JURNAL EMACS
e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.5 No.3 September 2023: 169-174 DOI: 10.21512/emacsjournal.v5i3.10602

Predictive Modeling of Jakarta's Social Cohesion: GBDT Leads Comparative Analysis

Muhammad Rizki Nur Majiid[1], Karli Eka Setiawan[2*]
[1,2] Computer Science of Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
muhammad.majiid@binus.ac.id; karli.setiawan@binus.ac.id

*Correspondence: karli.setiawan@binus.ac.id

Abstract *– In this study, we address the challenge of predicting the Social Cohesion Index in Jakarta through a comprehensive analysis of machine learning models. Finding the most accurate and effective predictive model for this crucial urban evaluation task is the primary goal of our research. We use a variety of machine learning algorithms, comparing their performance using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and computational cost. These algorithms include Gradient Boosted Decision Trees (GBDT), Polynomial Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). It should be noted that GBDT stands out as a top performer, regularly displaying outstanding accuracy with a competitive MAE of 0.692, RMSE of 0.887, and MAPE of 25.59%. The computational efficiency of GBDT is also impressive, with predictions taking only 0.05 seconds. These results underscore the potential of GBDT as a practical and precise tool for real-time assessments of social cohesion in large urban environments like Jakarta. The findings offer a data-driven way to guide policy decisions and community development activities, with important implications for urban planning and governance. Overall, this research emphasizes the promise of GBDT in boosting social cohesion evaluation approaches and increases our understanding of the application of machine learning in addressing complex urban difficulties.*

Keywords: *Social Cohesion; Machine Learning; Urban Assessment; GBDT; Predictive Accuracy*

I. INTRODUCTION

Social cohesion is crucial in determining any urban community's harmony, resilience, and development (Jewett et al., 2021) (Steiner et al., 2018). Its evaluation provides crucial insights into individuals' collective health and interdependence within a city's diverse fabric. In megacities like Jakarta, where population density is surging, it is challenging to collect comprehensive data on social cohesion using conventional survey techniques (Rybak, 2023). The overwhelming number of individuals, geographical complexities, and logistical constraints have necessitated assessing social cohesion.

This study employs a data-driven methodology to forecast and examine the Social Cohesion Index of Jakarta's neighborhoods. Using machine learning capabilities, we intend to revolutionize the method by which we gain insight into the complex social dynamics of this enormous metropolis. Despite their value, traditional survey techniques frequently need help to accommodate the vast size of Jakarta's population (Andariesta & Wasesa, 2022) (Viljanen et al., 2022) (Sarker, 2021). Consequently, our research arises as a timely response, offering an alternative method for data acquisition that is both scalable and efficient..

We predicted the Social Cohesion Index using a variety of well-known machine learning algorithms and a large dataset containing numerous socioeconomic and demographic characteristics. This study compared the effectiveness of Polynomial Regression (Narayan & Daniel, 2022), Decision Tree (Pekel, 2020), Random Forest (Tzenios, 2020), Support Vector Machine (SVM) (Parbat & Chakraborty, 2020), and Multi-Layer Perceptron (MLP) models (Ouma et al., 2020). Notably, within the

scope of our investigation, the Gradient Boosted Decision Tree (GBDT) model surfaces prominently, demonstrating remarkable predictive abilities (Zhang & Jung, 2021). This accomplishment has the potential to advance the field of predictive modeling considerably and highlights the expanding role of machine learning techniques in addressing significant urban challenges comprehensively.

This research serves as a beacon of innovation for urban planners, policymakers, and public data management entities grappling with the complexity of accurately documenting social cohesion in Jakarta. Our findings could facilitate more enlightened decision-making processes and targeted interventions that promote social harmony. By investigating the relationship between data science and social dynamics, this study represents a proactive step toward developing comprehensive and effective urban assessment methods.

II. METHODS

This study aimed to develop a machine learning model for predicting Jakarta's Social Cohesion Index (SCI), which assesses the degree of social integration and solidarity among city residents. Figure 1 depicts the six primary stages involved in developing and evaluating our model. First, this study conducted a literature review to comprehend the concept and dimensions of SCI and identify the variables and indicators that can be used to measure it. Secondly, we gathered data from a publicly available dataset paper to ensure validity. Based on the data analysis and the literature review, we then planned the design and architecture of our model. Our primary machine learning technique is a polynomial regression model with optimal degree selection because it can capture the nonlinear and complex relationship between the SCI and its predictors. Fourth, we constructed and trained our model with the collected data using Python and Sci-Kit-Learn libraries. Fifth, we utilized K-Fold Cross-Validation to assess the performance and precision of our model. Several metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Average Percentage Error (MAPE), were utilized to evaluate the model's fit. In conclusion, this research compared our model to existing models that employ various machine learning techniques, such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Multi-Layer Perceptron (MLP).

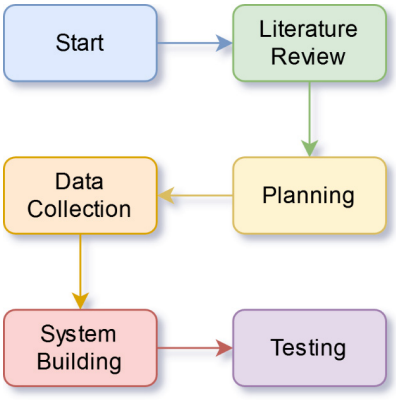

Figure 1. Machine Learning Model Research and Development Plan

2.1 Jakarta Social Cohesion Dataset

The dataset by (Amir et al., 2023) is an extensive compilation of factors associated with social cohesion in Jakarta's urban life. The dataset comprises information regarding 2,052 respondents from 44 Jakarta districts. Various questions were posed to the respondents regarding their social interactions, participation in community activities, and perceptions of trust and reciprocity. The dataset also contains information on the demographic characteristics of the respondents, such as their age, gender, level of education, and income.

The dataset was collected using a technique of stratified random sampling. The respondents were selected from different districts in Jakarta in proportion to the population of each district. The data compilation occurred between January and February of 2022. The dataset includes eight cohesion factor variables, including gender, age, level of education, income, two district variables, religion, and residential location. There are three variables for each: Trust, Recognition, Participation, Reciprocity, and Insertion among the cohesion variables. All data types are categorical or ordinal, and no complex data cleansing is required. The model will be trained to predict all 15 variables of cohesion.

2.2 K-Fold Cross-Validation

We have used the K-Fold Cross-Validation method in all of our experimental settings to reduce the potential for variance bias caused by randomly splitting the test set. This bias may result in inaccurate results. Specifically, we have divided our dataset into five folds, resulting in a test data ratio equaling twenty percent of the overall dataset for each split. The purpose of this method is to systematically evaluate the performance of our predictive models across several different data folds. The test data error values were recorded throughout each iteration of the K-Fold Cross-Validation procedure. After all folds have been completed, these values are then averaged in order to produce an overall evaluation of the model's performance for the scenario that was provided.

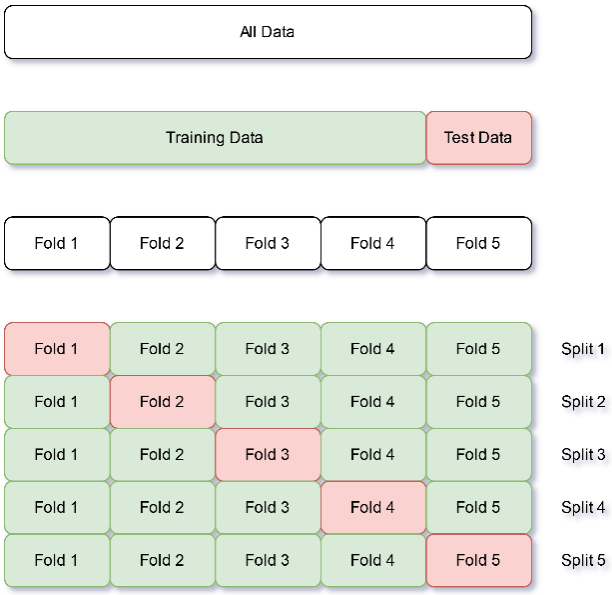

Figure 2. K-Fold Cross-Validation Principle

In this study, the application of K-Fold Cross-Validation, depicted in Figure 2, is one of the most critical factors contributing to improving our experimental findings' reliability and validity. This method ensures that our models are carefully tested across various subsets of the data, delivering a more comprehensive evaluation of their ability to predict future outcomes as a result. We can evaluate the model's generalization performance more accurately by averaging the test error values from numerous folds. This method also lessens the impact of the data's variability on our findings and increases the overall reliability of our conclusions.

2.3 Gradient Boosted Decision Tree

Gradient boosting develops sequentially weaker (simpler) prediction models, each attempting to predict the error left over by the previous model. Weak learners who perform slightly better than random chance is used in boosting. Gradient Boosting focuses on adding these weak learners one at a time and eliminating the observations a learner gets right at every step. The focus is on teaching new, weaker learners how to handle the remaining difficult observations at each step.

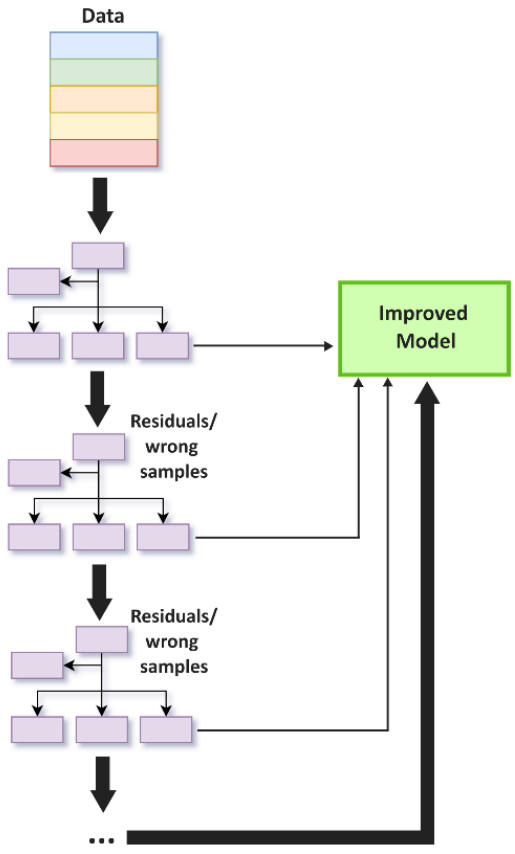

Figure 3. Gradient Boosted Decision Tree (GBDT)

As shown by Figure 3, GBDT combines several weak learners into a single strong learner. The weak learners, in this case, are the individual decision trees. Each tree attempts to reduce the error of the one that comes before it, and all the trees are connected in series. Boosting algorithms are typically slow to learn but also highly accurate because of this sequential connection. Slower learning models outperform faster learning ones in statistical learning. The weak learners are adjusted so that each new learner fits into the leftovers from the step before. As the model gets better over iteration, the weak learners fit better. Each step's results are combined in the final model to produce a strong learner. The residuals are found using a loss function. For example, logarithmic loss (log loss) can be used for classification tasks and mean squared error (MSE) can be used for regression tasks. It is important to note that adding a new tree does not impact any existing trees in the model. The additional decision tree fits the current model's residuals.

2.4 Experimental Setup

This comprehensive experiment uses several different machine-learning models, which is a diversity that inevitably creates issues when attempting to develop a uniform parameter situation relevant to all models. To address this level of complexity, a more subtle technique has been employed, wherein each model has been painstakingly fine-tuned over a specific number of trials. By making this tactical change, we are allowed to maximize the parameters of each model and, as a result, make full use of the inherent benefits and distinguishing qualities of these models. We ensure a solid foundation for our comparative research by adapting the tuning process to the specific requirements of each model. This decision enables us to identify the intricacies of the performance of each model.

The replies to eight survey questions regarding respondents' places of residence make up the primary dataset used for all machine learning models. These characteristics, which function as independent variables, are essential components of the predictive models we use. The answers to multiple-choice questions go through an essential step called one-hot encoding so machines may process them more easily. This transformation guarantees that the data are compatible with the algorithms used in machine learning and maintains the critical nuances embedded in the responses. In the meantime, our research depends on several aspects of social cohesion, represented by the dependent variables. Each of these dimensions consists of two parameters, and the machine learning models' overall goal is to forecast both characteristics to contribute to a comprehensive knowledge of the dynamics of social cohesiveness.

In the framework of our experiment, the evaluation set comprises around 410 rows (or 20% of the overall dataset), whereas the training set comprises the remaining 1642 rows. This segmentation ensures a thorough examination of the predicted performance of each model on data that has yet to be observed. Performance metrics, including the Mean Average Error (MAE), Root Mean Squared Error (RMSE), Mean Average Percentage Error (MAPE), and model processing duration in seconds, are systematically recorded to assess each model's efficacy comprehensively. In addition, it is of the utmost importance to point out that all tests are carried out on the Google Colab platform without employing GPUs. This strategic decision ensures the validity and comparability of our metric measurements while emphasizing the applicability and accessibility of our research findings for a broad audience of users and scholars.

III. RESULTS AND DISCUSSIO

Examining the performance between the Gradient Boosted Decision Trees (GBDT) model and the Decision Tree (DT) model, which is more generic, was the first stage in designing our experiment. The necessity of laying a basic standard for our predictive modeling system led to the conscious decision to use this benchmark. When we first evaluate how well a traditional DT performs, we can get essential insights into the intrinsic complexity of the social cohesiveness prediction problem. Using this comparison to establish a baseline, we can evaluate the additional value that the GBDT, in its capacity as an ensemble learning approach, brings to the table. In addition, it offers a helpful point of reference for assessing the efficiency of more sophisticated machine learning models that were subsequently included in the scope of our investigation. This method of strategically comparing models enables us to determine that GBDT is superior and can improve prediction accuracy considerably. As a result, verify its function as the proposed model for forecasting the Social Cohesion Index in Jakarta.

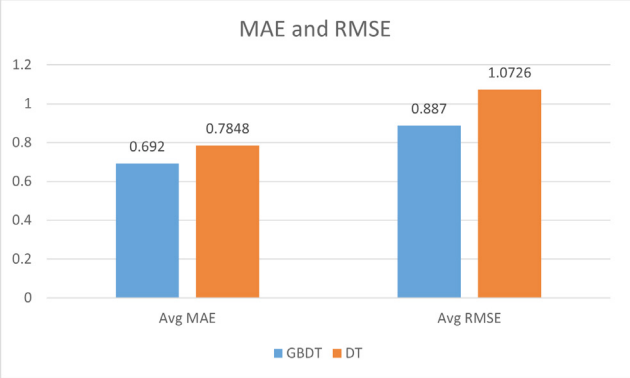

Figure 4. GBDT and Decision Tree (DT) Error Comparison

In Figure 4, we present a comparative analysis of the performance of Gradient Boosted Decision Trees (GBDT) and traditional Decision Trees (DT) based on their average Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) across cross-validation folds. The findings show that the GBDT model has a significant advantage over its DT counterpart. Remarkably, GBDT outperforms DT in terms of accuracy in predicting social cohesion scores, with an average MAE of 0.692 compared to 0.7848 for DT. This result shows that GBDT captures more intricate correlations and patterns in the data, leading to more accurate predictions. Similarly, GBDT surpasses DT in terms of RMSE, with an average RMSE of 0.887 as opposed to DT's 1.0726, indicating that GBDT's predictions are more accurate and have less volatility. These results highlight the effectiveness of ensemble approaches, such as GBDT, in improving predictive performance, which is crucial in applications where precision is crucial, such as evaluating social cohesion in urban situations.

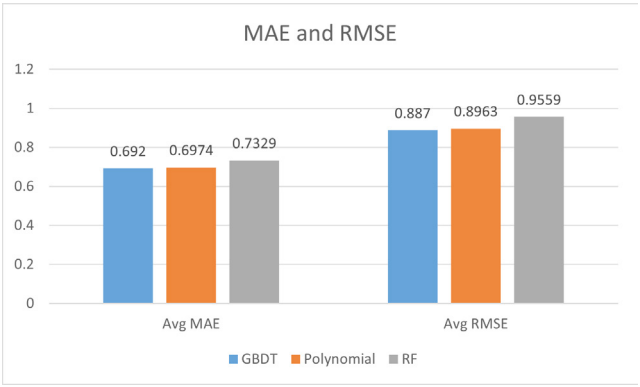

Figure 5. GBDT, Polynomial Regression, and Random Forest (RF) Error Comparison

Polynomial Regression and Random Forest (RF) are three additional machine learning models that we include in Figure 5 to further our comparative research. The graph shows the typical Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for these models across cross-validation folds. With an average MAE of 0.692 and RMSE of 0.887, GBDT remains the best-performing model. With an average MAE of 0.6974 and RMSE of 0.8963, Polynomial Regression comes in second place, demonstrating its proficiency in capturing complex relationships. Random Forest performs admirably while somewhat trailing with an average MAE of 0.7329 and RMSE of 0.9559. These outcomes demonstrate how GBDT consistently provides higher accuracy and precision, reiterating its effectiveness in predicting social cohesion scores. Furthermore, the competitive performance of Polynomial Regression and Random Forest highlights the significance of investigating several modeling strategies in tackling complex urban challenges, where nuanced insights and reliable forecasts are of utmost relevance.

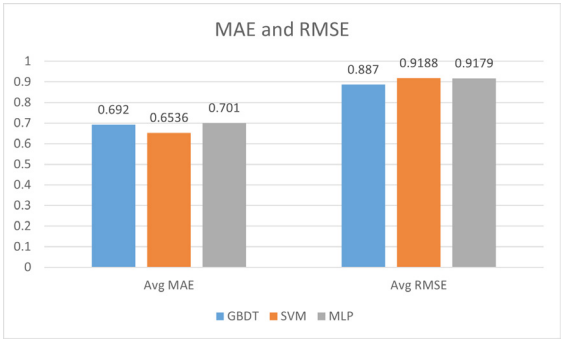

Figure 6. GBDT, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) Error Comparison

Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) were two additional machine learning models that were included in our comparison study, as shown in Figure 6. The average Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for these models throughout cross-validation folds are shown in the graph. The average MAE and RMSE for GBDT are noteworthy at 0.692 and 0.887, respectively. However, SVM shines out because of its exceptional ability to identify underlying patterns in the data, as evidenced by its incredibly low average MAE of 0.6536 and RMSE of 0.9188. With an average MAE of 0.701 and RMSE of 0.9179, MLP, despite slightly lagging, nevertheless exhibits good predictive performance. These

findings highlight the adaptability of machine learning methods; of notice are the excellent accuracy of SVM and the robustness of GBDT. In the case of forecasting social cohesiveness scores, the choice of the most relevant model may depend on certain application needs, such as the need for precision or computational efficiency, underlining the significance of carefully selecting the right method for the task at hand.

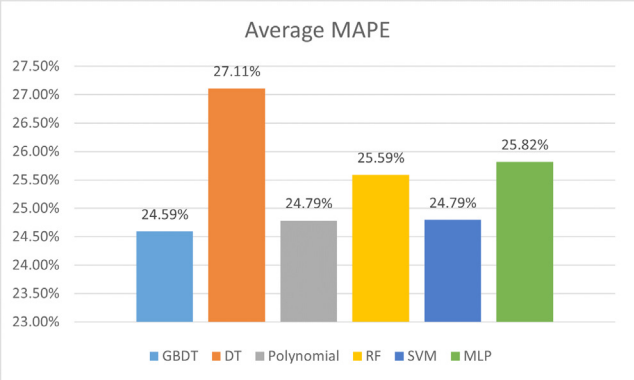

Figure 7. Models Error Percentage Comparison

We analyze the average Mean Absolute Percentage Error (MAPE) throughout cross-validation folds for all the models considered in this study, and the results are shown in Figure 7. With a MAPE of 25.59%, the Gradient Boosted Decision Trees (GBDT) model, which is the suggested model, stands out. This result shows that, on average, the social cohesion scores predicted by GBDT differ by about 25.59% from the actual values. Although GBDT has good predictive performance, placing this finding within the larger pool of models tested is essential. The MAPE for Decision Trees (DT) is somewhat higher at 27.11%, while the MAPE for Polynomial Regression and Support Vector Machine (SVM) are both competitive at 24.79%. With a MAPE of 25.59%, Random Forest (RF) exhibits a comparable level of accuracy, while Multi-Layer Perceptron (MLP) trails slightly with a MAPE of 25.82%. Different MAPE values provide helpful information about the relative accuracy of different models and can help determine which model is best for forecasting Jakarta's Social Cohesion Index. Despite performance differences, GBDT's steady performance upholds its reputation as a reliable and accurate option for this crucial forecasting activity.

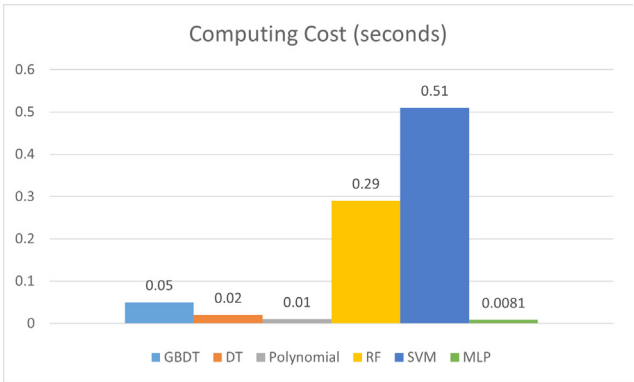

Figure 8. Models Computational Cost Comparison

By calculating the time each model took to forecast the whole evaluation set, Figure 8 provides essential insights into the computational effectiveness of the examined models. The Gradient Boosted Decision Trees (GBDT) model is an outstanding example of efficiency, which can provide predictions in about 0.05 seconds. This efficiency is awe-inspiring, given its strong prediction abilities and the competitive MAE, RMSE, and MAPE values. In comparison, the Decision Trees (DT) model requires only 0.02 seconds, showing quick processing. Polynomial Regression is the most time-efficient among the models, taking only 0.01 seconds, while Random Forest (RF) and Support Vector Machine (SVM) incur slightly longer processing times at 0.29 and 0.51 seconds, respectively. Incredibly, Multi-Layer Perceptron (MLP) performs predictions in just 0.0081 seconds. These results demonstrate a trade-off between computational expense and prediction precision. In real-time or large-scale applications where accuracy and speed are critical factors, GBDT appears attractive because it strikes a compromise between predictive strength and computational efficiency.

IV. CONCLUSION

In conclusion, this study has investigated in-depth machine learning models for estimating social cohesiveness scores in Jakarta's dynamic metropolitan environment. The Gradient Boosted Decision Trees (GBDT) model, put out as a strong candidate, has repeatedly shown exceptional prediction ability among the models tested. In terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), GBDT attained competitive results, demonstrating its competence in capturing the complex patterns underlying social cohesiveness dynamics. Additionally, GBDT demonstrated admirable computing efficiency, providing a helpful edge for large-scale and real-time applications.

Although GBDT stood out as a performer, it is crucial to recognize the larger context of our research. Alternative models with different strengths and weaknesses in terms of accuracy and efficiency included Polynomial Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Specific application needs and constraints should guide when selecting the best model.

Overall, the study emphasizes the critical role that machine learning plays in tackling complex urban issues like social cohesiveness evaluation and draws attention to the potential of data-driven insights to guide community development projects. As we look to the future, the continuous study of cutting-edge machine learning methods and their integration with urban planning and governance processes holds the possibility of producing more resilient and peaceful urban settings.

REFERENCES

Amir, S., Hidayana, I., Rahvenia, Z., & Haydar, S. (2023). Dataset on factors associated with social cohesion of urban life in Jakarta. *Data in Brief*, *49*, 109339. https://doi.org/10.1016/j.dib.2023.109339

Andariesta, D. T., & Wasesa, M. (2022). Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach. *Journal of Tourism Futures*, 1–17. https://doi.org/10.1108/JTF-10-2021-0239

Jewett, R. L., Mah, S. M., Howell, N., & Larsen, M. M. (2021). Social Cohesion and Community Resilience During COVID-19 and Pandemics: A Rapid Scoping Review to Inform the United Nations Research Roadmap for COVID-19 Recovery. *International Journal of Health Services*, *51*(3), 325–336. https://doi.org/10.1177/0020731421997092

Narayan, V., & Daniel, A. K. (2022). Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model. *Journal of Scientific and Industrial Research*, *81*(12), 1297–1309. https://doi.org/10.56042/jsir.v81i12.54908

Ouma, Y. O., Okuku, C. O., & Njau, E. N. (2020). Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya. *Complexity*, *2020*. https://doi.org/10.1155/2020/9570789

Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons and Fractals*, *138*, 109942. https://doi.org/10.1016/j.chaos.2020.109942

Pekel, E. (2020). Estimation of soil moisture using decision tree regression. *Theoretical and Applied Climatology*, *139*(3–4). https://doi.org/10.1007/s00704-019-03048-8

Rybak, A. (2023). Survey mode and nonresponse bias: A meta-Analysis based on the data from the international social survey programme waves 1996 2018 and the European social survey rounds 1 to 9. *PLoS ONE*, *18*(3 March). https://doi.org/10.1371/journal.pone.0283092

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Steiner, A., Woolvin, M., & Skerratt, S. (2018). Measuring community resilience: Developing and applying a "hybrid evaluation" approach. *Community Development Journal*, *53*(1). https://doi.org/10.1093/cdj/bsw017

Tzenios, N. (2020). Examining the Impact of EdTech Integration on Academic Performance Using Random Forest Regression. *ResearchBerg Review of Science and Technology*, *3*(1), 94–106. https://researchberg.com/index.php/rrst/article/view/84

Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, *21*(1), 1–18. https://doi.org/10.1186/s12942-022-00304-5

Zhang, Z., & Jung, C. (2021). GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(7), 3156–3167. https://doi.org/10.1109/TNNLS.2020.3009776