Optimizing Water Quality: A Machine Learning Approach for Safe Drinking and Efficient Irrigation

Abram C. Dorado[1], Kristine Jewel B. Malimban[1], Eleazar Jaren S. Sancio[1], Joana S. Tria[1], Ma Sheila A. Magboo[1][0000-0002-6221-7892] and Vincent Peter C. Magboo[1][0000-0001-8301-9775]

[1] Department of Physical Sciences and Mathematics, University of the Philippines Manila, Manila, Philippines
acdorado2@up.edu.ph, kbmalimban@up.edu.ph, essancio@up.edu.ph, jstria1@up.edu.ph, mamagboo@up.edu.ph, vcmagboo@up.edu.ph

Abstract. Access to water is essential for human survival, animal consumption, irrigation, and various domestic and commercial uses. Unfortunately, the quality of water resources in many areas has declined and has become increasingly contaminated brought by substantial number of human activities. The evaluation of water quality entails labor-intensive and time-consuming processes like the water quality index and are often susceptible to errors. This research study aims to evaluate water quality and to identify the feature selection method which provides the highest capability in predicting water quality. Four supervised machine learning models (random forest, logistic regression. Support vector machine, and gradient boosting) were evaluated to predict water quality in both drinking and irrigation water datasets in four various model configurations assessing the effect of integrating three different feature selection methods namely: correlation method, SelectKBest method and recursive feature elimination with cross validation method. Logistic regression and support vector machine both with correlation feature selection methods appeared to be the best performing models for the drinking water dataset with the metrics ranging from 97-98%. On the other hand, for the irrigation water dataset, support vector machine with recursive feature elimination with cross validation feature selection achieved the best performing model with indices ranging from 98-100%. Based on the high performance and ease of creation of these models, integration of these tools in decision support systems in the actual setting for the swift classification of water quality would then lead to a prompt intervention to address our water needs to sustain human lives.

Keywords: water quality, potability, irrigation, feature selection.

1 Introduction

Water is a crucial resource to sustain the life of humans. It has an essential role to achieve economic prosperity, ensures ecological security, and advancement of human development [1]. Unfortunately, the quality of water resources in many areas had declined and had become increasingly contaminated brought by substantial amount of human activities [2]. The evaluation of water quality entails labor-intensive and time-consuming processes like the water quality index and are often susceptible to errors [3]. Furthermore, these conventional methods often struggle to keep pace with the dynamic nature of water quality, making real-time monitoring and decision-making challenging. In response to these challenges, the incorporation of machine learning (ML) techniques offers an optimistic approach to amplify the efficiency and accessibility of water quality assessment. By leveraging advanced algorithms and large datasets, ML models can predict water quality parameters and assess overall suitability for drinking and irrigation purposes with greater speed.

This research undertaking is an attempt at utilizing data science techniques to evaluate water quality and to identify the feature selection procedure that provides the highest capability in predicting water quality. Through the implementation of various ML algorithms such as random forest (RF), logistic regression (LR), support vector machine (SVM), and gradient boosting (GB), this study endeavors to offer a reliable, efficient, and scalable alternative to traditional laboratory assessments. By utilizing the strength of data-driven insights, this study is envisioned to significantly contribute to the advancement of water quality management practices, clearing the way for informed decision-making and promoting the reliable use of the invaluable water resource.

The structure of this study is illustrated as follows: an introduction section is succeeded by a literature review section underscoring the pertinent research studies regarding the use of ML algorithms on assessing water quality. The next section lays out the particulars of the methodology including the description of the dataset, the preprocessing steps, and the ML algorithms utilized in the study. The main findings of the study and its associated in-depth analysis are expounded in the Results and Discussion section. Ultimately, the study concludes with a summary of the research and recommendations for future investigative work.

2 Literature Review

Makumbura et al. utilized several ML models coupled with explainable artificial intelligence (XAI) methods to evaluate water quality [4]. Results showed XGBoost generated the best predictive capability with the optimum R2 and RMSE. The authors also concluded that the findings offered a dependable and interpretable method for water quality prediction benefiting various stakeholders particularly water specialists and decision-makers. In the study by Jibrin et al., researchers employed a novel method incorporating non-parametric kernel Gaussian learning, adaptive neuro-fuzzy inference system, and decision tree algorithms for assessing water quality [5]. The study affirmed its significant contribution to a larger effort of supervising and safeguarding water resources in arid and semi-arid. In [6], authors employed ML techniques such as SVM, regression trees, linear regression, and neural networks to predict water quality index. SVM and Linear Regression bested other algorithms obtaining the highest R^2 value of 0.99. The findings highlighted the benefits of ML models for accurate water quality prediction aimed at enhancing pollution reduction strategies. In the review made by

Yan et al., recent updates in ML were made for water quality prediction into two segments, indicator prediction and water quality index prediction [7]. Authors analyzed technical details and challenges of the ML methods as well as the future research tasks in this domain.

In [8], supervised and unsupervised ML algorithms in predicting water quality as well as the suitability of incorporating various IoT wireless technologies. Authors emphasized the important role in policymaking processes to secure water quality, recognizing areas in need of prompt intervention measures to forestall contamination. Rajitha et al. [9] examines the incorporation of ML and artificial intelligence procedures in the water treatment methods, highlighting the capacity to precisely identify and reduce contaminants compared to traditional methods. Authors underscored the challenges on scalability and data security and the need for interdisciplinary collaboration to achieve sustainability. Yigit and Baransel studied an autoencoder-based feature selection procedure applicable to labeled and unlabeled data in the assessment of water quality [10]. The researchers concluded that proposed method for water quality assessment had contributed to the efficient management of complex datasets for achieving fruitful environmental supervision and sustainable water resource governance. In [11], ML methods were employed to predict water quality index using four regression models and water quality classification using four classification models. Gradient Boosting model obtained the best classifier with an 99.5% accuracy while multilayer perceptron regressor model generated the highest R2 value of 99.8%. Abbas et al. examined several ML classifiers for water quality index prediction with RF and GB obtaining the highest accuracy rate at 95% and 96%, respectively [3]. Authors believed that the approach contributed to a more extensive discernment of water quality and strengthened the soundness of policymaking as to groundwater utilization.

In summary, these studies collectively showcase significant feasibility of automating water quality monitoring and prediction. The contribution of feature selection methods to upgrade the predictive capability of the ML models in assessing water quality has yet to be fully explored.

3 Materials and Methods

This research study followed a structured pipeline comprising several stages. It commenced with the extraction of the dataset from a publicly available repository followed by data preprocessing steps, including thorough data cleaning. The cleaned dataset was divided into a training set and a testing set. Feature selection techniques were subsequently applied to amplify the predictive potential of the ML models. Finally, the models were assessed based on their predictive performance. The overall framework of this research is illustrated in Fig. 1.

3.1 Dataset Acquisition

The dataset used in this study was sourced from "Dataset for Assessing Water Quality for Drinking and Irrigation Purposes using Machine Learning Models" on IEEE

Dataport, curated from multiple sources [12]. The dataset involves water quality parameters essential for evaluating the appropriateness of water for both drinking and irrigation. The drinking water subset consists of 718 instances with 13 features, and with Potability as the target variable. In this drinking water dataset, a slight class imbalance in is evident with 61.44% labeled as potable and 38.56% as non-potable. On the other hand, the irrigation water subset includes 356 instances with 9 features, with “Usable” as the target variable.

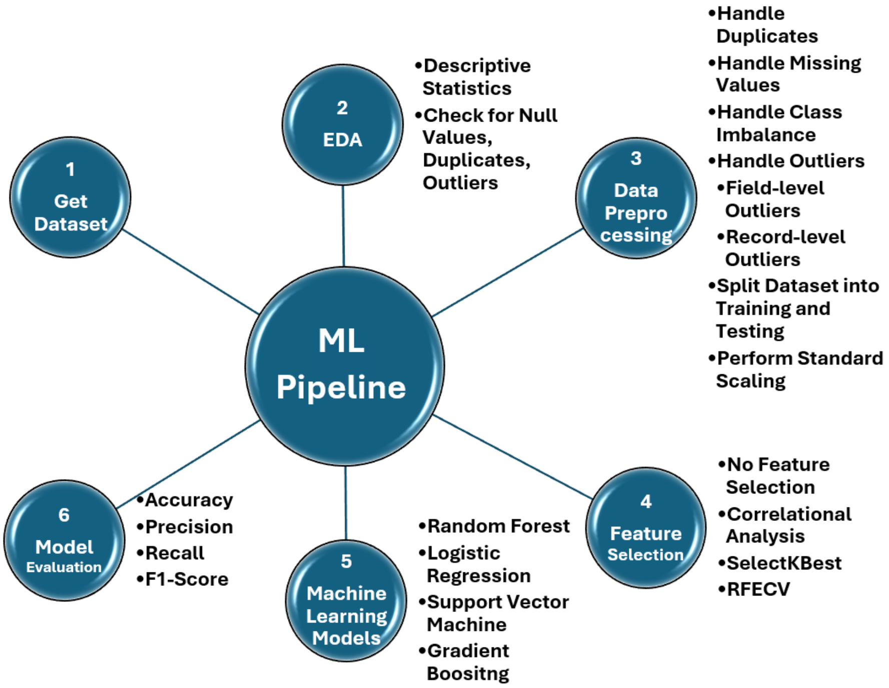

Fig. 1. Machine learning pipeline for Classification of Drinking and Irrigation Water

3.2 Data Pre-processing

Data preprocessing prepares the data for machine learning. Examination of the data for existence of irrelevant, inconsistent attributes, duplicate records, outliers, missing values, or class imbalance issues were handled during this stage. In this study, there were no duplicates nor missing values on both water datasets. Likewise, all outliers, redundant, and irrelevant attributes such as IDs were removed. The data underwent scaling and normalization to maintain consistency and improve model performance. To handle the mild class imbalance issue in the drinking water dataset, carefully selected indices will be given more importance, i.e. precision, recall, and F1-score instead of accuracy.

3.3 Feature Selection Methods

The feature selection procedure aims to determine the most relevant variables in the dataset for building an ML as it captures which attributes contribute highly towards the prediction. This step is crucial because not all features are relevant to the target variable, more particularly the redundant or irrelevant features which potentially lead to increased complexity, risk of overfitting, and reduced interpretability. In this study, three feature selection methods were applied namely: correlation analysis and SelectKBest, both filter-based techniques, and a wrapper method using Recursive Feature Elimination with Cross Validation (RFECV).

3.4 Model Configuration and Machine Learning Algorithms

The dataset was split into 70% training and 30% testing. Four supervised ML algorithms were employed in this study namely: RF, LR, SVM and GB in four model configurations. The configurations consisted of a baseline model which did not include any feature selection procedure while the other three configurations included feature selection using correlation analysis method, SelectKBased method and RFECV method. As there were two water datasets involved, each of the four ML algorithms were trained on each dataset without the application of any feature selection technique resulting in 8 models. As there are three other feature selection methods and four ML algorithms to be applied on two water datasets, additional 24 models were created.

3.5 Model Evaluation

Model evaluation involves assessing performance using applicable indices such as accuracy, precision, recall, and F1-score for classification problems. Accuracy measures how the model correctly predicts. This is acceptable for balanced classes but not applicable for datasets with high class imbalance. For highly imbalanced classes, other metrics should be considered such as precision, recall, and F1-score. Precision provides the percentage of the true positives among all predicted positives. In short precision assesses the accuracy of positive predictions made by the model. Recall, also known as sensitivity or true positive rate, provides the percentage of true positive instances among all the actual positive instances, i.e. it measures the model’s capability to accurately identify positive instances. F1-score combines precision and recall in one index giving equal importance to precision and recall. For datasets with class imbalance issues, a combination of precision, recall and F1-score is necessary to assess model performance. The aim is to obtain high scores in precision, recall, and F1-score.

4 Results and Discussion

The performance indices of the four model configurations for both drinking and irrigation water are presented and discussed in this section. Table 1 summarizes the results of the base model where no feature selection was incorporated. Subsequently, Tables 2, 3, and 4 present the outcomes of configurations employing three different feature selection techniques namely: Correlation Analysis, Select K-Best method, and RFECV method. As seen in Table 1, LR obtained the highest performance indices for the drinking water dataset ranging from 97-99% as well as in the irrigation water dataset with metrics ranging from 95-100%. In the drinking water dataset, SVM, GB and RF obtained mildly lower performance as compared to LR. In the irrigation water dataset, SVM had a comparable performance as compared to LR while GB and RF obtained lower performance results.

Table 1. Performance Indices for the Baseline Model Configuration (No Feature Selection)

Water Dataset	ML Model	Accuracy	Precision	Recall	F1-score
Drinking Water	RF	0.95	0.96	0.97	0.97
	LR	0.97	0.97	0.99	0.98
	SVM	0.96	0.97	0.98	0.97
	GB	0.96	0.96	0.98	0.97
Irrigation Water	RF	0.92	0.95	0.88	0.91
	LR	0.98	1.00	0.95	0.97
	SVM	0.97	1.00	0.93	0.96
	GB	0.95	0.93	0.95	0.94

To determine the effect on the classification performance of utilizing a correlation analysis feature selection, Table 2 shows the performance indices in this model configuration. Contrasted with the baseline model configuration, generally, there was no improvement in the indices for LR, RF, and GB while SVM generated a mild 1 percentage point increase in the drinking water dataset. Likewise, no discernible improvement is seen for all ML models in the irrigation water dataset. Nonetheless, in this configuration, SVM garnered the highest performance for the drinking water dataset while LR remained to be the best model for the irrigation water dataset. The performance indices shown in Table 3 reflect the classification capability for the ML models with SelectKBest feature selection method. In the drinking water dataset, LR and SVM generated a mild decrease of 1 and 1-2 percentage points, respectively in all indices when compared to the baseline configuration of no feature selection method. On the other hand, no improvement can be seen for RF and GB with this added feature selection method. Similarly, no added enhancement in the performance metrics for all ML models in the irrigation water dataset can be noted in this configuration of incorporating SelectKBest feature selection. In this configuration, GB and LR yielded the best performance indices for the drinking and irrigation water datasets, respectively.

Performance indices to assess the added enhancement of utilizing RFECV for all ML models are illustrated in Table 4. For the drinking water dataset, SVM obtained a mild increase of 1 percentage point for its indices (accuracy, recall and F1-score) while LR had a 1-2 mild decrease by 1-2 percentage points when compared to the baseline model configuration of having no feature selection. Generally, no improvement in the indices can be seen for RF and GB. For the irrigation water dataset, SVM obtained a 3-5 percentage point increase particularly in its recall and F1-score. Generally, there was no change in the performance indices for the rest of the ML models in irrigation water dataset in comparison to the baseline configuration. Nonetheless, SVM yielded to be the best performing ML model in both drinking and irrigation water datasets when RFECV is incorporated.

Table 2. Performance Indices for Model Configuration (Feature Selection using Correlational Analysis)

Water Dataset	ML Model	Accuracy	Precision	Recall	F1-score
Drinking Water	RF	0.96	0.97	0.97	0.97
	LR	0.98	0.98	0.98	0.98
	SVM	0.97	0.98	0.99	0.98
	GB	0.96	0.96	0.98	0.97
Irrigation Water	RF	0.92	0.95	0.88	0.91
	LR	0.98	1.00	0.95	0.97
	SVM	0.97	1.00	0.93	0.96
	GB	0.95	0.93	0.95	0.94

Table 3. Performance Indices for Model Configuration (Feature Selection using SelectKBest)

Water Dataset	ML Model	Accuracy	Precision	Recall	F1-score
Drinking Water	RF	0.95	0.96	0.97	0.97
	LR	0.96	0.96	0.98	0.97
	SVM	0.95	0.95	0.98	0.96
	GB	0.96	0.97	0.98	0.97
Irrigation Water	RF	0.92	0.95	0.88	0.91
	LR	0.98	1.00	0.95	0.97
	SVM	0.97	1.00	0.93	0.96
	GB	0.95	0.93	0.95	0.94

Table 4. Performance Indices for Model Configuration (Feature Selection using RFECV)

Water Dataset	ML Model	Accuracy	Precision	Recall	F1-score
Drinking Water	RF	0.95	0.96	0.97	0.97
	LR	0.95	0.95	0.98	0.96
	SVM	0.97	0.97	0.99	0.98
	GB	0.96	0.96	0.98	0.97
Irrigation Water	RF	0.92	0.95	0.88	0.91
	LR	0.98	1.00	0.95	0.97
	SVM	0.99	1.00	0.98	0.99
	GB	0.96	0.97	0.93	0.95

Overall, LR and SVM, both with correlation feature selection methods, appeared to be best performing ML models for the drinking water dataset. For the irrigation water dataset, SVM with RFECV achieved the best performing model.

The underlying principle behind the use of a feature selection step is to highlight only relevant and important data as well as obliterating the noise and redundancy in the dataset. Farghaly and El-Hafeez have reported that datasets with huge number of features are predisposed to overfitting with the ML model becoming very complicated and unable to make sound generalization, enlarging computational complexity, expanding risk of bias due to irrelevant features, and decreased interpretability [13].

Feature selection procedures are labelled as filter-based or wrapper-based methods. Filter-based feature selection takes advantages of data intrinsic features such as statistics or correlation in relation to how the feature is associated with the target variable. The features are ranked and gauged based on the relevance of attributes with the target while immaterial attributes are discarded based on correlation among the attributes [14]. As such, the ML model and the dataset are independent with respect to this feature selection. On the other hand, a wrapper-based feature selection examines the importance of a feature based on a specific ML algorithm by iteratively adding / removing an attribute based on its importance. Hence, the wrapper method is tailor-made to that specific ML algorithm and analyzes the quality of a subset of feature candidates [15].

In this study, correlation analysis and SelectKBest, both filter-based techniques, and a wrapper method using Recursive Feature Elimination with Cross Validation were employed. Correlation analysis is a filter method used to ascertain if there is a strong linear relationship between each pair of variables. Highly correlated attributes should be removed as the non-removal of these features can lead to less reliable statistical inferences. Correlation also gives more importance to features exhibiting moderate to strong correlation with the target variable which can boost predictive capability of the model and ensuring robustness of the model due to the method’s simplicity and interpretability [16]. SelectKBest is another filter method used to score and rank the features based on their relationship with the target. It adopts the attributes with the topmost scores to be included in the final attribute subset. As the features are all numeric in this study, the recommended scoring function is the ANOVA F-statistic of “f_classif” which identifies the features that can best separate the classes. This is a simple but efficient step providing a strategy to decrease dimensionality and acknowledging only the most predictive features while decreasing the computational complexity [17]. However, there is a need to take note that SelectKBest examines the individual feature relevance only and not the potential interactions which may exist between features [18]. In wrapper-based feature selection methods, a subset of attributes is selected, and the model’s performance is evaluated using different subsets of attributes. Wrapper methods are computationally expensive but can identify complex relationships between features [14]. RFECV, a wrapper method, recursively removes features from the dataset and then evaluates the performance of a machine learning model at each step. It starts with all features and ranks them based on their importance to the target then removes the least important feature(s) and repeats the process until the desired model performance is reached [19].

The foremost impediment of this research investigation lies in the chosen dataset as it only contained relatively fewer attributes in assessing water quality. It is therefore suggested to include other larger datasets including many other water sample attributes.

5 Conclusions and Recommendations

The quality of water resources in many areas had declined and had become increasingly contaminated brought by substantial amount of human activities. The evaluation of water quality entails labor-intensive and time-consuming processes like the water quality index and are often susceptible to errors. The utility of harnessing machine learning research is an optimistic approach to upgrade the efficiency and accessibility of water quality assessment. By harnessing advanced algorithms and large datasets, machine learning models can predict water quality parameters and assess overall suitability for drinking and irrigation purposes with greater speed. In this work, four supervised machine learning models were evaluated to examine water quality in both drinking and irrigation water datasets in four various model configurations assessing the effect of integrating three different feature selection methods namely: correlation method, SelectKBest method and recursive feature elimination with cross validation method. Logistic regression and support vector machine both with correlation feature selection methods appeared to be the best performing models for the drinking water dataset. On the other hand, for the irrigation water dataset, support vector machine with recursive feature elimination with cross validation feature selection achieved the best performance indices. Based on the high performance and ease of creation of these models, integration of these tools in decision support systems in the actual setting for the swift classification of water quality would then lead to a prompt intervention to address our water needs to sustain human lives.

For the succeeding research activity, it is recommended to use another dataset with the inclusion of more measures or attributes of water quality. Additionally, further research studies need to assess other feature selection steps and other machine learning classifiers resulting in more model configurations in the assessment of water quality for drinking and irrigation purposes.

References

1. X. Han, X., M.W. Boota, S. Soomro, et al.: Water strategies and management: current paths to sustainable water use. Appl Water Sci 14, 154 (2024). https://doi.org/10.1007/s13201-024-02214-2.
2. Murti M.A, Saputra, A.R.A., Alinursafa I., et al.: Smart system for water quality monitoring utilizing long-range-based Internet of Things. Appl Water Sci 14, 69 (2024). https://doi.org/10.1007/s13201-024-02128-z.
3. Abbas F., Cai Z., Shoaib M., et al.: Machine Learning Models for Water Quality Prediction: A Comprehensive Analysis and Uncertainty Assessment in Mirpurkhas, Sindh, Pakistan. Water 16, 941 (2024). https://doi.org/10.3390/w16070941.
4. Makumbura R.K., Mampitiya L., Rathnayake N., et al.: Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial

intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature. Results in Engineering, Volume 23, 102831 (2024). https://doi.org/10.1016/j.rineng.2024.102831.

5. Jibrin A.M., Al-Suwaiyan M., Aldrees A, et al.: Machine learning predictive insight of water pollution and groundwater quality in the Eastern Province of Saudi Arabia. Sci Rep 14, 20031 (2024). https://doi.org/10.1038/s41598-024-70610-4.
6. Doddabasappaar A.K.T., Yogendra B.E., Janardhan P., Siddegowda P.N.: Machine Learning for Water Quality Index Forecasting. EMSI, vol. 3, pp.43–53, (Apr. 2024). https://doi.org/10.46604/emsi.2024.12870.
7. Yan X., Zhang T., Du W., Meng Q., Xu X, Zhao Z.: A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. J. Mar. Sci. Eng. 12, 159 (2024). https://doi.org/10.3390/jmse12010159.
8. Essamlali I., Nhaila H., El Khaili M.: Advances in machine learning and IoT for water quality monitoring: A comprehensive review. Heliyon, Volume 10,Issue 6, e27920 (2024). https://doi.org/10.1016/j.heliyon 2024.e27920.
9. Rajitha A., K A., Nagpal A., et al.: Machine Learning and AI-Driven Water Quality Monitoring and Treatment. E3S Web Conf., volume 505, number 03012 (2024). https://doi.org/10.1051/e3sconf/202450503012.
10. Yiğit G.Ö., Baransel C.: Utilizing machine learning techniques for enhanced water quality monitoring. Water Quality Research Journal, 8, wqrj2024007 (2024). https://doi.org/10.2166/wqrj.2024.007.
11. Shams M.Y., Elshewey A.M., El-kenawy ES M, et al.: Water quality prediction using machine learning models based on grid search method. Multimed Tools Appl 83, 35307–35334 (2024). https://doi.org/10.1007/s11042-023-16737-4.
12. Olasupo Ajayi, Antoine Bagula, Hloniphani Maluleke, July 19, 2022 : Dataset for Assessing Water Quality for Drinking and Irrigation Purposes using Machine Learning Models. IEEE Dataport, doi: https://dx.doi.org/10.21227/trcf-1s03.
13. Farghaly H.M., El-Hafeez T.A.: A high-quality feature selection method based on frequent and correlated items for text classification. Soft Comput 27, 11259–11274 (2023). https://doi.org/10.1007/s00500- 023-08587-x.
14. Moslemi A.: A tutorial-based survey on feature selection: Recent advancements on feature selection. Engineering Applications of Artificial Intelligence, Volume 126, Part D (2023), 107136. https://doi.org/10.1016/j.engappai.2023.107136.
15. Noroozi Z., Orooji A., Erfannia L.: Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Sci Rep 13, 22588 (2023). https://doi.org/10.1038/s41598- 023-49962-w.
16. Mohtasham F., Pourhoseingholi M., Hashemi Nazari S.S., et al.: Comparative analysis of feature selection techniques for COVID-19 dataset. Sci Rep 14, 18627 (2024). https://doi.org/10.1038/s41598-024-69209-6.
17. Abid-Althaqafi N.R., Alsalamah H.A.: The Effect of Feature Selection on the Accuracy of X-Platform User Credibility Detection with Supervised Machine Learning. Electronics 13, 205 (2024). https://doi.org/10.3390/electronics13010205.
18. Tariq M.A.: A Study on Comparative Analysis of Feature Selection Algorithms for Students Grades Prediction," Journal of Information and Organization Sciences. 48 (1): 133-147 (2024). https://doi.org/10.31341/jios.48.1.7.
19. Awad M., Fraihat S.: Recursive Feature Elimination with Cross- Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems. J. Sens. Actuator Netw. 12, 67 (2023). https://doi.org/10.3390/jsan12050067.