Proceedings of ISFOG 2025
5TH INTERNATIONAL SYMPOSIUM ON
FRONTIERS IN OFFSHORE GEOTECHNICS
Nantes, France | June 9-13 2025

ISBN 978-2-85782-758-0

Application of Machine Learning to Predict Pile Driveability Performance in Offshore Southeast Asia Soils

N. N. Huang
PETRONAS, Kuala Lumpur, Malaysia

M. J. Rohani, A. A. Rahman, N. Yusoff, Z. A. M. Ali

PETRONAS, Kuala Lumpur, Malaysia

N. A. Osman, E. A. Patah Akhir

Universiti Teknologi PETRONAS (UTP), Perak Darul Ridzuan, Malaysia

noorizal.nasrih@petronas.com

ABSTRACT: Machine learning (ML) based modelling methods enable geotechnical engineers to leverage state-of-the-art tools to create predictive models to be applied in various complex geotechnical engineering problems. This paper investigates the ML techniques and algorithms that can be adopted to develop suitable ML models with actual pile installation and cone penetration test (CPT) data to predict blowcounts and hammer energy. For this study, ML models from scikit-learn (sklearn) libraries in Python such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Linear Regression (LR), and Polynomial Regression (PR) have been considered based on their ability to tackle regression problems. It is necessary to calculate the soil resistance during driving (SRD) using the CPT dataset to have a better ML model since offshore piles comes in many sizes. Then, SRD and actual pile installation datasets from sites around offshore Southeast Asia were resampled to a regular grid of 0.25m intervals to facilitate data handling prior establishing a ML model. Outcome from the ML models were interpreted in form of R-Square and Root Mean Square Error (RMSE). Two ML models were generated, a model to predict the blowcounts and a model to predict the hammer energy used based on the actual pile installation and CPTU datasets provided. Based on the five ML algorithms, RF gave the highest R-Square value for predicted blowcounts ML model with a value of 0.801 followed by DT, PR, LR and SVM. As for the hammer energy ML model, RF again showed better results with R-Square value of 0.911 followed by PR, DT, LR and SVM. Soil variability at each location around Offshore Southeast Asia dictates the result of obtaining a good ML model. Based on this assessment, Random Forest ML model has consistently appeared to be the best ML model to predict the blowcounts and energy required to install offshore piles to its design depth thus minimizing potential issues such as pile refusal.

Keywords: Machine learning; CPT; Predicted blowcounts; Hammer energy; Regression

1 BACKGROUND

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Supervised machine learning algorithms apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors to modify the model accordingly.

Recent advances in offshore foundations installation and operation monitoring, data transfer and storage and computational resources have led to an acceleration of the use of data-driven methods in geotechnical applications (Stuyts, 2020). While geotechnical engineers have already started to adopt these methods, there are still several challenges to overcome to allow routine use of machine learning. Buckley et al. investigate the application of a Bayes-

ian Optimization framework to enhance the prediction of Soil Resistance during Driving (SRD). This improved prediction model is subsequently utilized to forecast pile driveability, providing a more accurate and reliable method for assessing pile installation performance. This paper focuses on the application of machine learning in predicting pile driveability performance using cone penetration testing (CPTU) data.

2 MACHINE LEARNING MODELS

Five (5) models were considered to train and test the given dataset. These models can be found in scikit-learn, an open-source machine learning library for Python.

2.1 Linear regression (LR)

Linear regression is a statistical method used for predictive modelling. It establishes a linear relationship between one or more independent variables and a dependent variable. By fitting a line to the observed data points, it enables the prediction of the dependent variable based on the values of the independent variables. The model aims to minimize the difference between the observed and predicted values, making it a valuable tool for making forecasts, understanding relationships between variables, and identifying patterns in data.

2.2 Polynomial regression degree 2 (PR)

Polynomial regression of degree 2 is a predictive model that extends linear regression by allowing for a curved relationship between the independent and dependent variables. Instead of fitting a straight line, it fits a quadratic curve to the data. This model captures more complex patterns in the data and can provide a better fit than linear regression when the relationship between the variables is non-linear. By including squared terms of the independent variable(s), it allows for both upward and downward curves in the prediction. Polynomial regression of degree 2 is a simple yet powerful tool for capturing quadratic relationships in data and making predictions based on them.

2.3 Random forest (RF)

Random forest is a powerful predictive modeling technique that belongs to the ensemble learning family. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. Each tree in the forest is grown using a subset of the training data and a random subset of features. This randomness helps to decorrelate the trees, making the model robust to overfitting and highly accurate in making predictions. Random forest can handle large datasets with high dimensionality and is resistant to outliers and noise. It is widely used across various domains for tasks like classification, regression, and feature selection, owing to its simplicity, flexibility, and excellent predictive performance.

2.4 Support vector machine (SVM)

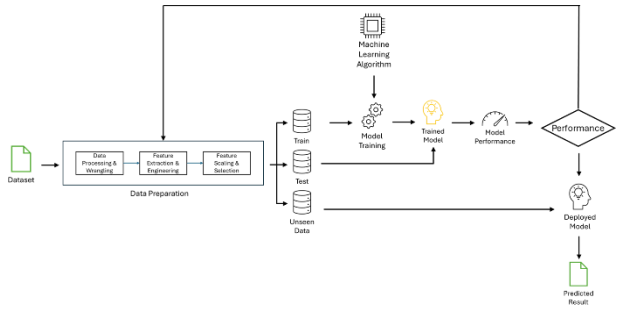
SVM is a powerful predictive modeling algorithm used for both classification and regression tasks. It works by finding the optimal hyperplane that best separates the data points into different classes or predicts continuous values. The "support vectors" are the data points closest to the hyperplane, which determine its position and orientation. SVM aims to maximize the margin between classes, making it robust to outliers and capable of handling high-dimensional data effectively. It can also utilize the kernel trick to transform the input space into a higher-dimensional space, enabling it to capture non-linear relationships between variables. SVM is widely used in various fields due to its versatility, effectiveness, and ability to generalize well to unseen data.

2.5 Decision tree (DT)

A decision tree predictive model is a versatile algorithm used for both classification and regression tasks. It works by recursively splitting the dataset into subsets based on the most significant attribute at each node, forming a tree-like structure of decisions. Each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or prediction. Decision trees are easy to interpret and visualize, making them useful for understanding the relationship between variables in the data. However, they are prone to overfitting, especially with complex trees, which can be mitigated by techniques like pruning. Decision trees are widely used in various fields due to their simplicity, interpretability, and ability to handle both numerical and categorical data.

3 MACHINE LEARNING MODEL DEVELOPMENT PROCESS

The machine learning (ML) model development process is a structured approach to building, training, and deploying models that can learn from data and make predictions or decisions. Figure 1 depicted the



system architecture for machine learning model development process.

Figure 1. System architecture for machine learning model development process

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Supervised machine learning algorithms apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system can provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors to modify the model accordingly. For this project, FIVE (5) models were used to train and test the given dataset as described further in section 4.3.

The overall machine learning development process was illustrated in Figure 2. Data normalization is a common technique employed in data preprocessing for machine learning. Its objective is to standardize the numeric values across columns in a dataset to a common scale, preserving differences in value ranges and information content. This transformation adjusts the mean of the data to zero and the standard deviation to one.

Data splitting, specifically the train-test split, is a methodology used to assess the efficacy of a machine learning algorithm. It is applicable to both classification and regression tasks and is compatible with various supervised learning algorithms. This procedure entails partitioning a dataset into two subsets: the training dataset, utilized for model fitting, and the test dataset, employed for validating model performance. A random split ratio of 80% to 20% is adopted for training and test data, respectively (Toleva, 2021 and Yun Xu et al, 2018).

Model training involves feeding a machine learning algorithm, also known as the learning algorithm, with training data to facilitate learning. The resulting model artifact, referred to as the ML model, captures patterns present in the training data that map input attributes to the target variable, also known as the target attribute. These patterns are utilized to make predictions on unseen data instances.

Model testing and validation are essential processes in machine learning aimed at assessing a model's ability to generalize beyond the training data. To achieve this, it is imperative to evaluate the model's performance on unseen data. Therefore, model validation involves estimating the generalization quality using data not utilized during training. Evaluation metrics such as R-Square, MAPE, RMSE, and MAE are commonly employed to gauge model performance, as detailed in Table 1.

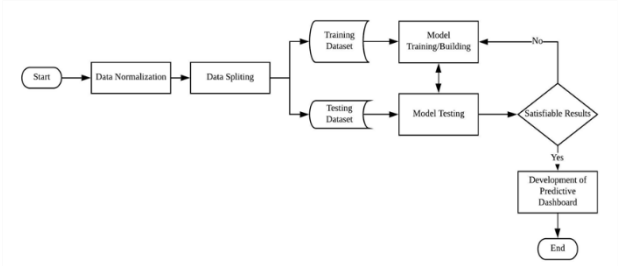


Figure 2. Machine learning development process

Table 1. Performance measurement metrics.

Metric	Description	Range
R-Square	The coefficient of determination quantifies the proportion of variability in one variable that is accounted for by another variable. A high R-Square signifies a robust positive linear correlation, meaning that as one variable increases, the other variable also tends to increase—a pattern indicative of a well-performing model.	> 0 to 100%
RMSE	Represents the standard deviation of the residuals, which measure the deviations between observed and predicted values. RMSE provides insight into the dispersion of these residuals, indicating how widely they are distributed.	Depends on the range of actual values. The lower the value, the better. The RMSE and MAE range has been widely discussed by Chicco et. al
MAE	Quantifies the average size of errors in a	

	prediction set, regardless of their direction.	2021 and Vujovic et. al. 2021.

4 RESULTS AND ANALYSES

This section provides a comprehensive overview of the results and analyses derived from the predictive models developed. The analysis is structured into three major steps: **1) Data Understanding, 2) Data Cleansing, Training & Output, 3) Predictive Model Development, and 4) Testing Dataset and Model Evaluation.**

4.1 Data Understanding

Data understanding is a crucial step that guides all subsequent processes in machine learning, from preprocessing to model evaluation, ensuring that the insights derived are reliable and meaningful. plays a vital role in machine learning for several reasons, as it forms the foundation of model development and successful outcomes. By understanding the nature of your data, including its distribution, outliers, and relationships among features, you can choose the most suitable machine learning algorithms. For example, linear models may work well for data with linear relationships, while non-linear models like decision trees or neural networks might be better for more complex data structures.

A typical CPTU dataset includes four parameters: testing penetration depth (z) [m], tip resistance (q_c) [MPa], sleeve friction (f_s) [MPa], and pore pressure (u_2) [MPa]. For pile driving data, the main parameters considered are pile penetration [m], pile diameter [m], blowcounts [Blows/0.25m], and hammer energy [kJ].

For data preparation, both feature extraction and feature engineering were performed. Feature extraction involves converting raw data into a set of features usable by machine learning algorithms, while feature engineering is the process of creating new features or modifying existing ones to enhance the performance of these models. During the feature engineering process, a new parameter called SRD was introduced. Using a predefined calculation, the q_c [MPa] values were combined with f_s [MPa] and pile diameter [m] to generate the SRD [MN] parameter.

4.1.1 Determination of SRD based on CPTU data

SRD is expressed using the following equation:

$$SRD = f_s \cdot a_{outer} + f_s \cdot a_{inner} + q_c \cdot a_{tip} \quad (1)$$

Where f_s is the sleeve friction, q_c is the end bearing and a = area

4.2 Data Cleansing, Training and Output

The data cleansing process was the initial step in preparing the dataset for training the predictive models. This ensured that the data used was accurate, consistent, and suitable for modelling. The following steps consists of data collection, handling missing data, outlier detection and treatment, normalisation and standardisation, feature selection and data splitting.

4.2.1 Data collection

The dataset utilized in was sourced from pile installation databases and soil investigation from sites around offshore Southeast Asia that monitored various variables such as pile driving blowcounts, hammer energy and CPTU data.

4.2.2 Handling missing data

Handling missing data is a common challenge in machine learning, and there are several techniques to address this issue, depending on the type and amount of missing data. For this study, a few methods were performed which is deletion, imputation and K-Nearest Neighbour (K-NN). Few sets of experiments were conducted using these methods and the results were compared. Among the methods tested, deletion seems to be the best method for handling inconsistencies and missing data for CPTU and actual data. We found that using K-NN imputation data, the results dropped due to large scale injection of synthetic data. The lack of realism and accuracy is perhaps the biggest limitation of synthetic data. Hence, deletion of the missing data and using only actual data provided was the best method to perform the model prediction

4.2.3 Normalisation and standardisation

To ensure that all input features contributed equally to the model training process, normalization (scaling features to a range) and standardization (scaling features to have a mean of zero and a standard deviation of one) were applied where necessary. CPTU data were resampled to a regular grid of 0.25m intervals to facilitate data handling. This step was crucial for algorithms sensitive to feature scales, such as Support Vector Machine (SVM)

4.2.4 Data splitting

The cleansed dataset was split into training and validation subsets. 80% of the data was allocated for

training, and 20% was reserved for validation. This split ensured that the models could be evaluated on unseen data to assess their generalization capabilities.

4.3 Training the predictive models

With the data cleansed and prepared, the next step involved training the predictive models. Separate models were developed for predicting blowcounts and hammer energy using the same set of input features. The training process consists of the following:

- Random Forest (RF)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- Linear Regression (LR)
- Polynomial Regression (PR)

The training process consisted of the following sub-steps:

- Model Initialization: Each algorithm was initialized with default parameters. Where necessary, hyperparameters were tuned to optimize performance. For example, the number of trees in the Random Forest or the kernel type in SVM.
- Model Training: The models were trained using the training subset of the data. This involved fitting the models to the input features to learn the underlying patterns that predict the outputs (blowcounts and hammer energy).
- Model Tuning and Optimization: Hyperparameter tuning was conducted using techniques such as Grid Search or Random Search in combination with cross-validation to find the optimal set of parameters for each model.
  For Random Forest, parameters like the number of trees (n-estimators), maximum depth (max_depth), and minimum samples per leaf (min_samples_leaf) were optimized.
  For SVM, parameters such as the kernel type (kernel), regularization parameter (C), and gamma (gamma) were fine-tuned to improve model performance.
- Model Validation: Cross-validation was performed to ensure that the models did not overfit to the training data and could generalize well to new, unseen data. Typically, a k-fold cross-validation approach was adopted.
- Separate Model Training for blowcounts and hammer energy: Although the training process was identical for both outputs, models were trained separately to account for any differences in data distribution and feature importance specific to each prediction task.

4.4 Prediction output

After training, the models were used to generate predictions for both blowcounts and hammer energy. This step involved evaluating the models' performance on the testing dataset and comparing the predicted values against the actual values. The prediction outputs and evaluation metrics were documented as follows:

- Generating predictions: For each trained model, predictions were made on the testing dataset. Separate predictions were generated for blowcounts and hammer energy.
- Comparison of predicted and actual values: The predicted values from each model were compared against the actual values in the testing dataset to assess accuracy and reliability.
- Performance measurement metrics calculation

These metrics were calculated for both blowcounts and hammer energy predictions to evaluate model performance comprehensively as shown in Table 2 and 3. Figure 3 shows the predicition of blowcounts and hammer energy from test and train data for RF model.

Table 2. Blowcounts prediction performance measurement metrics.

ML Model	R-Square	RMSE
RF	0.801	7.811
DT	0.690	9.743
SVM	0.188	15.776
LR	0.270	14.956
PR	0.355	14.068

Table 3. Hammer energy prediction performance measurement metrics.

ML Model	R-Square	RMSE
RF	0.911	25.233
DT	0.878	29.534
SVM	0.584	54.581
LR	0.800	37.838
PR	0.818	36.047

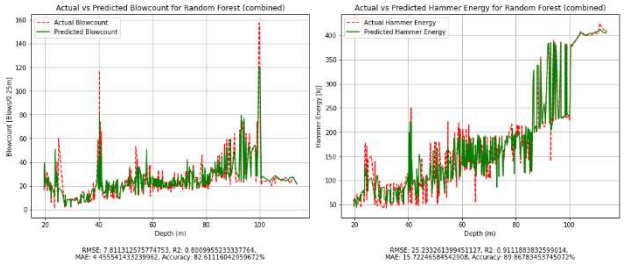


Figure 3. Prediction of blowcounts and hammer energy from test and train data for RF model

4.4 Testing dataset

The predictive models were evaluated based on their performance measurement metrics during the training phase. For blowcounts and hammer energy, the Random Forest (RF) model was chosen to generate predictions on the testing datasets using the CPTU data from site A. By applying the machine learning model to the site A datasets, predictions for blowcounts and hammer energy were generated. Figure 4 illustrates these predictions for site A.

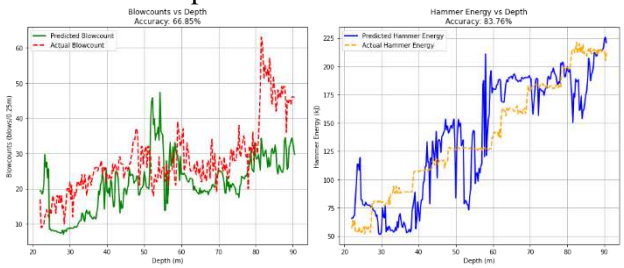


Figure 4. Prediction of blowcounts and hammer energy for site A

5 DISCUSSIONS

From the results, Random Forest (RF) and Decision Tree (DT) outperform other models like Support Vector Machine (SVM), Linear Regression (LR), and Polynomial Regression (PR) in terms of RMSE and R-Square values due to their inherent characteristics and capabilities. Random Forest is an ensemble method that combines multiple decision trees, reducing overfitting and improving generalization. By averaging predictions from many trees, RF achieves lower RMSE and higher R^2 compared to individual models like DT, SVM, or LR. Both RF and DT can automatically assess the importance of features, focusing on the most relevant ones for prediction. This leads to more accurate models with lower errors (RMSE) and better explanatory power (R^2). Moreover, RF and DT are less sensitive to outliers compared to linear models (LR, PR) and SVM, which can be heavily influenced by extreme values. This robustness contributes to better performance metrics. Unlike SVM, which relies on kernel functions and hyperparameters, or linear models that assume a specific relationship between variables, RF and DT adapt flexibly to the data structure, resulting in superior performance

6 CONCLUSIONS

In summary, the Random Forest model outperformed other machine learning models with strong validation results. The R-Square values are 0.801 for the blowcounts prediction model and 0.911 for the hammer energy prediction model. The Root Mean Square Error (RMSE) for blowcounts and hammer energy predictions are 7.811 and 25.233, respectively.

The prediction performance measurement metrics can potentially be improved by training with more datasets of this trend.

The application of machine learning has shown potential in assessing and predicting pile driveability performance, which helps optimize pile installation activities and minimize potential issues such as pile refusal.

AUTHOR CONTRIBUTION STATEMENT

Authors are asked to add an author contribution statement before the Acknowledgement section and following the recommendations of https://credit.niso.org/.

Sample CRediT author statement

First Author: Data curation, Formal Analysis, Writing - Original draft. **Other Authors**.: Conceptualization, Methodology, Supervision. **Additional Authors**: Formal Analysis, Visualization, Investigation. **Last Author**: Formal Analysis , Writing- Reviewing and Editing

ACKNOWLEDGEMENTS

The authors are grateful for the support provided by all parties, both from PETRONAS and UTP.

REFERENCES

Stuyts, B. and Suryasentana, S.K. Applications of data science in offshore geotechnical engineering: state of practice and future perspectives. in Session on 'AI applications' in 9th SUT OSIG conference, 2023.

Buckley, R.M, Chen, Y.M, Sheil, B.B, Suryasentana, S.K, Randolph, M.F and Doherty, J.P. Improving driveability predictions for offshore piles using

Bayesian optimization, in 9th SUT OSIG conference, 2023.

Stuyts, Bruno. Data science applications in geo-intelligence. In 4th International Symposium Frontiers in Offshore Geotechnics, p. 3432. Deep Foundations Institute, 2020.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj computer science, 7, e623.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.

Toleva, B. (2021). The Proportion for Splitting Data into Training and Test Set for the Bootstrap in Classification Problems. Business Systems Research Journal. Volume 12 (2021): Issue 1 (May 2021)

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Springer Nature Link