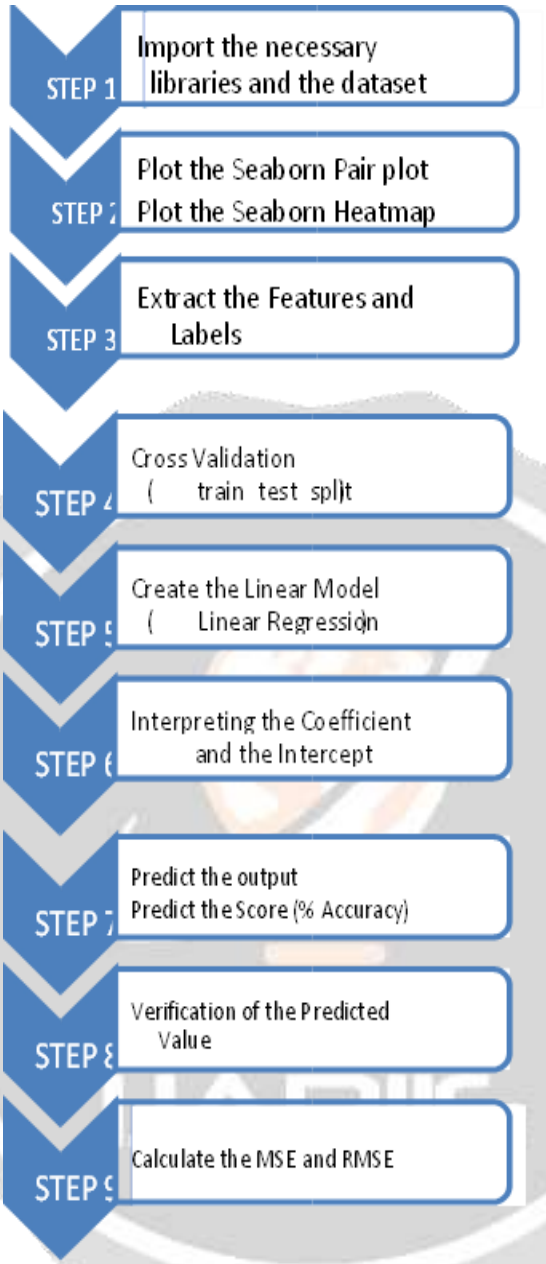


**Fig -1** shows a flowchart of the method used.

Following cross-validation, the data was divided for training and testing. For training and testing, the data were divided into 70% and 30%, respectively. Following the creation of a linear model, the coefficients and intercept were further analyzed using the equation for multiple linear regression. Later, forecast both the production and the accuracy rating. After carefully checking the predicted value, the mean square error (MSE) and root mean square error (RMSE) were calculated to assess the model's performance

## 6. RESULT

The findings demonstrate that both of the variables considered—population and power consumption—contribute to a specific amount of CO2 emissions. CO2 emissions in India will keep rising if they aren't stopped in time.

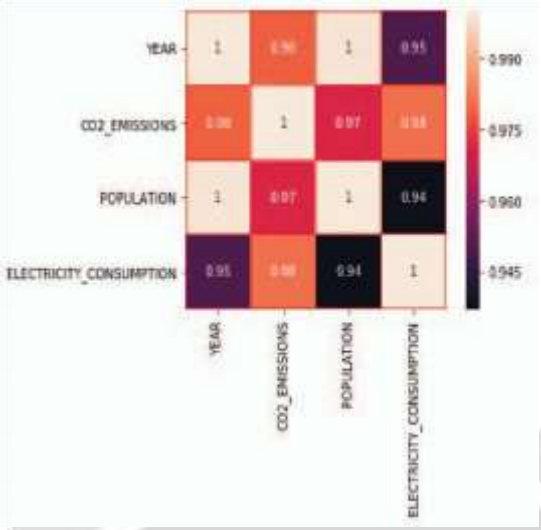


**Fig 2** shows the dataset's Seaborn Heatmap, which shows how the various components are correlated

In the beginning, there were no outliers detected in the graphs that were produced using the Seaborn pair plot to determine if the different variables were positively or negatively connected. A heatmap was created to represent the correlation values in mathematical terms. As can be seen in figure 2, all the values produced were within a narrow range of -1 and +1, suggesting that the distribution and correlation of the various fields under evaluation were appropriate.

After that, features and labels are separated from the dataset. The independent variables are delivered as features while the dependent variable is sent to the system as a label. Here, the year, India's population, and its features, such as consumption, are seen as the independent variables or characteristics, while CO2 emission is thought of as the dependent variable or label. The next step is data sampling. The model was trained using the multiple linear regression equation and the training dataset. Based on the training dataset, the result predicted the trend. For training and testing, the data is split into 16 and 8 rows, which are displayed in the size column.

| Name | Type | Size | |
|---|---|---|---|
| X1 | list | 1 | [[2019, 1370, 1673]] |
| X_test | float64 | (8, 3) | [[2007. 1179.68 640.99]<br>[2008. 1197.15 675.39]] |
| X_train | float64 | (16, 3) | [[2002. 1089.81 448.97]<br>[2010. 1230.98 791.43]] |
| features | float64 | (24, 3) | [[2018. 1352.61 1547.11]<br>[2017. 1338.65 1386.13]] |
| labels | float64 | (24, 1) | [[2299. ]<br>[2193.7 ]] |
| out1 | float64 | (1, 1) | [[2457.74912328]] |
| y_pred | float64 | (8, 1) | [[1325.43040562]<br>[1396.1758875 ]] |
| y_test | float64 | (8, 1) | [[1256.96]<br>[1334.82]] |
| y_train | float64 | (16, 1) | [[ 929.34]<br>[1580.62]] |

**Fig 3** shows the dataset's variable explorer.

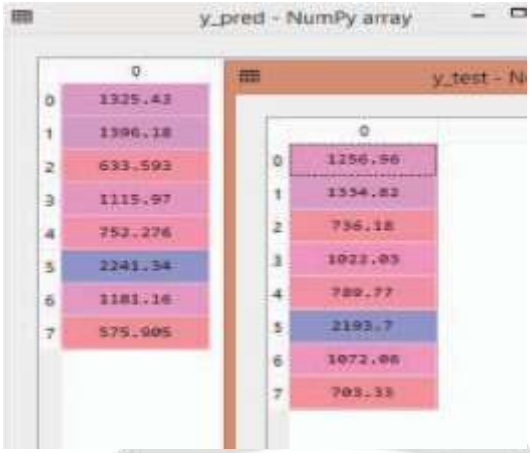


Figure 4 shows the model's predicted values (y_pred) following training and the actual values (y_test).

The train and test data obtained from the sample are then fitted to the linear regression model produced in Python using the scikit-learn module. The model was so trained. The y_pred was then projected using the results from the X_test. Figure 4 illustrates how the y_predicted values were found to be quite close to the y_test values.

The values of the intercept and slope were calculated to get a line that best fits the data using the earlier discussed equation y = b0 + b1*x1 + b2*x2 + b3*x3. Here, b0 is the intercept and b1, b2, and b3 are the regression coefficients based on the target and predictors of the model created. The coefficient came out as 198.5317606, -7.74019951, 0.21613391, and the intercept as -388135.39.

As a result, the equation for each row in the dataset is y_pred = -388135.39+ 198.5317606*x1 + (-7.74019951)*x2 + 0.21613391*x3. The actual values of y are shown to match those anticipated by the model. Here, x1 stands for the year, x2 for the population, and x3 for power usage. Figure 5, which displays the values for x1, x2, and x3, may be used to validate the equation.


**Fig 5** Values that meet the equations in the regression model

The resulting test and train scores were 98.62% and 96.40%, respectively. The RMSE number computed the difference between the actual and anticipated value, whereas the MSE value calculated the average of the square of errors. The RMSE value was 86.32, and the MSE value was 7451.92 as a result. According to the forecast, based on the premise that India's population will be 1370 million and its power consumption will be 1673 TWh, CO2 emissions for the year 2019 will rise to 2475.74 Mt (metric tonne) of CO2.

## 7. CONCLUSIONS

CO2 emissions have mostly had an impact on environmental change. Future measurements and figure layouts that can be used to manage carbon emissions can be created by using information mining to investigate significant data.

Global CO2 emissions are still growing, but they must be reduced before it's too late. As India's population and economy have grown, so have its carbon emissions. Most of the time, CO2 emissions have an impact on how the environment changes. Future measures and figure layouts that can be used to manage carbon emissions can be created by mining remarkable data for significant information. Despite the fact that CO2 emissions are still growing globally, they must be reduced before it is too late. With a growing population and economy, India's carbon emissions have been rising.

## 8. REFERENCES


[1] Margreet Wewerinke and Vicente Paolo Yu II, "Addressing Climate Change Through Sustainable Development and The Promotion Of Human Rights ", South Center, November, 2010

[2] Hervé Le Treut, Richard Somerville, "Historical Overview of Climate Change Science", Cambridge University Press, 2007

[3] Wang, H.; Ang, B.; Su,"B. A Multi-region Structural Decomposition Analysis of Global CO2 Emission
   a. Intensity", 2017

[4] Allen, M.R.; Frame, D.J.; Huntingford, C.; Jones, C.D.; Lowe, J.A.; Meinshausen, M.; Meinshausen, N., "Warming caused by cumulative carbon emissions towards the trillionth tonne", 2009

[5] Feng, Y. Y., S. Q. Chen, L. X. Zhang., "System Dynamics Modeling for Urban Energy Consumption and CO2 Emissions: A Case Study of Beijing, China", Ecological Modelling, 2013

[6] Murad, M. W., and R. I. Molla, M. Mokhtar, M. A. Raquib, "Climate Change and Agricultural Growth: An Examination of the Link in Malaysia", International Journal of Climate Change Strategies and Management, 2010

[7] Kone, A. C., and Buke, T., "Forecasting of CO2 Emissions from Fuel Combustion Using Trend Analysis", Renewable and Sustainable Energy Reviews, 2010

[8] Hajji, A. M., and P. Lewis., "Development of Productivity-based Estimating Tool for Energy and Air Emissions from Earthwork Construction Activities", Smart and Sustainable BuilT Environment, 2013

[9] Say, N. P., M. Yucel, "Energy Consumption and CO2 Emissions in Turkey: Empirical Analysis and Future Projection Based on an Economic Growth", Energy Policy, 2006

[10] Greta Thunberg, Ellen-Anne, Petition, Communication to The Committee on The Rights of The Child, United Nations Convention,2019

[11] Jameel R. Kaddo, "Climate Change: Causes, Effects, and Solutions", Parkland College, 2016

[12] Bill Brath, Todd Friesen, Yves Guerard, "Climate Change and Resource Sustainability an Overview for Actuaries", 2015

[13] Khushbu Kumari, Suniti Yadav, "Linear regression analysis study", University of Delhi, 2018

[14] Anisha Bhatnagar, Namrata Majumdar, Shipra Shukla; "BGP Anomaly Detection using Decision Tree Based Machine Learning Classifiers"; International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019

[15] Namrata Majumdar, Anisha Bhatnagar, Shipra Shukla; "Using Linear Discriminant Analysis for Dimensionality Reduction for Predicting Anomalies of BGP data"; International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019