Comparative Analysis: Machine Learning Algorithms for TOC Prediction in Pharmaceutical Water Treatment Systems

Dieki Rian Mustapa[1]*, Aris Tjahyanto[2]
Department Management Interdicipline dan Technology's School [1]
Department of Information Systems [2]
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dieki.rian.mustapa@gmail.com [1], aristj@its.ac.id [2]

Abstract— Water quality is crucial in pharmaceutical production, where it serves as a solvent and raw material. Contamination with organic compounds poses a risk to product integrity and safety. TOC serves as a key indicator for assessing organic pollution levels in water. An increase in TOC signals potential issues with water treatment systems. Machine learning prediction of TOC values is essential for preemptive monitoring and maintenance. This study aimed to compare three different machine learning algorithms - Linear Regression (RL), Random Forest (RF), and multilayer perceptron (MLP) - for predicting Total Organic Carbon (TOC) in pharmaceutical water treatment systems. By utilizing a dataset covering various operational conditions of pharmaceutical water treatment systems, the research conducted a comprehensive analysis. Each algorithm underwent evaluation using performance metrics like coefficient of determination (R-squared), and prediction accuracy to assess their effectiveness in predicting TOC concentrations. A correlation coefficient approaching 1 (100%) signifies a strong relationship between model predictions and actual target values (accuracy prediction), while a smaller Mean Absolute Error (MAE) indicates higher accuracy in predicting target values. The study found that the results of the correlation coefficient in order from highest to lowest are the RF, MLP, and RL models with values of 95.04%, 93.11%, and 80.27%, respectively. Likewise, additional metrics for evaluation, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE), exhibit a ranking from lowest to highest values across RF, MLP, and RL models. RF has a higher prediction accuracy of the TOC than other models (95%) and lowest MAE (3.9). This research offers valuable insights into utilizing machine learning algorithms for TOC prediction within pharmaceutical water treatment to make informed decisions, improving water treatment systems and overall product quality.

Keywords— Machine Learning, Total Organic Carbon (TOC), Pharmaceutical Water Treatment Systems, Algorithm Comparison, Water Quality Assessment

I. INTRODUCTION

The pharmaceutical industry is one sector that relies heavily on water quality in its production process. Water is widely used as a raw material and solvent in the processing, formulation and manufacture of pharmaceutical products, active pharmaceutical ingredient (API) and intermediates [1]. Therefore, water quality in the pharmaceutical industry plays a very important role. Contamination of water with unwanted organic components can threaten the integrity and safety of pharmaceutical products. Microbiological impurities are the most critical in water systems and are therefore of particular concern [2]. TOC has become one of the crucial parameters regulated under the Good Pharmaceutical Manufacturing Practice (CPOB) No. 34 of 2018. Consequently, it becomes mandatory for industries to adhere to these regulations.

Total organic carbon (TOC) serves as a dependable indicator for estimating the organic content quantity within water samples [3]. It stands as a primary parameter utilized for assessing organic pollution levels within water bodies. An increase in TOC results in a water treatment system is an important indicator that the number of organic molecules present in the water is increasing, or there are different, more complex molecules entering the water treatment system. Such changes in TOC results can be an indicator of the integrity of the water treatment system [4]. This organic carbon can be highly toxic and persistent, accumulate in the human body, and result in adverse health impacts [5].

Currently, machine learning method approaches are widely used to evaluate TOC content [6]. Supervised learning approaches with these methods have been widely used to predict TOC values [7]. In some studies, Supervised learning algorithms used are algorithms with Liner Regression, Random Forest and Multi-layer Perceptron methods. Research using machine learning to analyze water quality has been done, for example predicting the water quality value of the Ciliwung River, based on the results of evaluating the models used from 4 types of Machine Learning methods studied, it is known that Random Forest is the highest at 99.7% and JST 94.6% [8]. Further advancements in machine learning modeling pertaining to determining the total organic carbon (TOC) content in shale formations involves employing regression analysis. TOC content is established once both the Fisher distribution, indicating the significance of each model, and Student's t-distribution, representing the significance of variables within

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

the models, reach values equal to or surpassing their respective threshold values at a 95% confidence level. Through the utilization of 45 sets of logging measurements, the newly proposed correlation demonstrates an ability to replicate TOC values with a root mean-squared absolute difference (RMSAD) of 0.30 wt % and a root mean-squared relative difference (RMSRD) of 23.8% [9]. Assessing a source rock involves various parameters, including the determination of Total Organic Carbon (TOC). This determination traditionally relies on costly laboratory tests and is constrained by the availability of rock samples. However, TOC prediction using well log data, accessible from most oil and gas wells, offers a solution by providing continuous data on organic content. Hence, employing well log data for prediction emerges as an ideal method for TOC determination in source rock units. This study aims to forecast TOC values using well logs through the application of the Multi-Layer Perceptron Artificial Neural Network (ANN) technique. Eighteen data samples from the Talang Akar Formation were utilized to train and test the MLP-ANN model. The well log data employed for TOC prediction encompass density log (RHOB), transit time (DT), deep resistivity (ILD), gamma-rays (GR), and neutron porosity (NPHI), yielding a strong correlation (R2 0.87) and low mean absolute percentage error (MAPE) of 10% against the resulting MLP-ANN model. Such a TOC prediction technique holds promise for aiding geophysicists and reservoir geologists in source rock evaluation within oil and gas fields, obviating the necessity for extensive source rock sample datasets [10]. Another study mentioned the use of MLP-ANN to predict TOC values in well logs for oil and gas searches getting an accuracy rate of 87% [11]. Random forest has gained popularity as a machine learning technique for constructing predictive models across various research contexts [12]. In a research endeavor aimed at constructing a Total Organic Carbon (TOC) prediction model within the Rumaila oil field in Iraq, the findings validated the capacity of machine learning models to develop effective estimations of TOC utilizing accessible borehole log data, thereby obviating the need for costly coring procedures [13].

From several studies that have been conducted, the researcher takes a further approach related to the comparison of the modeling produced using the Liner Regression, Random Forest and Multi-layer Perceptron method approaches will show the prediction of the TOC value of the water treatment system. How do the predictive capabilities of Linear Regression, Random Forest, and Multi-layer Perceptron methods compare in forecasting Total Organic Carbon (TOC) values in water treatment systems, and how can the optimal machine learning model aid the pharmaceutical industry in efficient decision-making for water treatment process efficiency. The prediction of TOC values from the optimal machine learning model can later be used by the pharmaceutical industry for quick decision making in the efficiency of the water treatment process. In addition, the modeling results can identify critical process parameters related to TOC values, so that the pharmaceutical industry can determine effective water treatment system maintenance management related to these parameters.

The paper structure consists of several key sections aimed at comprehensively addressing the research objectives. Section 1 serves as the introduction, providing an overview of the study's significance, objectives, including discussions on previous studies related to machine learning methods such as Linear Regression, Random Forest, and Multi-layer Perceptron for TOC prediction in water treatment systems. This section also explores the importance of TOC prediction in water treatment processes and the significance of identifying critical process parameters for effective maintenance management. In Section 2, the paper delves into a detailed review of relevant literature. Section 3 presents the research methodology employed in the study, including problem identification, determine research objective, data collection, preprocessing techniques, and the implementation of machine learning algorithms. In Section 4, the results of the comparative analysis between the three machine learning methods are presented, along with insights into the predictive capabilities and performance metrics. This section also discusses the implications of the findings for the pharmaceutical industry, emphasizing how the optimal machine learning model can facilitate quick decision-making and efficient water treatment system maintenance management. Section 5 offers conclusions, summarizing result, and suggesting avenues for future research in this field.

The research contributes both theoretically and practically. Theoretical contribution lies to identify critical parameters influencing Total Organic Carbon (TOC) in water treatment systems, enriching the scientific understanding in water treatment and machine learning fields. Meanwhile, the practical contribution involves implementing Machine Learning models combining data acquisition and Linear Regression/Random Forest/Multi-layer Perceptron algorithms at pharmaceutical industry. ML development enabling the identification of critical TOC-affecting parameters. Consequently, the developed Machine Learning models empower the company to implement early warning systems based on TOC prediction values, facilitating reactive management to water quality changes and timely actions.

II. LITERATURE REVIEW

A. Linear Regression

Linear regression aims to uncover connections and interdependencies among variables by modeling the association between a single continuous scalar dependent variable, typically labeled as y or the target in machine learning, and one or more explanatory variables denoted as X, which could be a D-dimensional vector see Fig. 1. These explanatory variables are also referred to as independent variables, input variables, features, observed data, attributes, dimensions, data points, and so forth [14]. This modeling process utilizes a linear function to depict the relationship between these variables. In regression analysis, the goal is to predict a continuous target variable [15].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

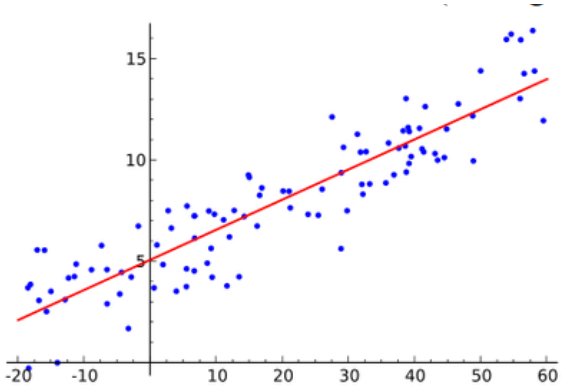

Fig. 1. Visual representation of the linear regression [16]

B. *Random Forest*

Random forest is an algorithm that combines many decision trees built randomly from labeled data. It can be used for classification or regression, by taking the average or mode of the results of the trees, see Fig. 2. Random forests can improve prediction accuracy and stability compared to using a single decision tree. A decision tree represents a classifier expressed as a recursive partition of the example space. A decision tree consists of nodes that form what is called a root tree [17].

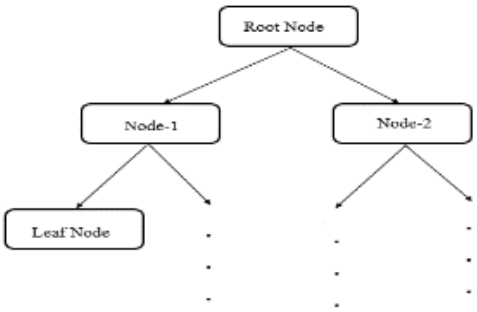

Fig. 2. Decision tree example [18]

C. *Multi-layer preceptor*

The multilayer perceptron (MLP) represents a neural network variant characterized by a supervised learning approach, employing the back-propagation method. As illustrated in Fig. 3, MLP is structured with three layers: an input layer, a hidden layer, and an output layer. In this architecture, every neuron is interconnected with all neurons within each layer. [19]. An MLP, short for multilayer perceptron, belongs to the category of artificial neural networks (ANNs) [20]. Its structure encompasses a minimum of three layers of nodes: input, hidden, and output layers. Each node, functioning as a neuron, applies a nonlinear activation function, excluding those in the input layer [21].

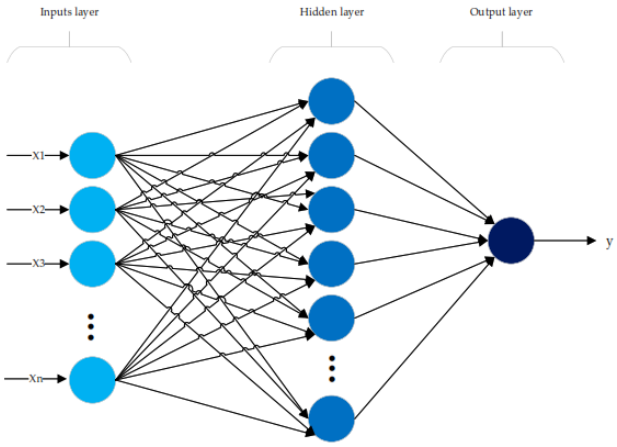

Fig. 3. Architecture MLP [22]

D. *Machine learning development using WEKA application*

WEKA is a machine learning application intended to assist in the application of machine learning techniques where it is capable of data preprocessing, visualization of results, database linkage, and cross-validation, and machine learning modeling [23]. WEKA accepts data in ARFF format which is an attribute relation file format, comma separated CSV format and other formats [24].

There are two categories of classifiers supervised and unsupervised. Three basic steps for classification in WEKA refer to Fig. 4

1. *Preparing the data*
2. *Choose a classification and apply the algorithm*
3. *Analyze the results or output [25]*

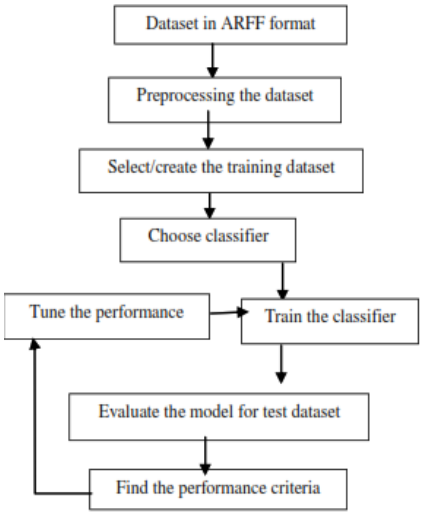

Fig. 4. Step processing in WEKA [26]

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

III. RESEARCH METHODOLOGY

The research methodology in this journal consists of 7 steps, which are described in the following *Fig. 5*

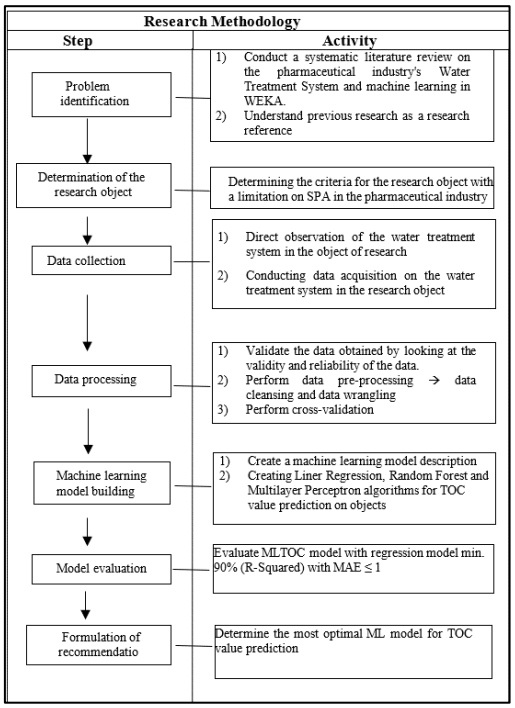

Fig. 5. Diagram of Research Methodology

A. *Problem identification*

In this context, machine learning approaches, such as linear regression, random forest, and multi-layer perceptron, provide a choice of solutions to model the complex relationship between system parameters and TOC values. This research aims to develop a machine learning model that can assist industry in identifying key parameters that affect TOC values in water treatment systems.

B. *Determination of the research object*

The research object used in this study is the water treatment system at one of pharmaceutical industry. This research will focus on aspects related to water quality management in the system, with the aim of developing Machine Learning modeling to identify critical parameters that affect Total Organic Carbon (TOC) refer to **Error! Reference source not found.**.

TABLE I. RESEARCH OBJECT CRITERIA

Criteria for Water Treatment System (WTP) at Pharmaceutical Industry
C.1 Water treatment system with performance parameter sensors.
C.2 The water treatment system has at least installation qualification, operational qualification and performance.
C.3 Data from the SPA from existing sensors can be transferred and stored in the SPA database.
C.4 Data transmission is real-time and data gathering can be done into the data analytics platform.
C.5 Clearly labeled data, not null, empty, accurate, does not contain errors, missing values.
C.6 Clear metadata, including variable definitions, data formats, units of measurement, categories, and data origin.
C.7 Data is collected with calibrated tools so that the resulting data is valid and reliable.

C. *Data collection*

This research uses two methods in collecting data, namely observation and data acquisition methods in the water treatment system of the pharmaceutical industries in Indonesia. Observation and data acquisition were carried out on the object of research on the industry's water treatment system. Observation is done by following the process of the water treatment system from the object of research to find out the real - time conditions of the system. The data acquisition method involves collecting data through sensors that automatically generate data.

The water treatment system has been qualified for design, installation, operational, and performance stages 1 and 2 conducted throughout the year 2023. This indicates that the data collected from the system are deemed valid and reliable for subsequent data processing stages. The dataset collected in 2023 comprises 78 attributes (columns) with 1414 instances (rows). The following are some examples of the names of the predictor parameters (columns) contained in the data, PWG_TOC_PWGEN_Value,PWG_CL_101,PWG_CM_104, PWG_Delta_PT03_PT04,PWG_Delta_PT05_PT06,PWG_TOC_Looping, PWG_TT_101, etc. The data types present in the dataset include numeric and nominal data. Numeric data represent variables with numerical values, while nominal data represent variables with specific categories or labels. The absence of missing values in the dataset indicates that each entry in the dataset is complete.

D. *Data Processing*

The first step is to validate the data by checking the data to ensure measurement accuracy, deleting invalid or outlier data, missing/null/zero data [27]. The next step is data pre-processing by normalizing the data if the units of measurement are different. Data is organized in a suitable format and compiled in a table or spreadsheet with columns representing observed

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

variables and rows representing observations.

Data Exploration (EDA) is performed to understand patterns, correlations, and trends in the observed data [27]. EDA can involve data visualization, descriptive statistical calculations, and graphical analysis [28]. Determining variables that are relevant and have a significant impact on TOC values. The TOC selection taken as the value to be predicted is TOC PW Gen (PWG_TOC_PWGEN_Value), this is because TOC supply is one of the values that determines whether the results of the water purification process are running well or not. The process of dividing datasets, known as data splitting, involves separating them into training data and testing data. The training data is utilized to train the model, whereas the testing data is employed to evaluate the model's performance. Option method will be used cross-validation with 10 folds.

E. Machine learning development

The Machine Learning models used in this research are generated by WEKA application with classifier specific to Liner Regression, Random Forest and Multi-layer Perceptron to predict the Total Organic Carbon (TOC) value of the water treatment system [26]. The selection of these models is based on the results of a literature review that has been conducted by researchers based on prediction results, model complexity, and implementation capabilities on other systems such as Arduino. Later the optimal model will be selected from the algorithms used. Liner Regression, Random Forest and Multi-layer Perceptron models are given training data to train them. The model will use various data samples from the training dataset to build decision trees. Once the model is trained, testing data is used to test the performance of the model.

F. Machine learning evaluation

Several assessment criteria that can be employed such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), Mean Squared Error (MSE), coefficient of determination (R-squared), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) [29]. A correlation coefficient approaching 1 (100%) signifies a strong relationship between model predictions and actual target values (accuracy prediction), while a smaller Mean Absolute Error (MAE) indicates higher accuracy in predicting target values [30]. This model evaluation helps understand the extent to which the model is able to accurately predict TOC values.

If the evaluation results are inadequate, the model can be optimized by adjusting the model-parameters or using other techniques such as feature selection or feature engineering [31]. The constructed model can be used to predict TOC values in water treatment systems. Interpretation of the results helps in understanding the factors affecting TOC and in taking appropriate measures to improve the water treatment system.

G. Formulation of recommendation

Recommendations are built based on the optimal results of machine learning modeling to predict TOC values in water treatment systems. For future research the use of the model can formulate predictor factors that can affect TOC values while model simulation can determine the limits of each predictor.

IV. RESULT AND ANALYSIS

A. Data Collection & Processing

Data collection is carried out by observation and using the SCADA system of water treatment process which will be used to develop machine learning models. The object of research already has compliance with the requirements where the system already qualified and has several sensors that are used to collect data on all operational parameters including sensors to collect real time TOC data. Here is an example of the value of the data. The type of data obtained is numeric and nominal data.

TABLE II. EXAMPLE OF DATASET FROM PWS SCADA SYSTEM

Date	Time	Conduct	Flow. Meter	Pressure	TOC Supply
9/18/ 2023	13:00	0.1503	20117.18	2.996961	8.247932
9/18/ 2023	14:00	0.1545	20139.79	3.091724	9.016540
9/18/ 2023	15:00	0.1593	20187.71	3.021918	9.623336
9/18/ 2023	16:00	0.1447	20137.98	3.062789	9.627831
9/18/ 2023	17:00	0.1595	20123.51	3.066044	9.475008
9/18/2023	18:00	0.1791	20128.03	3.040726	9.169363

The first stage in preprocessing is data cleansing. The data obtained from the SCADA system is analyzed more deeply to understand the structure of the information. Of the 78 attributes (columns) with 1414 instances (rows), there are only 49 attributes that will be used, because these data are directly related to the TOC value to be predicted (PWG_TOC_PWGEN_Value). Other columns that are not related will be "removed" in the WEKA application. Referring to table 3. These columns are attribute numbers 1-13, 59-68, and 73-78. The next process is Attribute Selection. This process can help in selecting the most relevant attributes or reducing the dimensionality of attributes thereby improving model performance and reducing overfitting. In this research, the attribute selection evaluator used is CfsSubsetEval (Correlation-based Feature Selection) and the BestFirst method available in WEKA. This method aims to select a subset of attributes that have a high correlation relationship with the target class, but low between each other.

From the results of the Attribute Selection process, 6 attributes are obtained that are highly correlated directly to the target class with a correlation of 100% or 90% and low correlation with each other, namely Purified water generator chlorine value (PWG_CL_101_Value), Purified water generator conductivity value 4 (PWG_CM_104_Value), Purified water generator delta pressure transmitter 2 (PWG_Delta_PT03_PT04_Value), Purified water generator delta pressure transmitter 3 (PWG_Delta_PT05_PT06_Value), Purified water generator TOC Looping Value (PWG_TOC_Looping_Value), Purified water generator TT Value 1 (PWG_TT_101_Value).

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

B. Machine learning development & evaluation

1) Regression Linear (RL)

The first machine learning modeling process is with the Linear Regression classifier. In using linear regression with cross validation 10 folds and the target class is PWG_TOC_PWGEN_Value. The model evaluation shows the following results. The correlation coefficient obtained from the test data is 80.27%, MAE is 6.0928, RMSE is 15.1646, RAE is 38.59%, and RRSE is 60.78%, and total number instance 1414.

2) Random Forest (RF)

The second machine learning modeling process is with the Random Forest (RF) classifier. In using RF with cross validation 10 folds and the target class is PWG_TOC_PWGEN_Value The model evaluation shows the following results. The correlation coefficient obtained from the test data is 95.04%, MAE is 3.9685, RMSE is 7.7581, RAE is 25.139%, and RRSE is 31.1042%, and total number instance 1414.

3) Multi-Layer Perceptron (MLP)

The third machine learning modeling process is with the Multi-Layer Perceptron (MLP) classifier. In using MLP with cross validation 10 folds and the target class is PWG_TOC_PWGEN_Value. The results are obtained as shown in the summary result image below. The correlation coefficient obtained from the test data is 93.11%, MAE is 5.8336, RMSE 9.2581, RAE 36.954%, and RRSE 37.1183%, and total number instance 1414.

4) Evaluation

Machine Learning models, the conditions used are data that have gone through preprocessing with the same test option conditions using 10 Folds Cross-Validation and several different calcifications, namely Linear Regression (RL), Random Forest (RF) and Multi-Layer Perceptron (MLP). Obtained 3 learning models as follows RL model, RF model, and MLP model. The conclusions of the evaluation results of several training model evaluations are listed in the following **Error! Reference source not found.**.

TABLE III

Evaluation Metrics	Model Name		
	RL	MLP	RF
Correlation coefficient	80.27%	93.11%	95.04%
Mean absolute error	6.0928	5.8336	3.9685
Root mean squared error	15.1646	9.2581	7.7581
Relative absolute error	38.5958%	36.9540%	25.1390%
Root relative squared error	60.7989%	37.1183%	31.1042%
Total Number of Instances	1414	1414	1414

From the **Error! Reference source not found.**, it is found that the results of the correlation coefficient in order from highest to lowest are the RF, MLP, and RL 2 models with values of 95.04%, 93.11%, and 80.27%, respectively. Similarly, other metric evaluation parameters such as MAE, RMSE, RAE, and RRSE which have the lowest to highest values are RF, MLP and RL models.

Researchers conducted a comparison again using new data (in Apr 2024 contain 102 instance) that was not used to train the model. This new data is used to test the reliability of the model, whether the models have evaluation results that are patterned the same as during model training. The following **Error! Reference source not found.** are the results of model evaluation using new data.

TABLE IV

Evaluation Metrics	Model Name		
	RL	MLP	RF
Correlation coefficient	81.29%	89.11%	95.12%
Mean absolute error	1.9353	1.9919	0.9877
Root mean squared error	2.8068	2.6713	1.473
Relative absolute error	53.3581%	54.9100%	27.2300%
Root relative squared error	60.8089%	57.8700%	31.9114%
Total Number of Instances	102	102	102

From the **Error! Reference source not found.**, it is found that the results of the correlation coefficient in order from highest to lowest are the RF, MLP, and RL models with values of 95.12%, 89.11%, and 81.29% respectively. Similarly, other metric evaluation parameters such as MAE, RMSE, RAE, and RRSE which have the lowest to highest values are RF, MLP and RL models. It is found that the RF model is better in evaluation compared to MLP and RL.

This RF model can be a reference to be used as a model that is able to predict the target class of PW generator TOC value where some of the related parameters that greatly affect the TOC value are Purified water generator chlorine value (PWG_CL_101_Value), Purified water generator conductivity value 4 (PWG_CM_104_Value), Purified water generator delta pressure transmitter 2 (PWG_Delta_PT03_PT04_Value),Purified water generator delta pressure transmitter 3 (PWG_Delta_PT05_PT06_Value), Purified water generator TOC Looping Value (PWG_TOC_Looping_Value), Purified water generator TT Value 1 (PWG_TT_101_Value). Researchers can suggest to the industry to monitor more closely the 6 parameters to better control the TOC value of the PW generator.

V. CONCLUSION

Analysis of the machine learning model evaluation results on new data, it can be concluded that the Random Forest (RF) model consistently demonstrates superior performance compared to the Linear Regression (RL) and Multi-Layer Perceptron (MLP) models. Evaluation conducted using several metrics, including correlation coefficient, MAE, RMSE, RAE, and RRSE. The study found that the results of the correlation

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

coefficient in order from highest to lowest are the RF, MLP, and RL models with values of 95.04%, 93.11%, and 80.27%, respectively. Likewise, additional metrics for evaluation, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE), exhibit a ranking from lowest to highest values across RF, MLP, and RL models. It indicates that RF provides better performance results than another model. RF shows the highest correlation coefficient compared to RL and MLP, with a value of 95.04%. Additionally, RF also has the lowest mean absolute error (3.9685), lowest root mean squared error (7.7581), as well as lower relative absolute error (RAE) and root relative squared error (RRSE) compared to the other two models. This indicates that the RF model provides more accurate predictions that closely approximate the true values in the new dataset.

In the context of its application in the industry, these results have significant implications. Parameters such as PWG_CL_101,PWG_CM_104,PWG_Delta_PT03_PT04,PWG_Delta_PT05_PT06, PWG_TOC_Looping, PWG_TT_101, which have been proven to significantly influence the TOC PW generator values, need to be closely monitored to enhance control over the TOC values. By using the RF model as a reference, industries can optimize monitoring and quality control strategies to ensure the reliability and stability of production processes.

Further research could involve refining the RF model or exploring advanced variations to improve model performance. Moreover, practical implementation studies across diverse industrial sectors could validate the efficacy of RF models in improving production process control and quality assurance. Thus, this research contributes valuable insights into the development and implementation of machine learning models in industry, particularly in the monitoring and quality control of production processes.

REFERENCES

[1] T. Sandle, "Chapter 14 - Assessment of pharmaceutical water systems," in Biocontamination Control for Pharmaceuticals and Healthcare (Second Edition), T. Sandle, Ed., Academic Press, 2024, pp. 313–327. doi: https://doi.org/10.1016/B978-0-443-21600-8.00014-2.

[2] F. Roeder and T. Sandle, "Microbial Contamination in Water Systems," PDA J Pharm Sci Technol, p. pdajpst.2021.012636, Jan. 2022, doi: 10.5731/pdajpst.2021.012636.

[3] H.-S. Lee, J. Hur, and H.-S. Shin, "Enhancing the total organic carbon measurement efficiency for water samples containing suspended solids using alkaline and ultrasonic pretreatment methods," Journal of Environmental Sciences, vol. 90, pp. 20–28, 2020, doi: https://doi.org/10.1016/j.jes.2019.11.010.

[4] A. Shetty and A. Goyal, "Total organic carbon analysis in water – A review of current methods," Mater Today Proc, vol. 65, pp. 3881–3886, 2022, doi: https://doi.org/10.1016/j.matpr.2022.07.173.

[5] Y. Huang, L. Zhang, and L. Ran, "Total Organic Carbon Concentration and Export in a Human-Dominated Urban River: A Case Study in the Shenzhen River and Bay Basin," Water (Basel), vol. 14, no. 13, 2022, doi: 10.3390/w14132102.

[6] L. Zhu, X. Zhou, W. Liu, and Z. Kong, "Total organic carbon content logging prediction based on machine learning: A brief review," Energy Geoscience, vol. 4, no. 2, p. 100098, 2023.

[7] L. Goliatt, C. M. Saporetti, and E. Pereira, "Super learner approach to predict total organic carbon using stacking machine learning models based on well logs," Fuel, vol. 353, p. 128682, 2023, doi: https://doi.org/10.1016/j.fuel.2023.128682.

[8] J. T. Lingkungan, M. Haekal, and W. C. Wibowo, "Prediksi Kualitas Air Sungai Menggunakan Metode Pembelajaran Mesin: Studi Kasus Sungai Ciliwung Prediction of River Water Quality Using Machine Learning Methods: Ciliwung River Case Study," vol. 24, no. 2, pp. 273–282, 2023.

[9] J. Wang, D. Gu, W. Guo, H. Zhang, and D. Yang, "Determination of Total Organic Carbon Content in Shale Formations With Regression Analysis," J Energy Resour Technol, vol. 141, no. 1, Jan. 2019, doi: 10.1115/1.4040755.

[10] R. C. Wibowo, O. Dewanto, and M. Sarkowi, "Total organic carbon (TOC) prediction using machine learning methods based on well logs data," in AIP Conference Proceedings, AIP Publishing, 2022.

[11] . C. Wibowo, O. Dewanto, and M. Sarkowi, "Total Organic Carbon (TOC) Prediction Using Machine Learning Methods Based on Well Logs Data," in AIP Conference Proceedings, American Institute of Physics Inc., Oct. 2022. doi: 10.1063/5.0103209.

[12] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," Expert Syst Appl, vol. 134, pp. 93–101, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.028.

[13] A. M. Handhal, A. M. Al-Abadi, H. E. Chafeet, and M. J. Ismail, "Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms," Mar Pet Geol, vol. 116, p. 104347, 2020, doi: https://doi.org/10.1016/j.marpetgeo.2020.104347.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis. John Wiley & Sons, 2021.

[15] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," Journal of Applied Science and Technology Trends, vol. 1, no. 2, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[16] S. Badillo et al., "An Introduction to Machine Learning," Clin Pharmacol Ther, vol. 107, no. 4, pp. 871–885, Apr. 2020, doi: https://doi.org/10.1002/cpt.1796.

[17] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," Stata J, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[18] S. S. Azmi and S. Baliga, "An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies," Int. Res. J. Eng. Technol, vol. 7, no. 5, pp. 6867–6870, 2020.

[19] S. Nosratabadi, S. Ardabili, Z. Lakner, C. Mako, and A. Mosavi, "Prediction of food production using machine learning algorithms of multilayer perceptron and ANFIS," Agriculture, vol. 11, no. 5, p. 408, 2021.

[20] F. Yang, H. Moayedi, and A. Mosavi, "Predicting the degree of dissolved oxygen using three types of multi-layer perceptron-based artificial neural networks," Sustainability, vol. 13, no. 17, p. 9898, 2021.

[21] E. R. AlBasiouny, A.-F. A. Heliel, H. E. Abdelmunim, and H. M. Abbas, "Multilayer Perceptron Generative Model via Adversarial Learning for Robust Visual Tracking," IEEE Access, vol. 10, pp. 121230–121248, 2022, doi: 10.1109/ACCESS.2022.3222867.

[22] M. I. C. Rachmatullah, J. Santoso, and K. Surendro, "A Novel Approach in Determining Neural Networks Architecture to Classify Data With Large Number of Attributes," IEEE Access, vol. 8, pp. 204728–204743, 2020, doi: 10.1109/ACCESS.2020.3036853.

[23] J. Pavic, "An Introduction to WEKA: The All-in-One Machine Learning Software in Java".

[24] A. Sadiq, "Intrusion Detection Using the WEKA Machine Learning Tool," 2021.

[25] B. Saleh, A. Saedi, A. al-Aqbi, and L. Salman, "Analysis of Weka Data Mining Techniques for Heart Disease Prediction System," International Journal of Medical Reviews, vol. 7, no. 1, pp. 15–24, 2020, doi: 10.30491/ijmr.2020.221474.1078.

[26] S. F. Mohd Radzi, M. S. Hassan, and M. A. H. Mohd Radzi, "Comparison of classification algorithms for predicting autistic spectrum disorder using

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024

WEKA modeler," BMC Med Inform Decis Mak, vol. 22, no. 1, p. 306, 2022, doi: 10.1186/s12911-022-02050-x.

[27] V. Da Poian et al., "Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry," Frontiers in Astronomy and Space Sciences, vol. 10, p. 1134141, 2023.

[28] D. T. H. S. Tariq and P. S. Aithal, "Visualization and Explorative Data Analysis," Int J Enhanc Res Sci Technol Eng, vol. 12, no. 3, pp. 11–21, 2023.

[29] Engr. Dr. F. Obodoeze, C. Nwabueze, and S. Akaneme, "Comparative Evaluation of Machine Learning Regression Algorithms for PM2.5 Monitoring," American Journal of Engineering Research, vol. 10, pp. 19–33, Dec. 2021.

[30] J. Rong et al., "Machine Learning Method for TOC Prediction: Taking Wufeng and Longmaxi Shales in the Sichuan Basin, Southwest China as an Example," Geofluids, vol. 2021, 2021, doi: 10.1155/2021/6794213.

[31] A. Apaza-Pinto, J. Esquicha-Tejada, P. López-Casaperalta, and J. Sulla-Torres, "Supervised Machine Learning Techniques for the Prediction of the State of Charge of Batteries in Photovoltaic Systems in the Mining Sector," IEEE Access, vol. 10, pp. 134307–134317, 2022, doi: 10.1109/ACCESS.2022.3225406.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2148, Copyright ©2024
Submitted : April 25, 2024, Revised : May 11, 2024, Accepted : May 13, 2024, Published : June 15, 2024