Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 3, September 2023, pp. 896~910
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i3.4892

Forecasting Carbon Dioxide Emission in Thailand Using Machine Learning Techniques

Siriporn Chimphlee[1], Witcha Chimphlee[2]
[1,2]Faculty of Sciene and Technology, Suan Dusit University, Thailand

Article Info

Article history:

Received May 30, 2023
Revised Aug 28, 2023
Accepted Sep 23, 2023

Keywords:

Thailand
CO2 emission
Carbon dioxide
Forecasting
Machine Learning

ABSTRACT

Machine Learning (ML) models and the massive quantity of data accessible provide useful tools for analyzing the advancement of climate change trends and identifying major contributors. Random Forest (RF), Gradient Boosting Regression (GBR), XGBoost (XGB), Support Vector Machines (SVC), Decision Trees (DT), K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), ensemble methods, and Genetic Algorithms (GA) are used in this study to predict CO2 emissions in Thailand. A variety of evaluation criteria are used to determine how well these models work, including R-squared (R2), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and correctness. The results show that the RF and XGB algorithms function exceptionally well, with high R-squared values and low error rates. KNN, PCA, ensemble methods, and GA, on the other hand, outperform the top-performing models. Their lower R-squared values and higher error scores indicate that they are unable to accurately anticipate CO2 emissions. This paper contributes to the field of environmental modeling by comparing the effectiveness of various machine learning approaches in forecasting CO2 emissions. The findings can assist Thailand in promoting sustainable development and developing policies that are consistent with worldwide efforts to combat climate change.

Corresponding Author:

Witcha Chimphlee,
Faculty of Science and Technology,
Suan Dusit University,
295 Ratchasima Road, Dusit, Bangkok 10300, Thailand.
Email: witcha_chi@dusit.ac.th

1. INTRODUCTION

Since global warning has developed into a problem that affects climate change and the environment [1][2][3]–[5] and cause the greenhouse effect [6], which raises global temperatures, causes sea levels to rise, and has detrimental consequences on ecosystems and public health. CO2 emissions have therefore become a global problem. This issue of greenhouse gases and increasing energy use has led to interest in predicting future carbon dioxide emissions. Like many other nations, Thailand struggles to successfully manage and lower CO2 emissions. The nation's CO2 emissions have significantly increased as a result of the quick industrialization, rising population, and rising energy needs [7]. To reduce the negative effects on the environment and encourage sustainable growth, it is imperative to address this issue and create effective systems for tracking, forecasting, and reducing CO2 emissions. Due to increased CO2 emissions [8], Thailand confronts serious environmental challenges [9][10]. Accurate emission forecasting can help stakeholders, academics, and policymakers create effective plans to reduce negative environmental effects and advance sustainable development. Machine learning algorithms provide an effective collection of tools for examining huge datasets, finding patterns, and producing accurate predictions [11][12]–[15].

Accurately forecasting carbon dioxide (CO2) emissions has become a crucial problem as concern over climate change and its effects on the environment has grown. It has been demonstrated that machine learning [6], [7], [16]–[19] approaches are useful for modeling complex relationships and producing precise

predictions. Using a variety of machine learning algorithms [7], such as Random Forest (RF) [19], Gradient Boosting Regression (GBR) , XGBoost (XGB), Support Vector Machines (SVC) [20][21], Decision Trees (DT), K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), ensemble methods, and Genetic Algorithms (GA).

From the literature mentioned above, forecasting CO2 emissions depends on many factors. For this reason, new models are being developed to predict carbon dioxide emissions. For starters, previous studies may have been limited by insufficient data samples, preventing a thorough grasp of the emissions landscape. Furthermore, several important elements influencing CO2 emissions in Thailand may have been neglected in previous studies, thus limiting the accuracy of forecast models customized to the country's specific environment. It is worth mentioning that Thailand badly needs the investigation of CO2 forecasting because, in 2022, it has produced around 19678.52 tons, which increased from last year by approximately 4.95%. Considering all those influential factors on CO2 emission, the problem formulation leads to analyze and compare how well machine learning algorithms, which are difficult to execute but provide more accurate results. Thailand, in contrast to countries with rich study in this field, may not have received the essential attention, although having significant emissions growth and environmental issues. In Thailand's CO2 prediction attempts, addressing these gaps, applying complex multivariate time series forecasting tools, and exploiting the potential of machine learning techniques remain largely untapped. Finally, undertaking such study in Thailand, a populous country coping with environmental issues, has the potential to greatly influence national CO2 emission policy and contribute to global sustainability efforts.

To fulfill the research gaps, this paper aims to forecast CO2 emissions in Thailand and explores how well different machine learning algorithms perform in forecasting CO2 emissions in Thailand with the goal of advancing the field of environmental modeling. The results will provide light on the benefits and drawbacks of various algorithms as well as their suitability for emissions prediction. The most effective models will be determined by comparing evaluation metrics based on Multiple evaluation indicators will be used to evaluate the predictive models. The coefficient of determination (R2) gauges how much of the variation in CO2 emissions is explained by the models, giving a sense of how well-fit the models are. The average and root mean squared disparities between projected and actual CO2 emissions are quantified by mean absolute error (MAE) and root mean squared error (RMSE), respectively [3]. The percentage difference between projected and actual emissions is calculated using the mean absolute percentage error (MAPE), allowing for a comparative evaluation of prediction accuracy. If categorizing emission levels is part of the prediction task, accuracy will be used as a measurement parameter [4].

In the end, the findings of this study will support stakeholders, policymakers, and researchers in developing ways to lower CO2 emissions in Thailand. We can pave the road for more precise and successful predictions by utilizing machine learning and employing strong assessment measures, which will support international efforts to battle climate change and advance sustainable development. The current challenge is to provide a reliable machine learning-based method to forecast CO2 emissions in Thailand. Utilizing historical data, identifying important contributing elements, and developing precise models that may predict future emission levels are the objectives. Establishing effective measures to tackle Thailand's CO2 emissions crisis can be facilitated by the use of machine learning approaches, as demonstrated in this study, which is shown below. The first portion is an introduction. Section 2 contains works that are related. Section 3 describes the strategy used in this paper, while Section 4 describes the experimental design. Section 5 presents the experimental data, while Section 6 ends the study.

2. RELATED WORKS

There is a large body of literature that discusses and reviews this issue. According to Pérez-Suárez et al. [22] focuses on the transition from Millennium Development Goals (MDG7) to Sustainable Development Goals (SDG), with an emphasis on environmental sustainability. It emphasizes the need of environmental forecasts, notably for carbon dioxide emissions, which have nearly doubled since 1990. The study assesses the Extended Environmental Kuznets Curve (EKC) and the Environmental Logistic Curve (ELC) for projecting CO2 emissions in 175 nations, providing useful insights into their appropriateness and predictive accuracy. These findings contribute to the ongoing global effort under the SDG framework to solve environmental concerns.

Phatchapa Boontome et al. [23] emphasizes the importance of CO2 emissions forecasting for improving government energy strategies. It finds significant long-term causal linkages using an autoregressive distributed lag approach: a 1% increase in renewable energy, energy consumption, and oil prices reduces CO2 emissions by 5.66%, 14.73%, and 5.07%, respectively. The prediction predicts a 30.17% reduction in CO2 emissions over the next 14 years, above the 2030 target of 20-25%. This emphasizes the necessity for quick changes in the energy consumption structure to prevent pollution.

Sutthichaimethee Pruethsan et al. [24] used a second-order autoregressive-structural equation model (second order autoregressive-SEM) to anticipate economic and environmental growth in accordance with the Thai government's 2020-2035 strategic plan. Alternative methods are outperformed by the model, which achieves a low mean absolute percentage error (MAPE) of 1.02% and a root mean square error (RMSE) of 1.51%. The findings show that, while present policy promotes economic growth, it results in a worrying increase in CO2 emissions that exceeds safety levels. The study underlines the need of changing policies for environmental sustainability, as well as the model's usefulness in guiding successful and long-term policy creation.

Kumari et al. [7] looks into the critical issue of CO2 emissions in India, a country with high per capita emissions. To forecast emissions over the next decade, it employs a variety of statistical, machine learning, and deep learning models. The LSTM model is identified as the most accurate in the performance study, making it a recommended choice for CO2 emission prediction in India, providing useful insights for environmental policy and planning.

Freitas et al. [8] focuses on reducing carbon dioxide (CO2) emissions from sugarcane production in Brazil. It predicts soil CO2 emissions using neural networks and a backpropagation method, which is an important aspect in greenhouse gas dynamics. The results show that the neural network can accurately estimate soil CO2 emissions, improving our understanding of geographical patterns and helping to more exact emission estimates in sugarcane fields, hence assisting environmental efforts.

Achiraya Chaichaloempreecha et al. [3] evaluates long-term energy policy in the industrial and building sectors (2005-2050), with a focus on energy savings and GHG mitigation using LEAP. The findings emphasize the efficacy of energy labeling and monetary incentives in lowering energy use and GHG emissions. According to the report, these initiatives can help Thailand fulfill its NDC commitments, with biogas deployment and CCS technology playing important roles in reducing emissions. Furthermore, the article emphasizes the relevance of economic and environmental factors in ensuring energy security.

Ma N et al. [11] investigates the use of a nonparametric kernel prediction technique in machine learning for predicting CO2 emissions. A thorough literature study guides the selection of independent variables. The study compares classic parametric models to Gaussian Process Regression (GPR) algorithms, demonstrating that GPR algorithms produce the most accurate CO2 emission forecasts.

Sutthichaimethee P. et al. [24] offers the Path Analysis-VARIMA-OVi Model, a powerful tool for long-term forecasting (2020-2034) to achieve sustainable development goals. It identifies the optimum way for government scenario programs by revealing the causal links between economic, social, and environmental growth elements. Using this model, the study discovers that reducing energy use can result in a considerable reduction in future CO2 emissions with a low error rate. When compared to other models, Path Analysis-VARIMA-OVi emerges as the best option for successful sustainability planning in Thailand.

A thorough evaluation of the literature reveals the critical need for CO2 emissions prediction research. Thailand's emissions are increasing, necessitating the use of more accurate forecasting models. Accurate forecasting is critical for aligning domestic policies with international environmental commitments. The difficulty arises in addressing sector-specific dynamics while also embracing changing socioeconomic conditions. Creating precise models that capture Thailand's specific context is critical for developing successful emissions reduction measures, guaranteeing environmental sustainability, and meeting global climate goals.

3. RESEARCH METHOD

The Thailand CO2 emissions dataset provides an in-depth study of emissions over 36 years with monthly data and source details. The available data includes fuel type, allowing a closer look at his CO2 emissions for the country. The source of this data is Thailand's Department of Energy Policy and Planning and was originally published in Excel[25]. Data were consolidated and cleaned up using Python scripts. The result is a simplified and user-friendly dataset containing 2,424 rows emissions from three distinct sources and employing three different types of fuels are included.. This dataset will serve as a valuable resource for researchers and analysts wishing to assess and understand Thailand's carbon emissions.

Thailand's carbon footprint estimation method used a variety of machine learning algorithms and scoring criteria. The algorithms/models listed below were used. Random Forest (RF), Gradient Boosted Regression (GBR), XGBoost (XGB), Support Vector Classifier (SVC), Decision Tree (DT), K Nearest Neighbors (KNN), Principal Component Analysis (PCA), Ensemble Methods and genetic algorithm (GA) are examples of machine learning techniques.

Several evaluation criteria were employed to evaluate the performance of these models, including R-squared (R2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Accuracy. These metrics provide information about the accuracy, precision, and dependability of the models' predictions [26].

This section describes the technique and experimental design used in this study. RF, GBR, XGB, SVC, DT, KNN, PCA, ensemble techniques, and GA are among the machine learning algorithms compared in the methodology.

3.1 Random Forest (RF)

Random Forest (RF) is an ensemble learning system that makes accurate predictions by combining numerous decision trees [27]. Because of its resilience and effectiveness, it is often utilized for prediction tasks. The Random Forest algorithm is explained in depth below:

- Random Forest Construction: Given a training dataset with N samples and M features, RF builds a preset number of decision trees, known as "n_estimators." Each tree is constructed independently using a randomly selected sample of the training data.
- Random Feature Selection: A random subset of features, designated by "mtry," is considered at each node of the decision tree to determine the best split. Typically, the number of features in the subset is substantially lower than the entire number of features in the dataset.
- Decision Tree Construction: The decision tree is built using a random selection of features by recursively splitting the data depending on various splitting criteria such as Gini impurity or information gain. The process is repeated until a preset stopping criterion, such as reaching a maximum depth or a minimum amount of samples in a leaf node, is fulfilled.
- Voting Mechanism: During prediction, each decision tree creates a forecast independently based on the input attributes. The final forecast for categorization tasks is decided by majority voting among the individual trees. The final prediction for regression problems is the average of the predictions from all trees.
- Ensemble Effect: The key notion underlying Random Forest is that combining numerous decision trees minimizes the tendency of individual trees to overfit the data. The ensemble effect improves generalization performance and handles noise in the dataset better.
- Feature Importance: Random Forest calculates feature importance based on how much each feature adds to model correctness. This is derived by taking the average decrease in impurity or information gain induced by each feature and dividing it by the number of trees in the forest.

Random Forest's forecast can be expressed as:

$$RF(X) = (1/n\_estimators) * \Sigma(tree(X)) \quad (1)$$

where RF(X) is the Random Forest prediction for input feature X, n_estimators is the number of decision trees in the forest, and tree(X) is the prediction of each individual tree.

3.2 Gradient Boosting Regression (GBR)

Gradient Boosting Regression (GBR) is a popular machine learning approach for regression applications [16]. It is a boosting algorithm that combines several weak learners, typically decision trees, to produce a powerful prediction model. Here is a brief summary of GBR:

- First Prediction: GBR begins by creating an initial prediction using a simple model, such as the target variable's mean.
- Residual Calculation: The algorithm computes the residuals, which are the differences between the actual target values and the initial predictions.
- Weak Learner Training: GBR creates a sequence of weak learners to forecast residuals. Using gradient descent, each weak learner is trained to reduce residual error.
- Model Update: The weak learners' predictions are merged to update the model. The learning rate parameter governs how much each weak learner contributes to the final prediction.
- Iterative Process: Steps 3 and 4 are iteratively repeated, with each new weak learner focusing on residuals left by prior learners.
- Final forecast: The final forecast is calculated by adding the predictions of all weak learners and weighting them by the learning rate.

GBR prediction can be expressed as:

$$GBR(X) = initial\_prediction + learning\_rate * \Sigma(weak\_learner(X)) \quad (2)$$

where GBR(X) is the prediction for input feature X, initial_prediction is the initial prediction, learning_rate is the learning rate parameter, weak_learner(X) represents each weak learner's prediction, and the total is calculated over all weak learners in the model.

3.3 XGBoost (eXtreme Gradient Boosting)

The XGBoost (eXtreme Gradient Boosting) machine learning technique is widely used for regression and classification tasks [28]. It is a tree-based implementation of the gradient boosting framework that has been optimized. Here is an overview of XGBoost.

- Initial Prediction: XGBoost begins with an initial prediction based on a simple model, such as the target variable's mean.
- Residual Calculation: The algorithm computes the residuals, which are the differences between the actual target values and the initial predictions.
- Tree Ensemble Training: To estimate the residuals, XGBoost creates an ensemble of decision trees. Each tree is trained to minimize a loss function plus a regularization term.
- Gradient-Based Optimization: The algorithm optimizes the objective function, which represents the total inaccuracy of the model, using gradient descent techniques.
- Model Update: The decision trees' predictions are pooled to update the model. With a weight specified throughout the optimization process, each tree contributes to the final prediction.
- Iterative Process: Steps 3–5 are iteratively repeated, with each new tree focused on the residuals left by prior trees.
- Final Prediction: To generate the final prediction, add the original predictions, the predictions of the decision trees, and a shrinkage factor (learning rate) to control the contribution of each tree.

XGBoost's forecast can be expressed as:

$$XGB(X) = initial\_prediction + learning\_rate * \Sigma(tree\_prediction) \quad (3)$$

where XGB(X) is the prediction for input feature X, initial_prediction is the initial prediction, learning_rate is the shrinkage factor, tree_prediction represents each decision tree's prediction, and the total is calculated over all decision trees in the ensemble.

3.4 Support Vector Machines (SVM)

A popular approach for classification and regression applications is Support Vector Machines (SVM) with a linear kernel [21][20]. In a high-dimensional feature space, it finds a hyperplane that divides the data points into various classes. An overview of the Support Vector Machines algorithm is provide below:

- Data Representation: SVM represents data as points in a high-dimensional space, with each feature corresponding to a coordinate axis.
- Hyperplane Construction: SVM seeks the hyperplane that best separates the classes in the feature space. The hyperplane in linear SVM is a linear decision boundary that separates the data into classes.
- Support Vectors: Support vectors are the data points nearest to the decision border. They are vital in defining the hyperplane.
- Margin Maximization: SVM aims to maximize the margin, which is the distance between the hyperplane and the support vectors. It seeks the hyperplane that optimizes the separation between the classes.
- Classification/Regression: Once the hyperplane has been found, SVM can classify new data points or generate predictions based on their position relative to the hyperplane.

The decision function of SVM with a linear kernel can be written as:

$$f(X) = sign(w \cdot X + b) \quad (4)$$

where f(X) signifies the predicted class or regression value for input features X, w denotes the weight vector, means the dot product, and b denotes the bias factor. Based on the position relative to the hyperplane, the sign function determines the class or value.

3.5 Decision Trees (DT)

Decision Trees (DT) are a common machine learning method that may be used for classification as well as regression [29][30]. Based on various input features, they build a model that predicts the value of a target variable. Here's a basic overview of the Decision Tree algorithm:

- Tree Structure: A Decision Tree is built in the form of a tree, with each internal node representing a feature or attribute and each leaf node representing a class or a regression value.
- Splitting Criteria: The tree is constructed by recursively splitting the data depending on the specified features, with the goal of maximizing information gain or decreasing impurity at each node. Gini impurity and entropy are two common splitting criteria.

- Leaf Nodes: The splitting process continues until a requirement, such as reaching a maximum depth or having a minimum amount of samples at a node, is reached. The predicted class or regression value is contained in the resulting leaf nodes.
- Predictions: New data points are traversed down the tree from the root node to a leaf node based on the feature values to produce predictions. As the output, the prediction at the leaf node is used.

A Decision Tree's decision process can be described mathematically as a set of if-else conditions based on the selected attributes. As an example:

```
If (feature1 = threshold1), then
    If (feature2 = threshold2), prediction = class1.
    otherwise, prediction = class2
otherwise, prediction = class3
```

Each node in the tree indicates a decision made on the basis of a feature and a threshold value. The ultimate prediction is determined by the path taken from the root node to a leaf node.

3.6 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet powerful machine learning technique that may be used for classification and regression problems [8]. It predicts the target variable by taking into account the training dataset's "k" nearest neighbors. Here's a basic of the KNN algorithm:

- Training Phase: KNN stores the whole training dataset with labeled data points during the training phase.
- Distance Calculation: In order to predict new data points, KNN computes the distances between the new point and all existing points in the training dataset. The Euclidean distance is the most often used distance metric, but other distance metrics can also be utilized.
- Neighbor Selection: Based on the estimated distances, KNN selects the "k" nearest neighbors. The value of "k" is a user-defined parameter that specifies the number of neighbors to take into account.
- Prediction: The class labels of the "k" nearest neighbors are reviewed for categorization, and the majority class label is assigned as the prediction for the new data point. The forecast for regression is the average or weighted average of the target values of the "k" nearest neighbors.

The KNN prediction method can be described mathematically as follows:

Prediction = mode(neighbors' class labels) for Classification
Prediction = mean(neighbors' goal values) for Regression

KNN is a non-parametric algorithm that makes no assumptions about the data's underlying distribution. It is, however, sensitive to the value of "k" and the distance metric used. To eliminate bias in the predictions, it is critical to choose a suitable "k" value and scale the characteristics suitably.

3.7 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction approach that is used to convert high-dimensional data into a lower-dimensional space while retaining the most significant information. It is often used in machine learning projects as a preprocessing step [31]. Here's a quick rundown of the PCA algorithm:

- Data Standardization: PCA necessitates data standardization by removing the mean and scaling it to unit variance. This phase guarantees that each characteristic makes an equal contribution to the analysis.
- Covariance Matrix Calculation: PCA computes the standardized data's covariance matrix. The covariance matrix depicts the interactions between several features.
- Eigendecomposition: To acquire the eigenvectors and eigenvalues, PCA performs an eigendecomposition of the covariance matrix. The eigenvectors reflect the principal components, and the related eigenvalues indicate how much variance each principal component explains.
- Principal Component Selection: The eigenvectors are sorted according to their eigenvalues. As the principal components that capture the most substantial variance in the data, the top "k" eigenvectors are chosen.
- Projection: To obtain the lower-dimensional representation, the original data is projected onto the selected major components. This is accomplished by multiplying the standardized data by the matrix of eigenvectors chosen at random.

The data projection onto the selected major components can be expressed mathematically as:

$$Y = X * W \quad (5)$$

Where Y is the transformed data, X is the standardized input data, and W is the eigenvector matrix.

PCA can be used to reduce data dimensionality, visualize high-dimensional data, and remove irrelevant or duplicate characteristics. It can increase machine learning model performance by minimizing noise and focusing on the most informative characteristics.

3.8 ENSEMBLE METHODS

Machine learning ensemble approaches integrate numerous distinct models to produce a more robust and accurate prediction [32][33]. Bagging and Boosting are the two most common ensemble approaches. Here is an overview of ensemble methods:

- Bagging: Bagging is an abbreviation for Bootstrap Aggregating. Multiple independent models are trained on various subsets of the training data. Each model is trained on a random sample of the original dataset with replacement (bootstrap). The final prediction is derived by averaging or voting on all of the individual model projections.
- Boosting: Boosting is an iterative technique that trains numerous weak models consecutively and then combines them into a strong predictive model. Each weak model is trained on a subset of the training data, with a particular emphasis on examples misclassified by prior models. The final prediction is generated by integrating all of the weak models' predictions and assigning greater weight to the models that perform better.

The ensemble prediction can be expressed as:

$$\text{Ensemble Prediction} = f(\text{Model1}, \text{Model2}, ..., \text{ModelN}) \quad (6)$$

Model1, Model2,..., ModelN are the individual models in the ensemble, and f is the aggregation function, such as averaging or voting.

Ensemble approaches aid in the reduction of overfitting, improvement of generalization, and improvement of prediction accuracy. Ensemble approaches can manage complicated interactions in data and produce more trustworthy predictions by integrating the capabilities of numerous models.

3.9 Genetic Algorithm (GA)

The Genetic method (GA) is a metaheuristic optimization method influenced by the natural selection process. It is frequently used for optimization problems, such as prediction jobs. Here's a quick rundown of GA:

- Initialization: A population of possible solutions (chromosomes) is formed at random.
- Fitness Evaluation: The fitness of each chromosome is evaluated, which signifies how well it performs on the prediction job. Fitness can be quantified in the context of CO2 emission prediction using assessment metrics such as R-squared, MAE, RMSE, MAPE, or accuracy.
- Selection: Chromosomes with higher fitness are more likely to be selected for reproduction. This mechanism is similar to natural selection's survival of the fittest.
- Crossover: Crossover occurs when genetic information is shared between pairs of chromosomes. This aids in the exploration of various feature or parameter combinations.
- Mutation: To increase diversity and prevent convergence to inferior solutions, random alterations are introduced into the chromosomes.
- Next Generation: The kids produced by crossover and mutation replace the population's least fit individuals, resulting in a new generation of potential solutions.
- Termination: When a stopping requirement is reached, such as reaching a maximum number of generations or obtaining a desirable fitness level, the algorithm terminates.

3.10 Forecasting evaluation metrics

To evaluate the performance of your model, select the relevant metric. Among the measures often utilized for regression tasks in CO2 emission prediction are:

- Mean Absolute Error (MAE): This metric computes the average absolute difference between anticipated and actual values. It calculates the average magnitude of the errors.
- Mean Squared Error (MSE): This metric computes the average squared difference between anticipated and actual data. It penalizes greater mistakes more severely than MAE.
- Root Mean Squared Error (RMSE): This metric computes the square root of MSE and provides a measure in the same unit as the target variable.

- R-squared (R2) Score: Calculates the proportion of the target variable's variance that can be explained by the model. It has a value between 0 and 1, with higher values suggesting a better fit.

Figure 1 displays a general technique for evaluating the performance of such algorithms, which may be expressed in the following steps.

- Statistics Collection: Compile a complete dataset containing historical CO2 emission statistics for Thailand, as well as essential variables such as energy consumption, industrial production, population, and economic indicators.
- Data Preprocessing: Handle missing values, remove outliers, and normalize the data to provide a consistent scale across features.
- Feature Selection/Engineering: Analyze the dataset to determine which features are most relevant for estimating CO2 emissions. Techniques such as correlation analysis, feature importance ranking, or domain knowledge-based selection may be used in this step.
- Dataset Splitting: Separate the dataset into training and testing sets. A popular technique is to employ a significant chunk (e.g., 70-80%) for training and the remainder for testing the model's performance.
- Algorithm Selection: Select relevant machine learning algorithms for regression tasks. Linear regression, decision trees, random forests, support vector regression (SVR), gradient boosting, and neural networks are some popular CO2 emission prediction algorithms.
- Model Training: Using the training dataset, train the selected algorithm(s). To improve model performance, adjust hyperparameters using strategies such as grid search or random search.
- Model Evaluation: Using the testing dataset, evaluate the trained model(s). Mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score are common evaluation metrics for regression problems.
- Cross-Validation: Use k-fold cross-validation to test the stability and robustness of the model. This technique entails dividing the data into k subsets and completing training and assessment k times, each time utilizing a different subset as testing data.
- Performance Comparison: Using the assessment measures, compare the performance of various methods. Choose the method that offers the best performance based on the metric(s) of interest.

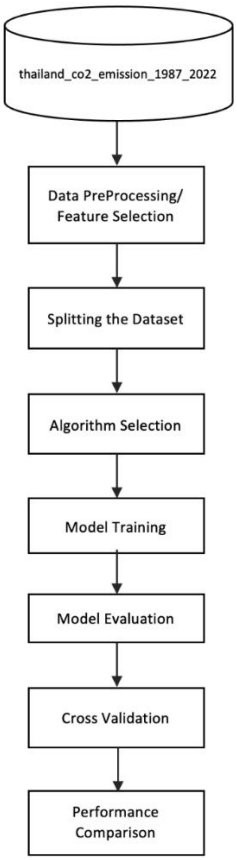

Figure 1. Workflow of Prediction Algorithm Evaluation and Comparison

4. EXPERIMENTAL SETUP

The computer utilized had a 64-bit version of ImacPro 2017 called 13.5.2 (22G91), and it had the following specs: 3.2 GHz 8-Core Intel Xeon W, with 32 GB 2666 MHz DDR4 memory. For implementation and evaluation of the suggested model, the Python (version 3.9) environment was used with the numpy, pandas, and sklearn tools for data processing. We did preprocessing, exploratory data analysis by using Thailand CO2 Emission dataset [34] provides a comprehensive monthly breakdown of carbon dioxide (CO2) emissions for the country over the course of 36 years, includes information on emission sources (Industry, Transportation and others) and fuel types (Oil, Natural gas and coal), allowing for a detailed analysis of the country's carbon footprint. The data provided by the Thailand Energy Policy and Planning Office (EPPO) were in multiple Excel files which later merged and cleaned by my python script, making it more accessible in the appropriate format. The data is clean and ready to use, including 36 Years, 3 Sources, and 3 Fuel Types totaling 2,424 rows. Sample of data show in Figure 2.

	year	month	source	fuel_type	emissions_tons
0	1987	1	transport	oil	1588.61
1	1987	2	transport	oil	1428.29
2	1987	3	transport	oil	1581.16
3	1987	4	transport	oil	1557.40
4	1987	5	transport	oil	1513.35
5	1987	6	transport	oil	1465.69
6	1987	7	transport	oil	1542.77
7	1987	8	transport	oil	1493.17
8	1987	9	transport	oil	1406.57
9	1987	10	transport	oil	1503.36

Figure 2. Sample of data [34]

5. RESULTS AND DISCUSSION

In the Exploratory Data Analysis step, we obtained the following result using figure 3-6, which is detailed below.

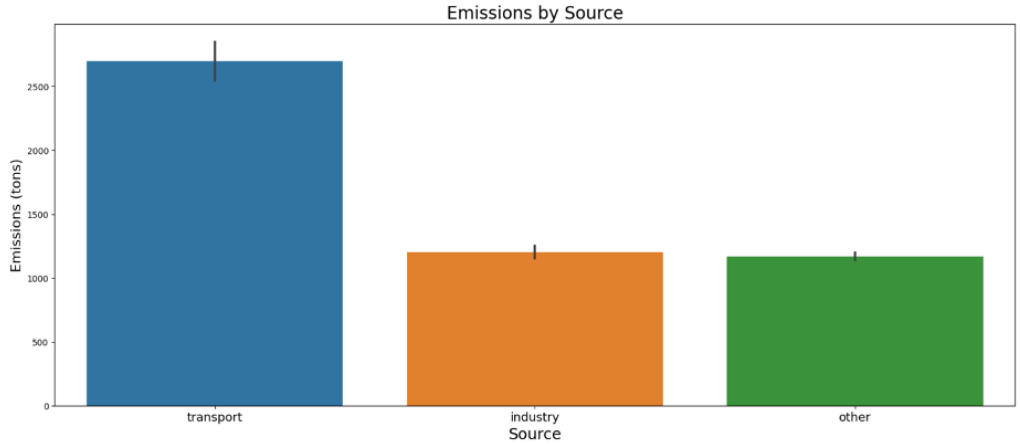

Figure 3. Emission source

Based on the provided values for transport, industry, and other as 2694.20, 1203.12, and 1168.37 respectively, the resulting bar plot will show three bars (Figure 3) corresponding to these source types: transport, industry, and other. The bar representing the transport source type will have a height that represents the emissions of 2694.20 tons, and a value label of "2694.20" will be displayed above the bar. The bar

representing the industry source type will have a height that represents the emissions of 1203.12 tons, and a value label of "1203.12" will be displayed above the bar. The bar representing the other source type will have a height that represents the emissions of 1168.37 tons, and a value label of "1168.37" will be displayed above the bar.

The labels displaying values show the emissions values linked with each type of source, making it simple to compare between the various categories. This information allows examination and understanding of the emissions levels of each source type in the dataset.

The program creates a bar chart showing the emissions in tons for each fuel type, with corresponding value labels displayed above each bar. Value labels show the exact value of each bar with two decimal places. The resulting bar graph (Fig. 4) shows his three bars for these fuel types.

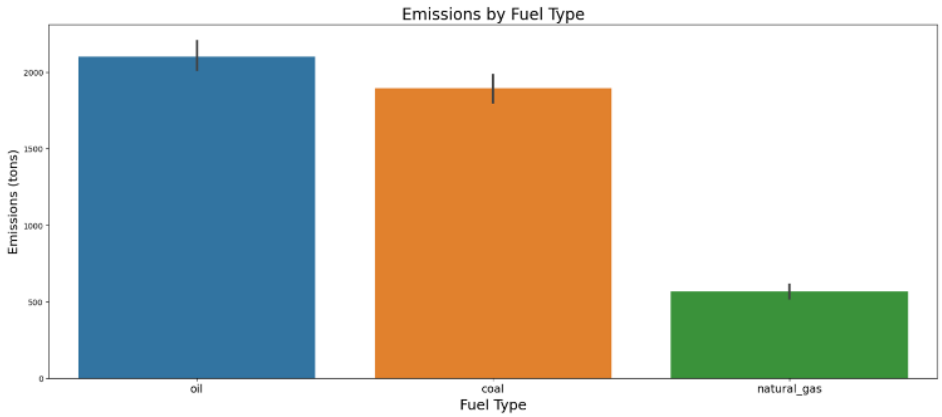

Figure 4. Fuel type

The oil, coal, and natural gas emissions are 2108.45 tonnes, based on the "oil column" with the entered values of 2108.45, 1895.10, and 566.65, or the higher value of '2108.45'. The carbon stick has a height of 1895 emissions of 10 tons and displays the value "1895.10" above the bar. The height of the natural gas column is 566.65 tons, above which the value is '566.65'. Value labels clearly express the emission levels associated with each fuel type, facilitating comparisons between different categories. This information will help you analyze and understand the relative emission levels of different fuel types in your dataset. The resulting line graph (Figure 5) provides a visual representation of emission trends for each fuel_type category over time. Emission patterns are possible Compare across different sources and identify trends and patterns within each category identified.

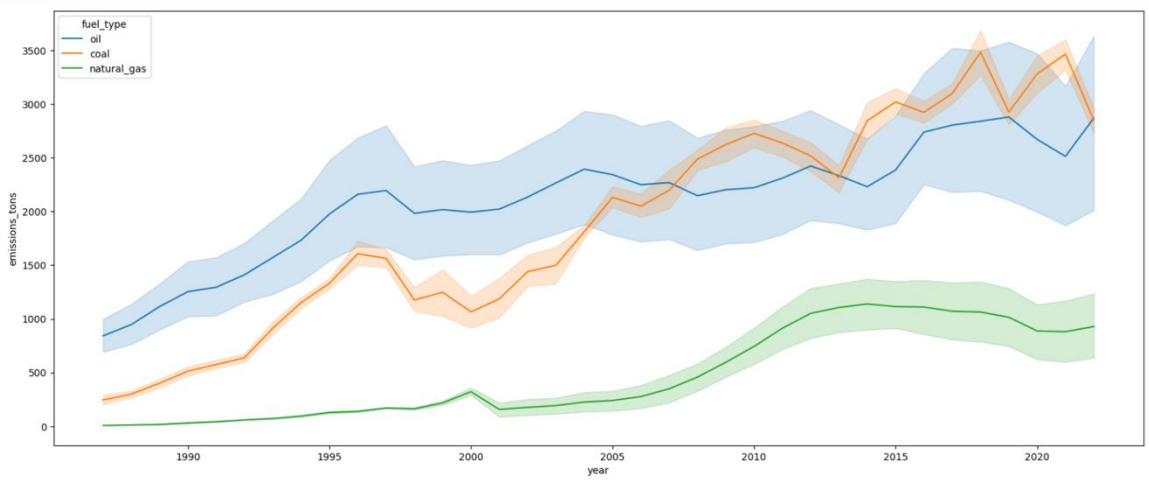

Figure 5. Emission trends

Figure 5 depicts the total amount of emissions from 2018 to 2022, broken down by year, source, and fuel_type. It demonstrates that, with the exception of coal, practically all fuel_types from various sources have increased.

year	source	fuel_type	emissions_tons
2018	industry	coal	41784.15
		natural_gas	20856.74
		oil	15805.86
	other	oil	15902.14
	transport	natural_gas	4665.54
		oil	70501.87
2019	industry	coal	35093.49
		natural_gas	20242.50
		oil	15102.48
	other	oil	15984.48
	transport	natural_gas	4114.47
		oil	72557.22
2020	industry	coal	39352.20
		natural_gas	18340.95
		oil	11874.15
	other	oil	13359.39
	transport	natural_gas	2958.52
		oil	70918.22
2021	industry	coal	41575.48
		natural_gas	18750.84
		oil	10926.33
	other	oil	12648.55
	transport	natural_gas	2388.66
		oil	66876.69
2022	industry	coal	34209.49
		natural_gas	19678.52
		oil	12557.62
	other	oil	13715.90
	transport	natural_gas	2598.09
		oil	76974.46

Figure 6. Emissions from 2018 through 2022

Using the evaluation measures of choice, compare the performance of several models or algorithms. Choose the model with the lowest MAE or RMSE and the highest R2 score. In our CO2 emission analysis, we use the sklearn library for data preparation for each method. For the Random Forest (RF) model, we use the RandomForestRegressor with the following parameters: n_estimators is set to 1000, and random_state is set to 42. We use the GradientBoostingRegressor technique with the n_estimators set to 1000. For our investigation, we use a variety of machine learning models and approaches, including the XGBRegressor, SVC (Support Vector Classifier), DecisionTreeRegressor, KNeighborsRegressor, PCA (Principal Component investigation), and GA (Genetic Algorithm), all with their default parameter settings.

The evaluation metrics for each algorithm/model used in estimating CO2 emissions in Thailand are shown in Table 1. According to the evaluation metrics, certain algorithms/models outperform others in estimating CO2 emissions in Thailand. The Random Forest (RF) technique achieves a high R-squared value of 0.998637, suggesting that the model fits the data very well. It also has the best predictive performance, with the lowest MAE, RMSE, MAPE, and maximum accuracy among the models. With a near-perfect R-squared value of 0.9999975 and reasonably low MAE, RMSE, and MAPE scores, the XGBoost (XGB) algorithm also exhibits strong predictive skills. However, it is slightly less accurate than the Random Forest model.

Table 1. Summary of experimental results

Algorithm	R-Squared	MAE	RMSE	MAPE (%)	Accuracy (%)
RF	0.998637	41.65	72.4	5.16	99.95
GBR	0.9926592	47.79	77.13	10.86	89.14
XGB	0.9999975	38.95	70.29	7.59	92.41
SVC	0.99587264	96.63	205.15	14.79	85.21
DT	0.98433061	52.32	86.56	6.41	95.55
KNN	0.835	200.13	280.88	79.82	82.97
PCA	0.205596	503.36	616.31	257.46	57.17
Ensemble Methods	0.8177274	217.99	295.21	83.45	81.45
GA	-11697.76573	85.88	108.17	32.17	92.69

Other algorithms, such as Gradient Boosting Regression (GBR), Decision Trees (DT), and Support Vector Machines (SVC), also perform reasonably well, with respectable MAE, RMSE, MAPE, and accuracy scores and relatively high R-squared values.

Algorithms such as K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), Ensemble Methods, and Genetic Algorithms (GA) perform poorly in estimating CO2 emissions. These models have lower R-squared values and greater MAE, RMSE, MAPE, and accuracy scores when compared to the top-performing models.

When choosing a decent algorithm/model for computing CO2 emissions in Thailand, it is critical to consider these evaluation factors. Because of their high R-squared values, low error scores, and high accuracy, the Random Forest (RF) and XGBoost (XGB) models appear to be the most promising alternatives for accurate CO2 emission calculations. These statistics provide an evaluation of the performance of each algorithm/model in computing CO2 emissions. Models with greater R-squared values, lower MAE, RMSE, MAPE, and better Accuracy are thought to be more accurate and dependable.

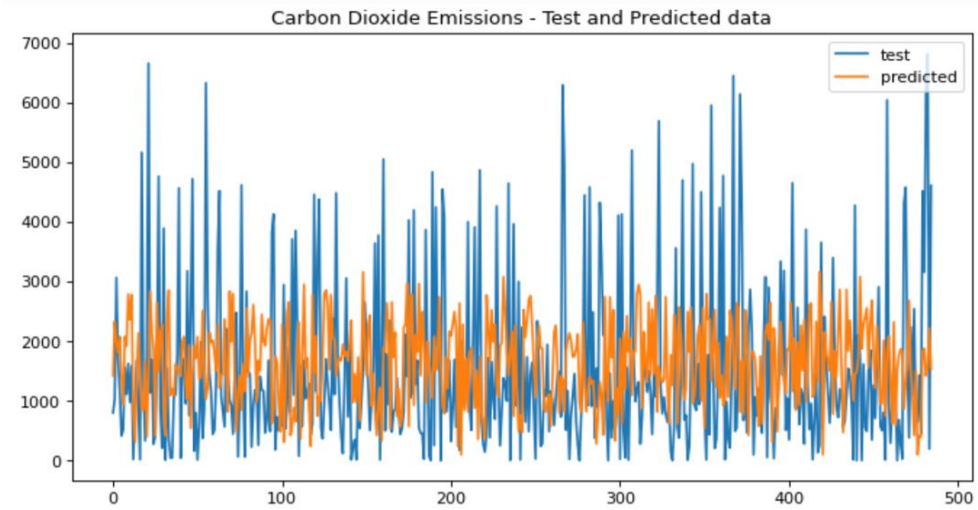

Figure 7. Emission value measured and forecasted

We choose to use XGBoost (XGB) models for CO2 forecasting and have obtained the results. The blue line in Figure 7 reflects the measured emission values, whereas the orange line indicates the anticipated emission values. This graph illustrates the difference between the actual measured emissions and the projected levels.

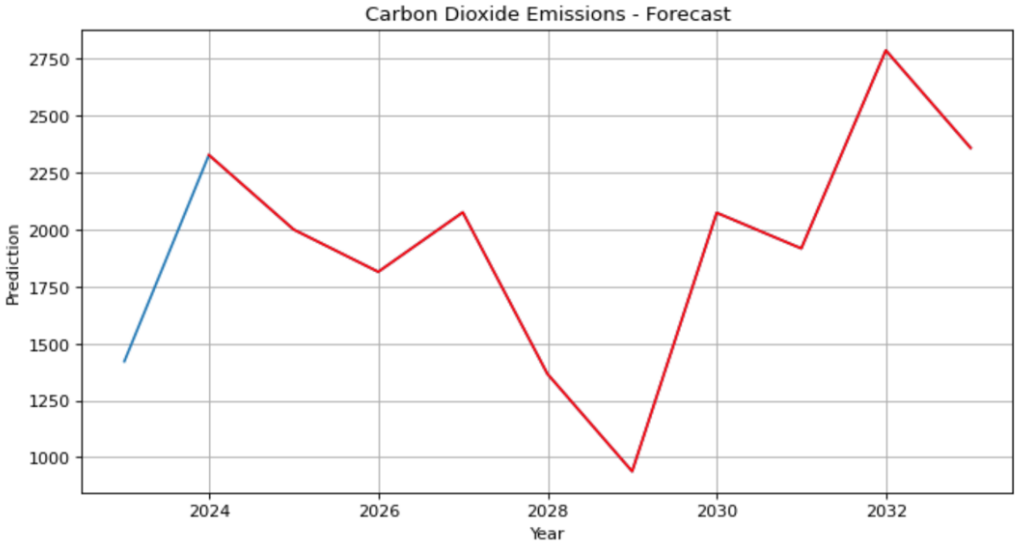

Figure 8. CO2 emissions in Thailand by XGBoost model for 2023-2033.

The forecasting model represented in Figure 8 has been used to project increases in CO2 emissions in Thailand over the next decade, from 2023 to 2033. This forecast is based on government policies. The graph clearly shows a falling trend from 2024 to 2029, followed by a rise in emissions from 2029 to 2032.

Overall, as indicated by high R-squared values and low error metrics, the Random Forest (RF), XGBoost (XGB), and Decision Tree (DT) models predict CO2 emissions in Thailand with good accuracy and reliability. These models, which employ machine learning techniques, are suitable for estimating CO2 emissions in Thailand.

6. CONCLUSION

This research projected CO2 emissions in Thailand using a number of machine learning algorithms/models and assessed their effectiveness using a variety of metrics. The findings shed information on how well these models predict CO2 emissions in the country. Random Forest (RF) and XGBoost (XGB) outperformed the other algorithms/models examined. They achieved high R-squared values while exhibiting low mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and great precision. These findings demonstrate the efficacy of ensemble techniques like RF and XGB in capturing complex relationships and projecting CO2 emissions in Thailand. Other algorithms/models did well as well, with reasonable high R-squared values and acceptable error scores, including Gradient Boosting Regression (GBR), Decision Trees (DT), and Support Vector Machines (SVC). While they may not beat RF and XGB in terms of performance, depending on the requirements and constraints, they can still be viable alternatives. However, algorithms/models such as K-Nearest Neighbors (KNN), Principal Component Analysis (PCA), Ensemble Methods, and Genetic Algorithms (GA) performed poorly in estimating CO2 emissions. They produced lower R-squared values and greater error scores, indicating difficulties in capturing the data's underlying patterns and relationships. The superior performance of RF and XGB suggests that they have the ability to guide policy decisions and promote sustainable development in Thailand.

Our research is limited to CO2 emissions data and does not account for many confounding factors such as gas and oil prices, annual car sales, pandemic effects, legislation changes, and so on. These variables can have a substantial impact on CO2 emissions. While our current study may provide some insights into the underlying trends in CO2 emissions, I believe that its practical utility for generating accurate predictions or policy recommendations is limited. These extra elements must be considered and incorporated in order to construct a more thorough and realistic model. Their inclusion may result in more accurate estimates and a better understanding of the numerous dynamics that influence CO2 emissions.

Further research could concentrate on improving the forecasting models by including new variables such as weather patterns and technological improvements. Furthermore, studying the effects of certain policy actions on CO2 emissions could help us better understand successful mitigation techniques.

ACKNOWLEDGMENTS

We would like to thank Suan Dusit University for their assistance and resources. We also like to thank the blind reviewers for their crucial efforts and feedback. Their knowledge and insights have considerably improved the quality of our research.

REFERENCES

[1] M. Mirzaei *et al.*, "Assessment of soil CO2 and NO fluxes in a semi-arid region using machine learning approaches," *J Arid Environ*, vol. 211, Apr. 2023, doi: 10.1016/j.jaridenv.2023.104947.

[2] T. T. Nguyen, "Reduction of Emission Cost, Loss Cost and Energy Purchase Cost for Distribution Systems With Capacitors, Photovoltaic Distributed Generators, and Harmonics," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 11, no. 1, Feb. 2023, doi: 10.52549/ijeei.v11i1.4103.

[3] Achiraya Chaichaloempreecha, Puttipong Chunark, and Bundit Limmeechokchai, "Assessment of Thailand's Energy Policy on CO2 Emissions : Implication of National Energy Plans to Achieve NDC Target," *International Energy Journal*, pp. 47–60, 2019.

[4] P. Sutthichaimethee and H. A. Wahab, "A forecasting model in managing future scenarios to achieve the sustainable development goals of Thailand's environmental law: Enriching the path analysis-VARIMA-OVi model," *International Journal of Energy Economics and Policy*, vol. 11, no. 4, pp. 398–411, 2021, doi: 10.32479/ijeep.9693.

[5] S. Phoualavanh, "LOW CARBON EMISSIONS IN THE TRANSPORT SECTORS IN THAILAND AND LAO P.D.R," bangkok, 2017.

[6] Y. Meng and H. Noman, "Predicting CO2 Emission Footprint Using AI through Machine Learning," *Atmosphere (Basel)*, vol. 13, no. 11, Nov. 2022, doi: 10.3390/atmos13111871.

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

[8] L. P. S. Freitas *et al.*, "Forecasting the spatiotemporal variability of soil CO2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks," *Environ Monit Assess*, vol. 190, no. 12, Dec. 2018, doi: 10.1007/s10661-018-7118-0.

[9] V. Ratanavaraha and S. Jomnonkwao, "Trends in Thailand CO2 emissions in the transportation sector and Policy Mitigation," *Transp Policy (Oxf)*, vol. 41, pp. 136–146, Jul. 2015, doi: 10.1016/j.tranpol.2015.01.007.

[10] I. I. Monisha, N. Mehtaj, and Z. I. Awal, "A STEP TOWARDS IMO GREENHOUSE GAS REDUCTION GOAL: EFFECTIVENESS OF MACHINE LEARNING BASED CO2 EMISSION PREDICTION MODEL," 2022. [Online]. Available: https://ssrn.com/abstract=4445120

[11] N. Ma, W. Y. Shum, T. Han, and F. Lai, "Can Machine Learning be Applied to Carbon Emissions Analysis: An Application to the CO2 Emissions Analysis Using Gaussian Process Regression," *Front Energy Res*, vol. 9, Sep. 2021, doi: 10.3389/fenrg.2021.756311.

[12] H. J. Touma, M. Mansor, M. S. A. Rahman, H. Mokhlis, and Y. J. Ying, "Influence of Renewable Energy Sources on Day Ahead Optimal Power Flow Based on Meteorological Data Forecast Using Machine Learning: A Case Study of Johor Province," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 11, no. 1, pp. 225–240, Mar. 2023, doi: 10.52549/ijeei.v11i1.4115.

[13] Y. Benlachmi, A. El Airej, and M. L. Hasnaoui, "Fruits Disease Classification using Machine Learning Techniques," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 917–929, Dec. 2022, doi: 10.52549/ijeei.v10i4.3907.

[14] I. M. K. Karo, M. F. M. Fudzee, S. Kasim, and A. A. Ramli, "Sentiment Analysis in Karonese Tweet using Machine Learning," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 219–231, Mar. 2022, doi: 10.52549/ijeei.v10i1.3565.

[15] P. S. Kammath, V. V. Gopal, and J. Kuriakose, "Detection of Bundle Branch Blocks using Machine Learning Techniques," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 559–566, Sep. 2022, doi: 10.52549/ijeei.v10i3.3852.

[16] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments," *Energies (Basel)*, vol. 14, no. 16, Aug. 2021, doi: 10.3390/en14165196.

[17] Y. Meng and H. Noman, "Predicting CO2 Emission Footprint Using AI through Machine Learning," *Atmosphere (Basel)*, vol. 13, no. 11, Nov. 2022, doi: 10.3390/atmos13111871.

[18] M. R. Joel, M. V Srinath, and M. R. Joel, "Optimizing profit by retaining customers using machine learning techniques," *Sci Trans Environ Technovation*, vol. 14, no. 4, 2021, doi: 10.56343/stet.116.014.004.007.

[19] K. Indira and U. Sakthi, "A hybrid intrusion detection system for sdwsn using random forest (RF) machine learning approach," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 275–284, 2020.

[20] N. Rezaeian and G. Novikova, "Persian text classification using naive bayes algorithms and support vector machine algorithm," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 8, no. 1, pp. 178–188, Mar. 2020, doi: 10.11591/ijeei.v8i1.1696.

[21] C. Saleh, N. R. Dzakiyullah, and J. B. Nugroho, "Carbon dioxide emission prediction using support vector machine," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Mar. 2016. doi: 10.1088/1757-899X/114/1/012148.

[22] R. Pérez-Suárez and A. J. López-Menéndez, "Growing green? Forecasting CO2 emissions with Environmental Kuznets Curves and Logistic Growth Models," *Environ Sci Policy*, vol. 54, pp. 428–437, 2015, doi: 10.1016/j.envsci.2015.07.015.

[23] P. Boontome, A. Therdyothin, and J. Chontanawat, "Forecasting Carbon Dioxide Emission and Sustainable Economy: Evidence and Policy Responses," *International Journal of Energy Economics and Policy*, vol. 9, no. 5, pp. 55–62, 2019, doi: 10.32479/ijeep.7918.

[24] P. Sutthichaimethee, A. Chatchorfa, and S. Suyaprom, "A forecasting model for economic growth and CO2 emission based on industry 4.0 political policy under the government power: Adapting a second-order autoregressive-SEM," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 5, no. 3, 2019, doi: 10.3390/joitmc5030069.

[25] T. R, "Thailand CO2 Emission dataset." [Online]. Available: https://www.kaggle.com/datasets/thaweewatboy/thailand-carbon-emission-statistics

[26] M. Xu, S. Grant-Muller, and Z. Gao, "Evolution and assessment of economic regulatory policies for expressway infrastructure in China," *Transp Policy (Oxf)*, vol. 41, pp. 42–49, Jul. 2015, doi: 10.1016/j.tranpol.2015.03.007.

[27] J. Mei, D. He, R. Harley, T. Habetler, and G. Qu, "A Random Forest Method for Real-Time Price Forecasting in New York Electricity Market."

[28] Zeal Education Society, C. A. and R. Zeal Institute of Business Administration, Institute of Electrical and Electronics Engineers. Pune Section, and Institute of Electrical and Electronics Engineers, "2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) : Zeal Education Society, Pune, India, Feb 24-26, 2017.,"

[29] S. O. Abdulsalam, M. O. Arowolo, Y. K. Saheed, and J. O. Afolayan, "Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 431–440, Jun. 2022, doi: 10.52549/ijeei.v10i2.2985.

[30] Lior. Rokach and Oded. Maimon, *Data mining with decision trees : theory and applications*. World Scientific, 2008.

[31] C. Skittides and W. G. Früh, "Wind forecasting using Principal Component Analysis," *Renew Energy*, vol. 69, pp. 365–374, 2014, doi: 10.1016/j.renene.2014.03.068.

[32] M. A. Al-Hagery, E. I. Al-Fairouz, and N. A. Al-Humaidan, "Improvement of alzheimer disease diagnosis accuracy using ensemble methods," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 8, no. 1, pp. 132–139, Mar. 2020, doi: 10.11591/ijeei.v8i1.1321.

[33] H. Wu and D. Levinson, "The ensemble approach to forecasting: A review and synthesis," *Transp Res Part C Emerg Technol*, vol. 132, Nov. 2021, doi: 10.1016/j.trc.2021.103357.

[34] THAWEEWAT R, "Thailand CO2 Emission [1987-2022]," Thailand 36-Year CO2 Emission Monthly Breakdown by Sources and Fuel Types.

BIOGRAPHY OF AUTHORS

Siriporn Chimphlee holds a PhD in Computer Science from Malaysia's University of Technology and is currently an Assistant Professor in the Data Science and Analytics department at Thailand's Suan Dusit University. Data mining, intrusion detection, web mining, and information technology are among her research interests. Dr. Siriporn has multiple papers published in prestigious publications and has actively participated in international conferences. She is enthusiastic in exploring new areas in computer science and is always looking for ways to contribute to the field through her research.

Witcha Chimphlee received his PhD in Computer Science from Malaysia's University of Technology. He is currently an Assistant Professor in the Data Science and Analytics department at Suan Dusit University in Thailand. Data mining, intrusion detection, and soft computing are among his research interests. He has multiple papers published in peer-reviewed journals and has attended numerous international conferences. He is a motivated researcher with a strong desire to advance computer science through his efforts.