Figure 6. Safety Management Systems.
Source: https://ehsdailyadvisor.blr.com

Types of undertakings and the number of workers for which an overall safety and health controller(Soukatsu-Anzen-Eisei-Sekininsha) must be appointed

- Undertakings involving tunnel construction, bridge construction (limited to construction on roads or sites close to roads, or on rails or sites close to rails, within areas of concentrated population) and work using compressed-air methods, in which more than 30 workers are regularly employed, including subcontractor workers.

- Other undertakings in which more than 50 workers are regularly employed, including subcontractor worker.

Employers who must appoint a safety and health supervisor(Anzen-Eisei-Sekininsha)

- All subcontractors must appoint a safety and health supervisor when they participate in a construction undertaking for which an overall safety and health controller (Soukatsu-Anzen-Eisei-Sekininsha) has been appointed.

Employers who must appoint a master safety and health supervisor (Motokata-Anzen-Eisei-Sekininsha)

- Master employers must appoint a master safety and health supervisor for work undertaken by said employer itself if an overall safety and health controller (Soukatsu-Anzen-Eisei-Sekininsha) has been appointed for the undertaking.

Employers who must appoint a safety supervisor, health supervisor and industrial physician

- Master employers and each subcontractor must appoint a safety supervisor, health supervisor and industrial physician when such master employer and subcontractor employ more than 50 workers at the workplace.

Employers who must appoint an operations-chief

- All employers must appoint an operations chief to any undertaking for which the appointment of an operations-chief (Sagyo-Shuninsha) is legally required.

Risk Management Optimization:

Implement comprehensive risk management, from assessing and identifying potential hazards to developing prevention and response plans. Diligent risk management helps minimize workplace accidents and incidents during construction.

Understand the risks

Begin by identifying hazards in your workplace. A hazard is anything that may cause harm, such as chemicals, electricity, or equipment.

After you determine what hazards exist in your workplace, the next step is to assess the risk these hazards pose to workers, so you can dedicate the appropriate

(xem tiếp trang 80)

Figure 7. Professional Safety Training. Source: https://imcinstitute.ae; https://mtvco.vn/

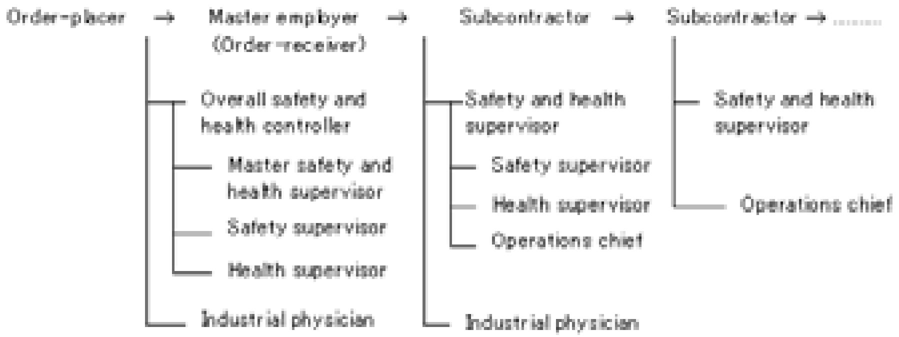


Figure 8. Safety and health management systems. Source:https://www.jisha.or.jp

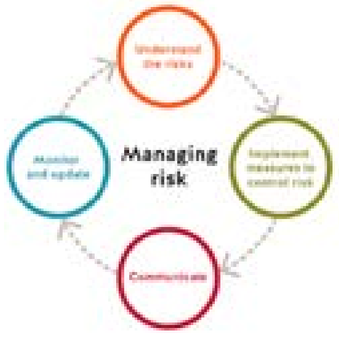


Figure 9. Safety and health management systems.
Source:https://www.jisha.or.jp

Evaluation of the ability of the random forest algorithm in machine learning for studying construction hydraulics

Nguyen Minh Ngoc[1], Bui Hai Phong[2]

Abstract

The Decision Tree and Random Forest algorithm is a "black box" prediction model, this algorithm is formed based on the "binary tree" structure. The study conducted an analysis of the structure of the Random Forest algorithm, built a process for analyzing and predicting a hydraulic factor using the regression algorithm. In particular, Pi theory is used to analyze and determine the objective function, thereby determining the input data fields for the Machine Learning model, coordinating experimental data with physical experimental model of the hydraulic jump in the trapezoidal channel. The study analyzed machine learning models according to Decision Tree and Random Forest algorithms, the research results showed good computational efficiency, strong correlation coefficient (R2 ≥ 0.9), other statistical indicators are very close to the ideal point (zero), the MAPE is from 3% to 6%. The study also shows that the Random Forest model has better prediction performance than the Decision Tree for the hydraulic factors of water jumping in an horizontal trapezoidal channel.

Key words: Machine Learning, Decision Tree, Random Forest, Pi theory, Hydraulic jump

1. Introduction

Machine Learning (ML) is a powerful branch of artificial intelligence (AI) research that has been developed since the 1980s. Machine Learning is a field of Computer Science that has the ability to learn on its own based on input data without requiring specific programming algorithms for the research subject. Instead of writing programming lines for software manually with a specific set of instructions to complete a specific task, the machine is "trained" using large amounts of data and algorithms that allow it to learn how to perform tasks [1].

Machine learning has a very close relationship with statistical theory. Machine learning uses statistical models to "remember" the distribution of data, simulating the ability to generalize and infer, and then produce forecast results based on optimal statistical indicators (such as MSE index). Machine learning can only predict accurately within the input data range, while predicting outside the analyzed data range will give results with low accuracy. Therefore, empirical formulas still have certain advantages, especially forecasting trends outside the data range or areas with sparse data. Machine Learning is used for studies on object classification or regression prediction.

Machine Learning is a solution applied to many fields and industries, from social sciences, information technology, finance, medicine, remote sensing, electronics, robotics and other engineering fields. In this scope, analyzing the studies that have applied Machine Learning algorithms in the field of water resources and hydraulic engineering. This study will be based on machine learning algorithms on regression.

2. Related works

Study on applying Machine Learning algorithms to engineering, hydraulics have been synthesized and analyzed, the results have shown the usefulness of applying Machine Learning in the flow research, as Steven L. Brunton has overviewed Machine Learning and basic applications in fluid mechanics. It's shown that Machine Learning applications give good efficiency in the hydraulic research, it can support in evaluating physical models and sharing data better [2]. Melhem has analyzed Machine Learning solutions and applied in engineering with suitable databases, Machine Learning algorithm gives fast results and high accuracy [1]. In the analysis and calculation of flow for river basins, Corentin J. Lapeyre presented a machine learning solution in analyzing hydrological and hydraulic factors of river basins, to forecast water level and flow of Garonne River, France using Random Forest (RF) algorithms, 2D hydraulic model, multi-layer neural network (MLP), research to establish the relationship between Machine Learning model and 2D hydraulic model, to reduce costs for research and analysis of hydraulic characteristics of flow [3]. Elaheh White studied flow forecasting in some California rivers and streams, USA using RF model and evaluated by R2 > 0.8 [4]. Granata predicted spring flows in Rasiglia Alzabove, Umbria, Central Italy with DT and RF machine learning models for analysis and used the R2 evaluation index to achieve quite good results with R2 = 0.991 [5] or Ziyao Xu's water price analysis [6] predicted water price growth using the RF model in the US and used the R2 evaluation index for a value greater than 0.8, showing that Machine Learning helps in deciding on appropriate and rapid water price adjustments.

Studies show that the application of Machine Learning algorithms is still very limited, and its application has not been widely deployed like the application of neural networks (ANN) of Artificial Intelligence systems. Especially in the field of hydraulics and hydraulic engineering, there are also very few studies, which limits the ability to apply and expand the research database system. Making new research

[1] Department of Water technology, of Urban Environmental and Infrastructural Engineering, Hanoi Architectural University, Hanoi, Vietnam, Email:ngocnm@hau.edu.vn; Tel: 0396050595
[2] Department of Computer networks and information systems, Faculty of Information Technology, Hanoi Architectural University, Hanoi, Vietnam, Email: phongbh@hau.edu.vn; Tel: 0915594033

Date of receipt: 25/10/2024
Editing date: 26/11/2024
Post approval date: 04/11/2024

eqautions only applicable to each specific research case, there are no comprehensive analyses and construction of general calculation methods in the field of hydraulic engineering.

In Machine Learning, there are two basic algorithms, Decision Tree (DT) and Random Forest (RF), which have a wide range of applications and do not require too strict input data, and the results ensure accuracy for scientific research on hydrodynamics.

Therefore, this research focuses on the evaluation and analysis of DT and RF models by regression algorithms, the research includes theory, data characteristics in Machine Learning, the determination of data fields in Machine Learning from Buckingham's Pi theory and analytical examples of geometric characteristics (the sequent depths and the jump length) of the hydraulic jump in a horizintal trapezoid channel.

3. Characteristics of data used in the machine learning

The database used in the machine learning must satisfy the following requirements [4,6,8]:

+ The data set must be of sufficient quality when analyzed, and must be divided into a training data-set and a test set, in which the training data-set must have complete classification variables and target variables, while the test data-set does not have target variables.

+ The training data-set must be abundant and diverse in terms of variables and data attributes so that the training process for the model takes place optimally and the analysis results are accurate.

+ The classes, groups or values of the target variable must be discrete, and analysis cannot be applied to continuous target variables (ensuring that the analysis process is not affected by values that follow a certain function, then the training results will not be accurate).

Thus, the data used for machine learning solutions must ensure independence and have a relationship with the target variable. In the study of the hydraulic jump, the factors of the sequent depths and the jump length considering friction factors are all independent and satisfy the database characteristics applied to the machine learning process.

4. Theoretical basis of research

Regression algorithms based on the “binary tree” theory include the DT and RF models, in which the RF algorithm is an upgrade of the DT algorithm with the “bagging” technique to increase the model’s prediction efficiency. The research algorithms are described as follows:

4.1 Decision Tree algorithm

Decision Tree (DT) [6,10] is one of the predictive modeling approaches used in statistics, data mining, and machine learning. Using DT (as a predictive model) to go from observations of a target value of that parameter (represented in the branches) to conclusions (represented in the leaves). DT is the most powerful algorithm in the supervised algorithm category. The main entities of DT are “Nodes” and “Leaves”, which are where the data is split and where the prediction results appear. The decision of the split strategy greatly affects the accuracy of the tree. The regression theory of DT is to use the Mean Square Error (MSE) to decide whether to split a node into two or more “child” nodes.

Considering a binary DT, the algorithm will first select a value and divide the data into two subsets. For each subset, it will calculate the MSE value separately. The "tree" selects the value that results in the smallest MSE value. DT uses the method of dividing into different regions, each region uses the smallest MSE criterion to evaluate:

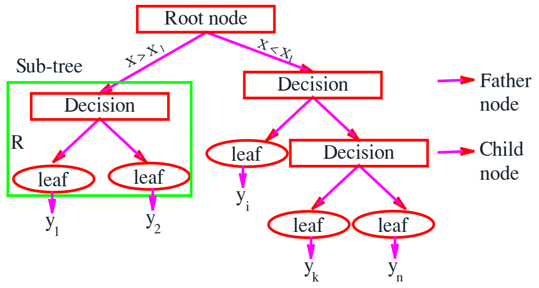


Figure 1. Structure of the DT model

$$\min_R\left[\sum\left(y_i-\widehat{y_R}\right)^2\right] \tag{1}$$

where: yi is the actual predicted value in the region R; $\widehat{y_R}$ is the average value of the region R.

In practice, it is difficult to split the data into regions R, so the alternative is to use the binary DT method. That is, each node has only 2 branches/leafs and analyzes all variables in the data field in turn.

4.2 “Random Forest” algorithm

Random Forest (RF) [6,8] is a supervised learning method, so it can handle problems of classification and prediction of values (Regression). The mathematical explanation of the algorithm is as follows: RF is a set of many DTs, in which each DT is randomly created from resampling (randomly selecting a part of the data to build) and randomly selecting variables from all variables in the data. With such a mechanism, RF gives us a very high accuracy result, but the trade-off is that we cannot grasp the operating mechanism of this algorithm due to the overly complex structure of the model.

This algorithm can be considered one of the Black Box modeling methods and it can only draw results but cannot explain the operating mechanism of the model. The RF methods are usually trained using a “bag-based” approach, i.e. each “Tree” is analyzed by a separate regression function to predict the analysis result.

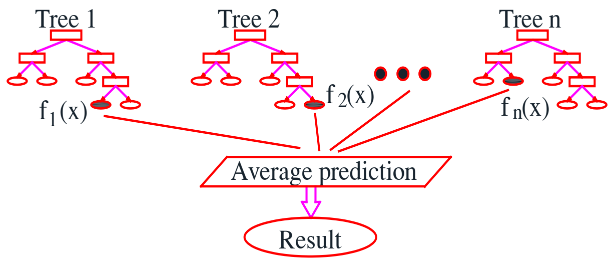


Figure 2. Model of RF's forecasting algorithm

The algorithm of the RF model can be described as follows:

$$g(x)=\sum_{i=1}^{n}\alpha_i f_i(x) \tag{2}$$

Where: $f_i(x)$ is the base model of the DT.

α_i is Weight of the DT model in the RF algorithm.

Table 1. Analyzing the objective function of the jump according to studies

a. For the study of the sequent depth of the jump	b. For the study of the jump length
+ Frictional force	
The friction force (F_{ms}) can be ignored, only the surface roller characteristics are considered.	Using the bottom stress theory (τ_o) and average wetted perimeter ($\overline{P}$) to analyze the friction force: $F_{ms} = \tau_o L_j \overline{P}$ (4)
+ The equation of influencing factors is as follows:	
$f(y_2,\ y_1,\ Q,\ V_1,\ V_2,\ ,\ m,b,a_o,r,m,\ g) = 0$ (5)	$f\left(L_J,y_1,y_2,V_1,V_2,m,\rho,\mu,g,b,\alpha_o,e\right)=0$ (6)
+ Applying Buckingham's Pi theory to determine the objective function in studying the intinial depth of the jump (y1), the second depth of the jump (y2) and the length of the jump (L_J) according to the upstream Froude number of the jump (Fr_1), as follows:	
$\frac{y_2}{y_1} = \Psi\left(\frac{my_1}{b},Fr_1\right)$ (7)	$\frac{L_J}{y_1} = \Phi\left(\frac{y_2}{y_1},Fr_1,\frac{my_1}{b}\right)$ (8)

Table 2. Experimental data from physical models

Authors	Bed width b (cm)	Length of channel (m)	Q (l/s)	y_1 (m)	y_2 (m)	Lj (m)
Ngoc N.M. et al.	55	4	60 ÷ 201	0.033 ÷ 0.081	0.182 ÷ 0.488	1.0 ÷ 2.87
	33.5		65 ÷ 167	0.041 ÷ 0.092	0.258 ÷ 0.490	1.5 ÷ 3.18
Wanoscheck R. et al.	20	6	20.1 ÷ 98	0.040 ÷ 0.081	0.142 ÷ 0.441	0.55 ÷ 3.0

This technique is widely used and has good prediction performance, also known as ensemble model. In the RF model, the DT base models are built independently with a separate data sample.

4.3 Data structure of machine learning algorithms

The Machine learning model is a training process, this process requires input data set and output data to analyze according to different algorithms. The input data of the influencing variable will be divided into "nodes" (the DT algorithm), divided into "Trees" (the RF algorithm) etc. Thus, the basic structure of the machine learning model is as follows:

+ Training process: Training = training (data 1, data 2,... data n)

+ Forecasting process: Forecasting result = f (Training, test data)

Determining data fields is a difficult problem, because these data fields cannot be proposed randomly or empirically, but need to have a clear scientific basis. The method of dimensional analysis and applying Buckingham's Pi theory is a suitable method in building data fields in machine learning models.

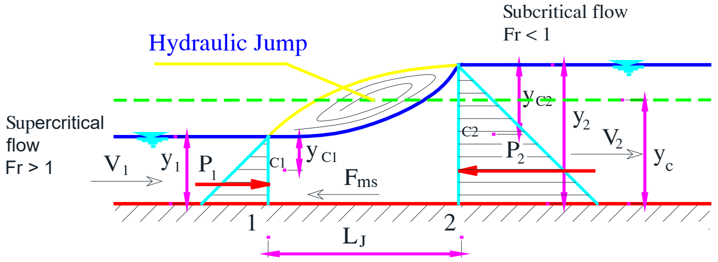


Figure 3. The structure of the jump in the horizontal trapezoidal channel

In addition, the predicting results of the machine learning algorithms will be evaluated through statistical indicators, such as correlation coefficient (R2), mean square error (MSE), absolute percentage error (MAPE, %), mean absolute error (MAE), and error between mesured and predicted values (ε, %). These indicators are used to evaluate the forecasting performance of the models.

5. Illustrating the machine learning problem through the study of predicting the hydraulic characteristics of a hydraulic jump

5.1 Theoretical equation of the jump

Considering a structure of the jump in a horizontal trapezoidal channel. Study of the momentum equation for the roller zone of the jump:

$$P_1 - P_2 - F_{ms} = \rho Q_2 \alpha_{02} V_2 - \rho Q_1 \alpha_{01} V_1 \tag{3}$$

where:

P_1, P_2 is the hydrostatic pressure at sections 1 and 2 (N)

V_1, V_2 is the average velocity at cross-sections 1 and 2(m/s)

α_{01}, α_{02} is the momentum coefficient

Q_1, Q_2 is the dischagre at upstream and downstream of the jump (m^3/s)

Based on the objective function, determining the forecast target data (y2/y1 or LJ/y1) and the data fields affected in the training process of the machine learning model.

5.2 Experimental model and research data fields

The study used 3 physical models with the same conditions (trapezoidal channel, stilling basin, channel slope m = 1 and smooth bed) to collect data.

Separating about a part of the research data to make test data, should be randomly separated and ensure that the test data is spread evenly within the research data range (It can use the INDEX function in Excel software to separate the data for objectivity). The selection principle is based on the coverage of data within the research scope, here considered through the upstream Froude number, the test data has a Froude number from 4.0 to 7.78 (steady jump), with a total of 109 data sets, separate 12 data sets (11% of the total data) for testing (Table 4).

Table 3. Research data fields

Values	y_2/y_1	Fr_1	M_1=my_1/b	L_J/y_1
Min	2,347	2,125	0,060	9,091
Max	9,926	10,570	0,406	62,50

Table 4. Test data for the machine learning model

Values	y_2/y_1	Fr_1	M_1	L_J/y_1
Min	4.587	4.062	0.087	28.125
Max	7.879	7.776	0.406	48.966

5.3 Deploying and testing the machine learning models

In the study of applying the machine learning algorithms in the library system of Matlab R2022b software, the algorithms have the following structure (Figure 5).

Predicted values of the ML models (Figure 6, 7, 8).

Analyzing of statistical indicators in evaluating predicted values of the Machine learning models

For the the DT model, the arrangement of data fields to separate at different nodes will give the model different forecasting efficiency, as seen when forecasting the jump length, Fr_1 or y_2/y_1 used as data division at the nodes, gives different statistical indicators when comparing the measured and predicted values.

For the RF model, the data fields are not meaningful when arranged to be included in the training model. Because with the RF, the division of data into DTs will be randomly arranged and evenly distributed, so all cases will be overviewed for study.

6. Conclusion

The machine learning model (DT, RF) has an algorithm suitable for the characteristics of predicting hydraulic factors of construction works. When studying Machine Learning, it is necessary to clearly identify the influencing variables to build data fields. The study proposed the dimensional analysis method and Pi theory for analysis, the results showed that this method is very suitable when combined with machine learning algorithms.

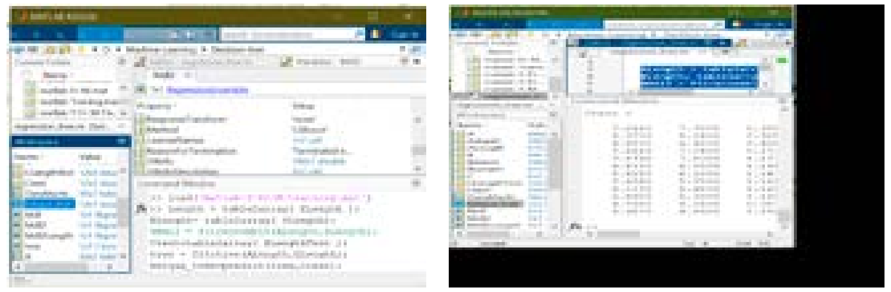
The machine learning problem with good forecasting efficiency, in addition to fully identifying the influencing factors, must also have a rich data set.

The study has only initially analyzed, proposed and tested the forecast for hydraulic factors of water jumps with a short data set. Therefore, if applied in practice, it is necessary to compare with empirical formulas to select suitable hydraulic factors.

Comparing the predicted data of the RF and DT, it shows that the efficiency of RF is better, which is clearly shown in the forecast with the same test data series, such as the high correlation coefficient ($R^2 \geq 0.9$), other statistical indicators of the RF are much smaller than the DT.

This study is limited to the use of "binary tree" algorithm with DT and RF models to study and evaluate the suitability of Machine Learning Model in regression prediction of hydraulic factors. Upgrading studies will be carried out in subsequent evaluations.

Regression machine learning model is an algorithm with rich applications, especially when used in the R ecosystem (such as in Matlab software), it is very effective, the research method is convenient and especially the change of analysis methods is quick and gives good calculation efficiency./.



a. codes of the DT

b. Codes of the RF

Figure 4. Codes of the DT and RF model in the Matlab R2022b

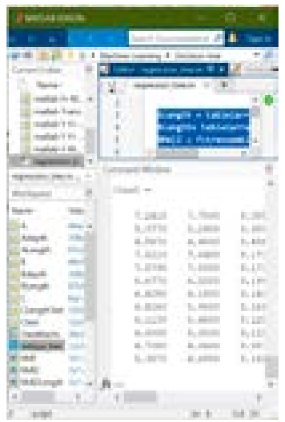


Figure 5. Test data on predicting the jump length

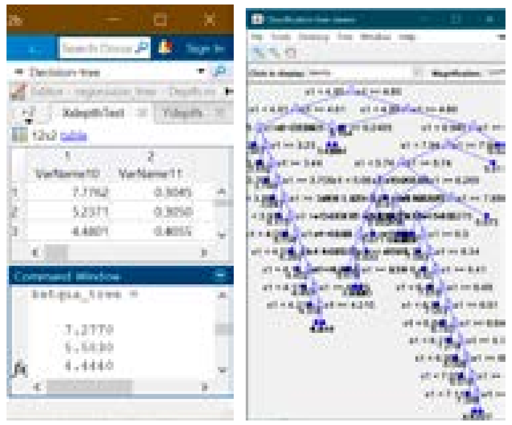


Figure 6. Predicted values and "Tree" diagram of the DT algorithm on the sequent depth

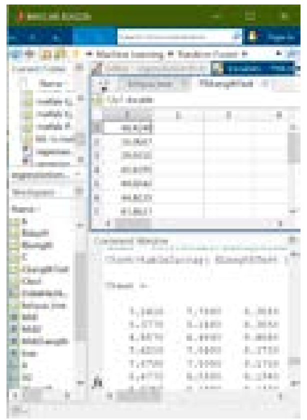


Figure 7. Predicted value of the jump length of the RF model

Table 5. Statistical indicators of predicting the sequent depth

No.	Models	Influence variable	MEA	MSE	R^2	MAPE (%)	ε_{max} %
1	DT	Fr_1, M_1	0,380	0,212	0,798	5,856	13,2
2	RF	Fr_1, M_1	0,206	0,112	0,893	3,177	13,9

Table 6. Statistical indicators of predicting the jump length

No.	Models	Influence variable	MEA	MSE	R^2	MAPE (%)	ε_{max} %
1	DT	y_2/y_1, Fr_1, M_1	1,263	2,169	0,949	3,466	7,25
2		Fr_1, y_2/y_1, M_1	2,251	7,197	0,832	5,998	13,46
3	RF	y_2/y_1, Fr_1, M_1	1,144	2,023	0,953	2,988	6,95

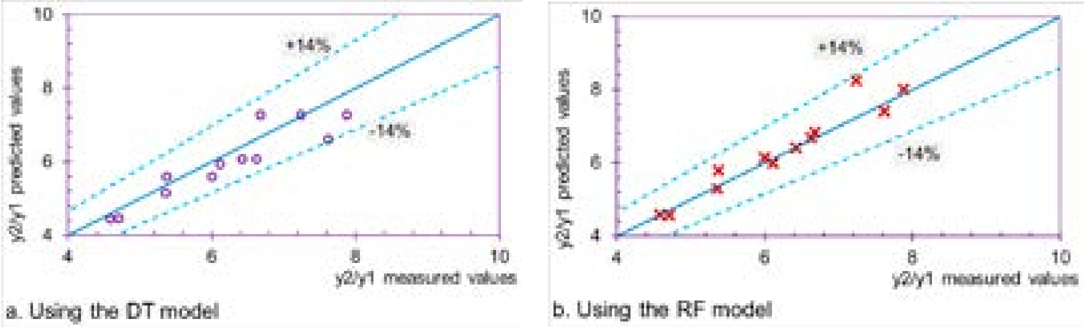


Figure 8. Comparison of measured and predicted valuess of the sequent depth by Machine Learning models

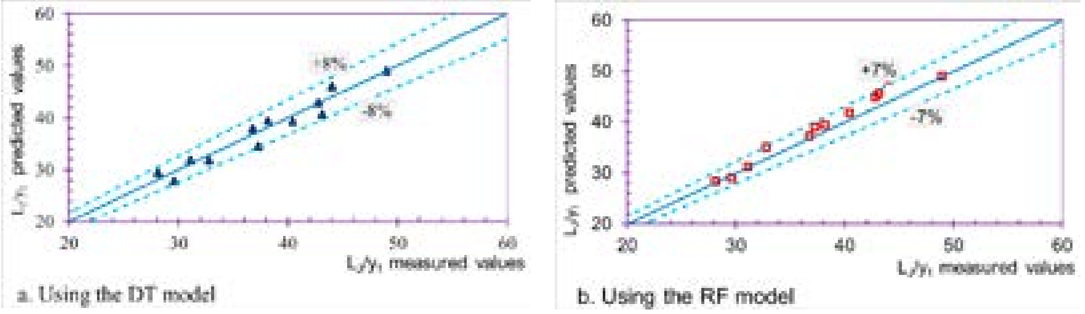


Figure 9. Comparison of measured and predicted valuess of the jump length by Machine Learning models

References

1. *Hani G. Melhem and Srinath Nagaraja. Machine learning and its application to civil engineering systems. Civil Engineering Systems, Vol. 13 (4), 259-279, 1996.*
2. *Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine Learning for Fluid Mechanics. Annu. Rev. Fluid Mech, Vol. 52, 477–508, 2020.*
3. *Corentin J. Lapeyre, Nicolas Cazard, Pamphile T. Roy, Sophie Ricci and Fabrice Zaoui (2019). Reconstruction of hydraulic data by machine learning. SimHydro 2019 conference held in Sophia Antipolis, Nice, France, 701-715, 2019.*
4. *Elaheh White. Predicting Unimpaired Flow in Ungauged Basins: Random Forests Applied to California Streams. Master's thesis of Science in Civil and Environmental Engineering, University of California, US, 2017.*
5. *Francesco Granata, Michele Saroli, Giovanni de Marinis, Rudy Gargano. Machine learning models for spring discharge forecasting. Geofluids, Volume 2018, 21 – 34, 2018.*
6. *Ziyao Xu, Jijian Lian, Lingling Bin, Kaixun Hua, Kui Xu 1 and Hoi Yi Chan. Water Price Prediction for Increasing Market Efficiency Using Random Forest Regression: A Case Study in the Western United States. Water, 11, 228, 2019.*
7. *Makridakis, Spyros. Accuracy measures: theoretical and practical concerns. International Journal of Forecasting, Vol. 9(4), 527-529, 1993.*
8. *Scott Hartshorn. Machine Learning with Random Forests and Decision Trees: A Visual Guide for Beginners. Amazon, 2016.*
9. *Leo Brei man, Jerome Friedman, Charles J.Stone and Richarcd A. Olshen. Classification and Regression Trees (1st Edition). Chapman and Hall/CRC, 2017.*
10. *Chris Smith, Mark Koning. Decision Trees and Random Forests: A Visual Introduction For Beginners: A Simple Guide to Machine Learning with Decision Trees. Blue Windmill Media, 2017.*