EPBM Advance Rate Prediction Using Hybrid Feature Selection and Support Vector Regression Modeling

Shengfeng Huang(✉), Misagh Esmaeilpour, Pooya Dastpak, and Rita Sousa

Department of Civil, Environmental, and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ, USA
shuang54@stevens.edu

Abstract. Advance rate (AR) prediction is crucial for optimal mechanized tunneling performance. However, the type of input features used when developing AR prediction models vary greatly from study to study. In this paper, a hybrid automatic feature selection method is presented and demonstrated through the development of a support vector regression (SVR) model for AR prediction in Earth pressure balance machine (EPBM) tunnel construction. EPBM datasets are collected from a tunnel project in the city of Porto, Portugal. Irrelevant features whose values are constant for most of the time were first removed via constant and quasi-constant detection method (CQD). Sequential forward selection (SFS) was then performed to determine the best subset of features to develop the best performed model. The results showed that the SVR model successfully predicted AR using the selected features with squared correlation coefficient (R^2) of 0.919 and 0.884 for training and testing, respectively. The efficiency of the feature selection method is demonstrated by comparing the results of the SVR model with feature selection and the one without. It is proved that proposed method helps improve the accuracy of the predictions by 8% and 17% for training and testing, respectively.

Keywords: EPBM · tunnelling performance · advance rate · hybrid feature selection · SVR

1 Introduction

The growing traffic in metropolitan areas and the need for high-performance rail systems has driven cities to use their underground space for transport infrastructure [1,2]. This resulted in increased construction of tunnels and continuous development in tunnelling technology in the past years. In urban environments, tunnel boring machines (TBM) are widely used due to their suitability to drive in a broad range of geological conditions and their increased safety. Given the risks of tunnelling in metropolitan areas, making an accurate prediction of TBM

S. Javankhoshdel and Y. Abolfazlzadeh (Eds.): TMIC 2022, AHE 13, pp. 253–264, 2023.
https://doi.org/10.2991/978-94-6463-104-3_22

performance is of great importance to ensure a safe and economical construction [3,4].

To properly asses TBM performance, one needs to accurately forecast advance rate (AR) or penetration rate (PR), i.e., the speed of tunnel excavation, since a realistic prediction of AR directly affects our ability to estimate cost, times of completion and assess risk [5]. However, it is difficult to make accurate predictions of AR as AR is not directly controlled by operator but a result of a complex interaction between the ground and the TBM [6,7]. Operators, particularly their experience and ability to react to observed parameters, have also a great influence on TBM driving performance. During the design phase, thresholds are set for operational TBM parameters. These values are estimated based on data from site survey which is often scarce and uncertain. The operators monitor certain parameters like the quantity of muck extracted and earth pressures and adjust these parameters as the excavation progresses to maintain their values within the thresholds set during design to maintain a stable AR and assure safety. Moreover, complex operations, which affect AR, such as maintaining the appropriate foam lance pressure, maintaining required earth pressure at the face, selecting effective controllable parameters, and processing the muck through a depressurizing screw conveyor, are still performed intuitively and thus are not optimized [6,8]. Therefore, accurate models to predict AR given ground conditions based on monitored data from TBM are urgently needed.

There have been several attempts to predict AR. In earlier stages, traditional statistical methods as well as experimental methods were widely adopted for AR prediction. However, the highly complex and non-linear interactions between different TBM parameters cannot be captured by these methods [7,9]. Recently, machine learning algorithms have been implemented to develop AR prediction models for TBMs based on machine monitored data. The results have confirmed that machine learning algorithms are able to capture the complex behavior of TBMs and make accurate AR predictions [10,11].

Several machine learning models have been used to predict AR by utilizing great amount of recorded data by TBM during tunnelling, including artificial neural networks (ANN) [11,12], support vector regression (SVR) [6,13], random forest (RF) [14,15], fuzzy logic [16,17], and classification and regression tress [18,19]. Among them, two models, SVR and RF, became the most widely adopted for AR prediction due to their robustness and high accuracy. For example, Mokhtari and Mooney [6] developed a SVR based model capable of successfully predicting earth pressure balance TBM (EPBM) AR. Zhou [13] used hybrid SVR models to predict AR with high accuracy, aiming at minimizing the financial and scheduling risks for tunneling projects. Yang [14] developed RF models for TBM performance prediction, showing greater accuracy than numerical regression method. In this paper, SVR is used to model EPBM AR. SVR was chosen in this paper since it is a well-established and accepted machine learning technique, showing promising results.

Based on recent research [5,7,20,21], most of the features used in previous studies on AR prediction are determined by researchers' experience rather than

systematically, sometimes failing to consider the most influential and efficient parameters. Simply selecting parameters based on previous studies, sometimes referring to tunnels under different conditions, is not rigorous and may result in loss of accuracy of the model. Addressing this need to systematically select the adequate model features, we developed an hybrid automatic feature selection method which combines constant and quasi-constant detection method (CDQ) and sequential forward selection (SFS) [22]. The developed method allows for a better selection of the model input features, leading to a more accurate model and prevents redundancy, thus increasing computational speed.

The datasets used for the modeling were collected from the 3.95 km long S line from the light metro project in the city of Porto, Portugal. The excavation method adopted in this project was a EPBM which is rarely a target of research in AR prediction. To identify and select the best model features, a new hybrid automatic feature selection method was developed. CDQ was employed first to filter out parameters with constant and quasi-constant values. Then, best feature subsets describing EPBM AR were determined by SFS. Stratified split from verstack package was introduced to ensure training and testing data represent entire range of AR. To find out the optimum values of the parameters and prevent over-fitting, cross-validation and grid search were performed during the process of model development. The model performance was evaluated through R^2 and RMSE. Another SVR model was also developed through the features selected based on previous research (i.e. selecting the typical features used in AR models) and compared with SVR model developed using hybrid automatic features selection method to highlight the importance of feature selection.

2 Data Description

2.1 Projects Description

The monitored data are collected from the Porto light metro project in Porto, Portugal. The tunnel used to demonstrate the methodology is 3.95 km long and it runs between the Salgueiros and Sao Bento stations, as shown in Fig. 1.

The excavation method adopted was an EPBM, capable of both closed and open mode excavation in mixed face conditions. A schematic of the EPBM used in Porto is shown in Fig. 2.

2.2 Data Preparation

The EPBM used in the Porto metro tunnel of Line S contained numerous sensors which recorded a total of 195 features every 10 s, including data from both excavation and halt (e.g. segment assembly) phases of the tunnel construction. Since the main focus of our work was to predict AR, only the data that corresponded to the excavation phase was included in the learning of the models, and all data corresponding to the halt phase was excluded. For simplicity, all parameters and AR during one ring excavation (i.e. a length of 1.4 m) was averaged.

Fig. 1. Project Location

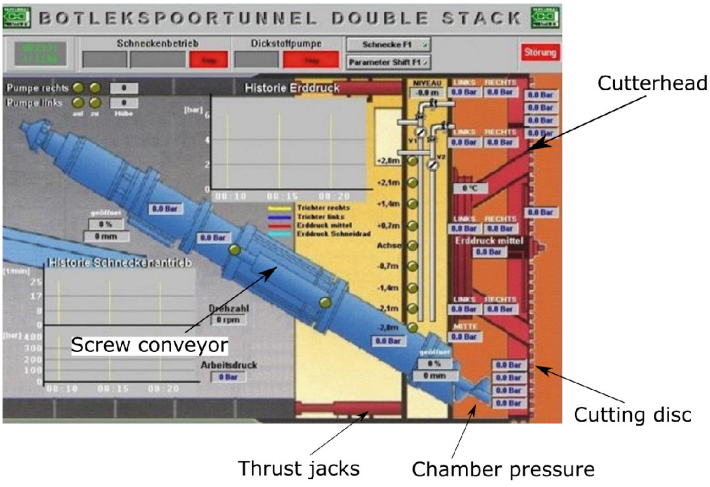


Fig. 2. TBM structure

3 Feature Selection and Analysis

The high dimensionality of the data collected by the EPBM sensors may lead to a poor performance of the machine learning model. More data leads to a robust model, but this is true for the number of instances and not for the number of features. For this reason, dimensionality reduction is needed to exclude redundant and irrelevant parameters that provide little information but affect the efficiency

of the model. In this section, a hybrid feature selection method is presented by the combination of CDQ and SFS.

3.1 Constant and Quasi-constant Detection Method

The monitored data involve two types of irrelevant parameters: constant parameters and quasi-constant parameters. The constant parameters which are the ones with constant values, such as feed line bentonite, were eliminated as they had no influence in the predicting AR. The second type of parameters, the quasi-constant, have only limited impact on the target variable, as they contain insufficient information, and including them in the model may even lead the model to learn from the fringe cases and cause overfitting. Thus, these were also eliminated.

One common way to perform the elimination is to measure its variance by low variance filter method. However, sometimes useful parameters may be eliminated via this method, for example, cutting wheel speed of rotation, a potentially useful variable, has small values and thus low variance. This issue can be avoided by adopting the constant and quasi-constant detection method (CDQ) from fast_ml package which can detect and remove constant and quasi-constant parameters while keeping important parameters. CDQ excludes quasi-constant parameters by setting up cut-off value of percentage of constant values. In this case, thirty (30) constant parameters and thirty-eight (38) quasi-constant parameters were detected by setting the threshold of constant parameter percentage to 50%. After the application of the CDQ method to the datasets, a total of 127 parameters remained.

3.2 Sequential Forward Selection Method

Feature selection is the process of selecting the most relevant and non-redundant features to use in model development. As a first step, we used the constant and quasi-constant detection to reduce the number of features from 195 to 130. To further reduce dimensionality (i.e. the number of input features), Sequential forward selection (SFS), a more intelligent model-performance based method, was implemented.

SFS is a part of stepwise algorithm [23] by which a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion. With SFS, one feature is added in each iteration until no further improvement of model performance is possible, resulting in a optimized feature subset, as shown in Fig. 3. SFS first runs the model using each feature and determines the most relevant (i.e. the variable which results in best model performance). In the case of the example in Fig. 3, Feature 3 is selected. Then the algorithm pairs Feature 3 with the remaining features, one at a time, and determines which pair performs the best for the model. In this example, Feature 3 and Feature 2 are confirmed. This process keeps going, another feature is added to the previous "best" feature set until the desired number of features is reached.

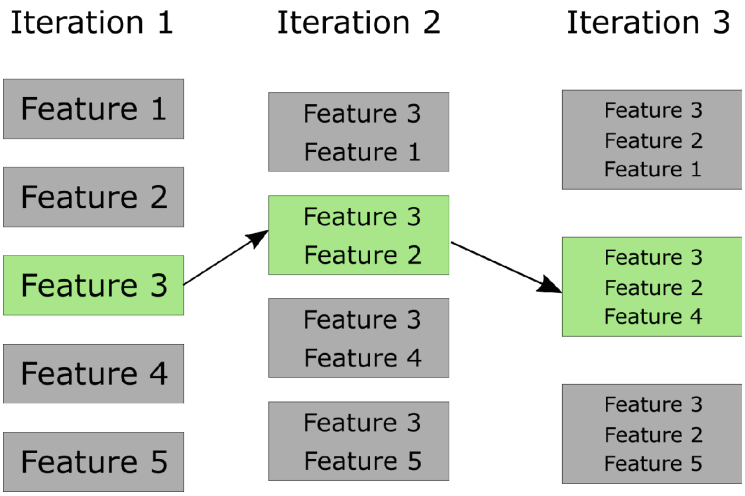


Fig. 3. Sequential forward selection

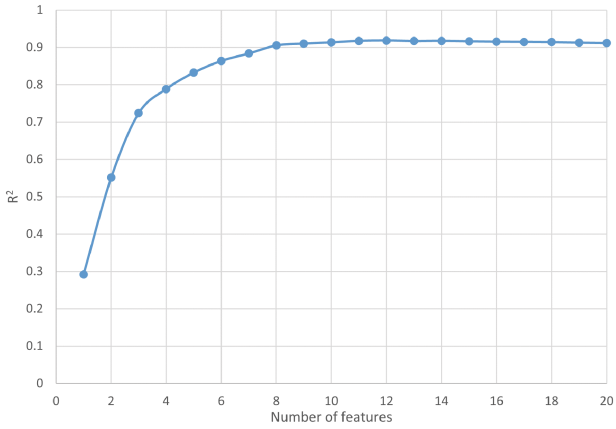


Fig. 4. Sequential forward selection results

3.3 Results of Feature Selection

In our analysis, the desired number of features was set to 20 for high computational efficiency. The relationship between model performance and the number of features is shown in Fig. 4. The SVR model performance increases first with the increasing number of features and peaks when the number is 12, then decreases slightly after that. When the number of features is beyond 8, there is a steady trend of model performance. It is worth noting that R^2 during that range are larger than 0.90, indicating a good performance of SVR models.

Table 1. Selection order through sequential forward selection

Selection order	Features
1	Polymer or Bentonite injected q.ty
2	Water injected quantity
3	Foam lance 2 liquid flow
4	N° of strokes tailskin grease back 8.00 oclock
5	Grout injection group A3 pressure
6	Liquid pump water flow
7	Thrust pressure ZYL.6 group C
8	Stroke measuring shield articulation cylinder 02
9	Pressure grease chamber front 11.00 oclock
10	Pressure force cutting wheel

To reduce computational burden and ensure that all relevant parameters are included in the analysis, the 10 first "best" features are chosen to develop the machine learning model, as listed in Table 1.

4 Comparison Between Hybrid and Experience-Based Feature Selection

4.1 Methodology

Support vector regression (SVR) is a extension of support vector machine (SVM) which is a well known classification algorithm. It has been widely adopted for AR prediction due to its high efficiency and accuracy [5,6]. In this part, two SVR models we developed for comparison. The first one was developed using the input features selected by proposed feature selection method while the second one used input features selected based on authors' experience and previous research (without feature selection), as summarized in Table 2. The first 8 features are also averaged during one ring excavation as before. It should be noted that earth pressure and foam lance pressure are measured by several sensors at different locations of the cutterhead (e.g. 7 sensors for earth pressure and 4 sensors for foam lance pressure). In this work, the average value of all earth pressure and the average value of all foam lance pressure sensor data was used for model development.

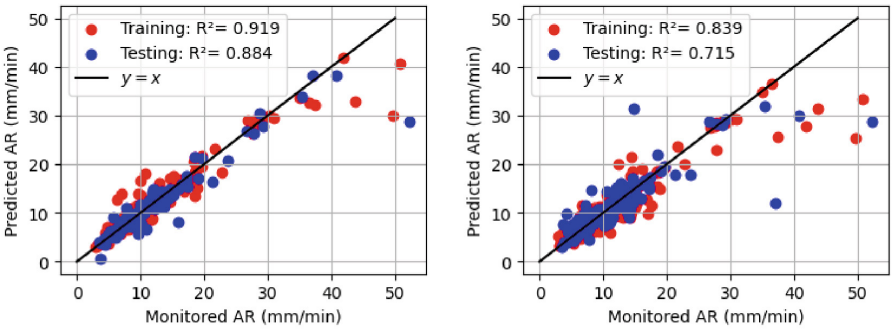
The split of training and testing data sets is a key factor that can affect the prediction performance of SVR models. If entire range of AR is not represented in both training and testing datasets, correct relationship between features and AR is not guaranteed and bias may be introduced. In such situations, predictions made by SVR models may not be accurate when it comes to the range which is not represented in the training dataset. In this paper, stratified split is introduced to ensure that the training and testing datasets are representative of the entire

Table 2. Features selected based on experience

Feature number	Features
1	Torque cutting wheel
2	Pressure force cutting wheel
3	Thrust force
4	Torque screw
5	Cutting wheel speed of rotation
6	Cutting wheel high pressure
7	Thrust pressure
8	excavated material flow
9	Earth pressure
10	Pressure foam lance

Table 3. SVR Model performance with and without feature selection

Model	Training dataset (80%)		Testing dataset (20%)	
	R^2	RMSE (mm/min)	R^2	RMSE (mm/min)
With feature selection	0.919	2.794	0.884	4.651
Without feature selection	0.839	5.558	0.715	11.409



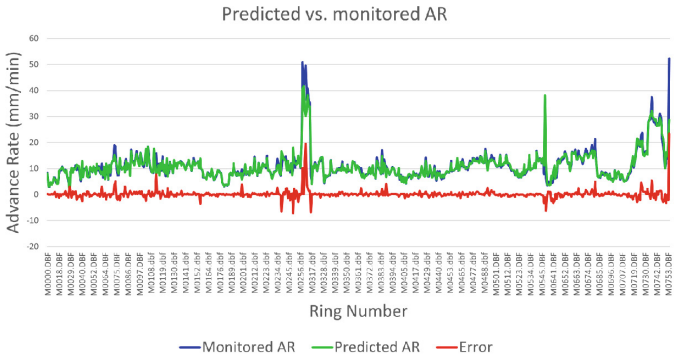
(a) With feature selection (b) Without feature selection

Fig. 5. Performance of SVR models

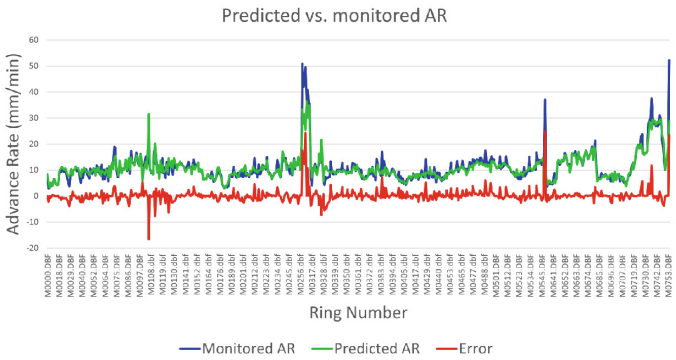
range of AR. As such, 80% of dataset was assigned for training and 20% for testing the model. For training dataset, 20% out of 80% is used for validation in the cross validation process to determine the optimum hyperparameters. Finally, both models were evaluated using R^2 and root mean squared error (RMSE) and compared. The results are presented in the next section.

4.2 Results of Support Vector Regression

The performance of the models is summarized in Table 3. In both models, the prediction accuracy in the training datasets is higher than in the testing datasets, verifying the effectiveness of the developed model in predicting AR. The model with feature selection explains 92% of AR variation (i.e. the R^2 value is 0.92), leaving only 8% of AR variation unexplained. Besides, this model performs well also in the prediction, with a R^2 value of 0.88. The model without feature selection performs worse than the model developed using feature selection. Only 84% of the AR variation is explained by the model (training data) and only 72% of the AR variation can be predicted (testing data). The performance of the model with feature selection has an increase in R^2 of 8% and 17% in training and testing, respectively. Model errors, expressed in RMSE, are low in the model with feature selection for both training and testing. In contrast, RMSE is much higher in the model without feature selection, showing an increase of 98.9% and



(a) With feature selection



(b) Without feature selection

Fig. 6. Predicted AR, monitored AR, and error over rings

145.3% for training and testing, respectively, when compared with the RMSE of the model with feature selection.

Figure 5 and Fig. 6 show the output of each model versus the monitored sensor data, where line y = x means perfect prediction. It is clear that the performance of the model with feature selection is better than the one without feature selection. The difference in performance of the two models is further confirmed by the results in Fig. 6a, where one can clearly observe the error curve of the model with feature selection is quite stable and mostly around 0. Whereas in Fig. 6b, which shows the results of the model without feature selection, the discrepancy between predicted and monitored data is larger and error curve less stable. Accordingly, the developed model using features selected by proposed method is able to better capture EPBM performance, revealing and confirming the importance of feature selection in building a robust prediction model. Besides, it should be noted that for both models, large errors often occur beyond 20 mm/min and generally concentrate in the range beyond 35mm/min, which is worthy for further research.

5 Conclusions

In this paper, we proposed a hybrid automatic feature selection method to select features and developed support vector regression (SVR) models for EPBM advance rate prediction. Constant and quasi-constant detection method (CDQ) is used to filter out parameters with constant and quasi-constant values and sequential forward selection (SFS) is then used to select the best feature subset for developing the SVR model. The model was trained and tested using the 10 first "best" features as input. Based on the model results, the following conclusion can be drawn:

(1) Hybrid automatic feature selection method is developed by combining CDQ and SFS, which is able to reduce dimensionality and consequently computational burden while improving model accuracy.
(2) The developed SVR model using automatically selected features can effectively predict AR, better than the one developed without feature selection. The prediction accuracy of the model with feature selection was 92% and 88% for training and testing, respectively. When compared with the performance of model without feature selection, the accuracy of the model with feature selection improved by 8% and 17% for training and testing, respectively.

References

1. Sousa, R.L., Einstein, H.H.: Risk analysis during tunnel construction using Bayesian Networks: Porto Metro case study. Tunnelling Underground Space Technol. 27(1), 86–100 (2012). https://doi.org/10.1016/j.tust.2011.07.003

2. Huang, S., Chen, Z., Xie, Y., Lin, Z.: A variational approach to the analysis of excavation-induced vertical deformation in a segmental tunnel. Tunnelling Underground Space Technol. 122, 104342 (2022). https://doi.org/10.1016/j.tust.2021.104342
3. Huang, S., Chen, Z., Jiang, T., Xie, Y., Lin, Z., Deng, Y.: Basement excavation influences zones for deformation of the adjacent side tunnel. Arabian J. Geosci. 15(11), 1066 (2022). https://doi.org/10.1007/s12517-022-10214-2
4. Sousa, R.L., Einstein, H.H.: Lessons from accidents during tunnel construction. Tunnelling Underground Space Technol. 113, 103916 (2021). https://doi.org/10.1016/j.tust.2021.103916
5. Pan, Y., Fu, X., Zhang, L.: Data-driven multi-output prediction for TBM performance during tunnel excavation: An attention-based graph convolutional network approach. Autom. Constr. 141, 104386 (2022). https://doi.org/10.1016/j.autcon.2022.104386
6. Mokhtari, S., Mooney, M.A.: Predicting EPBM advance rate performance using support vector regression modeling. Tunnelling Underground Space Technol. 104, 103520 (2020). https://doi.org/10.1016/j.tust.2020.103520
7. Sheil, B.B., Suryasentana, S.K., Mooney, M.A., Zhu, H.: Machine learning to inform tunnelling operations: Recent advances and future trends. Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction 173(4), 74–95 (2020). https://doi.org/10.1680/jsmic.20.00011
8. Mokhtari, S., Navidi, W., Mooney, M.: White-box regression (elastic net) modeling of earth pressure balance shield machine advance rate. Autom. Constr. 115, 103208 (2020). https://doi.org/10.1016/j.autcon.2020.103208
9. Zhang, W., Zhang, R., Wu, C., Goh, A.T.C., Lacasse, S., Liu, Z., Liu, H.: State-of-the-art review of soft computing applications in underground excavations. Geosci. Front. 11(4), 1095–1106 (2020). https://doi.org/10.1016/j.gsf.2019.12.003
10. Mahdevari, S., Shahriar, K., Yagiz, S., Akbarpour Shirazi, M.: A support vector regression model for predicting tunnel boring machine penetration rates. Int. J. Rock Mech. Min. Sci. 72, 214–229 (2014). https://doi.org/10.1016/j.ijrmms.2014.09.012
11. Koopialipoor, M., Fahimifar, A., Ghaleini, E.N., Momenzadeh, M., Armaghani, D.J.: Development of a new hybrid ANN for solving a geotechnical problem related to tunnel boring machine performance. Engineering with Computers 36(1), 345–357 (2020). https://doi.org/10.1007/s00366-019-00701-8
12. Salimi, A., Rostami, J., Moormann, C., Delisio, A.: Application of non-linear regression analysis and artificial intelligence algorithms for performance prediction of hard rock TBMs. Tunnelling Underground Space Technol. 58, 236–246 (2016). https://doi.org/10.1016/j.tust.2016.05.009
13. Zhou, J., Qiu, Y., Zhu, S., Armaghani, D.J., Li, C., Nguyen, H., Yagiz, S.: Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate. Eng. Appl. Artif. Intell. 97, 104015 (2021). https://doi.org/10.1016/j.engappai.2020.104015
14. Yang, J., Yagiz, S., Liu, Y.J., Laouafa, F.: Comprehensive evaluation of machine learning algorithms applied to TBM performance prediction. Underground Space 7(1), 37–49 (2022). https://doi.org/10.1016/j.undsp.2021.04.003
15. Tao, H., Jingcheng, W., Langwen, Z.: Prediction of hard rock TBM penetration rate using random forests. In: The 27th Chinese Control and Decision Conference (2015 CCDC), pp. 3716–3720 (2015). https://doi.org/10.1109/CCDC.2015.7162572

16. Ghasemi, E., Yagiz, S., Ataei, M.: Predicting penetration rate of hard rock tunnel boring machine using fuzzy logic. Bull. Eng. Geol. Environ. 73(1), 23–35 (2014). https://doi.org/10.1007/s10064-013-0497-0
17. Minh, V.T., Katushin, D., Antonov, M., Veinthal, R.: Regression Models and Fuzzy Logic Prediction of TBM Penetration Rate. Open Eng. 7(1), 60–68 (2017). https://doi.org/10.1515/eng-2017-0012
18. Xu, H., Zhou, J., G. Asteris, P., Jahed Armaghani, D., Tahir, M.M.: Supervised Machine Learning Techniques to the Prediction of Tunnel Boring Machine Penetration Rate. SN Appl. Sci. 9(18), 3715 (2019). https://doi.org/10.3390/app9183715
19. Salimi, A., Rostami, J., Moormann, C.: Application of rock mass classification systems for performance estimation of rock TBMs using regression tree and artificial intelligence algorithms. Tunnelling Underground Space Technol. 92, 103046 (2019). https://doi.org/10.1016/j.tust.2019.103046
20. Gao, X., Shi, M., Song, X., Zhang, C., Zhang, H.: Recurrent neural networks for real-time prediction of TBM operating parameters. Autom. Constr. 98, 225–235 (2019). https://doi.org/10.1016/j.autcon.2018.11.013
21. Wang, Y., Gao, X., Jiang, P., Guo, X., Wang, R., Guan, Z., Chen, L., Xu, C.: An extreme gradient boosting technique to estimate TBM penetration rate and prediction platform. Bull. Eng. Geol. Environ. 81(1), 58 (2022). https://doi.org/10.1007/s10064-021-02527-5
22. Čehovin, Luka, and Zoran Bosnić. "Empirical evaluation of feature selection methods in classification." Intelligent data analysis 14.3 (2010): 265-281. https://doi.org/10.3233/IDA-2010-0421
23. Draper, Norman R., and Harry Smith. Applied regression analysis. Vol. 326. John Wiley & Sons, 1998.