**PREDICTION AND ANALYSIS OF CO2 EMISSION**

A. DRUVA KUMAR (2103A52121)

G. SRI HARSHINI (2103A52137)

G. SRAVYA (2103A52134)

##### ABSTRACT

Now-a-days emission of the Co2 is the major global issue in the society. Co2 is nothing but a greenhouse gas and the co2 emissions are growing rapidly because of the human activities, combustion of fossil fuels such as coal, oil and the natural gases. These emission of the co2 has much impact on the earth climate change, biodiversity loss, agricultural field, shortage of the water and more dangerous consequences such ultraviolet radiations etc. Here it presents the overview of the recent prediction of the co2 emission. It includes the factors including the historical emission data and environmental factors provided in the dataset.

At last, it concludes that which machine learning algorithm has given best accuracy in predicting the co2 emission and by which factor or factor is more affecting in the emission of the co2.And it also helps in the future study So by this everyone can get the awareness about the emission of the co2 and also its side effects in the society.

##### INTRODUCTION

Emissions of carbon dioxide (CO2) are a significant aspect of our modern society and it has much impact on the environmental concerns on a global scale [1]. The main causes of these emissions are burning of fossil fuels industrial processes, deforestation [3]. A greenhouse gas, CO2 that tries to catch heat in the earth’s environment, which has a huge negative impact on the ongoing problem of climate change [5]. International efforts to reduce climate change and make the transition to more environmentally friendly and sustainable activities now place a high priority on understanding and reducing CO2 emissions [7]. One of the main objectives in promoting a healthier world and ensuring a sustainable future for future generations is the decrease of CO2 emissions [11]. Concerns about carbon dioxide (CO2) emissions are crucial in the context of today's environmental problems. These emissions are mostly caused by people [12].

In this by applying some of the Algorithms for machine learning extremely useful for forecasting CO2 emissions and comprehending their dynamics [2]. ML models can identify difficult patterns, connections by examining large datasets that include historical emissions, climate variables, economic activity, and other aspects [3]. These models can predict future emissions, giving governments, companies, and researchers crucial information they may use to predict emission trends and create mitigation plans and it helps a lot [6]. Additionally, Machine Learning models supports in making knowledgeable decisions by estimating the prospective effects of policy modifications, technical improvements, and sudden changes on the environment [9].

Additionally, it optimizes energy use, improves the effectiveness of transportation, and helps to build climate models for more precise long-term climate projections [12]. Overall, ML's data-driven methodology enables a thorough understanding of CO2 emissions and makes it easier to adopt successful initiative [13].

CO2 emissions are a global concern and must be addressed as soon as possible in order to mitigate its negative impacts on our planet [3]. Machine learning algorithms provide us with a powerful tool to understand, forecast, and reduce CO2 [6]. This innovative technology helps us make better decisions, use resources more efficiently, and set us on the path to a greener and more sustainable future [9]. CO2 emissions have a direct impact on the health and wellbeing of the planet and our children and grandchildren [11].

To be sure, machine learning’s role in CO2 emissions forecasting is crucial and multi-faceted [4]. Machine learning algorithms help us anticipate and manage CO2 emissions better []. These algorithms dive deep into the complex data sets of historical emissions, climate variability, economic cycles, and many contributing factors, uncovering hidden patterns and connections that may otherwise go undetected by human analysis [13]. In short, machine learning is a powerful tool in our collective effort to fight climate change [17]. It helps us anticipate CO2 emissions better, take action, optimize resource use, and work toward a greener future for us and future generations [20].

**Literature Survey**

**Survey Details**

Surbhi Kumari and Sunil Kumar Singh (July 2, 2022) developed an ML model for CO2 forecasting in India. LSTM appears to be the optimal choice with a MAPE of 3.101%, 60.635 RMSE, 28.898 MedAE, reflecting its accuracy and robustness in predicting CO2 emissions [6]. Shanshan Li, Yam wing Siu, Guoqin Zhao (2021) developed a model that can predict co2 emissions. The models used are linear regression, lasso, elastic net, classification, T-regression, Svm regression, k-nearest neighbour, random forest. The KNN parameters are used and their values ​​are rmse, i.e. 0.1750 and 0.3641 for the two data when zero. there are two neighbours [7]. Yang Meng and Hossain Noman (November 9, 2022) used artificial intelligence to develop the model through machine learning. The research methodology includes applying AI and machine learning techniques, especially SARIMA models, to know CO2 emissions. It exhibits higher accuracy with a MAPE of 0.09, allowing a robust model to predict co2 [8]. Md. Omer Faruque (November 2022) developed a hybrid approach to machine learning model development. Forecasting co2 emissions using deep learning and regression algorithms, with the DNN model giving the highest accuracy [9]. Abbas Mardani (December 1, 2020) developed a stepwise method to predict co2 emissions using dimensionality reduction, clustering and machine learning. algorithm. The method includes many steps such as ANFIS, ANN, SOM and SVD. Notably, the combined technique achieved an impressive accuracy of 0.065 MAE [10]. Shreeya Namboori (December 12, 2019) developed a forecast of carbon dioxide emissions in the United States using machine learning. ARIMA, SVM, SVM-PSO and Prophet models for prediction. The results indicate that the prediction performance of the Prophet model is the best. [11]. Ümit Ağbulut (January 2022) developed forecasts of energy demand and transportation-related CO2 emissions in Turkey using various machine learning algorithms. Artificial neural networks, deep learning and support vector machines are the three machine learning algorithms. The parameters used for model robustness and accuracy are R2, Rmse, Mape, Mbe. [12Omer Saud Azeez, Biswajit Pradhan, Helmi Z.M. Shafri (September 26, 2018) developed prediction of vehicle CO emissions using support vector regression model and GIS. The hybrid model integrates three separate components CFS, SVR, and geographic information systems (GIS). Achieved 80.6% validation accuracy, high correlation coefficient of 0.9734, and minimal prediction error [13]. Chairul Saleh et al (2016) have developed co2 emission prediction using svm. The experimental results are SVM model with optimized parameters achieved a minimal RMSE of 0.004 [14]. Majid Emami Javanmard (December 2017) developed Co2 emissions forecasting using regression algorithms. This optimization model increased the prediction accuracy compared to the machine learning algorithms applied in the project The least improvement is 31.7 percent with the Pso algorithm and 12.8 percent with the Gwo algorithm. [15]. Ren Cheng Zhu, (R.Z) (2016) developed prediction of Co2 emissions using deep learning. The R2 between the CO2 emission values ​​of Lddt1, Lddt2 predicted by DL-Dtcem and the value monitored by Pems are 0.986 and 0.990, consequnetly. [16] Professor Dr. Xiaodong Li (November 2019) developed the exploration of Co2 emission models using ML modeling. This paper used three methods to predict transportation-related CO2 emissions using three ML methods - Ols, Svm and Gbr. Three kinds of features, Tran (traffic-related features only), Soeco (socio-economic features only) and All (combination of TRAN and So Eco features) [17]. Arva Arsiwala (May 7, 2019) has developed a digital twin with ML technology to forecast and cover the CO2 equivalent of existing buildings. This study uses a dts solution to monitor and control carbon dioxide equivalent (eCO2) emissions from IoT, Bim, AI through a comprehensive solution, further validating its feasibility through real-life utilize case breakdown [18]. Nishant Raj Kapoor (July 7, 2018) developed machine learning-based CO2 forecasting for offices. The working room is evaluated using Ann, Gpr, Dt, El, Svm, Lr algorithms. ML Models,The accuracy rates of ML models Rmse, Mae, Ns and a20 index values ​​are 0.988, 4.200ppm, 3.35ppm, 0.9817, respectively. and 1. [19]. Shukhi Tan, Cheng Zhang, Milan Saha (2019). Has developed the reduction of carbon emissions from building energy consumption through machine learning. Based on the principle of downscaling, we proposed a new gridded BECCE mapping method that integrates remote sensing and socioeconomic data into flexible PLS and Cubist regression models, built to captures the complex relationships between covariates (e.g., GDP, POP, GST, HDD18 and CDD26) and BECCE intensity at the provincial and corresponding levels [20]. In the below table-1 the summary is represented in the form of tabled data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AUTHORS &**  **YEARS** | **MODEL USED** | **PARAMETERS USED** | **MERITS** | **DEMERITS** | **FUTURE SCOPE** |
| Tomio Miwa 2010[1] | Cmem, Vsp | Maximum Likelihood Estimation | Comparing CO2 emission models | Unlike traditional networks where only a unique attribute is allotted to each link Rephrase | Eco-routing method to access the problem of searching the most ecofriendly way in terms of min Co2 releasing conditioned by a travel time budget. |
| Chairul Saleh et al 2016[2] | SVM | RMSE | Optimal Parameter Search, Decision-making | Model Complexity, Hardware and Software Limitations, Data Scope and Generalizability, | parameters of Svm can instantaneously chose by collaborating optimization techniques such as genetic algo |
| Caifei Wang Wei Sun  2017[3] | PLO-ELM  Particle Swarm  Optimization - Extreme Learning machine, | Activation function,  Convergence  criteria, learning rates. | Comparison of Models,  Forecasting, Factor  Analysis, Machine  Learning Techniques | Data Quality, Model Complexity, Bias and Generalization. | refine modelling techniques, explore different algorithms  Expand the analysis to consider multiple influencing factors for more accurate and comprehensive predictions. |
| Biswajit Pradhan 2018[4] | Machine Learning | Architecture of the model, the choice of algorithms, hyperparameter  settings | Real-world Application  Validation,  Machine Learning  Techniques, Emission | Generalization,  Interpretability, Data  Quality Model, Complexity | Explore different algorithms  ·Expand the analysis to consider various influencing factors for more accurate |
| Z.M. Shafri  2018[5] | hybrid model by integrating three distinct components: correlation-based feature selection (CFS), support vector | RMSE, correlation, heuristic ‘merit’ | Hybrid Model, Flexibility and Adaptability, Decision Support, Applicability  · Cost-Efficiency, Relevance-Driven Selection, Spatial | Data Complexity, Spatial Constraints, Model Complexity, Parameter Sensitivity, Limited Predictive Scope, | · The integration of deep learning algorithms could enhance prediction accuracy and capture intricate emission patterns  ·While further advancements in |
| Shreeya Namboori  2019[6] | Arima, Svm, Svm-PSO, and Prophet | RMSE, MAPE, and MAE | Handles seasonality, Well-suited for stationary data, Simple and interpretable. | Limited flexibility, Sensitive to outliers, may require differencing | CNN-LSTM (Convolution Neural Network Long Short-Term Memory), LSTM (Long Short-Term Memory), (GM) can be used for the time series forecasting. |
| Abbas Mardani  2020[7] | ANN, SVD | Activation function, Loss function, Grid size, Neighbourhood functions | Insights for Policy-Making, Accurate Predictions, Multi-Stage Approach | Computational Resources, Data Sensitivity, Complexity | · Can use Dynamic Models, Ensemble Methods, Long-Term Predictions  · Incorporating External Factors and Enhanced Clustering Algorithms |
| John Kenya  2021[9] | ARIMA, SARIMAX, linear regression | MSE, RMSE, MSE, MSLE, MAE, MAPE | · Deep learning model for complex time-series patterns  · Captures long-range dependencies and intricate trends | · Requires data to be stationary; struggles with non-linearity  · Limited to linear relationships, may not capture complexities. | integrating external factors like economic indicators and policy changes for more accurate CO2 emission predictions. |
| Guoqin Zhao  2021[10] | Linear Regression, Lasso, Elastic Net, Tree, Support Vector Regression, k-Nearest Neighbours Regression, Networks. | RMSE | Used all types of models and selected the model with least rmse. | Alternate models like LSTM (Long Short-Term Memory), singular value decomposition | The future scope could be using more deep learning. making the model with robust and best fit. |
| Melahat Sevgül Bakay 2021[11] | [deep learning](https://www.sciencedirect.com/topics/engineering/deep-learning) , SVM | Rmse, mae | R2 value for the emissions between 0.861 and 0.998 | Data Scarcity, quality,  overfitting,  complexity. | these data can be used to forecasted for the following four years |
| Dr. Alina Cristina Nuta 2021[11] | XGBoost, and AdaBoost, Random Forest | RMSE, MSE | modelling the city and the affect of energy intensity on the Co2 | · After sampling the best fitted features using the machine learning algorithms, we developed explanatory linear regression models. | The non-renewable energy influence Carbon dioxide. |
| Maomao Zhang  2021[12] | PLS, Cubist Regression models. | No of talent variables for PLS, Tree complexity for Cubist. | Sustainability Insights,  Complex Relationships,  Downscaling  Comprehensive Mapping | Generalization, Data  Limitations, Model Complexity, Data Quality. | Refine the modelling techniques, incorporate additional data sources  Expand the analysis to provide more comprehensive insights |
| Prof. Dr. Florian Nuta  2021[13] | Linear  Regression,  Decision Tree, support vector machine. | Depth of the decision tree, no of estimators of the random forests. | Potential Linear Models,  Consistency, Ensemble  Methods,  Comprehensive Analysis. | Model Selection, Data  Quality,  Interpretability, Model  Complexity | Ensemble methods to analyses the impact of urbanization on  Carbondioxide emissions provides valuable insights. |
| Umit Ablute  2022[14] | Deep Learning (DL), SVM, ANN | Activation function, Loss function, kernel, max iterations. | Accurate Predictions, Robust Comparison, High Prediction Accuracy, Low Error Rates | Complexity, Data Dependency | Extend the analysis beyond 2050 to provide insights into the long-term implications of transportation-related energy. |
| Yang meng  2022[15] | SARIMA | AR order, differencing order, MA order, Seasonal AR order. | Limited External Factors, Reliance on Historical Data, Uncertainty | In tree-based classifier,  data standardization cannot provide any guarantee to improve the performance. | · Exploring the integration of additional external factors that influence CO2 emissions, enhancing the model's accuracy and predictive capabilities. |
| Md. Alamgir Hossain  2022[16] | convolutional neural networks, dense neural network, long short-term memory, Fully Modified Ordinary Least Squares | MAPE, MAE, RMSE | · Suitable for capturing complex and nonlinear relationship  · Can handle multivariate time series | Huge amount of data is needed for training and tuning.  · Interpretability can be challenging | · Incorporating more advanced techniques  · Autoregressive Distributed Lag (ARDL) models kernel could provide deeper insights |
| Ali Almusalam2022[19] | [deep learning](https://www.sciencedirect.com/topics/engineering/deep-learning) (DL), [ANN](https://www.sciencedirect.com/topics/engineering/artificial-neural-network), SVM | Activation function, Loss function, kernel | The model is best for ANN and SVM respectively. | The paper might not address potential ethical issues related to using government data | It compares predicted data with the actual one to test the efficiency of the models that were used. |
| Prof. Mohammad Abdelkareem2022[22] | FFNN, ANFIS and LSTM | MSE, RMSE | · Calculated percentage error was found to be 6.8675% with an accuracy of 93.1325 percent. | · After sampling the best fitted features using the machine learning algorithms, we developed explanatory linear regression models. | Clearly, the non-renewable sources of energy influence CO2, but also relate to the urbanization processes. |
| F. Ghaderi2022[26] | Artificial  Neural Network  (ANN),  Stepwise  Regression  Algorithm. | Inclusion and  Exclusion Criteria,  Significance  Level, Model Evaluation Criterion. | Feature Selection,  Simplicity, Efficiency, Variable Ranking. | Overfitting,  Assumptions, Lack of Explorations, Model Complexity. | Expanding and refining the predictive model, considering different factors  · Additionally, exploring real-world implications and collaborating across disciplines. |
| Menglei Wang 2022[27] | Deep Learning  Dynamic  Tailpipe  (DL-DYCEM) | Neural Network  Architecture, Activation function, Learning Rate. | Complex Pattern  Recognition, Feature Learning, Flexibility, Generalization. | Model Complexity,  Data Preprocessing, Data Requirements. | To predict other tailpipe pollutants |
| Helmi Z.M. Shafri 2018[28] | CFS, SVR, and GIS | RMSE | Resarchers Demonstrated potential for real-world prediction. | Overfitting,  Assumptions, Lack of Explorations, Model Complexity. . | It is basically on the combination of data sampling techniques and spatial interpolation methods |
| Prof. Dr. Ai Ren  2022[29] | OLS, SVM | Kernel Type,  Learning Rate, Loss Function. | Predictive Power, Feature Importance, Ensemble Learning | Data Quality,  Overfitting,  Hyperparameter  Tuning, Model Complexity | Exploring the external factors that influence CO2 emissions  · Enhancing the model's accuracy and predictive |
| **Sangsoon Lim**  2022[29] | Artificial  Neural Network  (ANN), Linear Regression,  Decision Tree. | Network  architecture for  ANN,  hyperparameters for GPR and SVM, tree depth for DT | Multi-Algorithm  Comparison,  Optimized Algorithms Real-Time Data, Data Preprocessing. | Generalization, Data Quality, Hyper parameter Tuning,  Algorithm Complexity. | · The future scope of the research is  Integration with Building Systems Long-  Term Monitoring, Ensemble Methods, Feature Engineering |
| [Anand Paul](https://sciprofiles.com/profile/100461)  2023[30] | Ensemble methods using weak learners and boosting, bagging, boost. | RMSE, MAE, MSE | Many boosting and regression models for CO2 emission | ·Model Complexity, while boosting and regression models like Cat boost can offer | Clearly, the non-renewable sources of energy influence CO2, but also relate to the urbanization processes. |
| Dr. Theyazn H.H Aldhyani  2023[30] | Ensemble | RMSE, MSE | This model can make predictions on CO2 emissions in a single dimension. | Data Scarcity and quality,  overfitting,  complexity,  Interpretability | · We are not using methods svm,deep learning |
| Mohammed Zoher  2023[30] | IoT Sensor Optimization, BIM. | MSE, RMSE,  MSE, MSLE,  MAE, MAPE | Environmental Impact,  Digital twin Integration. | Data Quality, Model Complexity, Bias. | Potential to revolutionize the monitoring and control of CO2 |

**Table-1**

**Problem Statement:**

The main problem is to Predict the amount of co2 emissions.so that if we could identify then we can easily find the solution that how to control amount of co2 .Develop a system to predict, analyze CO2 emissions, focusing on understanding, predicting and also minimizing their environmental impact.

This involves developing accurate prediction models based on historical data and assessing impacts on industrial, transport and energy sources. Deploy real-time monitoring using IoT and remote sensing, conduct geographic and industry analysis, evaluate the effectiveness of policies, and recommend strategies to reduce carbon emissions. Additionally, evaluate the economic and environmental consequences of CO2 emissions and raise public awareness through user-friendly interfaces. This initiative will promote informed decisions, sustainable practices and efforts to mitigate global climate change.

**Objectives**

* In this paper we are going to know how the co2 is going to be increased in the following years by using machine learning algorithms.
* Ensuring the model performs good on unseen data. Implement cross-validation and rigorous evaluation techniques to assess the model's generalization. Regularize the model to prevent overfitting.
* Lack of interpretability in complex models may hinder policy decision-making. Prioritize model transparency and interpretability, using techniques like feature importance analysis or model visualization to explain predictions.

**Proposed Methodology Architecture**.

VZ.

Data Collection

Regularization

Hyper parameter tuning

Data Preprocessing

Feature Selection

Models Used (linear regression, Support vector regression, Decision tree regression, Ridge, Lasso regression.

Predicting the output using Model

Model Selection

Model Evaluation (RMSE, Accuracy, MSE, MAE)

Cross Validation

**Fig-1**

From the above fig-1 we can know how our model is working

**Analysing Data set**

Before starting the project, we need to understand the data contained in our dataset. Our dataset consists of 12 columns and 1584 rows and has no null values ​​because all null values ​​are represented as “..”. and except for year, each column is a data type object.

**Data Collection and Preprocessing**

We collected our data from Kaggle.

First we need to check if there are any duplicate values ​​and if there are then we need to remove those duplicate values ​​(.drop\_duplicates()).

In this data set, null values ​​are represented by ".." but the system recognizes these values ​​as string values ​​and does not treat these value types as null values. Our main job is to replace all “..” values ​​with pandas.NA. The NaN value should be replaced by the mean, mode or median, otherwise if there are more NaN values ​​in a row then it is better to remove that row.

The total number of rows with missing values ​​is 534 rows, and the total number of missing values ​​in the data set is 920.

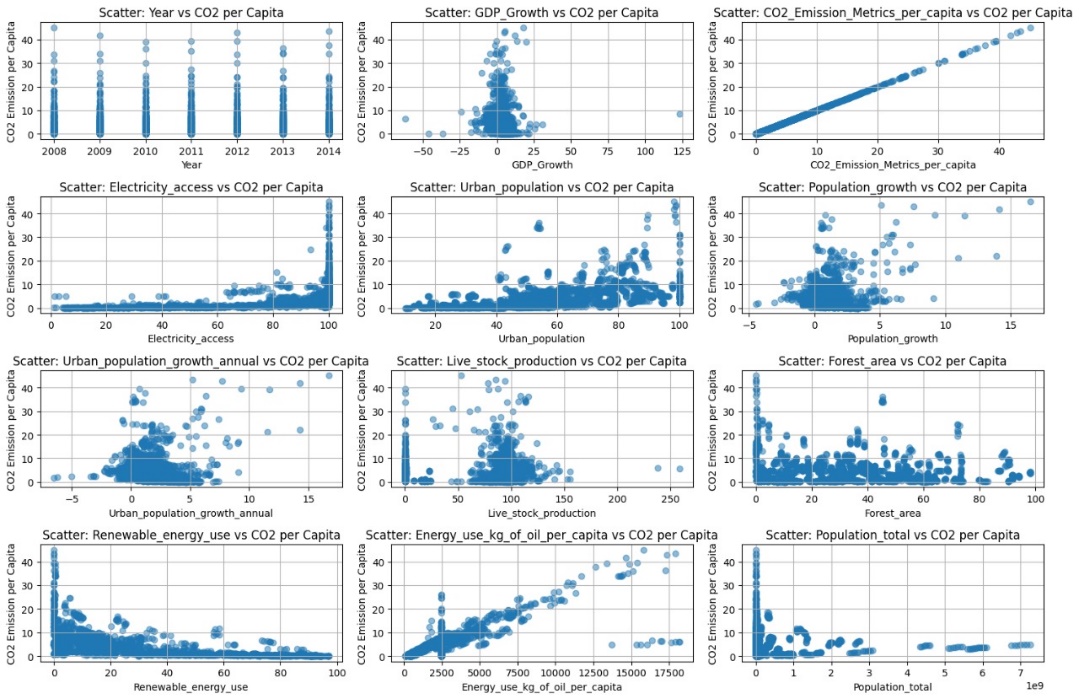
Note:

Since we have two different datasets for training and testing. so we need to apply all the preprocessing techniques on both the datasets.

**Data Visualization**

Data visualization is an important part of the machine learning (ML) process. It helps data scientists and analysts discover, understand, and communicate patterns, relationships, and insights in data.

The choice of visualization procedure depends on the sort of information and the particular objectives of your ML venture. Interactive visualization libraries such as Matplotlib, Seaborn, Plotly, and libraries for domain-specific visualizations (e.g., Geopandas for geospatial data) are commonly used in ML to create visualizations. Visual images are informative and insightful. Below fig-2 shows the relation between co2\_emission\_per\_captia and all the features.



**Fig-2**

**Data Splitting**

The data needs to be split into training and test datasets, but here we have two different CSV files, so they need to be split. For evaluation, we can apply it directly on the test data set.

**Models Used**

Since the dataset consists of outputs and outputs can range from -infinity to +infinity, different regression techniques have been applied such as linear regression, decision tree regression, support vector regression, ridge regression and lasso regression. The evaluation is performed using these models on the dataset**.**

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 observed data. It is one of the simplest and most widely used techniques in statistical modelling and machine learning to predict continuous outcomes.

yi=β0+β1x1+β2x2+...+βxi+ϵ

Support Vector Machine:

Support vector machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks.It is especially suitable for classification problems where the goal is to separate data points into different classes based on their characteristics.SVM is effective in both linear and non-linear classification tasks and has become popular in various fields such as image classification, text classification, and bioinformatics.

Minimize: 0.5 \* ||w||^2 + C \* Σ[max(0, |yi - f(xi)| - ε)]

Subject to: ∀i, f(xi) - ε ≤ yi ≤ f(xi) + ε

Decision Tree:

In a decision tree, like a diagram, an internal node represents a variable (or characteristic) of the data set, tree branches represent decision rules, and each leaf node represents the outcome of a particular decision.The first node from the top of the decision tree is the root node. We can divide data based on attribute values ​​corresponding to independent characteristics.

y = Σ(yi) / N

Ridge Regression**:**

Ridge regression is a model fitting method used to analyze any data that exhibits multicollinearity. This method performs L2 regularization.Ridge regression is a technique used for multiple regression analysis where the data exhibits multicollinearity.The problem that arises due to multicollinearity is that the basic linear regression model (least squares estimation) becomes unbiased and the variance becomes so large that the predicted values ​​are far from the actual values.

minimize: ||y - Xβ||^2 + α||β||^2

Lasso Regression:

Lasso means “Least absolute selection and elimination operator”. This is a type of linear regression technique used for both regression and feature selection. Lasso adds a regularization term to the linear regression model to prevent overfitting and encourage feature selection. Lasso introduces a regularization term into the linear regression equation, which is added to the ordinary least squares (OLS) cost function.

minimize: ||y - Xβ||^2 + α||β||

**Model Training**

Model training means we train the models that we choose. It tries to get optimal values ​​for variance and bias. At this stage, it understands all the features of the model and captures the hidden structure of the given data. Model training is mainly done to obtain a mathematical function based on the given data, and this function is used to take input and give output, that is, it makes predictions using the function This..

**Model Evaluation**

Model evaluation is performed to know the predictive ability of the model. It can be used to know the performance of the model. This is done on the test dataset. The metrics used are precision, recall, F1 score, area under the curve, confusion matrix, and mean square error. It performs error analysis to identify common errors and areas for improvement**.**

**Model Selection**

Model selection is made based on comparison of metrics and the best model is used to predict the outcome.

**EXPERIMENTAL WORK**

**Dataset Description:**

The dataset for the co2 emission cause prediction is taken from the Kaggle. This dataset is used to predict the co2 emissions in the atmosphere and also used to know which factors affecting most for the increase in carbon dioxide in nature and which factors effecting the least. From the table 2 we can understand the details about the data set.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute name** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Year** | 1848.000000 | 2011.000000 | 2.000541 | 2008.000000 | 2009.000000 | 2011.000000 | 2013.000000 | 2014.000000 |
| **CO2\_Emission\_Metrics\_per\_capita** | 1848.000000 | 4.868034 | 5.777943 | 0.000000 | 0.900000 | 3.500000 | 6.500000 | 45.100000 |
| **Electricity\_access** | 1848 | 80.714773 | 28.171213 | 1.200000 | 68.025000 | 98.000000 | 100.000000 | 100.000000 |
| **Urban\_population** | 1848.000000 | 57. 798647 | 22.955891 | 10.100000 | 39.000000 | 57.000000 | 76.725000 | 100.000000 |
| **Population\_growth** | 1848.000000 | 1.396953 | 1.481911 | -4.500000 | 0.500000 | 1.200000 | 2.300000 | 16.500000 |
| **Urban\_population\_growth\_annual** | 1848.000000 | 2.079654 | 1.869745 | -6.500000 | 0.700000 | 1.900000 | 3.200000 | 16.700000 |
| **Live\_stock\_production** | 1848.000000 | 72.896591 | 41.242117 | 0.100000 | 66.425000 | 92.100000 | 98.700000 | 258.200000 |
| **Forest\_area** | 1848.000000 | 32.239123 | 22.885908 | 0.000000 | 12.600000 | 30.600000 | 46.325000 | 98.100000 |
| **Renewable\_energy\_use** | 1848.000000 | 29.425866 | 28.332251 | 0.000000 | 5.700000 | 19.100000 | 47.850000 | 97.000000 |
| **GDP\_Growth** | 1848.000000 | 3.195886 | 5.519357 | -62.100000 | 1.200000 | 3.197943 | 5.600000 | 123.100000 |
| **Population\_total** | 1.848000e+03 | 2.818474e+08 | 8.903332e+08 | 9.891000e+03 | 1.418084e+06 | 9.526260e+06 | 5.930292e+07 | 7.254227e+09 |
| **Energy\_use\_kg\_of\_oil\_per\_capita** | 1848.000000 | 2468.622231 | 2344.184563 | 63.700000 | 1003.775000 | 2468.622231 | 2468.622231 | 18178.100000 |

**Table-2**

**Performance parameters with Formulas**:

**Mean square error (MSE):**

MSE defined as the average of the squared differences between the predicted value and the actual value. This gives higher weight to larger errors.

Formula:

MSE = (1/n) \* Σ(predicted value - actual value)^2

**Mean absolute error (MAE):**

MAE defined as the average of the absolute difference between the predicted value and the actual value.

It treats all errors equally and is less sensitive to outliers than MSE.

Formula:

MAE = (1/n) \* Σ(predicted value - actual value).

**Root mean square error (RMSE):**

RMSE is almost similar to MSE but takes the square root of the mean square of the difference between the predicted value and to the actual value.RMSE has the same units as the target variable, making it easier to understand than MSE.

Formula:

RMSE = sqrt((1/n) \* Σ(predicted value - actual value)^2)

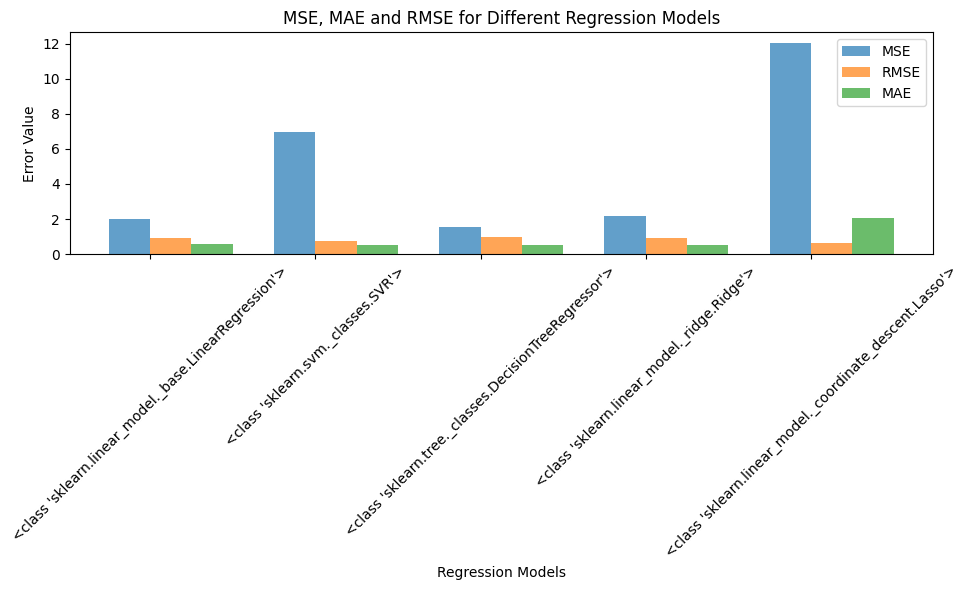
**RESULTS:**

TABLE:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Used** | **Mean Squared Error (MSE)** | **Root Mean Squared Error (RMSE)** | **Mean Absolute Error(MAE)** |
| Linear Regression model | 1.2223666689846053 | 0.9611421599517289 | 0.5670951087798872 |
| Support Vector Regression model | 2.7521816753284516 | 0.9125108381656613 | 0.5076464632939135 |
| Decision Tree Regression model | 0.7956433853579152 | 0.9747072754941978 | 0.47799126799126795 |
| Ridge Regression model | 1.2269738201209617 | 0.9609957031262304 | 0.5502026291791738 |
| Lasso Regression model | 12.136199920723524 | 0.614202082501929 | 0.6156173307052011 |

**Table-3**

GRAPH:

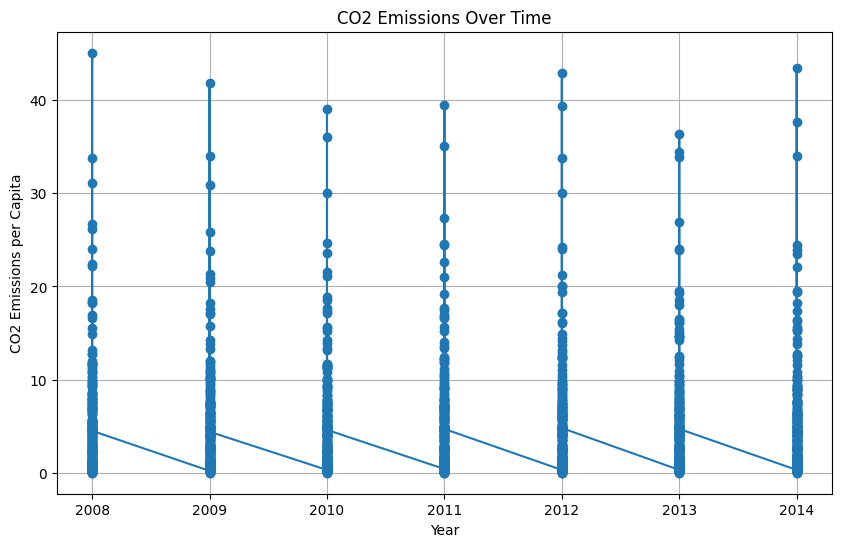
****

**Fig-3**

**Analysis about models:**

* Decision tree regression has the lowest MSE but slightly higher RMSE than linear regression.
* Linear regression and Ridge regression perform similarly with likelihood MSE and RMSE.
* Support vector regression maintains reasonable RMSE despite higher MSE.
* Lasso regression model has the highest MSE but lowest MAE, giving the best explanation of variance
* The linear regression model demonstrated moderate accuracy with a mean squared error (MSE) of 1.2224 and a mean absolute error (MAE) of 0.9611.
* Model selection must be appropriate to the specific analytical goals, taking into account interpretability, computational complexity, and trade-offs between bias and variance.

**Analysis**



**Fig-4**

The line graph(fig-4) is between years and co2 emissions, for any particular country the co2 emission is increasing year to year

In every year ‘Qatar’ country is the most polluted country because in average it releases 38 co2 emission per captia, and the resources used by it is more.

In every year ‘Burundi’ country is the least polluted country because on average it releases 0.01 co2 emission per captia, and the resources used by it is less.

|  |  |  |
| --- | --- | --- |
| Year | More Co2 Emitted Country (With percentage) | Less Co2 Emitted Country (With percentage) |
| 2008 | Qatar (45.1) | Burundi (0) |
| 2009 | Qatar (41.8) | Burundi (0) |
| 2010 | Qatar (39.1) | Burundi (0) |
| 2011 | Qatar (39.5) | Burundi (0) |
| 2012 | Qatar (42.9) | Burundi (0) |
| 2013 | Qatar (36.4) | Burundi (0) |
| 2014 | Qatar (43.5) | Burundi (0) |

**Table-4**

**CONCLUSION**

At finally, undertstanding and also predicting the co2 emissions in the environment is very crucial to fight against the climate change. This report given some brief information about co2 is emitting present and where it comes, how it is effecting our planet .We can also observe which country is most effecting with much release of co2 emissions and how it is effecting Global warming.

And also studied how we predict what will happen in future if it continues, What we can do to make better for the environment and for our future too. It is important to know about co2 emissions to face climate change effectively.

As we predicted the emissions of co2 we should look into the factors effecting the climate and should see and alternative. For example, If releasing of co2 from vehicles is more then we should try to decrease the using of vehicles as much as possibles instead go with electric vehicles or else use cycles for shorter distance which is also an exercise for our body.So,in this way we should see an alternatives to decrease the carbon dioxide.

**Limitations:**

* Developing comprehensive CO2 emissions requires more amount of resources
* To need the co2 amount we need the country name if it is of string type we need to use one hot encoding which makes our model complex

**FUTURE SCOPE**

We must create some cool devices and methods to decrease the CO2 emissions from things like automobiles, industries, and farms.

By using cutting-edge technology like artificial intelligence and more data, we can increase the accuracy of our projections in predicting CO2 emissions.We can build some new devices which can absorb CO2 from the atmosphere, something like a sky-sized vacuum.

There are a lot of opportunities for CO2 emissions research in the future. We can make our planet cleaner and safer for everyone, but it will take a lot of cooperation and imagination.

**Together, we can make a clean and safe environment**.

**REFERENCES**

# [1] Zeng, W., Miwa, T., & Morikawa, T. (2016). Prediction of vehicle CO2 emission and its application to eco-routing navigation. *Transportation Research Part C: Emerging Technologies*, *68*, 194-214.

[2] Azeez, O. S., Pradhan, B., & Shafri, H. Z. (2018). Vehicular CO emission prediction using support vector regression model and GIS. *Sustainability*, *10*(10), 3434.

[3] Nassef, A. M., Olabi, A. G., Rezk, H., & Abdelkareem, M. A. (2023). Application of Artificial Intelligence to Predict CO2 Emissions: Critical Step towards Sustainable Environment. *Sustainability*, *15*(9), 7648.

[4] Nuta, F. M., Nuta, A. C., & Zamfir, C. G. (2021). National carbon accounting—analyzing the impact of urbanization and energy-related factors upon CO2 emissions learning algorithms and panel Data analysis. *Energies*, *14*, 2775.

[5] Al-Nefaie, A. H., & Aldhyani, T. H. (2023). Predicting CO2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model. *Sustainability*, *15*(9), 7615.

# [6] Kumari, S., & Singh, S. K. (2022). Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research*, 1-16.

[7] Li, S., Siu, Y. W., & Zhao, G. (2021). Driving factors of CO2 emissions: further study based on machine learning. *Frontiers in Environmental Science*, *9*, 721517.

[8] Meng, Y., & Noman, H. (2022). Predicting CO2 Emission Footprint Using AI through Machine Learning. *Atmosphere*, *13*(11), 1871.

[9] Faruque, M. O., Rabby, M. A. J., Hossain, M. A., Islam, M. R., Rashid, M. M. U., & Muyeen, S. M. (2022). A comparative analysis to forecast carbon dioxide emissions. *Energy Reports*, *8*, 8046-8060.

# [10] Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M., & Cavallaro, F. (2020). A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production*, *275*, 122942.

[11] Namboori, S. (2020). *Forecasting carbon dioxide emissions in the United States using machine learning* (Doctoral dissertation, Dublin, National College of Ireland).

# [12] Ağbulut, Ü. (2022). Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. *Sustainable Production and Consumption*, *29*, 141-157.

[13] Azeez, O. S., Pradhan, B., & Shafri, H. Z. (2018). Vehicular CO emission prediction using support vector regression model and GIS. *Sustainability*, *10*(10), 3434.

[14] Saleh, C., Dzakiyullah, N. R., & Nugroho, J. B. (2016, February). Carbon dioxide emission prediction using support vector machine. In *IOP Conference Series: Materials Science and Engineering* (Vol. 114, No. 1, p. 012148). IOP Publishing.

[15] Majid Emami Javanmard, S.F. Ghaderi ((25-08-2023)) A Hybrid Model with Applying Machine Learning Algorithms and Optimization Model to Forecast Greenhouse Gas Emissions with Energy Market Data

[16] Javanmard, M. E., & Ghaderi, S. F. (2022). A hybrid model with applying machine learning algorithms and optimization model to forecast greenhouse gas emissions with energy market data. *Sustainable Cities and Society*, *82*, 103886.

[17] Li, X., Ren, A., & Li, Q. (2022). Exploring patterns of transportation-related CO2 emissions using machine learning methods. *Sustainability*, *14*(8), 4588.

[18] Arsiwala, A., Elghaish, F., & Zoher, M. (2023). Digital twin with Machine learning for predictive monitoring of CO2 equivalent from existing buildings. *Energy and Buildings*, *284*, 112851.

[19] Kapoor, N. R., Kumar, A., Kumar, A., Kumar, A., Mohammed, M. A., Kumar, K., ... & Lim, S. (2022). Machine learning-based CO2 prediction for office room: A pilot study. *Wireless Communications and Mobile Computing*, *2022*, 1-16.

# [20] Zhao, Z., Yang, X., Yan, H., Huang, Y., Zhang, G., Lin, T., & Ye, H. (2021). Downscaling building energy consumption carbon emissions by machine learning. *Remote Sensing*, *13*(21), 4346.

# [21] Diebold F., Rudebusch G. Long memory and persistence in aggregate output. *J. Monetary Econ* 19-20

# [22] Granger C.W.J. Long memory relationships and the aggregation of dynamic models. *J.* Economy [5-13]

# [23] Hassler U., Rodrigues P., Rubia A. Quantile regression for long memory testing: a case of realized volatility. *J.* Financ.Econom. [8-10]

# [24] Andrews D., Ploberger W. Optimal tests when a nuisance parameter is present only under the alternative. Econometrica [11-15]

# [25] Barassi M., Cole M., Elliott R. The stochastic convergence of CO2Emissions: ALong memory approach. *Environ.* Resour. Econ. [25-30]

# [26] Apergis N., Tsoumas C. Long memory and disaggregated energy consumption: evidence from fossil fuels, coal and electricity retail in the US. Energy Econ. [22-24]

# [27] aillie R. Long-memory processes and fractional integration in econometrics. J. Econom. [19-22]

# [28] Shou Hsing Shih and Chris P. Tsokos, "Prediction models for carbon dioxide emissions and the atmosphere", Neural Parallel Sci. Comput[*12-16]*

# [29] Sandeep Kumar and Pranab K. Muhuri, "A novel GDP prediction technique based on transfer learning using CO2 emission dataset"[14-18]

[30] Samira Yeasmin, Sonia Noor Jahan Syed, Leena Abu Shmais and Aghad Al Dubayyan, "Artificial Intelligence-based CO2 Emission Predictive Analysis System [15-25]