International Journal of Analysis and Applications

Air Quality Forecasting Based on Socio-Economic and Environmental Indicators: Combining Statistical and Machine Learning Techniques

Ibrahim G. Khattab[1,2], Mohamed C. Ali[3], Mohamed R. Abonazel[4,*], Hany M. Elshamy[5], Abeer R. Azazy[6]

[1]*Department of Statistics, Mathematics, and Insurance, Faculty of Business, Alexandria University, Alexandria, Egypt*

[2]*Department of Business Administration, Gulf Colleges, Saudi Arabia*

[3]*Faculty of Business Administration, Deraya University, Minya, Egypt*

[4]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt*

[5]*Department of Economics, Faculty of Commerce, Tanta University, Tanta, Egypt*

[6]*Al-Alsun Higher Institute for Tourism, Hotels and Computer, Cairo, Egypt*

**Corresponding author:* mabonazel@cu.edu.eg

Abstract. This research aims to model and predict greenhouse gas (GHG) emissions in Saudi Arabia by examining their association with crucial socio-economic and environmental factors. Utilizing annual data from 1980 to 2023, the study focuses on three emission variables as dependent variables: carbon dioxide (CO_2) emissions from the power sector, methane (CH_4) emissions from the power sector, and nitrous oxide (N_2O) emissions from industrial activities. The independent variables include agricultural land area, urban population, GDP growth, exports, trade openness, foreign direct investment, and manufacturing output. A comparative assessment of various modeling approaches Ordinary Least Squares (OLS), Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net (Enet), Random Forest (RF), and a new hybrid method that merges Elastic Net and Random Forest (ENRF) was performed. The performance of the models was evaluated based on Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results indicated that the ENRF model consistently surpassed both traditional and machine learning

Received Jun. 21, 2025

2020 Mathematics Subject Classification. 62J07, 62P12, 91B76.

Key words and phrases. big data; OLS; ridge regression; LASSO; elastic net; environmental economics; random forests; Saudi Arabia; socio-economic factors.

https://doi.org/10.28924/2291-8639-23-2025-183

ISSN: 2291-8639

techniques, achieving the lowest MSE and RMSE values. The outcomes underscore the efficacy of hybrid statistical and machine learning models in reliably predicting emissions and informing environmental policy in complex, big data contexts.

1. Introduction

Air pollution has become a critical global issue due to its significant effects on human health, environmental conditions, and climate change. The rapid growth of cities, industrial activities, and vehicle emissions has resulted in dangerously poor air quality in numerous urban areas, especially in developing countries like China. Fine particulate matter (PM2.5) and trace gases such as nitrogen dioxide (NO_2), carbon monoxide (CO), and sulfur dioxide (SO_2) are strongly linked to respiratory and heart diseases, highlighting the urgent need for effective forecasting systems.

Koçak [1] carried out an in-depth assessment of five different machine learning models aimed at forecasting hourly levels of particulate matter (PM). By utilizing real-world data that included pollutant concentrations and meteorological factors, the research effectively captured short-term air pollution trends. The Ridge Regression model attained a moderate R^2 of 0.44 for $PM_{2.5}$ and a strong R^2 of 0.91 for PM_{10}. Support Vector Regression demonstrated superior performance in predicting $PM_{2.5}$ ($R^2 = 0.83$) but was less successful with PM_{10}. The Random Forest and Extra Trees Regression models showed strong performance, especially for PM_{10} ($R^2 = 0.75$). Extreme Gradient Boosting also produced competitive outcomes for both $PM_{2.5}$ and PM_{10}, achieving R^2 values of 0.80 and 0.81, respectively. Additionally, the study employed the AirQ+ model to evaluate the health impacts of $PM_{2.5}$ exposure, indicating an average attributable proportion of 10.2% (with a range from 6.5% to 13.2%) in long-term mortality rates. These results underscore the necessity for customized strategies in air quality management and safeguarding public health.

In their thorough review, [2] examined the efficacy and theoretical foundations of penalized regression and machine learning methods for high-dimensional data, highlighting standard techniques such as Ridge, LASSO, and Elastic Net, while emphasizing their importance in addressing multicollinearity and selecting variables. They also explored machine learning techniques, including random forests, support vector machines, and neural networks, showcasing their ability to model complex non-linear interactions. A crucial element of their research was the introduction of two hybrid methods, LASSOPBRF and EnetRARTEN, which skillfully combined statistical regularization with ensemble learning to enhance model accuracy and variable identification. Similarly, [3] presented these two innovative hybrid estimation methods—LASSOPBRF, which integrates LASSO with post-selection boosting of random forest trees, and EnetRARTEN, which utilizes elastic net to refine and consolidate random forest trees—demonstrating through Monte Carlo simulations and a case study on air quality that these models outperform traditional techniques in terms of MSE and RMSE, particularly in cases involving multicollinearity and outliers. Despite these advancements, challenges remain, such as inconsistent evaluations of hybrid models across different urban datasets, insufficient consideration of the temporal and spatial variations in pollutant levels, and limited interpretability for policy-related decisions. To address these issues, the current study proposes an enhanced hybrid model that merges elastic net with random forest to improve predictions of urban air quality. [4] provides a detailed examination of air pollution in China, outlining its complex origins, health impacts, and governmental responses. The authors highlight that China experiences both traditional and photochemical smog due to a combination of industrial emissions, dependence on coal for energy, rapid growth in vehicle usage, and the widespread adoption of solid fuels in households. Exposure to fine particulate matter ($PM_{2.5}$) has led to severe health consequences, with an estimated 1.5 million premature deaths reported in 2015 alone. Although there have been slight improvements in recent years, particularly in reducing sulfur dioxide and nitrogen oxides emissions, pollution levels in urban areas still exceed both national and international health standards. The study emphasizes significant regional disparities and warns that rural communities relying on biomass and coal remain particularly vulnerable. Despite ongoing challenges, the authors note that national policies, technological advancements, and growing public awareness have initiated a gradual shift towards improved air quality and environmental governance.

In 2020, [5] a significant achievement occurred, marking an extraordinary moment in Singgih's path. Comprehensive research has revealed a strong connection between exposure to air pollutants—such as nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), and particulate matter (PM)—and the onset of heart and lung diseases. In response, many local governments have implemented real-time air quality monitoring systems to inform public health initiatives, while global institutions like Peking University, Christchurch, and Los Angeles have effectively utilized data collection and analysis tools to understand and disseminate air quality information. To enhance the Post-Selection Boosting Random Forest (PBRF) algorithm, [6] proposed a new method called "Reducing and Aggregating Random Forest Trees Using an Elastic Net" (RARTEN), which integrates penalized regression techniques into a three-step approach: prediction with a random forest, optimization through elastic net regularization to minimize the number of trees, and aggregation of the selected trees. Simulations and validations with actual data demonstrated RARTEN’s effectiveness, achieving improvements of 7%, 5%, and 8.5% in linear, nonlinear, and noisy models, respectively, along with an approximate 16% reduction in error, exceeding traditional random forest and established penalized regression models. Meanwhile, [7] unveiled a groundbreaking ensemble learning framework designed to predict the binding strength and kinetic behavior of small molecules interacting with the HIV-1 TAR RNA structure. By generating a training dataset from small molecules tested against the RNA construct and employing surface plasmon resonance for binding assessment, they developed the first validated 2D QSAR model for RNA-ligand interactions. This innovation provides a crucial basis for developing RNA-targeted ligands with less reliance on high-resolution structural data. In a separate study, [8] addressed the inverse problem in geophysics using Random Forest Regression (RFR) to infer subsurface physical properties from synthetic magnetotelluric (MT) and DC resistivity data. By crafting multiple decision trees from equal subsets of data and avoiding iterative forward modeling, RFR produced predictions that closely aligned with the actual model parameters and outperformed other methods like Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Ridge Regression (RR), and Grey Wolf Optimization (GWO). Lastly, [9] developed a machine learning model based on crystallographic protein-ligand complexes to predict binding affinity, using energy terms derived from MolDock and PLANTS scoring systems and integrating IC50 data. The resulting polynomial scoring functions demonstrated superior predictive accuracy compared to traditional scoring methods, including AutoDock4, AutoDock Vina, MolDock, and PLANTS, particularly in predicting CDK2 binding affinities. Monitoring and anticipating air quality have grown increasingly crucial due to the rising health risks linked to fine particulate matter (PM2.5) and trace gases in urban settings. [10] examined long-term emission forecasts and their impact on PM2.5 pollution levels in India between 2015 and 2050, highlighting the importance of source-targeted strategies to mitigate air pollution. Aside from standard environmental modeling, advancements in technology have enabled real-time and localized air quality assessments. [11] introduced a mobile microscopy system combined with machine learning for easy air quality evaluation, while [12] employed Internet of Things (IoT) networks for continuous monitoring and forecasting through mobile and stationary sensors. A comparative study conducted by [13] underscored regional variations by analyzing trace gases and particulate matter in Delhi and Beijing, stressing the need for customized air pollution management strategies.

Both the industrial and agricultural sectors play significant roles in contributing to greenhouse gas (GHG) emissions in Saudi Arabia, posing a major challenge to the nation's sustainability objectives. [14] conducted an in-depth study of the Industrial Processes and Product Use (IPPU) sector, revealing that the cement, petrochemical, and iron and steel industries are the main contributors to CO_2 emissions, which together account for more than 80% of the sector's total emissions. Projections suggest that, without intervention, IPPU emissions could rise to between 199 and 426 $MtCO_2eq$ by the year 2050, highlighting the urgent need for mitigation strategies such as enhancing energy and material efficiency, implementing carbon capture technologies, and promoting recycling. In addition, [15] explored the long-term relationship between CO_2 emissions and economic indicators using ARDL and FMOLS models, finding that the expansion of agricultural land, energy usage, and economic growth all have significant positive impacts on CO_2 emissions. These results emphasize the necessity of aligning both industrial and agricultural advancement with the objectives of Saudi Vision 2030 and the Saudi Green Initiative to realize a low-carbon, sustainable future.

There are other studies that have studied the different factors that affect air quality and GHG emissions, such as [16, 17].

The main aim of this study is to evaluate the effectiveness of hybrid modeling techniques in predicting greenhouse gas (GHG) emissions in Saudi Arabia, using actual national-level data from 1980 to 2023. In particular, the research analyzes the efficacy of a proposed hybrid model that combines Elastic Net (Enet) with Random Forest (RF) to forecast three dependent variables: carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) emissions. These forecasts are compared with those generated by traditional statistical models, such as Ordinary Least Squares (OLS), Ridge, LASSO, as well as isolated machine learning methods like RF. Furthermore, the study aims to pinpoint the most significant socio-economic and environmental factors influencing emission levels for various emission types.

This research is driven by three key questions. Firstly, how does the proposed ENRF hybrid model stack up against traditional statistical and machine learning models in terms of accuracy and reliability for emission predictions? Our results reveal that the hybrid model surpasses the performance of individual methods by capitalizing on the advantages of both regularization and ensemble learning. Secondly, how effectively do these models tackle the difficulties posed by high dimensional data, multicollinearity, and dataset variability? The ENRF model exhibits a greater ability to address these challenges by employing feature selection, penalized regression, and tree-based interactions. Lastly, which independent variables such as urban population growth, agricultural land area, foreign direct investment, and manufacturing output have the most significant impact on each type of emission?

In order to answer these questions, we employed a 44-year panel dataset for Saudi Arabia, which included various economic, demographic, and environmental metrics. Following the application of preprocessing methods and selection of variables, we developed six different models: OLS, Ridge, LASSO, Enet, RF, and the newly proposed ENRF hybrid. We assessed model performance through Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and goodness-of-fit metrics. The ENRF model consistently demonstrated the lowest MSE and RMSE figures, highlighting its exceptional predictive ability and robustness against data imperfections in long-term emission forecasting.

The subsequent sections of this paper are organized as follows: Section 2 describes the materials and methods employed in this research, detailing the proposed hybrid approach. Section 3 thoroughly examines the results and explores their significance. Lastly, Section 4 summarizes the key findings of our research.

2. Material and methods

This section describes the materials, data sources, experimental design, and analysis techniques used to meet the goals of this research. The chosen methodology guarantees the reliability and accuracy of the findings.

2.1 Data Sources and Description

This study employed a longitudinal dataset that included 44 annual observations from 1980 to 2023, encompassing essential environmental, economic, and demographic indicators pertinent to greenhouse gas (GHG) emissions in Saudi Arabia. The main goal was to forecast the levels of three dependent emission variables carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) by utilizing a variety of socio-economic and industrial predictors. These predictors consist of agricultural land area, urban population size, GDP growth rate, trade openness, manufacturing value added, exports of goods and services, and foreign direct investment outflows. All variables were represented as continuous numeric data, and the initial preprocessing steps involved eliminating redundant or incomplete features to maintain data quality and consistency throughout the models. This well-structured dataset formed the basis for training and assessing various statistical and machine learning models designed to enhance the accuracy of national emission predictions.

The dataset for CO_2, CH_4, and N_2O emissions was provided by the International Energy Agency (https://www.iea.org/data-and-statistics), while other variables were provided through the official World Bank database (https://data.worldbank.org/).

2.2 Methodological Framework

The research evaluated various statistical and machine learning approaches, along with the suggested method.

2.2.1 Ordinary Least Squares (OLS)

OLS estimators are utilized to obtain or approximate numerical values, model a data set, and describe the statistical characteristics of the estimates. The least-squares estimator is expressed by

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y \tag{1}$$

Where

$\hat{\beta}^{OLS}$: The vector of estimated coefficients using OLS.

X: Matrix where each row is an observation, and each column is an independent variable

X': The transpose of X.

$(X'X)^{-1}$: The inverse of $(X'X)$

$(X'X)$: The non-singular matrix.

Y: The vector of observed dependent variable values.

2.2.2 Ridge Regression

Ridge regression decreases the magnitude of the regression coefficients by applying a penalty. The coefficients from ridge regression aim to minimize a residual sum of squares that includes this penalty [18].

$$\hat{\beta}_{ridge} = \overset{argmin}{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\,\beta_j\right)^2 + \lambda\sum_{j=1}^{P}\beta_j^2\right\} \qquad (2)$$

Where

y_i: The dependent variable for observation i.

x_{ij}: The value of the j^{th} independent variable for observation i.

β_0**:** The intercept term.

β_j: The coefficient for independent variable j.

λ: The regularization parameter since $\lambda > 0$

2.2.3 Least Absolute Shrinkage and Selection Operator (LASSO)

When p exceeds n, the ordinary least squares estimator lacks uniqueness and tends to overfit the data significantly. Therefore, it is essential to implement some form of complexity regularization. This discussion will center around regularization using the L_1-penalty. The LASSO estimator is [19]:

$$\hat{\beta}(\lambda) = \overset{\arg min}{\beta}\left(\frac{\|Y-\boldsymbol{X}\beta\|_2^2}{2n} + \lambda\|\beta\|_1\right) \qquad (3)$$

where $\|Y - \boldsymbol{X}\beta\|_2^2$=$\sum_{i=1}^{n}(y_i - (\boldsymbol{X}\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$ and where $\lambda > 0$ is make the estimator has the proper makes does variable selection in the sense that $\hat{\beta}(\lambda) = 0$ for some j's (depending on the choice of λ) and $\hat{\beta}_j(\lambda)$can be through it cans a shrunken least squares estimator; hence, the name Least Absolute Shrinkage and Selection Operator (LASSO

2.2.4 Naive Elastic Net

[20] presented the Elastic Net as an innovative method for regularization and variable selection in linear regression, which has significantly impacted the field. The Elastic Net is a mathematical framework that combines L_1 and L_2 regularization techniques in a linear fashion, effectively overcoming some limitations associated with Lasso and Ridge regression methods. This approach provides distinct advantages in situations where the number of predictors (p) far

exceeds the number of observations (n), a context where the Lasso method is not suitable. Results from simulation studies showed that the Elastic Net algorithm consistently outperformed the Lasso algorithm while achieving a similar degree of sparsity. Furthermore, the use of Elastic Net regularization leads to a phenomenon known as the "grouping effect," where strongly correlated variables are likely to be included or excluded from the model together. The authors of this study introduced an algorithm called LARS-EN to efficiently compute the regularization paths for the Elastic Net.

Suppose the data set has n observations with p predictors. Let $y = (y_1, \cdots, y_n)^T$ be the response and $X = [X_1| \cdots |X_n]$ be the model matrix, where $x_j = \left(x_{1j}, \cdots, x_{nj}\right)^T$, $j = 1, \ldots, p$ are the predictors. After a location and scale transformation, we can assume the response is centered and the predictors are standardized,

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for }, j = 1, \ldots, p \tag{4}$$

For any fixed non-negative λ_1 and λ_2, we define the naive elastic net criterion

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1, \tag{5}$$

Where

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \text{ and } |\beta|_1 = \sum_{j=1}^{p} |\beta_j| \tag{6}$$

The naive elastic net estimator $\hat{\beta}$ is the minimizer of (3):

$$\hat{\beta} = \arg\min L(\lambda_1, \lambda_2, \beta) \tag{7}$$

The above procedure can be viewed as a penalized least-squares method. Let $\alpha = \frac{\lambda_2}{\lambda_1 +, \lambda_2}$, then solving $\hat{\beta}$ in (3) is equivalent to the optimization problem:

$$\hat{\beta} = \arg\min L(\lambda_1, \lambda_2, \beta), \text{ subject to } (1 - \alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ for some t.} \tag{8}$$

The Elastic Net penalty function is represented as $(1 - \alpha)|\beta|_1 + \alpha|\beta|^2$,. In this framework, the parameter α determines the distribution of penalties between Lasso and Ridge regression. The choice of α impacts the balance between L_1 and L_2 regularization. When α is equal to 1, the Elastic Net method effectively becomes Ridge regression, while setting α to 0 transforms the approach into Lasso regression.

The previously mentioned phenomenon can be demonstrated through a contour plot that represents two dimensions. In Figure 1, the outer contour of this plot outlines the shape of the Ridge penalty, whereas the diamond-shaped curve represents the Lasso penalty. Furthermore, the red solid curve indicates the use of the Elastic Net penalty, with the coefficient α assigned a value of 0.5. The contour plot illustrates a considerable level of convexity along its edges, with the degree of convexity varying according to the parameter α. In this investigation, the researchers employed the "glmnet" package in R version 4.0.0 to implement the Elastic Net approach [21].

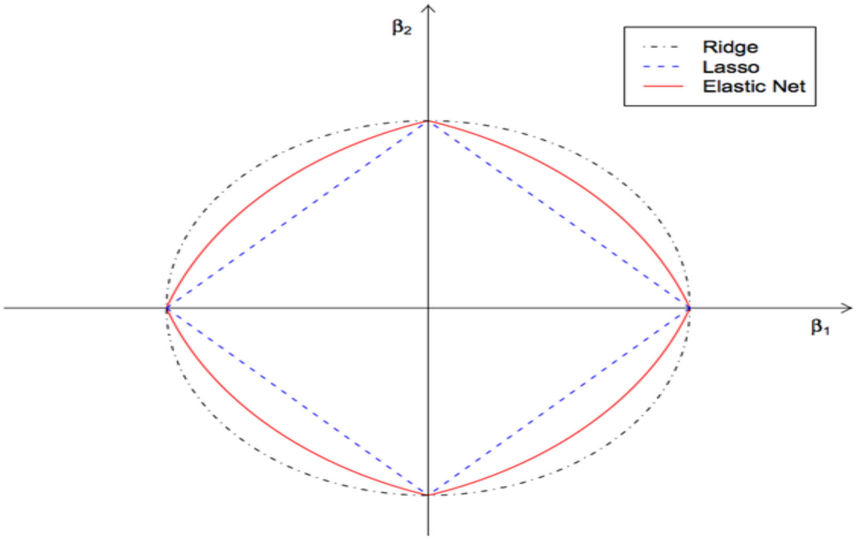


Fig 1: The geometric characteristics of the elastic net penalty

2.2.5 Breiman’s random forest mechanism (RF)

The Random Forest technique is highly esteemed in the realm of machine learning and has proven effective in tackling a range of real-world problems. Its applications span areas such as predicting air quality, cheminformatics, ecology, 3D object recognition, and bioinformatics, among others. Introduced by [22], this technique is classified as an ensemble learning method that utilizes numerous randomized decision trees and amalgamates their predictions through averaging. This approach is especially beneficial in scenarios where the number of variables surpasses the number of available data points.

The Random Forest algorithm is recognized as a powerful computational tool for addressing both regression and classification tasks. Ensemble methods involve the merging of various machine learning strategies to improve prediction accuracy.

The Random Forest strategy involves creating a collection of decision trees, with the entire dataset divided into subsets to assist in making predictions. Each subgroup results in the formation of a unique decision tree within the forest. In the field of machine learning, it is acknowledged that each individual decision tree produces a specific output. Therefore, within a random forest model, the ultimate decision is reached by selecting the most common results from the individual decision trees. For the current study, the authors utilized the Random Forest package in R version 4.0.0 to implement the random forest methodology [23].

Algorithm 1 outlines the process for the Random Forest algorithm mentioned in [24]:
Step 1: Start by building M decision trees.
Step 2: Use the data from the root node as the initial point.
Step 3: Select an attribute and create a logical condition based on that attribute.
Step 4: Direct each outcome of the test to its corresponding child node by passing a subset of examples that satisfy the criteria.
Step 5: Investigate every node within the child structures.
Step 6: Continue this process until the leaf nodes are considered 'pure.'
Step 7: Make the final decision based on the majority vote from the Decision Trees.

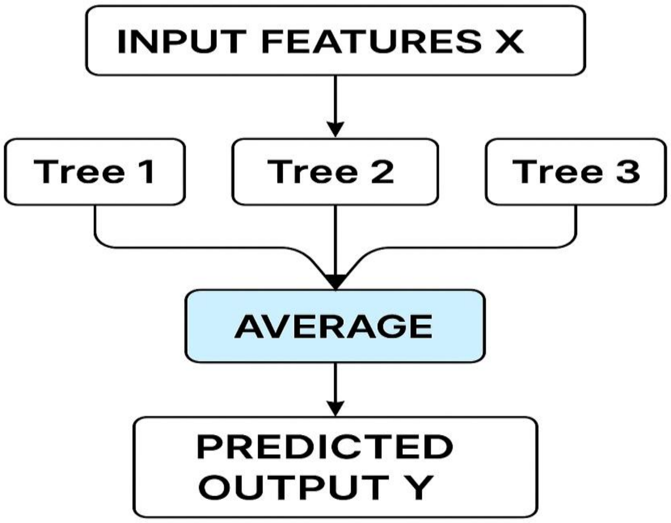


Fig 2: Random Forest applied to regression analysis.

Figure 2 illustrates the functioning of the Random Forest Regressor (RFR), which involves partitioning the training dataset into several subsets in a uniform and random manner. Each subset is subsequently processed by a decision tree that analyzes the data and generates its own prediction. The Random Forest Regression (RFR) algorithm aggregates the predictions from all the individual decision trees and ultimately derives a final outcome by averaging these forecasts.

2.2.6 Proposed Method

In this section, we present a hybrid estimator referred to as Elastic-Net Random Forest (ENRF), which integrates the variable-selection strengths of Elastic Net (ENet) with the non-linear predictive capabilities of Random Forests (RF). This method is akin to the approach adopted by [25], who merged LASSO with neural networks and also combined RF with neural networks to harness the complementary strengths of conventional regularization methods and machine-learning techniques [26].

Elastic Net is particularly advantageous in high-dimensional contexts where the predictor count (p) greatly exceeds the number of observations (n). Its combined ℓ_1–ℓ_2 penalty creates a grouping effect: highly correlated predictors tend to enter or exit the model together, leading to more stable and interpretable coefficient patterns than those achieved with LASSO or Ridge used separately. In contrast, RF improves the performance of individual decision trees by aggregating numerous uncorrelated trees, which enhances predictive accuracy and provides a natural defense against overfitting, even when working with smaller sample sizes. During each split, RF randomly chooses a subset of available variables, further minimizing variance and preventing a small number of influential predictors from dominating the model.

By integrating these two components, we expect that ENRF will outshine conventional statistical methods (Enet) and independent machine-learning algorithms (RF) in both predictive accuracy and stability.

Algorithm 2 outlines the steps for the proposed Elastic Net RF (ENRF) method:

Step 1: The analysis kicks off by implementing an Elastic Net model.

Step 2: The goal is to pinpoint and choose the most relevant variables based on the Elastic Net model.

Step 3: The identified variables are then to be input into the Random Forest algorithm.

2.3 Software and Implementation

This research utilized a variety of supervised statistical and machine learning methods, including Ordinary Least Squares (OLS), Ridge regression, LASSO, Elastic Net (Enet), Random Forest (RF), and a newly proposed hybrid model that combines Enet and RF (ENRF), to estimate greenhouse gas (GHG) emissions levels specifically CO_2, CH_4, and N_2O. The dataset was divided using a standard 80/20 train-test ratio, and hyperparameter optimization was performed through 10-fold cross-validation as part of the model tuning process. The efficacy of the models was assessed using conventional error metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), Additionally, the importance of various features was examined to determine the key socio-economic factors influencing emission fluctuations. All computational tasks were carried out using R and Python, which facilitated robust and reproducible modeling processes.

3. Results and Discussions

This study employs a nationally consolidated dataset that includes 44 annual records from Saudi Arabia, covering the years from 1980 to 2023. The dataset features three dependent variables that denote greenhouse gas emissions: carbon dioxide (CO_2) emissions from the power sector, methane (CH_4) emissions from the power industry, and nitrous oxide (N_2O) emissions from industrial activities. These emissions significantly contribute to environmental degradation and are integral to discussions surrounding global climate change.

Alongside the emission variables, the dataset includes an extensive array of macroeconomic and environmental metrics. These encompass agricultural land area (in square kilometers), urban population, GDP growth rate, trade as a percentage of GDP, manufacturing value added, net outflows of foreign direct investment (FDI), and total exports of goods and services. These variables were chosen based on their expected relevance to emissions patterns and sensitivity to policy changes. All variables underwent numerical encoding, and preprocessing measures were taken to eliminate multicollinear or redundant predictors, thus enhancing the robustness of the modeling process.

Even though the dataset is temporal, rather than spatial, it reflects long-term trends and the effects of policies, allowing for the examination of national patterns in emission generation and management. Its annual frequency supports macro-level forecasting, providing valuable insights for climate policy, industrial planning, and assessments of sustainability in the context of Saudi Arabia.

The main aim of this research is to assess the forecasting ability of independent variables in predicting GHG emission levels and to compare the efficacy of various predictive models. To tackle potential issues such as multicollinearity and outliers, the study employs several advanced modeling techniques, including OLS, Ridge Regression, LASSO, Elastic Net (Enet), Random Forest (RF), and the newly proposed hybrid model Elastic Net Random Forest (ENRF). The findings indicate that the ENRF model stands out as the most precise and dependable method, effectively capturing the intricate interactions among variables while minimizing prediction error.

Table 1 presents a description of 10 key variables used in this research to model and predict greenhouse gas (GHG) emissions in Saudi Arabia. These variables were chosen for their significance in economic, environmental, and policy contexts, along with their possible impact on emissions patterns. The dependent variables carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) account for crucial aspects of industrial and energy-related emissions within the nation. The other independent variables comprise agricultural land area, urban population, GDP growth rate, trade as a fraction of GDP, exports, foreign direct investment (FDI) outflows, and manufacturing value added. These indicators represent the macroeconomic framework and industrial activities of the country, which are tightly linked to emissions production. Altogether, this collection of 10 variables offers a solid foundation for analyzing and predicting emission trends at the national level in Saudi Arabia.

Table 2 provides a summary of descriptive statistics for the three main greenhouse gas (GHG) emission variables analyzed in this research, based on 44 annual data points. The mean carbon dioxide (CO_2) emissions from the power sector are roughly 153.26 Mt CO_2e, with a standard deviation of 103.88, indicating significant fluctuations over the years. Methane (CH_4) emissions from the power industry show a lower average at 0.19 Mt CO_2e, while nitrous oxide (N_2O) emissions from industrial activities average 2.19 Mt CO_2e. For all the emission variables, the median values are quite similar to their respective means, signifying approximately symmetrical distributions. The interquartile ranges (IQRs) reflect moderate variation, and there are no extreme outliers that could skew the analysis. These descriptive statistics provide essential insights into emission trends throughout the study period and affirm the reliability of the dataset for predictive modeling.

Table 3 displays the Variance Inflation Factor (VIF) values for the seven independent variables utilized in this research to forecast greenhouse gas emissions in Saudi Arabia. The analysis reveals notable multicollinearity issues [27-31], especially for X_2 and X_7, which show VIF values of 28.62 and 14.49, respectively—significantly exceeding the typical threshold of 10. Such values indicate strong linear correlations with other variables, which could misrepresent coefficient estimates in ordinary least squares regression. Furthermore, X_4 has a VIF of 10.84, indicating a moderate level of multicollinearity that could also affect the performance of the model. Conversely, X_1, X_3, X_5, and X_6 have VIF values below 6, reflecting acceptable levels of correlation. To mitigate these multicollinearity challenges, this research adopts penalized regression techniques like Ridge and Elastic Net, which are ideal for high-dimensional contexts with correlated predictors.

Figure 3 depicts the procedural workflow for projecting greenhouse gas (GHG) emissions in Saudi Arabia through the use of statistical and machine learning techniques. The process initiates with a dataset of national-level emissions, which is subjected to extensive preprocessing. This step includes addressing missing values, identifying and rectifying outliers, and ensuring data quality by removing any anomalies or NaN entries. Following this, the data undergoes normalization, and exploratory analysis is conducted—such as calculating statistical correlations and evaluating skewness among the variables—to improve the interpretability and stability of the models. The cleaned dataset is subsequently divided into training and testing subsets. Predictive models, such as Ordinary Least Squares (OLS), Ridge, LASSO, Elastic Net (Enet), Random Forest (RF), and the proposed hybrid Elastic Net–Random Forest (ENRF), are trained using the training set and assessed with the testing set. The final phase consists of forecasting emissions of CO_2, CH_4, and N_2O and carrying out a comparative analysis of model performance utilizing MSE, and RMSE to determine the most effective prediction approach.

Figure 4 illustrates the correlation matrix among greenhouse gas (GHG) emission variables (CO_2, CH_4, N_2O) alongside the socioeconomic and environmental predictors utilized in this study. The most substantial correlations are evident between the emission variables themselves (Y_1, Y_2, Y_3), with coefficients surpassing 0.99, indicating a high degree of mutual variability and parallel movements over time. Among the predictors, X_2 and X_7 demonstrate strong positive correlations with all emission variables ($r = 0.88$), suggesting that they are likely to have a significant influence on emission forecasting models. X_4 and X_6 also present moderately strong positive correlations ($r =0.57$) with emissions.

These results indicate that while certain predictors are strongly connected possibly leading to multicollinearity, others remain mostly independent. This supports the use of regularization techniques like Ridge and Elastic Net, which can adeptly manage collinear relationships while retaining important predictors for the development of robust models.

Table 1: Variables description

Variable	Description	Type
Y_1	Carbon dioxide (CO_2) emissions from Power Industry (Energy) (Mt CO_2e)	Dependent
Y_2	Methane (CH_4) emissions from Power Industry (Energy) (Mt CO_2e)	Dependent
Y_3	Nitrous oxide (N_2O) emissions from Industrial Processes (Mt CO_2e)	Dependent
X_1	Agricultural land (sq. km)	Independent
X_2	Urban population	Independent
X_3	GDP growth (annual %)	Independent
X_4	Exports of goods and services (current US$)	Independent
X_5	Trade (% of GDP)	Independent
X_6	Foreign direct investment, net outflows (% of GDP)	Independent
X_7	Manufacturing, value added (% of GDP)	Independent

Table 2: Descriptive statistics of each variable

Variable	Sample Size (n)	Min	Max	Mean	Q2 (Median)	Q3 (75%)
Y_1	44	24.727	262.369	129.241	106.000	213.319
Y_2	44	0.033	0.342	0.163	0.133	0.267
Y_3	44	1.013	3.827	2.446	2.266	3.454
X_1	44	869620	1737980	1495978.24	1733970	1736370
X_2	44	3983211	28258016	15148543.16	14014708	21909933.25
X_3	44	-16.109	10.99376	2.122	2.726	5.159
X_4	44	2.32E+10	4.46E+11	1.59794E+11	1.01683E+11	2.52447E+11
X_5	44	49.7135	96.103	72.31186411	68.835	82.103
X_6	44	-0.543	2.8226	0.562	0.208	0.756
X_7	44	3.985	14.788	9.735	9.665	10.693

Table 3: VIF values of each independent variable

Variable	VIF
X_1	5.413028
X_2	28.61843
X_3	1.445944
X_4	10.83749
X_5	4.220568
X_6	2.757556
X_7	14.49032

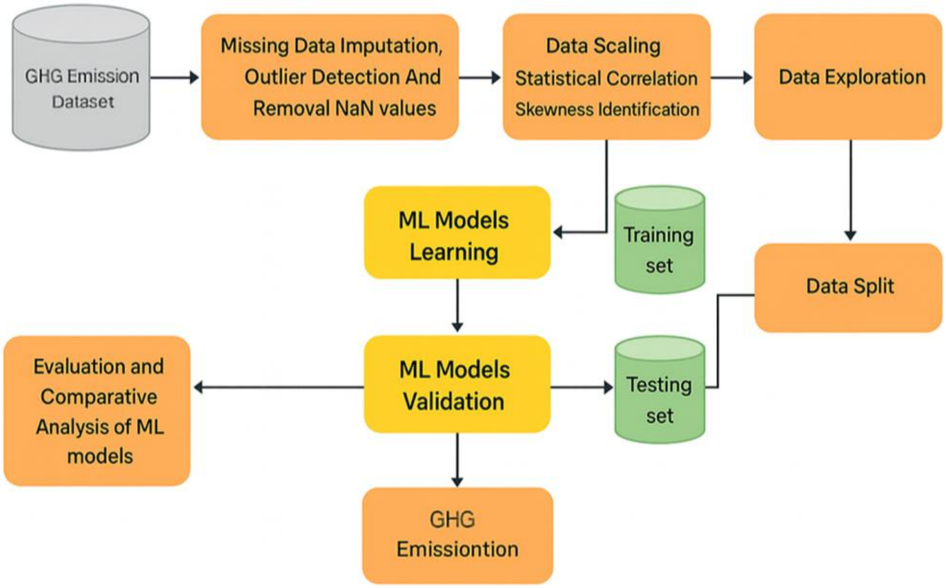
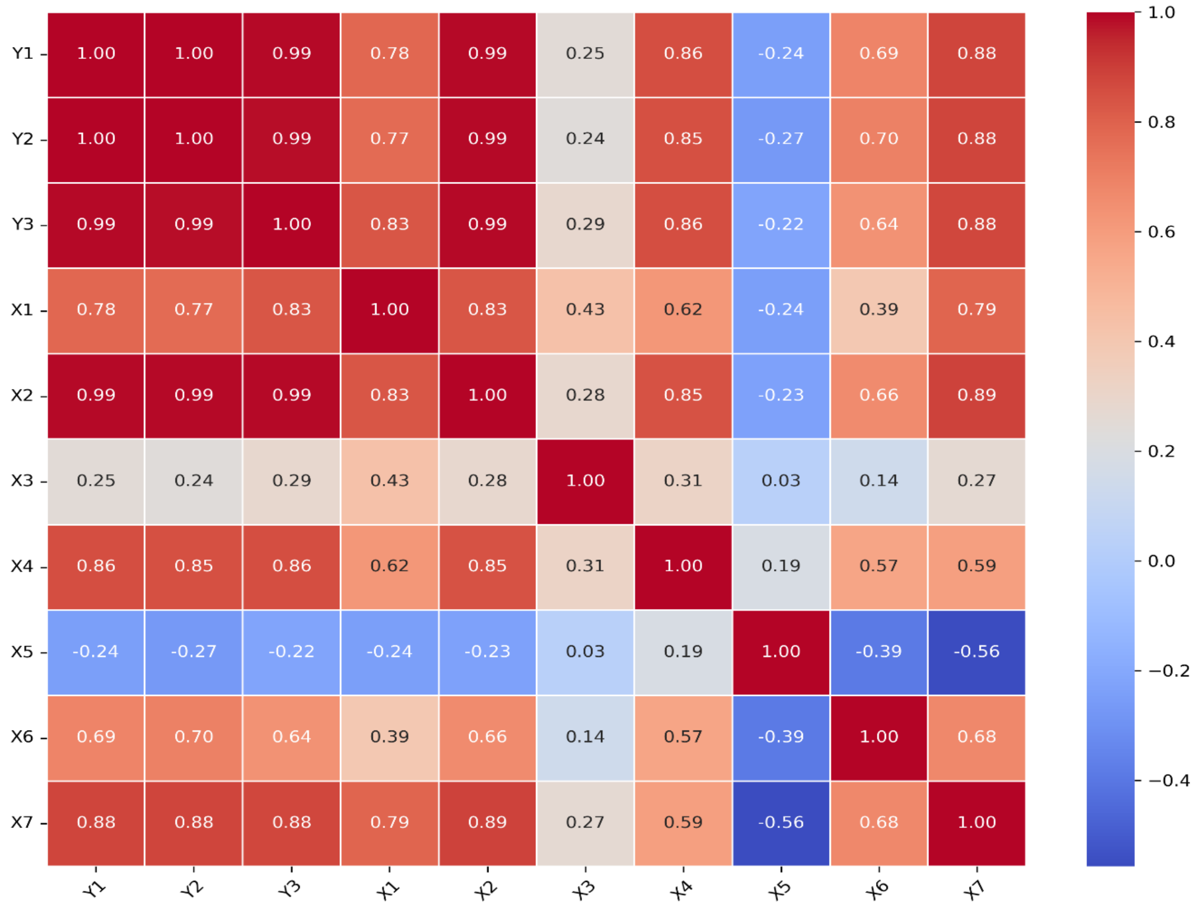

GHG Emission Dataset
Missing Data Imputation, Outlier Detection And Removal NaN values
Data Scaling Statistical Correlation Skewness Identification
Data Exploration
ML Models Learning
Training set
Data Split
Evaluation and Comparative Analysis of ML models
ML Models Validation
Testing set
GHG Emissiontion

Fig 3: Flowchart of proposed methods



	Y1	Y2	Y3	X1	X2	X3	X4	X5	X6	X7
Y1	1.00	1.00	0.99	0.78	0.99	0.25	0.86	-0.24	0.69	0.88
Y2	1.00	1.00	0.99	0.77	0.99	0.24	0.85	-0.27	0.70	0.88
Y3	0.99	0.99	1.00	0.83	0.99	0.29	0.86	-0.22	0.64	0.88
X1	0.78	0.77	0.83	1.00	0.83	0.43	0.62	-0.24	0.39	0.79
X2	0.99	0.99	0.99	0.83	1.00	0.28	0.85	-0.23	0.66	0.89
X3	0.25	0.24	0.29	0.43	0.28	1.00	0.31	0.03	0.14	0.27
X4	0.86	0.85	0.86	0.62	0.85	0.31	1.00	0.19	0.57	0.59
X5	-0.24	-0.27	-0.22	-0.24	-0.23	0.03	0.19	1.00	-0.39	-0.56
X6	0.69	0.70	0.64	0.39	0.66	0.14	0.57	-0.39	1.00	0.68
X7	0.88	0.88	0.88	0.79	0.89	0.27	0.59	-0.56	0.68	1.00

Fig 4: Correlation Matrix

3.4 Discussion

The analysis conducted in this study included the following steps:

1. Outlier Detection

The dataset's anomalies were carefully examined through statistical summaries. The analysis showed that the variables display approximately symmetric distributions with uniform ranges, and no significant outliers were identified. Consequently, no observations were discarded or altered due to extreme values.

2. Multicollinearity Assessment

An assessment of multicollinearity among the independent variables was performed utilizing both a correlation matrix and Variance Inflation Factor (VIF) analysis. The findings indicated that multicollinearity was not a major issue within the dataset. The majority of the variables had VIF values significantly lower than the usual threshold of 10, suggesting a minor level of linear dependency. While a couple of predictors, specifically X_2 ,X_4 and X_7, showed somewhat higher VIF values, these levels were not critical enough to necessitate the exclusion of any variables. Consequently, all predictors were kept for model development. This choice is further supported by the application of penalized regression methods (e.g., Ridge and Elastic Net), which are specifically intended to reduce the effects of multicollinearity in predictive modeling.

3. Missing Value Analysis

A completeness assessment revealed that the dataset had no missing or null values across any of the variables. Therefore, there was no need for imputation or removal processes, maintaining the dataset's integrity and temporal consistency.

4. Data Partitioning

For the purpose of model development and evaluation, the dataset was randomly split into two groups: a training set consisting of 50% of the observations, and a testing set that included the remaining 50%. This division ensures model performance can be evaluated on unseen data while still having enough observations to train robust models.

5. Comparative Model Evaluation

A comparative assessment was conducted across various modeling techniques, including Ordinary Least Squares (OLS), Ridge Regression, LASSO, Elastic Net (Enet), Random Forest (RF), and the proposed hybrid Elastic Net–Random Forest (ENRF). The performance of each model was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Among all tested models, the ENRF consistently yielded the lowest error metrics, confirming its superior predictive ability for estimating CO_2, CH_4, and N_2O emissions.

Table 4: Goodness of fit measures for methods applied to Y_1, Y_2 and Y_3

Response	Criteria	OLS	Ridge	Enet	LASSO	RF	ENRF
Y_1	**MSE**	108.160	195.440	84.824	84.456	219.632	21.437
	RMSE	10.40	13.98	9.21	9.19	14.82	4.63
	#SV	7	7	6	6	7	6
Y_2	**MSE**	0.0185	0.0149	0.0004	0.0002	0.0005	0.0001
	RMSE	0.136	0.122	0.020	0.0128	0.0224	0.011
	#SV	7	7	3	5	7	5
Y_3	**MSE**	0.0231	0.0182	0.0144	0.0151	0.0256	0.0032
	RMSE	0.152	0.135	0.120	0.123	0.160	0.057
	#SV	7	7	4	4	7	4

Table 4 clearly illustrates that the hybrid method of ENRF (Elastic Net and Random Forest) surpasses all other techniques across the three response variables: ***Y_1*, *Y_2* and *Y_3***. In particular, ENRF consistently recorded the lowest Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, showcasing its exceptional predictive accuracy and reliability. Moreover, ENRF employed either fewer or the same number of selected variables (#SV) as the other methods, highlighting its capability in variable selection while maintaining model sparsity. Conversely, both OLS and RF displayed the poorest results especially for ***Y_1*** demonstrating their limited ability to handle complex data structures without the advantages of regularization or ensemble techniques. Although traditional penalized regression methods like Enet and LASSO exhibited some improvements over OLS and Ridge, they were consistently outperformed by the ENRF model. These results emphasize the importance of combining penalized regression with ensemble approaches to improve model generalization and minimize overfitting, particularly in scenarios involving high-dimensional and complex data.

4. Conclusion

Assessing and predicting greenhouse gas (GHG) emissions poses significant yet intricate challenges owing to the evolving interactions among environmental, economic, and industrial factors over time. This study aimed to forecast CO_2, CH_4, and N_2O emissions in Saudi Arabia over a span of forty years by utilizing an integrated modeling framework that merges statistical and machine learning approaches. After conducting thorough data preprocessing, which included cleaning, and identifying outliers, several models were applied, such as Ordinary Least Squares (OLS), Ridge Regression, Elastic Net, Random Forest (RF), and an innovative hybrid model: Elastic Net Random Forest (ENRF). The results confirmed the presence of multicollinearity among certain predictors but no significant outliers. These challenges were effectively addressed using the proposed hybrid ENRF (Elastic Net–Random Forest) method. This approach demonstrated superior predictive performance across multiple metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), outperforming traditional statistical and standalone machine learning models. The findings highlight the robustness of hybrid models in managing high-dimensional and noisy environmental data. By combining the variable selection strength of Elastic Net with the nonlinear learning capabilities of Random Forests, the ENRF model achieved improved accuracy and model stability.

This modeling framework holds significant potential for use by environmental agencies and policymakers in Saudi Arabia, particularly for national-level emissions monitoring and forecasting. Its adaptability makes it well-suited for integration into sustainable development and climate strategy planning. However, this study does present certain limitations. The analysis is geographically restricted to Saudi Arabia and may not generalize directly to other regions with different emission sources, policy landscapes, or climatic conditions.

Future investigations can expand on this research in various significant ways. Firstly, enhancing the temporal resolution of the dataset such as utilizing quarterly or monthly emissions data would enable the analysis of seasonal and short-term emission trends that are frequently overlooked in yearly summaries. Secondly, integrating supplementary data sources like satellite-based remote sensing, sector-specific emission inventories, and real-time industrial output could augment both the spatial and temporal details of emission estimates. Testing the ENRF model in different geographic areas, especially within the Middle East and North Africa (MENA) region, would also be beneficial for assessing its applicability across diverse environmental and economic scenarios. Additionally, incorporating IoT-enabled real-time monitoring systems and advanced deep learning models (e.g., LSTM, CNN) could further improve forecasting accuracy and adaptability. These improvements would render the hybrid modeling framework more resilient and versatile for implementation in national emission monitoring systems and intelligent environmental infrastructure.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] E. Koçak, Comprehensive Evaluation of Machine Learning Models for Real-World Air Quality Prediction and Health Risk Assessment by AirQ+, Earth Sci. Informatics 18 (2025), 447. https://doi.org/10.1007/s12145-025-01941-7.

[2] A. El-Sheikh, M.R. Abonazel, M.C. Ali, A Review of Penalized Regression and Machine Learning Methods in High-Dimensional Data, Egypt. Stat. J. 69 (2025), 250-261. https://doi.org/10.21608/esju.2025.368665.1080.

[3] A.A. El-Sheikh, M.C. Ali, M.R. Abonazel, Development of Two Methods for Estimating High-Dimensional Data in the Case of Multicollinearity and Outliers, Int. J. Anal. Appl. 22 (2024), 187. https://doi.org/10.28924/2291-8639-22-2024-187.

[4] K. Aunan, M.H. Hansen, S. Wang, Introduction: Air Pollution in China, China Q. 234 (2017), 279-298. https://doi.org/10.1017/s0305741017001369.

[5] I. Kristianto Singgih, Air Quality Prediction in Smart City's Information System, Int. J. Informatics, Inf. Syst. Comput. Eng. 1 (2020), 35–46. https://doi.org/10.34010/injiiscom.v1i1.4020.

[6] Z. Farhadi, H. Bevrani, M. Feizi-Derakhshi, Improving Random Forest Algorithm by Selecting Appropriate Penalized Method, Commun. Stat. - Simul. Comput. 53 (2022), 4380-4395. https://doi.org/10.1080/03610918.2022.2150779.

[7] Z. Cai, M. Zafferani, A. Hargrove, Ensemble Learning-Based Quantitative Structure-Activity Relationship Platform Predicts Binding Behavior of RNA-Targeted Small Molecules, ChemRxiv (2021). http://doi.org/10.33774/chemrxiv-2021-czl9p.

[8] A.P. Singh, D. Vashisth, S. Srivastava, Random Forest Regressor for Layered Earth Data Inversion, in: Fall Meeting 2019, American Geophysical Union, #S53D-0483, (2019). https://ui.adsabs.harvard.edu/abs/2019AGUFM.S53D0483V/abstract.

[9] M.B. de Ávila, M.M. Xavier, V.O. Pintro, W.F. de Azevedo, Supervised Machine Learning Techniques to Predict Binding Affinity. A Study for Cyclin-Dependent Kinase 2, Biochem. Biophys. Res. Commun. 494 (2017), 305-310. https://doi.org/10.1016/j.bbrc.2017.10.035.

[10] C. Venkataraman, M. Brauer, K. Tibrewal, P. Sadavarte, Q. Ma, et al. Source Influence on Emission Pathways and Ambient PM2.5 Pollution Over India (2015–2050), Atmospheric Chem. Phys. 18 (2018), 8017-8039. https://doi.org/10.5194/acp-18-8017-2018.

[11] Y. Wu, A. Shiledar, Y. Li, J. Wong, S. Feng, et al. Air Quality Monitoring Using Mobile Microscopy and Machine Learning, Light: Sci. Appl. 6 (2017), e17046-e17046. https://doi.org/10.1038/lsa.2017.46.

[12] D. Zhang, S.S. Woo, Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network, IEEE Access 8 (2020), 89584-89594. https://doi.org/10.1109/access.2020.2993547.

[13] S. Zheng, R.P. Singh, Y. Wu, C. Wu, A Comparison of Trace Gases and Particulate Matter Over Beijing (China) and Delhi (India), Water Air, Soil Pollut. 228 (2017), 181. https://doi.org/10.1007/s11270-017-3360-2.

[14] M.M. Rahman, M.S. Rahman, S.R. Chowdhury, A. Elhaj, S.A. Razzak, et al. Greenhouse Gas Emissions in the Industrial Processes and Product Use Sector of Saudi Arabia—An Emerging Challenge, Sustainability 14 (2022), 7388. https://doi.org/10.3390/su14127388.

[15] J. Binsuwadan, L. Alotaibi, H. Almugren, The Role of Agriculture in Shaping CO_2 in Saudi Arabia: A Comprehensive Analysis of Economic and Environmental Factors, Sustainability 17 (2025), 4346. https://doi.org/10.3390/su17104346.

[16] E.E.M. Ebrahim, M.R. Abonazel, A.E.A. Ahmed, S. Abdel-Rahman, W.A.A. Albeltagy, Analysis of the Economic and Environmental Factors Affecting Co2 Emissions in Egypt: A Proposed Dynamic Econometric Model, Int. J. Energy Econ. Polic. 15 (2025), 152-165. https://doi.org/10.32479/ijeep.19222.

[17] E.E.M. Ebrahim, M.R. Abonazel, O.A. Shalaby, W.A.A. Albeltagy, Studying the Impact of Socioeconomic and Environmental Factors on Nitrogen Oxide Emissions: Spatial Econometric Modeling, Int. J. Energy Econ. Polic. 15 (2025), 248-259. https://doi.org/10.32479/ijeep.18300.

[18] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12 (1970), 55-67. https://doi.org/10.2307/1267351.

[19] H. Zou, T. Hastie, Regularization and Variable Selection via the Elastic Net, J. R. Stat. Soc. Ser. B: Stat. Methodol. 67 (2005), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[20] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Stat. Soc. Ser. B: Stat. Methodol. 58 (1996), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[21] J.H. Friedman et al., glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, R Package Version 1-1, (2010). https://cran.r-project.org/web/packages/glmnet/glmnet.pdf.

[22] L. Breiman, Random Forests, Mach. Learn. 45 (2001), 5-32. https://doi.org/10.1023/a:1010933404324.

[23] S. RColorBrewer, M.A. Liaw, Package 'randomForest', Univ. Calif., Berkeley, (2018). https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[24] Y.M. Abd Algani, M. Ritonga, B.K. Bala, M.S. Al Ansari, M. Badr, A.I. Taloba, Machine Learning in Health Condition Check-Up: An Approach Using Breiman's Random Forest Algorithm, Measurement: Sensors 23 (2022), 100406. https://doi.org/10.1016/j.measen.2022.100406.

[25] M.C. Ali, E.E.M. Ebrahim, M.R. Abonazel, Air Quality Forecasting Using a Modified Statistical Approach: Combining Statistical and Machine Learning Methods, Int. J. Innov. Res. Sci. Stud. 8 (2025), 1321-1335. https://doi.org/10.53894/ijirss.v8i4.8061.

[26] A.A. EL-Sheikh, M.R. Abonazel, M.C. Ali, Proposed Two Variable Selection Methods for Big Data: Simulation and Application to Air Quality Data in Italy, Commun. Math. Biol. Neurosci. 2022 (2022), 16. https://doi.org/10.28919/cmbn/6908.

[27] M.R. Abonazel, A New Biased Estimation Class to Combat the Multicollinearity in Regression Models: Modified Two--Parameter Liu Estimator, Comput. J. Math. Stat. Sci. 4 (2025), 316-347. https://doi.org/10.21608/cjmss.2025.347818.1096.

[28] M.R. Abonazel, I. Dawoud, M.N. Al-Ghamdi, R.A. Farghali, Developing the Generalized Dawoud-Kibria Estimator for the Multinomial Logistic Model: Simulation Study and Application, Sci. Afr. 29 (2025), e02803. https://doi.org/10.1016/j.sciaf.2025.e02803.

[29] M.R. Abonazel, New Modified Two-Parameter Liu Estimator for the Conway–maxwell Poisson Regression Model, J. Stat. Comput. Simul. 93 (2023), 1976-1996. https://doi.org/10.1080/00949655.2023.2166046.

[30] M.N. Al-Ghamdi, M.R. Abonazel, I. Dawoud, Z.Y. Algamal, A.R. Azazy, A New Estimator of the Gamma Regression Model: Theory, Simulation, and Application to Body Fat Data, Commun. Math. Biol. Neurosci. 2025, (2025), 53. https://doi.org/10.28919/cmbn/9149.

[31] M.R. Abonazel, E.E.M. Ebrahim, A.A. Saber, A.R. Azazy, On New Ridge Estimators of the Conway-Maxwell Poisson Model in the Case of Highly Correlated Predictor Variables: Application to Plywood Quality Data, Int. J. Innov. Res. Sci. Stud. 8 (2025), 603–614. https://doi.org/10.53894/ijirss.v8i5.8775.