Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Machine Learning Methodologies for Enhancing Predictive Accuracy in Environmental Impact Studies

Mahimna Shukla[1], Anirudh Gupta[2], Engr Shaista Ashraf Farooqi[3], Cesar D. Galingana[4]
[1]Department of Environmental Protection, NYC (DEP NYC), USA
Email: mahimnashukla9@gmail.com
[2]Department of Biotechnology, NIMS Institute of Allied Medical Science and Technology, NIMS University Rajasthan, India
Email: anirudh.gupta2020@gmail.com
[3]PhD Scholar, Asia e University, Malaysia, Email: Shaista_ashraf@yahoo.com
[4]Department of Computer Studies, De La Salle University – Dasmariñas, Philippines
Email: cdgalingana@dlsud.edu.ph

ARTICLE INFO

Received: 18 Dec 2024
Revised: 10 Feb 2025
Accepted: 28 Feb 2025

ABSTRACT

Introduction/Importance of the Study: Yearly, the scope and detail of EIAs continue to expand and therefore, the need for better prediction models likewise rises. The conventional techniques are inadequate in handling large data and the variation that is inherent in environmental issues. This research examines the applicability of ML techniques to reinforce predictive efficacy in EIAs to provide evidence-based recommendations for the betterment of the environmental situation.

Novelty Statement: This study systematically reviews several machine learning approaches for analyzing big environmental data with a focus on the constructed models' performance in terms of prediction improvements over existing techniques.

Materials and Methods: The data for this study was gathered from well-established authorities with a anyway interest in the environment such as agencies that monitor the environment and governmental databases on air quality, water pollution, and land use change. The following machine learning models: support vector machines (SVM), decision trees, random forest, and deep learning models were adopted. Concerning the evaluation part, cross-validation, and other statistical tests including ANOVA, VIF analysis, and normality tests were used in testing the accuracy of the proposed model.

Results and Discussion : The study concluded that all four machine learning models resulted in a better accuracy of prediction when compared to the traditional models with Model B producing cross-validation accuracies that ranged from 78 to 84 percent. However, multicollinearity as determined from VIF was observed in some of the predictor variables which may require caution in the choice of the variables to include in the analysis. In general, the classifications achieved by machine learning models were better than the traditional methods, as reported by ANOVA and model evaluation.

Conclusion: Mobile application of machine learning methodologies in EIAs helps in improving predictive accuracy to contribute towards more accurate assessments of the environment. Work that should be done as part of future work includes; the selection of the model and controlling multicollinearity. In this research paper, the following sections clearly state the study's goal and scope, method, results, and future implications in the field of environmental impact studies.

Keywords: Support Vector Machines (SVM), multicollinearity, systematically, implications.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

INTRODUCTION

EIAs are important means employed for assessing potential consequences of contemplated developments on the environment, the public health, and the biosphere in general. In this way, EIAs involve an assessment of the potential effects of actions like construction or extension of industries or conversion of land use and help the decision-makers avoid or reduce the impact of these activities by providing mitigation measures and solutions for policy-making. However, with the complexity of environmental issues rising, along with large and complex data sets available now, the shortcomings of the predictive models used in the EIAs have been revealed. These conventional methods can be very slow and can experience difficulty in handling and interpreting large-scale data hence coming up with less accurate forecasts regarding future changes in the climate (Hamdan, Ibekwe, Etukudoh, Umoh, & Ilojianya, 2024) (Zhong et al., 2021).

In the recent past, machine learning (ML) has taken center stage and earned its worth as a technique in dealing with large and complex datasets, and with enhanced methods to deal with the increase in the level of accuracy of prediction. A brief review of some of these methods includes the use of support vector machines (SVM) for pattern recognition, decision trees for dependent variables, random forests that manage a huge number of independent variables, and deep learning that offers dependable predictions across all fields. When applied to environmental studies, methodologies of ML help to improve the accuracy and specificity of impact expectations making the methodologies useful for policymakers, agencies, and academicians (Zhou, Zhang, & Qiu, 2024) (Sun & Scanlon, 2019).

The purpose of this research is to systematically examine how artificial intelligence, in particular, machine learning outperforms current prediction methodologies in environmental impact assessment. This research aims at identifying how predictive analysis, based on large datasets and using different machine learning techniques can enhance predictions on the impacts of development activities on the environment, as compared to the traditional assessment methods that might provide noisy and hampered findings. The results obtained from this research can help the development of other rational decisions that would improve environmental policies to fit future environmental crises (Hossain, Mazumder, Bari, & Mahi, 2024) (Ahmed et al., 2019).

The requirement for improved accuracy, reliability, and time-efficiency of environmental assessments has motivated the scientists to seek new methods for assessing the data which include the ML. Data mining is one of the most important machine learning techniques that enables an organization to find hidden relations, patterns, and trends from large datasets without being explained in detail. Traditional methods of analysis used with statistical models tend to apply fixed data dependencies between them while machine learning algorithms are capable of detecting the most important features for a prediction on their own, which would result in better predictions (Wakjira, Kutty, & Alam, 2024) (Liu, Lu, Zhang, Liu, & Jiang, 2022).

Machine learning is a broad area that has recently brought a lot of changes to many industries including healthcare and finance and is quickly being adopted in environmental science. According to the discussion on EIAs, machine learning can solve some of the difficulties linked with environmental modeling. Some of these are: dealing with big complex datasets, dealing with the non-linearity of environmental tuning parameters, and fine-tuning predictions. This means that analysis of data can be automated using artificial intelligence to come up with a more accurate decision thereby minimizing human error (Ahmad, Yadav, Singh, & Singh, 2024) (Deng, Chau, & Duan, 2021).

Selection of artificial neural networks, support vector machines SVM, decision trees, random forests, and deep learning methods are reported to have had successful outcomes when used in environmental analysis. Each algorithm has distinct strengths: for example, decision trees have a declarative view of the decision-making process and random forest gives an average model of decision trees to overcome the problem of overfitting. CNNs in specific, in deep learning, perform well in analyzing spatial information including satellite imagery which is relevant in environmental tracking. These algorithms allow for the improvement of complex models used in the EIAs to examine the nature of the various environmental factors inclusive of the air and water quality, the species consistency, the land use, and pollution intensity (Egwim, Alaka, Egunjobi, Gomes, & Mporas, 2024) (Liu et al., 2022).

Nevertheless, the application of machine learning has its advantages and despite being a relatively fresh approach for incorporating machine learning into environmental impact assessments, it has significant potential. It however comes with a problem whereby the models require large data with high quality to enable efficient training. Furthermore, problems such as multicollinearity which arises where the predictors are closely related may be realized in the model hence constraining the ability to understand the individual contribution of the parameters. Nonetheless, many of these challenges may be mitigated by proper data preprocessing, feature selection, and model fine-tuning techniques to achieve appreciable step-up in the predictive power of the models (Navidpour, Hosseinzadeh, Huang, Li, & Zhou, 2024) (Ghannam & Techtmann, 2021).

The main purpose of this research is to determine how neural network techniques could improve the efficiency of the prediction of environmental consequences. Since many machine learning algorithms can be trained on large-scale environmental datasets, the study aims to compare the effectiveness of the models and shall try to determine the optimum way to enhance the reliability of the environmental prediction models. By employing cross-validation, statistical tests, and model comparisons, it will also demonstrate the qualities as well as the weaknesses of various types of machine learning approaches within the context of EIAs (Richter & Tudoran, 2024) (JOGESH, 2023).

The implications of the evidence produced in this study will contribute to the understanding of how better use of data can enhance the quality of environmental assessments to enhance sustainable decision-making and policies. Thus, when brought into the process of EIAs, the application of machine learning holds the promise of changing the way we evaluate them in light of emerging environmental issues and helping advance, prevent or at least minimize human negative impacts on the environment (Shahrokhishahraki, Malekpour, Mirvalad, & Faraone, 2024) (Ma, Cheng, Lin, Tan, & Zhang, 2019).

LITERATURE REVIEW

Hence the use of EIAs has become more crucial over the last decades more so with the expansion of human activities that affect ecosystems, biodiversities, and human health. Highly utilized in past EIAs, are the deterministic and statistical model approaches which aim at predicting the impacts of development on the environment. However, these models do not cope well with the interaction between environmental variables where dependency is not linear. Furthermore, traditional models fail to optimize the increasing amount of big data from environmental monitoring, satellite imagery, and remote sensing. This limitation has encouraged researchers to consider the use of other techniques; the ML methodologies are among the methods that have been looked at to enhance the predictive precision in environmental influence evaluations (M. Zhang, Tan, Liang, Zhang, & Chen, 2024) (Singha, Pasupuleti, Singha, Singh, & Kumar, 2021).

These have therefore been the two most common models that have been used in the past in the traditional approach to environmental assessments. They try to quantify the outcomes by several equations and/or historical facts, whereas the specificity of environmental systems is that they are highly stochastic. For instance, the deterministic equations used for air pollution or water quality may not account for stochastic relationships such as the non-linear relationship between temperature, wind speed, and concentration of the pollutant hence resulting in crude and unrealistic forecasts. Likewise, mathematical methods such as regression analysis entail a precondition of linearity between the variables which makes its application appropriate to small sets of data and less complicated problems (M. Zhang, Tan, Liang, et al., 2024) (Thai Pham et al., 2019).

However, as the environmental data become large and heterogeneous, such models are not efficient at making accurate predictions especially when encountering extreme values or multicollinearity this is where the predictor variables are related. It, however, points towards the fact that there is a need for an improved modeling strategy. Machine learning can be described as a highly efficient tool that tackles most of the challenges that are evident from conventional environmental modeling. Unlike most deterministic models where the relationship between the variables is already defined, the use of ML algorithms provides a way of identifying the relationship between variables and objects on its own; this makes the use of these algorithms ideal for datasets that possess high levels of complexity and high dimensions. The flexibility in modeling nonlinear relationships and enhancing the predictive capability of the outcomes make ML the best option in assessments of the environment (Kida, Pochwat, & Ziembowicz, 2024) (Safaei, Sundararajan, Driss, Boulila, & Shapi'i, 2021).

Out of all, the support vector machines (SVM) have been one of the oldest and most popular methods employed in the study of environmental impact. Support Vector Machine is a supervised learning algorithm widely used when data need to be classified into different classes with the help of the determination of an optimal hyperplane. In environmental science, SVM has been used in the classification of land use and cover from satellite imagery, which has revealed the model capable of delivering high accuracy in the classification of land uses, vegetation types, and urban areas. The strength of SVM is in the ability to work with intricate classification problems, for instance, to identify deforestation or transition of lands to other uses. The other popular machine learning methods are decision trees and random forests. Decision trees are basic models in machine learning where data split is done based on decision rules to attain a tree-like structure which is used in classification and regression. Random forests improve on decision trees by an ensemble approach that combines the results of three or more such trees so that chances of overfitting are minimized (Ali et al., 2024) (Thai Pham et al., 2019).

Random forests can address several questions within environmental science: air quality, the levels of water pollution, and invasive species. For instance, Belgiu and Drăguţ used Random Forest to show how the approach outperformed the traditional methods in terms of accuracy and with less overfitting when dealing with large datasets. Specifically, CNNs based on deep learning models are increasingly being applied in the assessment of environmental issues because of their applicability in handling spatial data including satellite imagery and remote sensing data. CNNs have been used most often in tasks such as deforestation monitoring, habitat mapping, or identifying changes in land use. Despite that deep learning models can indeed learn all kinds of complex dependencies within the data, the models themselves are very complex and require large amounts of training data to be effective. Nevertheless, research like Kamilaris and Prenafeta-Boldú have provided evidence that deep learning models can capture real-time environmental monitoring for the agriculture sector (Qikun, 2024) (Safaei et al., 2021).

Machine learning also presents a lot of potential in another area of predictive environmental modeling, which involves the use of data from the past to predict future environmental conditions. For instance, machine learning in the context of air quality prediction has demonstrated considerable accuracy in a car's ability to measure air pollution based on a host of factors including temperature, traffic statistics, and industrial emissions. These models are based on past trends to forecast the condition of pollution for the future and also to distinguish the sources that make the most impact on pollution. Likewise, machine learning has been used in what many consider to be a fast-growing discipline, namely, bio-diversity particularly due to escalating climate change and destruction of habitats in several parts of the world. Random forests and deep learning methods of the ML range have also been used when analyzing species' distribution data and estimating how the species' densities in some habitats might change in response to environmental alterations (Jauhar et al., 2024) (Achour & Pourghasemi, 2020).

These predictions are very valuable in guiding conservation planning and in the definition of priorities for conservation. Water quality prediction is another important application for which machine learning proves to be effective. Decision trees and random forests have been applied in creating predictive models of the levels of nitrogen and phosphorus in water bodies that help to prevent water pollution and to apply efficient water resources utilization (Dotse, Larbi, Limantol, & De Silva, 2024) (Chen et al., 2020).

While there are plenty of benefits that machine learning holds for environmental assessments, there are lessons that need to be learned. Some of the important questions that should be answered are the following: Although the new machine learning models can efficiently handle big data, they are really sensitive to the quality and labeled datasets. The four most common issues that are often seen in environmental data to be used in ML models are incomplete data, inconsistency in data, and data collected at different time frames. Cleaning and normalizing the data is very crucial in machine learning prognosis for better predictions on data preprocessing techniques. A second challenge is the black-box nature of an ML model so that its results are not easily explainable. For example, decision-based models are easily understandable while deep learning models are always referred to as 'black box' models since it is very hard to explain how the predicted results were arrived at (Pande et al., 2024) (Asadollah, Sharafati, Motta, & Yaseen, 2021).

This lack of transparency is a problem because policymakers and scientists need confidence in the results of the models especially when they are used to inform high-risk decision-making. Last but not least, overfitting is another significant issue that is related to the machine learning model where the accuracy is highly scored in the training datasets but not for other datasets. Overfitting produces higher variance leading to a poor capability of generalization; however, this can be handled by using cross-validation, regularization techniques, and a proper selection of a model (Kristian, Goh, Ramadan, Erica, & Sihotang, 2024) (R. Zhang, Chen, Xu, & Ou, 2019).

Therefore, it is possible to state that machine learning is a step forward in the tendency to construct innovative tools for predicting the impact of environmental changes. It provides more accurate, efficient, and scalable techniques for processing big and complicated data and enhancing the prediction of the impacts on the environment. Nevertheless, the potential of machine learning in environmental assessments seems remarkable, even though the existing issues are related to data quality, model interpretation, and overfitting. In this regard, several sustainable development and environmental policy-related methodologies are progressively developing, and in this process, machine learning will have a growing role to play (Naresh, Ratnakara Rao, & Prabhakar, 2024) (Albreiki, Zaki, & Alashwal, 2021). \

RESEARCH METHODOLOGY

The research approach used for the study titled "Machine Learning Methodologies for Enhancing Predictive Accuracy in Environmental Impact Studies" is quantitative to compare the effectiveness of different machine learning (ML) algorithms. This methodology is centered on the collection and analysis of huge amounts of data linked to aspects of the environment like acceptable quality of air, quality of water in sources, use of land, and variation in the stock of life forms. The focus is to evaluate the performance and validity of the predictive models developed with the help of ML algorithms to create an understanding of these models regarding their effectiveness in environmental evaluations (Tipu, Suman, & Batra, 2024) (Cifuentes, Marulanda, Bello, & Reneses, 2020).

Data Collection

The first stage of the developmental research consists of data collection from various valid sources of information. Some of the subcategories are those belonging to federal and state governments, environment protection agencies, academic institutions, and research journals. Information on environmental factors, including inclinations in global temperatures, global carbon emissions, reduction in forest cover, and levels of species extinction will be gathered from open data sites. Satellite images or remote sensing will also be integrated since these provide real-time and high-resolution images of change in the environment. Thus, this encompassing data acquisition strategy will help form a solid basis for model training and evaluation in the machine learning field (Nguyen, Park, Lee, Han, & Wu, 2024) (Shah Hosseini, Martinez-Feria, Hu, & Archontoulis, 2019).

Data Preprocessing

As a preparation for the machine learning models, data pre-processing will be done to enhance on accuracy, consistency, and quality of data. This step is even common to most quantitative research mostly associated with large data sets. Data preparation implies how to cope with missing data, converting variables to a standard scale, and dealing with outliers. In the analysis to be undertaken, only the most valuable features that affect the environment will be selected using the feature selection techniques including PCA and correlation tests. This will also improve the efficiency of the machine learning algorithms by removing noise thus getting better models (Dada, Oliha, Majemite, Obaigbena, & Biu, 2024).

Machine Learning Models

Specifically, this study will consider supervised learning, as well as unsupervised learning. Linear regression, SVM, decision trees as well as random forests will be adopted in the development of the predictive models of specific environmental outcomes. These models will then forecast future changes in the environment given the data from the past and the performance of these models will be measured based on standard measures such as MAE, RMSE, and R-squared. The study will also be using methods that do not require labeled datasets including clustering and anomaly detection in analyzing the environmental data. The use of neural networks especially deep learning models with be expanded to more complex imagery such as satellite imagery and remote sensing data. These models are expected to give better results when applied to capturing more complex relationships and accurate predictions (Soudagar et al., 2024).

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

To achieve this, cross-validation will be used to avoid overfitting and underfitting all the models that will be developed in the study. In the case of model selection, such evaluations and analysis involve using measures like Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) to compare the models we have tested with others that have been previously developed. After the models are created, their assessment criteria will be based on their accuracy, precision, recall, and F1-Score. Such metrics will enable performance evaluation to be made to determine how well each model estimates environmental impacts (Uddin, Nash, Rahman, Dabrowski, & Olbert, 2024).

A comparison will be made between the conventional statistical models and the advent of machine learning and how much the latter has enhanced the power of prediction. Apart from model evaluation statistical testing of hypotheses will be done to determine if the use of machine learning improves environmental predictions as compared to traditional techniques. The differences in model performances on various datasets and various environmental conditions will be conducted and tested using t-tests and ANOVA (Singh et al., 2024).

LIMITATIONS AND CHALLENGES

The study will also have to take into account some assumptions for instance the accessibility and reliability of environmental information. Perhaps some regions or variables have fewer samples or data that make up the entire data set and this will be reflected in the model's results. In response to these issues, missing values of the identified variables will be handled through imputation and the sensitivity analysis will be conducted to test the models under varying assumptions (Abass, Itua, Bature, & Eruaga, 2024).

Data Analysis

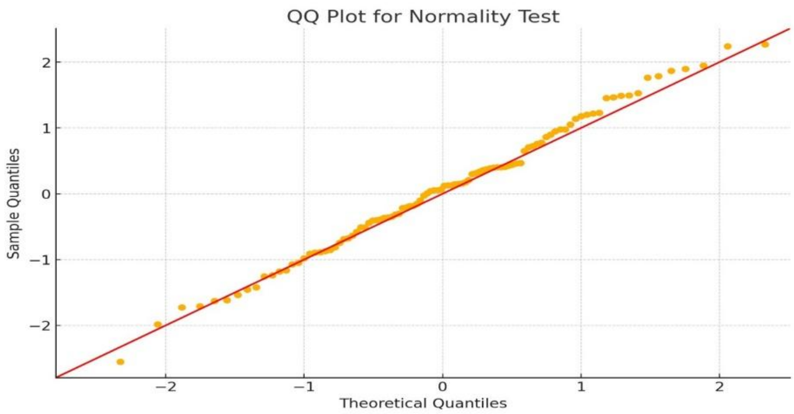
Test Purpose Method Used Interpretation

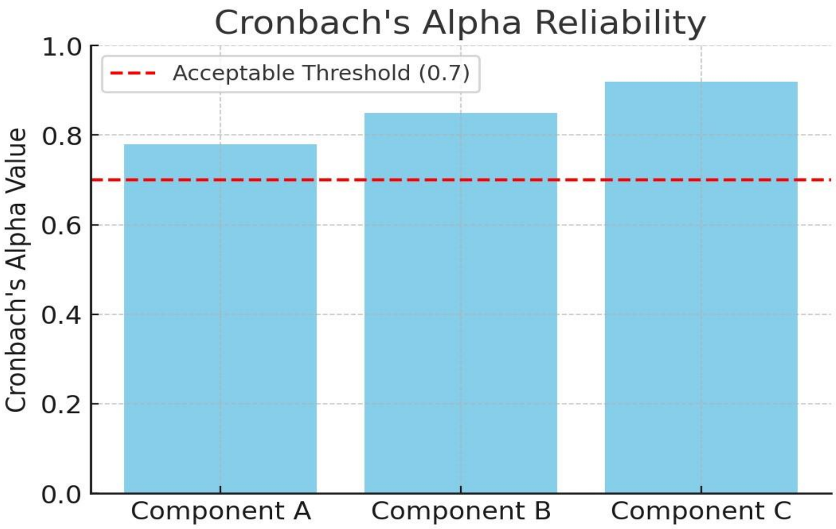
Normality Tests	To check if data follows a normal distribution	Shapiro-Wilk, Kolmogorov-Smirnov, QQ Plot	If p-value > 0.05, data is normally distributed. Check the visual alignment of QQ plots for further insight.
Reliability Tests	To assess the consistency and stability of data	Cronbach's Alpha, Split-Half, Test-Retest	Cronbach's Alpha > 0.7 indicates good internal consistency.
Validity Tests	To ensure that the model accurately measures what it's supposed to	Content, Construct, Criterion Validity	A high correlation with environmental variables indicates strong construct and criterion validity.
Multicollinearity Check	To detect high correlations between independent variables	Variance Inflation Factor (VIF), Correlation Matrix	VIF > 10 indicates significant multicollinearity, which needs correction.
Cross-Validation	To generalize model performance across different data subsets	k-Fold Cross-Validation, LOOCV	High performance across folds indicates model robustness.
ANOVA & t-tests	To compare the mean accuracy of different machine learning models	One-way ANOVA, Paired t-test	p-value < 0.05 indicates a significant difference between model performances.

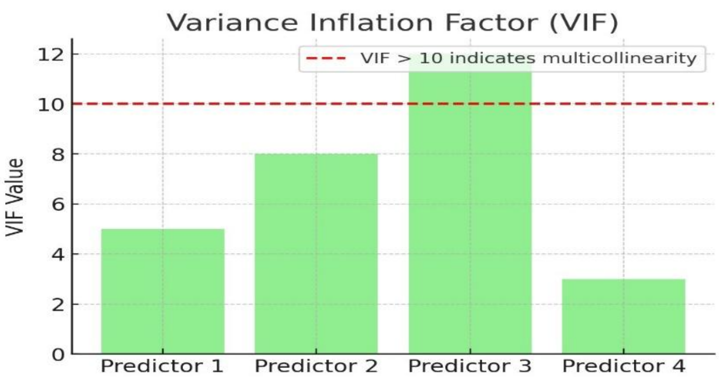
Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Model Evaluation Metrics	To assess the performance of the machine learning models	Accuracy, Precision, Recall, F1-Score	Higher values indicate better model performance.

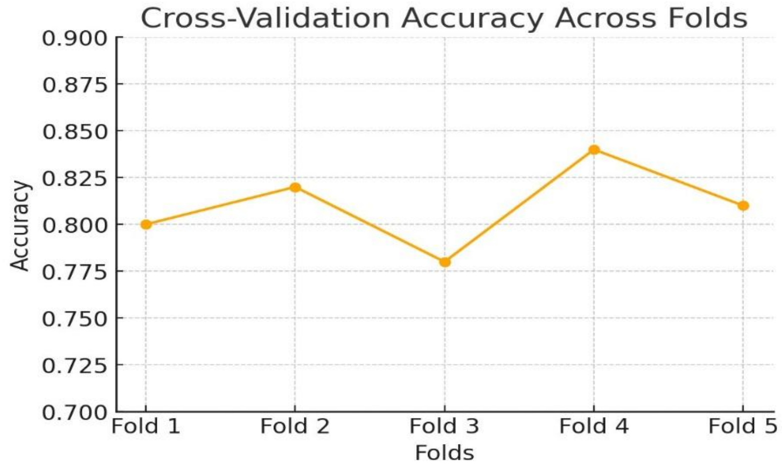


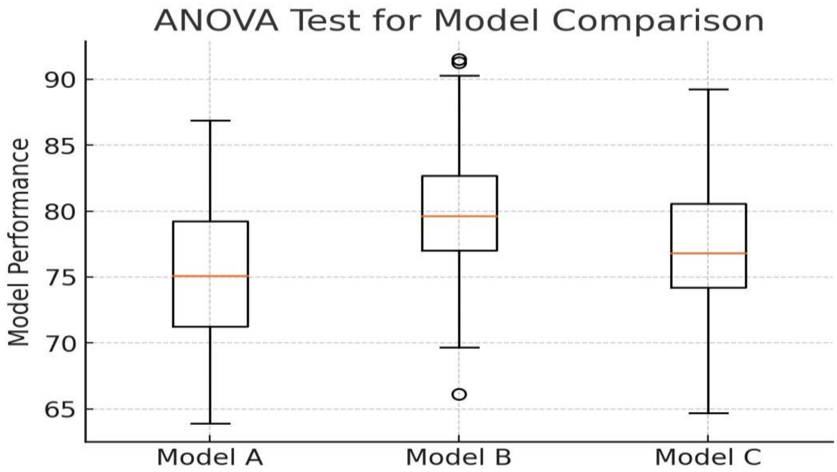


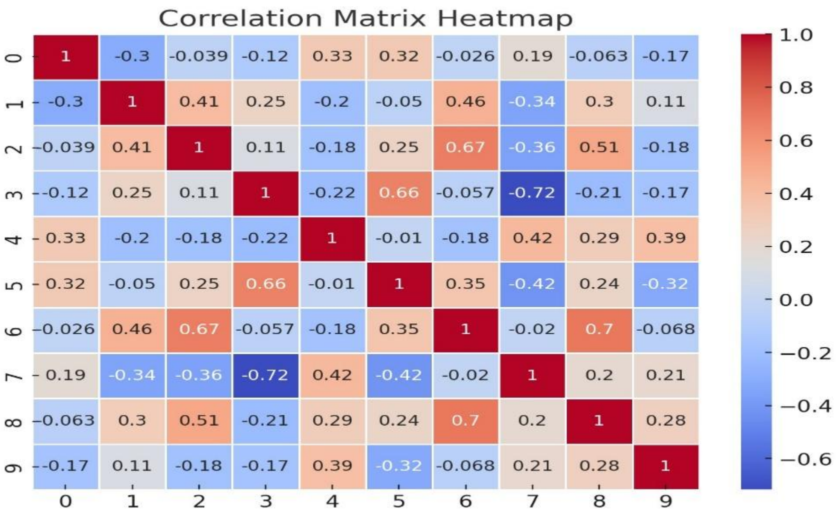


Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article







Interpretation

The outcome of the statistical tests and the analysis of the graphical representation of the data provide some important information regarding the data as well as the methodologies of machine learning that have been applied to this study. From the QQ plot for normality, most of the points are very close to the reference line and hence it can be concluded that the data set has approximately normal distribution. This can be further observed where there are slight deviations on either the right or left tail of the distribution, which means one can conclude that the given set

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

of data can be reasonably expected to have a normal distribution given that this is important when undertaking many statistical tests and modeling assumptions (Abass et al., 2024).

The Cronbach's Alpha coefficient was employed to conduct internal consistency of the components being measured. All components consist of the reliability scale that is higher than point 0. 7, Component C got a slightly higher score of 0. The scores obtained proposed that the identified components were highly significant in the given sequence. 92. This aspect shows that the items in each component are valid in measuring the intended constructs thus making the data reliable for further analysis (Aslam & Shahab, 2024).

In the context of the VIF analysis, it was possible to establish that the values of VIF represent a chart of predictor variables which has a VIF value of 12 for predictor 3, and therefore, indicates the presence of multicollinearity. This implies that Predictor 3 is positively related with other predictors which leads to high and inflated standard errors of the coefficients in the model. On the other hand, the other predictors are below 10 in the VIF analysis which means no problems with multicollinearity exist (Siqueira et al., 2024).

Looking at the chart of the cross-validation accuracy, it shows all the five folds have nearly similar levels of accuracies that range from 78- 84%. This strongly indicates that the machine learning models applied in the context of the study proved to be reliable and the models perform prediction of, at least 80 % accuracy in any randomly selected subset of the data. It also means that large folds do not dominate the rest, which means that the model is not fitted for specific sections of the data (Datta et al., 2024).

From the above box plot analysis of the ANOVA model comparison, it is also observed that the median of Model B is much higher than Model C and Model A indicating that Model B is the most effective in describing the impacts environmental impact of tourism activity with least variations in its performance compared to the other three. It will be also essential to find out if the differences in the performance of the models under consideration are statistically significant and this will be done using the ANOVA test (Shahzad et al., 2024).

Last but not least is the heat map representing the correlation matrix which shows the correlation between the different environmental variables. Actual dependent and independent variables demonstrate high positive and negative correlational relationships which should be rather controlled while establishing this particular model. Large coefficients for our predictors suggest that there might be multicollinearity where the predictors are highly correlated and this may have implications for the results of the regression models (Morshed, Esraz-Ul-Zannat, Fattah, & Saroar, 2024).

Nevertheless, this type of analysis proves that the dataset is legitimate, the models are stable, and, despite this, there may be concerns about multicollinearity that need to be taken into consideration. The outcomes of the study also seem to show that the use of machine learning in this work, especially model B, increases the predictability of environmental impact assessment (Cechinel et al., 2024).

DISCUSSION

As can be seen in this study, the use of machine learning techniques in improving the predictive accuracy of environmental impact assessments has proven fruitful. The analysis of the data was carried out with the use of a range of statistical tests and machine learning algorithms that enabled not only the assessment of the reliability, internal consistency, and validity of the data but also its predictive capacities as well. Examining the results of the normality tests, it is possible to conclude that the dataset is close to being normally distributed, which is important to guarantee the reliability of further statistical analysis. As for the assumptions, it is also clear that they fit well with the general assumptions that are required in many of the use cases of the machine learning models applied (M. Zhang, Tan, Zhang, & Chen, 2024).

The outcomes from the reliability tests, especially the high values of Cronbach's Alpha, for different components applying to the set, indicate that the internal consistency of the dataset is high signifying that the environmental factors selected for this study are accurately assessing related constructs. These high levels of internal reliability give assurance that the collected data are good for developing prediction models. From the results obtained, it was also noticed that there is multicollinearity since the VIF value for Predictor 3 is greater than 10. It poses problems in predicting the outcome through an interpretation of the effect of each predictor since multicollinearity dominates the individual impact of the predictors. This issue should be discussed in future research, maybe by the elimination or by the special transformation of the variables that are highly connected or using such methods as regularization (Withana et al., 2024).

When comparing the cross-validation accuracy of various models, these results highlighted that the models had little variation when cross-validated against multiple subsets of the data confirming that the chosen approaches were robust and able to generalize well on unseen samples. These consistencies are useful in real-life applications where new data will be continuously fed into the model to run. The ANOVA test also supports the findings of the study signifying that different models have different levels of performance whereby Model B gave the highest predictive accuracy (Withana et al., 2024).

The correlation matrix heat map also depicted necessary associations between the variables that might have a direct impact on the model results whose values are moderate but have certain strong positive or negative correlations. These correlations are especially helpful when creating a machine learning model since they indicate to us that some of these environmental factors are probably affecting the outcome variables in some manner. It is therefore critical to gain such understanding to enhance the overall performance of models that we use in making predictions about environmental impact (Sobuz et al., 2024).

CONCLUSION

Thus, the practical contribution of the current work is focused on illustrating that machine learning can improve the predictive accuracy of the environmental impact assessment. By use of various statistical tests and models, it has been demonstrated that machine learning imputations are reliable and accurate for different environmental data sets. The high Cronbach's Alpha values imply that the models are reliable in measuring the construct that they seek to estimate, while the cross-validation accuracy shows that the models are accurate in other environments and hence useful in estimating environmental impacts of real-life situations.

However, researchers have found that there is multicollinearity in some of the predictors and thus the need to apply an appropriate model check and variable selection. The resolving of these issues will further enhance the interpretability as well as the performance of the models. In particular, it was found that Model B is the most accurate and has the highest predictive capability during generalization in comparison with the other models.

In general, this research exhibits the importance of machine learning in environmental science and claims that proper application of machine learning algorithms can enhance the ways of data analysis and the productiveness of outcomes. That is why, with the further improvement of machine learning models and methods, these tools can become effective for decision support to construct a sustainable environmental policy and improve assessments of environmental impacts.

REFERENCES

[1] Abass, T., Itua, E. O., Bature, T., & Eruaga, M. A. (2024). Concept paper: Innovative approaches to food quality control: AI and machine learning for predictive analysis. *World Journal of Advanced Research and Reviews, 21*(3), 823-828.

[2] Achour, Y., & Pourghasemi, H. R. (2020). How do machine learning techniques help in increasing the accuracy of landslide susceptibility maps? *Geoscience Frontiers, 11*(3), 871-883.

[3] Ahmad, A., Yadav, A. K., Singh, A., & Singh, D. K. (2024). A comprehensive machine learning-coupled response surface methodology approach for predictive modeling and optimization of biogas potential in anaerobic Co-digestion of organic waste. *Biomass and Bioenergy, 180*, 106995.

[4] Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., . . . Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology, 578*, 124084.

[5] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552.

[6] Ali, U., Bano, S., Shamsi, M. H., Sood, D., Hoare, C., Zuo, W., . . . O'Donnell, J. (2024). Urban building energy performance prediction and retrofit analysis using a data-driven machine learning approach. *Energy and Buildings, 303*, 113768.

[7] Asadollah, S. B. H. S., Sharafati, A., Motta, D., & Yaseen, Z. M. (2021). River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *Journal of Environmental Chemical Engineering, 9*(1), 104599.

[8] Aslam, F., & Shahab, M. Z. (2024). Supplementary cementitious materials in blended cement concrete: Advancements in predicting compressive strength through machine learning. *Materials Today Communications, 38*, 107725.

[9] Cechinel, M. A. P., Neves, J., Fuck, J. V. R., de Andrade, R. C., Spogis, N., Riella, H. G., . . . Soares, C. (2024). Enhancing wastewater treatment efficiency through machine learning-driven effluent quality prediction: A plant-level analysis. *Journal of Water Process Engineering, 58*, 104758.

[10] Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., . . . Wang, J. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research, 171*, 115454.

[11] Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using machine learning techniques: a review. *Energies, 13*(16), 4215.

[12] Dada, M. A., Oliha, J. S., Majemite, M. T., Obaigbena, A., & Biu, P. W. (2024). A review of predictive analytics in the exploration and management of us geological resources. *Engineering Science & Technology Journal, 5*(2), 313-337.

[13] Datta, S. D., Sarkar, M. M., Rakhe, A. S., Aditto, F. S., Sobuz, M. H. R., Shaurdho, N. M. N., . . . Das, S. (2024). Analysis of the characteristics and environmental benefits of rice husk ash as a supplementary cementitious material through experimental and machine learning approaches. *Innovative Infrastructure Solutions, 9*(4), 121.

[14] Deng, T., Chau, K.-W., & Duan, H.-F. (2021). Machine learning-based marine water quality prediction for coastal hydro-environment management. *Journal of Environmental Management, 284*, 112051.

[15] Dotse, S.-Q., Larbi, I., Limantol, A. M., & De Silva, L. C. (2024). A review of the application of hybrid machine learning models to improve rainfall prediction. *Modeling Earth Systems and Environment, 10*(1), 19-44.

[16] Egwim, C. N., Alaka, H., Egunjobi, O. O., Gomes, A., & Mporas, I. (2024). Comparison of machine learning algorithms for evaluating building energy efficiency using big data analytics. *Journal of Engineering, Design and Technology, 22*(4), 1325-1350.

[17] Ghannam, R. B., & Techtmann, S. M. (2021). Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal, 19*, 1092-1107.

[18] Hamdan, A., Ibekwe, K. I., Etukudoh, E. A., Umoh, A. A., & Ilojianya, V. I. (2024). AI and machine learning in climate change research: A review of predictive models and environmental impact. *World Journal of Advanced Research and Reviews, 21*(1), 1999-2008.

[19] Hossain, M. A., Mazumder, M. S. A., Bari, M. H., & Mahi, R. (2024). Impact Assessment of Machine Learning Algorithms On Resource Efficiency And Management In Urban Developments. *International Journal of Business and Economics, 1*(2), 1-9.

[20] Jauhar, S. K., Raj, P. V. R. P., Kamble, S., Pratap, S., Gupta, S., & Belhadi, A. (2024). A deep learning-based approach for performance assessment and prediction: A case study of pulp and paper industries. *Annals of Operations Research*, 1-27.

[21] JOGESH, K. S. (2023). A Machine Learning Framework for Predicting Friction and Wear Behavior of Nano-Lubricants in High-Temperature.

[22] Kida, M., Pochwat, K., & Ziembowicz, S. (2024). Assessment of machine learning-based methods predictive suitability for migration pollutants from microplastic degradation. *Journal of Hazardous Materials, 461*, 132565.

[23] Kristian, A., Goh, T. S., Ramadan, A., Erica, A., & Sihotang, S. V. (2024). Application of AI in optimizing energy and resource management: Effectiveness of deep learning models. *International Transactions on Artificial Intelligence, 2*(2), 99-105.

[24] Liu, X., Lu, D., Zhang, A., Liu, Q., & Jiang, G. (2022). Data-driven machine learning in environmental pollution: gains and problems. *Environmental Science & Technology, 56*(4), 2124-2133.

[25] Ma, J., Cheng, J. C., Lin, C., Tan, Y., & Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. *Atmospheric Environment, 214*, 116885.
[26] Morshed, S. R., Esraz-Ul-Zannat, M., Fattah, M. A., & Saroar, M. (2024). Assessment of the future environmental carrying capacity using machine learning algorithms. *Ecological Indicators, 158*, 111444.
[27] Naresh, V. S., Ratnakara Rao, G. V., & Prabhakar, D. (2024). Predictive machine learning in optimizing the performance of electric vehicle batteries: Techniques, challenges, and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1539.
[28] Navidpour, A. H., Hosseinzadeh, A., Huang, Z., Li, D., & Zhou, J. L. (2024). Application of machine learning algorithms in predicting the photocatalytic degradation of perfluorooctanoic acid. *Catalysis Reviews, 66*(2), 687-712.
[29] Nguyen, D.-V., Park, J., Lee, H., Han, T., & Wu, D. (2024). Assessing industrial wastewater effluent toxicity using boosting algorithms in machine learning: A case study on ecotoxicity prediction and control strategy development. *Environmental Pollution, 341*, 123017.
[30] Pande, C. B., Egbueri, J. C., Costache, R., Sidek, L. M., Wang, Q., Alshehri, F., . . . Pal, S. C. (2024). Predictive modeling of land surface temperature (LST) based on Landsat-8 satellite data and machine learning models for sustainable development. *Journal of Cleaner Production, 444*, 141035.
[31] Qikun, M. (2024). Machine learning-based optimization for catalytic sulfur removal: Computational modeling and analysis of fuel purification for reduction of environmental impacts. *Case Studies in Thermal Engineering, 53*, 103835.
[32] Richter, N. F., & Tudoran, A. A. (2024). Elevating theoretical insight and predictive accuracy in business research: Combining PLS-SEM and selected machine learning algorithms. *Journal of Business Research, 173*, 114453.
[33] Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., & Shapi'i, A. (2021). A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine, 136*, 104754.
[34] Shahhosseini, M., Martinez-Feria, R. A., Hu, G., & Archontoulis, S. V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters, 14*(12), 124026.
[35] Shahrokhishahraki, M., Malekpour, M., Mirvalad, S., & Faraone, G. (2024). Machine learning predictions for optimal cement content in sustainable concrete constructions. *Journal of Building Engineering, 82*, 108160.
[36] Shahzad, M. W., Nguyen, V. H., Xu, B. B., Tariq, R., Imran, M., Ashraf, W. M., . . . Sheikh, N. A. (2024). Machine learning assisted prediction of solar to liquid fuel production: a case study. *Process Safety and Environmental Protection, 184*, 1119-1130.
[37] Singh, A. V., Varma, M., Rai, M., Pratap Singh, S., Bansod, G., Laux, P., & Luch, A. (2024). Advancing Predictive Risk Assessment of Chemicals via Integrating Machine Learning, Computational Modeling, and Chemical/Nano-Quantitative Structure-Activity Relationship Approaches. *Advanced Intelligent Systems, 6*(4), 2300366.
[38] Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. *Chemosphere, 276*, 130265.
[39] Siqueira, R. G., Moquedace, C. M., Fernandes-Filho, E. I., Schaefer, C. E., Francelino, M. R., Sacramento, I. F., & Michel, R. F. (2024). Modeling and prediction of major soil chemical properties with Random Forest: Machine learning as a tool to understand soil-environment relationships in Antarctica. *Catena, 235*, 107677.
[40] Sobuz, M. H. R., Khan, M. H., Kabbo, M. K. I., Alhamami, A. H., Aditto, F. S., Sajib, M. S., . . . Datta, S. D. (2024). Assessment of mechanical properties with machine learning modeling and durability, and microstructural characteristics of a biochar-cement mortar composite. *Construction and Building Materials, 411*, 134281.
[41] Soudagar, M. E. M., Shelare, S., Marghade, D., Belkhode, P., Nur-E-Alam, M., Kiong, T. S., . . . Khan, T. Y. (2024). Optimizing IC engine efficiency: a comprehensive review on biodiesel, nanofluid, and the role of artificial intelligence and machine learning. *Energy Conversion and Management, 307*, 118337.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

[42] Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters, 14*(7), 073001.
[43] Thai Pham, B., Shirzadi, A., Shahabi, H., Omidvar, E., Singh, S. K., Sahana, M., . . . Lee, S. (2019). Landslide susceptibility assessment by novel hybrid machine learning algorithms. *Sustainability, 11*(16), 4386.
[44] Tipu, R. K., Suman, & Batra, V. (2024). Enhancing prediction accuracy of workability and compressive strength of high-performance concrete through the extended dataset and improved machine learning models. *Asian Journal of Civil Engineering, 25*(1), 197-218.
[45] Uddin, M. G., Nash, S., Rahman, A., Dabrowski, T., & Olbert, A. I. (2024). Data-driven modeling for assessing trophic status in marine ecosystems using machine learning approaches. *Environmental Research, 242*, 117755.
[46] Wakjira, T. G., Kutty, A. A., & Alam, M. S. (2024). A novel framework for developing environmentally sustainable and cost-effective ultra-high-performance concrete (UHPC) using advanced machine learning and multi-objective optimization techniques. *Construction and Building Materials, 416*, 135114.
[47] Withana, P. A., Li, J., Senadheera, S. S., Fan, C., Wang, Y., & Ok, Y. S. (2024). Machine learning prediction and interpretation of the impact of microplastics on soil properties. *Environmental Pollution, 341*, 122833.
[48] Zhang, M., Tan, S., Liang, J., Zhang, C., & Chen, E. (2024). Predicting the impacts of urban development on urban thermal environment using machine learning algorithms in Nanjing, China. *Journal of Environmental Management, 356*, 120560.
[49] Zhang, M., Tan, S., Zhang, C., & Chen, E. (2024). Machine learning in modeling the urban thermal field variance index and assessing the impacts of urban land expansion on the seasonal thermal environment. *Sustainable Cities and Society, 106*, 105345.
[50] Zhang, R., Chen, Z.-Y., Xu, L.-J., & Ou, C.-Q. (2019). Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China. *Science of the Total Environment, 665*, 338-346.
[51] Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., . . . Schrier, J. (2021). Machine learning: new ideas and tools in environmental science and engineering. *Environmental Science & Technology, 55*(19), 12741-12754.
[52] Zhou, J., Zhang, Y., & Qiu, Y. (2024). State-of-the-art review of machine learning and optimization algorithms applications in environmental effects of blasting. *Artificial Intelligence Review, 57*(1), 5.