ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23
American Academic publishers, volume 05, issue 12,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai

FORECASTING CO_2 EMISSION IN UZBEKISTAN USING MACHINE LEARNING TECHNIQUES

Nurbek Khalimjonov
Tashkent State University of Economics, Tashkent, Econometrics department,
10066, Uzbekistan
nurbekkhalimjonov070797@gmail.com

Abstract: Greenhouse gas increases and climate change are both influenced by human-caused carbon dioxide emissions. A key component of the fight against climate change and global warming is the regulation and reduction of carbon dioxide emissions. The transition to renewable energy sources and the reduction of emissions of greenhouse gases are topics of active discussion on a global and national scale. This is why it's critical to predict future greenhouse gases emissions in order to plan accordingly. This research uses two separate machine learning algorithms to effectively predict Uzbekistan's CO_2 emissions. R^2, MSE, and MAE were the three statistical metrics used to assess the study's efficacy. Artificial neural networks had an R^2 of 93.8 percent, an MSE of 0.007, and an MAE of 0.005. Decision trees had an R^2 of 90.1 percent, an MSE of 0.011, and an MAE of 0.009. By comparing the two models, we find that ANN outperforms decision trees and produces more accurate predictions.

Keywords: Carbon emission, ANN, The Decision tree, Uzbekistan

1. Introduction

Air pollution is one of humanity's most pressing issues in the modern world. Air pollution is becoming more of a problem as the population grows, cities expand, and industries expand. Pollutants in the air may have devastating impacts on people, animals, and ecosystems, and these impacts can vary greatly over time, geography, exposure duration, concentration, and other factors. Pollutant patterns and trends, as well as human exposure levels, are challenging to predict and assess due to their complexity. Preventing air pollution requires first accurately modeling and assessing the pollutant level [1].

It is feasible to ascertain the extent of contamination in the area by routinely measuring the air quality. This allows for the development of distribution models, air pollution maps, and programs to mitigate air pollution. Here, healthier, more practical, and less complicated strategies to enhance air standards and quality may be developed based on data from air quality monitoring [2].

Findings on air pollution have been made more objective and exact by machine learning, a cutting-edge AI technology. The overarching term for algorithms that can figure out a solution to a problem on their own via data-driven decision-making and intricate pattern recognition is machine learning. Achieving optimal performance was the guiding principle for the model's creation utilizing pre-existing data sets and ML techniques. Research into air pollution forecasting has shown a number of approaches that use a mix of NO_2, NO, O_3, carbon monoxide, SO_2, $PM_{2.5}$, and PM_{10} information sets [3,4,5,6,7,8].

Using machine learning methods, this research estimates Uzbekistan's carbon footprint from 1990 to 2023, which is different from previous literature. Methods using artificial neural networks and decision tree regression were used for this objective.

The researchers in this work estimated Uzbekistan's carbon footprint using ML models. Methods using artificial neural networks and decision tree regression were used for this objective. The research examined 33 yearly data points spanning from 1990 to 2023.

Methodology

Population, GDP, consumption of electricity per capita (kwh), green energy percentage, coal percentage, natural gas percentage, fluid fuels (fuel, diesel, etc.) percentage, total energy, number of vehicles powered by internal combustion engines, and the amount of forest (hectare) are the independent variables in this study. We chose factors that might impact the change in Uzbekistan's carbon emissions as input units when we were constructing the model for predicting carbon emissions.

To begin, it is widely believed that rising populations and economies lead to more carbon emissions. Energy usage and usage over time both rise in tandem with population, leading to higher output. Because renewable energy sources are not yet completely used, nations with big populations, like Uzbekistan, India, and China, etc., are unable to reduce carbon emissions. Consequently, the model incorporates population, GDP, and electrical usage per capita as input units. A country's carbon emissions are positively correlated with its energy consumption, regardless of whether that consumption is fueled by renewable sources or fossil fuels.

Worldwide, nations release a great deal of carbon dioxide into the atmosphere from their power plants, industries, and homes because of the widespread use of coal, an energy source derived from fossil fuels that is one of the most polluting substances in the world. For the purpose of reducing nations' carbon emissions over time, the model incorporates the idea of increasing the use of sources of renewable energy, which are clean energy sources. Carbon emissions are also affected by the pace of rise in the number of cars powered by internal combustion fossil fuels. This rate varies with factors such as population size, the expansion of logistics businesses using heavy trucks as a result of economic development, and the degree of wealth. Despite advancements in engine technology and the rise of electric powered automobiles, the model accounts for the growing number of I.C.E. vehicles on the road because of the assumption that they contribute to higher levels of carbon emissions. The greatest known locations for storing carbon are forests. Whether caused by humans or by nature, the atmospheric release of the carbon that is stored is inevitable. Hence, forest cover has a direct correlation to atmospheric carbon levels; more trees mean less carbon released into the atmosphere, and vice versa. Consequently, the model was updated to include forest area coverage as the final output unit.

The accuracy of machine learning outcomes is highly dependent on the thoroughness and precision of data processing and preparation. It is common to find data in an unstructured format when it is gathered from many sources. The accuracy of the models' predictions is impacted by this. For this reason, cleaning up the raw data is an essential first step in training, testing, and using ML models. Data pre-processing entails a sequence of steps to clean, convert, organize, and get data ready for use. Improper values, such as those caused by corruption, duplication, misspelling, etc., may occur in raw data. Finding and fixing inaccurate or defective data, particularly in huge data sets, is known as data cleaning [9]. Because it fixes the ensuing defective data, it becomes valuable again. Following the identification of dispersed, noisy, corrupted, or inaccurate observations, the following procedures are implemented [10, 11]:

*Find out whether your data is typical or if there are any outliers by using summary statistics.

*Sorting columns by their shared variance or value and removing them

*Find and delete rows of duplicate data.

*Empty values are shown as missing.

*Use data analysis or a pre-trained model to complete the missing variables.

A appropriate data collection for training may be prepared by applying the normalization technique to the raw data. The network's performance is significantly impacted by normalization, the process of scaling the model's inputs and outputs. Extremely sluggish performance may result from not normalizing the raw data collection. Excessively big or tiny numbers could show up in the data collection due to normalization regularizing the distribution of values. These numbers could steer the network astray by producing implausibly huge or little net input calculations. These gaps between the data, particularly the extreme data, will have a greater impact on the outcomes since some values in the same information set have numbers less than 0 and others have bigger values. When the data is normalized, every parameter in the initial input set is given an equal chance to influence the model's predictions [12,13].

By uniformly scaling all inputs within a specified range, often between 0 and 1, we may remove the impact of accidentally entering very big or tiny numbers and also standardize the information that comes from various settings. Various methods may be used in the process of normalizing. Data normalization comes in several forms as discussed in the literature. Methods like the Z-score, sigmoid, median, and min-max rule may be used to catalog them. This research used the min-max approach to standardize the data from 0 to 1 [14].

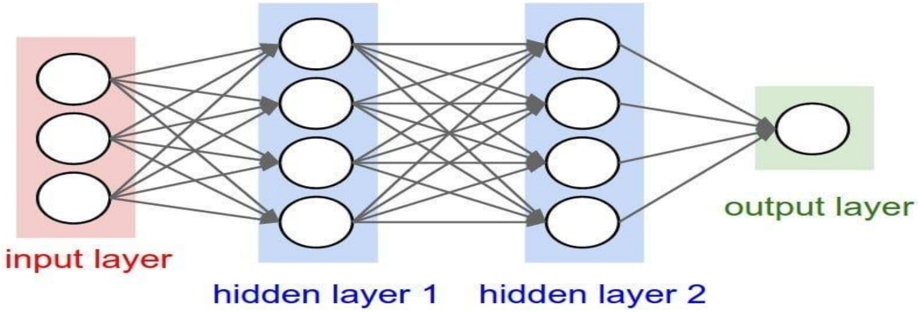
$$x' = \frac{x_1 - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

In this context, x_1 is the data that was normalized, x_1 represents the input worth, x_{min} is the input set's minimum value, and x_{max} is its maximum value.

When labels aren't present in the data, an unsupervised artificial intelligence approach is used. This algorithm examines the data with the intention of discovering correlations. Data is the primary emphasis of the system architecture in unsupervised machine learning [15, 16].

Among supervised learning algorithms, tree-based techniques see extensive use. The decision tree provides a data structure that uses a predefined set of rules to divide a big dataset into more manageable chunks. The two main applications of algorithmic decision trees are regression and classification. For numerical target data, regression is used, but for categorical data, classification is employed [17]. Every node within a decision tree stands for an attribute or characteristic, every branch for a rule of thumb, and every leaf node for the outcome; the structure resembles a flowchart. The root node is the highest node in an option tree. The expected value of the output is represented by its value at the node that represents the leaf [18].

They already know what the dependent variable will be. Based on what is known, it sorts the factors that are autonomous into intervals. [19]. The three main components of an artificial neural network are the input layer, the hidden layer (or layers), and the output layer (fig. 1).



Source: https://medium.com

Fig 1. Artificial Neural Network (ANN) model

Optimal values should be used for all of the network's weights. The term used to describe the steps used to get to these weights is "training the network." So, according to a particular criterion, weight values need to be dynamically adjustable for the network to be adaptable [22,23]. When the outputs of one layer of a network are sent back into the input units or earlier intermediary layers, we say that the model is a feedback neural network. Consequently, there is a two-way flow of inputs [25].

For the purpose of comparing and evaluating the models, the research favored the three most popular metrics found in the literature. We determined the best and most effective models by analyzing their determination coefficients (R^2), mean squared errors (MSEs), and mean absolute errors (MAEs).

$$R^2=1-\sum_{i}^{n} \frac{\left(y_{i_{actual}} - y_{i_{prediction}}\right)}{\left(y_{i_{actual}} - \text{Average}\left(y_{i_{artu\ al}}\right)\right)^2} \qquad (2)$$

In Equation 2, we can see R^2. Mean squared error, or MSE, is a metric for evaluating success. It allows you to compare the actual numbers with the anticipated values and see how they vary. It is a variable that is optimized in an effort to develop better models, particularly when optimization techniques are used. This metric measures the typical discrepancy between what a model developed using machine learning anticipates and what the world really looks like. Predictors whose MSE values are near to zero tend to have superior performance, and it is always favorable [26]. MSE may be shown in the third equation.

$$MSE=\frac{1}{n}\sum_{i=1} e_i \qquad (3)$$

When comparing actual and anticipated values, the mean absolute error (MAE) is the average of all the absolute figures of the errors. When applied to absolute error numbers, this measure produces the desired result. This form more accurately depicts the total of the mistake terms. The projected MAE is modest. As the MAE gets closer to zero, it is believed that the model is performing better [27]. In Equation 4, MAE is shown.

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad (4)$$

2. Results

The study's data set spans the years 1990 through 2023. The 33 data points in the dataset are all derived from yearly statistics. The yearly carbon footprint of Uzbekistan was estimated using machine learning models that included 8 independent factors and 1 dependent variable. Decision tree models and neural networks with artificial intelligence were the two machine learning approaches used in the research.

An analysis was conducted to determine the degree of correlation between the variables. One statistical tool for learning about the nature, strength, and direction of relationships between variables is correlation analysis.

The best values for the models' parameters, including their weights and coefficients, are determined here. To put the model that was trained on the training set to the test, it is necessary to have access to the test set. The model's learning performance and the reliability of the assessment metrics are both enhanced by increasing the size of the training set and the test set, respectively. The research found that the best effective rate was reached when 70% of the training and thirty percent test were separated, while other rates of separation (60–40%, 70–30%, 80%–20%) were also considered. We used a random sample technique to choose our test and training information in an effort to make the model more accurate. The research used a variety of training trials to determine the ANN model's parameters. The best performance was achieved by a feedback artificial neural network (ANN) model that has two layers that are concealed and three neurons per layer.

A number of methods have been proposed to address the issue. Chi-Square contingency table statistics, the towing rule, the Gini index, and the information return and information gain ratio are the most crucial ones. To decide which characteristic to branch the decision tree according to, the theory of information, including entropy principles, is used, in accordance with the data gain and knowledge gain proportion approaches. A system's entropy may be thought of as a measure of its disorder or uncertainty [29].

Entropy:

$$\mathrm{Entropy}(T)=\sum_{i=1}^{n} p_i\log_2(p_i) \qquad (5)$$

Partitioning a dataset according to a characteristic and then removing it from the total entropy is the foundation of information gain. The characteristic becomes less significant as it gets closer to 1. But the inverse is true when considering the benefit of knowledge gained [30].

$$\mathrm{Gain}(T,X)=\mathrm{Entropy}(T)-\mathrm{Entropy}(T,X) \qquad (6)$$

The research employed R2 (Coefficient of determination), MSE (Mean square error), and MAE (mean absolute error) metrics for evaluating and analyzing the models, as well as for errors. The models' evaluations are shown in Table 1.

Table 1. Evaluation of models

	ANN	**The Decision Tree**
R^2	0.938	0.901
MSE	0.007	0.011
MAE	0.005	0.009

The research found that ANN achieved an R2 of 93.8% and decision tree 90.1%. These findings suggest that the model's independent variables significantly explain the variable that is dependent and provide optimal values. Decision trees had an MSE score of 0.011 and ANNs of 0.007 in the research. These numbers are perfect since they are so near to zero. All of these numbers point to the model's accuracy in predicting the dependent variable and its low error rate. Decision trees had a mean absolute error (MAE) of 0.005 in the research, whereas artificial neural networks had an MAE of 0.009. These results demonstrate the model's efficacy as they are near the optimal MAE.

Conclusion

A rise in atmospheric greenhouse gasses and subsequent climate change are both aided by human-caused carbon dioxide emissions. A key component of the fight against climate change and global warming is the regulation and reduction of carbon dioxide emissions. The transition to renewable energy sources and the reduction of emissions of greenhouse gases are the targets of several national and international initiatives.In order to decide what actions to take, it is necessary to have an estimate of the emissions of carbon dioxide in the future years. Certain human activities, such as manufacturing, generating energy, and transporting goods, are often linked to emissions of carbon dioxide.

In order to shed light on the environmental implications and future of climate change, it is crucial to estimate these emissions. Nonetheless, there are a lot of moving parts in the process of producing a trustworthy prediction. Two separate machine learning models were able to accurately forecast Uzbekistan's CO2 emissions in this research.

In terms of accuracy and success rate, ANN outperforms decision trees when comparing the two models. This finding is consistent with previous research [31,32]. These findings suggest that machine learning approaches may accurately predict Uzbekistan's carbon dioxide emissions. In this way, the causes and consequences of carbon monoxide emissions may be identified. Those in positions of power will find it to be an invaluable resource. We may improve the success rate of future experiments by adding more factors that influence carbon dioxide emissions to the dataset and by testing other machine learning models. Furthermore, the research may be modified to fit other nations. As a result, other nations' actions will benefit from it.

References:

1. Alimissis, A., Philippopoulos, K., Tzanis, C.G., & Deligiorgi, D. (2018). Spatial estimation of urban air pollution with the use of artificial neural network models, Atmospheric Environment, 191, 205–213. DOI: 10.1016/j.atmosenv.2018.07.058

2. Aydınlar, B., Güveni, H., & Kırksekiz, S. (2009). Hava kirliliği ve modellenmesi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Çevre Mühendisliği Bölümü Yüksek Lisans Rapor.

3. Hu, K., & Rahman, A. (2017). HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation from Fixed and Mobile Sensors, IEEE Sensors, 17(11), 3571–3525. DOI: 10.1109/JSEN.2017.2690975

4. Huang, C-J., & Kuo, P-H. (2018). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities, Sensors.

5. Khalimjonov, N. (2024). Evaluating the environmental sustainability of Uzbek firms in the green economy, Yashil Iqtisodiyot va Taraqqiyot, 6(1), 142–147.

6. Khalimjonov, N. (2023). Foreign direct investment and electricity consumption during Uzbekistan’s green transition, Yashil Iqtisodiyot va Taraqqiyot, 1(11–12).

7. Khalimjonov, N. (2023). Ways to improve the green recovery in Uzbekistan by investment and trade, Yashil Iqtisodiyot va Taraqqiyot, 1(10), 97–101.

8. Khalimjonov, N., Allayarov, P., & Tajibaeva, K. (2023). Analysis and evaluation of worldwide carbon emission topics in economic journals: A review on articles published during 2016–2021, E3S Web of Conferences, 419, 01027.

9. Khalimjonov, N., & Rikhsimbaev, O. (2023). The analysis of economic growth and energy use and CO_2 emission in Uzbekistan, Economics and Innovative Technologies, 11(5), 247–257.

10. Khalimjonov, N., & Sharipov, B. (2024). An econometric analysis of FDI and trade in Uzbekistan, Mehnat Iqtisodiyoti va Inson Kapitali, 8(2), 163–171.

11. Kunt, F. (2007). Hava kirliliğinin yapay sinir ağları yöntemiyle modellenmesi ve tahmini, Selçuk University Graduate School of Natural and Applied Sciences, M.Sc. Thesis, Environmental Engineering Department, Konya.

12. Kwak, S.K., & Kim, J.H. (2017). Statistical data preparation: Management of missing values and outliers, Korean Journal of Anaesthesiology, 70(4), 407–411.

13. Kuhn, M., & Johnson, K. (2013). Data pre-processing, Applied Predictive Modeling, 27–59.

14. Mazziotta, M., & Pareto, A. (2022). Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the Min-Max method, Environmetrics, 33(5), e2730.