

Establishing the co-relations between different geotechnical parameters of Bangladesh coastal soil using machine learning techniques

Tashfiqur Rahman Shakeen

Md. Latul Ibn Amin

Fayaz Rohan



**Department of Civil and Environmental Engineering
Islamic University of Technology
2021**



**Establishing the co-relations between different
geotechnical parameters of Bangladesh coastal soil
using machine learning techniques**

Tashfiqur Rahman Shakeen (160051028)

Md. Latul Ibn Amin (160051025)

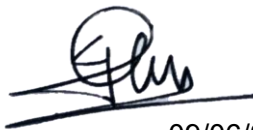
Fayaz Rohan (160051072)

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

**Department of Civil and Environmental Engineering
Islamic University of Technology
2021**

APPROVAL

This is to certify that the dissertation entitled “**Establishing the co-relations between different geotechnical parameters of Bangladesh coastal soil using machine learning techniques**”, by Tashfiqur Rahman Shakeen, Md. Latul Ibn Amin, and Fayaz Rohan has been approved fulfilling the requirements for the Bachelor of Science Degree in Civil & Environmental Engineering.



09/06/2021

Supervisor:

Istiakur Rahman

Assistant Professor

Department of Civil & Environmental Engineering

Islamic University of Technology (IUT)

Board Bazar, Gazipur-1704, Bangladesh.

DECLARATION

We declare that the undergraduate research work reported in this thesis has been performed by us under the adept supervision of Assistant Professor Istiakur Rahman. We have taken appropriate precautions to ensure that the work is original. We can corroborate that the work has not been plagiarized. We can also make sure that the work has not been published for any other purpose (except for publication).

Tashfiqur Rahman Shakeen

Student Id. 160051028

June, 2021

Md. Latul Ibn Amin

Student Id. 160051025

June, 2021

Fayaz Rohan

Student Id. 160051072

June, 2021

ACKNOWLEDGEMENT

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

All praises belong to the Almighty Allah (SWT) for giving us the strength and courage to successfully complete our Undergraduate Thesis. We would like to extend our special gratitude to our parents for being the constant source of inspiration and support.

We would like to express our heartfelt gratitude and sincere appreciation to our Supervisor **Mr. Istiakur Rahman**, Assistant Professor, Department of Civil and Environmental Engineering, Islamic University of Technology, for his gracious guidance, adept advice and continuous encouragement in supervising us. His technical and editorial advice was essential for the completion of this academic research. Without his assistance and guidance, the paper would have never been accomplished.

Furthermore, we would like to extend our gratitude to all the faculty members for their thoughtful recommendations during our study. We are thankful to our friends, juniors, seniors, and batch mates of our departments for their valuable suggestions and cordial assistance.

We would also like to thank Bangladesh Water Development Board and DevConsultants Ltd. for helping us by providing geotechnical soil investigation data of coastal region of Bangladesh which was required for our research project to complete it.

DEDICATION

We devote this thesis to our parents, who for many years sacrificed their valuable time, livelihood and effort to ensure that we could be who we are today. They motivated and supported us to follow our passion of engineering without ever looking back. Our parents have taught us that perseverance is one of the key traits to achieve success in life. We will forever be indebted to them.

ABSTRACT

The digital revolution is currently leaving no sector untouched. The combination of data and digital technologies opens up a multitude of opportunities in the geotechnical sector and Machine Learning is undeniably one of the most innovative applications in predicting soil parameters. Even so, owing to the uncertainty regarding the accuracy of the prediction models, traditional methods of determining soil parameters are still being used, which are both costly and time consuming. The purpose of this study is to correlate the different soil parameters of Bangladesh coastal soil, such as SPT N value, Shear wave velocity, Fine Content, Cohesion, and stiffness, and then use the correlation to predict the Angle of Friction. To predict the angle of friction of the coastal soil, six machine learning techniques were used: Simple linear regression model, Multi polynomial Regression, Support Vector Regression, Random Forest, Multivariate Adaptive Regression Splines, M5 Model Tree, and Artificial Neural Network. About 58 data sets were collected and used for this research project. Among 58 data sets, 48 were used to correlate the soil parameters and 10 data sets were used for testing and validation. Furthermore, all the machine learning methods were compared in terms of prediction accuracy. Finally, a validation of the predicted result has been conducted using PLAXIS 2D Software. In general, the Random Forest and M5 model tree regression models generated the best results, as the R^2 value (97.95% & 95.23% respectively) is the highest among all of the models and the error-values are also lower, reflecting better accuracy. Moreover, it is more evident from the study that conventional machine learning technique shows better performance than ANN where there is data scarcity.

Keywords: Conventional machine learning, prediction of Soil parameter, Simple linear regression model, Multi polynomial Regression, Support Vector Regression, Random Forest, Multivariate Adaptive Regression Splines, M5 Model Tree, and Artificial Neural Network (ANN), SPT N Value, Shear wave velocity, Fine Content, Cohesion, Stiffness, Angle of Friction, PLAXIS 2D, Embankment model.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1	GENERAL	1
1.2	STATISTICAL PREDICTION MODELS	2
1.3	USE OF STATISTICAL PREDICTION MODELS IN PREDICTING GEOTECHNICAL PARAMETERS	2
1.4	IMPORTANCE OF THIS STUDY.....	3
1.5	AIMS OF THIS STUDY.....	3
1.6	OBJECTIVES	3
1.7	WHY THE STUDY IS DIFFERENT FROM OTHERS	4
1.8	RESEARCH FLOW DIAGRAM.....	4
2	LITERATURE REVIEW	5
2.1	SUB-SOIL EXPLORATION METHODS.....	5
2.2	ESTABLISHING THE CORRELATION BETWEEN SOIL PARAMETERS TO PREDICT	6
2.3	APPLICATION OF MACHINE LEARNING IN GEOTECHNICAL ENGINEERING	6
2.4	MACHINE LEARNING TECHNIQUES SELECTED TO APPLY IN THIS STUDY FOR PREDICTION.....	7
2.4.1	SIMPLE LINEAR REGRESSION	8
2.4.2	MULTI POLYNOMIAL REGRESSION.....	9
2.4.3	SUPPORT VECTOR REGRESSION (SVM)	9
2.4.4	RANDOM FOREST	10
2.4.5	MULTIVARIATE ADAPTIVE REGRESSION SPLINES	11
2.4.6	M5 MODEL TREE.....	11
2.4.7	ARTIFICIAL NEURAL NETWORK (ANN).....	12
2.5	PLAXIS FOR VALIDATION	13

2.6	SUMMARY OF LITERATURE REVIEW	13
3	METHODOLOGY.....	14
3.1	STUDY AREA.....	15
3.2	DATA COLLECTION PROCEDURE.....	17
3.3	PREPARATION OF TEST DATA & TRAIN DATA	18
3.4	RESEARCH METHODOLOGY	20
4	RESEARCH ANALYSIS	21
4.1	SIMPLE LINEAR REGRESSION	21
4.2	MULTI POLYNOMIAL REGRESSION	22
4.3	SUPPORT VECTOR REGRESSION.....	24
4.4	RANDOM FOREST	25
4.5	MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)	27
4.6	M5 MODEL TREE	29
5	RESULTS & DISCUSSION.....	31
5.1	RESULT SUMMERY.....	31
5.2	ACTUAL VALUE VS PREDICTED VALUE.....	32
5.2.1	SIMPLE LINEAR REGRESSION	32
5.2.2	MULTI-POLYNOMIAL REGRESSION.....	34
5.2.3	SUPPORT VECTOR REGRESSION	35
5.2.4	RANDOM FOREST REGRESSION	37
5.2.5	MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS).....	38
5.2.6	M5 MODEL TREE.....	40
5.2.7	ARTIFICIAL NEURAL NETWORK	41
5.2.8	ORIGINAL VS PREDICTED (ALL MODELS)	42
5.3	R ² VALUE COMPARISON\.....	44
5.4	ERROR VALUE COMPARISON.....	45
5.5	CONCLUSION OF THE RESULTS	45
6	RESULT VALIDATION.....	46

6.1	DESIGN DATA USED FOR MODELING.....	46
6.2	PLAXIS MODELLING	47
6.3	DIFFERENCE OF SETTLEMENT VALUE	48
7	CONCLUSION & FURTHER RECOMMENDATION	49
7.1	CONCLUSION	49
7.2	FURTHER RECOMMENDATION	49
8	REFERENCES.....	50

LIST OF TABLES

Table 1 : Train Data	18
Table 2: Test data.....	19
Table 3 : Result Summary of all machine learning models	31
Table 4: Actual Value vs Predicted value in Simple Linear Regression Model.....	32
Table 5 : Actual Value vs Predicted Value in Multi-polynomial Regression.....	34
Table 6 : Actual Value vs Predicted Value in Support Vector Regression.....	36
Table 7: Actual Value vs Predicted Value in Random Forest Regression model.....	37
Table 8 : Actual Value vs Predicted Value in Multivariate Adaptive Regression Splines model.....	39
Table 9 : Actual Value vs Predicted Value in M5 Model Tree.....	40
Table 10: Actual Value vs Predicted Value in Artificial Neural Network	41
Table 11: Original vs Predicted Friction Value for All Models	42
Table 12: Design Data for Embankment Modelling in PLAXIS 2D	46
Table 13 : Difference in Settlement Value for all models in PLAXIS	48

LIST OF FIGURES

Figure 1: Research flow diagram source : Author	4
Figure 2 : Categorization of machine learning techniques Source: Author	8
Figure 3: Location-01 of collecting soil data	15
Figure 4: Location 02 of collecting soil data	16
Figure 5: Location 03 of collecting soil data	16
Figure 6: Python programming language for simple linear regression.....	21
Figure 7: Regression graph for simple linear regression	22
Figure 8: Python programming language for Multi polynomial Regression	23
Figure 9: Regression graph for Multi Polynomial Regression	23
Figure 10: Python programming language Support Vector Regression	24
Figure 11: Regression graph for Support Vector Regression Model.....	25
Figure 12: Python programming language Random Forest Model.....	26
Figure 13 : Regression Graph for Random Forest Model.....	26
Figure 14: Python Programming Language for Multivariate Adaptive Regression Splines Model ..	27
Figure 15 : Regression Graph for Multivariate Adaptive Regression Splines Model	28
Figure 16: Python Programming Language M5 Model Tree.....	29
Figure 17: Regression Graph for M5 Model Tree	30
Figure 18 : Original vs Predicted graph in Simple Linear regression Model	33
Figure 19 : Original vs Predicted graph in Multi-polynomial Regression.....	35
Figure 20 : Original vs Predicted graph in Support Vector Regression Model	36
Figure 21: Original vs. Predicted graph in Random Forest Regression model.....	37
Figure 22: Original vs Predicted graph in Multivariate Adaptive Regression Splines model.....	37
Figure 23 : Original vs Predicted graph in M5 Model Tree.....	37
Figure 24 : Actual vs Predicted Friction value Graph for all models	37
Figure 25 : R2 Value Comparison for All Models.....	37
Figure 26 : Error Value Comparison for All models	37
Figure 27 : Embankment Modeling for Actual Friction Value.....	37
Figure 28 : Embankment Modeling for Predicted friction Values.....	37

1 INTRODUCTION

1.1 GENERAL

Soil parameters play a vital role in designing and estimating the foundation stability of the structure. As the stability of a structure is fully dependent on the strength of soil, it is important to find out the strength of the soil before any design and construction of a structure. Soil strength mainly depends on the properties of the soil. Generally, soil exhibits two kinds of properties which are physical property and mechanical property. Physical properties like moisture content, liquid limit, plastic limit, void ratio, etc. helps to provide information about the type and nature of the soil whereas mechanical properties like unconfined compression strength, standard penetration value helps to estimate the bearing capacity of the soil. Among the mechanical properties of soil, cohesion and angle of internal friction the two internal parameters of soil that are directly related and used to calculate the shear strength of the soil. Therefore, it is essential to find out the cohesion and angle of friction of a certain soil before starting any construction of structure. But to know exactly about these parameters, a sub-soil investigation of a certain soil must be operated which is expensive and also time-consuming. This research is trying to use various statistical models and machine learning techniques to develop a model that can predict a parameter (angle of friction) of a certain area (coastal) of soil with the help of correlating the parameters of already existing data.

Sub-soil exploration is required to find out the strength and bearing capacity of soil and thus type and depth of foundation are calculated. A usual method of sub-soil exploration operation consists of three steps namely boring, sampling, and testing. First, a test hole is created and then a soil sampler (split-spoon, Shelby tube) is driven into the soil to take the samples. After collecting the samples, it is taken to the laboratory to observe and examine the sample for performing the various test and finding the parameters of the soil. This procedure is expensive and time-consuming to operate. Moreover, there are some similarities and correlation exist for a certain area of a soil. Therefore, this study is intended to find out the similarities and correlation of a coastal area soil and with the help of these correlation and machine learning techniques, a model will be developed which will predict the soil parameters using these machine learning models.

1.2 STATISTICAL PREDICTION MODELS

Statistical prediction models are those which utilize techniques such as machine learning to predict what might happen next. Machine learning is the method of teaching a computer to think like a human. Machine learning models are mainly used to make a prediction based on historical data. The performance of machine learning model's performance is assessed by observing how well it predicts by new data that hasn't been learned yet. Predictive analytics algorithms can be divided into two groups: Machine learning and deep learning. Machine learning algorithms comprise both linear and non-linear data and deep learning models are a subset of machine learning that is more popular to deal with audio, video, text, and images. By using these statistical predictive models, the cost can be reduced drastically but it doesn't mean benefits appear aimlessly, even predictive modelling shows the number of challenges also. Large and comprehensive data are tough to handle and often data needs cleansing.

1.3 USE OF STATISTICAL PREDICTION MODELS IN PREDICTING GEOTECHNICAL PARAMETERS

Statistical Prediction models have been used to develop correlations for predicting geotechnical parameters for Civil Engineering design. Machine learning techniques can predict the fairly accurate value of various geotechnical parameters. In geotechnical engineering, empirical correlations are used to evaluate various properties of soil. Correlations are developed with the help of statistical methods using data from extensive laboratory or field testing. The conventional machine learning techniques used in this study are Linear Regression (LR) Analysis, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and M5 model trees (M5P). These models learn from the data that are presented to them and try to find a pattern among the dataset even if the fundamental relationships are unknown or the physical meaning is tough to explain. Machine learning is well suited for most Geotechnical Engineering materials because it doesn't need any prior information about the data.

1.4 IMPORTANCE OF THIS STUDY

In the 21st century, machine learning is a modern popular trend that aims to emulate human intelligence into machines and transform it to be more efficient and accurate. While artificial intelligence (AI) is the broad science of mimicking human abilities, machine learning is a specific subset of AI that trains a machine how to learn. So far, machine learning is much better, accurate, time & cost convenient in terms of the application of Artificial Intelligence. Especially, it can show a great deal of potentiality if it is applied in Geotechnical Engineering. Besides, soil tests are expensive and time-consuming. Therefore, if this research can correlate the soil parameters and predict the results through machine learning, it will be able to discover new possibilities for geotechnical design.

1.5 AIMS OF THIS STUDY

The main aim of this research is to correlate the coastal soil parameters and predict them through machine learning and also to compare and find out the most effective machine learning techniques. The parameter angle of friction (ϕ) is selected to predict in this research using the correlation and the machine learning techniques.

1.6 OBJECTIVES

The research will try to achieve the following objectives-

- To find, extract and summarize the relevant coastal soil data.
- To establish the correlation between soil parameters.
- To use machine learning techniques in order to predict parameters.
- To compare the techniques and find out the most effective one.
- To save a great amount of time and money by avoiding excessive soil tests.

1.7 WHY THE STUDY IS DIFFERENT FROM OTHERS

In previous, many researches were conducted to correlate the parameters and derive formulas with simple statistical equations but in this study, it is intended to apply advanced machine learning techniques to get more accurate prediction results. This study is also intended to show a comparison result among all the conventional machine learning and deep learning techniques and find the most accurate one based on their accuracy.

1.8 RESEARCH FLOW DIAGRAM

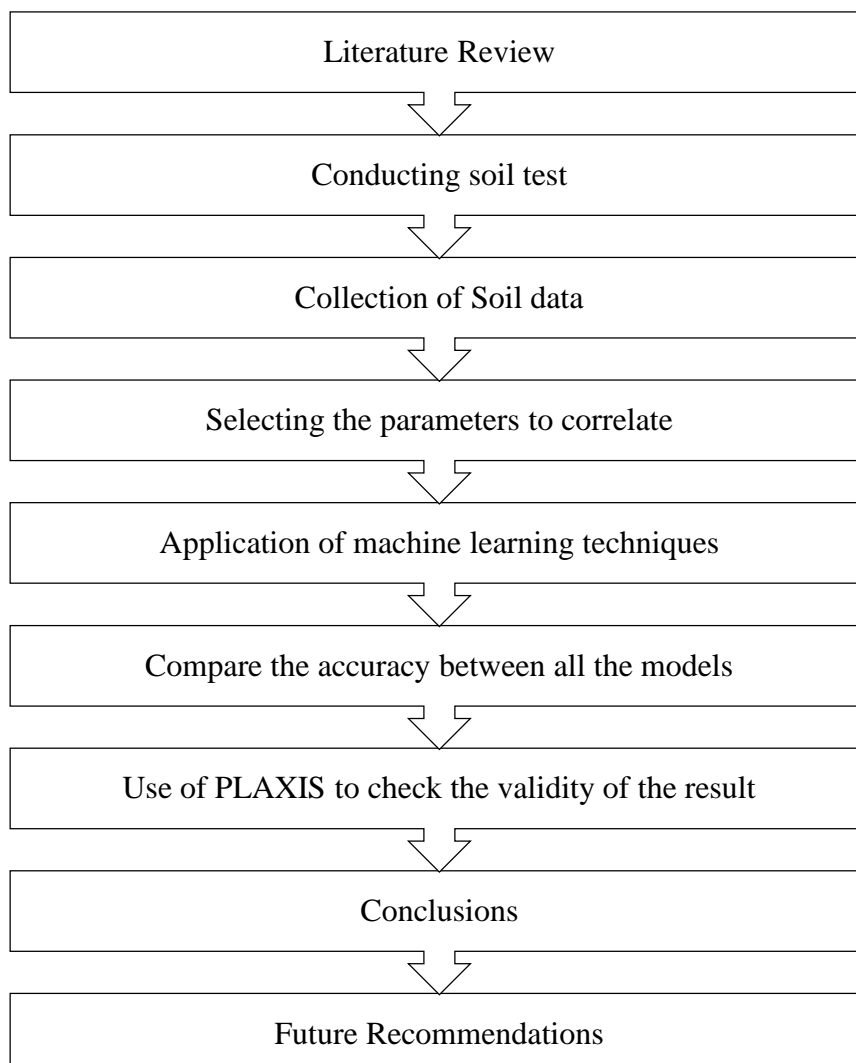


Figure 1: Research flow diagram source: Author

2 LITERATURE REVIEW

The main focus of this study is to predict soil parameters using machine learning techniques and apply the predicted parameters in geotechnical design consideration. Therefore, in this chapter, related literature in regard to the study has been discussed in details. The total literature review is divided into following sections where sections are categorized based on the main objectives and outcomes of this research.

2.1 SUB-SOIL EXPLORATION METHODS

Time and cost are the two vital considerations while doing sub-soil exploration process. According to Teng (1983), some standard rules are maintained all over the world for operating the sub-soil exploration before any construction process and the three common steps boring, sampling, and testing are commonly followed. There are different types of boring suitable for different types of soil collection. Nakhforoosh (2021) described about the vital point of a sub-soil exploration program includes location, depth of borings, test pits and the methods of boring chosen for that sub-soil exploration. The study described about the factors that to be considered before a sub-soil exploration which are stratification & engineering properties of soil underlying the site, shear strength, deformation and hydraulic characteristics of soil. Another study was conducted by Kamal (2016) where the main objective of a sub-soil exploration is considered to collect maximum amount of information at minimum cost. Ngah (2013) described about four steps of a soil exploration program which are collecting information, reconnaissance, preliminary exploration and detailed exploration. To complete the sub-soil exploration program, these four steps is required to complete which takes a good amount of time & money. When it comes about any mega construction process, sub-soil exploration is required to do between the short spacing distance. Supporting this doctrine, Ramabodu et al. (2013) also explained that during a mega construction project a huge number of soil tests in similar types of soil to know the parameters. This study also concluded that if soil parameters of a certain geological location can be correlated and used to predict the parameters in a similar geological location with the help of a highly accurate statistical prediction process, a huge number of soil test, cost, and time can be saved during the mega construction project.

2.2 ESTABLISHING THE CORRELATION BETWEEN SOIL PARAMETERS TO PREDICT

To predict parameters and save, time & cost, many researches has been conducted since long ago. Kahdaar et al. (2010) described the correlation between different physical properties (LL, PI, LI, water content, density, void ratio) with different mechanical properties (q_u , C_c , C_s , SPT) by using the simple linear regression model. The study concluded by using the correlation with some information, preliminary investigation stages and studies of any structure can be performed to find indicative design parameters of soil. Shaha (2013) determined the compressibility properties of soil by modified Shelby tubes (automatic SPT hammer) and correlated the properties with standard penetration resistance. In this study a linear relationship between unconfined compressive strength, dry density, initial void ratio with SPT N value as well as a relationship between compression index, liquid limit and initial void ratio has been established. It is concluded from the study that except for these correlations there is always possibility of developing correlation using other parameters of soil. Another study was conducted by Makoto et al. (2013) where the study delineated the procedure to determine the SPT N value using the other design parameters of soil. In that study, it was experimented with the formula of correlation and compared the estimated N value with the actual N value in different sites. The correlation between friction angle Depth, Compressive strength, water content with SPT N value has been established in that study. But at the end of the analysis of the study, it was concluded that the application of the generalized formula of one area in another area is hard as correlated formula in Japan was not sufficiently effective in the Ho Chi Minh of Vietnam.

It is generalized from the literature reviews regarding the establishment of the correlation between soil parameters that for a certain geological location, correlation can be developed between the soil parameters using existing soil data and these correlations can also be used to predict and calculate parameters for that certain geological location. Hence in this research, it was chosen the coastal area of Bangladesh as a particular geological condition to establish the correlation between soil parameters.

2.3 APPLICATION OF MACHINE LEARNING IN GEOTECHNICAL ENGINEERING

Machine Learning is undeniably one of the most innovative applications in geotechnical engineering. Several researches have been conducted to check the efficacy of machine learning techniques. Martens (2018) defined machine learning as the practice of using algorithms to parse data, learn from it, and then decide or predict about something in the world. Moreover, the nearly limitless quantity of available data, affordable data storage, and the growth of less expensive and more powerful processing has propelled the growth of machine learning. Lindvall et al. (2018) supported the previous statement and added that many industries have developed more robust machine learning models capable of analysing bigger and more complex data while delivering faster, more accurate results on vast scales and machine learning tools enable organizations to more quickly identify profitable opportunities and potential risks. Hamidi et al. (2017) conducted a research to develop models that can predict the snowfall by using random forest time series, support vector regression, and multivariate adaptive regression splines models and also compared the effectiveness of the techniques. The study found effective results of predicting snowfall using these machine learning techniques. Pirnia et al (2018) applied the machine learning techniques in geotechnical engineering and generated a large dataset using numerical models for machine learning applications in geotechnical engineering. In this study a Discrete Element Modelling named YADE (DEM) code, was used to produce assemblages of spherical particles and the Microsoft Cognitive Toolkit (CNTK) was used to analyse the images and the percentages passing for five sieves. It is concluded from the study that Convolutional Neural Networks (Conv.Net) can predict the PSD with a Root Mean Square Error (RMSE) of around 4 %.

It is generalized from this section of literature reviews that machine learning has been used effectively in geotechnical research and the effectiveness of machines learning techniques is well documented. Therefore, in this research machine learning techniques are selected to use to predict the soil parameters.

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

There are many machine learning techniques that has been well established from a very long time and the dynamic field of machine learning has been changed frequently. Researchers have been

working relentlessly to develop new and upgraded models. According to Chauhan et al (2018), machine learning techniques can be categorized into two sections - conventional machine learning and deep learning. Among the large number of conventional machine learning technique and deep learning techniques, this study has chosen Multi polynomial Regression, Support Vector Regression (SVR), Random Forest, Multivariate Adaptive Regression Splines (MARS) and M5 Model Tree as conventional machine learning techniques and Artificial Neural Network (ANN) as deep learning techniques. The following **Figure 2** can explain the categorization of machine learning techniques –

The following **Figure 2** shows the categorization of machine learning techniques source.

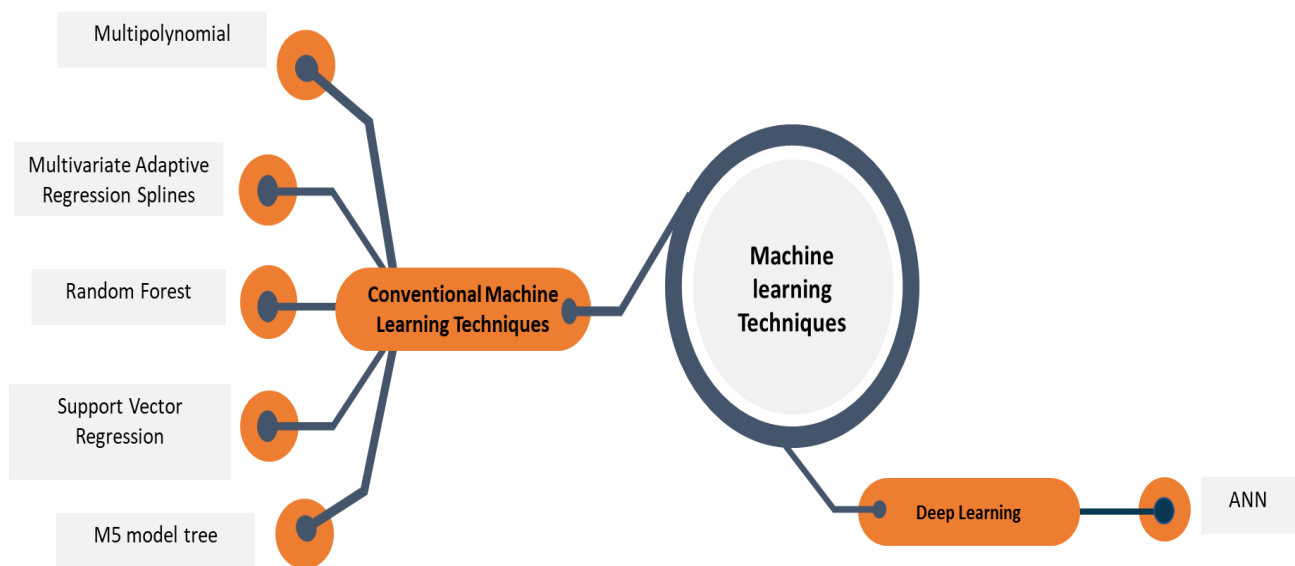


Figure 2 : Categorization of machine learning techniques Source: Author

2.4.1 SIMPLE LINEAR REGRESSION

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. Zou et al. (2003) described that a simple linear regression model can be used to analyse the specific relationship between the two or more variables and they also added that this model is mainly used to gain information about one through knowing values of the others. Kahdaar et al. (2010) used a simple linear regression model to correlate the physical and mechanical properties of soil. Moreover, Shaha (2013) also used the simple linear regression model to correlate between unconfined compressive strength, dry density, initial void ratio, compression index and liquid limit with SPT N value and found linear empirical relationship. Both researchers found effective results in correlating the parameters with a simple linear regression model. It can be generalized from the literature reviews about simple linear regression that all geological

locations may not show the linear relationship as the behaviour of soil comparing is always unpredictable one geological location to another.

2.4.2 MULTI POLYNOMIAL REGRESSION

The Multivariate Polynomial Regression is used for value prediction when there are multiple values that contribute to the estimation of values. These may be related to each other and can be converted to independent variable set which can be used for better regression estimation using feature reduction techniques. Khuri et al. (1981) described about the multi polynomial regression model comparing to the simple linear relationship, this model has some modification to gain more accuracy. It is observed from the study that the training data sets used in this model will have non-linearity in their relation after the correlation. Yin et al. (2016) conducted research on developing the compressibility of clay by using the polynomial model. In this study, multi-polynomial regression model showed the correlation between compressibility, void ratio, and plastic index with liquid limit. It is concluded from the study that polynomial regression showed better results with greater accuracy than the simple linear regression model in correlating these parameters.

2.4.3 SUPPORT VECTOR REGRESSION (SVM)

Support vector regression (SVR) is characterized by the use of kernels, sparse solution, and VC control of the margin and the number of support vectors. Although less popular than SVM, SVR has been proven to be an effective tool in real-value function estimation Yan et al. (2011) described in their study that the support vector regression model uses the basic idea of mapping the input vector of x space into a space with higher dimensions using an appropriate nonlinear kernel function, $\phi(x)$. This study focused that the kernel trick is the real strength of SVM and an appropriate kernel function can solve any complex problem. Aljanabi et al. (2017) conducted an experiment to show the stone column parameters in soft clay under highway embankment using the SVM model. This study showed that SVM can be equipped with the ability to generalize the statistical learning goal. Also, the SVM model for predicting liquefaction has been developed by using cone resistance. This study concluded that to achieve the best result of SVM regression, selection of tenfold cross-validation is required. Another research was conducted by Kirits et al. (2019) where the soil compressibility models were enhanced by the support vector machine, and later on, the performance of correlation was tested by

field verification in terms of settlement calculation. It was found from the study that SVM has shown better accuracy in both cases of correlation. Goh et al. (2007) demonstrated that the SVM has been successfully applied to assess liquefaction potential. This study has also demonstrated the viability of the SVM to model the complex relationship between the seismic and soil parameters, and the liquefaction potential using in situ measurements based on the CPT. Ma et al. (2018) have also concluded that SVM can be effectively used to classify soil, which is vital for geotechnical engineering and SVM can be used to establish models to forecast the deformation of rock and soil validly, as well as correctly predict the properties of rock.

All the literature reviews in this section have recommended SVM as a good performance showing model based on accuracy while doing the correlation.

2.4.4 RANDOM FOREST

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Zhang (2014) has described that random forest is a supervised learning algorithm that is flexible and an easy-to-use machine learning algorithm. Supporting these statements Follett (2016) emphasized the random forest model as a machine learning classifier that contains several decision trees and targets the class by individual trees. Singh et al. (2017) conducted a research study about modeling the impact of water quality on infiltration rate of soil where they used the random forest model in correlating the water quality parameters. The study showed that the random forest model can show a good performance based on accuracy in the case of correlating parameters. Another study was conducted by Grömping (2009) where the random forest model was compared with linear regression in the case of variable importance assessment. In this study, it was found that both models are similarly successful while doing the importance assessment but the random forest model showed slightly greater performance in respect to the accuracy and error.

This literature review section described the Random forest model that Random forest can show better performance in respect to correlate and predict with less error and greater accuracy.

2.4.5 MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Multivariate adaptive regression splines (MARS) is an algorithm that essentially creates a piecewise linear model which provides an intuitive stepping block into nonlinearity after grasping the concept of linear regression and other intrinsically linear models. Breiman (1991) stated in his research study about multivariate adaptive regression splines (MARS) that this model introduces new flexible modeling of high dimensional data and this model takes the form of an expansion in product spline basis functions. It was also described in the study that the number of basic functions, as well as the parameters associated with each one (product degree and knot locations), is automatically determined by the data. Another study was conducted by Austin, (2007) about the comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality where he stated about this model that this model provides a convenient approach to capture the nonlinear relationships in the data by assessing (knots) and also added that this model explains the complex nonlinear relationship of the inputs and the outcome variable. MARS was used in geotechnical engineering by Zhang et al. (2013) where the study has shown that this model can approximate the relationship between the inputs and outputs, and express the relationship mathematically. It was also mentioned that the main advantages of MARS are its capacity to produce simple, easy-to-interpret models, its ability to estimate the contributions of the input variables, and its computational efficiency.

Literature reviews from this section can conclude by comparing to regression models that MARS has also performed a good result with better accuracy to develop the relationship between input and output data.

2.4.6 M5 MODEL TREE

M5 model tree is a decision tree learner for regression task which is used to predict values of numerical response variable Y, which is a binary decision tree having linear regression functions at the terminal (leaf) nodes, which can predict continuous numerical attributes. Pal et al. (2009) researched about evapotranspiration system where the M5 model tree was used for prediction analysis. The study stated about this model that M5P is a decision tree learner for regression problems

and this tree algorithm assigns linear regression functions at the terminal nodes and fits a multivariate linear regression model to each subspace by classifying or dividing the whole data space into several subspaces. This study also added about this model that this M5 tree method deals with continuous class problems instead of discrete classes and can handle tasks with very high dimensionality. Solomatine et al. (2004) conducted a research on the application of flood forecasting in the upper reach of the Huai River in China and demonstrated that the M5 model tree showed a good result of accuracy in predicting the flood based on the known recent hydrological data. It was also concluded from the study that the conventional machine learning model has shown good performance with greater accuracy in predicting the flood using the correlation between hydrological data. Another study was conducted by Naeef et al. (2016) about predicting hydraulic conductivity prediction based on grain-size distribution using the M5 model tree where the study showed that the M5 model tree was easily used to represent the mathematical equations and performed the prediction analysis with less error value.

2.4.7 ARTIFICIAL NEURAL NETWORK (ANN)

Artificial neural networks (ANN) comprise hundreds or milliners of artificial neurons, which are interconnected by nodes, called processing units. All processing units consist of inputs and outputs. The input units obtain different information types and structures based on an internal weighting scheme and the neural network tries to understand the data provided in the development of a single output report. ANN needs rules and instructions for producing outcomes or outputs and ANN uses a range of learning rules called backpropagation to perfect their performance results. Shahin et al (2001) described the applications of ANN in solving geotechnical engineering problems and discussed the strengths and limitations of ANNs compared to the other modeling approaches. This study also applied the application of ANN in predicting pile capacity prediction, settlement of foundations, soil properties and behaviour, liquefaction, site characterization, earth retaining structures, slope stability, and the design of tunnels and underground openings. Also, the study concluded that the most successful and well-established applications are the capacity prediction of driven piles, liquefaction, and the prediction of soil properties and behaviour. Another study was conducted by Shooshpashaa et al (2015) related to the investigation of the correlation between SPT N value and soil properties such as effective stress, fine content, and moisture content by developing a polymer model based on the Group Method of Data Handling (GMDH) type Neural Network (NN). Approximately 195 data sets were used to develop the model and correlate the soil parameters. This study depicted an

identification system technique to develop correlations with soil geotechnical properties and assessed their influence on friction angle. The evolved GMDH type neural network was used to obtain a model for friction angle prediction.

2.5 PLAXIS FOR VALIDATION

PLAXIS is a computer programme that performs finite element analyses (FEA) within the realm of geotechnical engineering, including deformation, stability and water flow. The input procedures enable the enhanced output facilities provide a detailed presentation of computational results. PLAXIS enables new users to work with the package after only a few hours of training. Ponomarev et al. (2015) conducted a study about PLAXIS 2D modeling where the study stated about software that PLAXIS is a leading geotechnical engineering simulation software. This study also added about PLAXIS that which is renowned for ease of use and accuracy and this software can Optimize the designs more effectively than applying traditional conservative calculation methods. Many researches were conducted to check the validity of the result value using PLAXIS. Raja S. Madhyannapu et al conducted a research about the analysis of geotextile reinforced embankment over deep mixed soil columns where the study compared the settlement value of an actual embankment and predicted value experimented from finite element based PLAXIS program. This study concluded that the predicted settlement value from PLAXIS modeling is very close to the actual value and PLAXIS modeling can be used to verify the settlement value. Another study was conducted by Konyushkov (2020) about comparing the results of numerical modeling of slope stability in the PLAXIS program with analytical calculations where PLAXIS has performed better performance in comparing the predicted value and actual value of slope stability.

2.6 SUMMARY OF LITERATURE REVIEW

It can be generalized about all the literature reviews that the number of researches was conducted about correlating the soil parameters using different statistical prediction models is few in volume. About selecting the statistical models there are plenty of options to use in correlation. In this study, the literature review has picked the machine learning techniques that have already been used in predicting the value based on the input data either in geotechnical sectors or any other sectors. From the above literature review, conventional machine learning techniques and deep learning techniques

are selected to predict the soil parameters. The models that are selected to predict the soil parameters are –

- Simple linear regression model
- Multiple Polymer regression model
- Support Vector Regression Model
- Random Forest
- Multivariate Adaptive Regression Splines (MARS)
- M5 Model Tree
- Artificial Neural Network (ANN)

And from the literature review, this study has selected the parameters that will be used to build the correlation will be –

- SPT,
- Shear wave velocity,
- fine content
- cohesion and
- elasticity

and the parameter which will be predicted is the **Angle of friction**.

3 METHODOLOGY

In this section, the description of the study area, data collection methodology, and the preparation of the data for applying on the machine learning models have been described thoroughly.

3.1 STUDY AREA

In this study, the soil investigated the data were collected mainly from the coastal side of Bangladesh. The coastal region mostly covers 29,000 km² or about 20% of the country. Again, the coastal areas of Bangladesh cover more than 30% of the cultivable lands of the country. For this study purpose, the first location was selected at Upazila- Paikgacha, District – Khulna, the second location was selected at Upazila- Patharghata, District – Borguna, and lastly the third location was selected at Upazila – Bagerhat Sadar, District – Bagerhat. The following **Figure - 3, Figure-4, Figure-5** shows the geological location of the study area.

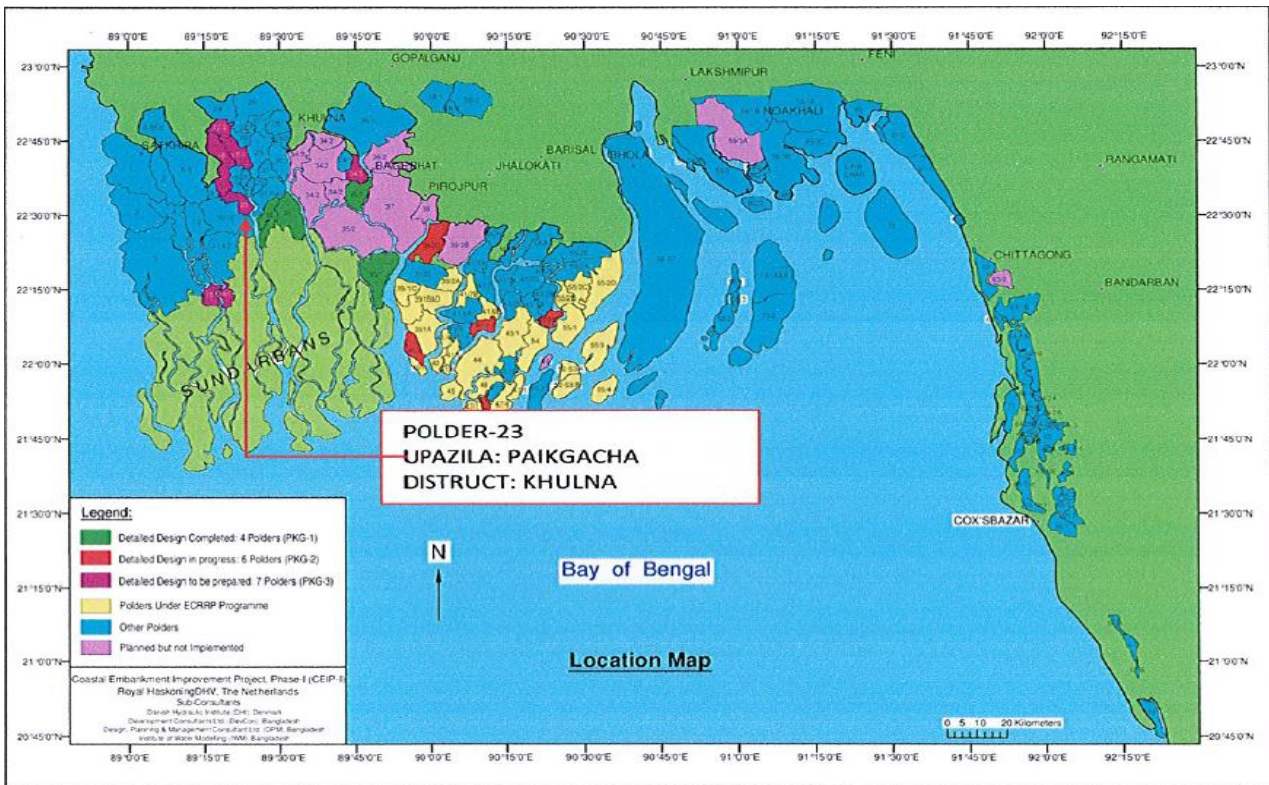


Figure 3: Location-01 of collecting soil data

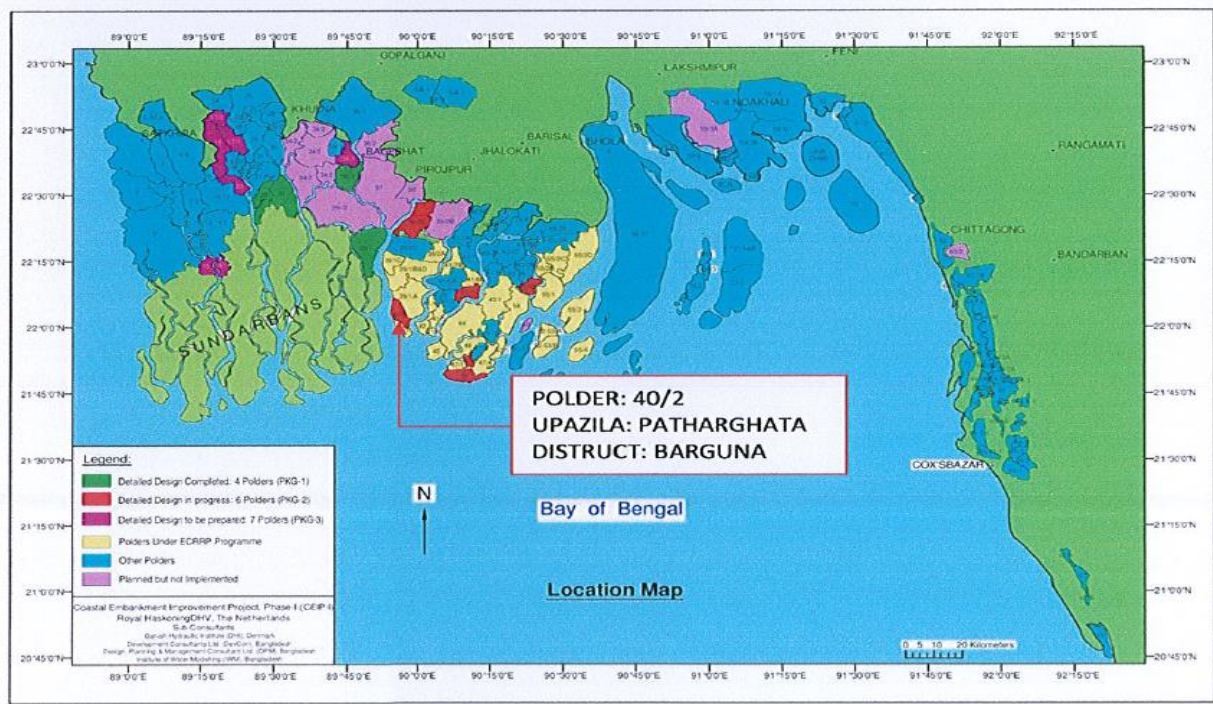


Figure 4: Location 02 of collecting soil data

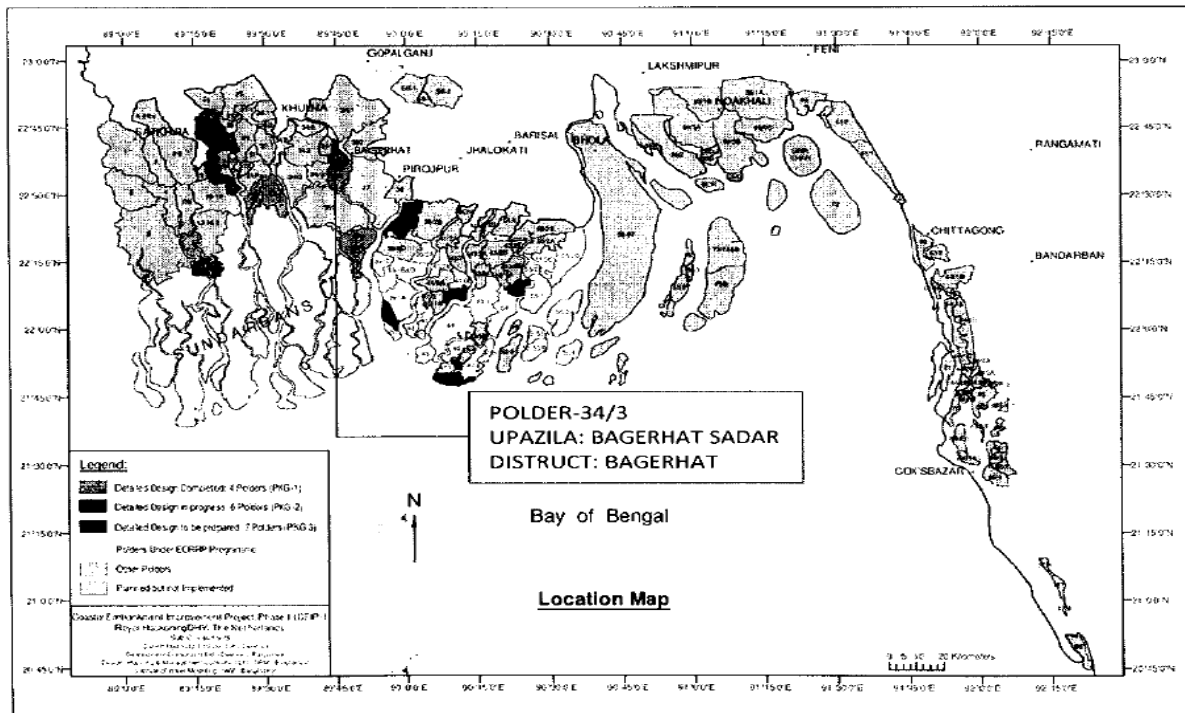


Figure 5: Location 03 of collecting soil data

The first location for collecting the soil data is situated at Upazila- Paikgacha, District – Khulna. There was in total three Drainage sluice where the soil data are collected. The Drainage sluice

01 is situated about 23.69m west side of the existing structure on Chok Horikhali Khal at Chok Horikhali, Paikgacha, Khulna under the Bangladesh Water Development Board. The drainage sluice 02 situated about 27.640m north-west side of the existing structure on Soladana Khal at Soladana, Paikgacha, Khulna under Bangladesh Water Development Board. The Drainage Sluice 03 is situated about 17.792m south-east side of the existing structure on Boroitola Khal at Boroitola, Paikgacha, Khulna under Bangladesh Water Development Board.

The second location for collecting the soil data is situated at Upazila- Patharghata, District – Borgia. There was in total two drainage sluice. The Drainage Sluice- 01 is situated about 23m north side of the existing structure on Nijladiamar Khal at Village-Tangrabazar, Upazila- Patharghata, Barguna. The drainage sluice-02 is situated about 16.713m north side of the existing structure on Adharia Khal at Village-Gaharpur, Upazila- Patharghata, Barguna. The third location from where soil data are collected is situated at Upazila – Bagerhat Sadar, District – Bagerhat. There was in total three drainage sluice from where soil data are collected. The Drainage Sluice-01 is situated about 21.830m north-west corner of the existing structure on Govordia Khal at village-Govordia, Bagerhat Sadar Upazila, Bagerhat under Bangladesh Water Development Board.

The Drainage Sluice-02 is situated about 21.170m north-east corner of the existing structure on Jiratola Khal at village-Jiraola, Bagerhat Sadar Upazila, Bagerhat under Bangladesh Water Development Board. The drainage sluice 03 is situated about 27.150m west side of the existing structure on Sosi Khal at village-Purba Sharia, Bagerhat Sadar, Bagerhat under Bangladesh Water development Board.

3.2 DATA COLLECTION PROCEDURE

The geotechnical investigation reports were collected from the data storage of the Bangladesh Water Development Board (BWDB). For the Coastal improvement embankment project CEIP (phase- I), Bangladesh Water Development Board (BWDB) gave the contract of preparing the geotechnical investigation report to DevConsultants Limited and Geo Profile soil investigation Limited. These professionals of soil investigators visited the coastal side of Bangladesh such as Khulna, Borguna, Bagerhat and collected all the forms of soil condition and other important information about that soil. These professional soil investigation companies used to collect the data by maintaining all the professional rules. Therefore, the data used in this study can be verified that all the data are error-free and suitable to use for research purposes.

3.3 PREPARATION OF TEST DATA & TRAIN DATA

In this study the prime source of the data was the Coastal improvement embankment project CEIP phase- I of Bangladesh Water Development Board (BWDB). Three locations from the district Khulna, Bogura and Bagerhat was selected as the study area for this research. In total there were 39 boreholes the number of total data set was 58. To conduct the research, the total data set were required to divide into 2. The larger portion of the data were used to train the research and the smaller portion of the data were used to check the result. In this study, among 58 data sets, the train data set were about the number of 48 and the test data were set about the number of 10.

The following **Table 1** and **Table 2** were used as train data and test data for this research.

Table 1 : Train Data

Train Data					
dependent	independent	independent	independent	dependent	independent
Angle of Friction	SPT N Value	Shear wave velocity	Fine Content	Cohesion	Elasticity
Y	X	X	X	X	X
7	11	159.3465	67	30	10000
10	36	301.8933	38	31	16500
8	11	187.8839	44	27	13000
7	23	252.3618	44	32	12000
8	10	156.7855	58	30	11600
7	9	154.0023	58	29	10500
8	10	156.7855	90	29	11000
7	9	173.3922	27	28	12000
9	13	163.9367	63	29	10500
9	15	212.7007	27	28	15000
5	13	163.9367	90	30	5700
7	11	159.3465	90	30	9500
5	20	176.3928	71	30	7800
4	10	156.7855	71	30	4800
7	18	173.2615	83	30	7200
10	23	180.634	67	32	20000
7	12	161.7211	83	29	11500
8	18	173.2615	62	30	7200
8	17	171.5861	90	28	12000
3	3	127.7663	100	27	2700
3	3	127.7663	98	23	2700
3	5	139.3576	98	26	3300
5	10	156.7855	69	30	4800
3	4	134.1702	105	29	3000
2	1	106	99	28	2100

3	5	139.3576	99	27	3300
3	5	139.3576	100	27	3300
9	37	195.8394	96	32	12900
10	30	280.6603	27	31	22500
10	31	190.0366	95	31	21100
10	30	280.6603	27	30	22500
10	30	280.6603	51	29	22500
8	39	311.7155	51	31	16000
10	31	190.0366	97	30	11100
12	46	332.9934	28	29	30500
6	21	177.862	99	30	8100
11	40	314.8883	31	30	27500
8	32	191.0651	83	30	11400
13	50	344.2869	30	30	32500
6	6	147.4324	44	30	10500
7	12	194.5382	44	30	13500
4	7	156.8093	32	29	5000
9	19	174.8614	70	30	7500
7	8	165.4126	33	29	11500
8	16	218.2632	19	30	15500
8	26	184.4384	96	31	9600
7	18	228.7923	38	30	10000
7	16	169.8268	83	29	6600

Table 2: Test data

Test Data					
Dependent	Independent	Independent	Independent	Dependent	Independent
Friction	SPT	Shear wave velocity	Fine Content	Cohesion	Elasticity
13	47	335.8703	46	30	31000
6	16	169.8268	82	29	6600
10	22	179.2742	68	29	19500
10	39	311.7155	37	33	15000
12	37	305.2201	37	29	26000
11	30	188.9803	51	29	20500
11	27	269.0779	26	29	21000
9	24	256.6948	45	29	19500
10	30	280.6603	45	30	22500
9	29	187.8943	70	30	10500

3.4 RESEARCH METHODOLOGY

The whole research was completed in three phases. In the first phase the data were collected from field investigation and laboratory test result. In the second phase, different kinds of machine learning techniques - Simple linear regression model, Multi-polynomial Regression, Multivariate Adaptive Regression Splines, Random Forest, Support Vector Regression, M5 Model Tree and ANN were applied to predict the co-relation between angle of friction with SPT N Value, Shear wave velocity, fine content, cohesion and elasticity. In the third phase, an embankment model was developed in the PLAXIS 2D software in order to validate the machine learning models.

4 RESEARCH ANALYSIS

In this section, the results generated from different machine learning models has been discussed in detail. The following sections represent the discussion of the result of various models.

4.1 SIMPLE LINEAR REGRESSION

In this study, we have applied a simple linear regression model to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the simple linear regression model. The following **Figure 6** shows the scripted coding of the simple linear regression model.

```
X=x
poly = PolynomialFeatures(degree=1)
X = poly.fit_transform(X)

regression = linear_model.LinearRegression()

model = regression.fit(X, y)
print(model.score(X, y))

y_pred=model.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))
```

Figure 6: Python programming language for simple linear regression

In the coding section, X is denoted as an independent variable, and Y is denoted as a dependent (Friction value). This research used Polynomial Features (degree=1) module to perform the simple linear regression model using the python programming language.

From the analysis, it is found out that the R^2 value which is the coefficient of determination is 85.56%. The R^2 value of this model indicates a good amount of determination. The value obtained for Mean Absolute Error (MAE) of this model is 0.549, Mean Squared Error (MSE) is 0.496, Root Mean Square Error (RMSE) is 0.705, and Mean Absolute Percentage Error (MAPE) is 5.27%. The following **Figure 7** shows the regression graph obtained for the simple linear regression model.

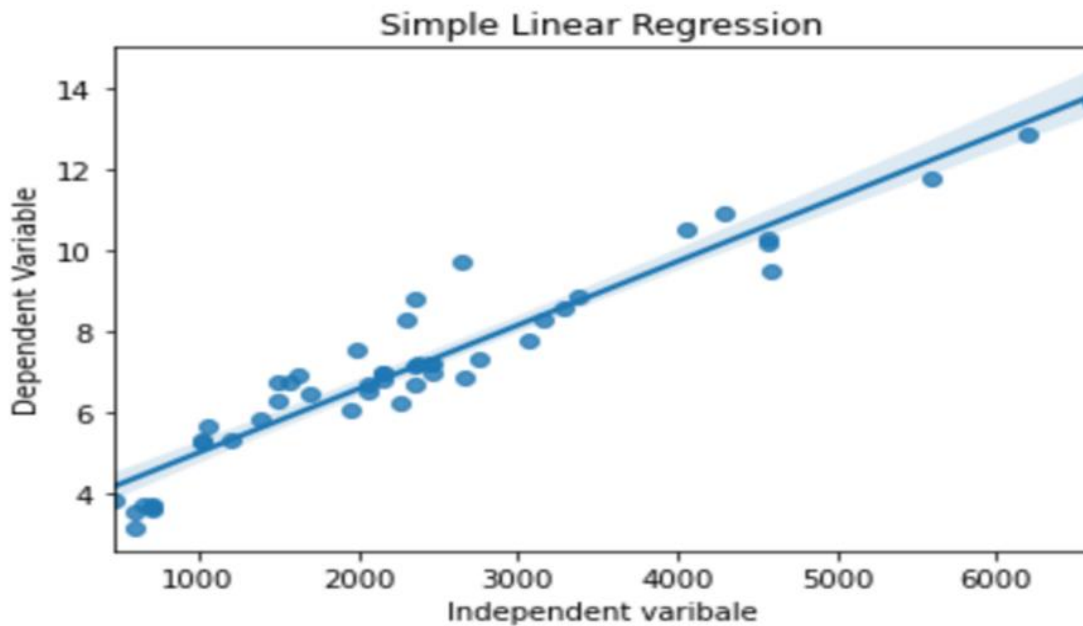


Figure 7: Regression graph for simple linear regression

4.2 MULTI POLYNOMIAL REGRESSION

In this study, we have applied a Multi polynomial regression model to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the Multi polynomial regression model. The following **Figure 8** shows the scripted coding of the Multi polynomial regression model.

```

X=x
poly = PolynomialFeatures(degree=2)
X = poly.fit_transform(X)

regression = linear_model.LinearRegression()

model = regression.fit(X, y)
print(model.score(X, y))

y_pred=model.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))

```

Figure 8: Python programming language for Multi polynomial Regression

In the coding section, X is denoted as an independent variable, and Y is denoted as a dependent (Friction value). This research used Polynomial Features (degree=2) module to perform the Multi Polynomial regression model using the python programming language. The following **Figure 9** shows the regression graph obtained for the multi polynomial regression.

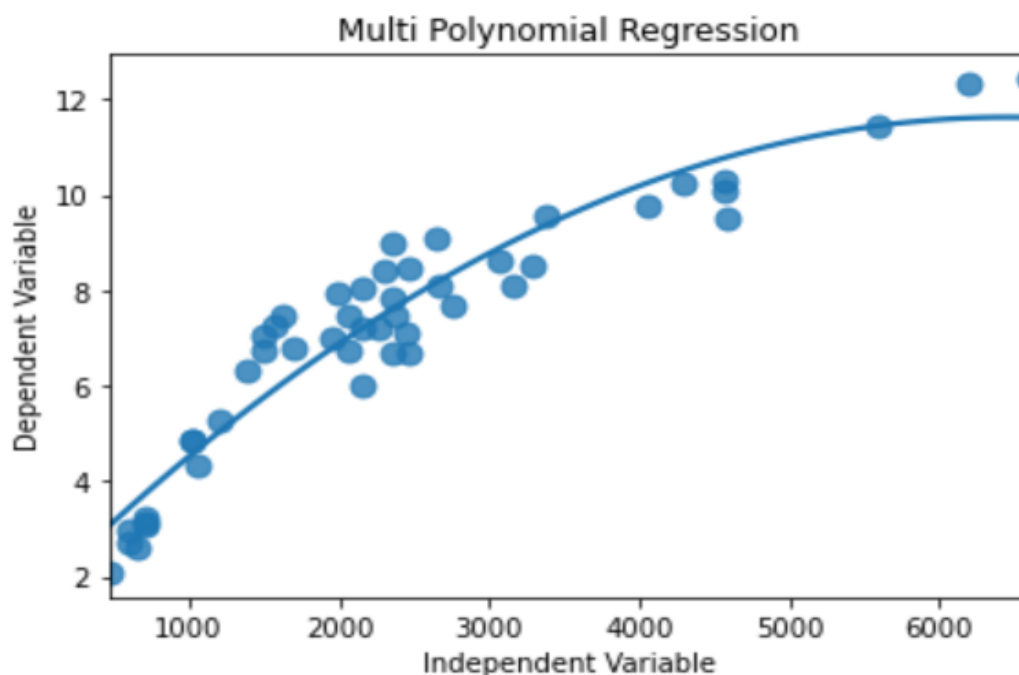


Figure 9: Regression graph for Multi Polynomial Regression

From the analysis, it is found out that the R^2 value which is the coefficient of determination is 93.24%. The R^2 value of this model indicates a good amount of determination. The value obtained for Mean Absolute Error (MAE) of this model is 0.807, Mean Squared Error (MSE) of 01.025, Root Mean Square Error (RMSE) of 1.013, and Mean Absolute Percentage Error (MAPE) of 7.59%.

4.3 SUPPORT VECTOR REGRESSION

In this study, we have applied the Support Vector regression model to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the Support Vector regression model. The following **Figure 10** shows the scripted coding of a simple Support Vector regression model.

```
X=x
regr = SVR(C=1.0, epsilon=0.2)
regr.fit(X, y.ravel())
print(regr.score(X, y))

y_pred=regr.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))
```

Figure 10: Python programming language Support Vector Regression

In the coding section, X is denoted as an independent variable, and Y is denoted as a dependent (Friction value). We used sklearn. SVM module to module to perform Support Vector Regression model using a python programming language. From sklearn module, we imported the SVR function and put $c=1.0$ and $\epsilon = 0.2$ to perform the regression model. The following **Figure 11** shows the regression graph obtained for the Support Vector Regression model.

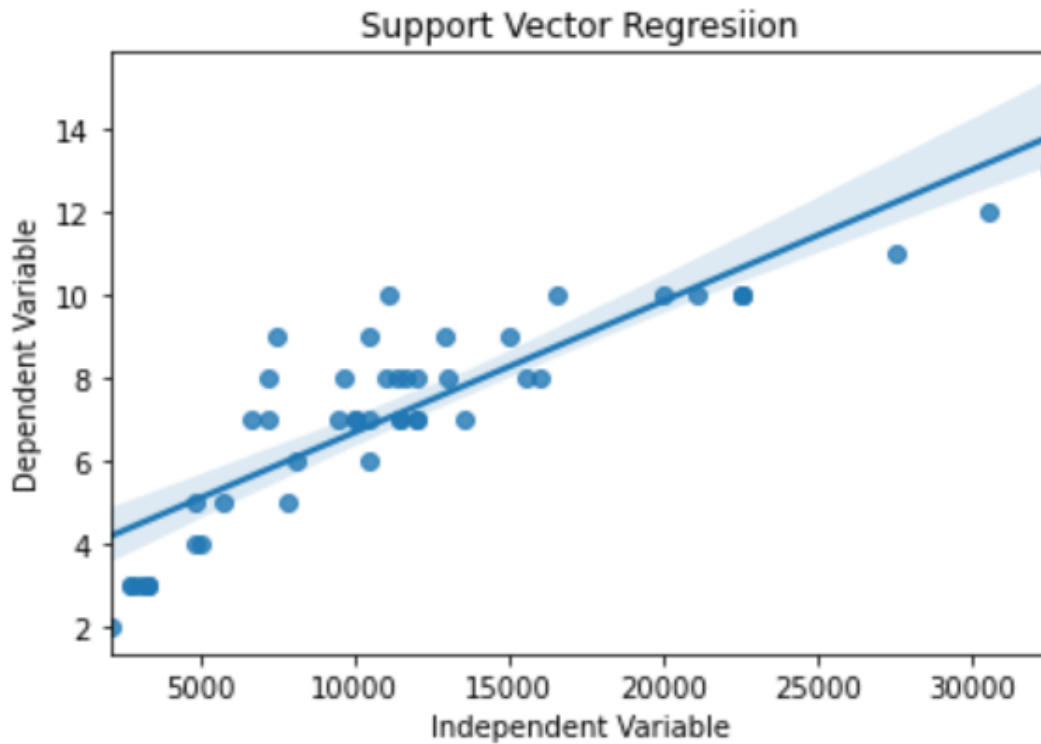


Figure 11: Regression graph for Support Vector Regression Model

From the analysis, it is found out that the R^2 value which is the coefficient of determination is 80.06%. The R^2 value of this model indicates a good amount of determination. The value obtained for Mean Absolute Error (MAE) of this model is 1.058, Mean Squared Error (MSE) of 1.539, Root Mean Square Error (RMSE) of 1.240, and Mean Absolute Percentage Error (MAPE) of 10.22%.

4.4 RANDOM FOREST

In this study, we have applied the Random Forest regression model to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the Random Forest regression model. The following **Figure 12** shows the scripted coding of simple Random Forest regression model.

```

X=x

rfr = RandomForestRegressor(max_depth=12, random_state=0)
rfr.fit(X, y.ravel())
print(rfr.score(X, y))

print(rfr.predict(X[0:1]))

y_pred=rfr.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))

```

Figure 12: Python programming language Random Forest Model

In the coding section, X is denoted as independent variable and Y is denoted as dependent (Friction value). We used Skill learn module to perform the Random forest Regression model using the python programming language. From skill learn module we imported Random Forest Regression function and put max depth = 12 and random state = 0 to perform the regression model. The following **Figure 13** shows the regression graph obtained for the random forest model.

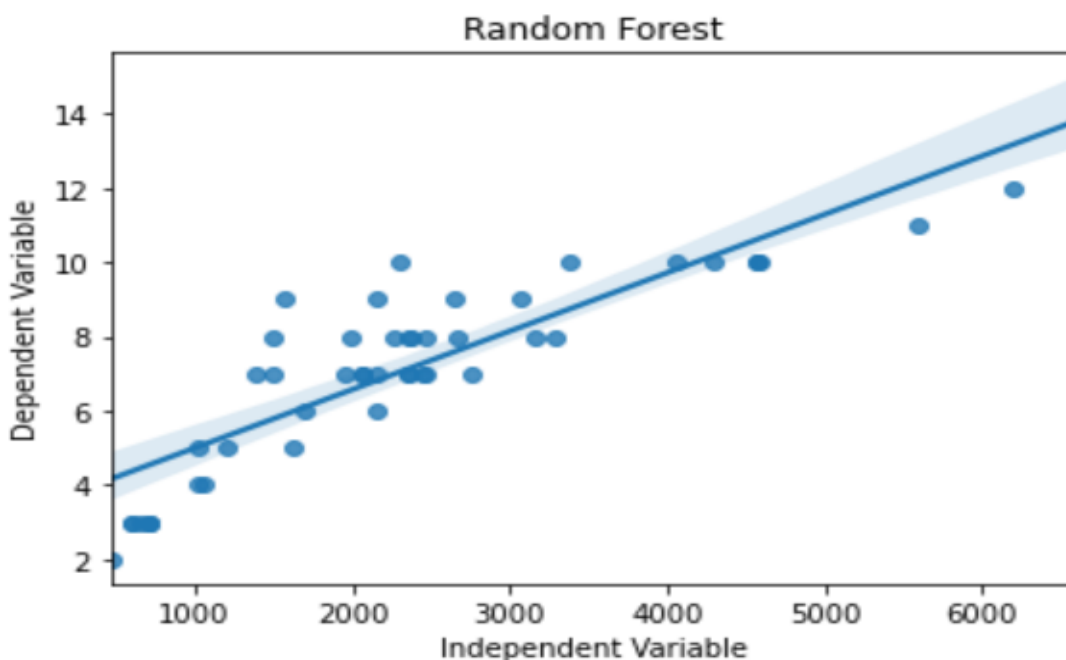


Figure 13 : Regression Graph for Random Forest Model

From the analysis, it is found out that the R^2 value which is the co-efficient of determination is 97.95%. The R^2 value of this model indicates good amount of determination. The value obtained for Mean Absolute Error (MAE) of this model is 0.879, Mean Squared error (MSE) of 1.016, Root Mean Square Error (RMSE) of 1.00, and Mean Absolute Percentage Error (MAPE) of 8.88%.

4.5 MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

In this study, we have applied Multivariate Adaptive regression model to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the Multivariate Adaptive regression model. The following **Figure 14** shows the scripted coding of Multivariate Adaptive regression model.

```
X=x

mars = Earth()
mars_model_fitted=mars.fit(X,y)
print(mars.score(X, y))

print(mars.predict(X[1:0]))

y_pred=mars.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))
```

Figure 14: Python Programming Language for Multivariate Adaptive Regression Splines Model

In the coding section, X is denoted as independent variable and Y is denoted as dependent (Friction value). We used py-earth module to perform the MARS Regression model using the python

programming language. From py-earth module, we imported Earth to perform the regression model. The following **Figure 15** shows the regression graph obtained for the multivariate adaptive regression model.

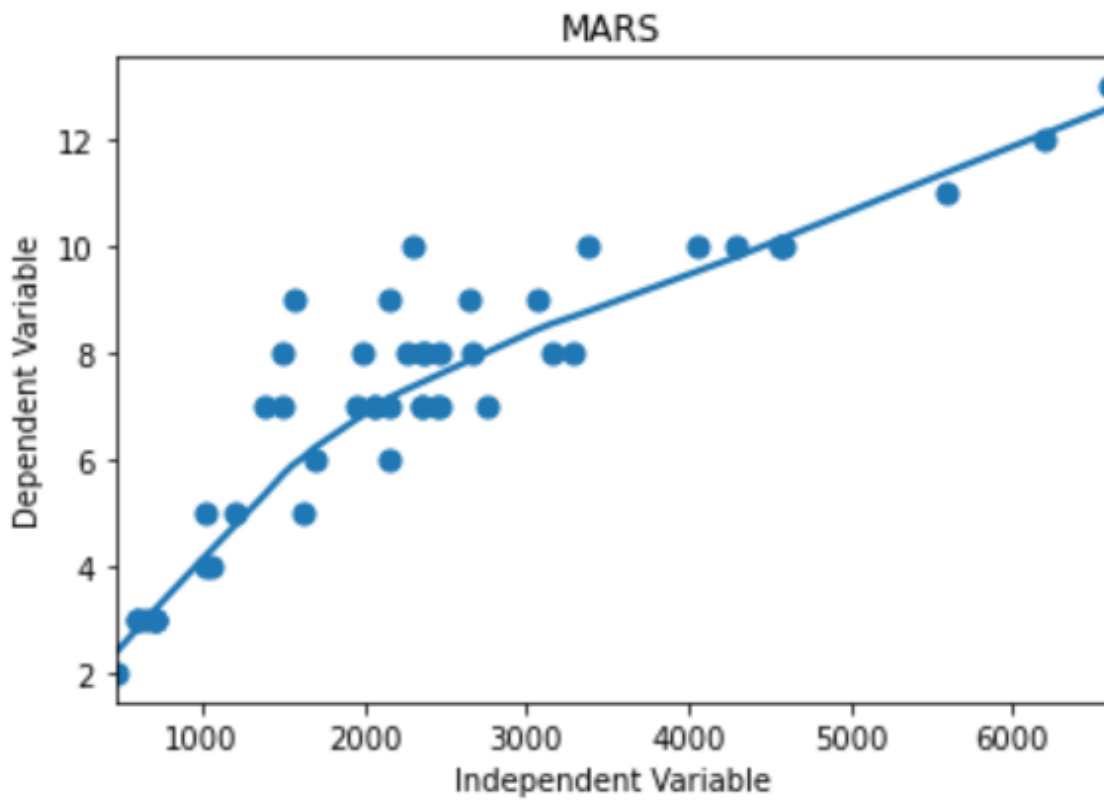


Figure 15 : Regression Graph for Multivariate Adaptive Regression Splines Model

From the analysis, it is found out that the co-efficient of determination (R^2 value) is 91.13%. The R^2 value of this model indicates good amount of determination. Mean Absolute Error (MAE) of 0.738, Mean Squared Error (MSE) of 0.696, Root Mean Square Error (RMSE) of 0.834, and Mean Absolute Percentage Error (MAPE) of 7.13%.

4.6 M5 MODEL TREE

In this study, we have applied M5 Model tree to predict the friction angle using 5 independent variables. The variables are SPT, Shear wave velocity, fine content, cohesion, and elasticity. The python programming language was used to code the M5 Model tree. The following **Figure 16** shows the scripted coding of M5 Model tree.

```
X=x

regressor = DecisionTreeRegressor(random_state=0)

regressor.fit(X,y)
print(regressor.score(X, y))

y_pred=regressor.predict(X)

print('R2 score: '+str(r2_score(y, y_pred)))
```

Figure 16: Python Programming Language M5 Model Tree

In the coding section, X is denoted as independent variable and Y is denoted as dependent (Friction value). We used skill learn tree module to perform M5 model tree using python programming language. From skill learn tree module we imported Decision Tree Regressor function to perform the regression model. The following **Figure 17** shows the regression graph obtained for the M5 Model Tree.

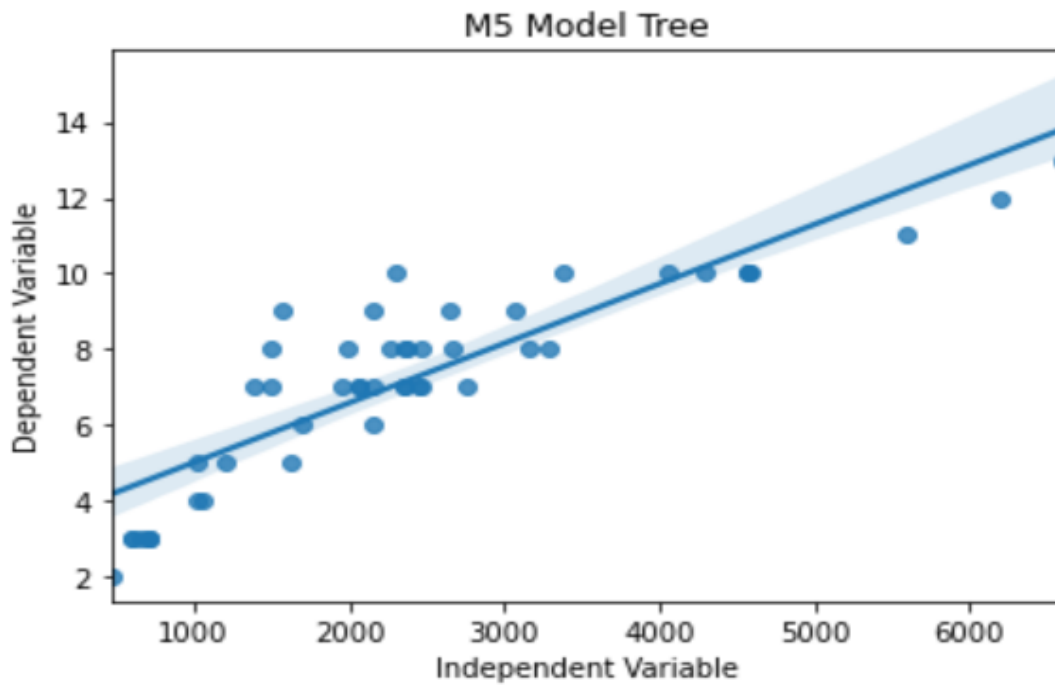


Figure 17: Regression Graph for M5 Model Tree

From the analysis, it is found out that the co-efficient of determination (R^2 value) is 95.23%, Mean Absolute Error (MAE) of 0.90, Mean Squared error (MSE) of 1.10, Root Mean Square Error (RMSE) of 1.049, and Mean Absolute Percentage Error (MAPE) of 9.31%.

5 RESULTS & DISCUSSION

In this section, the summary of the results obtained from all the models has been described in the following section.

5.1 RESULT SUMMERY

In our research we have applied all the models to correlate and predict the collected data and the result of research analysis is summarized in the following table. We have prioritized the co-efficient of determination (R^2 value) to compare the accuracy between all the models and also MAE MSE, RMSE, MAPE error value to compare the errors in predicting the values.

Here is a summary of the results obtained from all the models shown below in **Table 4**.

Table 3 : Result Summary of all machine learning models

Result Summary					
Model Name	R^2	MAE	MSE	RMSE	MAPE
Simple linear regression	85.56%	0.549	0.496	0.705	5.27
Multi-polynomial Regression	93.24%	0.807	1.02547	1.012655	7.585303
MARS	91.13%	0.738	0.69596	0.834242	7.13472
Random Forest	97.95%	0.879	1.01607	1.008003	8.881251
SVR	80.06%	1.058	1.53856	1.240387	10.2284
M5 Model Tree	95.23%	0.9	1.1	1.049	9.31
ANN	40.41%	0.15	0.18	0.82	0.77

From the above table, it is observed that conventional machine learning techniques obtained a greater than 80 & R^2 value except for the ANN. The table also summarizes that the error value of MAE, MSE, and RMSE is close to 1 or in the range of 0 to 1 which describes the better the performance for each of the models in correlating the parameters and predicting has less error.

5.2 ACTUAL VALUE VS PREDICTED VALUE

5.2.1 SIMPLE LINEAR REGRESSION

After the analysis, the model was run by the test data and a comparison was developed between the actual value with the predicted value. For a simple linear regression model, the actual value vs predicted value chart and graph is shown below:

Model Name	R^2	MAE	MSE	RMSE	MAPE in percentage
Simple linear regression	85.56%	0.549	0.496	0.705	5.27

The following **Table 5** shows the difference between the actual value and predicted value in the simple linear regression model.

Table 4: Actual Value vs Predicted value in Simple Linear Regression Model

Predicted Friction	Real Friction
12.74	13
5.87	6
9.91	10
8.87	10
11	12
11.41	11
9.57	11
8.68	9
9.74	10

The following **Figure 18** shows the original vs predicted graph in simple linear regression model.

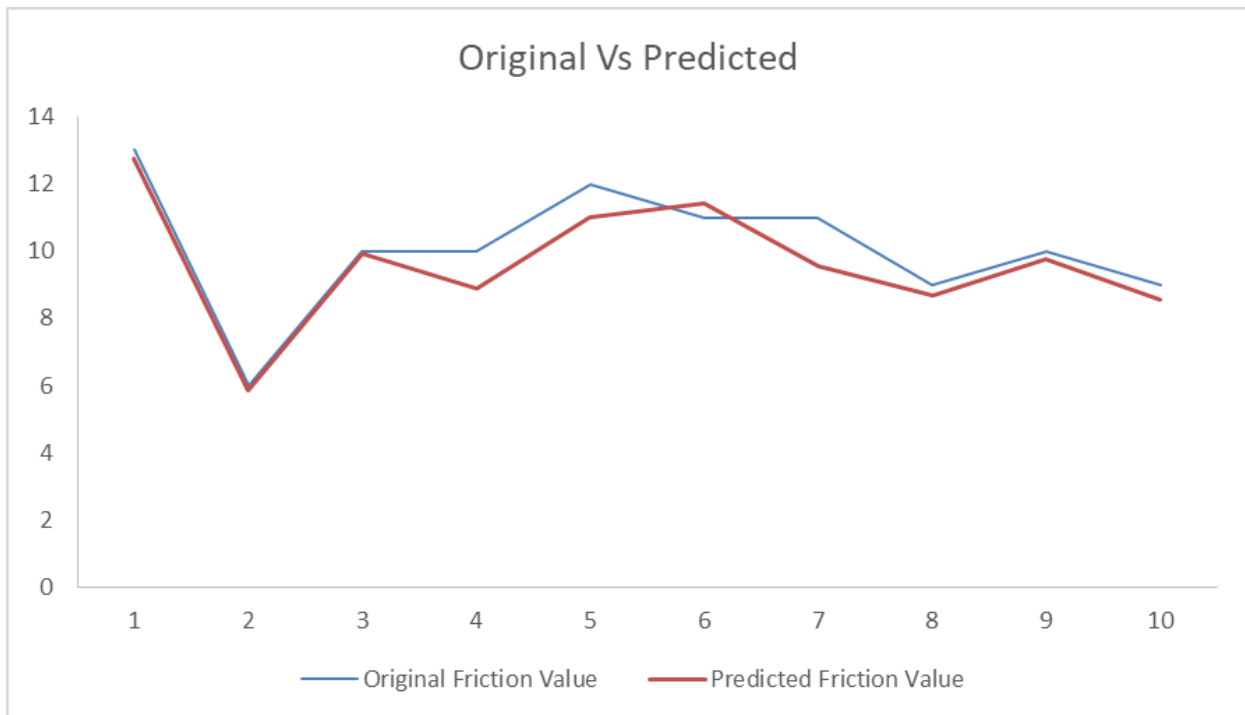


Figure 18 : Original vs Predicted graph in Simple Linear regression Model

5.2.2 MULTI-POLYNOMIAL REGRESSION

For Multi-polynomial Regression model, the actual value vs predicted value chart and graph is shown below:

Model Name	R ²	MAE	MSE	RMSE	MAPE
Multi polynomial Regression	93.24%	0.807	1.025	1.013	7.59

The following **Table 6** shows the difference between actual value and predicted value in multi-polynomial regression model.

Table 5 : Actual Value vs Predicted Value in Multi-polynomial Regression

Predicted Friction	Real Friction
11.18	13
6.35	6
9.35	10
8.07	10
10.99	12
10.11	11
10.17	11
9.22	9
9.65	10

The following **figure 19** shows the original vs predicted graph in multi-polynomial regression model.

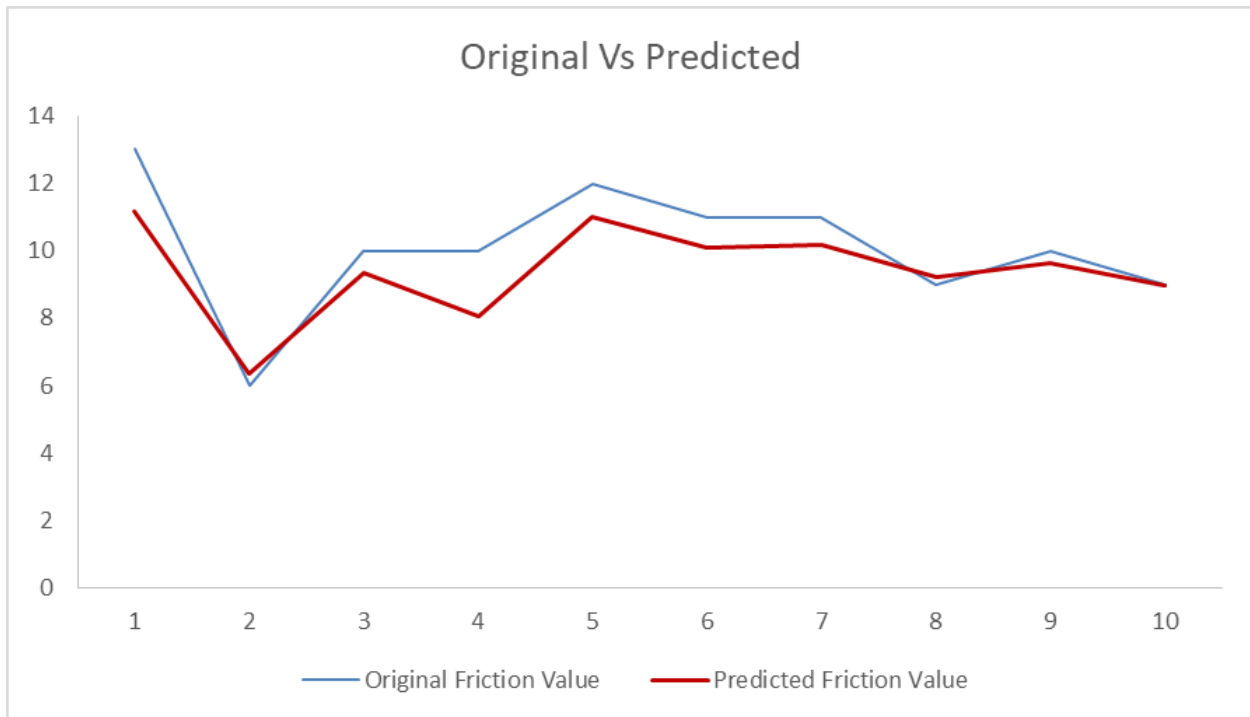


Figure 19 : Original vs Predicted graph in Multi-polynomial Regression

5.2.3 SUPPORT VECTOR REGRESSION

For Support Vector Regression model, the actual value vs predicted value chart and graph is shown below:

Model Name	R ²	MAE	MSE	RMSE	MAPE
Support Vector Regression	80.06%	1.058	1.539	1.24	10.228

The following **Table 7** shows the difference between actual value and predicted value in support vector regression model.

Table 6 : Actual Value vs Predicted Value in Support Vector Regression

Predicted Friction	Real Friction
10.65	13
5.5	6
9.76	10
8.79	10
10.56	12
9.93	11
10	11
9.77	9
10.2	10

The following **figure 20** shwos the original vs predicted graph in support vector regression model.

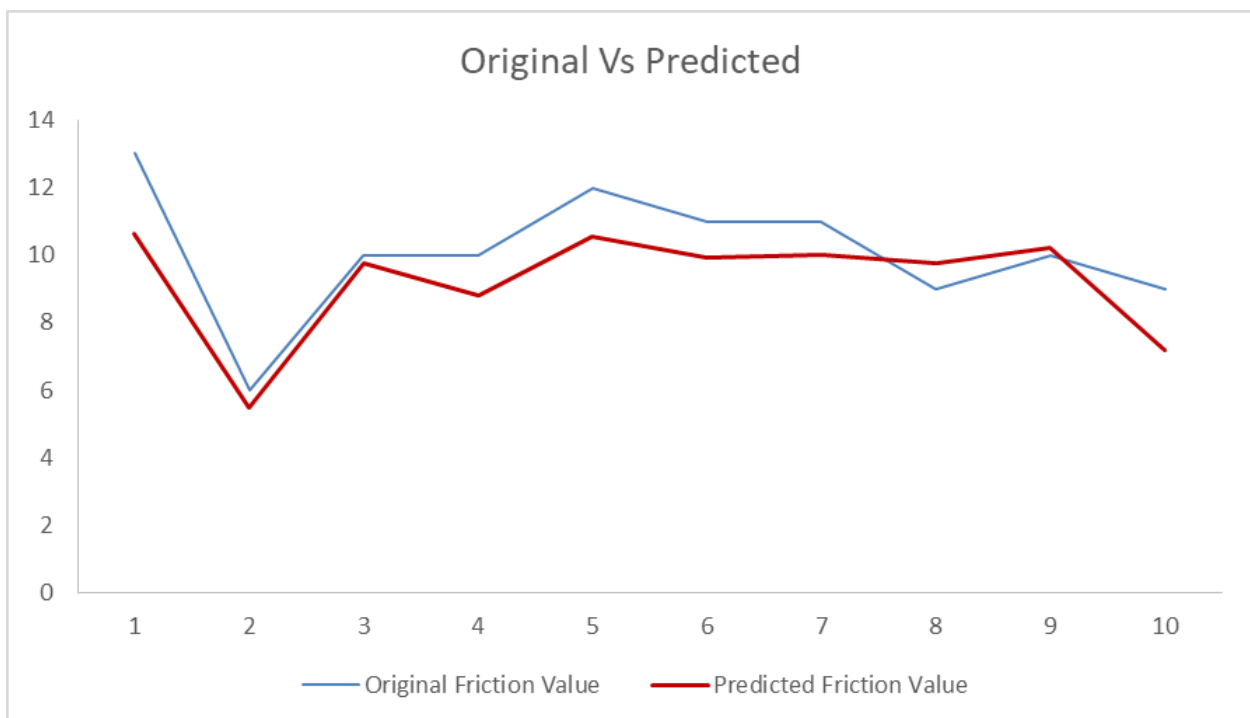


Figure 20 : Original vs Predicted graph in Support Vector Regression Model

5.2.4 RANDOM FOREST REGRESSION

For the Random Forest model, the actual value vs predicted value chart and graph is shown below:

Model Name	R ²	MAE	MSE	RMSE	MAPE
Random Forest	97.95%	0.0879	1.016	1.00	8.88

The following **Table 8** shows the difference between actual value and predicted value in random forest regression model.

Table 7: Actual Value vs Predicted Value in Random Forest Regression model

Predicted Friction	Real Friction
12.13	13
7.02	6
9.4	10
8.51	10
10.28	12
10.02	11
9.79	11
9.52	9
10.02	10

The following **Figure 21** shows the original vs predicted graph in random forest regression model.

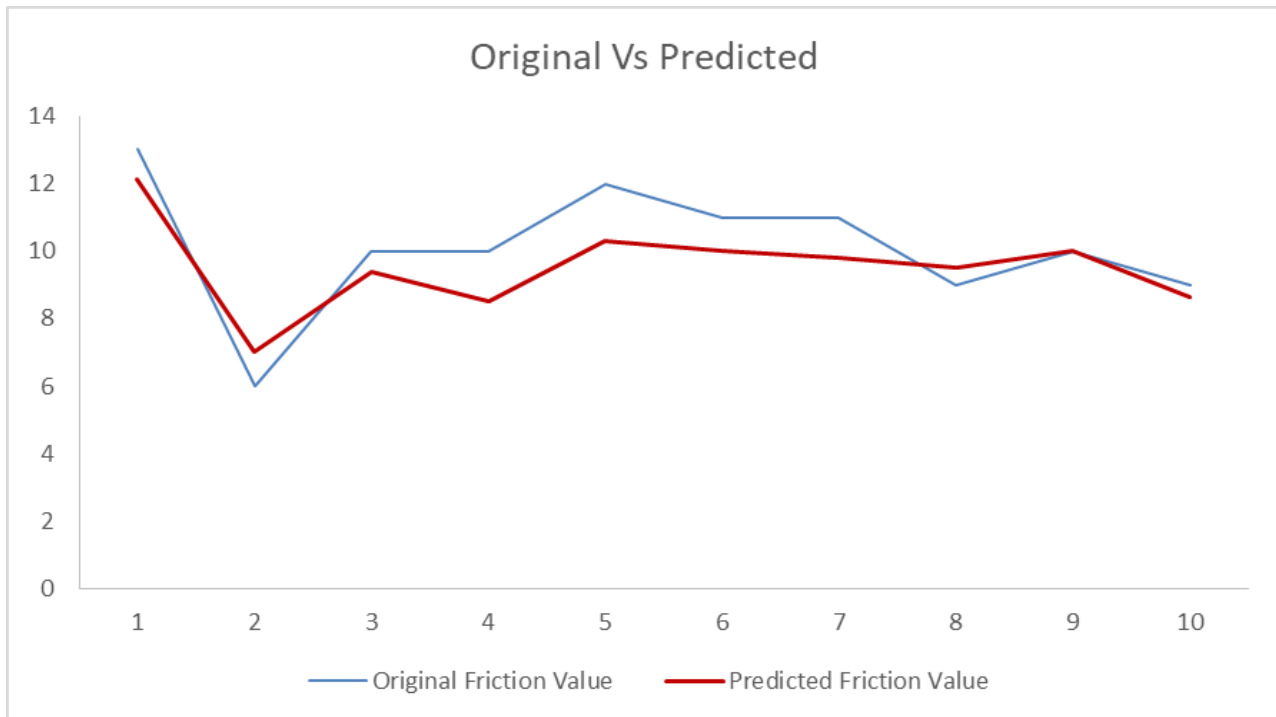


Figure 21: Original vs. Predicted graph in Random Forest Regression model

5.2.5 MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

For Multivariate Adaptive Regression Splines (MARS) model, the actual value vs predicted value chart and graph is shown below:

Model Name	R ²	MAE	MSE	RMSE	MAPE
MARS	91.13%	0.738	0.696	0.834	7.13

The following **Table 9** shows the difference between actual value and predicted value in multivariate adaptive regression splines model.

Table 8 : Actual Value vs Predicted Value in Multivariate Adaptive Regression Splines model

Predicted Friction	Real Friction
12.27	13
6.28	6
9.22	10
9.28	10
10.99	12
9.75	11
9.7	11
9.31	9
10.08	10

The following **figure 22** shows the original vs predicted graph in multivariate adaptive regression model.

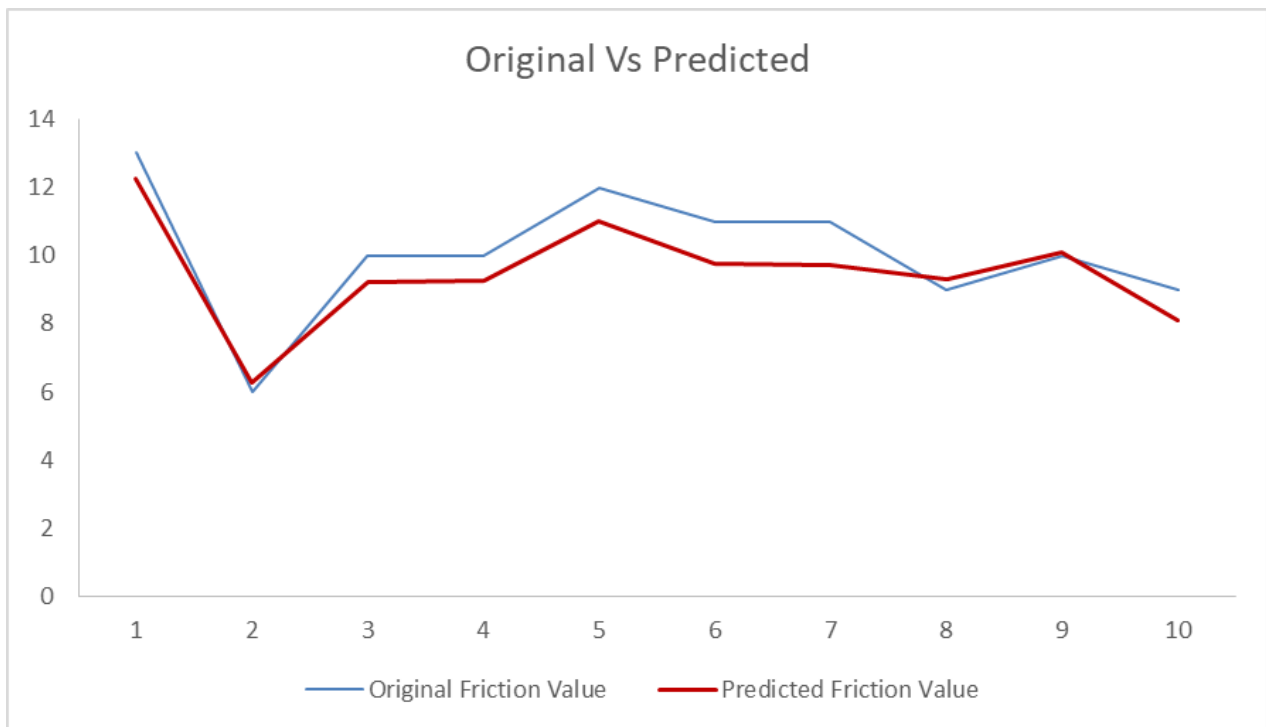


Figure 22: Original vs Predicted graph in Multivariate Adaptive Regression Splines model

5.2.6 M5 MODEL TREE

For M5 Model Tree model, the actual value vs predicted value chart and graph is shown below:

Model Name	R ²	MAE	MSE	RMSE	MAPE
M5 model tree	95.23%	0.9	1.1	1.049	9.31

The following **Table 10** shows the difference between actual value and predicted value in M5 model tree.

Table 9 : Actual Value vs Predicted Value in M5 Model Tree

Predicted Friction	Real Friction
8	13
3	6
4	10
5	10
5	12
4	11
5	11
2	9
9	10
1	9

The following **Figure 18** shows the original vs predicted graph in M5 model tree.

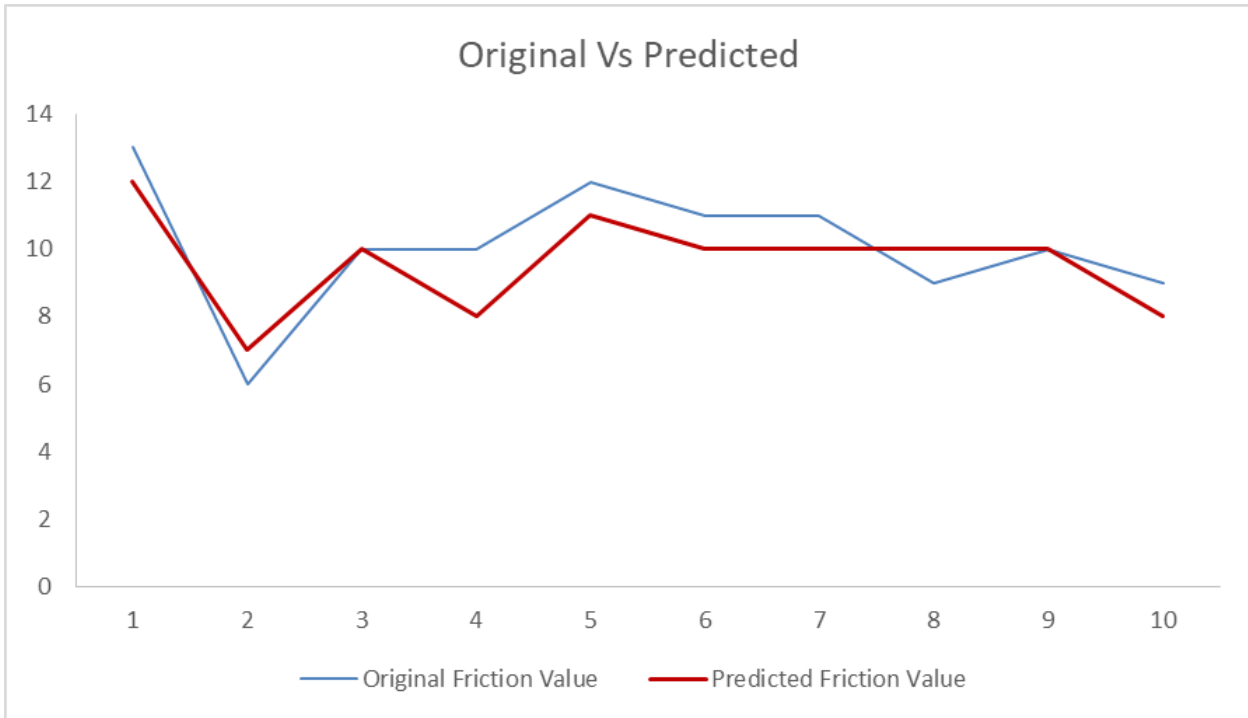


Figure 23 : Original vs Predicted graph in M5 Model Tree

5.2.7 ARTIFICIAL NEURAL NETWORK

Model Name	R ²	MAE	MSE	RMSE	MAPE
ANN	40.41%	6.3	34.3	5.85	55.16

The following **Table 11** shows the difference between actual value and predicted value in artificial neural network.

Table 10: Actual Value vs Predicted Value in Artificial Neural Network

Predicted Friction	Real Friction
12	13
7	6
10	10
8	10
11	12
10	11
10	11
10	9
10	10
8	9

5.2.8 ORIGINAL VS PREDICTED (ALL MODELS)

The original friction value of test data vs predicted value of all models is shown below in charts and graph:

The following **Table 12** shows the difference between original and predicted friction value for all models.

Table 11: Original vs Predicted Friction Value for All Models

Real Friction	Predicted Friction						
Test data	Simple Linear Regression	Multi-polynomial Regression	Support Vector Regression	Random Forest	MARS	M5 model tress	ANN
13	12.74	11.18	10.65	12.13	12.27	12	8
6	5.87	6.35	5.5	7.02	6.28	7	3
10	9.91	9.35	9.76	9.4	9.22	10	4
10	8.87	8.07	8.79	8.51	9.28	8	5
12	11	10.99	10.56	10.28	10.99	11	5
11	11.41	10.11	9.93	10.02	9.75	10	4
11	9.57	10.17	10	9.79	9.7	10	5
9	8.68	9.22	9.77	9.52	9.31	10	2
10	9.74	9.65	10.2	10.02	10.08	10	9
9	8.54	8.98	7.2	8.64	8.08	8	1

The following **Figure 24** shows the original vs predicted value graph for all models.

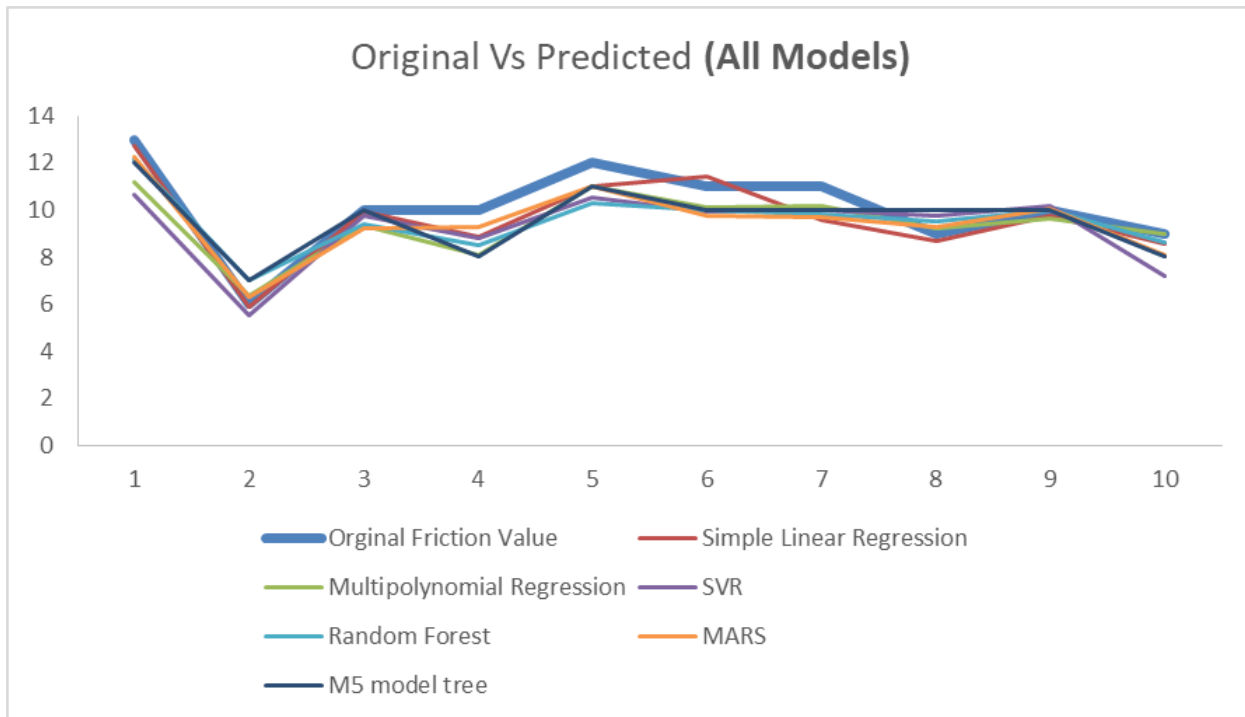


Figure 24 : Actual vs Predicted Friction value Graph for all models

The bold blue line indicates the original friction value and other colours shows the predicted friction value of the models.

5.3 R² VALUE COMPARISON

A comparison between all the models based on the calculated coefficient of determination (R² value) to find the models that shown better accuracy. A comparison of the R² value obtained from all models is described in the following **Figure 25**.

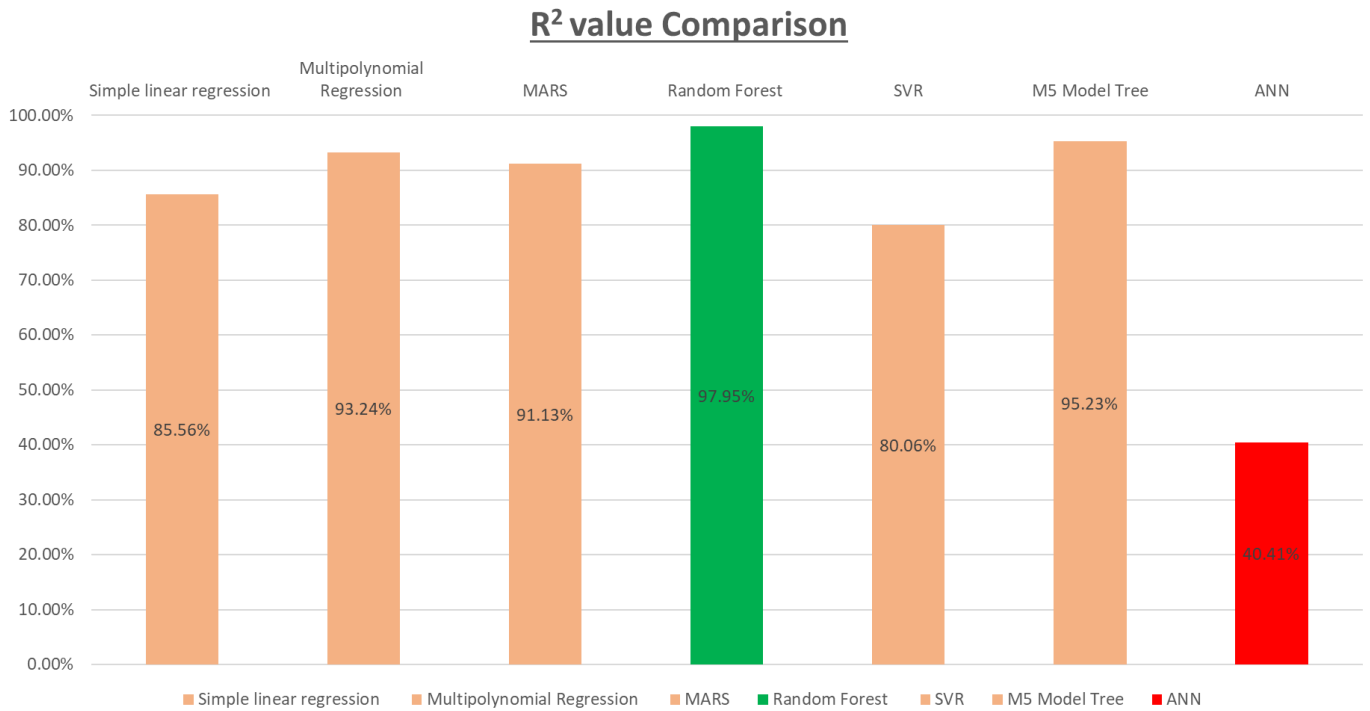


Figure 25: R² Value Comparison for All Models

5.4 ERROR VALUE COMPARISON

The following **Figure 26** shows a comparison of error values obtained from all models.

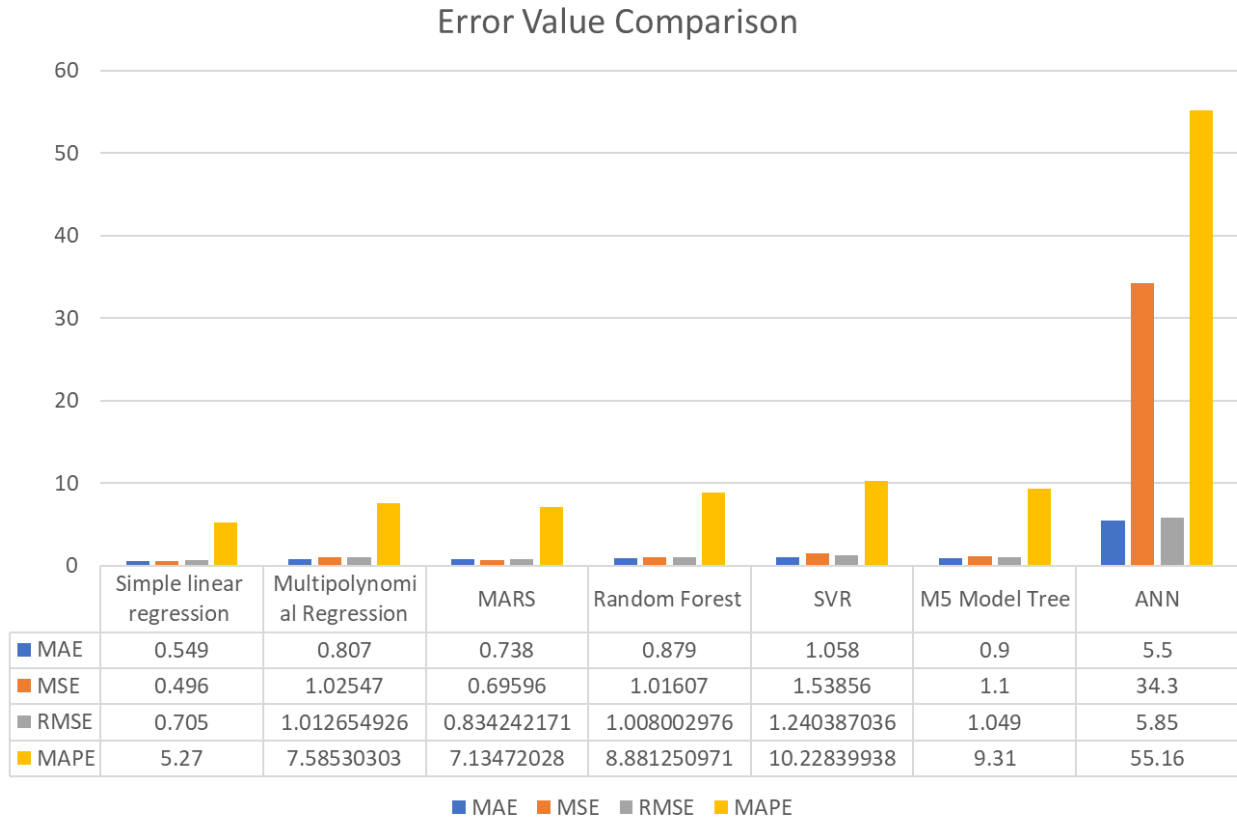


Figure 26: Error Value Comparison for All models

5.5 CONCLUSION OF THE RESULTS

We have developed the comparison of R^2 value and error values. All the conventional machine learning models obtained greater than 80% value but the deep learning ANN failed to obtain a good value of R^2 . The reason behind it might be running the model with less amount of data. The error value of all conventional machine learning models is also close to 1 or in the range between 0-1 whereas the error value of the ANN model is higher than the range. The reason behind it is also the same. From our result, it is concluded that the Random Forest model is correlated better than other models in terms of R^2 value. So, it can be said from the analysis of results and explanation of graphs that while predicting the angle of friction with the correlation of soil parameters, the Random Forest, and M5 model tree have shown the best result with higher accuracy and less error value.

6 RESULT VALIDATION

Any model loses its credibility if it fails to show accuracy through validation. In order to validate the research of the results, we have used PLAXIS 2D modelling software for the validation process. We have chosen PLAXIS because PLAXIS is a Leading geotechnical engineering simulation software and renowned for ease of use and accuracy. First, we have developed a model of an embankment by collecting the data from the Bangladesh Water Development embankment design manual and used the actual friction value from test data. In the second step, we developed the same embankment model with the same collected data and used the predicted friction value as a replacement for the actual friction value. We used the predicted friction value generated from the random forest model. The result of the settlement value for both of the models was very much close. As the difference of settlement value is close to 0, we can conclude that our model is validated with a higher range of accuracy.

6.1 DESIGN DATA USED FOR MODELING

The following **Table 12** shows the design data used in our research for embankment modeling in PLAXIS 2D.

Table 12: Design Data for Embankment Modelling in PLAXIS 2D

Design data used for modeling			
Parameters	For Ground Soil (actual data)	For Ground soil (predicted data)	For Embankment Soil (collected from embankment design manual of BWDB)
Dry density, γ (KN/m ²)	17	17	16.5
Saturated density, γ (KN/m ²)	20	20	19
Void ratio (e) initial	0.5	0.5	0.5
Cohesion (c')	29	29	19.1
Angle of Friction (ϕ)	13	12	10
Dilatancy Angle (Ψ)	2	2	0

6.2 PLAXIS MODELLING

- **Figure 27** shows the settlement for actual value in PLAXIS modeling:

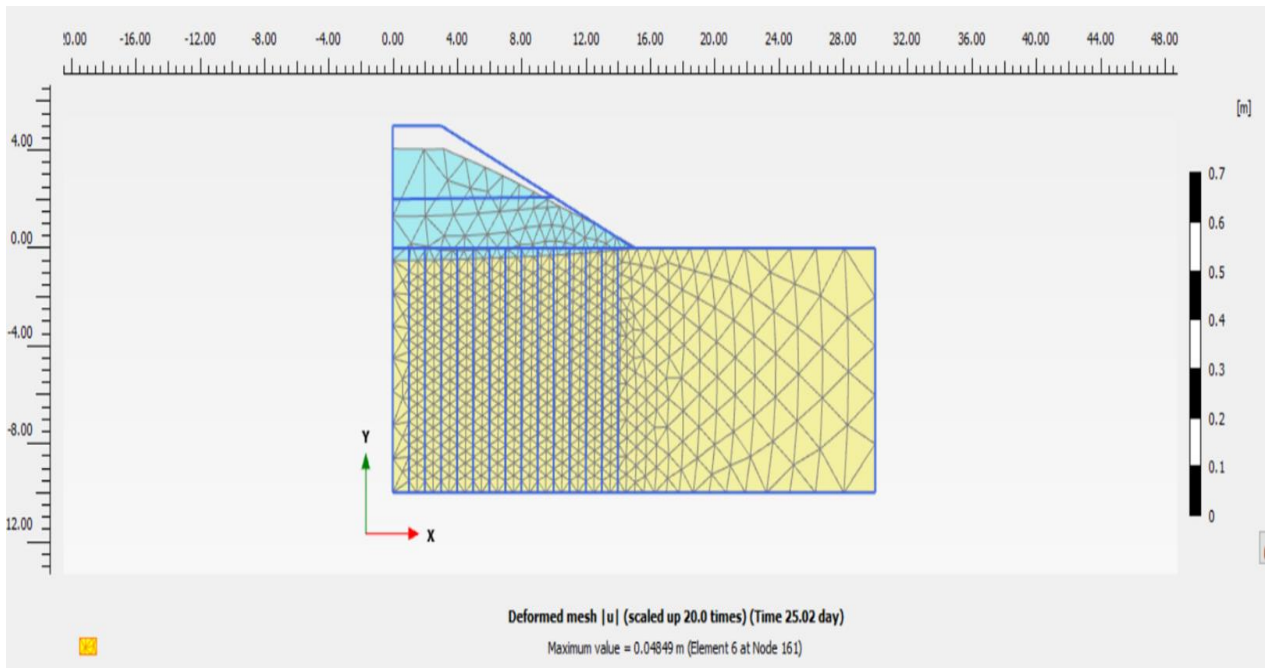


Figure 27 : Embankment Modeling for Actual Friction Value

- **Figure 28** shows the settlement for predicted value in PLAXIS modelling:

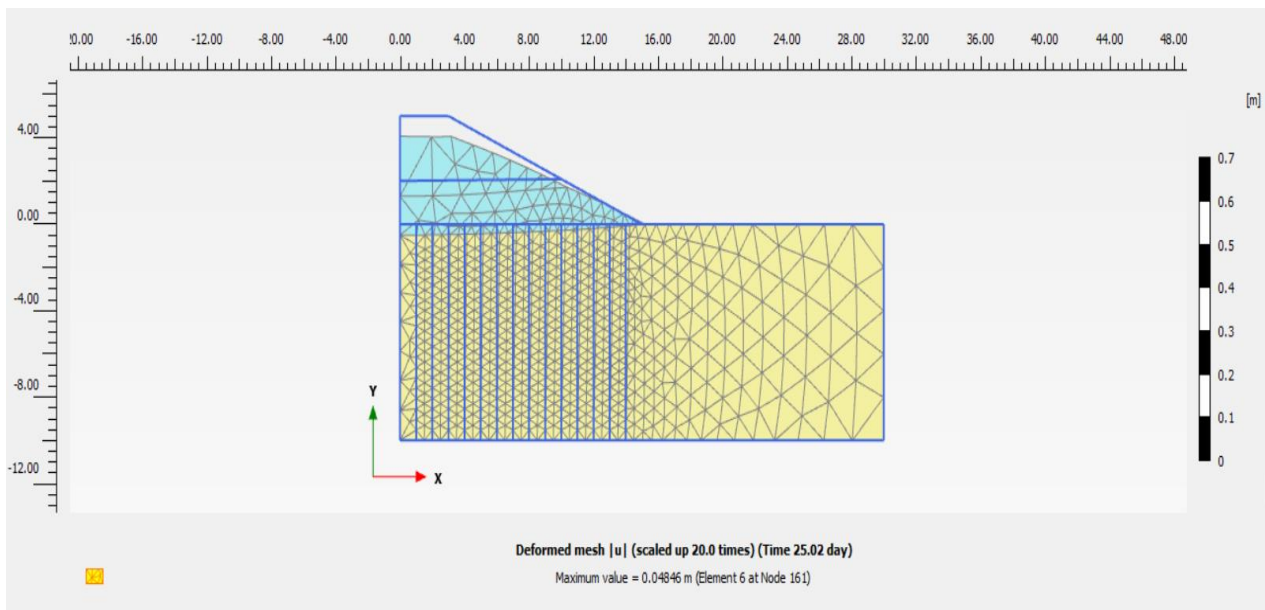


Figure 28 : Embankment Modeling for Predicted friction Values

6.3 DIFFERENCE OF SETTLEMENT VALUE

The following **Table 13** shows the difference in settlement value for all models in PLAXIS.

Table 13 : Difference in Settlement Value for all models in PLAXIS

Model Name	Actual value of Friction	The predicted value of friction	Settlement with actual value (m)	Settlement with predicted value (m)	Difference in Settlement (m)
Simple Linear Regression	13	12.74	0.04849	0.049	0.00051
Multi-polynomial	13	11.18	0.04849	0.05282	0.00433
MARS	13	12.27	0.04849	0.05002	0.00153
Random Forest	13	12.13	0.04849	0.05034	0.00185
SVR	13	10.65	0.04849	0.05455	0.00606
M5 Model Tree	13	12	0.04849	0.05065	0.00216
ANN	13	8	0.04849	0.06664	0.01815

Except for the ANN model, it can be concluded that all conventional machine learning models have shown a very less difference in settlement value which concludes that the models have predicted the angle of friction with higher accuracy.

7 CONCLUSION & FURTHER RECOMMENDATION

7.1 CONCLUSION

From all the research it can be concluded that:

- Soil Parameters can be correlated and predicted using machine learning techniques
- Random Forest and M5 model tree regression model gave us the best result, as the R^2 value is highest among all of the models. Error-values are also less which proves the higher accuracy.
- Conventional machine learning is more effective and accurate than ANN when we have a data shortage.
- The use of these advanced prediction models will decrease the need of performing the excessive number of soil tests which will save time & money.

7.2 FURTHER RECOMMENDATION

For further recommendation, it can be said that

- This research will facilitate the base for further researches in geotechnical engineering through the use of machine learning
- In the future, these prediction processes can be used not only for coastal areas but also for other areas.
- Different data augmentation techniques (Smote Analysis) can be applied to mitigate the limitation of data shortage in future

8 REFERENCES

- 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.
- Khuri, A.I. and Conlon, M., 1981. Simultaneous optimization of multiple responses represented by polynomial regression functions. *Technometrics*, 23(4), pp.363-375.

- Kirts, S., Nam, B., Panagopoulos, O. and Xanthopoulos, P., 2019. Settlement Prediction Using Support Vector Machine (SVM)-Based Compressibility Models: A Case Study. *International Journal of Civil Engineering*, 17(10), pp.1547-1557.
- Konyushkov, V., 2020. Comparing the results of numerical modeling of slope stability in the Plaxis program with analytical calculations using the simplified method. *Вестник гражданских инженеров*, 17(3), pp.108-115.
- Lindvall, M., Molin, J. and Löwgren, J., 2018. From machine learning to machine teaching. *Interactions*, 25(6), pp.52-57.
- 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 ground strength 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.
- Ponomarev, A. and Sychkina, E., 2015. The application of research results anisotropic deformability sandstones for numerical modeling in PLAXIS. *PNRPU Construction and Architecture Bulletin*, (1), pp.21-36.
- 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.
- 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.
- 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*.

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.

