

MACHINE LEARNING INTEGRATED COMPARATIVE STUDY OF SHEAR STRENGTH PREDICTION USING GEOTECHNICAL PARAMETERS

**Tahmeed Ahmed
Tanzila Mostafa Momo
Md. Abu Asif Shamsuddin Naveed
Faraque Hossain Pias**



**Department of Civil and Environmental Engineering Islamic
University of Technology
2022**



MACHINE LEARNING INTEGRATED COMPARATIVE STUDY OF SHEAR STRENGTH PREDICTION USING GEOTECHNICAL PARAMETERS

Tahmeed Ahmed (170051004)

Tanzila Mostafa Momo (170051019)

Md. Abu Asif Shamsuddin Naveed (170051039)

Faraque Hossain Pias (160051018)

**A THESIS SUBMITTED FOR THE DEGREE OF BACHELOR OF
SCIENCE IN CIVIL ENGINEERING (GEOTECHNICAL)**

**Department of Civil and Environmental Engineering
Islamic University of Technology
2022**

APPROVAL

This is to certify that the dissertation entitled “MACHINE LEARNING INTEGRATED COMPARATIVE STUDY OF SHEAR STRENGTH PREDICTION USING GEOTECHNICAL PARAMETERS”, by, Tahmeed Ahmed, Tanzila Mostafa Momo, Md. Abu Asif Shamsuddin Naveed, and Faraque Hossain Pias a certificate of compliance with the requirements of the Bachelor of Science Degree in Civil & Environmental Engineering.



Supervisor:

Istiaur Rahman

Assistant Professor

Department of Civil & Environmental Engineering

Islamic University of Technology (IUT)

Board Bazar, Gazipur-1704, Bangladesh.



Prof. Dr. Hossain Md. Shahin
Head
Dept. of Civil and Environmental Engineering
Islamic University of Technology (IUT)
Gazipur-1704, Bangladesh.

DECLARATION

We hereby certify that the study presented in this thesis was conducted by us as undergraduates, under the expert guidance of Assistant Professor Istiakur Rahman. We've taken the necessary efforts to ensure that the work is unique and original. We are able to verify that the job has not been plagiarized. The work might also be checked to make sure that it has not been published for any other reason (except for publication).

Tahmeed

Tahmeed Ahmed
Student Id. 170051004
May, 2022

Naveed

Md. Abu Asif Shamsuddin Naveed
Student Id. 170051039
May, 2022

momo

Tanzila Mostafa Momo
Student Id. 170051019
May, 2022

Pias

Faraque Hossain Pias
Student Id. 160051018
May, 2022

ACKNOWLEDGEMENT

“In the name of Allah, the Most Gracious, the Most Merciful.”

All glory and thanks are due to the Almighty Allah (SWT), for bestowing upon us the fortitude and bravery necessary to finish our undergraduate thesis. We would want to express our profound appreciation to our parents for being a never-ending wellspring of ideas and support throughout our lives.

We would like to convey our profound gratitude and appreciation to our Supervisor, Mr. Istiakur Rahman, Assistant Professor, Department of Civil and Environmental Engineering, Islamic University of Technology, for his gracious direction, competent advice, and constant support in supervising us. His assistance in terms of both technical and editorial aspects was crucial to the successful completion of this academic project. The paper wouldn't have been completed without his help and instruction.

In addition, we would like to use this opportunity to express our appreciation to all of the members of the faculty for the insightful recommendations they provided to us throughout the course of our research. We would like to express our gratitude to our friends, juniors, seniors, and batch mates within our departments for their insightful recommendations and kind support.

We would also want to extend our gratitude to Sthapati Associates Ltd. for their assistance, which consisted of supplying geotechnical soil investigation data from two distinct places in Bangladesh. This information was necessary for the conclusion of our study project.

DEDICATION

Our parents, who have invested a significant amount of their time, resources, and energy into assisting our development into the people we are today, are the recipients of the dedication that we have included in our thesis. They inspired and pushed us to pursue a profession in engineering without hesitating or making any reservations about doing so. When we were growing up, one of the most important lessons that our parents taught us was the need of having perseverance. They have our undying gratitude and appreciation.

ABSTRACT

In the world of Geotechnical engineering or Foundation engineering, the properties of soil become crucial to all the processes involved in determining whether it is suitable for supporting a given structure. Hence it is of essence that the significance of such parameters is predetermined to see which have the greatest effect. The information era is increasingly influencing every industry. Machine Learning is undoubtedly one of the most novel applications in forecasting soil parameters, and the integration of data and digital technologies opens up a plethora of options in the geotechnical field.

The utilization of artificial intelligence as an inexpensive yet accurate model has become a bright prospect since the efficiency of machine learning approaches has been proved for modeling various geotechnical parameters including the soil shear strength (SSS). Despite this, conventional techniques of estimating soil properties, which are both pricey and time intensive, are still utilized given the uncertainty around the accuracy of the prediction models. Thus, a soil shear strength predictive formula is presented for use as an alternative to the difficult traditional methods.

The focus of this research is to carry out comparative machine learning based Study of Shear Strength Prediction of Soil through Correlation Analysis of Geotechnical Parameters and also to observe how the previous models had fallen short and how to enhance the prediction of a parameter as important as soil shear strength.

A total of 5 different machine learning models were applied in this particular endeavor, namely Support Vector Machine (SVM), K Nearest Neighbor, Decision Tree and Ridge Regression which consisted of Linear Regression and Lasso Regression.

We used a total of 164 boreholes data for our research work. We used two different location our soil type was clay we used 67% of our data as training data and 33% of our data as testing data set.

The two evaluation metrics used in this paper were Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Pearson Correlation Analysis was also carried out as it measures linear correlation between two sets of data.

We also calculated our results in two ways, once before feature selection and another after feature selection. For both of these processes, we attained different results. Among these two if we see at given charts below we can see that after feature selection models performs better than before feature selection.

Keywords: Simple linear regression model, Support Vector Regression, Random Forest, Decision Tree, RMSE, MAE, K Nearest Neighbor, Ridge Regression, Linear Regression and Lasso Regression.

TABLE OF CONTENTS

Contents

APPROVAL	3
ACKNOWLEDGEMENT	5
DEDICATION	6
ABSTRACT	7
TABLE OF CONTENTS	9
LIST OF TABLES	11
1 INTRODUCTION	1
1.1 GENERAL	1
1.2 STATISTICAL PREDICTION MODELS.....	2
1.3 USE OF STATISTICAL PREDICTION MODELS IN PREDICTINGGEOTECHNICAL PARAMETERS.....	2
1.4 IMPORTANCE OF THIS STUDY	3
1.5 AIMS OF THIS STUDY	3
1.6 OBJECTIVES	4
1.7 WHY THE STUDY IS DIFFERENT FROM OTHERS.....	5
1.8 RESEARCH FLOW DIAGRAM.....	5
2 LITERATURE REVIEW	6
2.1 SUB-SOIL EXPLORATION METHODS.....	6
2.2 ESTABLISHING THE CORRELATION BETWEEN SOILPARAMETERS TO PREDICT.....	7
2.3 APPLICATION OF MACHINE LEARNING IN GEOTECHNICALENGINEERING	8
2.4 MACHINE LEARNING TECHNIQUES SELECTED TO APPLY INTHIS STUDY FOR PREDICTION	9
2.5 SUMMARY OF LITERATURE REVIEW	13
3 METHODOLOGY.....	14
3.1 STUDY AREA.....	14
3.2 DATA COLLECTION PROCEDURE	20
3.3 RESEARCH METHODOLOGY	21
4 RESULTS	22

4.1	RESULT SUMMERY.....	22
5	Discussion.....	25
6	CONCLUSION & FURTHER RECOMMENDATION.....	26
6.1	CONCLUSION.....	26
6.2	FUTURE RECOMMENDATION.....	26
7	REFERENCES.....	27

LIST OF TABLES

Table 1: Before feature selection (RMSE)	22
Table 2: Before feature selection (MAE)	23
Table 3: After feature selection (RMSE)	23
Table 4: After feature selection (MAE)	24

LIST OF FIGURES

Figure 1: Research Methodology	5
Figure 2: BH ID- C302-01	14
Figure 3: BH ID C305-01 & 02.....	16
Figure 4: BH ID C307-01& C308-01.....	16
Figure 5: BH ID C311-01.....	16
Figure 6: BH ID C312-01.....	17
Figure 7: BH ID C316-01 & 02.....	17
Figure 8: BH ID C317-01 & 02.....	17
Figure 9: BH ID C318-01 & 02.....	18
Figure 10: BH ID C307CPA-01	18
Figure 11: BH ID C8ST-01 & C3ST-01	18
Figure 12: BH Location in Hazrat Shah Jalal Int. Airport	19
Figure 13 Pearson Correlation Analysis.....	21
Figure 14: Effects of Feature Selection.....	25

\

1 INTRODUCTION

1.1 GENERAL

The primary objective of this study is to employ machine learning strategies in order to make a prediction of shear strength by making use of geotechnical factors, and then to make use of the results of that prediction as a geotechnical design consideration. Machine learning, often known as ML, is an empirical method in which a computer program learns from a dataset without the need to first code the problem and a strategy on how to solve it. There is a fast expansion occurring in the use of machine learning applications across the board in the engineering industry.

In the process of developing a structural foundation and determining its level of stability, shear strength is an essential criterion.

When discussing the mechanics of soil, the term "shear strength" refers to the amount of shear stress that a given soil can bear without failing. The resistance of the soil to shear is caused by the particles rubbing against one another and interlocking, as well as cementation or bonding at the points where the particles touch. Because of the interlocking that occurs as a direct result of shear strains, the volume of particulate material can either grow or shrink. If the volume of the soil increases, the particle density will drop, which will result in a reduction in the soil's strength; in this scenario, the point of maximum strength will be followed by a reduction in the soil's shear stress. The relationship between stress and strain will become stable after the material stops growing or shrinking, as well as when interparticle links are severed. The crucial state, also known as the steady state or residual strength, is a theoretical state in which the shear stress and density stay the same despite an increase in the shear strain. Physical qualities and mechanical properties are the two categories of properties that can be found in soil. Mechanical properties, such as unconfined compression strength and standard penetration value, help in determining an estimate of the soil's bearing capacity, whereas physical properties, such as moisture content, liquid limit, plastic limit, void ratio, and so on, help to provide information about the type and nature of the soil. Shear strength can be calculated by taking into account two internal properties of soil that are closely related to one another: cohesion and the angle of internal friction. As a consequence of this, it is essential, before to beginning the building of any structure, to identify the cohesion and angle of friction of the particular soil that will be used. It is necessary to undertake a subsurface examination of a particular soil in order to acquire additional information regarding these parameters; however, this type of investigation is not only expensive but also time-consuming. In this study, an attempt is made to construct a model that can predict the cohesiveness of a particular region of soil by connecting the parameters of previously gathered data using a variety of statistical models and machine learning approaches.

1.2 STATISTICAL PREDICTION MODELS

Statistical prediction models make forecasts about the future using methods like machine learning to determine what will occur next. The process of teaching a computer to think in the same way that a person would is referred to as machine learning. The primary function of machine learning models is to make predictions about future events by analyzing data from the past. The performance of a model for machine learning is evaluated based on how accurately it forecasts new data that the model has not yet been trained to recognize. Machine learning and deep learning are the two categories of predictive analytics algorithms that are available. Deep learning models are a subfield of machine learning that are becoming increasingly prevalent for applications involving the processing of audio, video, text, and images. However, this does not mean that benefits materialize arbitrarily; predictive modeling also displays the amount of problems that need to be overcome. The costs of adopting these statistical prediction models can be dramatically lowered. It is challenging to maintain large and complete data sets, and the process of data cleansing is necessary often.

1.3 USE OF STATISTICAL PREDICTION MODELS IN PREDICTING GEOTECHNICAL PARAMETERS

Using statistical prediction models, correlations for predicting geotechnical parameters for use in civil engineering design have been constructed. These models are used. Using approaches based on machine learning, it is possible to predict, with a certain level of accuracy, the values of a variety of geotechnical parameters. In the field of geotechnical engineering, empirical correlations are a common method for assessing the qualities of soil. Statistics and a significant amount of data gathered in a laboratory or in the field are used to develop correlations. Ridge Regression (RR), Lasso Regression (LR), K Nearest Neighbor (KNN), Support Vector Regression (SVR), Random Forest, and Decision Tree are some of the regression techniques that can be utilized (DT). These models acquire knowledge from the data that is provided to them and make an effort to discover a pattern within the data set, despite the fact that the underlying linkages are not known and the physical meaning is difficult to define. Machine learning is a useful methodology for our research because it does not presuppose any prior information about the data being analyzed.

1.4 IMPORTANCE OF THIS STUDY

When it comes to geotechnical engineering, the qualities of the soil play a significant role in determining whether or not the soil can adequately sustain a certain construction. The shear strength of the soil is a crucial metric that is utilized in the design and auditing of a wide variety of geoenvironmental and geotechnical structures, such as road foundations and pavements, earth dams, and retaining walls, amongst others. In order to obtain the findings of the shear strength test, it is necessary to obtain the results of many laboratory tests. Direct methods for measuring shear strength have a number of drawbacks, including the fact that laboratory tests for shear strength computation are a difficult and time-consuming task that primarily require performing destructive tests; high spending costs; and the requirement to learn specialized skills for working with complex apparatus. Other drawbacks of direct methods for measuring shear strength include the fact that these methods require the individual to acquire these skills in addition to performing destructive tests. As a consequence of this, if this research is successful in correlating soil properties and predicting shear strength through the application of machine learning, then it will be an important contribution to the field of geotechnical engineering.

1.5 AIMS OF THIS STUDY

The purpose of this investigation is to utilize machine learning in order to correlate soil factors and predict shear strength, as well as to analyze, contrast, and determine which machine learning strategies are the most successful. Additionally, the results of this study will assist in cutting expenses and shortening the amount of time required for shear strength testing.

1.6 OBJECTIVES

Objectives for the research project include achieving the following -

- Automated prediction of shear strength parameters using minimal geotechnical parameters.
- Comparative analysis of machine learning methods adopting the geotechnical engineering domain at every stage of the research.
- Reducing the time and cost for predicting shear strength further by excessive soil tests.
- Establish the correlation among cohesion and soil parameters.
- Compare the Machine Learning techniques and find out the most effective solution.

1.7 WHY THE STUDY IS DIFFERENT FROM OTHERS

In the past, numerous studies have been carried out to correlate parameters and derive formulas by making use of straightforward statistical equations; however, the purpose of this study is to make use of more sophisticated machine learning techniques in order to obtain more accurate prediction results. With the help of linked geotechnical characteristics and as few resources as possible and in an automated fashion, the purpose of this study is to make a prediction about the value of the shear strength parameter.

1.8 RESEARCH FLOW DIAGRAM

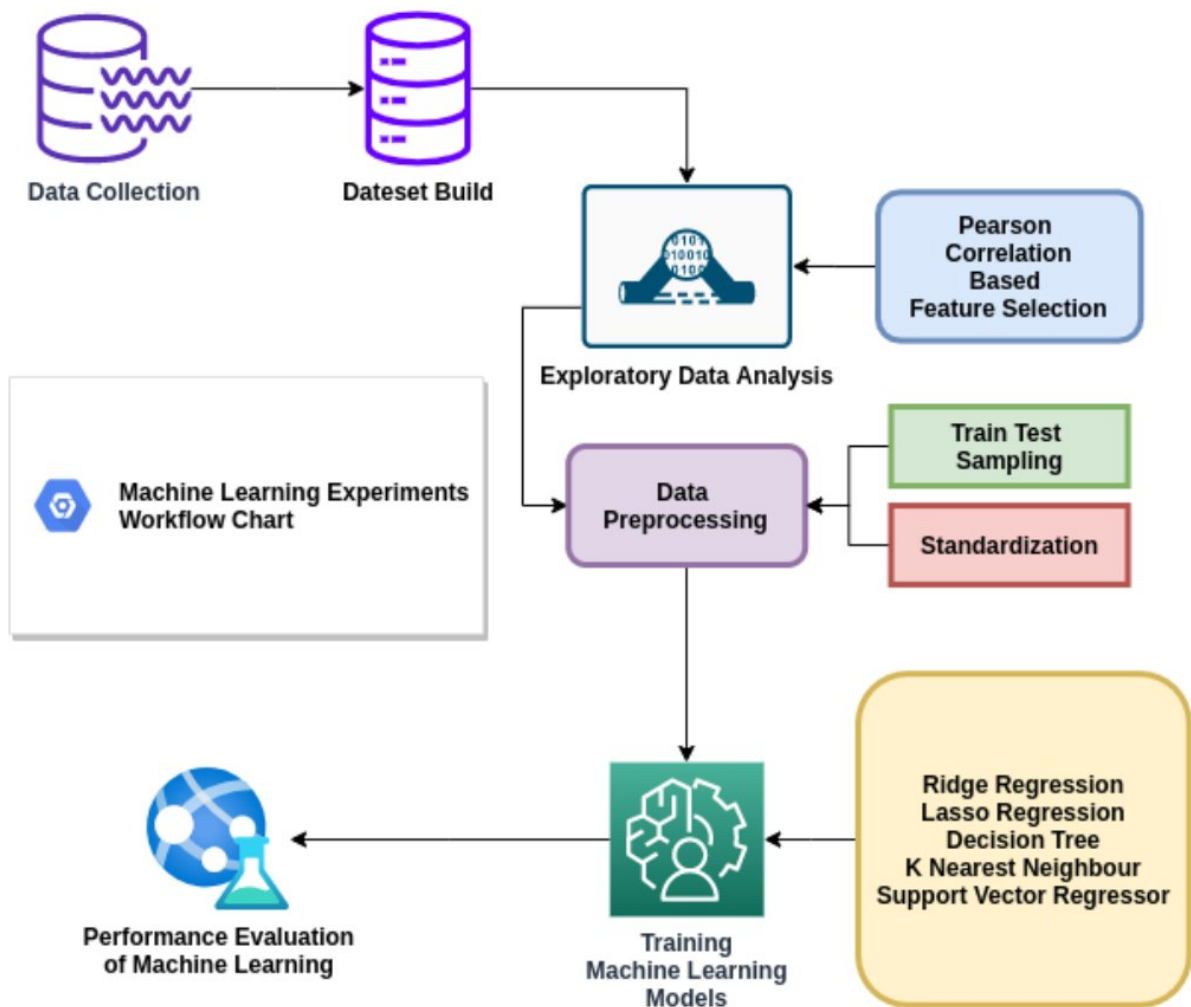


Figure 1: Research Methodology

2 LITERATURE REVIEW

The main focus of this study is the automated prediction of shear strength parameters with minimal geotechnical parameters as well as comparative analysis of machine learning methods adopting the geotechnical engineering domain at every stage of the research. Determining which machine learning model has the best performing algorithm helps determine if we can predict the value of Shear Strength parameter using correlated geotechnical parameters at minimal resources in an automated manner.

Therefore, in this chapter, related literature in regard to the study has been discussed in details. The total literature review is divided where they are categorized based on the main objectives and outcomes of this research.

2.1 SUB-SOIL EXPLORATION METHODS

When it comes to subsoil investigation, duration of testing and costs involved with testing are the two most important factors to consider. The three major phases of boring, sampling, and testing are generally adopted worldwide for running the sub-soil investigation before any building operation (Teng, 1983). Location, depth of borings, test pits, and boring technologies used for that sub-soil investigation are all important aspects of a sub-soil evaluation project.

The stratification and engineering features of the soil underneath the site, shear strength, deformation, and hydraulic characteristics of soil are all aspects to consider before a sub-soil investigation. According to Ngah (2013), collecting information, reconnaissance, preliminary exploration, and comprehensive exploration are the four phases of a soil exploration program. It is evident that for the completion of these phases and obtaining required data, a lot has to be invested into the traditional testing practices.

Whenever it pertains to just about any massive building project, subsoil research is essential between the tiny spacing distances. This research also found that if soil characteristics from one geographical setting can be strongly associated and used to forecast parameters in another geological location using a reasonably precise statistical prediction process, a huge proportion of soil tests, costs, and time can be recovered during the massive construction project.

2.2 ESTABLISHING THE CORRELATION BETWEEN SOIL PARAMETERS TO PREDICT

Many studies have been undertaken throughout the years to anticipate factors and save time and money. The use of a simple linear regression model allowed Kahdaar et al. (2010) to show the relationship between distinct physical variables (liquid limit, plastic limit, liquidity index, water content, density, void ratio) and different mechanical characteristics (q_u , C_c , C_s , SPT). The work came to a conclusion by applying the correlation with several early inquiry phases and assessments of any structure to uncover suggestive soil design characteristics. In some of the more recent works, like that of Phama (2018), relatively new fuzzy inference systems called PANFIS and GANFIS soil were used and result shows that among the four models, PANFIS and GANFIS have the highest prediction performance and among these two, PANFIS performed better.

There were more notable works and studies that delved in machine learning supported soil parameter prediction such as that of Moayedi (2019), where the applicability of two novel optimization methods ALO (ant lion optimization) and SHO (spotted hyena optimizer), was investigated for shear strength prediction when incorporated into an artificial neural network. The generalization capability of the ANN (artificial neural network) reinforced by the ALO is also helped produce more consistent results in the testing phase. However, at the conclusion of the study's research, it was determined that applying a standardized formula from one field to another is difficult. According to the literature reviews on the establishment of soil's parametric connections can be developed between soil properties using existing soil data for a specific geographical area, and these correlations can also be used to predict and calculate parameters for that specific geological location.

These approaches target to reduce the duration of lengthy field tests as well as the associated costs and also aims to use minimal features in the setup of models to define predictive formula of soil shear strength as an alternative to traditional methods. The comparison of such methods is vital for determining an effective prediction model that can be used in practical scenarios.

The ultimate goal is to enrich the use of automated methods for long-term, industry level solutions. Which led us to the reasoning or objective behind this research work on the specific locations as mentioned in previous parts of the paper.

2.3 APPLICATION OF MACHINE LEARNING IN GEOTECHNICAL ENGINEERING

Soil shear strength (SSS) needs to be meticulously evaluated for stability analysis of soil in comprehensive geotechnical engineering projects, especially since it indicates the internal resistance of soil against failure and sliding along any internal plane. Determining the SSS-related parameters by laboratory tests is a complicated, time-consuming task and mainly requires performing destructive tests.

Moreover, high spending costs, as well as learning special skills for working with complex apparatus are other disadvantages of direct approaches for shear strength measurement. Thus, alternatives such as newer digital solutions need to be found to automate the processes.

Numerous studies have been carried out to assess the effectiveness of machine learning techniques and according to Martens (2018), machine learning is the process of parsing data, learning from it, and afterwards deciding or predicting anything in the real world. Furthermore, machine learning has accelerated because to the practically infinite amount of available data, inexpensive information storage, and the development of less costly and more sophisticated computing.

Many sectors have created more powerful machine learning models capable of analyzing greater and more complicated data while delivering faster, more accurate results on large scales, according to Lindvall et al. (2018). Machine learning technologies help enterprises to more rapidly discover profitable possibilities and possible hazards. It allowed for newer approaches to be taken such as those seen in recent papers of Cao (2020). A metaheuristic-optimized meta-ensemble learning model (MOMEM) was established with the integration of the artificial electric field algorithm (AEFA) to dynamically blend the radial basis function neural network (RBFNN) and multivariate adaptive regression splines (MARS). This study is the first to put forward a novel stacking techniques-based ensemble model of hybridizing MARS and RBFNN that represent different types of learners for sharing reciprocal merits in the meta-ensemble model.

However, during the reviewing of all these qualified previous papers, some discrepancies could be noticed of which the most noticeable was that most papers did not relate to or adopt geotechnical parameters in models. If it is to be explained more elaborately, most of the algorithms did not carry out feature selection. This has left room for more diverse possibilities regarding incorporating machine learning algorithms into more heavily used geotechnical parameters.

Hence, the niche that this research paper fills is attaining close to accurate predictions with as few features (geotechnical parameters that are usually generated by traditional testing procedures) input into algorithms as possible.

2.4 MACHINE LEARNING TECHNIQUES SELECTED TO APPLY IN THIS STUDY FOR PREDICTION

Several machine learning approaches have been well developed over a long period time, and the vibrant application of machine learning is always changing and thus investigators have been hard at work developing new and improved models. Machine learning approaches may be divided into two categories, as per Chauhan et al (2018): standard machine learning and deep learning, both of which have a working principle of pattern recognition. This means the more trained the algorithms are, the more accurate predictions of soil parameters they will provide. Amongst the various models of machine learning algorithms, the particular models used in this paper included 5 variants. These include the widely used Support Vector Regression model (SVR) which is a linear support vector regressor, k Nearest Neighbors, Decision Tree and Linear Regression with two of its variants that are Lasso Regression and Ridge Regression. Once the models had been settled upon, some programming tools required to run the analysis were also chosen. These tools included Anaconda (for development environment), Pandas (for data manipulation and analysis), Scikit-Learn (machine learning models development) and Seaborn (for data visualization).

2.4.1 SIMPLE LINEAR REGRESSION

Lasso Regression

The lasso regression is an example of a linear regression algorithm that utilizes L1 regularization. It's a regression analysis technique that combines variable selection and normalization to improve the statistical model's prediction accuracy and interpretability. To prevent overfitting and make them operate better on diverse datasets, you may use the lasso regression to reduce or standardize these coefficients. When the dataset has a lot of multicollinearities or there is a need to automate variable removal and feature selection, this sort of regression is utilized. Lasso is a variant of linear regression in which the model is compensated for the summation of the weights' raw numbers. As a result, the critical values of weight will be lowered (in general), and most will be zeros.

Ridge Regression

This is another variant of linear regression algorithm that utilizes L2 regularization. In situations when linearly independent variables are heavily correlated, ridge regression is a technique of calculating the coefficients of multiple-regression models. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. The cost function is changed in this method by including the punishment term and so ridge regression consequence is the degree of bias introduced into the model.

2.4.2 SUPPORT VECTOR REGRESSION (SVR)

This model draws decision support vectors on multi-dimensional hyperplanes using a linear support vector regressor. Although it works on the same principle as support vector machine (SVM), it is slightly less widespread in its application to geotechnical engineering fields. It is a supervised learning algorithm that is used to predict discrete values and the basic idea behind SVR is to find the best fit line. The use of kernels, sparse solutions, and VC control of the margin and amount of support vectors characterize support vector regression (SVR). In SVR, the best fit line is the hyperplane that has the maximum number of points. The

SVR, unlike some of the other regression models, aims to fit the best line within a threshold value, rather than minimizing the error between the actual and projected value. The kernel function changes the data to a higher dimensional and accomplishes the linear separation in a non-linear regression. As a result, we can conclude that support vector regression is useful both in terms of regression and classification. The SVR's usefulness in modeling the complicated link between seismic and soil properties, as well as the liquefaction potential utilizing in situ data depending on the CPT, was also proved in through the various works before.

2.4.3 K Nearest Neighbor

This machine learning model utilizes a comparative distance metrics-based technique that is heavily reliant on the training and testing sets' variance similarity. The k nearest neighbor (KNN) method is a monitored machine learning technique that may be used to handle classification and regression issues. It's simple to set up and comprehend, but it has the disadvantage of being substantially slower as the amount of data in use rises. Here, determining the value of k is not as easy since a lower k number indicates that noise will have a greater impact on the outcome, whereas a big value indicates that it will be computationally costly. Its aim is to estimate the categorization of a new sample point using a database with data points divided into various groups. This is why it is usually written as a non-parametric, lazy-learning algorithm.

However, it saves the training dataset and only gains from it when producing real-time estimations. This makes the KNN method significantly quicker than other training-based algorithms like SVM and Linear Regression.

2.4.4 DECISION TREE

A decision tree learner for regression is a binary decision tree with linear regression functions at the terminal (leaf) nodes that can estimate continuous valued characteristics and is used to estimate the values of quantitative dependent variables Y . A decision tree is a decision-making aid that employs a tree-like representation of selections and their potential results, such as chance event outcomes, resource costs, and utility. It's one approach to show an algorithm made up entirely of dependent control statements. The algorithm is supervised learning, which means it is trained and evaluated on a set of data containing the intended categorization. Ultimately, it can be said that the decision tree is a form of probability tree that allows one to make a selection on a given procedure. A study was conducted by Naeef et al. (2016) about predicting hydraulic conductivity prediction based on grain-size distribution using the M5 model tree. The study yielded results such that the M5 model tree was effortlessly characterized the mathematical equations and executed the prediction analysis with less error value.

2.5 SUMMARY OF LITERATURE REVIEW

The literature review served as a generalized overview to the machine learning related works that have taken place so far in geotechnical engineering. Although by no means is this compilation enough to showcase the research work being done in this sector, it gives a slight idea of what has been attempted in the past and what more can be done in the future in the field. The machine learning models specifically discussed in the literature review were handpicked based on the input parameters and the parameter to be predicted (soil shear strength). The models discussed are therefore examples of conventional machine learning techniques and have been used together with statistical tools such as root mean square error. Machine learning models that this research paper particularly incorporated include:

- Support Vector Regression Model
- K Nearest Neighbor
- Decision Tree
- Lasso Regression
- Ridge Regression

Therefore, using these models the soil characteristic parameters that were investigated to establish correlation and eventually lead to prediction include Unconfined compressive strength, SPT-N value, cohesion, elasticity and more that have been shown later in results and analysis. These shall help train the algorithm on how to predict Soil Shear Strength. The major goal of this research is to use machine learning approaches to estimate soil characteristics and then apply the anticipated parameters to geotechnical design requirements.

3 METHODOLOGY

This part provides a comprehensive description of the study area, the process for collecting the data, and the preparation of the data for use with machine learning models.

3.1 STUDY AREA

For the purpose of this analysis, data on soil investigations were obtained from two separate initiatives. One of these is the Dhaka-Chittagong-Bazar Cox's Rail Project Preparatory Facility, which can be found in the area stretching from Chinki Astana to Chittagong (Component-2, 3 & 8). The inquiry plan called for boring holes in the ground and taking samples at intervals of 1.5 meters, which were then observed and subjected to laboratory analysis to determine the characteristics of the soil. In order to provide a clearer picture, the dull spots on the map were pinpointed with a hand-held GPS device and then placed on Google Maps. The field exploration program was carried out between the dates of May 3rd and June 24th of this year, as well as between June 18th and July 24th of the following year. Personnel from "Sthapati Associates Ltd" were the ones in charge of carrying out the program. They were responsible for conducting Standard Penetration Tests (SPT) and obtaining disturbed and undisturbed samples of the subsurface soil for laboratory testing and reporting. Another project that was carried out was the Hazrat Shahjalal International Airport Expansion Project. The geological position of the research region is depicted in the figures that follow below.



Figure 01: Boring location of C302-01

Figure 2: BH ID- C302-01



Figure 02: Boring locations of C305-01 & C305-02

Figure 3: BH ID C305-01 & 02



Figure 03: Boring locations of C307-01 & C308-01

Figure 4: BH ID C307-01 & C308-01



Figure 04: Boring locations of C311-01

Figure 5: BH ID C311-01



Figure 05: Boring locations of C312-01

Figure 6: BH ID C312-01



Figure 06: Boring locations of C316-01 & C316-02

Figure 7: BH ID C316-01 & 02



Figure 07: Boring locations of C317-01 & C317-02

Figure 8: BH ID C317-01 & 02



Figure 08: Boring locations of C318-01 & C318-02

Figure 9: BH ID C318-01 & 02



Figure 09: Boring location of C307CPA-01

Figure 10: BH ID C307CPA-01



Figure 10: Boring locations of C8ST-01 & C3ST-01

Figure 11: BH ID C8ST-01 & C3ST-01

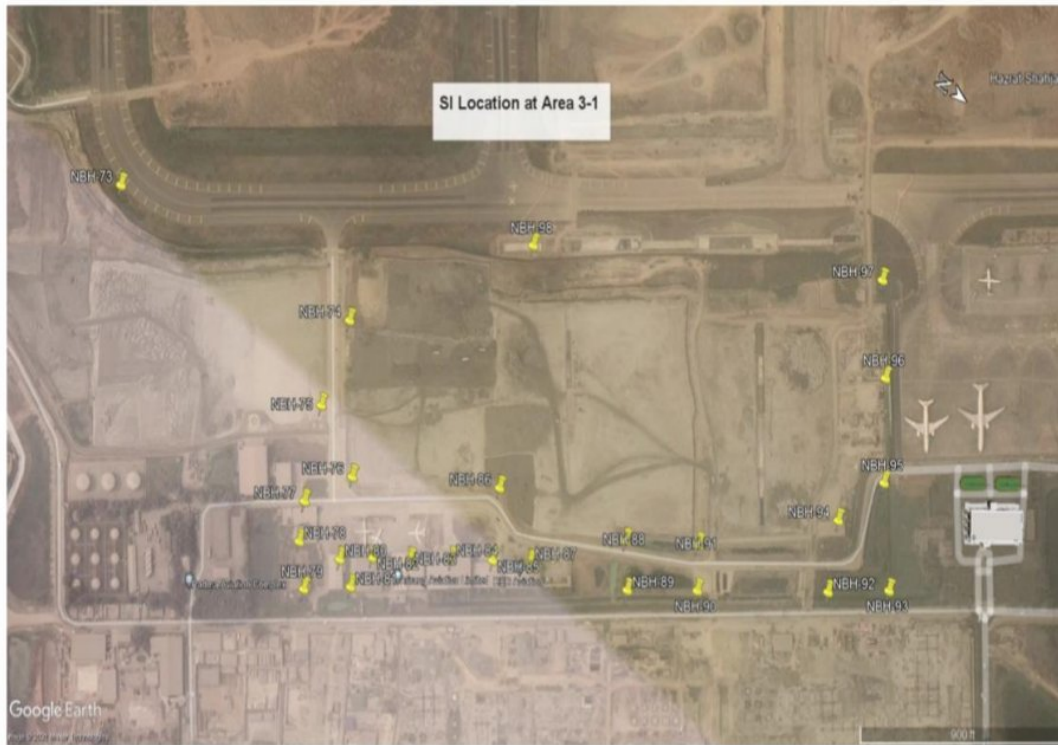


Figure: Soil Investigation Location Map [Area 3-1]

Figure 12: BH Location in Hazrat Shah Jalal Int. Airport

3.2 DATA COLLECTION PROCEDURE

The contract of two projects to collect and prepare the geotechnical investigation report was given to a company named "Sthapati Associates Ltd". The soil investigation experts visited selected places of the projects and later collected the forms of soil condition and other important information about that soil. Professional soil investigation companies used to collect data while adhering to all professional guidelines. As a result, the data used in this study can be verified as error-free and acceptable for research purposes

3.3 RESEARCH METHODOLOGY

The research project was broken down into three distinct stages for its whole. During the initial phase of the research, the data sets were obtained from two separate studies. The preparation of the datasets took place during the second phase of the project. And finally, in order to carry out our research, we utilized a variety of machine learning strategies, including Ridge Regression, Lasso Regression, K Nearest Neighbor Regression, Support Vector Regression, Random Forest, and Decision Tree.

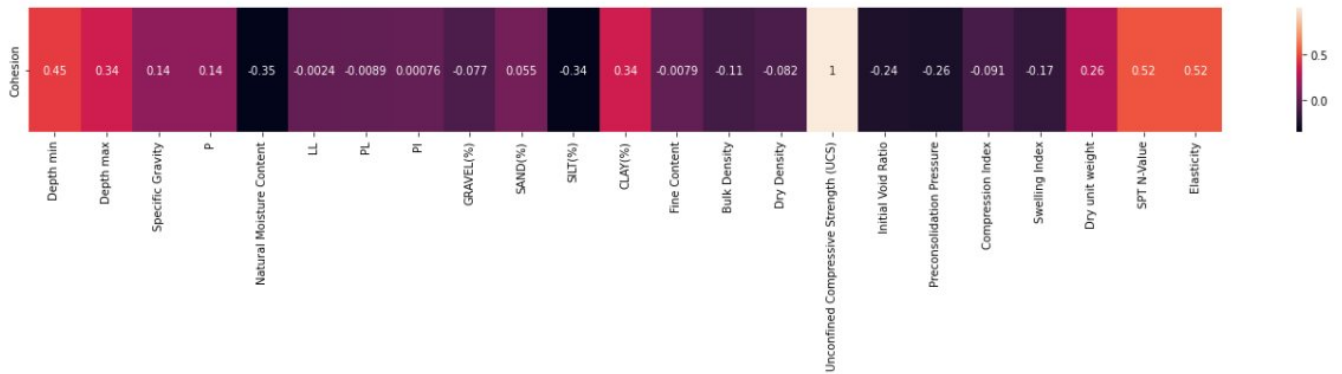


Figure 13 Pearson Correlation Analysis

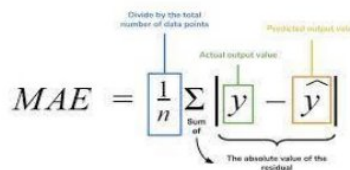
4 RESULTS

The summary of the results collected from all of the models is provided in this section.

4.1 RESULT SUMMERY

To conduct our research we have applied all our models to predict the collected data and the following table summarizes the findings of the research. For evaluating we used two evaluation metrics which are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). We also did Pearson Correlation Analysis which measures linear correlation between two sets of data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

$$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$$


We also calculated our results in two ways one is before feature selection and another is after feature selection for both of these processes we got different results. Among these two if we see at given charts below we can see that after feature selection models performs better than before feature selection.

The following is a summary of the results from all of the models presented in all the tables given below.

Before feature selection (RMSE)

Model Name	RMSE
Ridge Regression	0.4579
Lasso Regression	0.7797
Decision Tree	1.6424
K Nearest Neighbors	12.2258
Support Vector Regressor	0.1205

Table 1: Before feature selection (RMSE)

Before feature selection (MAE)

Model Name	MAE
Ridge Regression	0.3193
Lasso Regression	0.5792
Decision Tree	0.9515
K Nearest Neighbors	8.6444
Support Vector Regressor	0.0875

Table 2: Before feature selection (MAE)

After feature selection (RMSE)

Model Name	RMSE
Ridge Regression	0.4402
Lasso Regression	0.7797
Decision Tree	1.6616
K Nearest Neighbors	11.4251
Support Vector Regressor	0.1001

Table 3: After feature selection (RMSE)

After feature selection (MAE)

Model Name	MAE
Ridge Regression	0.2958
Lasso Regression	0.5792
Decision Tree	0.9462
K Nearest Neighbors	8.8314
Support Vector Regressor	0.0680

Table 4: After feature selection (MAE)

These are the tables consisting results of all the models we have used and different evaluating metrics we used to conduct our research. Among all the models Support vector regressor performed best both before and after feature selection.

5 Discussion

For our research we used Root mean square error and Mean absolute error as evaluating metrics and we used ridge regression, lasso regression, decision tree, k nearest neighbors and support vector regressor . We calculated the results both before and after feature selection. From the result we can see that Support vector regressor performed best among all the machine learning models we have used and gave better result after feature selection .

Given chart describes and differentiate between before and after feature selection.

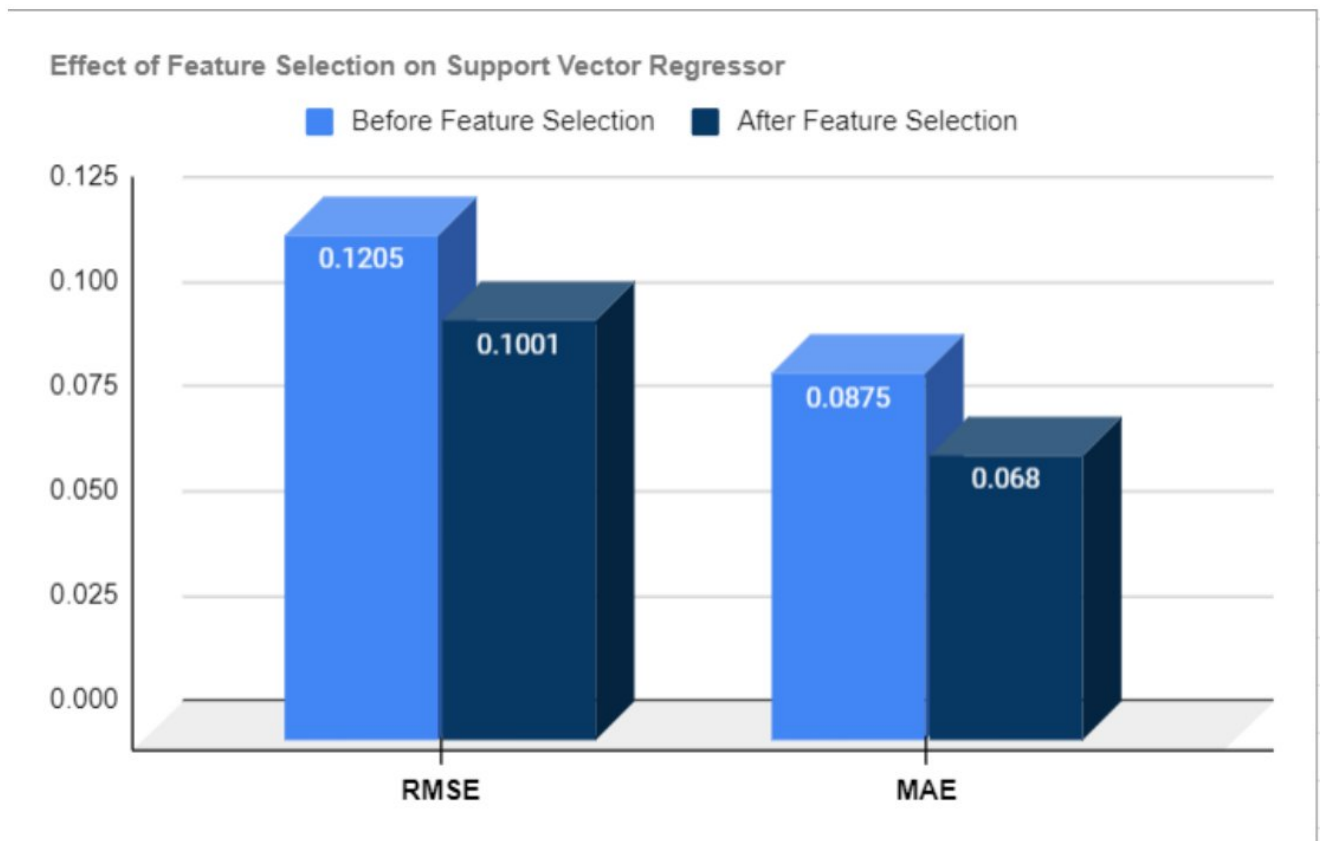


Figure 14: Effects of Feature Selection

6 CONCLUSION & FURTHER RECOMMENDATION

6.1 CONCLUSION

It is possible to draw the following conclusion from all of the research:

- Improved Feature Selection by identification of significant geotechnical parameters for Shear Strength prediction using statistical data mining techniques.
- Proposed integration of Support Vector Regressor as the most suitable machine learning algorithm for Shear Strength parameter prediction.
- Achieved better performance with simple and more interpretable machine learning methods to resolve the Blackbox issue in data science.
- Machine learning techniques can be used to correlate and predict soil parameters.
- The use of advanced prediction models will reduce the number of soil tests required, saving time and money.

6.2 FUTURE RECOMMENDATION

As a supplementary piece of advice, it should be mentioned that

- More data collection to support better prediction.
- Intensive feature selection through statistical and machine learning based analysis.
- Integration of Artificial Neural Networks to improve the predictions.
- Experimentation with a large range of machine learning models adopting the Geotechnical Engineering domain.
- Using machine learning, this research will provide a foundation for future geotechnical engineering research.
- Various data augmentation techniques can be used to alleviate future data scarcity.

7 REFERENCES

- Al-Kahdaar, R.M. and Al-Ameri, A.F.I., 2010. Correlations between physical and mechanical properties of Al-Ammarah soil in Messan Governorate. *Journal of Engineering*, 16(4), pp.5946- 5957
- Bui, D.T., et al.. A swarm intelligence-based machine learning approach for predicting soil shear strength for road construction: a case study at Trung Luong National Expressway Project (Vietnam). *Engineering with Computers*, Springer-Verlag London Ltd., Springer Nature 2018. <https://doi.org/10.1007/s00366-018-0643-1>
- Cao, M.T., et al.. An advanced meta-learner based on artificial electric field algorithm optimized stacking ensemble techniques for enhancing prediction accuracy of soil shear strength. *Engineering with Computers*, The Author(s) 2020. <https://doi.org/10.1007/s00366-020-01116-6>
- Chauhan, N.K. and Singh, K., 2018, September. A review on conventional machine learning vs deep learning. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 347-352). IEEE.
- Kamal, M.A., Arshad, M.U., Khan, S.A. and Zaidi, B.A., 2016. Appraisal of geotechnical characteristics of soil for different zones of Faisalabad (Pakistan). *Pakistan Journal of Engineering and Applied Sciences*.
- Lindvall, M., Molin, J. and Löwgren, J., 2018. From machine learning to machine teaching. *Interactions*, 25(6), pp.52-57.
- Martens, B., 2018. The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning. *SSRN Electronic Journal*,.
- Moayedi, H., et al.. Spotted Hyena Optimizer and Ant Lion Optimization in Predicting the Shear Strength of Soil. *Appl. Sci.* 2019, 9, 4738; doi:10.3390/app9224738, www.mdpi.com/journal/applsci
- Naej, M., Naej, M., Salehi, J. and Rahimi, R., 2016. Hydraulic conductivity prediction based on grain-size

distribution using M5 model tree. *Geomechanics and Geoengineering*, 12(2), pp.107-114.

Nakhforoosh, A., Nagel, K.A., Fiorani, F. and Bodner, G., 2021. Deep soil exploration vs. topsoil exploitation: distinctive rooting strategies between wheat landraces and wild relatives. *Plant and soil*, 459(1), pp.397-421.

Ngah, S.A. and Nwankwoala, H.O., 2013. Evaluation of sub-soil geotechnical properties for shallow foundation design in onne, Rivers state, Nigeria. *The International Journal of Engineering and Science (IJES)*, 2, pp.8-15.

Pham, B.T., et al.. Prediction of shear strength of soft soil using machine learning methods, 2018. *Catena* 166, 181–191. <https://doi.org/10.1016/j.catena.2018.04.004>

Ramabodu, M. and Verster, J., 2013. Factors that influence cost overruns in South African publicsector mega-projects. *International Journal of Project Organisation and Management*, 5(1/2), p.48.

Teng, W., 1983. *Foundation design*. New Delhi: Prentice-Hall.

Aljanabi, Q., Chik, Z., Allawi, M., El-Shafie, A., Ahmed, A. and El-Shafie, A., 2017. Support vector regression-based model for prediction of behavior stone column parameters in soft clay under highway embankment. *Neural Computing and Applications*, 30(8), pp.2459-2469.

Al-Kahdaar, R.M. and Al-Ameri, A.F.I., 2010. Correlations between physical and mechanical properties of Al-Ammarah soil in Messan Governorate. *Journal of Engineering*, 16(4), pp.5946- 5957

Anonymous, 2016. Peer review report 1 on “Evaluation of a random displacement model for predicting particle escape from canopies using a simple eddy diffusivity model”. *Agricultural and Forest Meteorology*, 217, p.290.

Austin, P., 2007. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, 26(15), pp.2937-2957.

Breiman, L., 1991. Discussion: Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1).

Chauhan, N.K. and Singh, K., 2018, September. A review on conventional machine learning vs deep learning. In *2018 International Conference on Computing, Power and Communication Technologies*

(GUCON) (pp. 347-352). IEEE.

CVS, R. and Pardhasaradhi, N., 2018. Analysis of Artificial Neural-Network. International Journal of Trend in Scientific Research and Development, Volume-2(Issue-6), pp.418-428.

Goh, A. and Goh, S., 2007. Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. Computers and Geotechnics, 34(5), pp.410-421.

Grömping, U., 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63(4), pp.308-319.

Hamidi, O., Tapak, L., Abbasi, H. and Maryanaji, Z., 2017. Application of random forest time series, support vector regression and multivariate adaptive regression splines models in prediction of snowfall (a case study of Alvand in the middle Zagros, Iran). Theoretical and Applied Climatology, 134(3-4), pp.769-776.

Ma, G., Chao, Z., Zhang, Y., Zhu, Y. and Hu, H., 2018. The application of support vector machine in geotechnical engineering. IOP Conference Series: Earth and Environmental Science, 189, p.022055.

Madhyannapu, R.S., Puppala, A.J., Hossain, S., Han, J. and Porbaha, A., 2006. Analysis of geotextile reinforced embankment over deep mixed soil columns: using numerical and analytical tools. In GeoCongress 2006: geotechnical engineering in the information technology age (pp. 1-6).

MAKOTO, K. and KHANG, T.T., Relationships between N value and parameters of groundstrength in the South of Vietnam.

Martens, B., 2018. The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning. SSRN Electronic Journal,.

Naeef, M., Naeef, M., Salehi, J. and Rahimi, R., 2016. Hydraulic conductivity prediction based on grain-size distribution using M5 model tree. Geomechanics and Geoengineering, 12(2), pp.107-114.

Pal, M. and Deswal, S., 2009. M5 model tree based modelling of reference evapotranspiration. Hydrological Processes, 23(10), pp.1437-1443.

Pirnia, P., Duhaime, F. and Manashti, J., 2018. Machine learning algorithms for applications in geotechnical engineering. Geo Edmonton, pp.1-7.

- Ramabodu, M. and Verster, J., 2013. Factors that influence cost overruns in South African public sector mega-projects. *International Journal of Project Organisation and Management*, 5(1/2), p.48.
- Shaha, N.R., 2013. Relationship between penetration resistance and strength compressibility characteristics of soil.
- Shahin, M.A., Jaksa, M.B. and Maier, H.R., 2001. Artificial neural network applications in geotechnical engineering. *Australian geomechanics*, 36(1), pp.49-62.
- Shooshpasha, I., Amiri, I. and MolaAbasi, H., 2015. AN INVESTIGATION OF FRICTION ANGLE CORRELATION WITH GEOTECHNICAL PROPERTIES FOR GRANULAR SOILS USING GMDH TYPE NEURAL NETWORKS (RESEARCH NOTE).
- Singh, B., Sihag, P. and Singh, K., 2017. Modeling of impact of water quality on infiltration rate of soil by random forest regression. *Modeling Earth Systems and Environment*, 3(3), pp.999-1004.
- Solomatine, D. and Xue, Y., 2004. M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. *Journal of Hydrologic Engineering*, 9(6), pp.491-501.
- Teng, W., 1983. *Foundation design*. New Delhi: Prentice-Hall.
- Yan, Q., Guo, M. and Jiang, J., 2011. Study on the Support Vector Regression Model for Order's Prediction. *Procedia Engineering*, 15, pp.1471-1475.
- Yin, Z.Y., Jin, Y.F., Huang, H.W. and Shen, S.L., 2016. Evolutionary polynomial regression-based modelling of clay compressibility using an enhanced hybrid real-coded genetic algorithm. *Engineering Geology*, 210, pp.158-167.
- Zhang, H., 2014. A Random Forest Approach to Model-based Recommendation. *Journal of Information and Computational Science*, 11(15), pp.5341-5348.
- Zhang, W. and Goh, A., 2013. Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics*, 48, pp.82-95.
- Zou, K.H., Tuncali, K. and Silverman, S.G., 2003. Correlation and simple linear regression. *Radiology*, 227(3), pp.617-628.

