buildings

MDPI

Article

A Machine Learning-Based Approach for Predicting Installation Torque of Helical Piles from SPT Data

Marcelo Saraiva Peres, José Antonio Schiavon and Dimas Betioli Ribeiro *

Civil Engineering Division, Aeronautics Institute of Technology, Praça Marechal Eduardo Gomes, 50, São José dos Campos 12228-900, SP, Brazil; marcelosrvperes@gmail.com (M.S.P.); schiavon@ita.br (J.A.S.)
* Correspondence: dimasbetioli@gmail.com

Abstract: Helical piles are advantageous alternatives in constructions subjected to high tractions in their foundations, like transmission towers. Installation torque is a key parameter to define installation equipment and the final depth of the helical pile. This work applies machine learning (ML) techniques to predict helical pile installation torque based on information from 707 installation reports, including Standard Penetration Test (SPT) data. It uses this information to build three datasets to train and test eight machine-learning techniques. Decision tree (DT) was the worst technique for comparing performances, and cubist (CUB) was the best. Pile length was the most important variable, while soil type had little relevance for predictions. Predictions become more accurate for torque values greater than 8 kNm. Results show that CUB predictions are within $[0.71, 1.59]$ times the real value with a 95% confidence. Thus, CUB successfully predicted the pile length using SPT data in a case study. One can conclude that the proposed methodology has the potential to aid in the helical pile design and the equipment specification for installation.

Keywords: helical piles; machine learning; installation torque; sandy soil; installation feasibility

Citation: Peres, M.S.; Schiavon, J.A.; Ribeiro, D.B. A Machine Learning-Based Approach for Predicting Installation Torque of Helical Piles from SPT Data. *Buildings* **2024**, *14*, 1326. https://doi.org/10.3390/buildings14051326

Academic Editor: Bingxiang Yuan

Received: 10 April 2024
Revised: 2 May 2024
Accepted: 5 May 2024
Published: 8 May 2024

1. Introduction

Helical piles are steel helical plates welded to a central shaft, usually with a tubular or a square solid cross-section. Usually, the helical plates are equally spaced, with equal or increasing diameters from the tip to the top of the pile. Machinery installs helical piles into the ground at a constant rotation rate associated with a vertical downward (crowd) force applied mainly at the initial depths. The pile penetration is recommended to be conducted at a vertical advance of approximately one helix pitch length per rotational cycle, aiming to minimize soil disturbance [1–4]. A torque indicator is coupled between the hydraulic motor and pile to measure the required installation torque. Thus, the pile capacity can be estimated via torque-to-capacity correlations [5].

The pile axial capacity can be correlated to the torque required to install a helical pile [2]. Therefore, usually, one establishes a minimum final torque as a termination criterion for the pile installation associated with a minimum embedded length [6]. The empirical correlation between torque and uplift capacity (K_t method, defined in Equation (1)) is usually adopted in validating the pile installation depth [7]. The geometry of the pile shaft has a major influence on K_t [5]. For piles with the standard square shaft section produced in the USA, these authors recommend using $K_t = 33\ \mathrm{m}^{-1}$, whereas $K_t = 23\ \mathrm{m}^{-1}$ and $K_t = 9.8\ \mathrm{m}^{-1}$ are recommended for shaft circular sections with diameters of 89 mm and 219 mm, respectively.

$$K_t = \frac{Q_u}{T}. \quad (1)$$

where Q_u is the uplift pile capacity and T is the final torque installation.

Considering the discussion of the influence of the shaft geometry on installation torque, Ref. [6] presents a proposition that includes an effective shaft diameter (d_{eff}) in the K_t correlation (Equation (2)).

$$K_t = \frac{\lambda_k}{{d_{eff}}^{0.92}}. \quad (2)$$

where λ_k is a fitting factor equal to 1433 $mm^{0.92} \times m^{-1}$.

Different theoretical models are available in the literature to correlate installation torque with helical pile uplift capacity in the sands. Though an equilibrium of forces acting both on helices and shaft at the pile screwing, Ref. [2] deduced a theoretical equation comprising eight components of torque derived from the moment mobilized at the shaft, from the moments acting on the helix and the moments caused by friction of the helix surface areas with the soil (upper, bottom, pitch, external perimeter of the plate thickness). Ref. [8] proposed a theoretical relationship between the installation torque and uplift capacity of deep helical piles in sand, validating via centrifuge tests and conducting both in-flight installation and load testing. Ref. [9] proposed a theoretical model that considers the effective stress approach rather than the total stress considered by [2]. This theoretical model enables the prediction of the K_t factor for both tensile and compressive cases. A total of 74 installation records validated the model.

Although one finds torque–uplift capacity correlations in the literature, few approaches that anticipate the installation torque in a design phase are available. The previous estimation of the installation torque is essential to evaluate the feasibility of the foundation to the design depth and to define the pile shaft's characteristics to avoid damage caused by torsional forces.

Some approaches are proposed in the literature to predict torque based on direct correlation with cone penetration testing. Ref. [10] used the theoretical model proposed by Ref. [8] to correlate the resisting portions of the helical pile uplift capacity with parameters from CPT tests and estimate the installation torque. Similarly, Ref. [11] proposed a CPT-based approach to predict installation torque using as background theoretical equations to calculate five components of torque (torque associated with the shaft, pile shaft tip, friction with upper, bottom, and peripheral surfaces of the helix, and the helix cutting into the soil).

Nonetheless, the SPT is the only in situ test for most Brazilian projects. Ref. [12] proposed a semiempirical approach to estimate the final installation torque of multi-helix piles in clayey sand using SPT data. In this approach, torque measurements at the end of the installation of 571 multi-helix piles were used in a series of multivariate linear regressions to fit a semiempirical model based on a theoretical model that correlates installation torque and uplift capacity [8]. However, this approach was built only considering the data concerning the torque at the final installation depth and, therefore, is not devoted to predicting the entire torque profile with depth.

The literature concerning using ML to design helical pile foundations is limited. Ref. [13] used artificial neural networks (ANN) to predict the uplift capacity of helical piles in sand based on 36 small-scale laboratory tests. Ref. [14] employed a similar approach, using gradient-boosting decision trees tunned with particle swarm optimization and applied to a dataset obtained from centrifuge tests. Nonetheless, no study has used ML techniques to predict the installation torque of helical piles.

In this context, the present work uses the same dataset as [12] to propose an approach to predict the torque over the complete installation depth, extending the study of [15]. It subjects the dataset to eight ML techniques to produce models capable of predicting helical pile installation torque from basic preliminary design information, which includes only SPT as an in situ test. The procedure obtains new variables from raw data to improve the model's accuracy. The mean absolute error (MAE), the root of the mean squared error ($RMSE$), and the coefficient of determination (R^2) are the references to compare the performance of the different ML techniques. The variable importance is measured using the two best-performing models, which are also evaluated using the concept of a confidence interval.

2. Materials and Methods

Three basic steps compose the methodology proposed to obtain the final ML models, as follows:

- Pre-processing;
- ML techniques calibration;
- Evaluation of models.

The strategies employed in each of these steps are summarized hereafter.

2.1. Pre-Processing and Database Information

The raw dataset used in this study contains information from the foundation of transmission towers constructed in a 350 km stretch of a power transmission line located between the towns of Paranatinga and Cláudia, Mato Grosso State, Midwest Region of Brazil. Figure 1 illustrates the State. The area presents a very diversified geological context. Sandy sediments, partially covered by a sand–clay layer, predominate the northern portion of this region. In the southern portion, old rocks, and recent sediments are found [16]. The predominant soil types in the study area are latosols, eutrophic podzolic, cambisol, hydromorphic laterite, low humic glei, quartz sands, alluvial soils, litholic soils, and concretion soils. The great variety of soils results from the geomorphological unit's diversity and lithologies found in the study area [17].

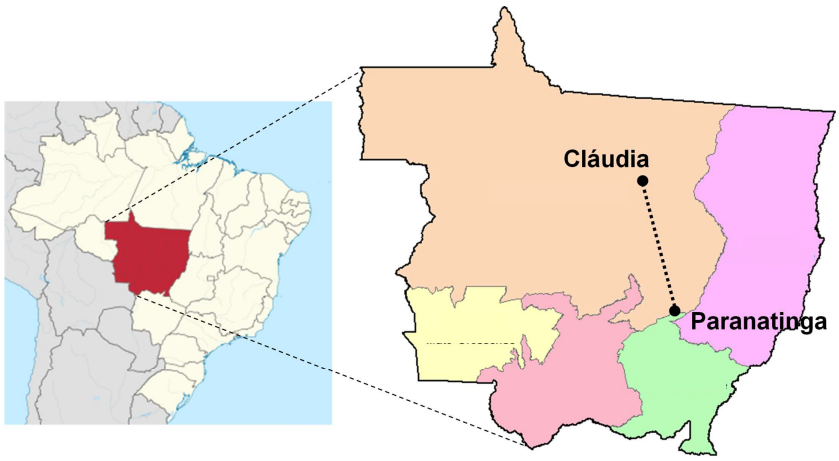

Figure 1. Study area. The transmission towers were constructed between Paranatinga and Cláudia, located in the Midwest region of Brazil. The geological context is diverse, with a predominance of clayey sand.

The information on soil type was provided on the SPT reports first used for the transmission line design. The soil classification was based upon information on particle size (sieve analysis), soil plasticity, color, and soil formation (e.g., sedimentary, residual), with such procedures being established in the Brazilian standard for SPT. Most samples were identified as clayey sand (65.8%), followed by clayey silt (13.8%), sandy silt (9.6%), sandy clay (5.9%), and sand (0.3%). Figure 2 presents the SPT blow counts (from now on referred to as SPT index) with depth. The SPTs were carried out according to the Brazilian Standard ABNT NBR 6484:2001 [18], which describes a testing procedure similar to the ASTM Standard D1586-08 [19]. For the Brazilian standard equipment, the SPT efficiency ranges between 70% and 80%, with 72% the most common value [20]. Therefore, this is the value considered in this work.

The helical piles installed in the study area were comprised of a 101.6 mm diameter steel tubular shaft and six helical plates with diameters of 254 mm, 305 mm, 356 mm, 356 mm, 356 mm, and 356 mm, with increasing diameters from the tip to the top of the pile. The helix spacing was three times the diameter of the smallest helix [15]. The piles were installed using a hydraulic drive head attached to a backhoe loader. The torque

measurement at each meter of pile penetration was performed using a torque indicator coupled between the driving head and the pile top. The pile advance rate was not included in the installation reports. In addition to soil type and SPT information, the raw dataset used in this study contains 707 installation torque measurements, pile installation angle, and pile final length.

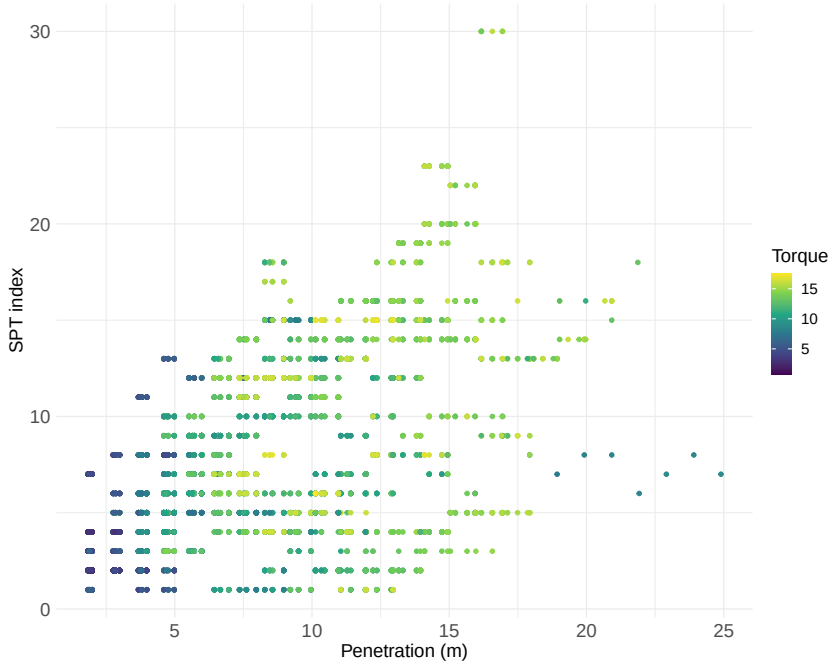

Figure 2. SPT index values with depth, with darker dots representing higher torque. Procedures follow Brazilian standards, and the most frequent hammer efficiency was 72%.

The first dataset assembled, which totaled 9355 samples, was pre-processed to improve the performance of the ML techniques. Each sample corresponds to a torque value for a specific pile number (identification) with one particular penetrated length, which can be associated with a soil type and SPT index. This dataset was cleaned from outliers in previous work. Data was normalized to the $[0, 1]$ interval.

The first concern was cleaning data to eliminate errors like missing values and inconsistencies, a procedure that reduced the original dataset to 7632 samples. Next, variables were chosen as inputs for predicting the installation torque. P refers to the pile tip depth corresponding to a given measured torque, as presented in Figure 3, and $NSPT_{tip}$ represents the SPT index at the pile tip level. The so-called Initial Dataset comprised only these two inputs plus the measured torque. Figure 3 presents other proposed inputs: $N1$–$N4$, representing the initial four m-depth SPT index after eliminating the first meter. It is well known in Brazilian practice that the pile penetration causes a gap in the pile shaft surrounding soil to at least 1 m depth. Therefore, the soil contribution on torque at 1 m depth is disregarded.

The soil type was also included using six predefined classes, represented by the binary variables $S1$–$S6$. This approach enables these qualitative variables to become equally spaced in the input space. Table 1 lists these soil types and their encoding. The mean SPT index for each depth of helical plates was also included as the input $NSPT_{helix}$. The variable N_{shaft} is the summation of all SPT indexes measured above the helical plates, representing the shaft resistance. The Initial Dataset combined with these new inputs constitutes the so-called Complete Dataset, with 14 inputs. All these variables and the output torque are summarized in Table 2.

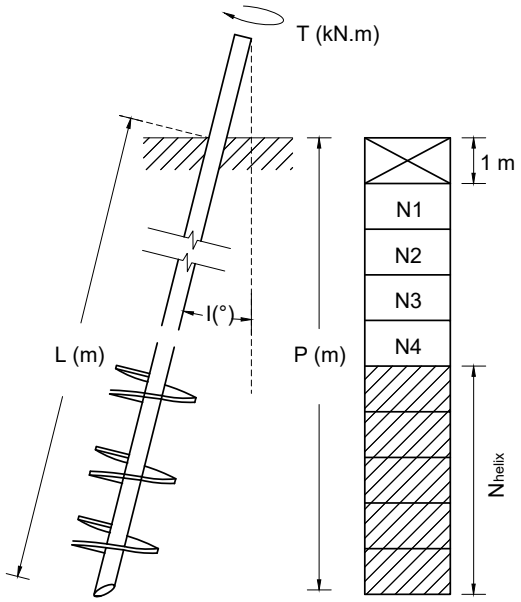

Figure 3. Schematic representation of a helical pile. A torque T is applied to the pile top. The pile has length L that vertically corresponds to P. The first soil layer is disregarded, $N1$–$N4$ is the SPT index of the following layers, and N_{helix} is the mean SPT index for the rest of the soil in contact with the pile.

Table 1. Types of soil and encoding. All soil types become equally spaced in input space with the proposed binary approach.

Soil Type	$S1$	$S2$	$S3$	$S4$	$S5$	$S6$
Clay	1	0	0	0	0	0
Silty sand	0	1	0	0	0	0
Silty clay	0	0	1	0	0	0
Sandy clay	0	0	0	1	0	0
Sand	0	0	0	0	1	0
Clayey sand	0	0	0	0	0	1

Table 2. Variables of the Complete Dataset. It includes a geometrical input (P), values related to the SPT index ($NSPT_{tip}$, $N1$–$N4$, $NSPT_{helix}$, and N_{shaft}) qualitative inputs encoded as binary ($S1$–$S6$), and the output (Torque).

Variable	Meaning
P	Vertical penetration in m
$NSPT_{tip}$	SPT index at pile tip depth
$N1$, $N2$, $N3$, $N4$	SPT index in top layers
$S1$, $S2$, $S3$, $S4$, $S5$, $S6$	Soil type
$NSPT_{helix}$	Mean SPT index at helical plates
N_{shaft}	Accumulated SPT index at shaft
Torque	Output variable in $kN \times m$

The third proposed dataset, the Contribution Dataset, was based on variable importance and used the same inputs as the Complete Dataset, excluding variables $S1$–$S6$. This exclusion was made after measuring variable importance and concluding they had little effect on the regressions. This low importance can be explained by the homogeneity of the used dataset, with a predominance of clayey sand.

The final step of pre-processing consists of evaluating input correlation. This work uses Pearson's correlation, which measures the association between variable pairs. Values close to ± 1 indicate perfectly related variables, while 0 means they are independent. Highly correlated inputs should be avoided because they cause redundancies that can destabilize

regression techniques [21]. In this work, all correlation indexes were in the $[-0.9, 0.9]$ range, which can be considered reasonable [22]. Figure 4 presents the so-called correlation matrix for the Complete Dataset.

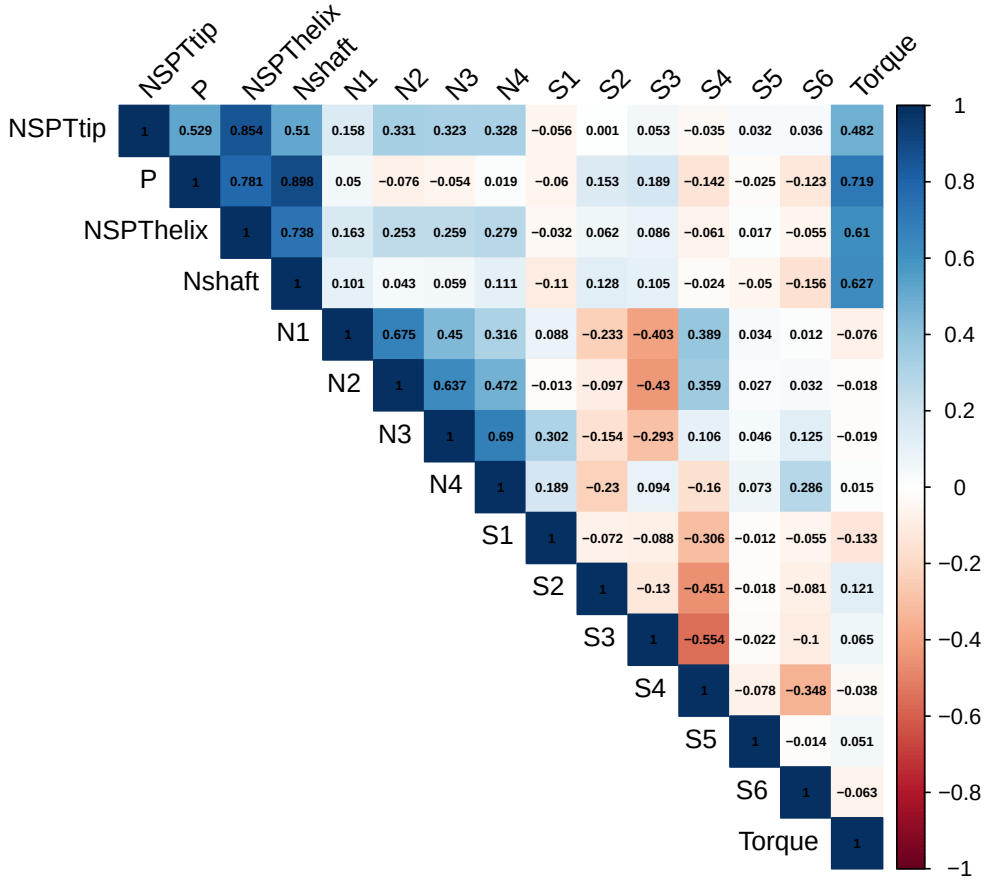

Figure 4. Correlation matrix considering all inputs considered in this study and the output. Darker colors indicate a stronger correlation.

Many other inputs were considered in this study; however, they are omitted here for conciseness. The ones selected to be presented in this section provided the most interesting conclusions, including the most accurate results. The study also considered that the three main parts of the helical pile (shaft above helical plates, helical plates region, and pile tip) should be well represented. Further details, such as different helix diameters, were not considered because this would require knowledge of the soil disturbance mechanism during installation.

2.2. Calibration of ML Techniques

The dataset was divided into two sets, namely training and testing. The first was used to calibrate the ML techniques, and the second was used to measure their performance. Considering literature recommendations (e.g., [22]), 80% of all examples were used for training and 20% for the testing stage, using random selection. The seed used to generate the random state was 998. Considering the size of the original dataset, its distribution tends to be replicated in the test and train datasets.

After calibration, the ML models were tested to estimate the installation torque for new data. The testing dataset, which represents 20% of the primary dataset, was kept apart from the training procedures to reproduce practical situations in which the data are entirely unknown. For training and testing, the metrics described below use inputs x_i to compare the predictions $f(x_i)$ with the known target values y_i.

The MAE was calculated using the absolute difference between known and predicted values. For n examples, the MAE can be obtained by Equation (3). The $RMSE$ represents the standard deviation of the difference between the observed and predicted values (Equation (4)). Finally, R^2 is a dimensionless measure representing the correlation between predicted and observed values (Equation (5)).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - f(x_i)|}{n}. \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2}. \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}. \tag{5}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{6}$$

These parameters evaluated the models' performance during the testing stage and their generalization during the training stage. Compared to the test, exact values in training can indicate overfitting, which means that the model reproduces even noisy data from the training dataset. An over-specialized model tends to become inaccurate when applied to new data. On the other hand, the model losing trends during training indicates underfitting, meaning the model ignores essential information from the training dataset, compromising generalization.

Models were calibrated in training using hyperparameters, which are specific configurations for each machine-learning technique. One strategy to set these hyperparameters is training several models with random values and choosing the best accuracy. However, this strategy is impractical and tends to have high computational costs. An alternative employed in this work is the so-called "pick the best" technique [21], which randomly chooses different parameters in training, decreasing the computational cost to adjust the parameters.

In this work, linear multiple regression (LM) was used as a reference to evaluate the predictive performance of the ML models. Linear regressions work to minimize errors in approximating data using linear functions. Equation (7) represents its general form:

$$f(x_i) = \alpha + \beta x_i. \tag{7}$$

where x_i represents the input, $f(x_i)$ is the output and α and β represent parameters to be calibrated. The best fit corresponds to the solution of an optimization problem for minimizing error. Equation (8) presents the measurement of the error J considering the least squares technique:

$$J = \sum_{i=1}^{n}(y_i - f(x_i))^2. \tag{8}$$

The concept is easily extendable to a general number of inputs.

2.3. ML Techniques

This work used eight models for torque prediction: ANN for artificial neural networks, DT for the decision tree, KNN for the k-nearest neighbor, RF for the random forest, SVM for the support vector machines, BAG for the DT associated with bagging, CUB for cubist and BOO for the LM associated with boosting. Preliminary results showed that combining boosting with non-linear techniques produces no relevant improvement in performance and increases the computational cost significantly. Therefore, this work does not include such cases. The following sections provide a brief description of each of these algorithms.

2.3.1. DT and CUB

A DT is a unidirectional graph that starts at a root node, proceeds to decision nodes, which divide the dataset using predefined rules, and ends at leaf nodes, where a value is assigned to the output [23]. When used for regressions, the model uses error parameters during training to define the tree architecture and calibrate the rules for decision nodes. Figure 5 presents a DT example that uses inputs P, $N1$, and $N2$. Each node presents a predicted torque and a percentage of data representing it. Note that the sum for all nodes of any layer is 100%. Most DTs are binary, meaning that each decision node distributes examples exactly to two nodes. The rule used in each node should be calibrated to maximize the precision of the DT, considering first the leaf node and then the previous ones recursively.

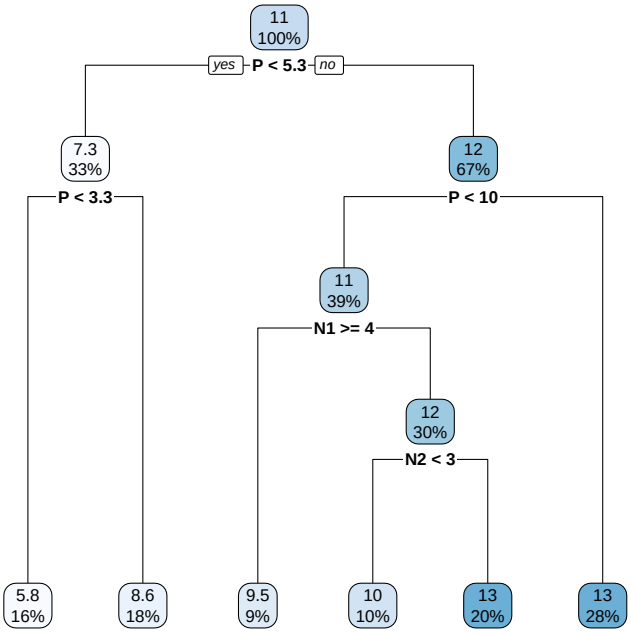

Figure 5. Example of a decision tree. Each node applies a rule to decide which other node should receive data. In this example, possible predictions for torque are within $[5.8, 13]$.

DT models have the advantage of being interpretable, and their main disadvantage is overfitting. In this work, overfitting is avoided by pruning the trees, which is a procedure that eliminates redundant and irrelevant branches. The DTs can also be associated with linear regressions for the tests performed at each node, which leads to the CUB model [24]. The linear regressions are calibrated considering the prediction by the previous nodes, recursively to the root node. Pruning also applies to the CUB model.

2.3.2. KNN

The KNN is a non-parametric ML technique, meaning the model is determined from the existing dataset structure [25]. It includes the following steps:

- The model starts with the training dataset.
- It includes value ranges to predict new points.
- It selects the k-nearest neighbors (in this example $k = 3$) from the range limits.
- The mean value of all neighbors is given to the new unknown point.

The inputs are coordinates to calculate the distance to the nearest neighbors. The Minkowsky metric is popular in the literature and uses the following expression:

$$d(A, B) = \left(\sum_{i=1}^{dim} |a_i - b_i|^{exp} \right)^{1/exp} \quad (9)$$

where A and B are points between which distance d is calculated, dim is the space dimension, a_i and b_i are the i^{th} coordinates of points A and B, respectively, and exp is a parameter to be chosen. This work uses the Euclidean distance, with $exp = 2$.

2.3.3. ANN

The human neurological system inspires ANN. The basic process of an artificial neuron is shown in Figure 6. The neuron receives x_i inputs, or signals, weighted by w_i. These contributions sum is the input of an activation function f_{ativ} that gives the neuron output y, which can be an input to another neuron. Connections between neurons and their weights determine the architecture of the ANN. One can demonstrate that one neuron can replicate any linear function.

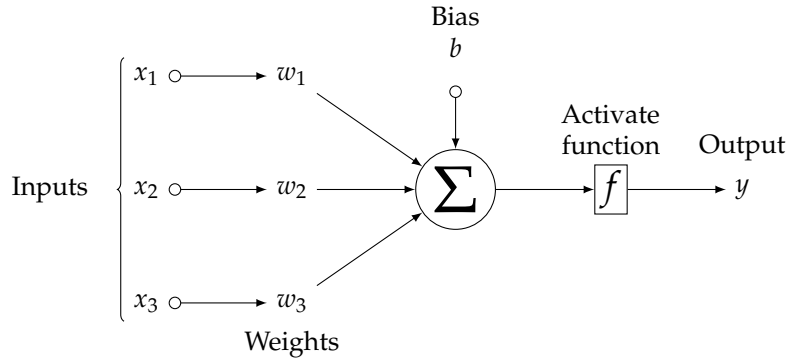

Figure 6. Isolated neuron. It produces the output using an activation function whose inputs can come from other neurons and are weighted.

The training of an ANN model is based on the adaptation of the architecture to minimize predictive error. Neurons are usually organized into layers, including an input layer, one or more hidden layers, and an output layer. Figure 7 illustrates a neural network. The universal approximation theorem states that an ANN with one hidden layer containing a finite number of neurons can approximate any continuous function. Thus, one can demonstrate that an ANN with two or more hidden layers containing finite numbers of neurons can approximate any function.

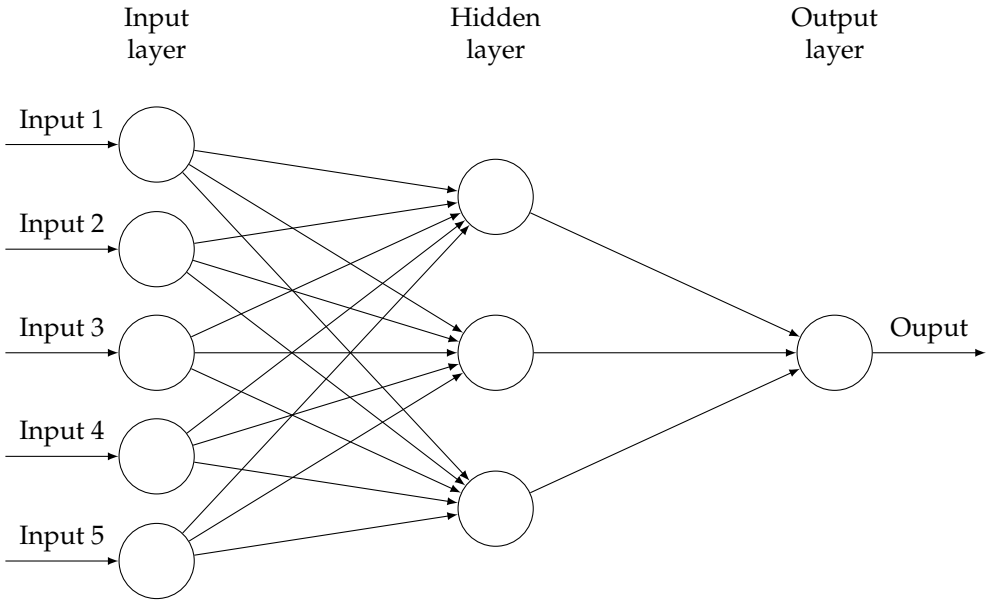

Figure 7. Neural network. The number of neurons of the hidden layers defines the architecture of the ANN, aiming to reduce error. Usually, every neuron of a hidden layer is connected to all neurons of the previous and the next layers.

The input layer contains one neuron for each input I_i and feeds the first hidden layer. There is no limit to the number of hidden layers or neurons H_i used for each, except computational cost. In classification problems, the output layer contains one neuron for each possible output O_i and only one for regression problems [26].

2.3.4. SVM

The SVM basic idea is constructing hyperplanes separating positive and negative data, ensuring global minimums, and maximizing the model generalization [27]. In regression problems, boundary lines create a margin around the hyperplane to contain all points of the training dataset, and support vectors are the data points placed at the boundary lines. Considering a margin ϵ, the hyperplane and boundary lines can be formulated as

$$f(x) = wx + b \pm \epsilon \tag{10}$$

The solution of the optimum hyperplane that minimizes ϵ is unique. It is possible to smooth the margins, allowing some training points to be outside the margin limits, which contributes to avoiding overfitting. It is also possible to embrace non-linear problems, such as the one presented in Figure 8, using kernels: functions that map the input space into a higher dimension space. The objective is to obtain a problem in this higher dimension that can be evaluated using a linear SVM. Figure 9 presents an example of kernel mapping data from a 2D space to a 3D space.

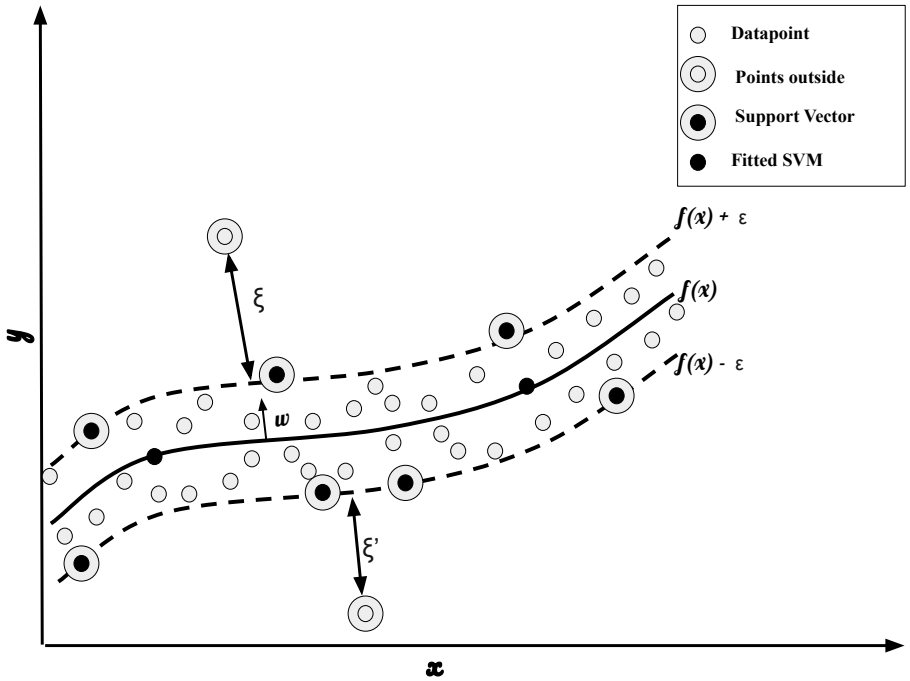

Figure 8. Two-dimesional non-linear problem evaluated with an SVM, using a kernel. Support vectors are points placed over the margins. Smooth margins penalize points outside range $[f(x) - \epsilon, f(x) + \epsilon]$.

Penalties are imposed upon training data outside the limits, considering their distances to the margin (ξ and ξ' in Figure 8). The most popular kernels in the literature use linear, polynomial, or radial functions. In this work, after preliminary tests, the polynomial kernel was chosen.

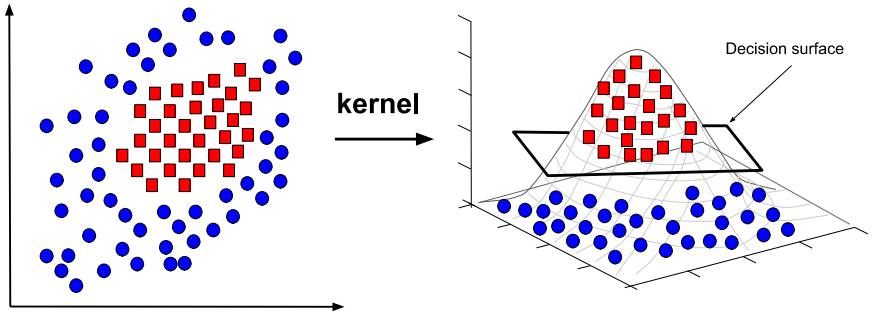

Figure 9. Using a kernel to map data from 2D to 3D space. A non-linear problem in 2D becomes linear in 3D.

2.3.5. Ensemble Techniques

Ensemble techniques work by combining two or more predictors. Such a definition embraces three techniques used in this work: BAG, RF, and BOO.

Ref. [28] presented BAG using a bootstrapping approach, which collects samples from the training dataset with replacement and trains the model with each instance. The mean value obtained from all trained models is the result of regression. This procedure tends to improve accuracy by reducing variation within unstable models. BAG can be combined with any regression model.

The RF technique uses DT, which BAG improves. The difference involves selecting random input combinations for each BAG technique model [29,30]. The number of random inputs is a calibration parameter.

The basic idea of the BOO technique is to focus on the model's weaknesses to make it stronger [31]. The model is trained repetitively. During each run, samples in which the model is inaccurate are identified, and then additional weight is given to these samples for the next run. The final model is a linear combination of all trained models, giving extra weight to more accurate models. The BOO technique can also be combined with any regression model; however, in this work, BOO references mean boosting applied to LM.

3. Results

All results presented in this section refer to the regression techniques applied to the Initial Dataset, Complete Dataset, and Contribution Dataset. They resulted in 7632 observations after the pre-processing stage, divided into a training dataset (6108 observations, 80%) and a test dataset (1524 observations, 20%). All models are evaluated using MAE, $RMSE$, and R^2 (see Section 2.2), along with the "pick the best" approach.

3.1. General Analysis

The left part of Figure 10 presents results obtained with the Initial Dataset. This figures show the evaluation metrics obtained with the eight ML techniques at the training and testing stages. The performance of each method is similar when the training and test stages are compared, which evidences that no overfitting occurred. In both scenes, LM was inferior to all ML techniques, which suggests torque non-linearity and, therefore, justifies the use of ML techniques. The DT model showed the second-poorest performance, which its simple form can explain since RF and CUB (improved techniques also based on trees) resulted in the second and third greater R^2, respectively. Finally, the LM associated with the BOO model completes the top three models with the best performance in the training and testing stages.

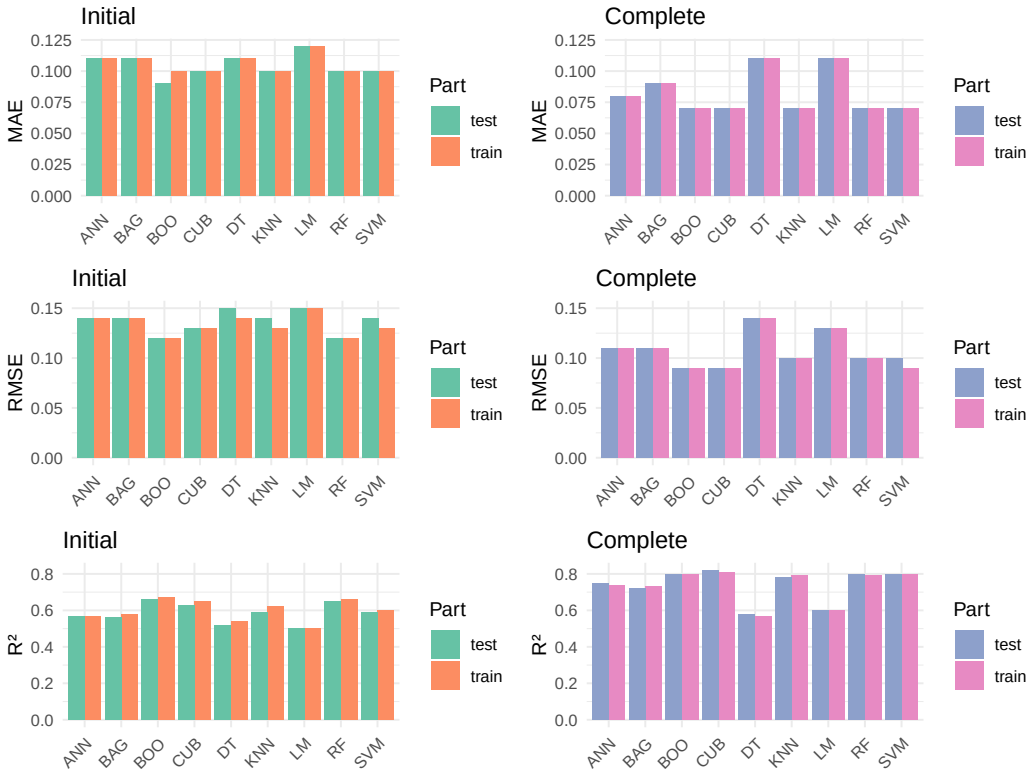

Figure 10. Metrics for training and testing stages for Initial Dataset (**left**) and Complete Dataset (**right**). LM and DT are the worst techniques, while BOO and CUB present the best performance.

The right part of Figure 10 presents the performance of the models in the training and testing stages regarding the Complete Dataset. Model performance was again similar when results of both training and testing stages were evaluated, evidencing no overfitting. The poorest performance occurred for LM and DT models, exhibiting very close values for the evaluation parameters. The other tree-based models (RF and CUB) performed well again, which emphasizes that the poor performance of the DT model was due to its simple form. The CUB model showed the best performance in both stages, followed by SVM in the training stage and BOO in the test stage. More importantly, comparing metrics from the Initial and Complete Datasets reveals that all models improved performance when the Complete Dataset was used.

This work employs the "pick the best" technique to calibrate ML techniques' hyperparameters. To verify if their full potentialities are being explored, an additional analysis considering a more careful hyperparameter selection was performed only for the two best-performing models in the testing stage using the Complete Dataset: CUB and BOO. Table 3 presents the hyperparameter grids used in the study. CUB training varied the number of committees from 1 to 100 and neighbors from 0 to 9, totaling 1000 combinations. BOO combinations used seven types of rounds, five lambdas, five alphas, and eight etas, totaling 1400 combinations.

Figure 11 presents the variable importance determined for the two best-performing models in the test stage with the Complete Dataset (BOO and CUB). The importance of a given variable is a performance measurement that determines how much the random chance of a given variable deteriorates the predicted result. It was initially suggested by [29] and enables the comparison and selection of input variables to produce accurate models with less computational cost. In this work, the percentage of predictions in which the variable was used results in its importance. Therefore, the importance is a value between 0 and 100%.

Table 3. Parameter grids for the best ML models. For CUB, the procedure tested all combinations within specified ranges. This procedure is impossible for BOO, which uses continuous parameters.

Model	Hyperparameter	Values
CUB	Committees	1 to 100
CUB	Neighbors	0 to 9
BOO	Rounds	50, 100, 150, 200, 250, 300, 350
BOO	Lambda	0, 0.5, 1, 1.5, 2
BOO	Alpha	0, 0.35, 0.7, 1, 1.5
BOO	Eta	0, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 1

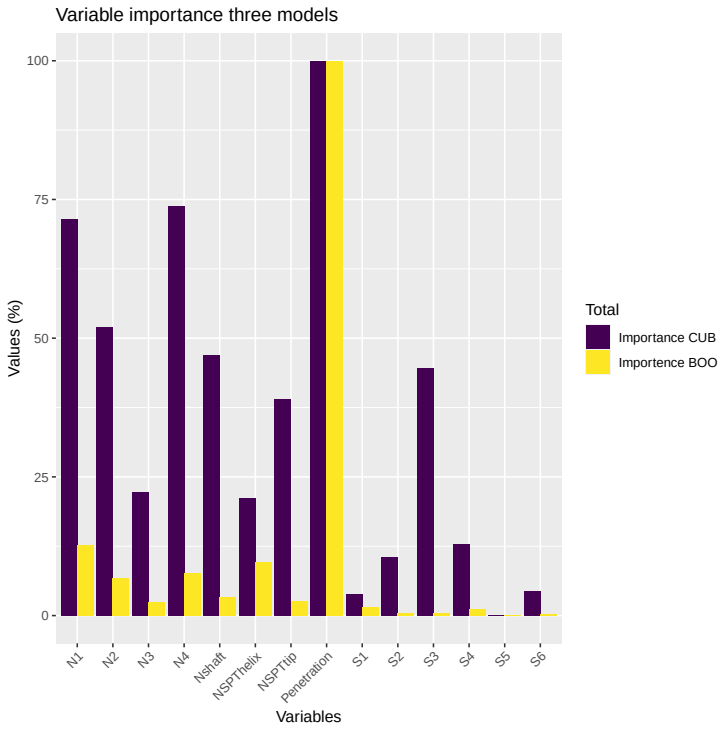

Figure 11. Variable importance for the best two models. CUB tends to explore better dataset information with a balanced usage of the variables. Both techniques used penetration in all predictions (100%), and no technique used *S*5.

The pile depth (*P*) was used in all predictions for the three models and is the most crucial variable. The prevalence of the importance of *P* is more apparent for BOO, considering that the second most important variable for this model (*N*1) was used in approximately 13% of all predictions. On the other hand, the CUB model showed the most balanced usage of variables, with seven variables being used in more than 40% of all projections.

The soil type, represented by *S*1–*S*6, had little importance for the two models, except for S3 for CUB. Considering that irrelevant data can jeopardize the model performance, a new dataset named Contribution Dataset was created, excluding the soil type variable from the Complete Dataset. Table 4 presents the accuracy of the model's BOO and CUB obtained with the new dataset for the training and testing stages. The metrics obtained using the Complete Dataset (right side of Figure 10) reveal similar accuracies, with a slight improvement in CUB and a slight deterioration in BOO. Despite the negligible gain in accuracy with the Contribution Dataset, it is worth noting that the computational cost can reduce significantly with the reduction in the number of inputs. Moreover, the negligible gain in accuracy further indicates that the soil type information is irrelevant to the analyzed data (predominantly sandy soils), which justifies the removal from the dataset.

Table 4. Metrics for the two best-performing models using the Contribution Dataset. Performance remains similar, with the advantage of reducing the computational cost.

	MAE		*RMSE*		R^2	
Algorithm	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
BOO	1.20	1.16	1.63	1.60	0.79	0.80
CUB	1.15	1.08	1.52	1.47	0.81	0.83

Figure 12 shows torque-measured values compared with the predictions made with BOO and CUB models for the test stage using the Contribution Dataset. In this figure, the grey dashed line represents predictions equal to measured values, and the black lines represent limiting prediction values equal to measured values ±50%. A reasonable accuracy is observed for the predictions with the two models. Less deviation is observed for torque values greater than ∼eight kNm, generally not associated with shallow depths.

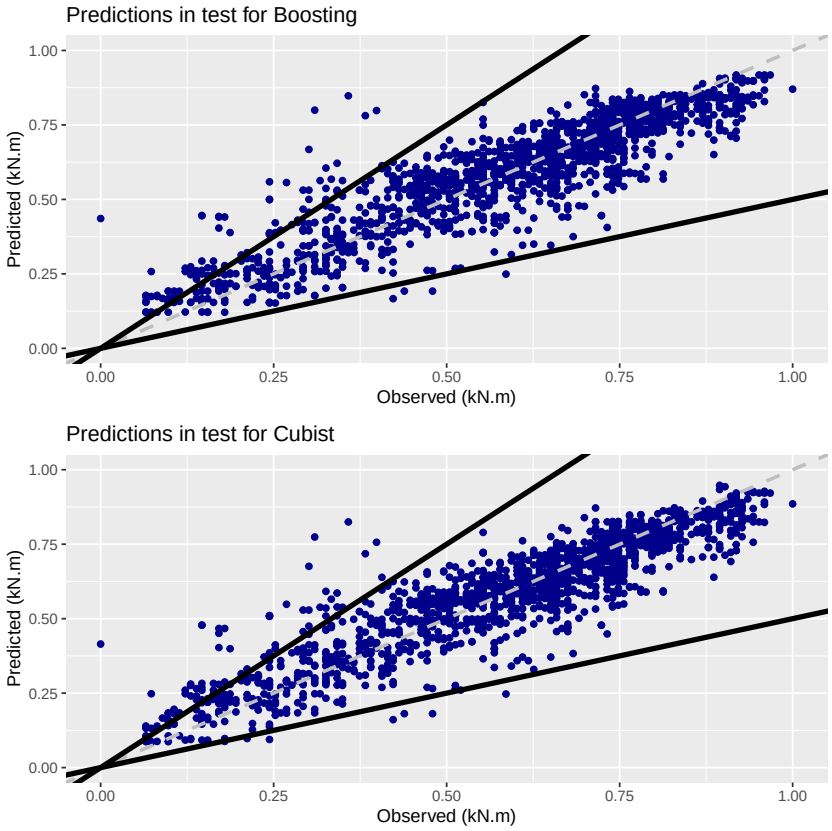

Figure 12. Comparison between observed and predicted installation torque. Considering a predicted torque T, the area between black lines represents values within $[0.5T, 1.5T]$. One can observe that this area contains the most points, showing that predictions are reliable.

A complement to the above analysis is given using confidence intervals. A confidence interval is a data percentage expected to be within a given value range. This range is proposed here using a factor given by the ratio of the predicted torque to the measured torque, as presented in Equation (11).

$$Factor = \frac{Predicted}{Observed} \tag{11}$$

This factor was calculated for all 7632 predictions obtained using BOO and CUB models, with the results sorted in ascending order to construct a histogram for each model. Figure 13 presents the histograms showing the frequency distribution of these data in

intervals of 0.01. The dashed line at the factor equal to 1 is the median, meaning that the predicted value equals the observed one. The left and right-handed dashed lines represent, respectively, the 2.5% and 97.5% percentiles; therefore, the range between these lines contains 95% of all factors, corresponding to a 95% confidence interval.

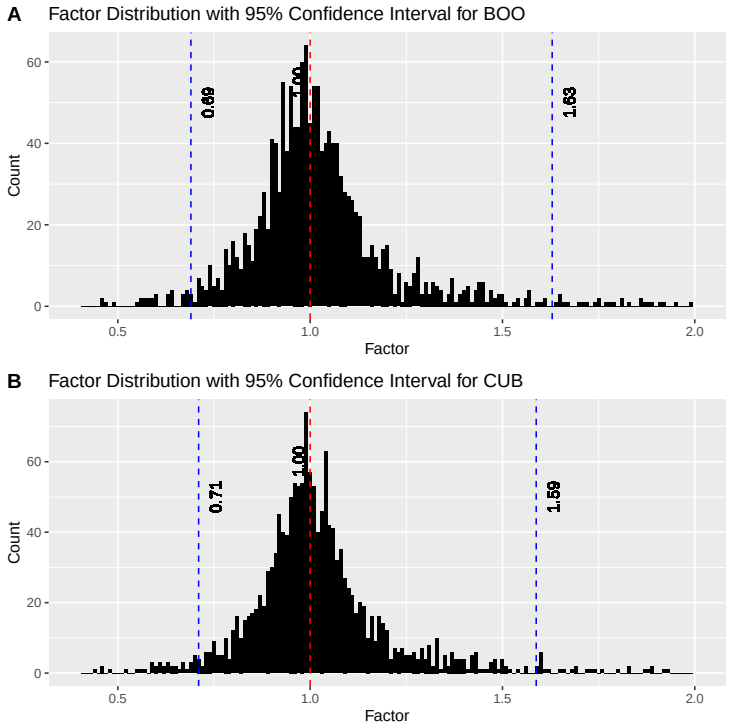

Figure 13. Comparison between observed and predicted installation torque. One can observe that distributions are approximately normal, although the mean is not well centered between the 2.5% and 97.5% percentiles.

For the CUB model, 95% of all predicted values are between 0.71 and 1.59 times the observed value, which means there is 95% confidence that the interval $[0.71, 1.59]$ will contain the factor for CUB. The confidence interval for BOO is $[0.70, 1.61]$. The CUB model has the smallest interval size. Therefore, it is the most accurate.

In this work, the 95% confidence interval was chosen because most statistics applications use this percentage. However, other rates could be used depending on the engineering application. For example, smaller confidence percentages, such as 90% or 80%, would lead to smaller interval sizes and vice versa.

3.2. Case Study

A case study using the current approach is presented to determine the maximum pile length based on a limiting torque value. The values of NSPT and measured torque were taken from the Contribution Dataset. A limit torque value of 14 kNm was chosen based on the low capacity of the machinery available on the Brazilian market for helical pile installation.

Table 5 presents the results obtained with the CUB, the ML model that showed the best results in previous sections. Notice that the SPT index and its related depth can deduce all required inputs. The limits of this interval were calculated using the values from Figure 13, with the lower limit being given by 0.71 times the predicted value and the upper limit by 1.59 times the expected value. The first column starts with $L = 2$ m since torque is not measured for $L = 1$ m.

Table 5. Results for the case study. T_{min} and T_{max} are the limits of a 95% confidence inteval. T_{pred} passes the limit of 14 kNm for $L = 8$ m. Therefore, the limit length is $L = 7$ m.

L (m)	$NSPT_{tip}$	T_{obs}	T_{pred}	T_{min}	T_{max}
2	2.00	6.10	4.63	3.28	7.36
3	3.00	7.86	8.06	5.72	12.82
4	6.00	9.08	10.43	7.40	16.58
5	6.00	9.63	11.46	8.14	18.23
6	7.00	11.66	12.50	8.87	19.87
7	6.00	13.15	13.50	9.58	21.46
8	7.00	13.69	14.65	10.40	23.29

In this example, the torque prediction is more accurate for pile lengths > 4 m (the shortest piles in the current dataset had 8 m lengths). Since the pile bearing capacity is usually estimated using the final installation torque, higher accuracy with the increase in pile length can be considered an advantage. Concerning the torque limitation, for a machinery torque capacity of 14 kNm, the installed size of the pile would reach $\sim$7 m for both predicted and measured values. Therefore, supposing that only SPT data are available in a preliminary design phase, torque values with depth could be estimated to assist in determining the maximum feasible pile length or in the specification of installation equipment.

3.3. Conclusions

One issue in transmission tower projects is the amount of material to be transported to the construction site, given that only SPT data are known in the first stages. This work addresses this issue, providing a tool to predict installation torque for helical piles. A second step enables the prediction of pile helical length from SPT data. Thus, it presents valuable information for real engineering problems, such as (a) obtaining confidence intervals, which are more informative than safety factors, and (b) observing that soil type was not helpful for the used datasets. Low soil variability within the used datasets explains the latter, which justified removing soil type from the datasets. One advantage of the non-specification of soil type is reducing subjectivity, since all inputs can be deduced from the SPT, as in the torque-to-capacity correlation used in practice (K_t method). Nevertheless, soil type is expected to become more critical for datasets containing soil type diversity. Studying efficient methodologies for this dataset type is a promising research field for future work, as well as investigating other types of problems that commonly use helical piles, like offshore constructions.

Author Contributions: Conceptualization, M.S.P., J.A.S. and D.B.R.; methodology, M.S.P., J.A.S. and D.B.R.; software, M.S.P.; validation, M.S.P.; formal analysis, M.S.P.; investigation, M.S.P.; data curation, M.S.P.; writing—original draft preparation, M.S.P., J.A.S. and D.B.R.; writing—review and editing, J.A.S. and D.B.R.; visualization, M.S.P.; supervision, J.A.S. and D.B.R.; project administration, J.A.S. and D.B.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Coordination of Superior Level Staff Improvement, grant number 001.

Data Availability Statement: All data and code used during the study are available in https://github.com/marcelosrvperes/helical-piles-torque-prediction (accessed on 1 May 2024).

Acknowledgments: The authors thank Bruno Oliveira da Silva and Cristina de Hollanda Cavalcanti Tsuha for making the dataset used in this work available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mitsch, M.P.; Clemence, S.P. Uplift capacity of helix anchors in sand. In *Uplift Behavior of Anchor Foundations in Soil*; American Society of Civil Engineers (ASCE): New York, NY, USA, 1985; pp. 26–47.

2. Ghaly, A.; Hanna, A.; Hanna, M. Uplift behavior of screw anchors in sand. I: Dry sand. *J. Geotech. Eng.-ASCE* **1991**, *117*, 773–793. [CrossRef]
3. Lutenegger, A. Screw piles and helical anchors—What we know and what we don't know: An academic perspective—2019. In Proceedings of the ISSPEA 2019: 1st International Symposium on Screw Piles for Energy Applications, Dundee, Scotland, 27–28 May 2019; p. 15.
4. Elsherbiny, Z.H.; El Naggar, M.H. Axial compressive capacity of helical piles from field tests and numerical study. *Can. Geotech. J.* **2013**, *50*, 1191–1203. [CrossRef]
5. Hoyt, R.; Clemence, S. Uplift capacity of helical anchors in soil. In Proceedings of the 12th International Conference on Soil Mechanics and Foundation Engineering, Rio de Janeiro, Brazil, 13–18 August 1989; pp. 1–7.
6. Perko, H.A. *Helical Piles: A Practical Guide to Design and Installation*; John Wiley & Sons: Hoboken, NJ, USA, 2009. [CrossRef]
7. Harnish, J.; El Naggar, M.H. Large-diameter helical pile capacity–torque correlations. *Can. Geotech. J.* **2017**, *54*, 968–986. [CrossRef]
8. Tsuha, C.d.H.C.; Aoki, N. Relationship between installation torque and uplift capacity of deep helical piles in sand. *Can. Geotech. J.* **2010**, *47*, 635–647. [CrossRef]
9. Sakr, M. Relationship between Installation Torque and Axial Capacities of Helical Piles in Cohesive Soils. *DFI J.-J. Deep Found. Inst.* **2014**, *7*, 44–58. [CrossRef]
10. Spagnoli, G. A CPT-based model to predict the installation torque of helical piles in sand. *Mar. Georesources Geotechnol.* **2017**, *35*, 578–585. [CrossRef]
11. Davidson, C.; Al-Baghdadi, T.; Brown, M.; Brennan, A.; Knappett, J.; Augarde, C.; Coombs, W.; Wang, L.; Richards, D.; Blake, A.; et al. A modified CPT based installation torque prediction for large screw piles in sand. In *Cone Penetration Testing 2018*; Taylor & Francis: Delft, The Netherlands, 2018; pp. 255–261.
12. da Silva, B.O.; Tsuha, C.H.C.; Beck, A.T. A Procedure to Estimate the Installation Torque of Multi-helix Piles in Clayey Sand Using SPT Data. *Int. J. Civ. Eng.* **2021**, *19*, 1357–1368. [CrossRef]
13. Wang, B.; Moayedi, H.; Nguyen, H.; Foong, L.K.; Rashid, A.S.A. Feasibility of a novel predictive technique based on artificial neural network optimized with particle swarm optimization estimating pullout bearing capacity of helical piles. *Eng. Comput.* **2020**, *36*, 1315–1324. [CrossRef]
14. Wang, L.; Wu, M.; Chen, H.; Hao, D.; Tian, Y.; Qi, C. Efficient Machine Learning Models for the Uplift Behavior of Helical Anchors in Dense Sand for Wind Energy Harvesting. *Appl. Sci.* **2022**, *12*, 10397. [CrossRef]
15. Silva, B.O. Estimation of the Installation Torque of Helical Piles Using SPT Data (Original Work Published in Portuguese). Master's Thesis, University of São Paulo, Sâo Carlos, Brazil, 2018. [CrossRef]
16. Ross, J.L.S.; Santos, L. *Geomorfologia*; IBGE: Rio de Janeiro, Brazil, 1982; Volume 21, p. 222.
17. Barros, A.; Silva, R.; Cardoso, O.; Freire, F.; Souza, J., Jr.; Rivetti, M.; Luz, D.; Palmeira, R.; Tassinari, C. *Projeto Radam Brasil -Folha SD. Cuiabá*; IBGE: Rio de Janeiro, Brazil, 1982; Volume 21.
18. *ABNT NBR 6484:2001*; Soil—Standard Penetration Test—SPT—Soil Sampling and Classification. ABNT: Rio de Janeiro, Brazil, 2001.
19. *ASTM D1586-08*; Standard Test Method for Standard Penetration Test (SPT) and Split Barrel Sampling of Soils. ASTM: West Conshohocken, PA, USA, 2008. [CrossRef]
20. Lukiantchuki, J.; Bernardes, G.; Esquivel, E. Energy ratio (E^R) for the standard penetration test based on measured field tests. *Soils Rocks* **2017**, *40*, 77–91. [CrossRef]
21. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; Volume 26. [CrossRef]
22. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112. [CrossRef]
23. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
24. Quinlan, J.R. Combining instance-based and model-based learning. In Proceedings of the 10th International Conference on Machine Learning, Amherst, MA, USA, 27–29 June 1993; pp. 236–243. [CrossRef]
25. Dudani, S.A. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Trans. Syst. Man Cybern.* **1976**, *SMC-6*, 325–327. [CrossRef]
26. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
27. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef]
28. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
29. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
30. Ho, T.K. Random decision forests. In Proceedings of the 3rd IEEE International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282. [CrossRef]
31. Abney, S.; Schapire, R.E.; Singer, Y. Boosting Applied to Tagging and PP Attachment. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, USA, 21–22 June 1999; pp. 38–45.