Prediction and Forecasting of CO2 Emissions with EDA using KNN and Random Forest

K. Vaishnavi, K. Asha, M.Raghu Vamshi.
Dr. Praveen Talari, Associate Professor, Department of Computer Science and Engineering.
Vignana Bharathi Institute of Technology- Hyderabad, Telangana, India.
Email: 21p61a05d8@vbithyd.ac.in, 21p61a05d9@vbithyd.ac.in, 21p61a05f5@vbithyd.ac.in.

ABSTRACT

CO2 emissions are one of the major sources of global warming and have proven challenges to be a significant issue in sustainable development. Therefore, it is very important to accurately predict and forecast CO2 emissions to develop effective mitigation strategies and policies. The authors apply ML models, such as KNN and RF, and combine EDA as a basis for predicting CO2 emissions. It uses a tough dataset consisting of the most critical economic factors and environmental determinants that impact the overall results. The respective model's performance is assessed using metrics like MAE, MSE, and R^2 scores. Results reveal that Random Forest surpasses KNN in the sense of accuracy and reliability, thereby providing a potential avenue for emission forecasting. This study offers useful inputs for policymakers and environmentalists in seeking sustainable development and mitigation from climate change.

Keywords: CO2 Emissions, Machine Learning, Exploratory Data Analysis, Random Forest, Forecasting Models.

1. INTRODUCTION

Global warming has been identified as the most serious threat of our time due to its sheer contribution through mainly CO2 emissions. Human activities, such as industrial processes, deforestation, and burning fossil fuels, have released higher levels of atmospheric carbon dioxide into the atmosphere, resulting in extreme consequences for the environment, including extreme weather conditions and sea level rise.

It focuses on studying the potential that these machine learning models have to predict and forecast CO2 emissions. While conventional statistical methods have been useful up till now, they are inefficient in handling large complex data sets. Machine learning offers a better promise: it can process vast amounts of data and uncover unknown patterns. The integration of EDA into the research is aimed at increasing knowledge and facilitating better accuracy in forecasts to be utilized by policymakers in developing efficient policies for decreasing emission levels.

2. LITERATURE SURVEY

In this chapter, a review of papers published on previous studies made about CO2 emissions prediction, the relationship between industrial structure, energy consumption, and CO2 emissions, and machine learning application (ML) for CO2 emissions prediction, focusing more on China. CO2 emissions play an important role in global warming; the global peak time emerged recently. Being a developing country, China has to find ways to solve the problem of balancing its high economic growth and environmental protection. Having already become the largest emitter in the world, more than 70% of emissions come from electricity generation and industry. Reducing emissions in these sectors would require a transition to clean energy and an adjustment of industrial structure. In the recent past, prediction models involving grey modeling, scenario analysis, and sometimes econometric techniques have been used for forecasting CO2 emissions. The regional-level emission forecasting uses GM(1, N) and time-delay grey models. Scenario analysis is another method of forecasting long-term emissions. This involves population, GDP, and energy consumption. These methods, however, have to cope with assumptions about data smoothness and scenario validity.

Recent studies on this topic have shown that industrial structure economic growth and renewable energy consumption are the most important factors leading to CO2 emissions.

Industrial structure adjustment, in particular from manufacturing-oriented toward service-based economies and clean energy transition, was of great importance for emission mitigation. Using LMDI and STIRPAT econometric models and panel data analysis, interactions are pointed out between urbanization, industrial structure, and energy sources. Renewable energy is beneficial in the long term but, meanwhile, has mixed impacts on emissions as they also depend on energy intensity and economic factors. Machine learning models have recently emerged as attention on these models as they can handle complex, nonlinear systems. Recently, approaches have started working with multivariate time-series models and use feature selection techniques like LASSO and PCA, and various predictive algorithms such as SVR, XGBoost, and LSTM. These methods get the most important drivers of emissions like energy consumption and industrial activity and therefore provide better accuracy for making the

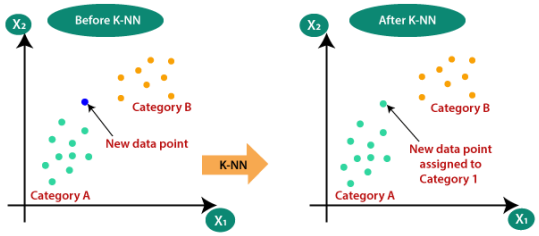
predictions. Three pertinent gaps in the research were also spotted.

First, grey models use smooth data, while scenario-based approaches use subjective inputs; this reduces forecasting accuracy.

There is a lack of research on the integration of industrial structure, renewable energy, and CO2 emissions into one framework. Further, there are not many works using ML for multivariate prediction of China's emissions. This paper fills in these gaps by presenting a multivariate forecasting model that integrates industrial structure with renewable energy. This approach is shallow learning; statistical assumptions are relaxed, and time series are transformed into a framework suitable for a better improvement in predictive capability and practicality for a supervised setting.

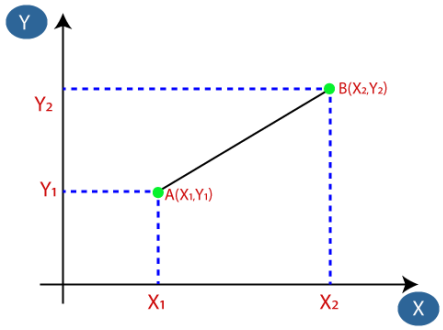
3. EXISTING SYSTEM

The current system is based on the KNearest Neighbors (KNN) algorithm, which is a well-established and widely applied supervised machine learning approach. KNN performs very well in classification or regression in cases when most of the data points have similar values and fall into the same category. It operates as an instance-based learning algorithm in which an explicit training process is not needed. Instead, it simply memorizes the whole training dataset and applies it during the time of prediction by comparing of similarity of data points with each other through distance metrics.



Working of KNN

KNN determines the distance between data points using metrics like Euclidean, Manhattan, and Minkowski distances. For classification, the label assigned is most frequently seen amongst the nearest neighbors, whereas for regression, it computes the mean or median value of the labels found on neighbors.



Challenges and Limitations

The selection of the hyperparameter 'k' is amongst the performance-determining factors. Small values of 'k' may make a model noisy and sensitive towards outliers while large values could smoothen the overlocal patterns and diminish the accuracies. In addition, KNN

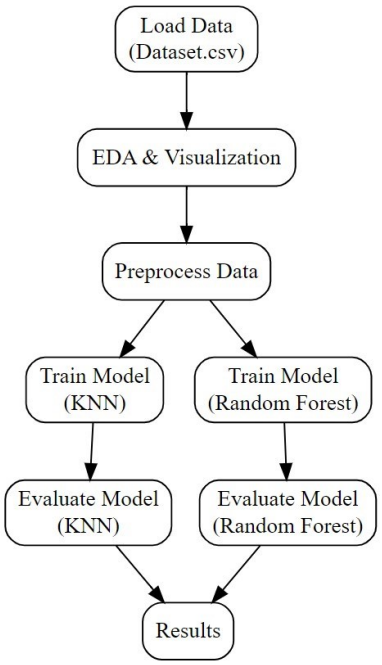
is computationally intensive as it stores and compares an entire dataset to make predictions, which makes it a less competitive approach for large-scale problems. Moreover, the distance metric on which the algorithm relies makes feature scaling an important step in making sure that all features contribute proportionally to the result.

Variants and Optimizations

To enhance its performance, several variants of KNN have been advanced. Weighted KNN uses more weights on the neighbors that are closer; it enhances the model's sensitivity to the data points nearby. Feature scaling, of which standardization is an example, is fundamental to normalization, thereby avoiding dominance by features of larger scales. Data structures like KD-Tree and Ball-Tree are applied to speed up nearest-neighbor searches against massive datasets. Despite these, KNNs' performance often depends on data preprocessing quality and careful choice of hyperparameters.

PROPOSED SYSTEM

The proposed system extends the traditional KNN framework by integrating advanced preprocessing techniques, feature engineering, and optimized modeling steps that enhance predictive accuracy, especially for CO_2 emissions forecasting. It also includes supplementary models like Random Forest (RF) to deal with the weakness of KNN and enhance robustness.



Exploratory Data Analysis (EDA)

The idea of the proposed system starts with an extensive analysis of the dataset; it observes the presence of patterns, trends, and even relationships between variables that exist in such a dataset. Visualizations involved are histograms, scatter plots, and even a correlation heatmap to study the distribution and interaction of features. Outliers and missing values are identified and managed during this phase for optimal data quality.

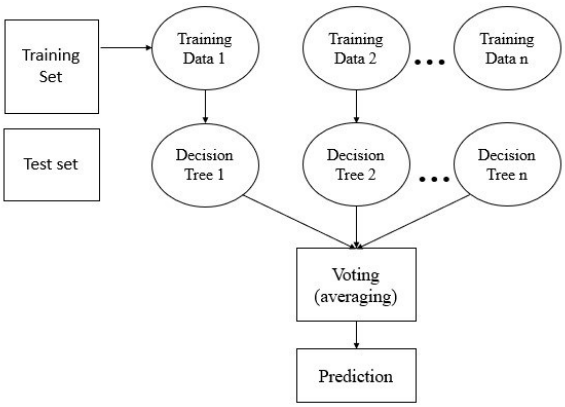
Advanced Preprocessing Techniques

Effective preprocessing is the backbone of the proposed system. Feature selection methods identify the most relevant variables for predicting CO2 emissions, and categorical variables are encoded using one-hot encoding to enable their smooth use in machine learning models. Numeric features are standardized techniques to bring them onto a uniform scale as per their dimension. The dataset is divided into subsets training and testing subsets to assess the performance of the model.

Improved KNN Model The presented KNN model extends the baseline by incorporating feature engineering to create new variables that are more informative. Hyperparameter tuning is done through grid search or cross-validation to find the best value of neighbors ('k') and other configurations. The improved KNN model is then trained and compared to the baseline to validate its effectiveness in predicting CO2 emissions.

Feasibility of Integration with Random Forest (RF)

To further improve the precision level, the Random Forest algorithm is integrated into the system. RF utilizes ensemble learning. It constructs multiple decision trees on different-sized subsets of examples and combines them either by majority voting or averaging to generate the output. This can reduce overfitting, and improve stability in the model, and it can, thus, cope better with missing data compared to KNN. Moreover, it computes feature importance scores that yield information on the relative influence of variables.



Performance Evaluation

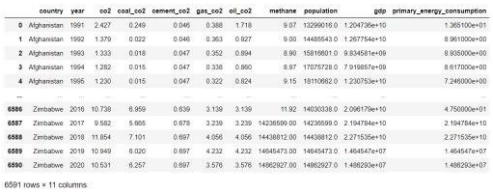
Both the current and proposed models use these metrics: MAE, MSE, RMSE, and R^2 scores to assess them. A comparison of these scores leads to the conclusion that the proposed system enhances the accuracy and predictability of CO2 emissions forecasting. The inclusion of RF adds further validation of the results, which ensures the robust nature of the approach. Advantages of the Proposed System

The proposed system bridges the gaps within the existing KNN framework with comprehensive data preprocessing, optimized modeling, and ensemble techniques. These allow for closer-

toactual, scalable, and reliable prediction for eventual deployment in real-world environmental forecasting applications.

4. RESULTS AND DISCUSSION

The CO2 emission prediction dataset has features like population, GDP, and energy consumption as shown in Fig. 7.1. The histogram for the emissions of CO2 (Fig. 7.2) highlights the different emission values along with the help of a KDE curve providing a smoothed version. The correlation heatmap in Fig. 7.3 depicts the interaction between all features visually and, therefore, signifies feature interactions. The pair plot in Fig. 7.4 further checks the interactions through scatter plots and identifies patterns and trends.



	country	year	co2	coal_co2	cement_co2	gas_co2	oil_co2	methane	population	gdp	primary_energy_consumption
0	Afghanistan	1991	2.427	0.249	0.046	0.388	1.718	9.07	13299016.0	1.204736e+10	1.365100e+01
1	Afghanistan	1992	1.379	0.022	0.046	0.363	0.927	9.00	14486543.0	1.267754e+10	8.961000e+00
2	Afghanistan	1993	1.333	0.018	0.047	0.352	0.894	8.90	15816601.0	9.834581e+09	8.935000e+00
3	Afghanistan	1994	1.282	0.015	0.047	0.338	0.860	8.97	17075728.0	7.919857e+09	8.617000e+00
4	Afghanistan	1995	1.230	0.015	0.047	0.322	0.824	9.15	18110662.0	1.230753e+10	7.246000e+00
...	...	...	...	...	...	...	...	...	...	...	...
6886	Zimbabwe	2016	10.738	6.959	0.839	3.139	3.139	11.92	14030338.0	2.096179e+10	4.750000e+01
6887	Zimbabwe	2017	9.582	5.966	0.878	3.239	3.239	14236599.00	14236599.0	2.194784e+10	2.194784e+10
6888	Zimbabwe	2018	11.854	7.101	0.897	4.056	4.056	14438812.00	14438812.0	2.271535e+10	2.271535e+10
6889	Zimbabwe	2019	10.949	6.020	0.897	4.232	4.232	14645473.00	14645473.0	1.464547e+07	1.464547e+07
6890	Zimbabwe	2020	10.531	6.257	0.897	3.576	3.576	14862927.00	14862927.0	1.486293e+07	1.486293e+07

6891 rows × 11 columns

Figure 7.1: sample dataset used for CO2 emission

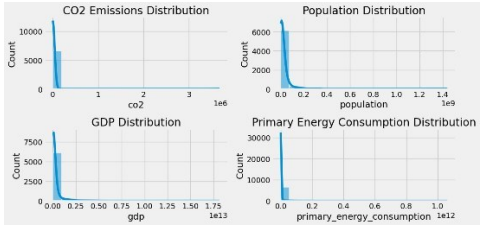


Figure 7.2: This subplot displays the distribution of CO2 emissions

After preprocessing, the data is cleaned and transformed - it is treated for missing values, scaled, and features selected (Fig. 7.5). Important features for prediction are shown (Fig. 7.6), and the target variable distribution, namely CO2 emission, postpreprocessing (Fig. 7.7).

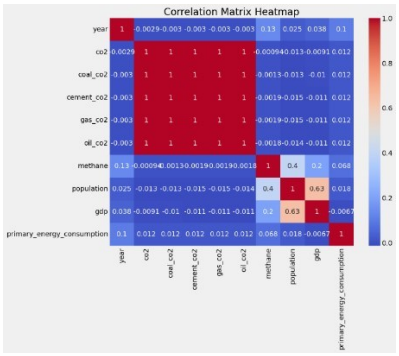


Figure 7.3: Heatmap of correlation of each variable

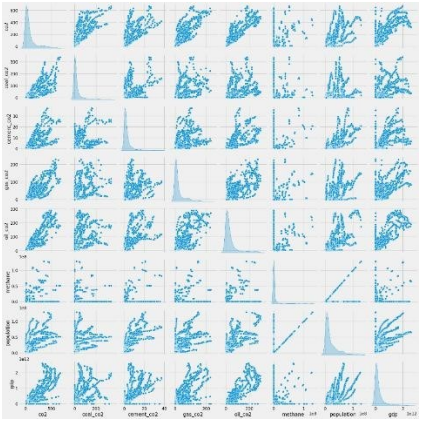


Figure 7.4: pair plot of features

	country	year	co2	methane	ccgo	gdp_per_capita
0	Afghanistan	1991	2.427	9.07	2.401	905.883692
1	Afghanistan	1992	1.379	9.00	1.358	875.185599
2	Afghanistan	1993	1.333	8.90	1.311	621.788531
3	Afghanistan	1994	1.282	8.97	1.260	463.807877
4	Afghanistan	1995	1.230	9.15	1.208	679.573506

Figure 7.5: dataset after preprocessing used for CO2 emission

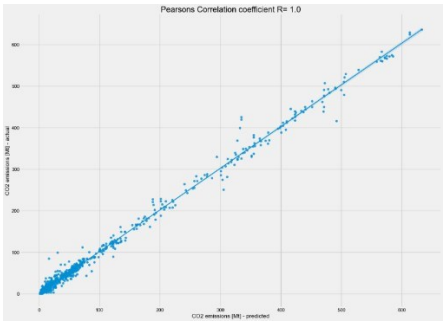


Figure 7.8: prediction results using KNN

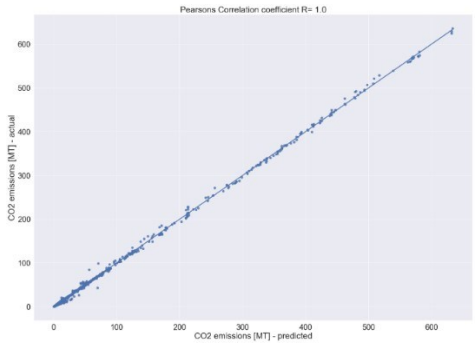


Figure 7.9: prediction results using Random Forest Classifier

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

Figure 7.10: Performance metrics of KNN & Random Forest classifier

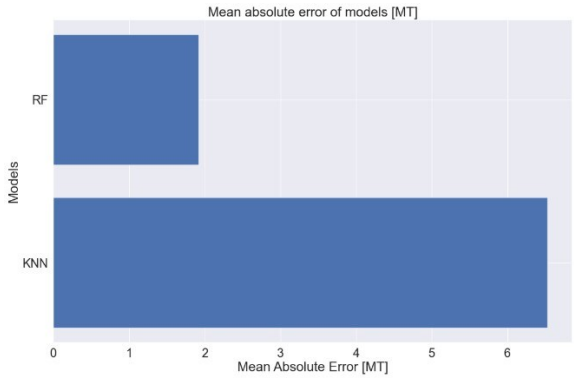


Figure 7.11: Bar plot of Mean absolute error of KNN &Random Forest Classifier

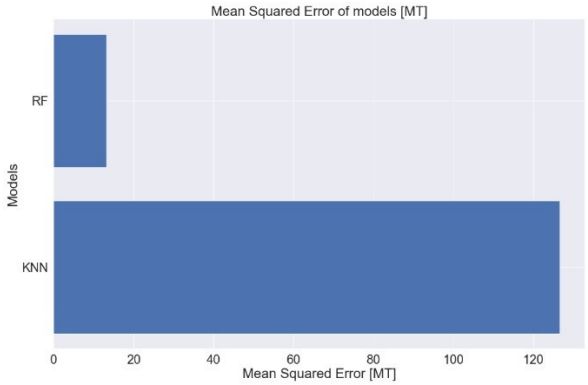


Figure 7.12: Bar plot of Mean Squared error of KNN &Random Forest Classifier

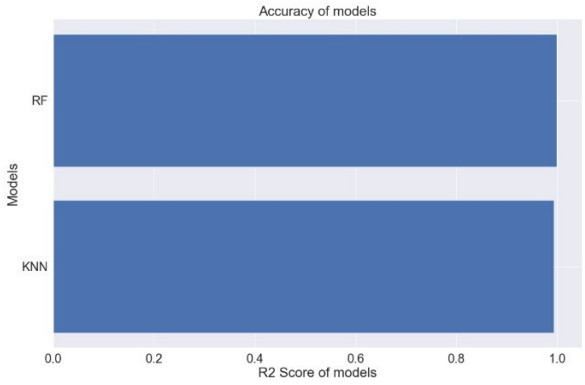


Figure 7.13: Bar plot of R2 Score of KNN &Random Forest Classifier

Different prediction results are compared to actual CO2 emission using KNN (Fig. 7.8), and RFC (Fig. 7.9) provides another alternative predictive approach. The performance metrics are the comparison of both models in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) scores, which are summarized in Fig. 7.10. Bar plots of MAE (Fig. 7.11), MSE (Fig. 7.12), and R^2 (Fig. 7.13) emphasize the comparative accuracy and goodness-of-fit achieved by RFC in comparison with KNN.

CONCLUSION

This research highlights the effectiveness of integrating machine learning models and exploratory data analysis (EDA) for predicting and forecasting CO2 emissions, addressing critical challenges posed by climate change. Machine learning proves adept at analyzing complex datasets, and uncovering patterns and relationships often missed by traditional methods. EDA complements this by offering deeper insights into data, identifying key features, and detecting outliers. Collectively, these methods yield precise and reliable CO2 emission forecasts, empowering policymakers to develop effective plans for reducing emissions and achieving sustainability. It contributes not only to the advancement of scientific knowledge but also offers practical applications in resource optimization, renewable energy promotion, and mitigation of global warming.

FUTURE SCOPE

Future advancements include refining machine learning models through deep learning and ensemble methods to improve prediction accuracy. Integrating real-time data and satellite imagery can enhance model responsiveness to dynamic environmental conditions. Including socioeconomic and policy-related variables can deepen insights into emission drivers, aiding in crafting targeted climate policies.

Collaboration among researchers, governments, and environmental organizations will be pivotal for data collection and policy implementation, driving impactful solutions to combat climate change.

REFERENCES

1. Intergovernmental Panel on Climate Change (IPCC), Sixth Assessment Report, 2022.
2. Song, M., et al. *Share Green Growth: Regional Evaluation of Green Output Performance in China*. Int. J. Prod. Econ., 2020.
3. Dong, B., et al. *Changes in Automobile

Energy Consumption during Urbanization*. Energy Policy, 2019.

4. Sun, T., Ren, Y. *Dynamic Spatial Spillover Effect of New Energy Vehicle Industry Policies*. Energy Policy, 2022.
5. Wang, M., Zhang, F. *Using Extended

LMDI Approach for Industrial CO2 Emissions Assessment*. Energy Econ., 2018.