Predicting CO2 Emission from Traffic Vehicles for Sustainable and Smart Environment Using Machine Learning Model

Amin Ocin[1][2021−3−60−135] Tasnova Haque Mazumder[2][2021−3−60−235] Dilruba Akter[3][2021−3−60−077] Akash Ahmed[4][2021−3−60−242]

[1] East West University, Dhaka
[2] lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. Protecting the environment is the most important thing we need to do to save the future of our upcoming generation. But CO2 emission is something that is going beyond our control. So, It is our responsibility to reduce it. To do so, we first have to detect the CO2 emission. And we rely on machine learning on for such detection nowadays. Our study shows a comprehensive comparison of several machine learning model like Linear Regression, Random Forest, Gradient Boosting, and Decision Tree on a dataset which focuses on CO2 emission depending on several features of vehicles. Among these models Random Forest out perform all the other models with accuracy 99.66% which is the highest achieved accuracy on this dataset so far. We performed VIF data processing to achieve this accuracy. We, also visualized the outcomes of the models using explainable AI like both SHAP and LIME. This helped us to identify the important features that contributes to the prediction. This study shows the most recent and most accurate result in this field and dataset both.

Keywords: Machine Learning · Random Forest Regressor · Emission.

1 Introduction

One of the most concerning issues in the world is the Global warming. Global warming causes rising temperature world-wide. The main cause of the global warming is the Green House Gas(GHGs),and CO2 is the main component of GHG[1]. The aftereffects and disadvantages of CO2 rising are uncountable, some of the mentionable damages are, Extreme weather (hurricane, wildfire etc.), Melting ice which is causing the rise in sea level. CO2 also hampers the agriculture and food supply, human health by causing Air-Pollution. Sustainability, green computing can be solution to this problem of increased carbon foot prints. Day by day the emission of CO2 is getting uncontrollable. In 2023 the emission of CO2 was 40.6 billion tons increasing up to 4%. [2][3]. Transportation vehicle

causes a great roll in increasing the CO2 emission. Among the transportation vehicles the road vehicle produces the most CO2. The number of road vehicle is also increasing rapidly[4] The amount of emission mostly depends on how much fuel it uses and how many kilometers it drives. The solution of this issue needs to be studied as which feature of which road vehicle is causing how much emission of CO2. Machine learning can focus on system to learn from data and predict pattern of the data. Machine Learning plays a significant role in monitoring CO2 emission by data driven analysis and predicting CO2 emission for vehicles [3]. Machine learning has various models to address this problem. This study aims to how well machine learning models work on predicting the co2 emission. These Machine learning models are trained to predict the co2 emission by each road vehicles based on the features of those vehicles. The models are Random Forest Regressor, Linear Regressor, Decision Tree Regressor, Gradient Boosting Regressor. Amongst all the models, The Random Forest regressor outperformed all the other models having a R2 score of 0.9964. This research is intended to identify the model that performed the best at prediction.

The rest of this paper is structured as follows: In Section 2, we'll take a look at previous research in the area of CO2 emission prediction and how Machine Learning has been utilized in environmental modeling. Section 3 presents the methodology, detailing the dataset and ML techniques employed and the evaluation criteria used to assess model performance. Section 4 discusses the results obtained from the experiments and provides insights into the relationship between vehicle attributes and CO2 emissions.Section 5 talk about the how the expalinable AI in implemented.

2 Related Work

Due to the globalization for last 2 decades, Co2 emission has increased to an alarming level. Study shows that both in the short run and long run, negative globalization and economic growth shocks positively and negatively influence CO2 emissions [11] . Six largest countries of the world United states of America, European Union, India, Russia and Japan are responsible for more than 67% of the co2 emission[12]. We know that population aggravates co2 emission and the trend of co2 emission construct an inverted-U shape[13]. Linear regression is a well known statistical model for understanding the relationships between the variables [14]. Random forest are also knowns for it's ease of use and ability to prevent over fitting [15]. Gradient boosting has shown exception capability in modeling complex systems and extremely powerful in terms of accurate prediction, interpretability and classification versatility [16]. Decision tree is known not only for it's widely usability but also for its robustness and interpretability. This statistical models are performing really good for last several years and have very less laggings as well. Machine learning is been using lately for predicting co2 emission. And it is hence a required step to identify co2 emission since it has reached to an alarming level. A study was performed for predicting co2 emission using several model. Among them linear regression and random forest based

LSTM model outperformed the others [17]. Other models like ARIMA (Autoregressive Integrated Moving Average), Shallow Neural Networks, and Deep Neural Networks are also being used for co2 emission prediction task [18] . For transportation related co2 emission these model are also being used and their performance were also between (80% to 90%) [19]. Co2 emission has emerged as a major concert in this century. Linear Regression, Random Forest Regression are some of the most frequently used model that are being used for co2 emission prediction in recent years.

3 Methodology

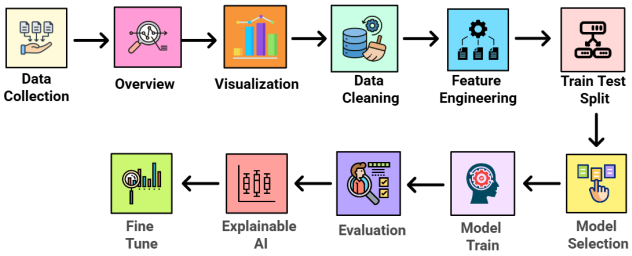


Fig. 1: Methodology Framework for CO2 Emission prediction

3.1 Data Collection

we have used a dataset from kaggle named, "CO2 Emission by Vehicles". This dataset has 7385 data of automobiles. It has valueable insights adding attributes. These attributes explains the technical and mechanical aspects of the vehicle, such as car model name, size, fuel type. Also, Fuel consumption measurements (L/100 km) for both city and highway. This dataset is well maintained and relevant to our study.

3.2 Dataset Overview

The dataset used in this study contains Ten feature columns and one target column. The features are the independent variable and the target column is the dependent variable. There are Five categorical feature and Five numerical features. There are 7385 distinctive data. Table 1 Represents the structure of the dataset.The Fig.3 shows that the data of the dependent variable is normally distributed making the result more reliable and interpretable. The bell curve shows no outliers in the column.

The categorical features are:

- **Make**
- **Model**
- **Vehicle Class**
- **Transmission**
- **Fuel Type**

The numerical features are:

- **Engine Size (L)**
- **Cylinders**
- **Fuel Consumption City (L/100 km)**
- **Fuel Consumption Hwy (L/100 km)**
- **Fuel Consumption Comb (L/100 km)**
- **Fuel Consumption Comb (mpg)**

The target column is **CO2 Emissions (g/km)**.

Table 1: Vehicle Data and Emissions

Make & Model	Class	Engine (L)	City/Hwy/Comb (L/100 km)	MPG	CO2 (g/km)
ACURA ILX	Compact	2.0	9.9 / 6.7 / 8.5	33	196
ACURA ILX	Compact	2.4	11.2 / 7.7 / 9.6	29	221
ACURA ILX Hybrid	Compact	1.5	6.0 / 5.8 / 5.9	48	136
ACURA MDX 4WD	SUV-Small	3.5	12.7 / 9.1 / 11.1	25	255
ACURA RDX AWD	SUV-Small	3.5	12.1 / 8.7 / 10.6	27	244

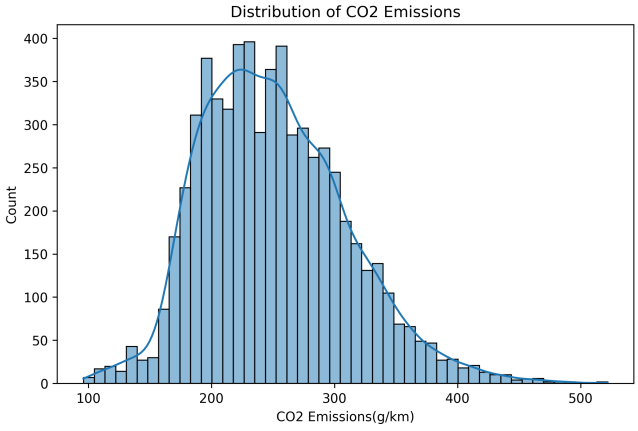


Fig. 2: Normal distribution of CO_2 emission(g/km)

3.3 Exploratory Data Analysis

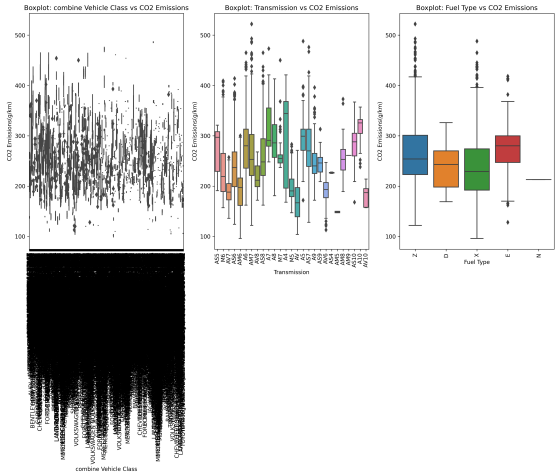


Fig. 3: Box Plot of the Categorical data

To get a proper understanding about the dataset we have done EDA. There are 7385 distinct data but among them there are 1103 duplicate data. There is no null value. After deleting duplicate data there are 6282 data.The box plot in the Fig.2 shows the data distribution of the categorical value with the target column. These Box plots of the categorical features show that the data transmission column and the fuel type column have some outliers.The box plot of the combine vehicle class shows very poor distribution with CO2 emission.In this study that dependent variable is CO2 Emissions (g/km). To have better understanding of the numerical data, we have seen the scatter plot in the Fig.4. The scatter plot for all the numeric feature shows the relation with the dependent column CO2 Emissions (g/km).

The engine size versus CO2 scatter plot describe the bigger the engine size the more CO2 is produced. Another numeric column Fuel consumption City shows linear relationship with the target variable. To train the model we need to select the most useful and related features. Correlation matrix shows the relation between the features. This helps to select the highly efficient features. In this dataset, the highly correlated features are Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Engine Size(L), Cylinders and least correlated is Fuel Consumption Comb (mpg), shown in Fig. . As, the highly correlated columns are of same type, a combine feature is made with the highly correlated features by taking the average of these feature. the numeric values are scaled with standard scalar and the categorical values are encoded with the label encoder.

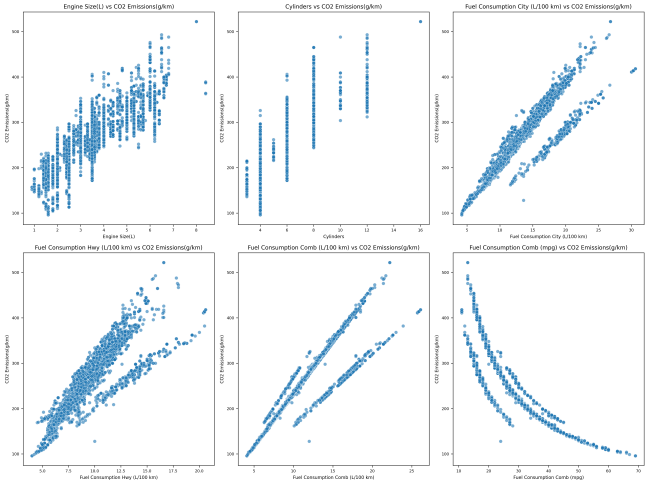


Fig. 4: Scatter Plot

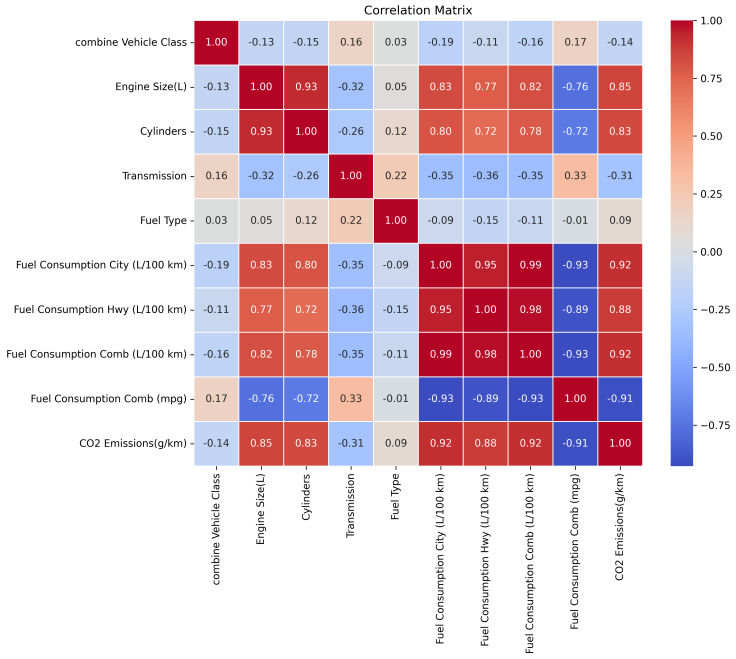


Fig. 5: Correlation Heatmap of the Features

3.4 Model Train

Linear Regression The Linear Regression model is a fundamental technic for predicting a dependent variable by fitting a linear relationship between one or more independent variables. In our research, we mainly focus on estimating CO2 emissions using machine learning approaches. The goal is to establish a linear relationship between the independent variables such as engine size, fuel efficiency, and vehicle class and the dependent variable, CO2 emissions. By reducing the variance between the predicted and actual CO2 emission, the method determines the coefficient of the linear equation that best fits the dataset [4]. In our study, this model is train with eighty percent of the data. this model does not show any impressive resut, as the R^2 score is is low and the loss is high.

Random Forest Regressor Random forest is an ensemble learning technic that produce multiple decision tree and combining their output to improve accuracy and reduce overfitting. [4] It is a modern decision tree that performs more accurate than bagged decision trees.[6] Bagging is an idea for minimizing the variance to ignore overfitting. [7] The bagging trees consistently produces similar type of output as the heavyweight variables will always attempt to remain on the top. So, the variance will be same. They will not have a noticeable change from the average of different trees due to similarities. To solve this problem, Random Forest is used.[20] To generate a final prediction using random forest, it is required the combination of prediction produced by each sub tree as shown in Fig.5 [9]. In this study, The Random Forest Regressor showed the best result, AFter k-fold cross validation with the Random Forest Regressor model to train and evaluate the model the loss has decreased. The model does not show overffitng behaviour.

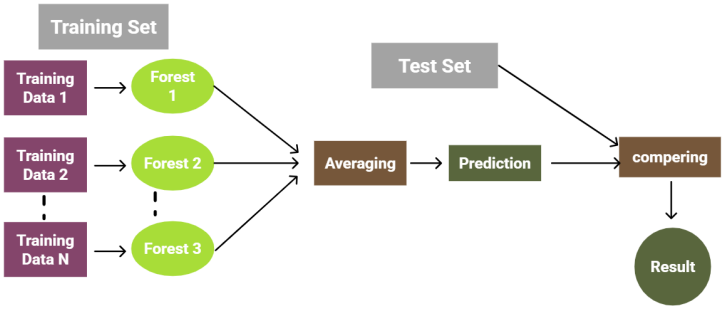


Fig. 6: Random Forest Regressor Model Architecture

Decision Tree Regressor In machine learning decision trees are essential for serving both classification and regression. It works by splitting the data

into smaller subsets based on specific features.[10] It makes a tree like structure where root node is the starting point that represents the entire dataset. internal Nodes are Points where decisions are made based on the input features.The terminal nodes at the end of branches that represent final outcomes or prediction called leaf nodes. decision tree model often shows overfitting result in predicting model.This model show over fitting result in this study too.

Gradient Boosting Regressor Gradient boosting is one of the most well known machine learning algorithms for tabular datasets. It is powerful to find nonlinear relationship between target and features of a model.[21] This difference is called residual .By mapping features to residuals ,gradient boosting regression trains a weak model. This residual predicted by a weak model is added to the current model input and thus this process moves the model towards the correct target. Repeating this step improves the overall model performance.[22] This model is train with customized parameters, such as, estimators=500, learning rate=0.01, validation fraction=0.2.

4 Result and Evaluation

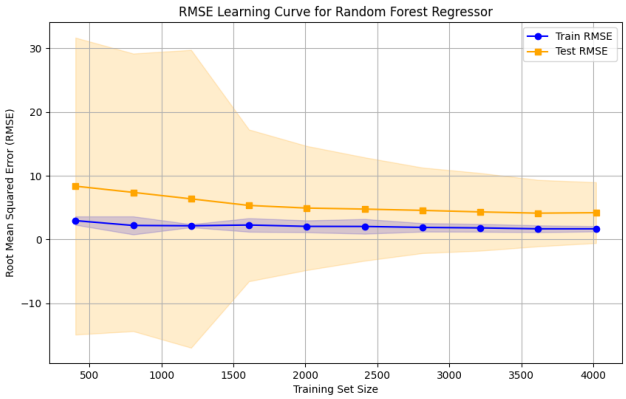
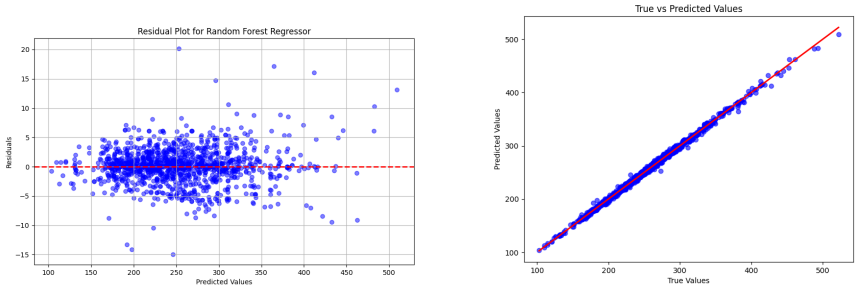


Fig. 7: Train Test Learning curve Random Forest Regressor

The result of our study is shown in this segment. In this study Four machine learning models are tested in the dataset. The evaluation metrics used in this study are R^2 value. The metrics RMSE, MAE represents the loss of the function.[7] The best performing model is the Random forest regressor model. It had $\mathbf{R^2}$ Score: 0.9964. In Fig.6 the learning curve of the Random Forest Regressor shows that the train loss and the test loss. The difference between these two

is very low, so, the model does not over fitt. The residual plot at Fig.7 shows most of the point are clustered around the zero, which means the model is a good fit and the model is capturing the pattern effectively.[24] The True vs Predict plot of Random Forest Regressor shows that all the points are tightly clustered around the diagonal line. The model is performing well, and not overfitting. The



(a) Residual Plot (b) True vs Predicted Plot

Fig. 8: True VS Predict and Residual Plot for the best performing Model

lowest performing model is the Linear Regressor with a R^2 Score of 0.9043. The table 2 shows the result of the four models.

Table 2: Performance Comparison of Trained Models

Model	**R^2 Score**	**MAE**	**RMSE**
Linear Regression	0.9043	12.1454	18.5845
Decision Tree Regressor	0.9950	0.0361	0.0714
Random Forest Regressor	0.996	2.0882	3.0389
Gradient Boosting Regressor	0.9924	0.0565	0.0882

5 Explainable AI

5.1 SHAP

Shap (Shapley Additive exPlanations) is a popular method to identify the importance of each feature in predicting the result. The Fig.8 shows the SHap scatter scatter summary plot for the features.[23] in our study, the best performing model is the Random Forest Regressor, from the Fig 6 it is defiined that the average fuel consumption named feature has the highest impact in predicting the co2 emission.

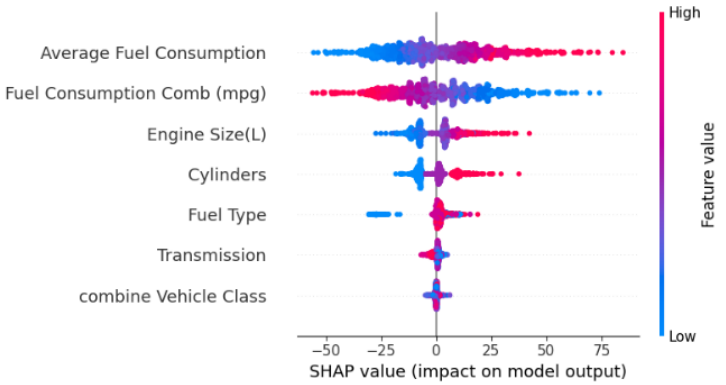


Fig. 9: SHAP plot of carbon emissions based on Random Forest Regressor

5.2 LIME

LIME (Local Interpretable Model-agnostic Explanations) is a method to expalin distimct prediction of a model. In fig 9 it shows that the feature named fuel consumption comb impact the resut negatively and the feature average fuel impacts positively.

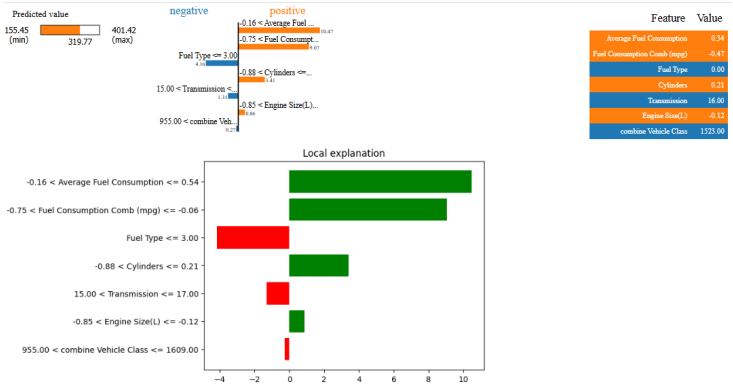


Fig. 10: Lime plot for understanding the distinctive prediction

6 Conclusion

This study especially focused on detecting the amount of CO2 emission by various types of automobile vehicles. We tried to predict the CO2 emission by a vehicle depending on the features of the vehicle, where it most likely moves

around, and the type of fuel it uses. We have used 4 machine learning models to predict the CO2 emission and compared their results. In which the model **Random Forest Regressor** showed outstanding performance with an R^2 score of 0.9964. Though the other models showed slightly overfitting behavior, the Random forest Regressor did not show any overfitting behavior as the train and test results converged at some point. We also tried to explain the results using explainable AI such as SHAP and Lime to get the most involved feature that has the most contribution to the prediction. Our results outcast most of the previous works on this dataset as well as this specific field. CO2 emission has been increasing drastically for the last 2 decades. It is high time some serious measures should be taken to avoid future CO2 emissions as little as possible. And to achieve this the first step is to detect the emissions. And our study would speed up this process. [**?**].

References

1. Costantini, L., Laio, F., Mariani, M. S., Ridolfi, L., Sciarra, C. (2024). Forecasting national CO2 emissions worldwide. *Scientific Reports*, **14**(1). DOI: 10.1038/s41598-024-73060-0
2. World Meteorological Organization: Record carbon emissions highlight urgency of Global Greenhouse Gas Watch. (2024, November 20). Available at: `https://wmo.int/media/news/record-carbon-emissions-highlight-urgency-of-global-greenhouse-gas-watch`
3. Statista: Topic: Transportation emissions worldwide. (2025, January 23). Available at: `https://www.statista.com/topics/7476/transportation-emissions-worldwide/`
4. Predictive Modeling of Vehicle CO2 Emissions Using Machine Learning Techniques: A Comprehensive Analysis of Automotive Attributes. (2023, November 1). Available at: `https://ieeexplore.ieee.org/document/10390183/`
5. R, D., Joy, H. K., R, S., A, E. A. J., Vanusha, V.: Machine learning and deep learning analysis of vehicle carbon footprint. *International Journal of Environmental Impacts* **7**(2), 287–292 (2024b). `https://doi.org/10.18280/ijei.070213`
6. Breiman, L.: CO2 emission. *Machine Learning* **45**(1), 5–32 (2001). `https://doi.org/10.1023/a:1010933404324`
7. Udoh, J., Lu, J., Xu, Q.: Application of machine learning to predict CO2 emissions in Light-Duty vehicles. *Sensors* **24**(24), 8219 (2024). `https://doi.org/10.3390/s24248219`
8. CO2 Emission Prediction. (n.d.). CO2 Emission Prediction and Identification of Relevant Factors to the Emission Based on Machine Learning Analysis: A Study in Bangladesh. Available at:Available at: PubMed
9. Tawsif, F. M., Mostafa, M. J. I., Hossain, B. M.: CO2 Emission Prediction and Identification of Relevant Factors to the Emission based on Machine Learning Analysis: A Study in Bangladesh. (2021, May 6). Available at: `https://www.ijcaonline.org/archives/volume183/number5/31922-2021921327/`
10. Redirect notice: Decision trees—How they work and practical examples. (n.d.). Available at: `https://www.google.com/amp/s/keylabs.ai/blog/decision-trees-how-they-work-and-practical-examples/amp/`

11. Rehman, A., Alam, M. M., Ozturk, I., Alvarado, R., Murshed, M., Işık, C., Ma, H.: Globalization and renewable energy use: how are they contributing to upsurge the CO2 emissions? A global perspective. *Environmental Science and Pollution Research* **30**(4), 9699–9712 (2023). https://doi.org/10.1007/s11356-022-22775-6
12. CO2 Emissions and Causal Relationships. (n.d.). CO2 Emissions and Causal Relationships in the Six Largest World Emitters [Preprint]. Available at: ScienceDirect
13. Lee, C.-C., Zhao, Y.-N.: Heterogeneity analysis of factors influencing CO2 emissions: The role of human capital, urbanization, and FDI. *Renewable and Sustainable Energy Reviews* **185**, 113644 (2023). https://doi.org/10.1016/j.rser.2023.113644
14. Roustaei, N.: Application and interpretation of linear-regression analysis. *Med Hypothesis Discov Innov Ophthalmol* **13**(3), 151-159 (2024, Oct 14). https://doi.org/10.51329/mehdiophthal1506, PMID: 39507810, PMCID: PMC11537238
15. Shastry, K., Aditya, and Sheik Abdul Sattar: Logistic random forest boosting technique for Alzheimer's diagnosis. *International Journal of Information Technology* **15**(3), 1719–1731 (2023).
16. Cicek, V., Orhan, A. L., Saylik, F., Sharma, V., Tur, Y., Erdem, A., Babaoglu, M., Ayten, O., Taslicukur, S., Oz, A., Uzun, M., Keser, N., Hayiroglu, M. I., Cinar, T., Bagci, U.: Predicting Short-Term Mortality in Patients With Acute Pulmonary Embolism With Deep Learning. *Circ J.* (2024, Nov 30). https://doi.org/10.1253/circj.CJ-24-0630, Epub ahead of print, PMID: 39617426.
17. Kumari, S., Singh, S. K.: Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research* **30**(55), 116601–116616 (2022). https://doi.org/10.1007/s11356-022-21723-8
18. Machine learning approaches for predictions of CO2 emissions in the building sector. (n.d.). *ScienceDirect* [Preprint]. Available at: ScienceDirect
19. ScienceDirect. (2022). Forecasting of Transportation-related Energy Demand and CO2 Emissions in Turkey With Different Machine Learning Algorithms [Preprint]. Available at: ScienceDirect
20. Prediction Model: CO2 Emission Using Machine Learning. (2018, April 1). Available at: https://ieeexplore.ieee.org/abstract/document/8529498/
21. Masui, T.: All You Need to Know about Gradient Boosting Algorithm Part 1. Regression. *Medium* (2024, February 18). Available at: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502
22. Implementing gradient boosting in Python | DigitalOcean. (n.d.). Available at: https://www.digitalocean.com/community/tutorials/implementing-gradient-boosting-regression-python
23. Yang, W., Chen, L., Ke, T., He, H., Li, D., Liu, K., & Li, H.: Carbon emission trend prediction for regional cities in Jiangsu Province based on the Random Forest model. *Sustainability* **16**(23), 10450 (2024). Available at: https://doi.org/10.3390/su162310450
24. How to Interpret a Residual Plot. (n.d.). Study.com. Available at: https://study.com/academy/lesson/how-to-interpret-a-residual-plot.html