Data-Driven Insights into Climate Change: A Study on Predicting CO2 Emissions with Advanced Analytics and Exploratory Data Analysis

E. Shravan Kumar[1], Mukkera Deepika[2], Gaddam Vinay[2], Ramini Laxma Reddy[2], Perum Rithin Yadav[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of CSE (Data Science)
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana

ABSTRACT

There is a major contribution that carbon dioxide (CO_2) emissions make to the phenomena of global warming. This phenomenon has serious consequences, including the occurrence of extreme weather events, the rise of sea levels, and the disruption of ecological equilibrium. In order to properly moderate and reduce CO_2 emissions in a sustainable manner, it is absolutely necessary for us to have a comprehensive grasp of the factors that influence them. Since this is the case, the purpose of this research is to evaluate various machine learning methods with the intention of predicting and projecting CO_2 emissions. Furthermore, we want to use exploratory data analysis (EDA) approaches, which will serve to facilitate the visualization and interpretation of the data in a manner that is both efficient and effective. Through the use of EDA, we are able to identify significant traits, interpret data distributions, and identify outliers that may have an effect on the performance of a model. The fact that our work can provide policymakers and environmentalists with new and insightful perspectives is the primary reason for its significance. It is possible for us to promote the establishment of effective policies that control and reduce emissions, optimize the allocation of resources, and encourage the transition to renewable energy sources if we are able to reliably anticipate CO_2 emissions. Furthermore, precise projections might be of assistance in the process of formulating different adaptation methods in order to mitigate the effects of climate change.

Keywords: Pollution controlling, CO_2 emissions, predictive analytics, machine learning, exploratory data analysis.

1. Introduction

Our environment requires a particular quantity of CO_2 in order to function properly. Emissions of carbon dioxide that are excessive have some effect on the ecosystem. A significant amount of CO_2 is being released into the atmosphere constantly as a result of industrialization and other human activities. When the COVID-19 pandemic began, the world had already been experiencing the largest quantity of CO_2 emission that had ever been recorded. In the course of the transmission (trans) time of COVID-19, the emission of carbon dioxide has decreased to 34.4 million tons (MT), which is a decrease from the previous peak of 36.1 MT [1]. There are numerous works that estimate CO_2 emission before the pandemic but there is no suitable work showing how CO_2 emissions will behave in the trans- and post-COVID-19 era, because most of the recent works either use data from before the pandemic, such as [2] (up to 2018), [3] (up to 2018), [4] (up to 2015), or they use data from during pandemic but with a local scope, such as [2] for India, [3] for Turkey, [4] for the UK, [5] for China and [6] for indoor environments, or they use different approaches for only near future (2 years) forecasting [7]. The purpose of this study is to construct an Artificial Intelligence (AI) model that is based on Machine Learning (ML) in order to forecast CO_2 emissions in the near and distant future. This prediction will take into account the decreased CO_2 emissions that have occurred as a result of lockdowns for the COVID-19 pandemic.

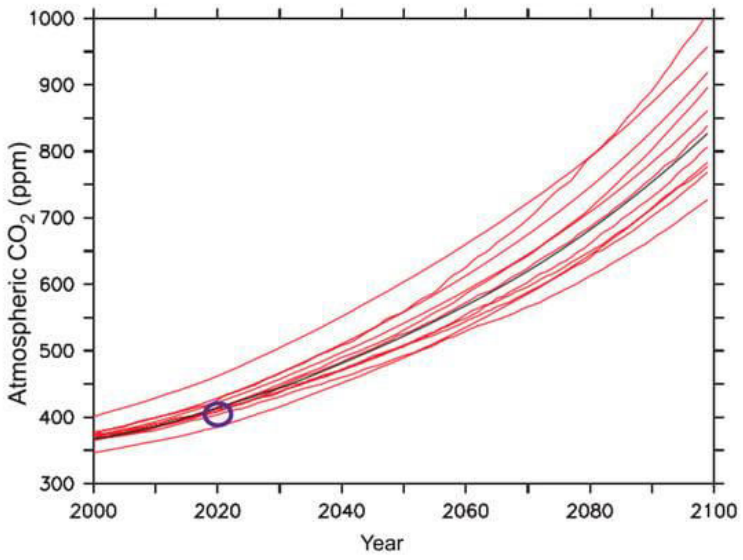

Figure 1: CO_2 emission forecasts by IPCC model [27].

1.1. Global CO_2 Emission Crisis

As a result of the greenhouse effect, it is well knowledge that carbon dioxide emissions are a significant contributor to this phenomenon [8]. There is a broad consensus, as evidenced by [9, 10], that CO_2 emission is the primary cause of global warming. This is despite the fact that there is a debate on whether or not carbon dioxide is responsible for the phenomenon of global warming [8]. As a consequence of this, estimates and projections of the carbon dioxide emissions footprint are essential for a number of reasons, including the following: The first step is to conduct an analysis of carbon dioxide emissions in order to determine the primary contributors to global warming. This is because carbon dioxide emissions are widely regarded as the primary cause of both climate change and global warming [11, 12]. A second objective is to get an understanding of the CO_2 emission footprint in order to formulate a strategy to combat it. Also, in order to pay for any financial or environmental losses that have been experienced as a result of CO2 emission. The fourth objective is to evaluate the rational impact of carbon dioxide emissions on the loss of gross domestic product [13], the casualty of the stock market [14], the upheaval of new or old diseases [15], the disruption of air quality [16], and the impact on the construction of a smart city that is greener and cleaner. First and foremost, it is crucial to have accurate forecasts of CO_2 emissions in order to measure and combat climate change that cannot be reversed [12].

2. Related work

Since the beginning of time, there have been a few studies that have attempted to estimate the global CO_2 emission footprint. These works include the COVID-19 transmission era, such as [7]. Most of the existing works either have a partial to local context, such as [17] in China, [18] in China, [19,20] in wheat fields, [21] in Iran, and [22] in the Middle East, or the modeling parameters and methodology are not appropriate for global CO2 emission prediction, such as [3] for Indian paddy fields and [2] for the Turkish transportation sector. These are just a few examples.

The following is a chronological presentation of the strengths and limits of the local works that are now available. Using machine learning and artificial neural network-based modeling, a forecast of the local CO2 emission for the Iranian domain was presented in [21]. There is a presentation of the CO2 emissions that are caused by the manufacture of cement and fossil fuels in [23]. A local county in China known as Changxing served as the basis for the CO2 driver and emission forecasting activity

that was carried out in [17]. Furthermore, the dataset from the Indian area that spans the years 1995 to 2018 is utilized in the analysis and forecasting of the CO2 emissions that are shown in reference [24]. [22] is a presentation of the CO2 emissions that occur in the Arabian region. The effect of CO2 emissions having a synergistic effect on the reduction of PM2.5 emissions in the Chinese region is provided in paragraph [25]. Regarding the construction of a CO2 emission model for a global scenario, reference [26] is missing, despite the fact that it offers a satisfactory overview of CO2 emission and the problems that are associated with it. The model that was established by the Intergovernmental Panel on Climate Change (IPCC) [27] is now the one that provides the most accurate predictions of CO2 levels. There are projections of CO2 emissions that are provided, including 398 parts per million (ppm) in 2019, 400 ppm in 2020, 402 ppm in 2021, and 405 ppm in 2022. As a unit for measuring CO2 emissions, ppm is an abbreviation that stands for parts per million. For the years that do not involve a pandemic, the results of the forecasting that is performed using the IPCC model are satisfactory; however, the model exhibits a declining behavior when it comes to projecting the values of emissions during the pandemic spike. As a consequence of this, a model that is more inclusive needs to be implemented.

For the purpose of predicting CO2 emissions, a variety of writers taking a variety of perspectives have experimented with a number of different modeling methodologies. Importantly, reference [28] offers some insight into the algorithm that uses machine learning to anticipate CO2 emissions. Throughout the period of China's economic development [29], the international community has exerted pressure on the country to address issues pertaining to CO2 emissions and environmental protection. Through the utilization of combined principal component analysis (PCA), a novel hybrid model was constructed for China in [19]. This model was based on data spanning from 1978 to 2014. In addition, the trends in carbon dioxide emissions from fossil fuels in Zambia from 1964 to 2016 are presented in reference [30]. An investigation into a prediction model for carbon dioxide emissions in the Chinese context was carried out in [31]. The model was based on multiple linear regression analysis. In addition, two models have been constructed for the purpose of creating a simulation of the CO2 emissions that wheat farms in New Zealand [20] produce. In addition to this, the SVM model was suggested as a means of forecasting the expenditure of carbon dioxide (CO2) emission in [32]. [33] presents a data mining approach that can be used to determine CO2 emissions from data collected from vehicles. For the purpose of predicting the amount of carbon dioxide (CO2) emissions that would be spent, the back-propagation artificial neural networks (ANN) model was introduced in [34]. A quantitative analysis that was published in [35] was conducted to analyze the impact that CO2 has on the fluctuation of temperature in five different regions. According to the findings, carbon dioxide is responsible for fifty-two percent of the increase in global temperature that occurred between the years 1990 and 2010 [35]. As can be seen in Figure 1 [27], a similar conclusion was also true for the subsequent decade, which lasted from 2010 to 2019, until the COVID-19 epidemic began to spread significantly over the globe.

3. PROPOSED SYSTEM

The research work starts with a discussion of the findings, including insights gained from EDA, the effectiveness of data preprocessing techniques, the performance of the existing and proposed KNN models, and any recommendations for improving CO_2 emission prediction and forecasting using machine learning. Additionally, the research work should discuss the limitations of the study and potential areas for future research. Figure 2 shows the proposed system model.

The detailed operation illustrated as follows:

Step 1. Exploratory Data Analysis (EDA):

- Data Collection: Gather the dataset containing historical CO2 emission data along with relevant features such as population, GDP, energy consumption, etc.
- Data Inspection: Examine the dataset's structure, including the number of rows and columns, data types, and any missing values.
- Data Visualization: Create various plots and visualizations to gain insights into the data's distribution, trends, and relationships. This may include histograms, scatter plots, correlation matrices, and bar charts.
- Outlier Detection: Identify and handle outliers in the dataset, as extreme values can adversely affect machine learning models.

Step 2. Data Preprocessing:

- Feature Selection: Choose the most relevant features for CO2 emission prediction. This step involves selecting a subset of features that have the most impact on the target variable.
- Handling Missing Data: Address any missing values in the dataset through techniques like imputation or removal of rows/columns with missing data.
- Normalization/Scaling: Scale numerical features to ensure they have similar scales, which can improve the performance of some machine learning algorithms.
- Encoding Categorical Data: If applicable, convert categorical data into numerical format using techniques like one-hot encoding.
- Data Splitting: Divide the dataset into training and testing sets for model development and evaluation.

Step 3. Existing KNN Model:

- Select Existing KNN Model: Choose a standard K-Nearest Neighbors (KNN) regression model as a baseline.
- Hyperparameter Tuning: Use techniques like grid search or cross-validation to find the best hyperparameters (e.g., the number of neighbors) for the KNN model.
- Model Training: Fit the selected KNN model to the training data.

Step 4. Proposed KNN Model:

- Feature Engineering: Create new features or combinations of features that may improve the prediction of CO2 emissions.
- Hyperparameter Tuning: Similar to the existing KNN model, optimize the hyperparameters for the proposed KNN model.
- Model Training: Train the proposed KNN model using the training data.

Step 5. Prediction:

- Predict CO2 Emissions: Use both the existing and proposed KNN models to predict CO2 emissions for the testing dataset.

Step 6. Performance Estimation:

- Mean Absolute Error (MAE): Calculate the MAE to quantify the average absolute difference between predicted and actual CO2 emissions.

- Mean Squared Error (MSE): Compute the MSE to measure the average squared difference between predicted and actual emissions.
- Root Mean Squared Error (RMSE): Calculate the RMSE by taking the square root of MSE, providing a measure in the original unit (e.g., Mt).
- R-squared (R2) Score: Determine the R2 score to evaluate how well the model explains the variance in CO2 emissions.
- Comparison: Compare the performance metrics between the existing and proposed KNN models to assess whether the proposed model provides better predictions.

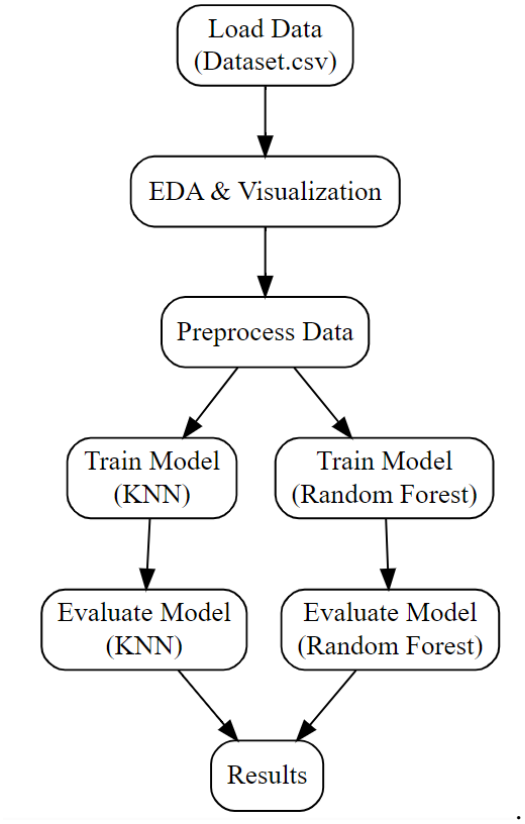

Figure 2: Block diagram of proposed system.

The following EDA performed in this work:

Histogram Plots: Histograms are generated for all numeric features in the dataset to visualize the distribution of each variable. This code snippet creates a figure with a size of 15x15, specifies the axis for plotting, and then generates histograms for each numeric feature in the DataFrame 'df.' These histograms help you understand the distribution of values within each feature.

Countplot for Target Variable: A countplot is created to check if the dataset is balanced or not with respect to the 'target' variable. The countplot visualizes the distribution of the 'target' variable, which is typically used in classification tasks to indicate the class labels. By examining the count of each class, you can assess whether the dataset is balanced or skewed towards certain classes.

Correlation Heatmap: A heatmap is generated to visualize the correlation between different features in the dataset. The code calculates the correlation matrix for all features in the DataFrame 'df' and selects the features with the highest correlation. It then plots a heatmap to display the pairwise

correlations between these selected features. The 'annot=True' parameter adds numerical values to the heatmap cells, providing insight into the strength and direction of correlations.

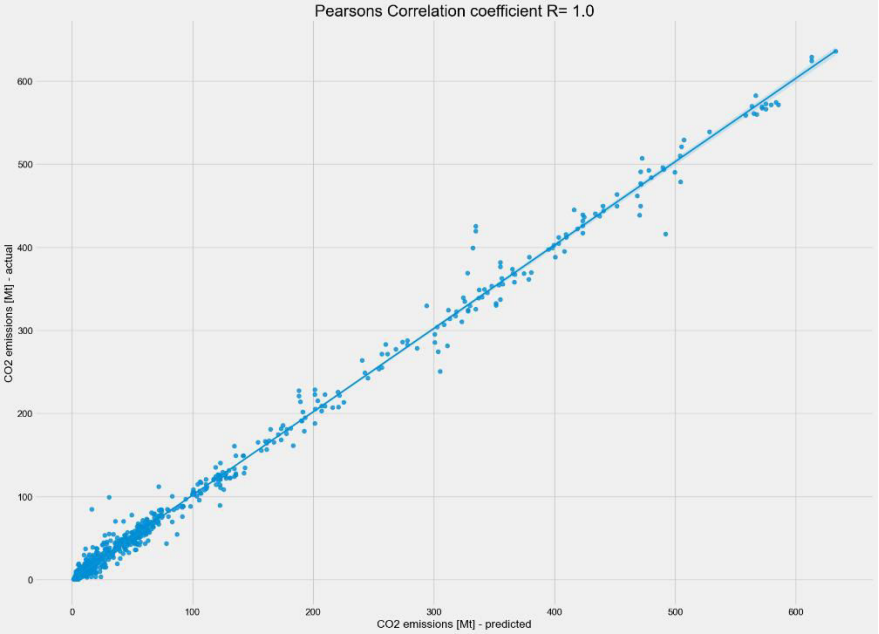

Figure 3: Prediction results using KNN.

4. SIMULATION RESULTS

Figure 3 presents the results of predictions made using the K-Nearest Neighbors (KNN) model. It may show a plot comparing the predicted CO2 emissions against the actual values. Figure 4 Similar to Figure 3, this figure displays the results of predictions. However, in this case, the predictions are generated using the Random Forest Classifier, a different machine learning model. Figure 5 provides a visual summary of the performance metrics (such as Mean Absolute Error, Mean Squared Error, etc.) for both the KNN and Random Forest Classifier models. It helps in comparing the effectiveness of the two models.

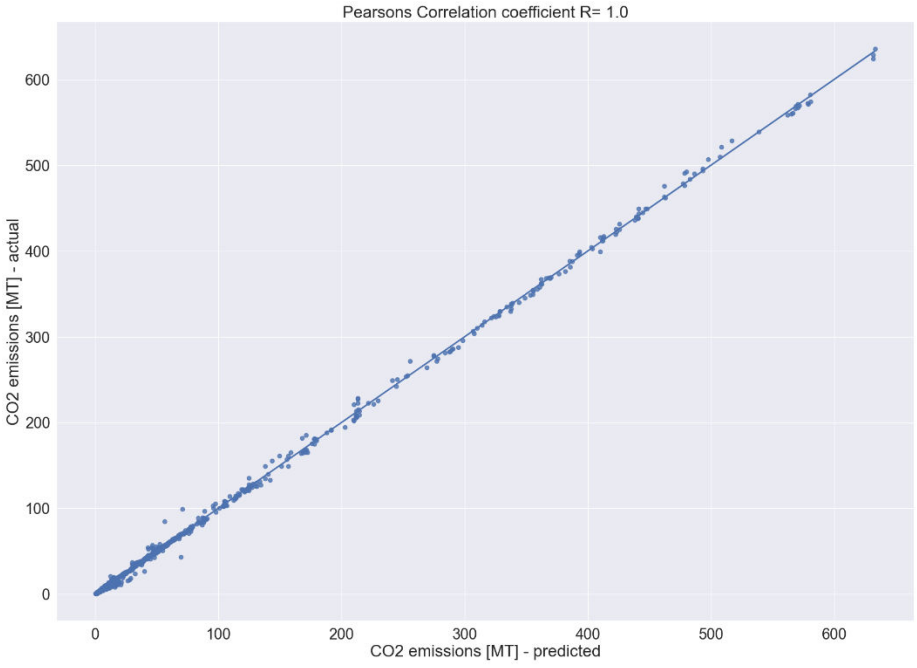

Figure 4: Prediction results using Random Forest Classifier.

	MAE	MSE	RMSE	R2_score
KNN	6.528102	126.592893	11.251351	0.993483
RF	1.919113	13.166444	3.628560	0.999322

Figure 5: Performance metrics of KNN & Random Forest classifier

Figure 11 displays a bar plot comparing the Mean Absolute Error (MAE) of the KNN and Random Forest Classifier models. It provides a visual representation of how well each model predicts CO2 emissions. Figure 12 Similar to Figure 11, this figure compares the Mean Squared Error (MSE) of the KNN and Random Forest Classifier models. It offers insights into the accuracy of the models' predictions. Figure 13 presents a bar plot comparing the R-squared (R2) scores of the KNN and Random Forest Classifier models. R2 score measures how well the model explains the variability in the data. This figure helps in understanding the goodness-of-fit of each model.

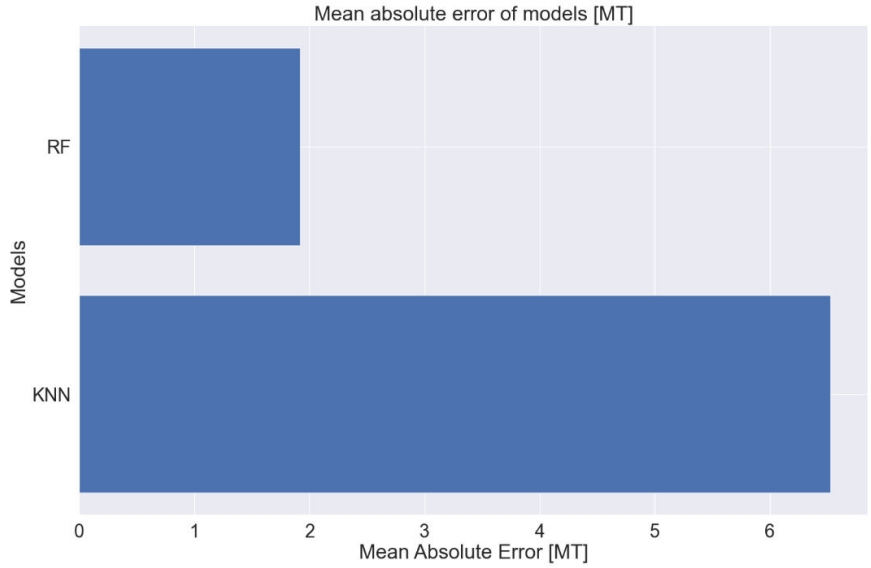

Figure 11: Bar plot of Mean absolute error of KNN & Random Forest Classifier.

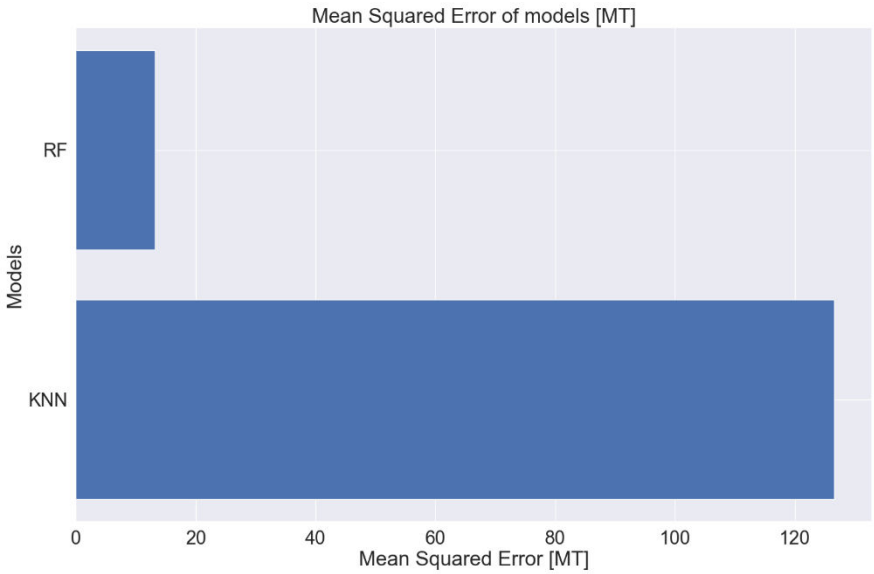

Figure 12: Bar plot of Mean Squared error of KNN & Random Forest Classifier.

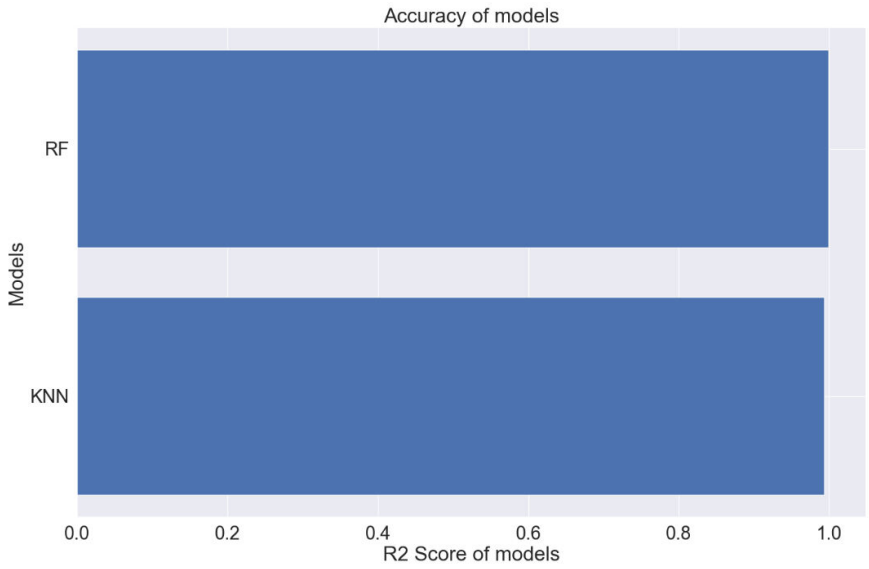

Figure 13: Bar plot of R2 Score of KNN & Random Forest Classifier.

5. CONCLUSION

In conclusion, the integration of machine learning models and exploratory data analysis (EDA) techniques offers a powerful approach for predicting and forecasting CO2 emissions, addressing the critical issue of climate change and its environmental consequences. Through this research, we have demonstrated the potential of machine learning to analyze large and intricate datasets, revealing hidden patterns and relationships that traditional statistical methods might miss. EDA has proven invaluable in providing a deeper understanding of the data, enabling the identification of influential features and outliers. By combining these two approaches, we can offer accurate and reliable predictions of CO2 emissions, empowering policymakers and environmentalists with valuable insights to develop effective strategies for emission reduction and sustainability. This work not only contributes to the scientific understanding of the factors driving CO2 emissions but also has practical implications in optimizing resource allocation, promoting renewable energy sources, and planning adaptation measures to mitigate the consequences of global warming.

REFERENCES

[1] Global Energy Review 2020. Available online: https://www.iea.org/reports/global-energy-review-co2-emissions-in-2021-2 (accessed on 16 October 2023).

[2] Singh, P.K.; Pandey, A.K.; Ahuja, S. Multiple forecasting approach: A prediction of CO2 emission from the paddy crop in India. Environ. Sci. Pollut. Res. Vol. 2022, 29, 25461–25472.

[3] Ağbulut, Ü. Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. Sustain. Prod. Consum. 2022, 29, 141–157.

[4] Demir, A.S. Modeling and forecasting of CO2 emissions resulting from air transport with genetic algorithms: The United Kingdom case. Theor. Appl. Climatol. 2022, 150, 777–785.

[5] Wang, H.; Zhang, Z. Forecasting CO2 Emissions Using a Novel Grey Bernoulli Model: A Case of Shaanxi Province in China. Int. J. Environ. Res. Public Health 2018, 19, 1–22.

[6] Ahn, K.U.; Kim, D.W.; Cho, K.; Cho, D.; Cho, H.M.; Chae, C.U. Hybrid Model for Forecasting Indoor CO2 Concentration. Buildings 2022, 12, 1540.

[7] Iania, L.; Algieri, B.; Leccadito, A. Forecasting Total Energy's CO2 Emissions, LIDAM Discussion Paper LFIN. 2022, pp. 1–58. Available online: https://ssrn.com/abstract=4116768 (accessed on 10 October 2023).

[8] Zhong, W.; Haigh, J.D. The greenhouse effect and carbon dioxide. Weather 2013, 68, 100–105.
[9] Cook, J.; Oreskes, N.; Doran, P.T.; Anderegg, W.R.; Verheggen, B.; Maibach, E.W.; Rice, K. Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. Environ. Res. Lett. 2016, 11, 048002.
[10] Myers, K.F.; Doran, P.T.; Cook, J.; Kotcher, J.E.; Myers, T.A. Consensus revisited: Quantifying scientific agreement on climate change and climate expertise among Earth scientists 10 years later. Environ. Res. Lett. 2016, 16, 104030.
[11] Florides, G.A.; Christodoulides, P. Global warming and carbon dioxide through sciences. Environ. Int. 2009, 35, 390–401.
[12] Solomon, S.; Plattner, G.K.; Knutti, R.; Friedlingstein, P. Irreversible climate change due to carbon dioxide emissions. Proc. Natl. Acad. Sci. USA 2009, 106, 1704–1709.
[13] Lane, J. CO2 emissions and GDP. Int. J. Soc. Econ. 2011, 38, 911–918.
[14] Chang, C.L.; Ilomäki, J.; Laurila, H.; McAleer, M. Causality between CO2 emissions and stock markets. Energies 2020, 13, 2893.
[15] Sharma, S.; Zhang, M.; Gao, J.; Zhang, H.; Kota, S.H. Effect of restricted emissions during COVID-19 on air quality in India. Sci. Total. Environ. 2020, 728, 138878.
[16] Franceschi, F.; Cobo, M.; Figueredo, M. Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá Colombia, using Artificial Neural Networks, Principal Component Analysis and k-means clustering. Atmos. Pollut. Res. 2018, 9, 912–922.
[17] Qian, Y.; Sun, L.; Qiu, Q.; Tang, L.; Shang, X.; Lu, C. Analysis of CO2 drivers and emissions forecast in a typical industry-oriented county: Changxing County, China. Energies 2020, 13, 1212.
[18] Zhou, J.; Yu, X.; Guang, F.; Li, W. Analyzing and predicting CO2 emissions in China based on the LMDI and GA-SVM model. Pol. J. Environ. Stud. 2018, 27, 927–938.
[19] Sun, W.; Sun, J. Prediction of carbon dioxide emissions based on principal component analysis with regularized extreme learning machine: The case of China. Environ. Eng. Res. 2017, 22, 302–311.
[20] Safa, M.; Nejat, M.; Nuthall, P.L.; Greig, B.J. Predicting CO2 Emissions from Farm Inputs in Wheat Production using Artificial Neural Networks and Linear Regression Models. Int. J. Adv. Comput. Sci. Appl. 2016, 7, 268–274.
[21] Shabani, E.; Hayati, B.; Pishbahar, E.; Ghorbani, M.A.; Ghahremanzadeh, M. A novel approach to predict CO2 emission in the agriculture sector of Iran based on Inclusive Multiple Model. J. Clean. Prod. 2021, 279, 123708.
[22] Ahmadi, M.H.; Jashnani, H.; Chau, K.W.; Kumar, R.; Rosen, M.A. Carbon dioxide emissions prediction of five Middle Eastern countries using artificial neural networks. Energy Sources Part Recover. Util. Environ. Eff. 2019, 1–13.
[23] Liu, Z.; Ciais, P.; Deng, Z.; Davis, S.J.; Zheng, B.; Wang, Y.; Cui, D.; Zhu, B.; Dou, X.; Ke, P.; et al. Carbon Monitor, a near-real-time daily dataset of global CO 2 emission from fossil fuel and cement production. Sci. Data 2020, 7, 2052–4463.
[24] Tanania, V.; Shukla, S.; Singh, S. Time series data analysis and prediction of CO2 emissions. In Proceedings of the Confluence 2020 10th International Conference on Cloud Computing, Data Science and Engineering, Noida, India, 29–31 January 2020; pp. 665–669.
[25] Dong, F.; Yu, B.; Pan, Y. Examining the synergistic effect of CO 2 emissions on PM 2.5 emissions reduction: Evidence from China. J. Clean. Prod. 2019, 223, 759–771.
[26] Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling Climate Change with Machine Learning. ACM Computing Surveys 2022, 55, 1–96.

[27] Meehl, G.A.; Stocker, T.F.; Collins, W.D.; Friedlingstein, P.; Gaye, A.T.; Gregory, J.M.; Kitoh, A.; Knutti, R.; Murphy, J.M.; Noda, A.; et al. Global Climate Projections. In Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change; Solomon, S., Qin, D., Manning, M., Averyt, K., Marquis, M., Eds.; Cambridge University Press: Cambridge, UK, 2007; Chapter 10; p. 790.

[28] Kadam, P.; Vijayumar, S. Prediction Model: CO2 Emission Using Machine Learning. In Proceedings of the 3rd International Conference for Convergence in Technology, I2CT 2018, Pune, India, 6–8 April 2018; pp. 1–3.

[29] Li, M.; Wang, W.; De, G.; Ji, X.; Tan, Z. Forecasting carbon emissions related to energy consumption in Beijing-Tianjin-Hebei region based on grey prediction theory and extreme learning machine optimized by support vector machine algorithm. Energies 2018, 11, 2475.

[30] Kunda, D.; Phiri, H. An Approach for Predicting CO2 Emissions using Data Mining Techniques. Int. J. Comput. Appl. 2017, 172, 7–10.

[31] Libao, Y.; Tingting, Y.; Jielian, Z.; Guicai, L.; Yanfen, L.; Xiaoqian, M. Prediction of CO2 Emissions Based on Multiple Linear Regression Analysis. Energy Procedia 2017, 105, 4222–4228.

[32] Saleh, C.; Dzakiyullah, N.R.; Nugroho, J.B. Carbon dioxide emission prediction using support vector machine. IOP Conf. Ser. Mater. Sci. Eng. 2016, 114, 012148.

[33] Deniz, S.; Gökçen, H.; Nakhaeizadeh, G. Application of Data Mining Methods for Analyzing of the Fuel Consumption and Emission Levels. Int. J. Eng. Sci. Technol. 2016, 5, 377–389.

[34] Saleh, C.; Chairdino Leuveano, R.A.; Ab Rahman, M.N.; Md Deros, B.; Dzakiyullah, N.R. Prediction of CO2 emissions using an artificial neural network: The case of the sugar industry. Adv. Sci. Lett. 2015, 21, 3079–3083.

[35] Chen, Y.; Li, B.; Li, Z.; Shi, X. Quantitatively evaluating the effects of CO2 emission on temperature rise. Quat. Int. 2014, 336, 171–175.