2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 979-8-3503-5035-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRITO61523.2024.10522152

Smart Forecasting: Harnessing Machine Learning for Accurate CO2 Emission Predictions

A. Hency Juliet
Department of Computer Applications
Saveetha College of Liberal Arts and Sciences
Saveetha Institute of Medical and Technical Sciences
Chennai – 602105, India
hencyjulieta.sclas@saveetha.com

P.Malathi
Department of Computer Applications
Saveetha College of Liberal Arts and Sciences
Saveetha Institute of Medical and Technical Sciences
Chennai – 602105, India
malathip113@gmail.com

N. Legapriyadharshini
Department of Computer Applications
Saveetha College of Liberal Arts and Sciences
Saveetha Institute of Medical and Technical Sciences
Chennai – 602105, India
sakthileha@gmail.com

***Abstract*—This The tenacity of this study is to improve a Machine Learning (ML) model to enhance the precision of Carbon Dioxide (CO2) emission predictions. This study utilizes the cutting-edge forecasting techniques for a more accurate understanding of environmental impact. This study will harness the power of smart forecasting to inform strategic decision-making in carbon mitigation efforts. In this research, the performance metrics of four commonly used ML classifiers, namely LR, Gaussian Process, MLP and SMOreg have been evaluated to foretell CO2 emissions using the dataset collected from Kaggle. The dataset was pre-processed, and all the algorithms were trained and tested. The number of instances used in this study is 935. The results of this investigation show that machine learning algorithms are capable of producing accurate CO2 emission forecasts. The findings suggest that the SMOreg Classifier is more accurate than the LR (LR), Gaussian Process Regression (GPR) and Multilayer Perceptron (MLP) classifiers for predicting CO2 emissions. This study emphasizes the possibility of using ML algorithms to predict CO2 emissions. The error values such as MSE, RMSE, MAE, Correlation Coefficient and Root relative squared error indicates the performance of SMOreg is a superior classifier for the forecasting, these results have an important implication in climate change for improving prediction models, which could assist in early detection of climate change.**

Keywords—Linear Regression, Gaussian Process, Multilayer Perceptron, SMOreg, Correlation Coefficient, RMSE, Machine Learning, CO2 emission.

I. INTRODUCTION

The Earth's population is experiencing exponential growth, leading to a continual rise in carbon dioxide depletion. This escalating trend is causing a pronounced surge in environmental overheating, emerging as a foremost provider to climate change. Universal initiatives aimed at addressing climate change are primarily directed towards minimizing the frequency of extreme environmental overheating in the future. Numerous studies and investigations undertaken by scientists, students, and other officials delve into the varied features causative to the heightened CO2 emissions across different nations.

The growing concern over the impact of greenhouse gas emissions, especially carbon dioxide (CO2), on the Earth's climate has elevated its significance in recent years. CO2 emissions are a result of social movement, comprising the sweltering of remnant fuels, deforestation, and industrial manoeuvres [1]. It is imperative to comprehend the extent of CO2 emissions from each country, as this knowledge is essential for formulating effective climate policies and alleviating the repercussions of climate change [2]. The rate at which we increase atmospheric carbon dioxide content in a given year is directly correlated with our ability to outpace natural carbon removal mechanisms. The proportion of increase in impressive CO2 worldwide throughout the 1960s was roughly 0.8 ± 0.1 PPM per year. This annual upsurge rate trebled during the next 50 years, reaching a peak of 2.4 ppm annually in the 2010s. The rate at which atmospheric CO2 has increased over the last 60 years is roughly 100 spells quicker than ordinary rises in the past, that occurred 11,000–17,000 years ago at the end of the last ice age.

Reducing the amount of heat-trapping greenhouse gases released into the atmosphere is necessary to combat climate change. Reducing greenhouse gas emissions from main sources, such as companies, power plants, cars, and agricultural practices, is necessary to achieve this. In addition too, forests, oceans, and soil play a crucial role by absorbing and storing these gases, contributing significantly to the overall solution. Effectively lowering and preventing emissions demands a comprehensive overhaul of our activities from restructuring the way we fuel our economy and cultivate our food, to altering our modes of transportation and lifestyle choices, as well as the products we use. This issue has repercussions both locally and globally.

II. RELATED WORKS

Recognizing the need for a comprehensive approach, it is crucial to examine the broader implications of ML, as emphasized in a recent Perspective article in Nature Climate Change (Lyn et al., 2023) [3]. Rather than solely prioritizing performance advancements in accuracy, [4] (Chen et al. 2023) advocate for a shift towards augmenting the balance between accuracy and CO_2 emissions. (Ning Ma et al. 2021) urged researchers to publish the influence of their models' carbon emissions in scientific journals, supporting transparency even when results are reported at the order-of-magnitude or qualitative evaluation levels [5]. Introduced the conventional parametric modelling approaches along with GPR algorithms, with a subsequent overview of their predictive performance. The dependability and effectiveness of the suggested processes were illustrated by contrasting actual and predicted outcomes. The findings underscored that the GPR method stands out in providing the most precise predictions for CO_2 emissions.

The determination of (Shanshan Li et al. 2021) is to examine the connection between CO2 discharges and variables including monetary expansion, industrial

composition, expansion, investment in R&D, real use of external capital, and the rate of increase in energy consumption [6]. The nonlinear KNN unit beats linear, nonlinear, ensemble, and ANN models for the provided dataset, according to model selection based on root mean square error (RMSE). A sensitivity study of CO2 releases around its centroid point was carried out using the KNN model.

Researchers have carried out a great deal of study to predict carbon emissions throughout time. In an investigation conducted by (Z. Xu et al. 2021) [7], CO2 emissions across 53 nations and numerous areas were predicted using a non-equigap grey technique. Carbon dioxide emissions were the experiment's output, and energy usage was its input. (Wang et al. 2020) used a Land Use Regression (LUR) option with a dataset containing a variety of variables, to forecast the levels of air pollution caused by traffic [8]. (Zhang et al, 2021) investigated different DL and ML algorithms to forecast CO2 emissions from agricultural soil; their LSTM model performed better in predicting CO2 and N2O [2]. Used cutting-edge ML methods, to forecast and examine the relationship between coal use, CO2 emissions, GDP, and the generation of solar and wind energy. Used a range of approaches, including dimensionality reduction, clustering, and ML, including singular value decomposition (SVD) and fuzzy neural networks, to project carbon emissions. In Changxing, China, used ML algorithms to forecast urban block carbon emissions (UBCE) based on ecological parameters [9]. Used models like LR, Random Forest Regression, and LSTM along with ML approaches on time series data to forecast CO2 emissions in India [4].

Autoregressive and Seasonal are the two Integrated Moving Average time series-based ML techniques that were employed (Mardani et al. 2018) [9]. CO2 emissions can be forecast with the use of AI and ML techniques. AI contributes significantly to mitigating weather change by optimising energy competence and providing precise data to decision makers. [10] (Kumar S. et al. 2022) Three statistical models were used: the Holt-Win. Method has average with exogenous factors (SARIMAX), autoregressive-integrated moving average (ARIMA), and two ML models: random forest and linear regression. Additionally, a DL-based LSTM model was also used. Results show that the LSTM model works better than models in estimating CO2 emissions. It has a 3.101% MAPE value, a 60.635 RMSE value, a 28.898 MedAE value, and it performs better than other models in other pertinent metrics (Kumari S. et al. 2022) [11].

III. MATERIAL AND METHODS

Recognizing The data for this study is collected from Kaggle dataset. There are 935 instances with 12 attributes. The attributes are Model_Year, Make, Model, Vehicle_Class, Engine_Size, Cylinders, Transmission, Fuel_Consumption_in_City (L/100 km), Fuel_Consumption_in_City_Hwy (L/100km), Fuel_Consumption_comb(L/100km), CO2_Emissions and Smog_Level. This paper utilizes a weka classification algorithm in ML to forecast CO2 releases. A comprehensive prose review was showed to identify appropriate autonomous variables. Initially the datasets were undergone the preprocess steps using Weka tool. In this study Interquartile range from unsupervised learning was applied. After preprocessing the attributes outlier and Extreme Value attributes were added with the existing attributes. 70% of data was occupied as training set and 30% of the data was taken as test data. Then the classification models such as LR, Gaussian Process, Multilayer perceptron and SMOreg algorithms were applied on the preparation data in direction to train the structure After Building the model on the training data. The models were tested with the test data and their prediction recital, along with effective evaluation criteria for model concert measurements, was shortened. The proposed algorithms' reliability and efficiency were demonstrated by comparing actual data with predicted results. The findings indicate that SMOreg yields the greatest precise forecasts for CO2 emanations. The flow diagram for this study is shown in Fig. 1.

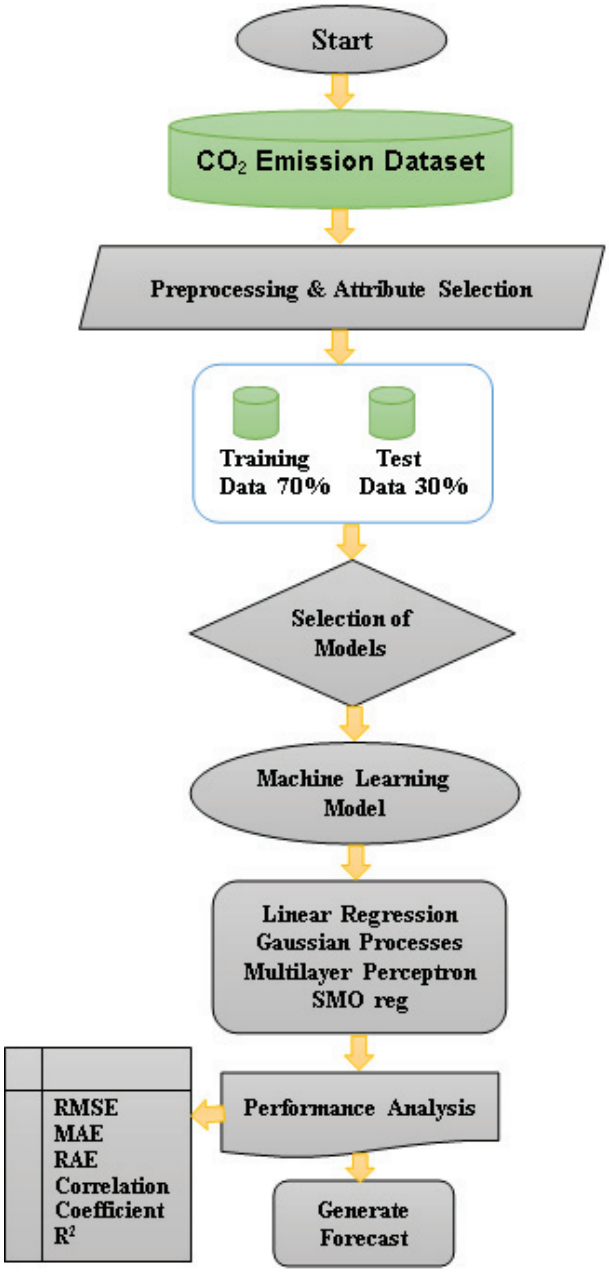

Fig. 1. Flow chart depicting the methodology adopted in the study

A. *Linear Regression*

LR, as a ML model for addressing regression glitches, resolves the task by positing a linear association between the provided input characteristics and the output. A class designed for prediction using LR. It employs the Akaike

standard for model assortment and has the capability to handle instances with different weights.

B. *Gaussian processes*

The Applies Gaussian courses for regression without hyper-parameter tuning. This implementation facilitates the selection of an optimal noise level by normalizing/standardizing both the target attribute and other attributes (when normalization/standardization is enabled) [12]. Global mean/mode replaces misplaced values, and insignificant attributes are transformed into binary format. It's important to memo that kernel hoarding is disabled when the employed kernel supports Cached-Kernel [1]

C. *Multilayer Perceptron*

A classification model employing back propagation for learning a multi-layer perceptron to categorize instances. The network can be manually constructed or established using a straightforward heuristic. Throughout the training process, the network strictures are subject to monitoring and modification. All nodes in this network operate as sigmoid units, with the exception of numeric classes, where the output nodes transform into unthresholded linear units.

D. *SMOreg*

SMOreg employs the SVM for regression, allowing the learning of strictures through diverse algorithms. The choice of the algorithm is determined by configuring the RegOptimizer [12]

IV. Statistial Analysis

Climate change is propelled by anthropogenic greenhouse gas (GHG) emissions, with approximately 60% originating from only 10 countries, while the 100 least-emitting nations contribute less than 3%. The top three counties are China, United States and India [13]. Figure shows the global historical emission by 3 countries. The energy sector accounts for approximately three-quarters of global releases, with agriculture following closely behind. Among energy-related activities, electricity and heat generation emerge as the leading contributors, trailed by transportation and manufacturing [14]. Additionally, the land use, and forestry (LULUCF) sector play a dual role as both a source and sink of emissions, making it a pivotal sector in achieving net-zero emissions [15]. Fig. 2 shows the global historical emissions.

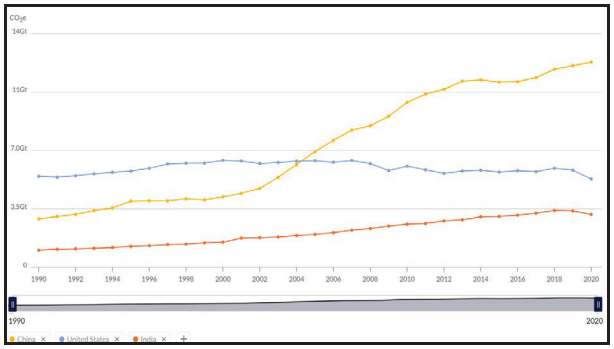

Fig. 2. Global Historical Emission

Fig. 3, 4 and 5 shows the Cumulative CO2 emission by China, United States and India.

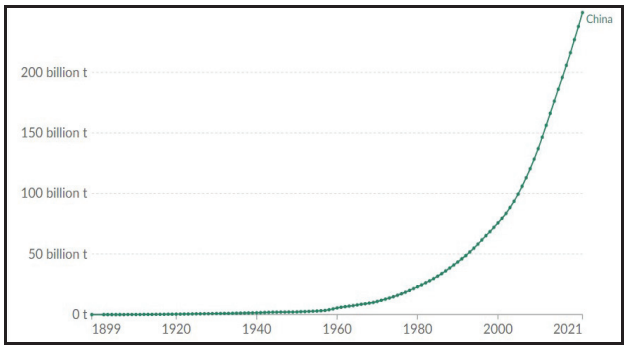

Fig. 3. CO_2 Emitted by China

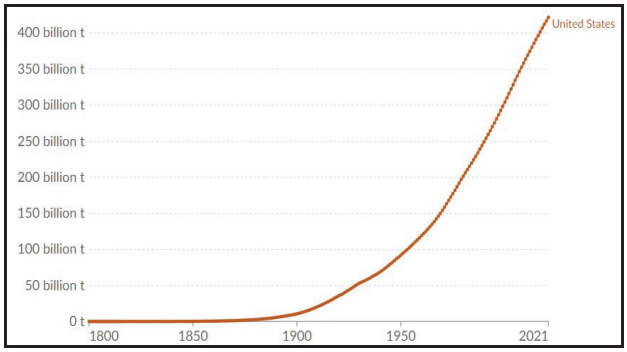

Fig. 4. United States CO2 Emission

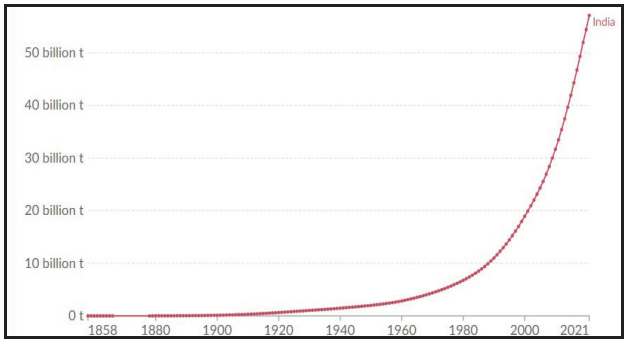

Fig. 5. Cumulative CO_2 Emission by India

The following figures 6, 7 and 8 shows the Production vs consumption-based CO_2 emissions in China, Us and India.

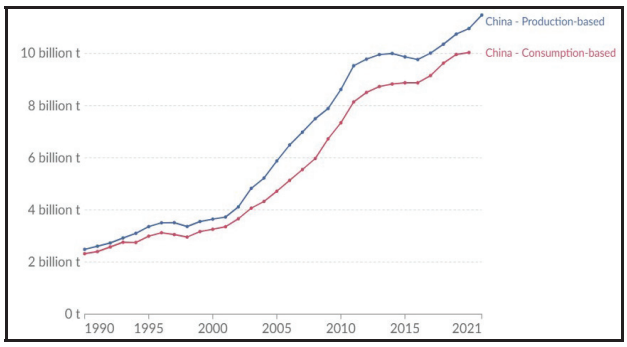

Fig. 6. Production vs consumption-based CO_2 emissions in China

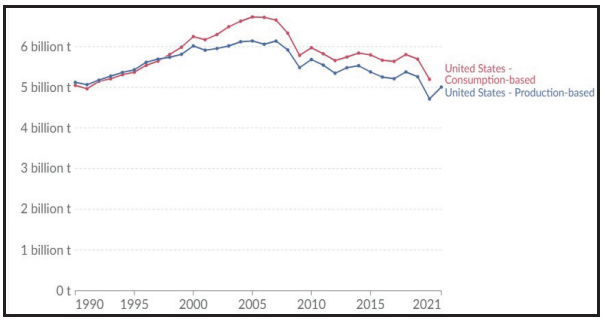

Fig. 7. United States Production vs Consumption-based CO2 emissions

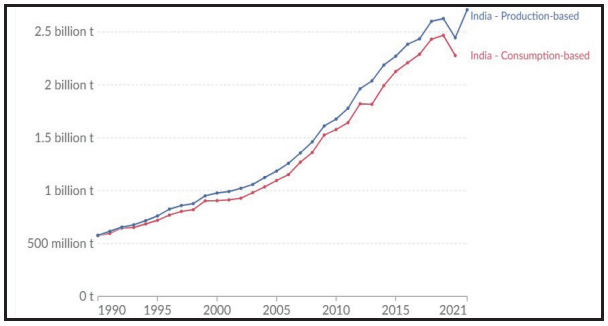

Fig. 8. India's Production vs consumption-based CO2 emissions

Using statistical Regression, the attributes such as Fuel Consumption in CityL100, Fuel Consumption in City HwyL100 and Fuel Consumption CombL100km are compared for the CO2 emission. Fig. 9, 10 and 11 shows the comparison result.

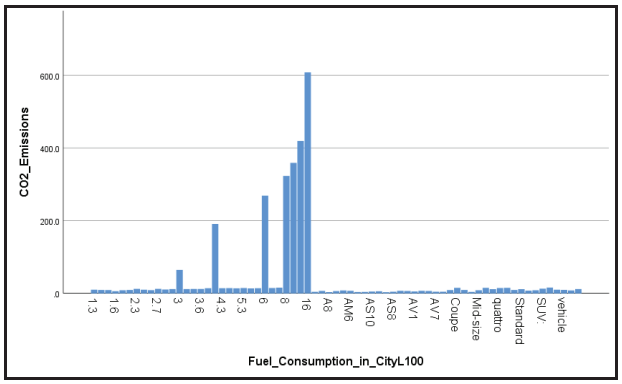

Fig. 9. CO_2 emission Vs Fuel Consumption in CityL100

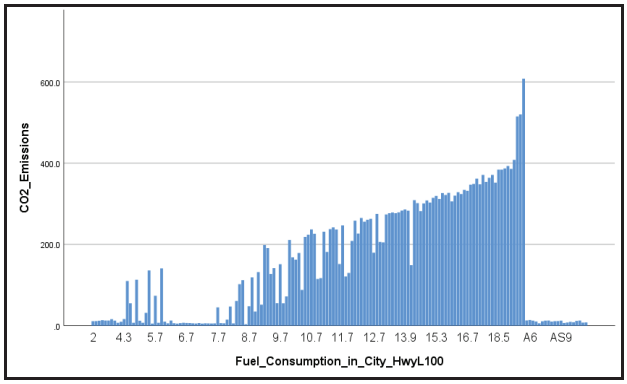

Fig. 10. CO_2 emission Vs Fuel Consumption in City HwyL100

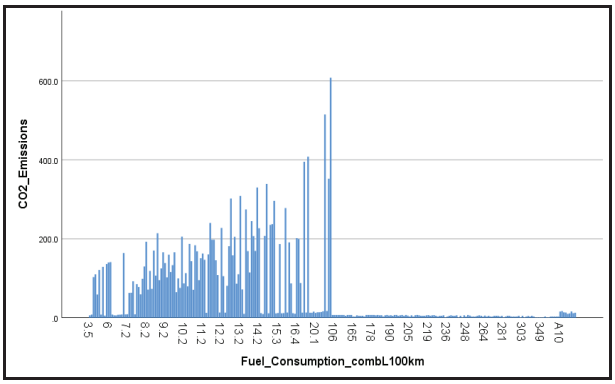

Fig. 11. CO_2 emission Vs Fuel Consumption CombL 100Km

V. RESULTS

The performance of the classifiers such as LR, Gaussian Process, Multilayer perceptron and SMOreg algorithms were measured for training and the test data. Amont the regression models the SMOreg classifier stands out in providing the most precise prediction for CO_2 emission. The performance of the classifiers for training data and test data are listed in table 1 and 2. The Classifiers performance for the preparation and test data using Scatter plot are shown in Fig. 12 and 13.

TABLE I. CLASSIFIERS PERFORMANCE FOR THE TRAINING DATA

0.9971	2.7739	4.8716	5.4038 %	7.5636 %
0.9906	6.2394	9.0038	12.1549 %	13.9792 %
0.9923	5.0311	7.991	9.801%	12.4067%
0.9726	6.3913	15.3038	12.4509 %	23.7606 %

TABLE II. CLASSIFIERS PERFORMANCE FOR THE TEST DATA

0.9992	**1.3093**	**2.1157**	**2.5202 %**	**3.6561 %**
0.9946	**5.01**	**6.0358**	**9.6437 %**	**10.4305 %**
0.9929	**5.259**	**6.6569**	**11.0218 %**	**12.1924 %**
1.0000	**0.3684**	**0.4028**	**0.7722 %**	**0.7378 %**

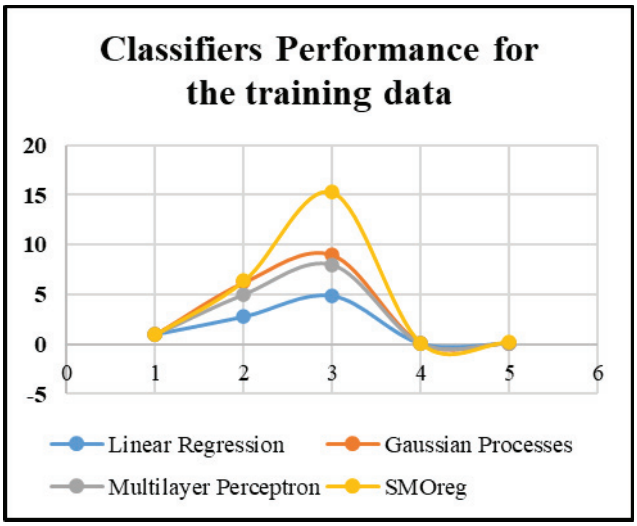

Fig. 12. CO_2 emission prediction using classifiers on training data

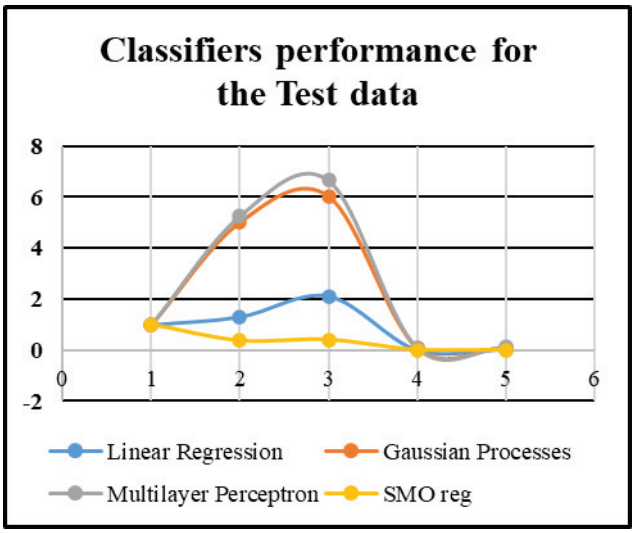

Fig. 13. CO_2 emission prediction using classifiers on test data

As per [5] GPR was the most precise predictions for CO_2 emissions. But according to the table 2 the performance of SMOreg is better when compare with LR and GPR and MLP. The metrics, which are displayed in 14 to 18, are plotted using bar graphs and include correlation coefficient, MAE, RMSE, RAE, and RRSE.

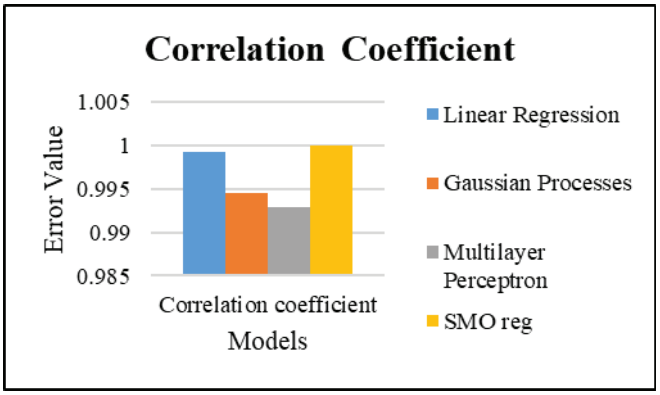

Fig. 14. Perfect Positive Linear Relationship

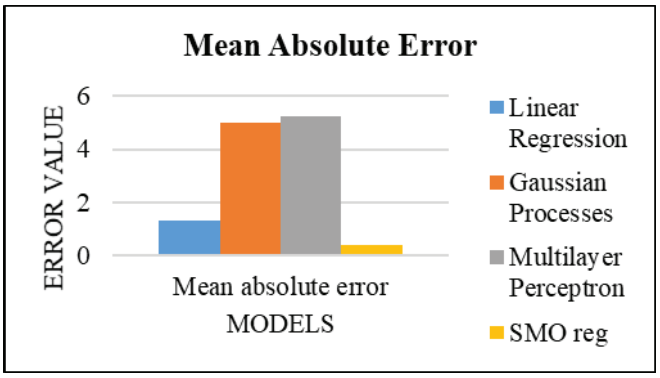

Fig. 15. Average absolute difference between predicted and actual values

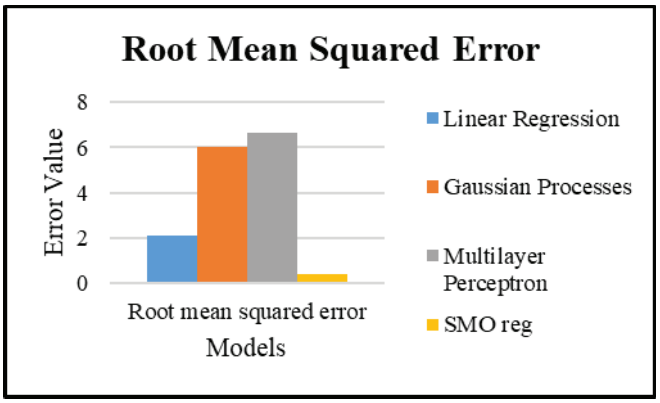

Fig. 16. Measure of the average magnitude of error

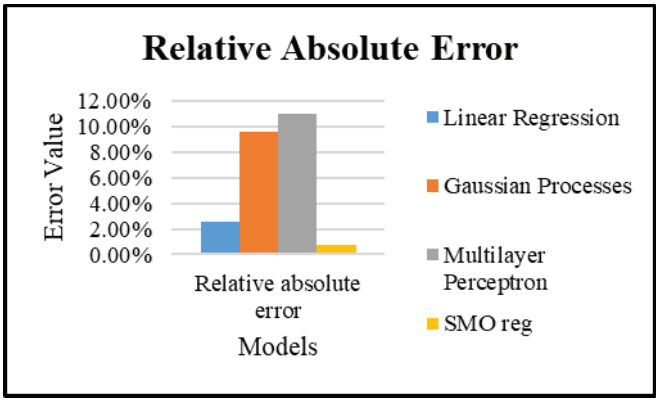

Fig. 17. performance of a predictive model

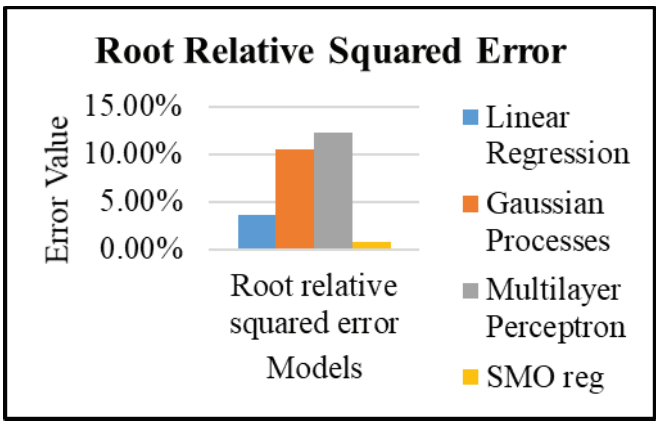

Fig. 18. Average of Actual Values

VI. Discussion

Metrics like RMSE, RRSE, MAE, and MSE are quantified. The average squared variations on real and the obtained parameters are compared by MSE. Better performance is designated by a lower MSE. The average absolute changes on actual and obtained parameters are validated using MAE [16]. Better presentation is designated by a lower MAE. The average size of error is measured by RMSE, which is the square root of the MSE. A lower RMSE indicates better performance [17]. R-squared values are different from the independent variables. It ranges from 0 to 1, and a higher R-squared indicates a better fit. RAE is a way to show how well a predictive model performs, and it's represented as a ratio. RRSE compares the accuracy of a model by considering how it would perform compared to a basic predictor, which is simply the average of the actual values [19]. For MSE, MAE, and RMSE, lower values are better. They represent smaller errors between predicted and actual values. For R-squared, higher values (closer to 1) are better, as they indicate a higher proportion of variance explained by the model [20]. The range of the correlation coefficient is -1 to 1. A flawless positive linear relationship is denoted by r=1, and a unspoiled negative linear relationship by r=-1. Better model performance is indicated by a lower RAE. Fig. 19 and 20 shows the CO_2 emission based on engine size and vehicle type.

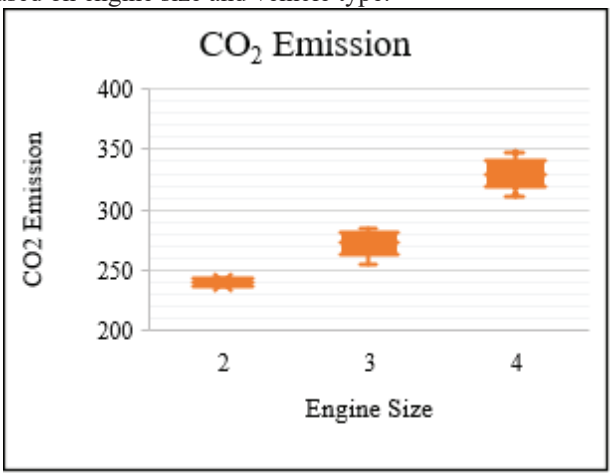

Fig. 19. Box plot for CO_2 Emission based on Engine Size

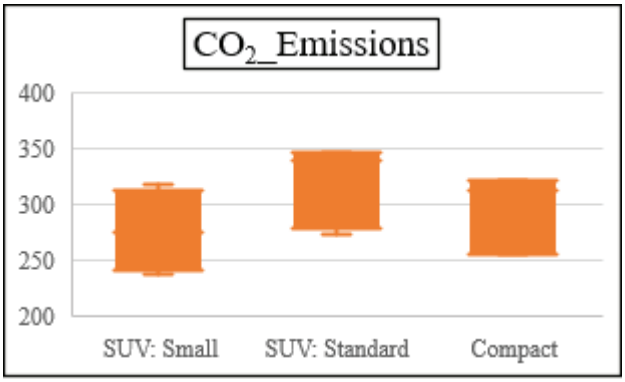

Fig. 20. Box plot for CO_2 Emission based on Vehicle Type

VII. Conclusion

It is imperative to accurately predict the future CO_2 emissions in India, as it holds significant implications for both the populace and the government. Currently, India ranks second among the largest contributors to CO_2 emissions, highlighting the urgency of our efforts. Our contribution lies in mitigating CO_2 emissions to safeguard human lives. Employing statistical, ML, and DL-based time series models, we examined the CO_2 emission patterns. The presentation of these models was rigorously assessed using nine appropriate metrics, aiding in the selection of the most suitable model for future forecasting. Moreover, a comparison was conducted with recent studies, evaluating the proposed model's performance using metrics like RMSE, MAE, RAE, RRSE, and correlation coefficient. The findings from the comparative analysis indicate that SMoreg stands out as the most fitting model, exhibiting the least errors in RMSE, MAE, RAE, RRSE and correlation Coefficient.

References

[1] Y. Meng, "Predicting CO2 Emission Footprint Using AI through Machine Learning," Atmosphere, vol. 13, p. 1871, 2022. DOI: 10.3390/atmos13111871.

[2] X. Zhang, F. Yan, H. Liu, and Z. Qiao, "Towards low carbon cities: a machine learning method for predicting urban blocks carbon emissions (UBCE) based on built environment factors (BEF) in Changxing City, China," Sustain. Cities Soc., vol. 69, p. 102875, 2021. DOI: 10.1016/j.scs.2021.102875.

[3] L. Chen, "Artificial intelligence-based solutions for climate change: a review," Environmental Chemistry Letters, vol. 21, pp. 2525–2557, 2023. DOI: 10.1007/s10311-023-01617-y.

[4] K. B., J. A. A., R. B., D. K. R., K. K., and M. J. N., "Implementation of Effective Rainfall Forecast Model using Machine Learning," in 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), pp. 1655-1660, 2023.

[5] N. Ma, "Can Machine Learning be Applied to Carbon Emissions Analysis: An Application to the CO2 Emissions Analysis Using Gaussian Process Regression," Frontiers in Energy Research, published: 24 September 2021, doi: 10.3389/fenrg.2021.756311.

[6] S. Li, "Driving Factors of CO2 Emissions: Further Study Based on Machine Learning," Frontiers in Environmental Science, published: 23 August 2021, doi: 10.3389/fenvs.2021.721517.

[7] Z. Xu, L. Liu, and L. Wu, "Forecasting the carbon dioxide emissions in 53 countries and regions using a non-equigap grey model," Environ. Sci. Pollut. Res., vol. 28, pp. 15659–15672, 2021. DOI: 10.1007/s11356-020-11638-7.

[8] A. Wang, J. Xu, R. Tu, M. Saleh, and M. Hatzopoulou, "Potential of machine learning for prediction of traffic related air pollution," Transp. Res. Part D Transp. Environ., vol. 88, p. 102599, 2020. DOI: 10.1016/j.trd.2020.102599.

[9] A. Mardani, D. Streimikiene, M. Nilashi, D. Arias Aranda, N. Loganathan, and A. Ju-soh, "Energy Consumption, Economic Growth, and CO2 Emissions in G20 Countries: Application of Adaptive Neuro-Fuzzy Inference System," Mdpi.Com, doi: 10.3390/en11102771, 2018.

[10] S. Kumar, "Machine learning-based time series models for effective CO2 emission prediction in India," Environmental Science and Pollution Research, published online: 2nd July 2022, published by Springer, DOI: 10.1007/s11356-022-21723-8.

[11] S. Kumari, "Machine learning-based time series models for effective CO2 emission prediction in India," Environmental Science and Pollution Research, June 2022, DOI: 10.1007/s11356-022-21723-8.

[12] D. J. C. Mackay, "Introduction to Gaussian Processes," Dept. of Physics, Cambridge University, UK, 1998. [Online]. Available: http://wol.ra.phy.cam.ac.uk/mackay/gpB.ps.gz

[13] A. Hamrani and A. Akbarzadeh, "Machine Learning for Predicting Greenhouse Gas Emissions from Agricultural Soils," Elsevier, 2020.

[14] A. T., A. Suganya, R. Muthalagu, M. Narender, and J. A. A., "Evaluation of a Wireless Sensor Network for the Detection of Forest Fires," in 2023 2nd International Conference on Edge Computing and Applications (ICECAA), pp. 1021-1027, 2023.

[15] L. H. Kaak Kaack, "Tackling Climate Change with Machine Learning," ACM Computing Surveys, vol. 55, no. 2, article 42, February 2022.

[16] C. Magazzino and M. Mele, "A Machine Learning Approach on the Relationship Among Solar and Wind Energy production, Coal consumption, GDP, and CO2 Emissions," Elsevier, 2021.

[17] Y. Meng and H. Noman, "Predicting CO2 emission footprint using AI through Machine Learning," Atmosphere, vol. 13, p. 1871, 2022. DOI: 10.3390/atmos13111871.

[18] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," IEEE Transactions on Neural Networks, 1999.

[19] H. Liang and W. Song, "Improved estimation in multiple linear regression models with measurement error and general constraint," J. Multivar. Anal., vol. 100, pp. 726–741, 2009.

[20] Z. Zuo, H. Guo, and J. Cheng, "An LSTM-STRIPAT model analysis of China's 2030 CO2 emissions peak," Carbon Manag, vol. 11, no. 6, pp. 577–592, 2020.