USING MACHINE LEARNING APPROACHES TO CONSTRUCT CORRELATIONS FOR COHESIVE SOILS USING IN-SITU AND LABORATORY DATA

by
Walid Khalid Mbarak
B.S., Civil Engineering, Gediz University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Civil Engineering
Boğaziçi University
2017

USING MACHINE LEARNING APPROACHES TO CONSTRUCT CORRELATIONS FOR COHESIVE SOILS USING IN-SITU AND LABORATORY DATA

APPROVED BY:

Assoc. Prof. Dr. Özer Çinicioğlu .....
(Thesis Supervisor)

Assoc. Prof. Nilüfer Özyurt Zihnioğlu .....

Assoc. Prof. Esma Nur Çinicioğlu .....

DATE OF APPROVAL: 15.08.2017

ACKNOWLEDGEMENTS

My deepest gratitude goes my supervisor Assoc. Prof. Dr. Özer Çinicioğlu, who guided me expertly through the research. His unwavering support and unmatched hospitality kept me enthusiastic for the duration of my study and together in making my years at Boğaziçi University all the worthwhile.

I would also like to thank Zemin Etüd ve Tasarım A.Ş and Geocon Zemin Uzmanlari Ve Mühendislik Ltd. Sti. for their non reluctance in granting me the data that has been used in this thesis study.

My appreciation extends to my long time friends Miza, Nasra and Said, without whose expertise, constant encouragement and intuitive insight contributed towards a part of this thesis.

Above all, I am indebted to my family, my aunties Farida, Hamida and Fadiya for their unwavering support, love and constant prayers through these grueling period, Without them I would not have been able to complete this research. Lastly, to my loving mother, it is her that has been the rock through my every storm,it is her that I have looked at for that extra motivation to complete my study. I love you, Mama.

ABSTRACT

USING MACHINE LEARNING APPROACHES TO CONSTRUCT CORRELATIONS FOR COHESIVE SOILS USING IN-SITU AND LABORATORY DATA

In a world where the sizes of construction sites are ever increasing and project deadlines ever reducing, the geotechnical engineer no longer has the time to properly conduct the necessary tests on the soil so as to come up with optimal soil properties that would as accurately as possible reflect the ones on site. Therefore, correlations equations together with in-situ tests and laboratory tests have formed the basis of geotechnical engineering design. The literature is filled with correlation equations developed by previous and present researchers. Some of these equations may or may not have any statistical background hence making them less reliable when used to estimate critical soil parameters. The goal of any correlation equation developed is to estimate as accurately as possible a response given a certain input. In this thesis, we aim at developing regression models using machine learning algorithms such as linear regression, Random Forest and Gradient Boosting so as to predict the undrained shear strength, c_u, the elastic modulus, E_m and the limit pressure, p_L. In order to further improve our prediction capabilities we can stack the aforementioned models using their weighted averages derived from their RMSE indices obtained from the test data. Finally, the best performing models are compared to the correlations equations found in the literature.

ÖZET

KOHEZYONLU ZEMİNLER İÇİN MAKİNE ÖĞRENMESİ YAKLAŞIMLARI KULLANILARAK ARAZİ VE LABORATUVAR TEST VERİLERİ İLE KORELASYON KURULMASI

Günümüz dünyasında inşaat faaliyet sahaları bu denli genişlemekte ve proje süreleri bu denli kısalmakta iken, geoteknik mühendislerinin sahadaki zemin özelliklerini mümkün olan en doğru şekilde yansıtacak saha deneylerini yapmak için yeterli zamanları bulunmamaktadır. Bu sebeple korelasyon denklemleri, saha deneyleri ve laboratuvar deneyleri ile birlikte geoteknik tasarım mühendisliğinin temelini oluşturmaktadır. Geoteknik literatürde geçmişteki ve günümüzdeki araştırmacılar tarafından geliştirilmiş bir çok korelasyon denklemi bulunmaktadır. Bu bağıntılardan bazılarının istatiksel bir temele sahip olmayışı, bu bağıntıları kritik zemin parametrelerinin belirlenmesinde daha az güvenilir yapmaktadır. Korealasyon denklemlerinin genel amacı elde etmek istediğimiz bir çıktıyı, belirli bir girdi ile mümkün olan en doğru bir biçimde tahmin edebilmektir. Bu tez çalışmasında, drenajsız kayma mukavemeti c_u, Elastisite modülü, E_m ve limit basınç, p_L paramatrelerini tahmin edebilmek için, "Doğrusal Regresyon", "Random Forest" ve "Gradient Boosting" gibi makine öğrenimi algoritmalarını kullanan regresyon modellerinin geliştirilmesi amaçlanmıştır. Tahmin kabiliyetimizi daha da geliştirmek için, bahsi geçen modellerin deney verilerinden elde edilmiş olan RMSE endeksleri ile hesaplanmış ağırlıklandırılmış ortalamaları kullanılarak, bu modeller bir arada değerlendirilmiştir. Sonuç olarak, en iyi performansı gösteren modeller literatürdeki korelasyon denklemleri ile karşılaştırılmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . ix
LIST OF TABLES . . . xii
LIST OF SYMBOLS . . . xiv
LIST OF ACRONYMS/ABBREVIATIONS . . . xv
1. INTRODUCTION . . . 1
1.1. Machine Learning Overview . . . 3
1.2. Supervised Learning . . . 3
1.3. An Overview of Bias and Variance . . . 4
1.4. Assessing Model Performance . . . 5
1.5. Resampling Techniques . . . 6
1.5.1. Cross-Validation . . . 6
1.5.2. Bootstrap . . . 7
1.6. Outline of Thesis . . . 8
2. LITERATURE REVIEW . . . 9
2.1. In-Situ Tests . . . 9
2.1.1. Advantages and Disadvantages of In-Situ Tests . . . 9
2.2. Standard Penetration Test (SPT) . . . 11
2.2.1. General Information . . . 11
2.2.2. Testing Procedure and Equipment . . . 11
2.2.3. Measured Parameters . . . 12
2.2.4. Preceding Correlations from the Standard Penetration Test . . . 14
2.2.4.1. SPT-N and Undrained Shear Strength (c_u) . . . 15
2.3. Pressuremeter Test (PMT) . . . 17
2.3.1. General Information . . . 17
2.3.2. Testing Procedure and Equipment . . . 17
2.3.3. Measured Parameters . . . 18

2.3.4. Preceding Correlations from the Pressuremeter Test 21
3. INTRODUCING MACHINE LEARNING ALGORITHMS 23
3.1. Linear Regression 23
3.1.1. Simple Linear Regression 23
3.1.2. Multiple Linear Regression 25
3.1.3. Determining Statistical Significance of a Linear Regression Model 26
3.1.4. Advantages and Disadvantages of Linear Regression 28
3.2. Random Forest 28
3.2.1. Decision Trees 29
3.2.2. Recursive Binary Splitting 30
3.2.3. Advantages and Disadvantages of Trees 30
3.2.4. Bagged Trees 31
3.2.5. Basics of Random Forest 32
3.3. Boosting 33
3.4. Stacking 33
4. DEVELOPING REGRESSION MODELS USING MACHINE LEARNING APPROACHES 35
4.1. Introducing the Datasets 35
4.2. Data Preprocessing 36
4.2.1. Uncovering Outliers 37
4.2.2. Parameter Selection and Correlation 38
4.3. Tuning the Models 38
4.3.1. Tuning the Random Forest Model 40
4.3.2. Tuning the Gradient Boosting Model 41
4.4. Results of Machine Learning Approaches 43
4.4.1. Results of Linear Models 43
4.4.2. Results of Random Forest Model 47
4.4.3. Results of Gradient Boosting Model 49
4.5. Developing the Stacked Model 51
4.6. Comparison of Performances of Machine Learning Approaches. 52
4.6.1. Comparing Models used to Predict Undrained Shear Strength . 53

4.6.2. Comparing Models used to determine Elastic Modulus and Limit Pressure 54
5. COMPARISON OF MACHINE LEARNING APPROACHES WITH EXISTING CORRELATIONS 57
5.1. Stacking against Existing Undrained Shear Strength Correlations . . . 57
5.2. Linear Model against Existing Undrained Shear Strength Correlations . 59
5.3. Random Forest against Existing Elastic Modulus and Limit Pressure Correlations 60
6. CONCLUSION 64
REFERENCES 66

LIST OF FIGURES

Figure 1.1. Simplified Display of Validation Approach. 6

Figure 1.2. 5 Fold Cross-Validation Approach. 7

Figure 1.3. Bootstrap Resampling Method. 8

Figure 2.1. Frequently Performed In-Situ Tests. 9

Figure 2.2. Standard SPT Sampler [1]. 12

Figure 2.3. Calibration Curves Obtained During Calibration Process [2]. . . . 19

Figure 2.4. Calibration Curves Obtained During Calibration Process. 19

Figure 3.1. Schematic Representation of the Intercept and Slope in a Simple Linear Regression. 24

Figure 3.2. Schematic Representation of a Linear Regression with Two Predictors in a Three Dimensional Space. 26

Figure 3.3. Schematic Representation of Segmentation of the Predictor Space in Decision Trees. 29

Figure 3.4. An Example of a Regression Tree. 31

Figure 3.5. Simplified Structure of Stacked Models 34

Figure 4.1. Histogram of Undrained Shear Strength. 37

Figure 4.2. Correlation Matrix of Parameters in Dataset A. 39

Figure 4.3. Correlation Matrix of Parameters in Dataset B. 39

Figure 4.4. Random Forest Tuning Results. 40

Figure 4.5. Gradient Boosting Tuning Results to Determine Undrained Shear Strength. 41

Figure 4.6. Gradient Boosting Tuning Results to Determine Elastic Modulus. 42

Figure 4.7. Gradient Boosting Tuning Results to Determine Limit Pressure. . 42

Figure 4.8. Relationship between Measured and Predicted Undrained Shear Strength. 45

Figure 4.9. Relationship between Measured and Predicted Elastic Modulus. . 46

Figure 4.10. Relationship between Measured and Predicted Limit Pressure. . . 47

Figure 4.11. Relationship between Measured and Predicted Undrained Shear Strength from Random Forest. 48

Figure 4.12. Relationship between Measured and Predicted Elastic Modulus from Random Forest. 48

Figure 4.13. Relationship between Measured and Predicted Limit Pressure from Random Forest. 49

Figure 4.14. Relationship between Measured and Predicted Undrained Shear Strength from Gradient Boosting. 50

Figure 4.15. Relationship between Measured and Predicted Elastic Modulus from Gradient Boosting Methods. 50

Figure 4.16. Relationship between Measured and Predicted Limit Pressure from Gradient Boosting Methods. 51

Figure 4.17. Correlation Plot of Individual Predictions of the Models. 52

Figure 4.18. RMSE Indices for Various Machine Learning Approaches in Predicting the Undrained Shear Strength. 53

Figure 4.19. RMSE Indices for Various Machine Learning Approaches in Predicting Elastic Modulus. 55

Figure 4.20. RMSE Indices for Various Machine Learning Approaches in Predicting Limit Pressure. 56

Figure 5.1. RMSE of Stacked Model and Existing Correlations. 58

Figure 5.2. RMSE of Linear Model Developed and Existing Correlations. . . . 59

Figure 5.3. RMSE of Random Forest and Existing Correlations of Elastic Modulus. 61

Figure 5.4. RMSE of Random Forest and Existing Correlations of Limit Pressure. 62

LIST OF TABLES

Table 2.1. Previous Correlations Presented By Different Researchers 16

Table 2.2. Correlations of PL and SPT 22

Table 2.3. Correlations of Em and SPT 22

Table 4.1. Summary of Dataset A. 36

Table 4.2. Summary of Dataset B 36

Table 4.3. Summary of Statistics of the Linear Model for Dataset A. 43

Table 4.4. Summary of Statistics of the Linear Model for Dataset A. 44

Table 4.5. Summary of Statistics of the Linear Model for Dataset B Equation 4.3. 45

Table 4.6. Summary of Statistics of the Linear Model for Dataset B Equation 4.4. 46

Table 4.7. Results of Models used to Estimate Undrained Shear Strength. . . 54

Table 4.8. Results of Models used to Estimate Elastic Modulus. 55

Table 4.9. Results of Models used to Estimate Limit Pressure. 56

Table 5.1. Comparison of Models used to Estimate Undrained Shear Strength. 58

Table 5.2. Comparison of Linear Models to Estimate Undrained Shear Strength. 60

Table 5.3. Comparison of Models used to Estimate Elastic Modulus. 61

Table 5.4. Comparison of Models used to Estimate Limit Pressure. 62

LIST OF SYMBOLS

c_u	Undrained shear strength
D_r	Relative density
E_m	Hammer efficiency
$\widehat{f}$	Mapping or Target function
N_{60}	SPT-N value corrected for field procedures
m_v	Coefficient of volume compressibility
q_u	Undrained compression strength
v_s	Shear wave velocity
w_n	Water content
x	Input variable
y	Output variable
$\widehat{y}$	Predicted Response
ϕ	Internal angle of friction

LIST OF ACRONYMS/ABBREVIATIONS

CARET	Classification and Regression Training
CPT	Cone Penetration Test
DMT	Flat Dilatometer Test
FVT	Field Vane Test
GBM	Gradient Boosting Method
LL	Liquid Limit
MLA	Machine Learning Algorithms
MRA	Multiple Regression Analysis
PI	Plasticity Index
PL	Plastic Limit
PMT	Pressuremeter Test
RF	Random Forest
RMSE	Root Mean Square Error
RSS	Sum of Squared Residuals
SPT	Standard Penetration Test
UCS	Unconfined Compression Test
UU	Unconsolidated Undrained Test

1. INTRODUCTION

In the fast paced world of construction, time has become a key factor in the design process. Gone are the days where geotechnical engineering designers would have ample time to properly carry out the necessary research, conduct tests punctiliously in order to give results that would correctly match what is encountered at the field of study. Understanding the behavior of soil has proven to be an immensely difficult prospect due to its peculiar nature of having the ability to reflect different properties at different depths. Thus the best a designer can do, is, to as to the best of his or her capability to estimate the soil behavior in order to put forth a safe, prudent and lasting design.

In order to properly define the soil profile and strength properties of soil media at a potential project site, conventional sub soil investigation methods are conducted. These methods include, drilling of boreholes, collection of disturbed and undisturbed samples and conducting of in-situ and laboratory tests on the collected disturbed and undisturbed samples. In-situ and field tests have been advanced where obtaining samples is difficult or sample disturbing is eminent. Incidentally, laboratory testing has become inadequate and time consuming as the size of project sites increased and project deadlines reduced.

Increase of new in-situ testing equipment and procedures have grown in popularity during this period due to their feasibility and practicality. Many in-situ tests have been developed over time with the most frequently used being the Standard Penetration Test (SPT) and the Cone Penetration Test (CPT). Other tests include the Pressuremeter Test (PMT) , Flat Dilatometer Test (DMT) and Field Vane Test (FVT).

The Standard Penetration Test (SPT), is the most commonly conducted test among the in-situ tests. The SPT can be practically explained as the resistance of the soil to vertical penetration. This resistance is measured by blow counts required to penetrate a certain depth of soil.

The Cone Penetration Test (CPT) has become quite popular as the SPT due to its ability to obtain continuous soil profiles and its ability to provide a reasonable estimate of a variety of geotechnical engineering parameters.

Differently from the SPT and CPT, is the Pressuremeter Test (PMT). With the PMT, soil resistance is measured radially by the use of an inflatable rubber membrane and frequently conducted on pre-bored holes. Strength characteristics can directly derived from the results of the PMT, thus making it an invaluable in-situ test. Furthermore, its relevancy in a wide range of soil and rock media is one of its main superiority over the other tests. Each of these in-situ tests uses different methods and parameters to predict soil behavior. However, unlike laboratory tests, none of the in-situ tests give the required geotechnical parameters directly. Countless empirical correlations have been developed in order to estimate these parameters from in-situ tests despite the enormous number of uncertainties involved with empirical approaches. These approaches combined with in-situ testing and laboratory testing form the basis of geotechnical design.

In the scope of this study, a substantial number of in-situ tests of cohesive soils including SPT and PMT were compiled from sub soil investigations executed in a number of projects across Turkey. In order to come up with correlation equations, laboratory test data from akin depths were also collected. Among the laboratory tests collected were the Atterberg limits; Liquid Limit (LL), Plastic Limit (PL) and Plasticity Index (PI), water content (w_n) and the undrained shear Strength (c_u).

In the past few years, there has been a steady increase in the interest of using machine learning algorithms in various fields of engineering [3–5]. This thesis aims to introduce correlation equations established by using multiple regression analysis (MRA) methods, both linear and non-linear. Among the non-linear regression methods, this study looked at the possibility of using machine learning algorithms (MLA) such as Random Forest(RF) and Gradient Boosting Method (GBM) to estimate the geotechnical parameters.

1.1. Machine Learning Overview

Machine Learning is essentially a subset of artificial intelligence, involving the usage of computer algorithms to autonomously learn from given data. This suggests that computers do not need to be rigorously programmed to put out the best results since machine learning algorithms have the ability to improve efficiency through their learning process.

The term machine learning is relatively new but most of its basic concepts were developed several years ago. Ever since the development of the method of the least squares which implemented the method that is now known as linear regression, researchers have developed various methods of analyzing data quantitatively and qualitatively. During the early 1980s, quite a number of approaches used to learn data had been developed. However, most of them were linear methods since fitting non-linear relationships was considered cumbersome. Briema, Stone, Olshen and Friedmann introduced decision trees for regression and classification purposes. This method demonstrated the capacity of a detailed practical implementation of a method which involves cross-validation for model tuning.

In recent times, machine learning progress has been helped by the availability of user-friendly and extremely powerful tools and softwares such R,Python among other coding languages. In this study, the user-friendly R program which is freely available for download was used to model the machine learning models as well as the multiple regression model.

1.2. Supervised Learning

Supervised learning generally means trying to map an outcome by using its input values. The mapping is essentially done by use of algorithms. The goal of supervised learning is to map the outcome of the input variables as accurately as possible so as to accurately predict the outcome on a different set of input data.

It has been given the term supervised learning because the whole process of the algorithm learning from the training data can be thought of as an instructor supervising a learning process. The correct answers are known, the algorithm makes predictions on the training data and it is corrected by the instructor until an acceptable level of performance is achieved.

1.3. An Overview of Bias and Variance

As mentioned above, a supervised machine learning algorithm learns from a training dataset. The objective of any supervised machine learning algorithm is to estimate the best mapping function,$\widehat{f}$, for an output variable,y, given input data x. The mapping function can also be termed as the target function as it is the function the supervised machine learning algorithm aims at approximating. Machine learning prediction error can be broken down into 3 parts:

- Bias Error
- Variance Error
- Irreducible Error

The irreducible error, from its name, is understood as an error that cannot be reduced regardless of what algorithm is used.

The bias error is simply explained as a model making assumptions of the mapping or target function. This makes the models exhibit low predictive performances on much more complex problems. An example of a high bias model is the linear regression as it assumes the input data and output data can be fitted through a linear model. However, models that make less assumptions about the form of the target function tend to have a high variance. This bring us to the variance error.

The variance error estimates how much the target function, $\widehat{f}$, varies if and when different training data is used. Essentially, the algorithm is expected to have some variance. A low variance suggests small changes to the target function with changes to the training with the vice versa being true.

The objective of any supervised machine learning algorithm is to achieve a bias-variance trade off that can map the target function as accurately as possible which in turn will enable us to obtain an acceptable prediction performance.

1.4. Assessing Model Performance

Assessing the performance of machine learning models on given data is straightforward. Since the problem at hand is a regression program, like previous researchers the root mean square error (RMSE) of the models is considered in determining model performance [4, 6]. An excellent prediction is represented with a 0 the RMSE values. The RMSE value is determined as following:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2} \tag{1.1}$$

In assessing the performance of our models, we would like to see how our models would fair when they are given data that has not been seen by the algorithms i.e test data. This is a much more non-biased way of evaluating how well the models have been able to map the target function. Finally, to determine the validity of the equations and models from the linear and machine learning models, equations developed by previous researchers are compared using the test data with the ones determined from this study.

1.5. Resampling Techniques

Resampling techniques are essential tools in modern statistics. They comprise of drawing repeatedly from a sample of training data and refitting a model so as to acquire more information of a model. An example can be trying to obtain the variability of a linear regression by repeatedly drawing different samples from the training data, fitting a linear regression and comparing the results obtained so as to see how much they differ. Such an approach may allow us to acquire information that would not be available from fitting the model only using the original training sample.

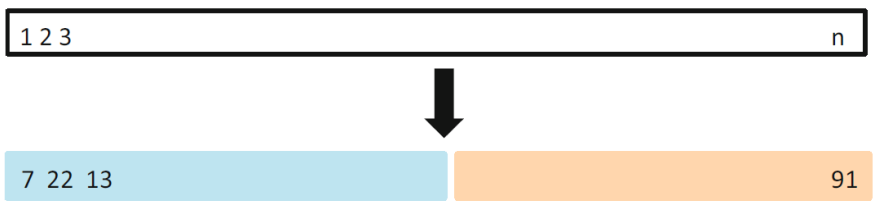


Figure 1.1. Simplified Display of Validation Approach.

The downside to using resampling approaches is that it can be computationally expensive, as it involves drawing data and fitting it on a model multiple times. However, thanks to recent improvements in the computational powers of computers resampling methods are no longer suppressive. Here we will discuss two of the most common resampling techniques which have been used in this thesis.

1.5.1. Cross-Validation

Cross-validation can be used to estimate the test error involved with any statistical learning method. Cross-validation involves splitting the data into two sets randomly, fitting the model on one and testing for its error in the other. The validation set as mentioned earlier is drawn from the training data set and is different from the test data set [7].

Figure 1.1 shows a data set that has been split randomly into two different parts. The dataset in blue is used to fit the model while the one in orange is used to check the accuracy of the fitted model. The validation error of such an approach may be highly variable hence less reliable. To mitigate such a problem, we can randomly split the data into a k separate fold, train on k-1 folds and test on the fold that has been held out. Then the average of the accuracies of the models can be taken into account as a more reliable estimate of the errors. Figure 1.2 shows an example of such an approach. In this schematic display a 5 fold cross-validation procedure was performed [7].

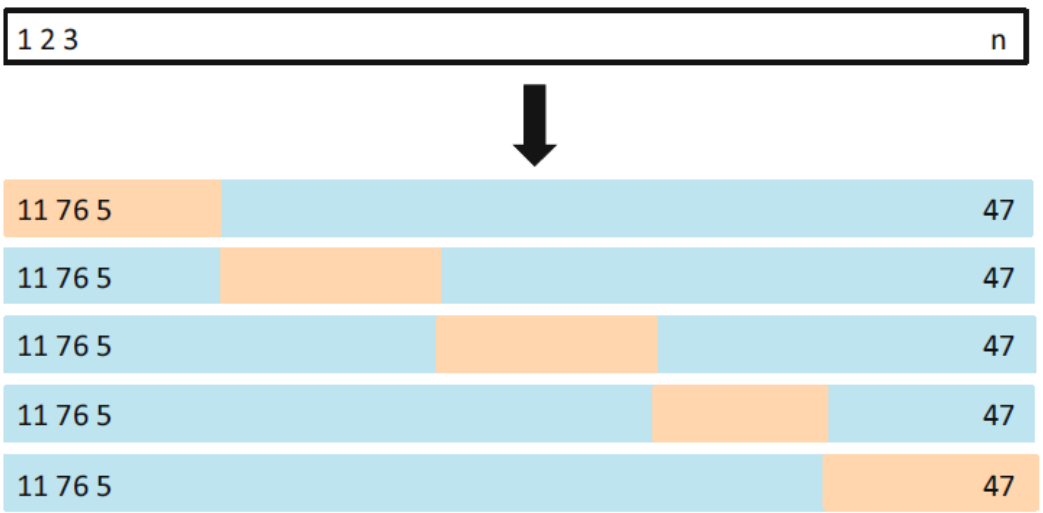


Figure 1.2. 5 Fold Cross-Validation Approach.

1.5.2. Bootstrap

Bootstrap is essentially a resampling technique which involves continuously taking random samples from a training data with replacement. This simply means that selected data may appear more than once in the selected subset [8].

A bootstrap sample will have the size of the original data set from which is being drawn. Consequently, various observations will be represented in the sample and others will not, these are called out of bag samples. A model will be built on the selected samples and evaluated on the out of bag samples.

Bootstrapping is an essential tool used in building the Random forest model which will be discussed in the upcoming chapters. An example of how bootstrap resampling is done is presented in Figure 1.3.

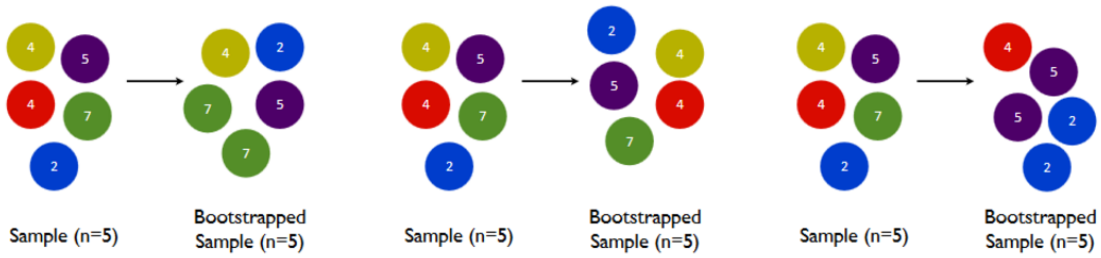


Figure 1.3. Bootstrap Resampling Method.

1.6. Outline of Thesis

This research aims to introduce new approaches and methodologies by using machine learning algorithms (MLA) and multiple regression analysis (MRA) models to predict geotechnicals strength parameters.

Chapter 2 includes a brief summary of the tests whose data has been used to determine the correlation equations developed in this research. Moreover, it also shows the correlation equations that have been developed over the years by different researchers in the field of civil engineering. Chapter 3 focuses on introducing the linear regression models and machine learning algorithms that have been used in this thesis. Additionally, properties of each model together with their advantages and disadvantages are discussed in this chapter. Furthermore, the tuning parameters for each model are explained and optimized. Chapter 4 is dedicated to introducing the dataset used in developing the machine learning models, giving results and comparing the results obtained from the different machine learning models utilized. Chapter 5 sees the comparison of the best performing models with the frequently used correlation equations from the literature. Lastly, the final chapter focuses on conclusions of the research and suggestions for future methodologies that could be applied in order to improve the performance the machine learning algorithms.

2. LITERATURE REVIEW

2.1. In-Situ Tests

In situ testing can be simply expressed as laying an instrument in a precise point in a borehole or on the ground surface so as to establish the properties of the soil or rock media in its natural stress condition. Most of the in-situ test are primarily penetration methods which allow them to be swift and cost effective. Frequently used in-situ tests in the world of geotechnical engineering are seen in Figure 2.1.

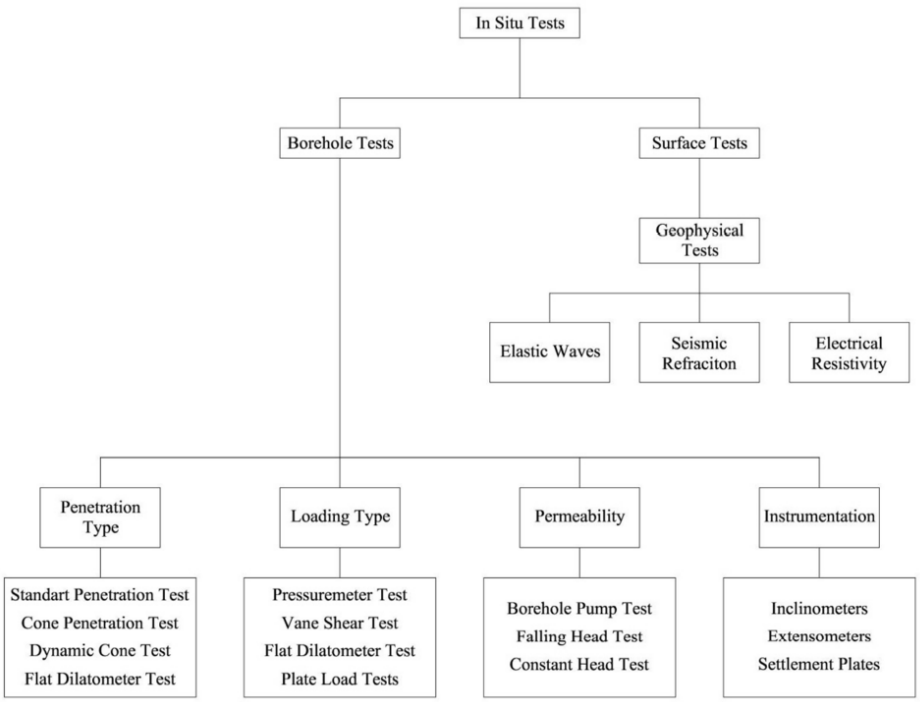


Figure 2.1. Frequently Performed In-Situ Tests.

2.1.1. Advantages and Disadvantages of In-Situ Tests

In-situ testing is an efficient way of determining different soil properties. Moreover, they do have significant advantages over laboratory testing, but, also hold some drawbacks which can greatly affect the design process.

Typical advantages and drawbacks of in-situ testing contrasted with laboratory testing are listed below:

(i) Advantages

- Larger percentage of soil is represented.
- Continuous soil profiling can be attained from a couple of the tests.
- They are applicable to both soil and rock media.
- Tests are conducted under natural stress environment which is significant in determining the parameters of the media on the potential project site.
- Most in-situ tests are cost effective and less time consuming.

(ii) Drawbacks

- Nature of the soil cannot be determined in all the tests while index properties can be only be determined from a disturbed sample of the Standard Penetration Test (SPT).
- Stress and deformation effect are not clear for most of the tests except for the Pressuremeter Test.
- Inconsistent results may be achieved for the same type of soils.
- Drainage conditions during the testing cannot be controlled.

With the help of correlation equations, results obtained from in-situ tests can be used to predict or estimate necessary geotechnical parameters that are to used in the design process.

Correlations used to determine the parameters together with brief introductions of the some of the in-situ tests are discussed in the following sections of this chapter.

2.2. Standard Penetration Test (SPT)

2.2.1. General Information

The Standard Penetration Test (SPT) is one of the oldest and most frequently used of the in-situ tests. Initially developed in the late 1920s and extensively used in North and South America, Great Britain, Japan and elsewhere. It is conducted in an exploratory boring using inexpensive, readily available equipment, hence adding little cost to the subsurface exploration program.

2.2.2. Testing Procedure and Equipment

The test was only standardized back in 1958 when the ASTM standards initially appeared [1]. It is typically:

- A standard sampler with dimensions as given in Figure 2.2 is driven into the ground by energy delivered from a 63.5kg weight hammer dropped from a distance of 760mm.
- The process is repeated until the sampler has penetrated a distance of 450mm.
- Hammer blows required to penetrate each interval of 150mm are recorded. The test is stopped if the blows required to penetrate a certain 150mm interval exceed 50, or if more than 100 blows are required for the entire 300mm.
- The SPT N or N_{30} value is calculated by adding up the sum of the blows required to penetrate the final 300mm.
- The procedure is repeated after boring to the next depth test is reached. Typically these tests are performed at intervals of 1.5 - 5m

One of the advantages of the SPT test, is that after performing the test, one extract the sampler, remove and save the soil sample for classification and conduction of index tests on it.

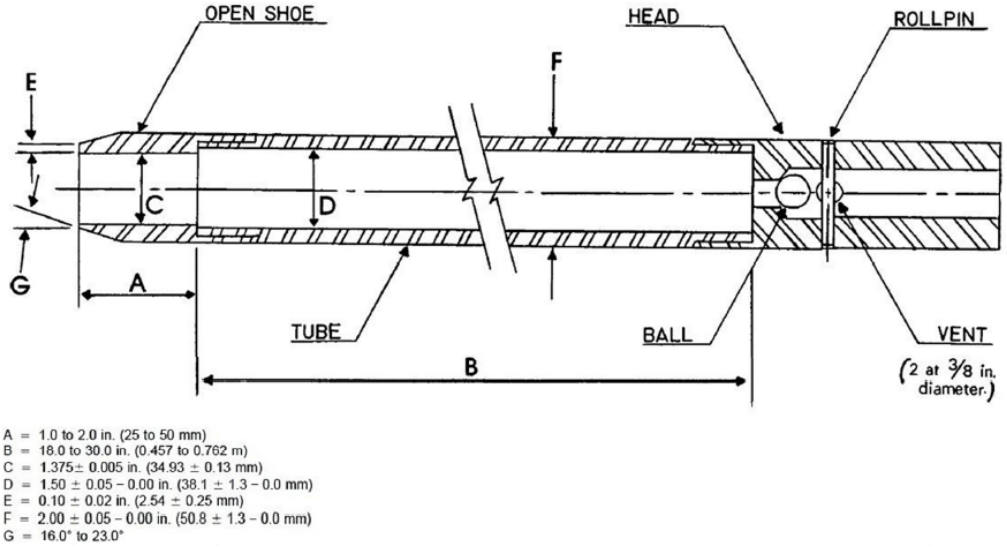


Figure 2.2. Standard SPT Sampler [1].

The reasons for the wide usage of the SPT in subsoil investigations can be accounted to many factors such as the already mentioned readily available equipment, directness of the operation, appropriateness in a variety of soil media and ability of sampling.

For its practical aspects, results of the SPT test can be drastically affected by drilling operations, competence of the operator, existence of coarse particles and ground water conditions.

2.2.3. Measured Parameters

Results of SPT-N are used to calculate imperative engineering properties of coarse-grained soils such as internal angle of friction (ϕ), relative density, D_r, bearing capacity and settlement. It can be used to calculate the shear wave velocity (v_s), liquefaction potential and also as a control for compacted fills.

Even though the SPT test was initially intended to be used in coarse-grained soils, it can also be used to determine certain properties of fine-grained soils such as undrained shear strength (c_u), undrained compressive strength (q_u), and coefficient of volume compressibility (m_v) [9].

Since the equipment and operating conditions vary, direct use of SPT results in design is not recommended [10]. However, ASTM standards recommend that the measured SPT-N value (N_{30}) should be standardized by ratio (C_E) between the energy measured transfered to the rod ($E_{measured}$) and 60% of the theoretical potential energy ($E_{theoretical}$).

$$C_E = (E_{measured}/E_{theoretical})/60 = ER/60 \tag{2.1}$$

This compensates for the different rods and different rigs used during the SPT test, hence, making the results more reliable when estimating parameters to be used in design.

There are a few number of factors that affect the validity of the SPT results [11], and for that matter the obtained penetration resistance may be too high or too low. High values result in estimating nonconservative results, while low values result in estimation of over conservative results of soil properties and bearing capacity. It is for this reason that corrections should be made to SPT results before using them to estimate engineering properties.

We can improve the raw SPT data by applying certain corrections,thus greatly improving its repeatability. The variations in testing procedure may be somewhat compensated by converting recorded SPT-N (N_{30}) values to N_{60} using equation 2.2 [12]:

$$N_{60} = \frac{E_M C_B C_S C_R N}{0.60} \tag{2.2}$$

where:

N_{60} = SPT-N value corrected for field procedures
E_M = Hammer Efficiency
C_B = Borehole Diameter Correction
C_S = Sampler Correction
C_R = Rod Length Correction
N = SPT-N value recorded in the field

Furthermore, to obtain the $N_{1,60}$ values, a correction of overburden pressure of 100kPa and a 60% of the theoretical free-fall hammer energy is applied. This can be simply explained by equation 2.3.

$$N_{1,60} = C_N N_{60} \tag{2.3}$$

where:

C_N = Overburden Correction Factor

The SPT-N value is used in a few number of empirical correlations to determine engineering properties of soil media to be used in design [13–16]. Although the equations present in the literature are known, little exists regarding what sort of SPT corrections were done or what sort of regression analysis method was undertaken.

2.2.4. Preceding Correlations from the Standard Penetration Test

In todays fast paced construction world, an engineer has to deal with two main factors, time and cost effectiveness. Getting soil the required soil parameters to start the design requires both time and money. Therefore, it is to the advantage of the engineer to use the correlations by using a small number of soil parameters that can be easily obtained.

Correlations are essential in estimating engineering properties of soils, particularly where projects are under a tight financial budget and need to completed within the shortest period. However, usage of correlations from the literature is not always clear. There are generally four uncertainties that arise from the use of the correlations [9]. These uncertainties include:

- whether the correlations have any corrections or not, and if they do,which corrections have been made.
- whether the correlation has a statistical background.
- which test results are to be used.
- which type of soil is the correlation credible for.

Therefore, when using a correlation equation, one must always question and answer the aforementioned uncertainties.

2.2.4.1. SPT-N and Undrained Shear Strength (c_u). It is essential to determine the undrained shear strength of fine-grained soils in order to calculate their bearing capacities as well as to calculate stability analysis for structures and slopes. c_u is determined primarily through laboratory tests such as unconsolidated undrained (UU) triaxial test. In addition for saturated fine-grained soils, the undrained shear strength can be obtained by taking half of the unconfined compressive strength from the unconfined compression test ($c_u = q_u/2$). Many researchers have over the years studied and tried to come up with different correlation between the SPT and C_u [9, 15, 17]. A summary of the correlations most frequently come across is presented in Table 2.1.

Table 2.1. Previous Correlations Presented By Different Researchers.

Researcher(s)	Explanations	c_u (kPa)
Sanglerat [18]	Clay	12.5N
	Silty Clay	10N
Terzaghi and Peck [15]	Fine-grained soil	6.25N
Nixon [19]	Clay	12N
Decourt [19]	Clay	12.5N
		$15N_{60}$
Sivrikaya and Togrol [9]	Highly Plastic Soil	$4.85N_{field}$
		$6.82N_{60}$
	Low Plastic Soil	$3.35N_{field}$
		$4.93N_{60}$
	Fine-grained Soil	$4.32N_{field}$
		$6.18N_{60}$
Ajayi and Balogun [9]	Fine-grained soil	1.39N+74.2
Hettiarachchi and Brown [20]	Fine-grained soil	$4.1N_{60}$
Sivrikaya [3]	UU Test	3.33N -0.75w_n + 0.20LL +1.67PI
	UU Test	4.43N_{60} - 1.29w_n + 1.06LL + 1.02PI

As seen from Table 2.1, uncertainties may arise when using some of the correlations. Researchers such as Sivrikaya and Togrol, Hettiarachchi and Brown and Decourt have explicitly stated what results of the SPT test have been used to determine the undrained shear strength (c_u). Furthermore, Sivrikaya offers a different approach when computing the undrained shear strength. He unlike the other researchers also decided to see how the water content (w_n) and Atterberg limits (LL, PL and PI) affect the estimation of the undrained shear strength.

This study attempts to come up with different correlation based data collected from projects completed around Turkey. In addition, it will compare the validity of some of the equations found in Table 2.1 to estimate the undrained shear strength (c_u) of the database used in this study.

2.3. Pressuremeter Test (PMT)

2.3.1. General Information

A pressuremeter can be defined as a cylindrical probe that has an expandable flexible membrane designed to apply a uniform pressure to the walls of a prebored borehole [21]. Invented in 1954 by Louis Menard, the pressuremeter test has become on the most sort after in-situ test during subsoil investigations. The initial concept developed by Menard was the inflation of a cylindrical balloon inside a pre-bored hole so as to measure the deformation properties of the soil media. The PMT is conducted in hard clays, dense sands and weathered rock. After developing the PMT device, Louis Menard attested to it being one of the most precise testing methods available for any type of soil [22].

2.3.2. Testing Procedure and Equipment

The Pressuremeter is made up of three main parts which are, a probe, a monitoring box and tubing for inflation.

- Probe: A conventional Menard pressuremeter probe include three separate cells; top cell, loading cell and bottom cell. The top and bottom cells are usually referred to as guard cells. These cells are filled with gas in order top protect the loading cell. The load cell is a flexible membrane that is filled with water after the guard cells are filled with gas.

- Monitoring box: This part of the pressuremeter is placed on the ground, preferably close to the borehole. Its main purpose is to regulate the pressure given to the probe inside the borehole, record and monitor the volume changes with respect to pressure increase by use of the dial gauges on it.
- Tubings:From the wording if this part one can easily guess that this part is responsible for delivering gas and water to the guard cells and loading cells respectively.

Aside from the originally developed Menard type pressuremeter, self boring pressuremeters and cone pressuremeters have been developed over time.

The Pressuremeter test is performed either by stress controlled method, where pressure application is applied in equal increments or by the strain controlled method, where volume application is of equal increment. Before the test begins, two main calibrations are performed. These calibrations include:

- Volume Calibration: This calibration is performed to check leaks in the system and to make the necessary adjustments required. The probe is usually placed in a steel tube before the volume calibration is done. The pressure is increased in steps. For a given pressure, the lost volume is determined since the probe is in a confined by the tubes.
- Pressure Calibration: This calibration is performed to determine the resistance of the rubber membrane to expansion. The probe is taken out of the steel tube and calibration performed under atmospheric conditions. A typical calibration graph is presented in Figure 2.3 [23].

2.3.3. Measured Parameters

After the test is conducted, volume changes recorded during the test are plotted against the pressure considering necessary corrections have been made based on the calibrations. The corrected pressuremeter graph usually obtained is given in Figure 2.4.

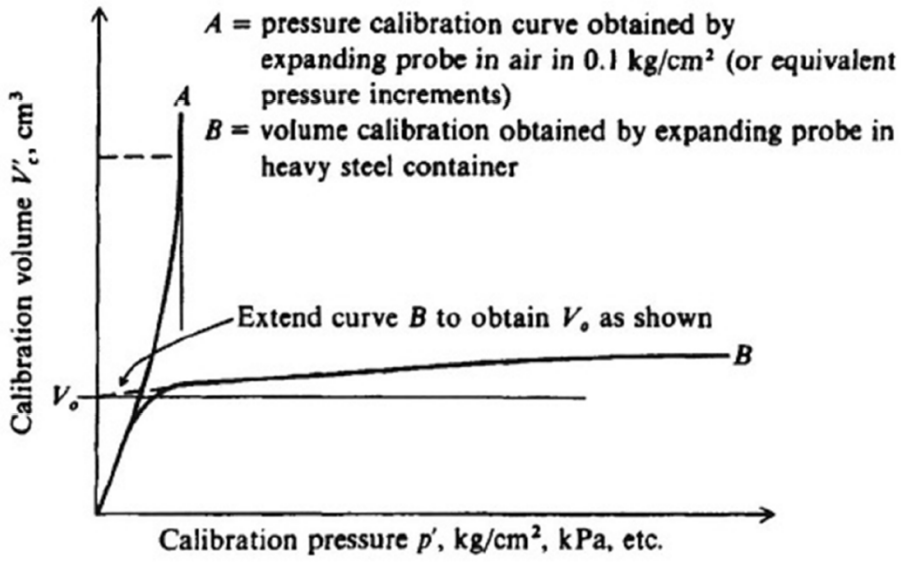


Figure 2.3. Calibration Curves Obtained During Calibration Process [2].

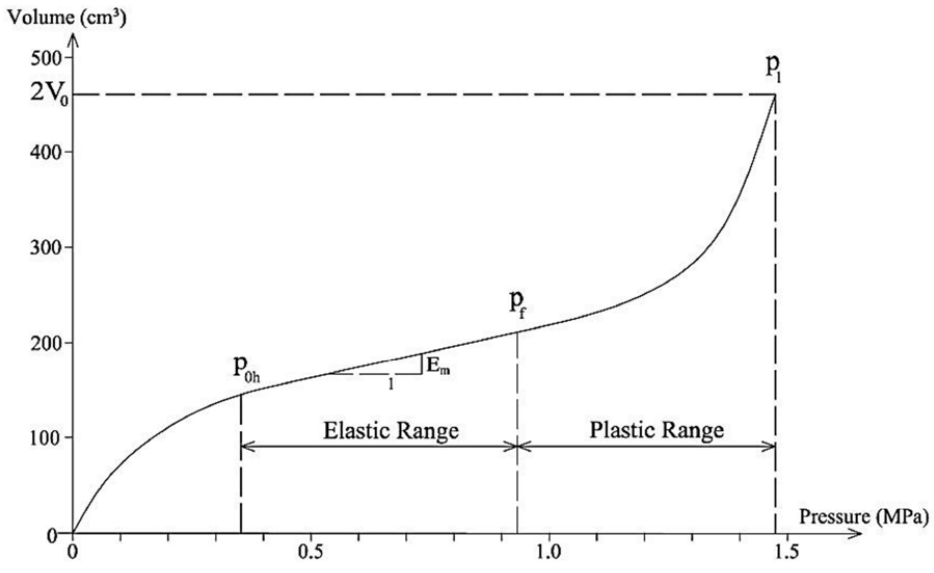


Figure 2.4. Calibration Curves Obtained During Calibration Process.

As seen from Figure 2.4, three values are recorded in order to obtain the graph. Pressure p_{ho}, is known as the initial horizontal pressure on the ground. At this pressure, it is assumed that the membrane is in full contact with the soil around the wall of the borehole.

As the pressure increases, the pressure-volume curve becomes almost linear, which is a result of the elastic behavior of soils, hence described on the graph as the elastic range. With further pressure increase, permanent deformations occur and volumetric expansion in the soil increases greatly.

Another measured parameter from the graph is the limit pressure, p_L. The limit pressure is defined theoretically as the pressure for which an infinite expansion is of the probe is expected [24]. It is assumed that soil failure occurs at this pressure point. This pressure is achieved at volume equal to $2v_0$, where v_0 is the initial volume required to inflate the pressuremeter.

Although the limit pressure p_L, defines the failure point that occurs in the soil, the net limit pressure p_{Ln} is frequently used in practice due to its crassness to disturbances in the borehole [24, 25]. The net limit pressure, p_{Ln}, is calculated as:

$$p_{Ln} = p_L - p_{ho} \tag{2.4}$$

Accurately determining the value of p_{ho} from the test data is difficult due to the disturbances in the borehole. For this reason, the following equation can be used as well:

$$p_{ho} = [(\gamma - u)z]K_o + u \tag{2.5}$$

where:

γ = Unit weight of soil being tested
u = Pore water pressure at testing depth
z = Depth of test level from ground surface
K_o = Earth pressure coefficient at rest

The Pressuremeter test is widely used for foundation designs because the method of the test resembles the behavior of a foundation. The settlements of foundations can be estimated by using the deformation modulus, E_{PMT}. This modulus can be determined from the elastic phase of the graph in Figure 2.4. Since the E_{PMT} is a function of Poisson's ratio, slope and cavity volume in the elastic range, it can be found using the equation below:

$$E_{PMT} = 2(1-\nu)(V_o + v_m)\frac{\Delta p}{\Delta v} \tag{2.6}$$

where:

ν = Poisson's ratio, typically taken as 0.33
V_o = Initial volume of probe
v_m = Average volume of probe over the considered stress range i.e $(v_o + v_f)/2$
Δp = Pressure change in the elastic range
Δv = Volume change in the elastic range

2.3.4. Preceding Correlations from the Pressuremeter Test

The literature is not filled with correlations of obtaining the Pressuremeter parameters from field tests. However, from the few studies that have been undertaken relationships between SPT parameters and PMT parameters have been determined [26–28].

Bozbey and Togrol in 2010 did present a relationship between N_{60}, E_{PMT} and p_L based on a study in Istanbul, Turkey [29].

They developed their relationships based on 182 tests carried out in both sandy and clayey soils. Gonin in 1992 also developed correlations between SPT, E_{PMT} and p_L for nine different French soils [30]. The literature does also show a non uniform relationship between the SPT and PMT parameters. This can be accounted for by factors such as type of soil, ranges of the N, E_{PMT}, p_L and the geological conditions of where the tests are conducted. Table 2.2 presents some of the correlations used to predict the limit pressure, p_L , from the SPT parameters. The equation proposed by Yagiz uses the N_{cor} to determine the limit pressure. However, the corrections made to the SPT-N value obtained are not mentioned. The same problem applies to the equations proposed the rest of the researchers, except for Bozbey and Togrol and Kayabasi. They explicitly state that N_{60} is used to estimate the limit pressure.

Table 2.2. Correlations of Limit Pressure and SPT From Different Studies.

Researcher(s)	Explanations	Proposed Correlation
Hobbs and Dixon [31]	Clay	$p_L = 0.021\text{N} + 0.33$ (kPa)
Waschkowski [6]	Clay	$p_L = 0.0561\text{N} - 0.092$ (kPa)
Yagiz [26]	Clay	$p_L = 29.45N_{cor} + 219.7$ (kPa)
Bozbey and Togrol [29]	Clayish Soil	$p_L = 0.26N_{60}^{0.71}$ (MPa)
	Sandy Soil	$p_L = 0.33N_{60}^{0.51}$ (MPa)
Kayabasi [6]	Clay	$p_L = 0.0425N_{60}^{1.1965}$ (MPa)

Table 2.3. Correlations of Elastic Modulus and SPT from Different Researchers.

Researcher(s)	Explanations	Proposed Correlation
Yagiz [26]	Clay	$E_m = 388.67N_{cor} + 4554$ (kPa)
Bozbey and Togrol [29]	Clayish Soil	$E_m = 1.61N_{60}^{0.71}$(MPa)
	Sandy Soil	$E_m = 1.33N_{60}^{0.77}$(MPa)
Kayabasi [6]	Clay	$E_m = 0.29N_{60}^{1.4}$ (MPa)

3. INTRODUCING MACHINE LEARNING ALGORITHMS

There are a wide variety of machine learning algorithms to use for regression purposes, this thesis uses only three of them. Linear regression model, Random forest and Gradient boosting algorithms.

3.1. Linear Regression

Linear regression is a very simple and old approach for supervised learning. It is particularly useful tool for predicting qualitative responses. One can describe this approach as dull compared to the more fancy statistical learning algorithms that are now available, but one can not overstate the importance of properly understanding the key ideas involved with linear regression before jumping off to the more eye catching supervised learning algorithms. Linear regression can be divided into two simple segments:

- Simple linear regression
- Multiple linear regression

3.1.1. Simple Linear Regression

Simple linear regression, from its name is a very straightforward method of predicting a quantitative response. It aims at predicting a response y, from a single predictor, x, i.e it maps the target function, $\widehat{f}$, of the input by looking at the output. Mathematically such a relationship can be expressed as:

$$\widehat{y} = \beta_1 + \beta_2 x \tag{3.1}$$

where $\widehat{y}$ indicates the prediction of response y based on input parameter x. The hat symbol here is used to denote the prediction made by our model and to denote an unknown coefficient or parameter.

Our aim is to obtain coefficient estimates β_1 and β_2 such that the model fits the available data as well possible. To phrase the previous statement more simply, we want to find an intercept β_1 and slope β_2 that results in a line that closely follows the data points.

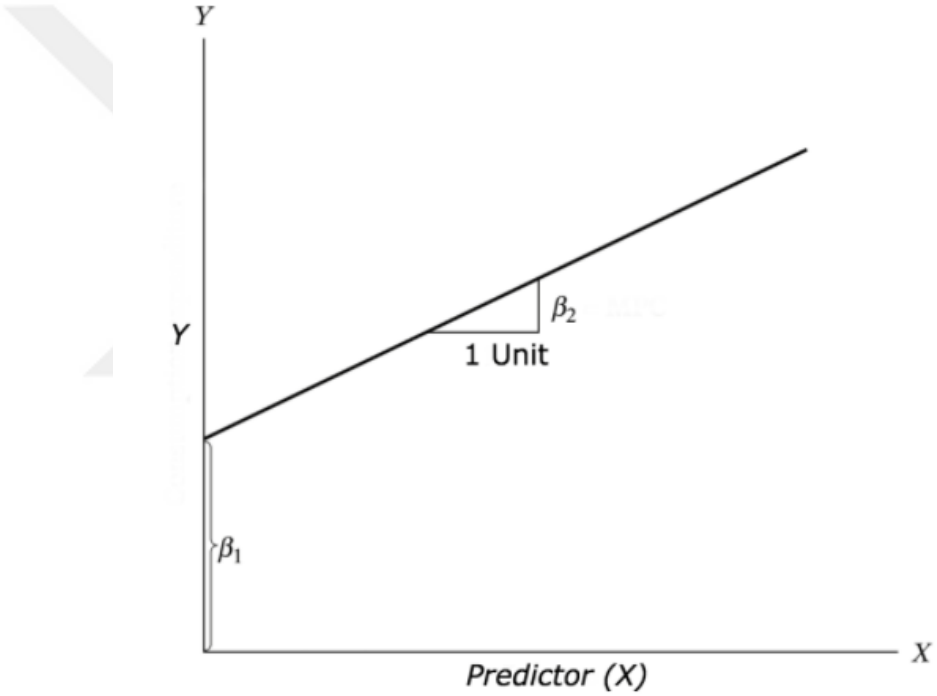


Figure 3.1. Schematic Representation of the Intercept and Slope in a Simple Linear Regression.

Having come up with a linear, one may assume that they could directly use the formula to estimate responses. However, there is the need to check the statistical significance of the coefficients that have been developed. These statistical significance checks will be discussed in the upcoming sections.

3.1.2. Multiple Linear Regression

Simple linear regression as mentioned earlier, is a useful approach for predicting a response on the basis of having a single predictor. However, in the real world and in practice this is rarely the case. To mitigate this problem, we can extend the simple linear regression to accommodate the multiple predictors available. Once this is achieved it is referred to as a multiple linear regression model. The multiple linear regression model will then take the form

$$\widehat{y} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p \tag{3.2}$$

To estimate the regression coefficients of both the simple and multiple regression models, we use the least squares approach. Essentially what we are trying to achieve can be expressed in equation 3.3.

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \tag{3.3}$$

where:

RSS = Sum of squared residuals
y_i = Actual response value
$\widehat{y_i}$ = Predicted response value

We would therefore like to choose coefficients that minimize the sum of squared residuals [7]. Figure 3.2 shows a three dimensional space of a multiple regression model. For such a model the least squares regression line becomes a plane. This plane is chosen in a way that is reduces the vertical distances of each points (shown in red) and the plane.

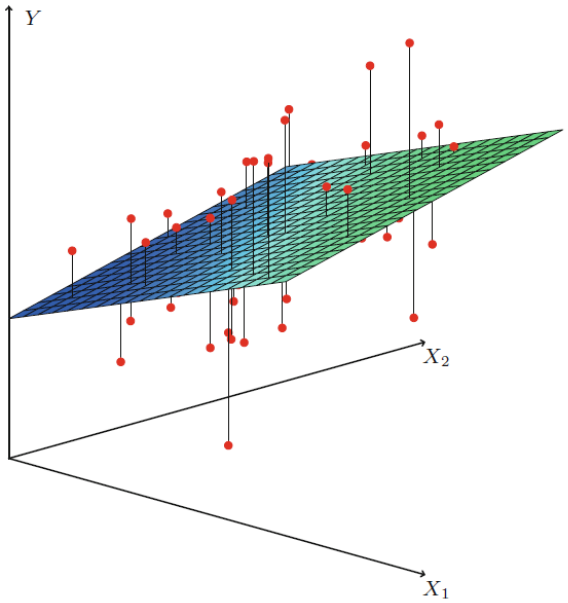


Figure 3.2. Schematic Representation of a Linear Regression with Two Predictors in a Three Dimensional Space.

3.1.3. Determining Statistical Significance of a Linear Regression Model

Recall that we mentioned that upon developing our models, we may not immediately dive in into applying it in our practices, we must first determine if the models are statistically significant. Basically this means that the possibility of a relationship between two or more variables is determined by something other than chance.

Statistical hypothesis testing among other significance tests are used to determine the significance of a relationship determined by a dataset. Here we shall discuss some of tests carried out to determine if a model is statistically significant.

The p-values are of great importance, a linear can only be considered as statistically significant when the p-values are less than the pre-determined significance level of 0.05.

When a p-value is involved, there is null and alternative hypothesis that comes with it. In Linear regression, the null hypothesis refers to the coefficients of the variables being equal to zero i.e $\beta 2 = \beta 3 = ... = \beta p = 0$. Whereas the alternative hypothesis refers to the coefficients not being equal to zero. The existence of an alternative hypothesis indicates that there exists a relationship between the predictor and the response.

How do we then determine if there is a null hypothesis or an alternative hypothesis. This is simply achieved by checking the p-value. If the p-value is less than 0.05 (p-value < 0.05), then we can safely reject the null hypothesis and conclude that our model is indeed statistically significant. It is vital that our model be statistically significant before going ahead and using it to predict future responses, otherwise, the confidence in the predicted responses is mightily reduced and may be described as an event of chance [32].

In linear regression, the coefficient of determination, R^2 , is mostly what people look at to see how well their model has performed. However, for this study we have chosen to use a different goodness of fit statistic that is the standard error of the regression. This can be more helpful and easily understandable. Furthermore, s gives us more valuable information than the coefficient of determination does.

The standard error for regression represents the average distance the observed values fall from the regression line or plane when there are multiple predictors. Smaller values of this statistic indicate that the observed values are closer to the fitted line. An advantage of the standard error of the regression has over the R-squared is the practicality and intuitiveness of using the natural units of the response variables. One can easily see how close or far are the predicted responses to the observed ones by simply checking the s statistic.

3.1.4. Advantages and Disadvantages of Linear Regression

As stated at the start of this chapter, an introduction of the models would be followed by their advantages and disadvantages. A few of the advantages and disadvantages of the linear regression will be listed here.

(i) Advantages
- It is a simple approach that is easy to understand
- It shows optimal results when the predictor and observed response are almost linear

(ii) Disadvantages
- It makes an assumption that the predictor and response have a linear relationship, this makes it prone to producing poor models if the the relationship is non-linear
- It is rigid way of producing prediction based models

3.2. Random Forest

The Random forest is a powerful machine learning algorithm used for both regression and classification problems. It essentially a tree-based method which involving breaking up the predictor space into a number of simple regions. So as to make a prediction of a certain observation, the mean of the training observation of the region to which it belongs is given. Since the splitting criterion used to break up the predictor space can be summarized in a tree, these methods are known as decision tree methods [7].

Decision tree methods are not competitive compared to the other supervised machine learning approaches. However, more competitive models such as bagging and Random forest, which involve producing multiple trees which are then combined to produce a single prediction have been developed [7]. We first look at the basics of decision trees.

3.2.1. Decision Trees

Decision trees try to simplify problems by segmentation. They build rectangular spaces in the predictor space and give region specific responses. What do we mean by region specific responses. Figure 3.3 shows how the predictor space is split into regions namely R_1, R_2, ... , R_5 using splitting criteria t_1, t_2, t_3 and t_4 [7]. For all the observations in a specific region R_k, a similar response is given.

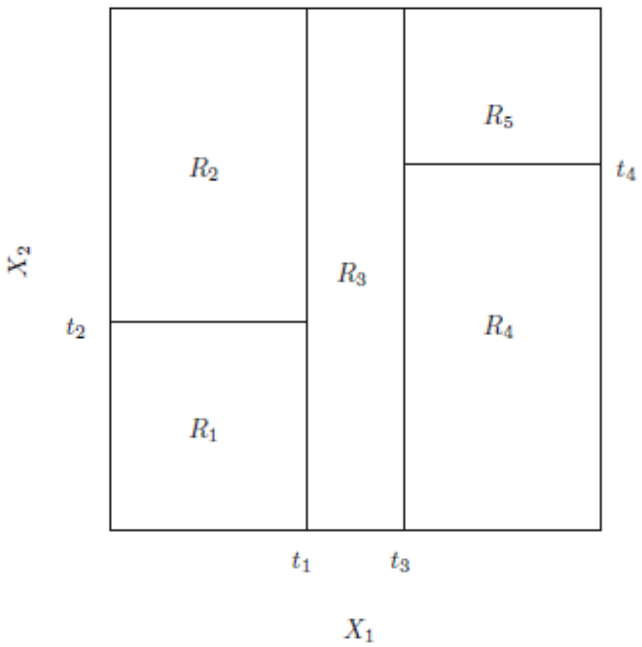


Figure 3.3. Schematic Representation of Segmentation of the Predictor Space in Decision Trees.

The regression tree is built by following the two steps explained below.

- Dividing the predictor space, that is for a possible set of variables X_1, X_2, ..., X_p, into K non-overlapping regions, R_1, ..., R_k.
- A similar prediction is made to all the observations that fall within the region. This simply means the mean of the response values for the training observations.

To elaborate more on these steps, suppose in the first step we obtain two regions, R_1 and R_2, and their means of the training responses are 10 and 20 respectively. Then for a given observation X=x, if x happens to fall in R_1, we will predict 10 as the response. Were it to fall in R_2, we will predict 20 as the response.

The question remains, how do we construct the regions. Similarly to how the coefficients of linear regression are determined, the regions are determined by minimizing the RSS of the region. A top-down greedy approach known as recursive binary splitting is used.

3.2.2. Recursive Binary Splitting

In order to perform recursive binary splitting, we select a predictor X_j and a cutting point s that will lead to least possible value of RSS in the segmented region. Next, we repeat the process looking for the cut point and best predictor of the available data so as to further reduce the RSS in the regions. This time, instead of splitting the entire space, we split one of already formed regions. We now have developed three regions. To further reduce the RSS we split any of these three regions. This process can continue until say no region contains no more than ten or even five observations.

3.2.3. Advantages and Disadvantages of Trees

Decision trees used for both regression and classification have advantages and disadvantages. We take a look at these here.

(i) Advantages

- Trees are easy to explain.
- They mirror human decision making.
- Trees can be graphically displayed and even a non-expert can interpret them.

(ii) Disadvantages

- The biggest disadvantage of trees is that they do not have the the predictive capabilities of other regression models.

However, by combining different decision trees using methods like bagging, random forests and boosting, the predictive capabilities of the trees can be drastically improved. We will now aim to shed some light on these methods.

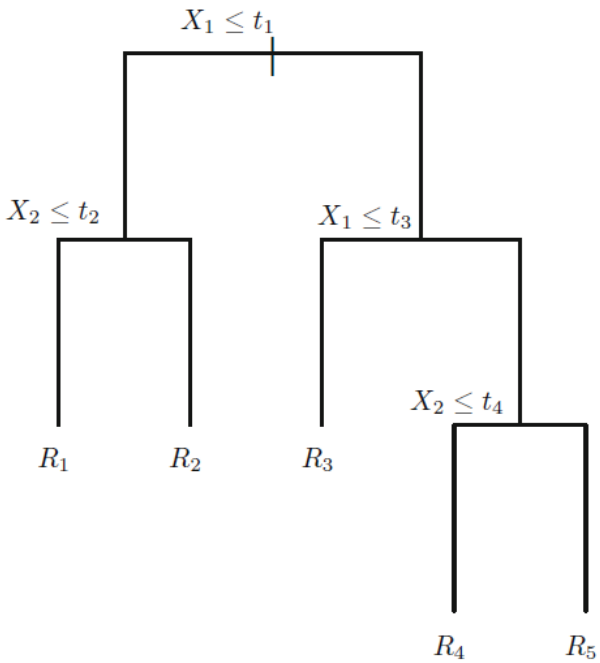


Figure 3.4. An Example of a Regression Tree.

3.2.4. Bagged Trees

The decision trees explained above suffer from high variance. This simply translates to that if we split the training data into two halves randomly and fit decision trees on them, the results could show significant differences. A conventional way of reducing the variance is to build multiple prediction models using multiple training data sets and average the resulting predictions.

This is not a very practical way of doing this since we do not have an infinite number of training sets. One way of mitigating this problem is by using the resampling method known as bootstrap. Recall the bootstrap method involves repeatedly taking samples from single training data set with replacement. We can create B different bootstrapped training datasets, train our model on each other bootstrapped datasets and average the predictions.

$$\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{b}(x) \tag{3.4}$$

where B is the number of bootstrapped samples, $\widehat{f}^{b}(x)$ is the result of the bth sample and $\widehat{f}_{bag}(x)$ is the average of all the predictions of the bootstrapped samples.

3.2.5. Basics of Random Forest

The Random forest is a tree-based model too that uses the bagged tree approach. Just like bagged trees it uses bootstrap resampling to create multiple training sets and fits trees to each bootstrapped training data set. The only difference between the bagged tree and random forest , is that at each split, it randomly picks m predictors and only searches within these randomly selected predictors for the best possible split [33].

Typically, different software packages have various default parameters for the m value. For regression purposes generally $m = p/3$, where p is the number of predictors. However, using the cross-validation introduced earlier one can tune to model to find the optimum m value. Note that the m value cannot exceed the number of predictors present in your training data set.

The reasoning behind using a randomly selected group of predictors rather than using the entire predictor space is that it is generally the case that one of the predictors is the best one to be used at the top of the tree. This results in having similar trees, and averaging similar trees does not reduce the high variance of the predictions.

In using a random set of predictors to decide the best split at the top of the tree, the Random forest enables trees to be formed using the more weaker predictors as well. Averaging these predictions results in less variable and more reliable predictions.

3.3. Boosting

Boosting is another avenue for improving the prediction results of a decision tree. Like bagging which has been discussed earlier, boosting can be applied to different statistical learning methods for both regression and classification problems .

Recall bagging involves generating multiple training data set from the original set using bootstrap. This is followed by fitting decision trees to each of these bootstrapped samples, and finally combining all these fitted trees to create a single model. Boosting works in a similar way, except this time each one of the trees grown is grown by using information from the previously grown trees. This is process can be termed as sequential growing. This sequential growing of the trees in the boosting approach allows the algorithm to learn slowly [7].

Let us break down this process a little further for us to form a clear picture. Given a model, we fit the residuals of the model rather than the response. Then we add to this a new decision tree so as to update the residuals. These new decision trees can be slow in nature, hence the residuals of the model are slowly improved. In the boosting algorithm, the shrinkage parameter, λ slows down the process even further allowing more smaller decision trees to improve the residuals hence improving the performance of the model.

3.4. Stacking

Stacking, can be simply understood as stacking multiple machine learning models on top of each other. The machine learning models pass their predictions to the upper layer and this layer makes decisions based on performances of the models in the layer below. A simple schematic in Figure 3.5 aims at simplifying this.

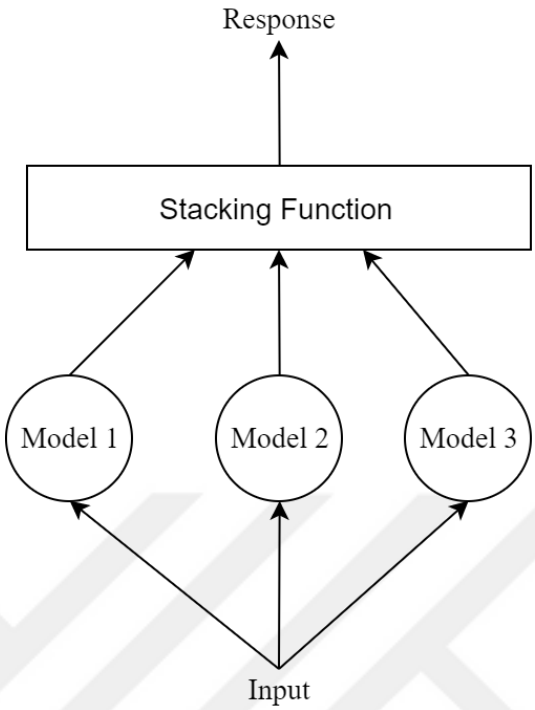


Figure 3.5. Simplified Structure of Stacked Models .

In Figure 3.5, we see two stacked layers of machine learning algorithms. For purpose of simplicity we only included two layers, but an arbitrary amount of layers can be arranged. The bottom layer of machine learning algorithms pass their predictions to the layer above. The layer above takes the outputs of the layers below as its input and produces an output. The main goal of using the stacked model is to increase the performance of our models, in order to achieve that, one must have a clear performance criterion that individual models need to achieve so as to part of the stack. Furthermore, the predictions of the individual models should not be highly correlated. If these predictions have high correlations then combining the models will not result in a better performance of the stacked model.

We can see from Figure 3.5 that the top layer has been labeled as stacking function. This simply means once can incorporate any machine learning algorithm to the stacking function or simpler functions such as weighted average of their performances.

4. DEVELOPING REGRESSION MODELS USING MACHINE LEARNING APPROACHES

In this chapter we shall try and develop the aforementioned machine learning approaches to create regression models. The models will be developed so as to predict the undrained shear strength, c_u, using in-situ parameter, N_{60}, and further by introducing the Atterberg limits (LL,PL, and PI) and water content, w_n. Furthermore, models will also be developed to predict elastic modulus, E_m, and limit pressure, p_L, of the PMT test again by using in-situ parameter N_{60}. It should be noted that all these models were built using the CARET (Classification and Regression Training) package found in R.

4.1. Introducing the Datasets

The dataset used to develop the various machine learning models was developed by requesting data from Zemin Etud ve Tasarim A.S and Geocon Zemin Uzmanlari Ve Mühendislik Ltd. Sti. Additionally, some data was adopted from the thesis of Kamil Özçelik previously of Istanbul Technical University [34]. The dataset consists of two portions. One of them contains parameters such as undrained shear strength, c_u, N_{60}, Atterberg limits and finally water content, w_n (from now on dataset A). While the second data set contains in-situ parameters N_{60}, the elastic modulus, E_m and the limit pressure, p_L (from now on dataset B). These datasets contain 231 and 110 observations respectively. These datasets are split into training and testing sets. The purpose of the training data is to model and tune our algorithm and then test it on the test data.

Table 4.1. Summary of Dataset A.

Variable	**Minimum**	**Maximum**	**Median**	**Mean**
N_{60}	1	53	15	18
w_n (%)	9	96	30	31.6
LL	23	104	56	55.62
PL	11	60	24	24.04
PI	7	62	30	31.8
c_u (kPa)	8	353	74	90.5

Table 4.2. Summary of Dataset B

Variable	**Minimum**	**Maximum**	**Median**	**Mean**
N_{60}	5	50	20	20.63
E_m (kPa)	1844	46540	20685	20583
p_L (kPa)	259	4470	1500	1543

4.2. Data Preprocessing

Before embarking on the building of the models, some preprocessing methods need to satisfied. There are some important aspects that can have significant brunt on our models. Such aspects include presence of non-informative parameters within our dataset. Another is the presence of highly correlated parameters which can result in developing unstable models. Furthermore, presence of outliers within a dataset can lead to drastic model inaccuracies. Outliers are defined as data points that do not conform to the general consensus of the sample data.

4.2.1. Uncovering Outliers

As mentioned above, the presence of outliers in the dataset will lead to building drastically inaccurate models which will furthermore lead to poor predictive performance. To detect the presence of outliers in our model we can refer to Table 4.1 and 4.2. A simple way of detecting presence of outliers can be seen in the values of the median and the mean of the variables. These values if close together indicate that there are no outliers present within the sample data. If these values are very different then it is possible that there are some outliers present. The mean and median values of the undrained shear strength are very far apart, this indicates that there could be some outliers that are augmenting the mean value. To further investigate this we plot a histogram of the undrained shear strength values in an attempt to further uncover these outliers.

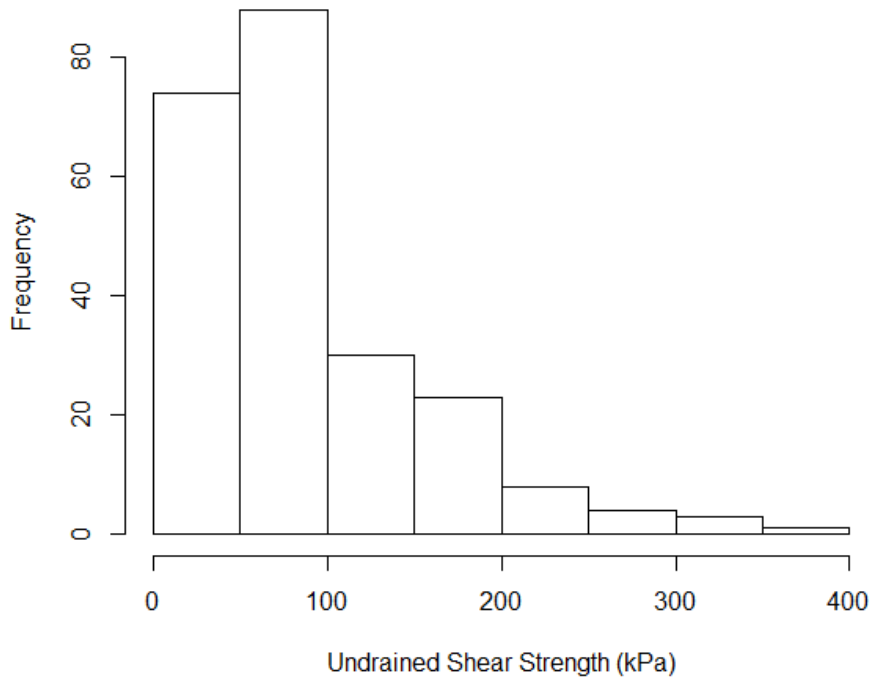


Figure 4.1. Histogram of Undrained Shear Strength.

From Figure 4.1 we can deduce that the bulk of undrained shear strength fall between 0 and 200kPa. Values above this limit are less frequent in the data, thus including them will make drastically decrease our models' predictive performances. Furthermore, SPT tests cannot be properly conducted on soils that have undrained shear strengths greater than 200kPa. It is for this reason that they are not included in the analysis.

4.2.2. Parameter Selection and Correlation

Another factor that may reduce the predictive performances of our models is the presence of non-informative parameters within the data. This however, is not a problem in our data as researchers have determined correlation equations based on the very same predictors before [16, 17]. The next obstacle we must tackle is to determine the correlation among our variables. The Pearson correlation coefficient is a good tool to use so as to determine how correlated our parameters are. The Pearson correlation can be calculated for all parameters and a correlation matrix obtained as seen in Figure 4.2. As seen from the correlation matrix above, none of parameters are highly correlated with one another. This means we use all the parameters as predictors to determine the undrained shear strength.

As seen as well from from Figure 4.3, none of the parameters are highly correlated with each other, hence reduction of the predictors is not necessary to obtain the best model for prediction.

4.3. Tuning the Models

Model tuning involves choosing the best parameters so as to best develop a model with the best possible prediction performance. The Linear regression has no model for tuning so it will not be included in this section.

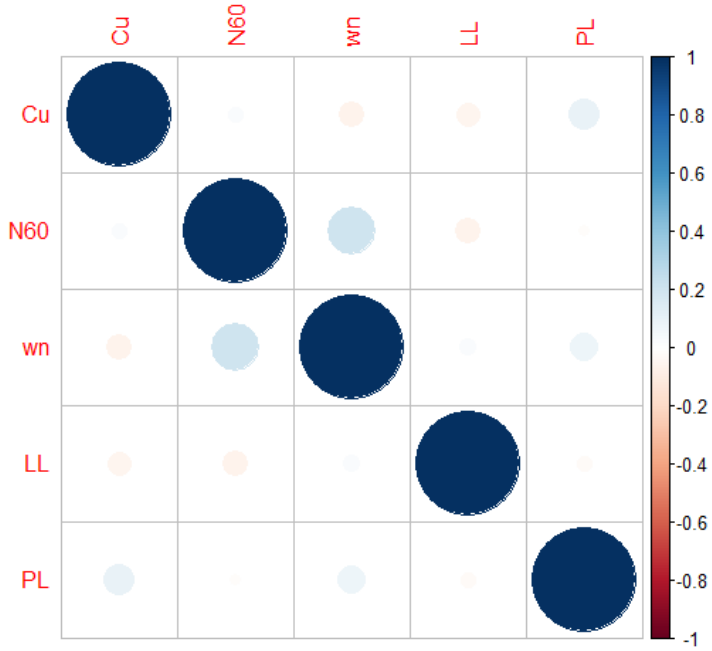


Figure 4.2. Correlation Matrix of Parameters in Dataset A.

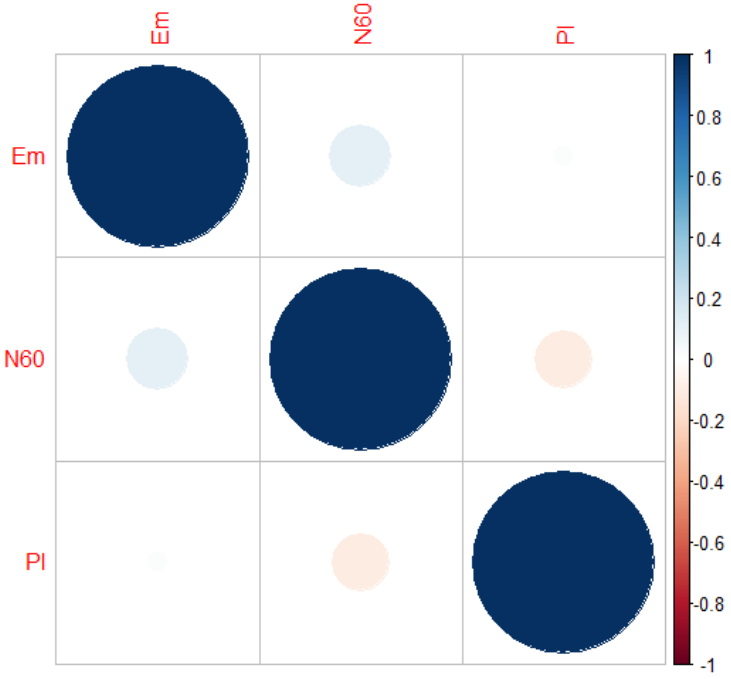


Figure 4.3. Correlation Matrix of Parameters in Dataset B.

4.3.1. Tuning the Random Forest Model

In the Random Forest model, two parameters need to be optimized so as to maximize prediction performance, these parameters are the *ntree* and *mtry* [35]. Though researchers have determined that prediction accuracy is more sensitive to *mtry* than *ntrees* [35, 36]. To determine the best *mtry* parameter, a simple cross-validation will determine the optimizing parameter of our model. Recall that cross-validation data is part of the training data and independent of the test data. Also recall that the *mtry* value cannot be higher than the available number of predictors. We will only need to determine the optimizing *mtry* for dataset A since it has a total of five possible predictors. A 10 fold cross-validation which is repeated 5 times is used to perform a search on a grid of multiple parameter values in order to determine the optimum parameters for the model [5]. Repeating this multiple times and taking the average of the cross-validation results ensures reliability of the cross-validation test. From the results seen in Figure 4.4, we can clearly see that the optimum *mtry* value is equal to 2.

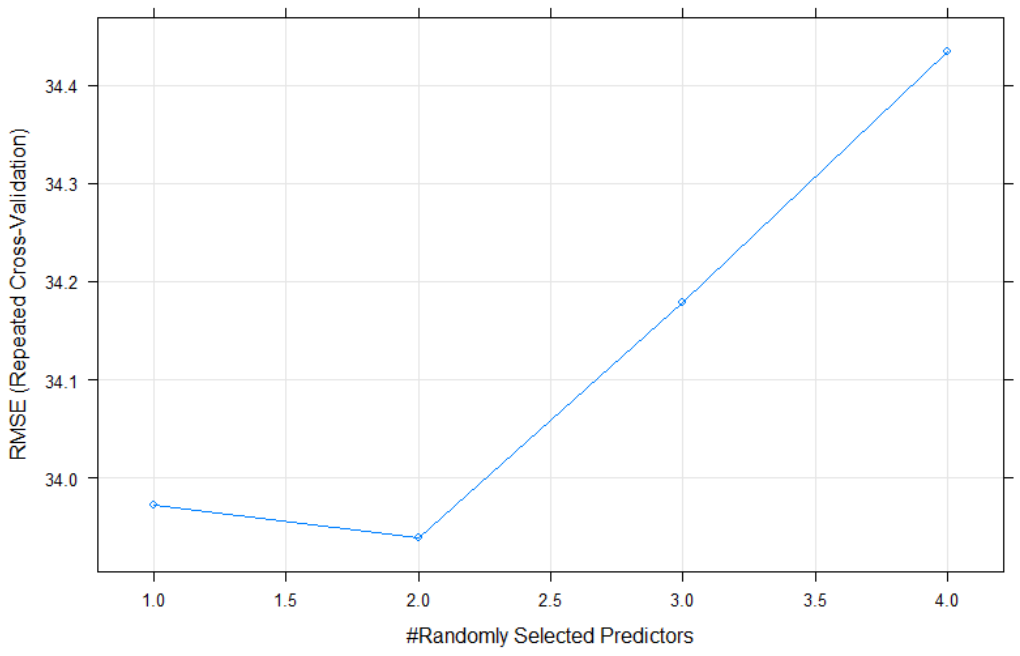


Figure 4.4. Random Forest Tuning Results.

4.3.2. Tuning the Gradient Boosting Model

Recall that boosting involves growing of trees sequentially and that newly grown trees are only grown considering information learned from previously grown trees. The shrinkage parameter λ, number of trees and interaction depth are what need to be tuned so as to obtain the best predictive model. Most machine learning experts set the number of trees to 100 and hence we will do the same for our problem. To determine the best shrinkage parameter and the interaction depth that will give us the optimal predictive model a 10 fold cross-validation repeated 5 times is applied as done for the Random Forest model. From the results presented in Figure 4.5, it can seen the best model parameters for the prediction of the undrained shear strength are λ of 0.11 and interaction depth of 1. Results of all these cross-validations are presented in Figure 4.6 and 4.7.

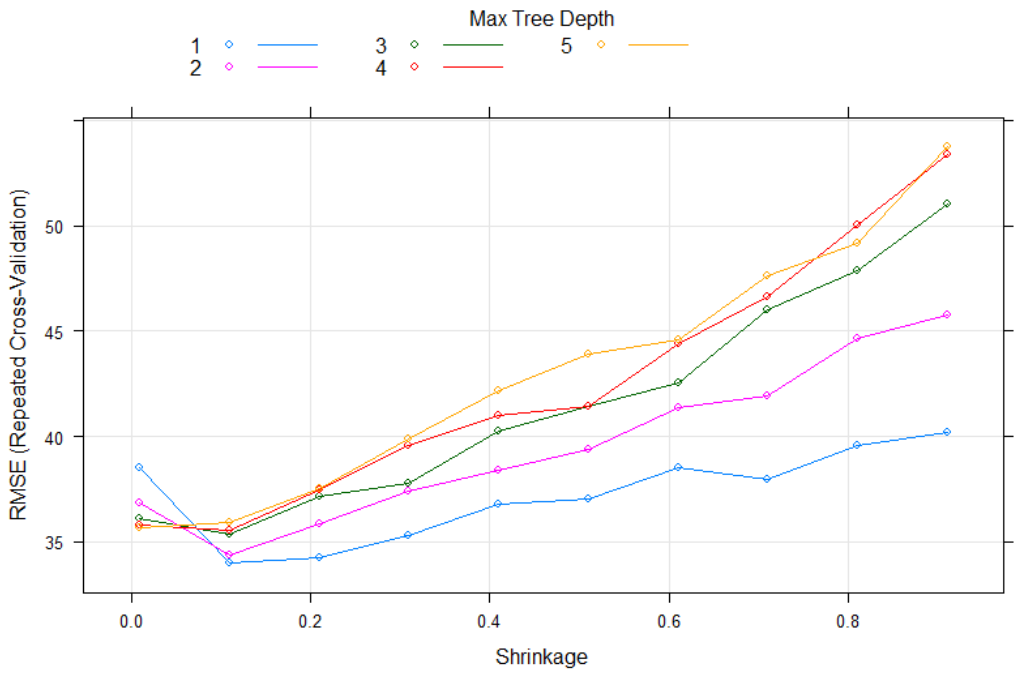


Figure 4.5. Gradient Boosting Tuning Results to Determine Undrained Shear Strength.

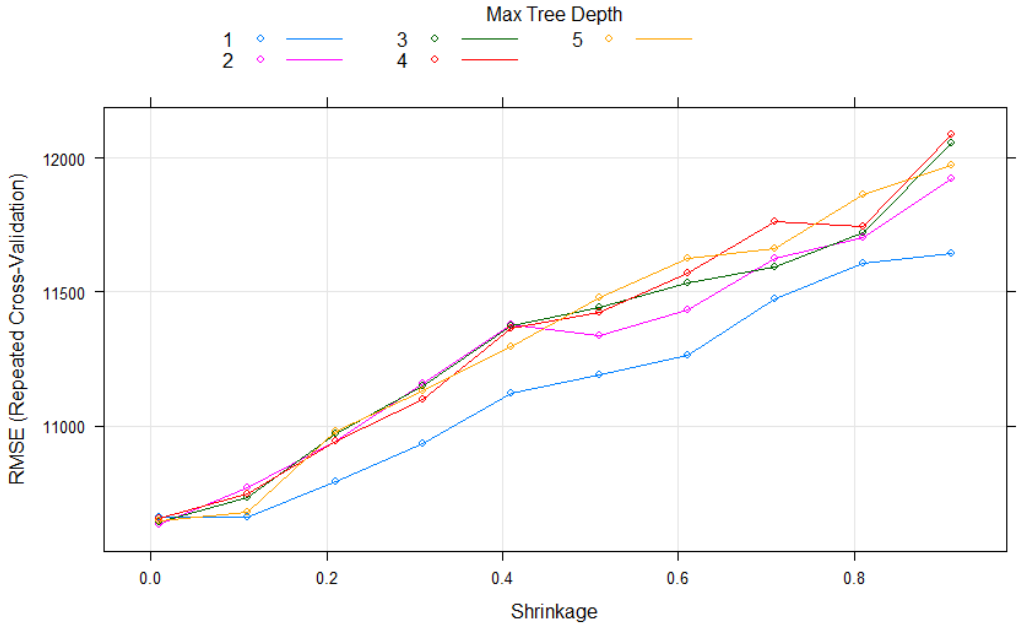


Figure 4.6. Gradient Boosting Tuning Results to Determine Elastic Modulus.

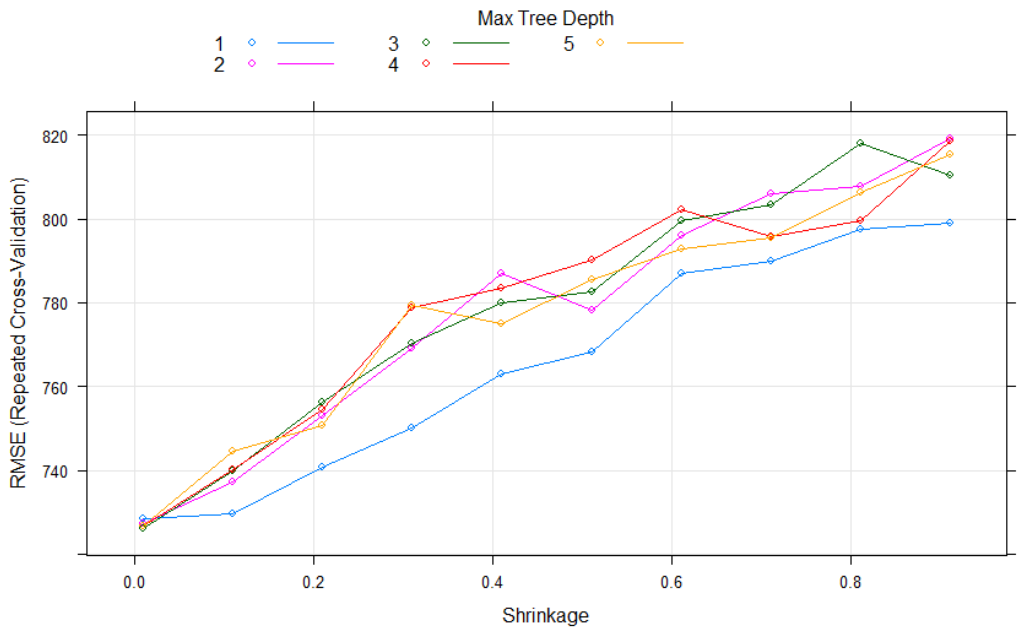


Figure 4.7. Gradient Boosting Tuning Results to Determine Limit Pressure.

4.4. Results of Machine Learning Approaches

Having tuned our models we are ready to see how they will perform when used to predict responses of unseen data i.e test data.

4.4.1. Results of Linear Models

From the analysis done on the R software a linear regression formula with a 95% confidence interval having the coefficients as given in equation 4.1 is developed.

$$c_u(kPa) = 45.86 + 2.61N_{60} - 0.84w_n + 0.40LL - 0.34PL \tag{4.1}$$

As explained in the chapters before, the resulting equation before being used for prediction has to be checked if it is statistically significant. This is simply achieved by printing the summary of the linear model when using R.

Table 4.3. Summary of Statistics of the Linear Model for Dataset A.

Coefficient	**Estimate**	**St.Error**	**t-value**	**Pr** $(>\|t\|)$
Intercept	45.86	10.69	4.29	3.01E-05
$\mathbf{N_{60}}$	2.61	0.245	10.659	2.00E-16
$\mathbf{w_n}$	-0.84	0.267	-3.139	0.028
LL	0.40	0.1332	2.618	0.00978
PL	-0.34	0.563	-0.606	0.5454

From the summary seen in table 4.3, we can clearly see that null hypothesis can be neglected for the coefficients of the intercept, N_{60}, w_n and LL whose p-value < 0.05. However, the same cannot be stated for the remainder of the coefficients. They show p-values that are greater than 0.05. This means that the null hypothesis for these coefficients is valid. This coefficient needs to be removed from the analysis completely and another regression model built.

From the results of the statistical significance of the model, equation 4.1 is therefore re-written as:

$$c_u(kPa) = 44.23 + 2.56N_{60} - 0.93w_n + 0.36LL \tag{4.2}$$

Table 4.4. Summary of Statistics of the Linear Model for Dataset A.

Coefficient	**Estimate**	**St.Error**	**t-value**	**Pr** $(>\|t\|)$
Intercept	44.22	10.322	4.29	3.08E-05
$\mathbf{N_{60}}$	2.56	0.229	11.136	2.00E-16
$\mathbf{w_n}$	-0.934	0.212	-4.404	1.89E-05
LL	0.356	0.164	2.168	0.003

The predicted values obtained by using equation 4.2 on the test data are plotted against their respective observed data in Figure 4.8. The resulting RMSE index is determined as 25.02.

From dataset B, we come up with two linear models, one to predict the elastic modulus, E_m, from in-situ parameter N_{60}, while the other is to predict the limit pressure, p_L, from in-situ parameter N_{60}. The resulting developed equations are shown below.

$$E_m(kPa) = 85340 + 576.7N_{60} \tag{4.3}$$

$$p_L(kPa) = 595.2 + 45.33N_{60} \tag{4.4}$$

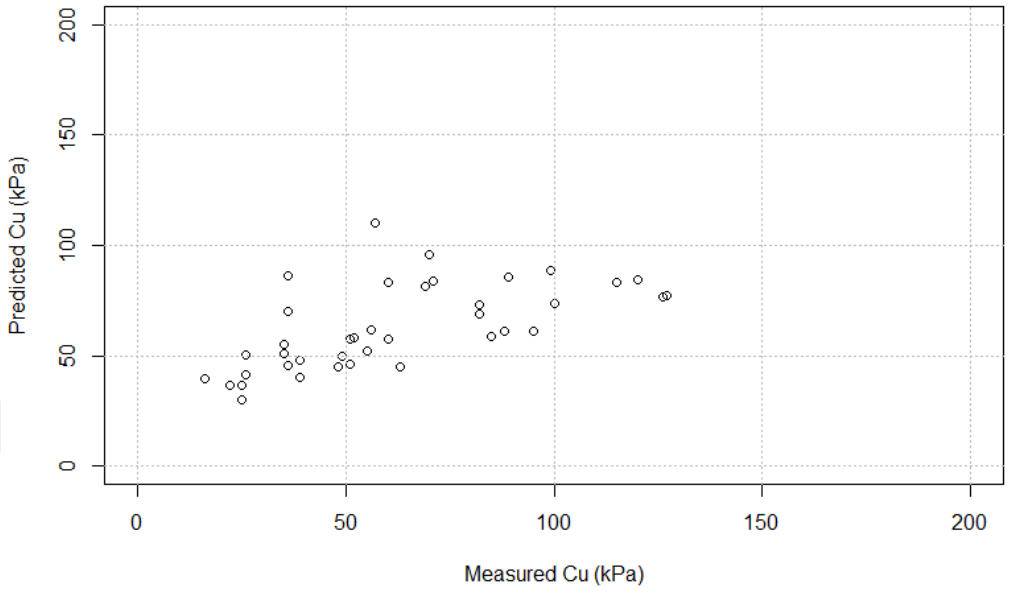


Figure 4.8. Relationship between Measured and Predicted Undrained Shear Strength.

Recall that before using any linear regression formula, the model has to be statistically significant. Again we print the summary of statistics for both models. Examining closer the summary of statistics from Table 4.5 and 4.6, clearly see that our p-value is below the pre-set threshold of 0.05, hence we can safely conclude that both our models are statistically significant.

Table 4.5. Summary of Statistics of the Linear Model for Dataset B Equation 4.3.

Coefficient	**Estimate**	**St.Error**	**t-value**	**Pr**$(>\|\mathbf{t}\|)$
Intercept	8534	2402.6	3.552	0.000605
$\mathbf{N_{60}}$	576.7	104.9	5.496	3.44E-07

Figure 4.9 and 4.10 show the linear relationships between the measured and the predicted responses of the elastic modulus, E_m and limit pressure, p_L. These equations yield RMSE indices of 7.61 and 0.46 respectively.

Table 4.6. Summary of Statistics of the Linear Model for Dataset B Equation 4.4.

Coefficient	**Estimate**	**St.Error**	**t-value**	**Pr**$(>|\mathbf{t}|)$
Intercept	595.29	176.050	3.381	0.00106
$\mathbf{N_{60}}$	45.33	7.703	5.884	6.41E-08

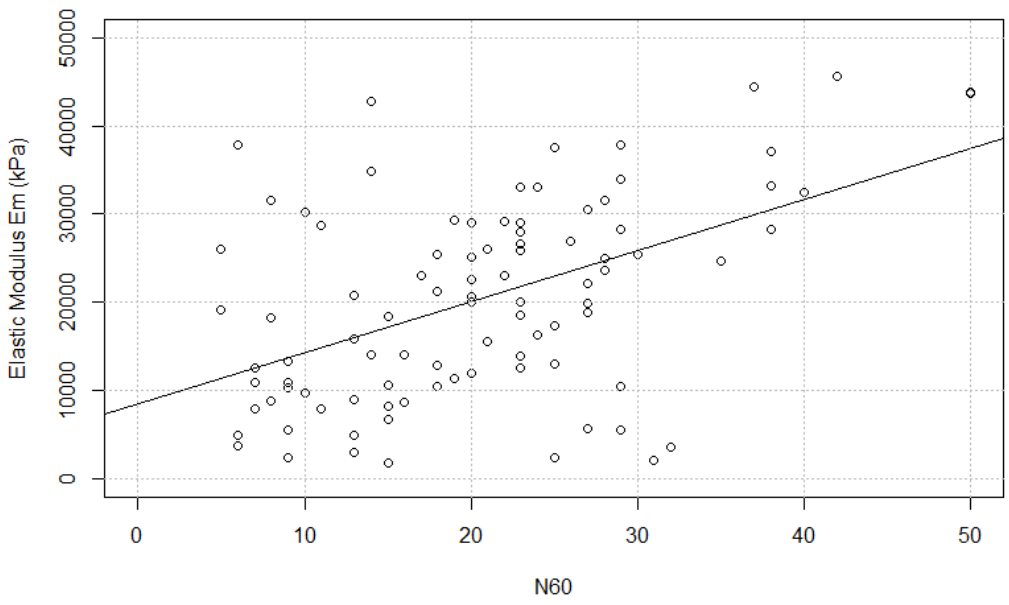


Figure 4.9. Relationship between Measured and Predicted Elastic Modulus.

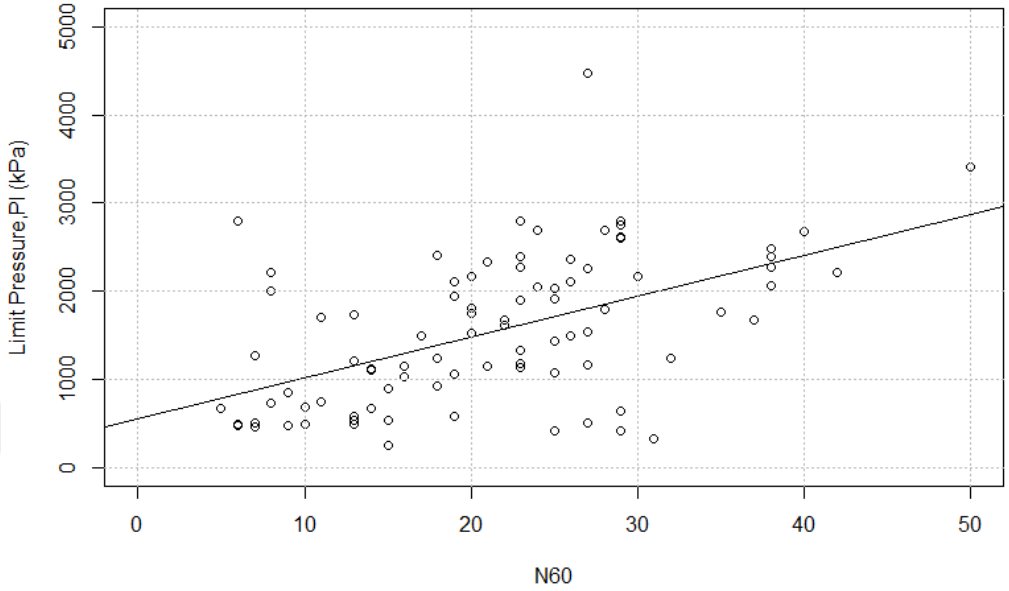


Figure 4.10. Relationship between Measured and Predicted Limit Pressure.

4.4.2. Results of Random Forest Model

Unlike the linear regression models, the Random Forest is unable to generate a tangible equation as to how prediction was achieved. Instead, as explained earlier, it splits the predictor space into regions while trying to minimize the RSS in each region as the splitting takes place. Figure 4.11, 4.12 and 4.13 show the relationships between the measured and predicted parameters as obtained from the Random Forest regression model. These models respectively give RMSE indices of 23.88, 5.8 and 0.34 respectively.

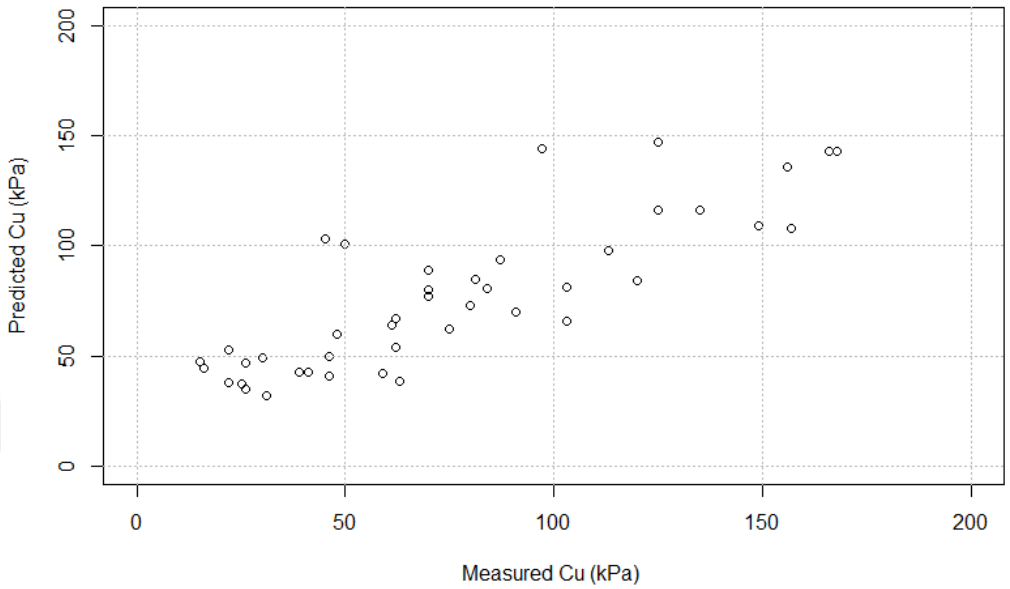


Figure 4.11. Relationship between Measured and Predicted Undrained Shear Strength from Random Forest.

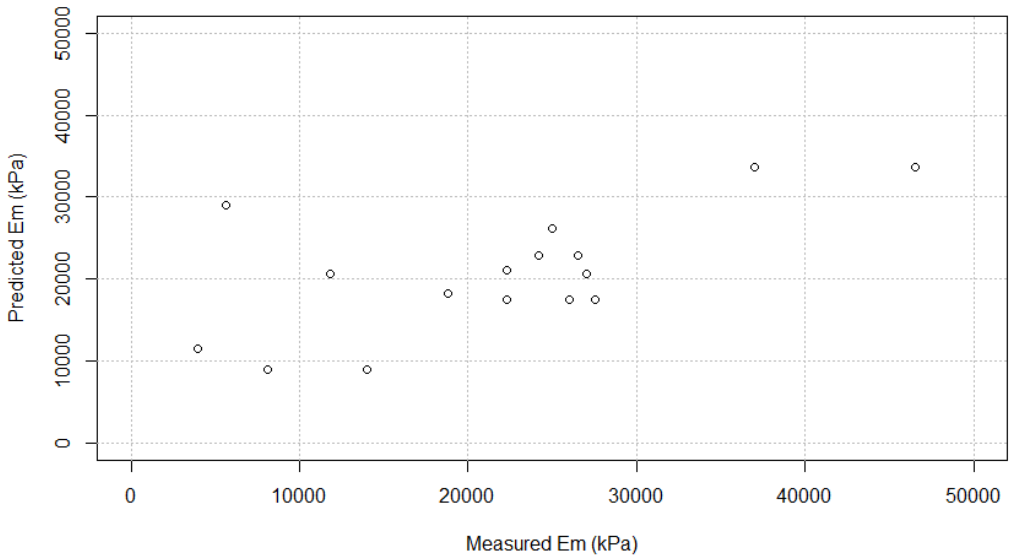


Figure 4.12. Relationship between Measured and Predicted Elastic Modulus from Random Forest.

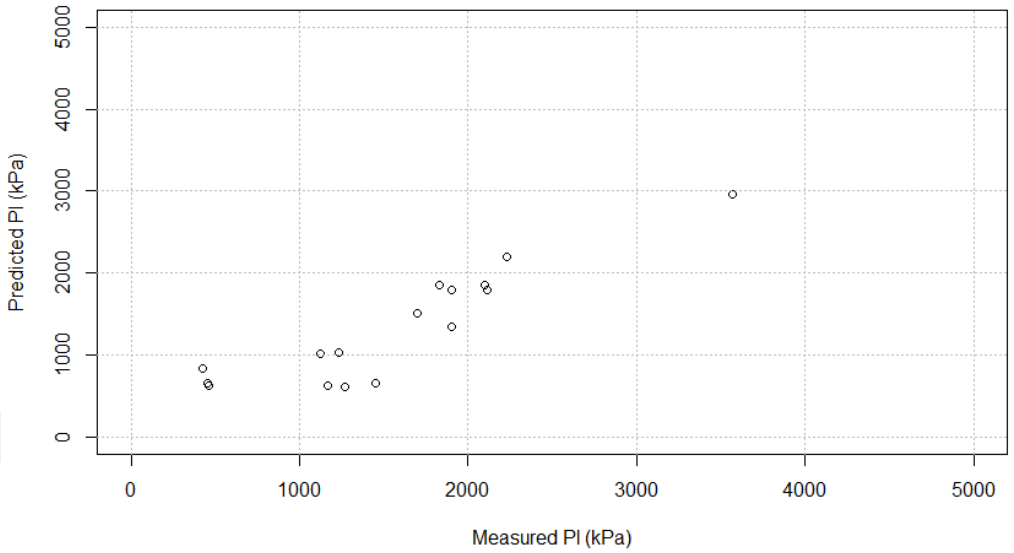


Figure 4.13. Relationship between Measured and Predicted Limit Pressure from Random Forest.

4.4.3. Results of Gradient Boosting Model

Like the Random Forest model and most of the supervised learning algorithms, a tangible equation is not possible to derive. Recall that boosting works by growing trees sequentially, meaning that new trees are grown by using information learned from the previously grown tree. As presented before, the results of the three models are presented. The relationships of the measured and predicted parameters are displayed in Figure 4.14, 4.15 and 4.16.

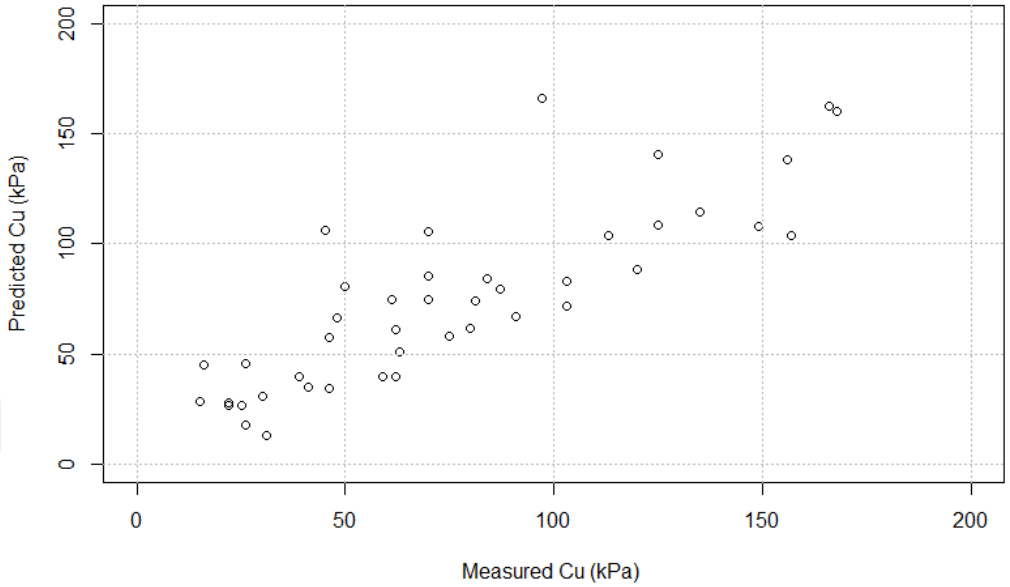


Figure 4.14. Relationship between Measured and Predicted Undrained Shear Strength from Gradient Boosting.

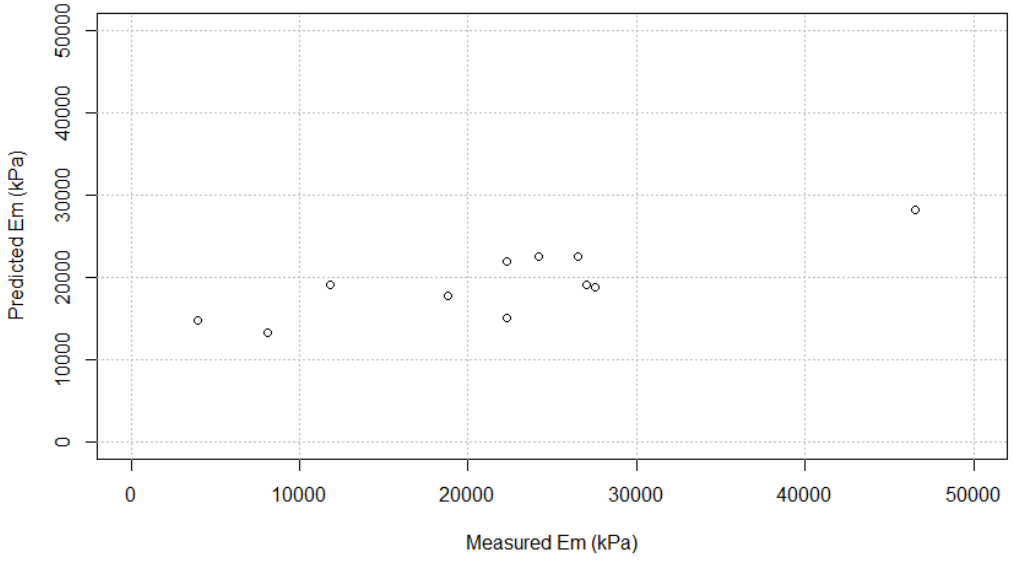


Figure 4.15. Relationship between Measured and Predicted Elastic Modulus from Gradient Boosting Methods.

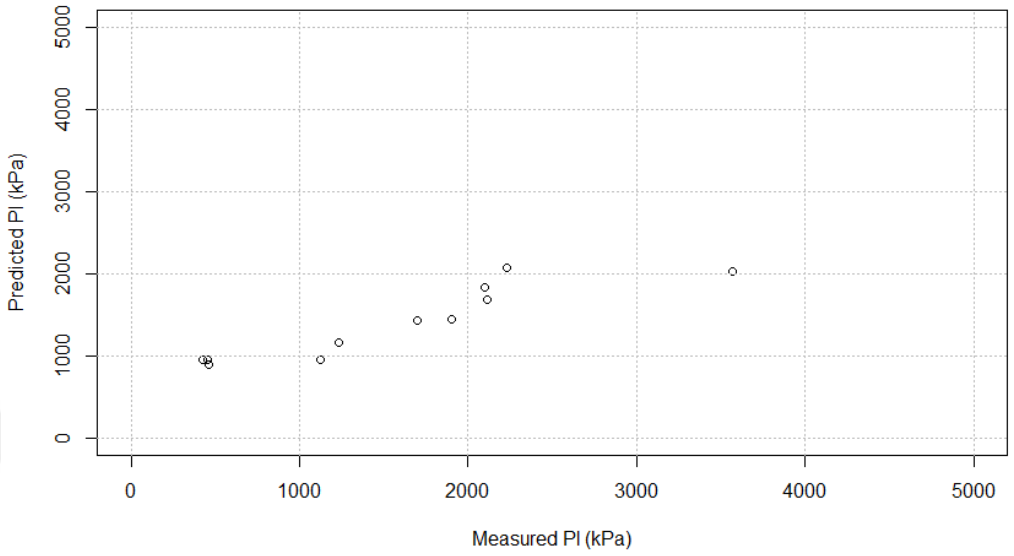


Figure 4.16. Relationship between Measured and Predicted Limit Pressure from Gradient Boosting Methods.

4.5. Developing the Stacked Model

Recall that if we want to furthermore improve the predictive capabilities of out models, we can stacked the already developed models [37]. Recall Figure 3.5, where the lower layers fed their predictions to an upper layer labeled as stacking function. In this thesis, the lower layers are linear regression, Random Forest and Gradient Boosting Models. As mentioned earlier, the weighted average of the prediction performances i.e RMSE indices are used to determine the weights given to the stacking function. Recall that before embarking on creating stacked models, we must ensure that the individual predictions of the models are not highly correlated, in this thesis we have set this correlation value at 0.70. For the predictions obtained from dataset A, the correlation plot is presented below.

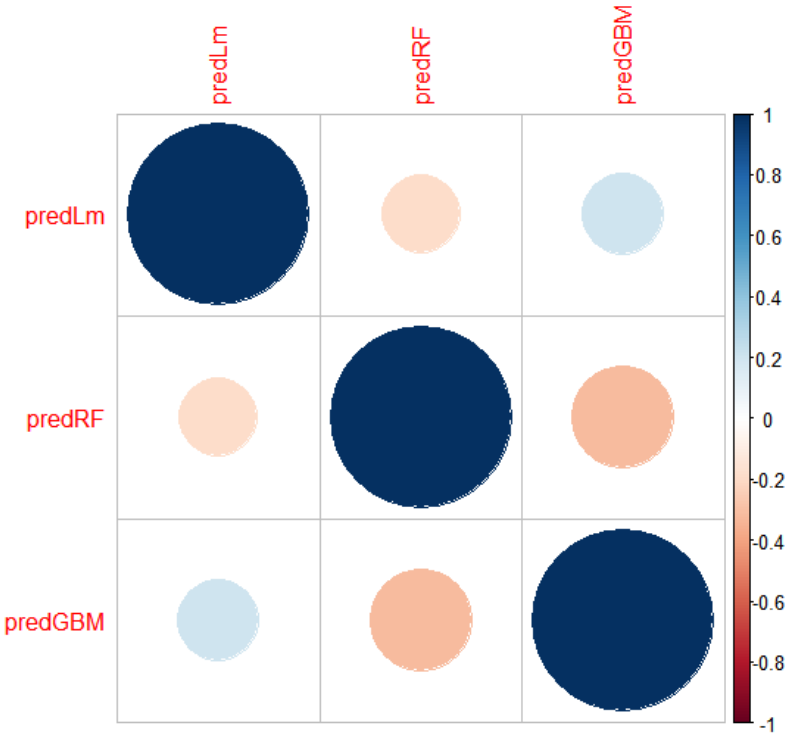


Figure 4.17. Correlation Plot of Individual Predictions of the Models.

It can be clearly seen from Figure 4.17 that the individual predictions of the undrained shear strength by the models are not highly correlated, hence a stacked model can be applied. A stacked model was only developed for models obtained from dataset A as models from dataset B only included one predictor, hence a stacked model would not increase the the prediction performance.

4.6. Comparison of Performances of Machine Learning Approaches.

In this section, we aim at selecting the machine learning approach that predicted the responses of the test data as accurately as possible. As mentioned before, through cross-validation the best possible parameters were selected for the Random forest and the Gradient Boosting approaches. To evaluate performance of the models, like previous researchers the root mean squared error (RMSE) of the observed and predicted values of the test data was evaluated [5, 38].

The purpose of selecting the best prediction performing model is to use this model in comparison with the most commonly used correlation equations from the literature. This comparison, however, is not in the scope of this chapter but the next. We shall start by comparing models developed from dataset A. Recall that dataset A was used to determine the undrained shear strength, c_u, using both in-situ parameters, N_{60}, and laboratory obtained parameters such as the Atterberg limits and water content.

4.6.1. Comparing Models used to Predict Undrained Shear Strength

From Figure 4.18, we can clearly see that all the models did perform reasonably. The Linear Model is seen to have performed the worst among the three supervised machine learning approaches. This can be attributed to the fact that the model assumes a linear relationship exists between the response and the predictors. The Random Forest and Gradient Boosting perform almost similarly since they do not make any assumption between the response and the predictors.

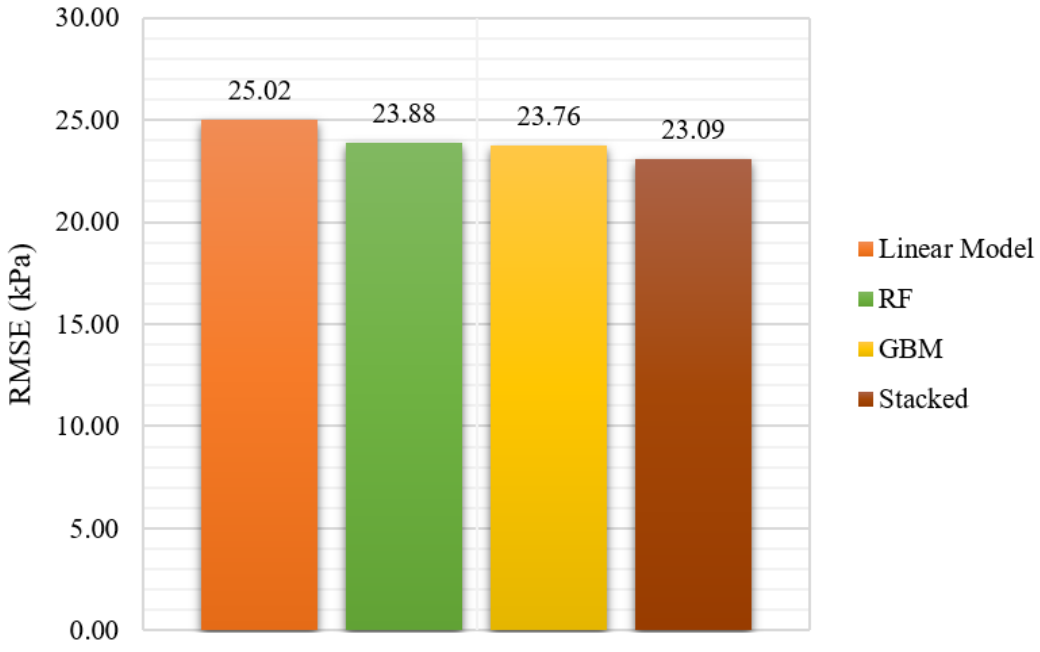


Figure 4.18. RMSE Indices for Various Machine Learning Approaches in Predicting the Undrained Shear Strength.

Additionally to the RMSE values of the models, the coefficients of determination were also computed.

Table 4.7. Results of Models used to Estimate Undrained Shear Strength.

Model	**RMSE (kPa)**	**R^2**
Linear Model	25.02	0.57
Random Forest	23.88	0.71
Gradient Boosting	23.76	0.71
Stacked	23.09	0.72

The Gradient Boosting Model shows the least value of the calculated RMSE of the individual models. Furthermore, we can see a lower RMSE index for the stacked model indicating that combining our three models together does give us a better prediction power. It should be noted that the test data upon which the RMSE values have been calculated is explicitly independent from the training data used to develop these approaches.

4.6.2. Comparing Models used to determine Elastic Modulus and Limit Pressure

Figure 4.19 clearly shows that the Random Forest outperforms the other models when predicting of the elastic modulus is taken into consideration. The Linear Model performs satisfactorily and this, as explained earlier is brought about by the model assuming a linear relationship between the response and set of predictors. Gradient Boosting goes from being the best performer to the worst. We can account this to the few number of predictors in this database. With one predictor, the Gradient Boosting model is unable to come up with a proper predictive model.

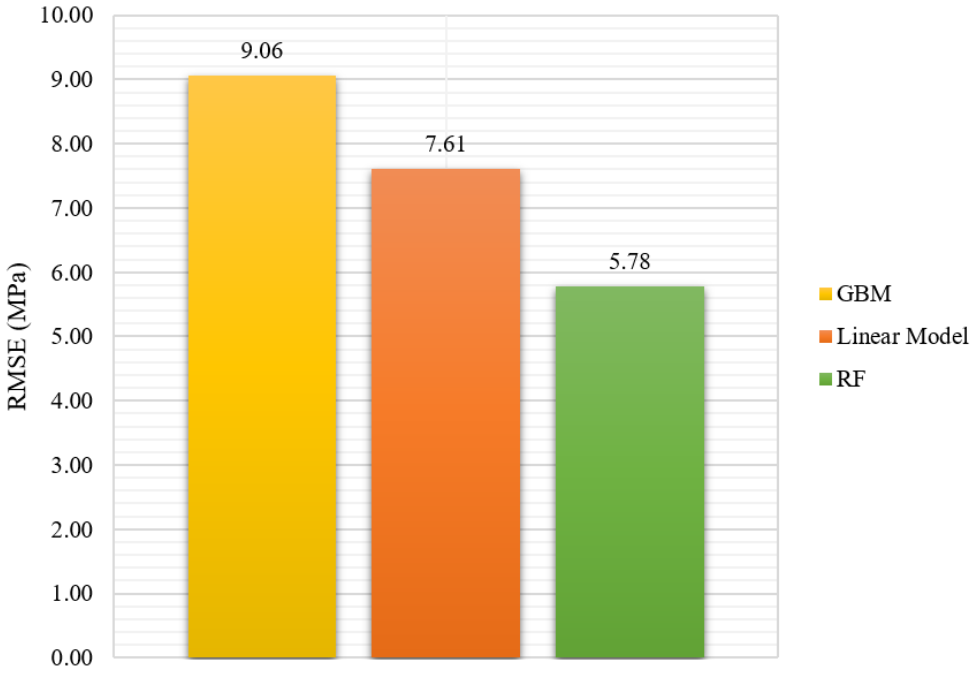


Figure 4.19. RMSE Indices for Various Machine Learning Approaches in Predicting Elastic Modulus.

Similarly, Figure 4.20 shows the Random Forest model again outperforming the other machine learning approaches. In conclusion, we have seen that to predict the elastic modulus and limit pressure from in-situ parameter N_{60}, the Random Forest model gives the best prediction performance. Similarly like presented earlier, the coefficients of determination are presented together with RMSE values.

Table 4.8. Results of Models used to Estimate Elastic Modulus.

Model	**RMSE (MPa)**	**R^2**
Linear Model	7.61	0.70
Random Forest	5.78	0.82
Gradient Boosting	9.06	0.71

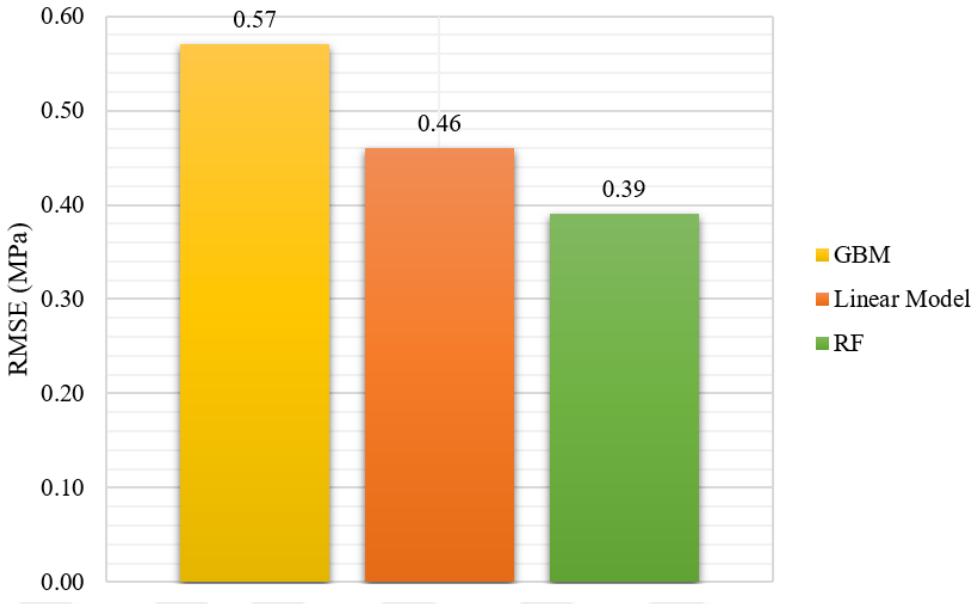


Figure 4.20. RMSE Indices for Various Machine Learning Approaches in Predicting Limit Pressure.

Table 4.9. Results of Models used to Estimate Limit Pressure.

Model	**RMSE (MPa)**	**R^2**
Linear Model	0.46	0.80
Random Forest	0.39	0.93
Gradient Boosting	0.57	0.80

From the results of the models used to predict the Elastic modulus and Limit pressure, it is evident that the Random Forest is a bit superior to the other developed models. To see the competitiveness of our models, we will in the next chapter put them to the test against correlations developed by researchers throughout the years. The stacked model having performed the best in predicting the undrained shear strength, while Random Forest performed the best in predicting both the elastic modulus and limit pressure are selected for comparison.

5. COMPARISON OF MACHINE LEARNING APPROACHES WITH EXISTING CORRELATIONS

In this chapter as it has been previously stated, we will now compare our best performing machine learning approaches with the correlations that exist in the literature. We saw in the previous chapter that the stacked model performed best when predicting the undrained shear strength and Random Forest performed best when predicting the elastic modulus and limit pressure.

5.1. Stacking against Existing Undrained Shear Strength Correlations

We start by comparing our best performing model in predicting the undrained shear strength. Many correlations of predicting c_u exist in the literature but very few exist where there is more than the N_{60} parameter as the predictor. Sivrikaya proposed an equation(Equation.5.1) where the predictors included, N_{60}, water content, w_n, liquid limit, LL, and plasticity index, PI [3]. Another equation that also uses the aforementioned parameters was put forth by researchers from Hormozgan University in Iran (Equation.5.2). These equation are listed below.

$$c_u(kPa) = 4.43N_{60} - 1.29w_n + 1.06LL + 1.02PI \tag{5.1}$$

$$c_u(kPa) = 2N_{60} - 0.4w_n - 1.1LL + 2.4PI + 33.3 \tag{5.2}$$

Using an independent test data that was not involved in the training of the stacked model, we compare the prediction performance of these three models.

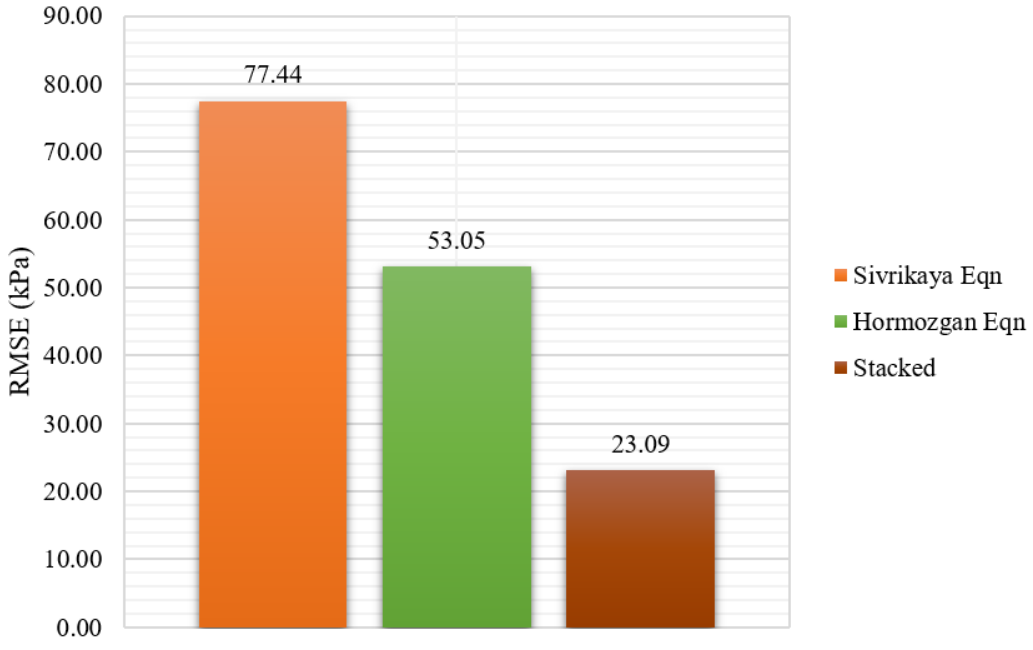


Figure 5.1. RMSE of Stacked Model and Existing Correlations.

It is evident from Figure 5.1, that the stacked model clearly outperforms the existing correlations from the literature. It can be concluded that using machine learning models can increase the estimation accuracy of undrained shear strength. The reason for such a better performance of the stacked model can be attributed to the fact that both Random Forest and Gradient Boosting models do not make any assumptions on the relationship between the predictor and its coinciding variables.

Table 5.1. Comparison of Models used to Estimate Undrained Shear Strength.

Model	**RMSE (kPa)**	**R^2**
Stacked	23.09	0.72
Hormozgan	53.05	0.40
Sivrikaya	77.44	0.52

5.2. Linear Model against Existing Undrained Shear Strength Correlations

The stacked model above is a combination of both the Random Forest and Gradient Boosting Models. However, we would also like to compare how the generated linear model from this study compares with the correlations that exist in the literature. For this, like similarly done throughout, an independent test data is used on both the correlations in the literature and the generated linear model presented in this thesis.

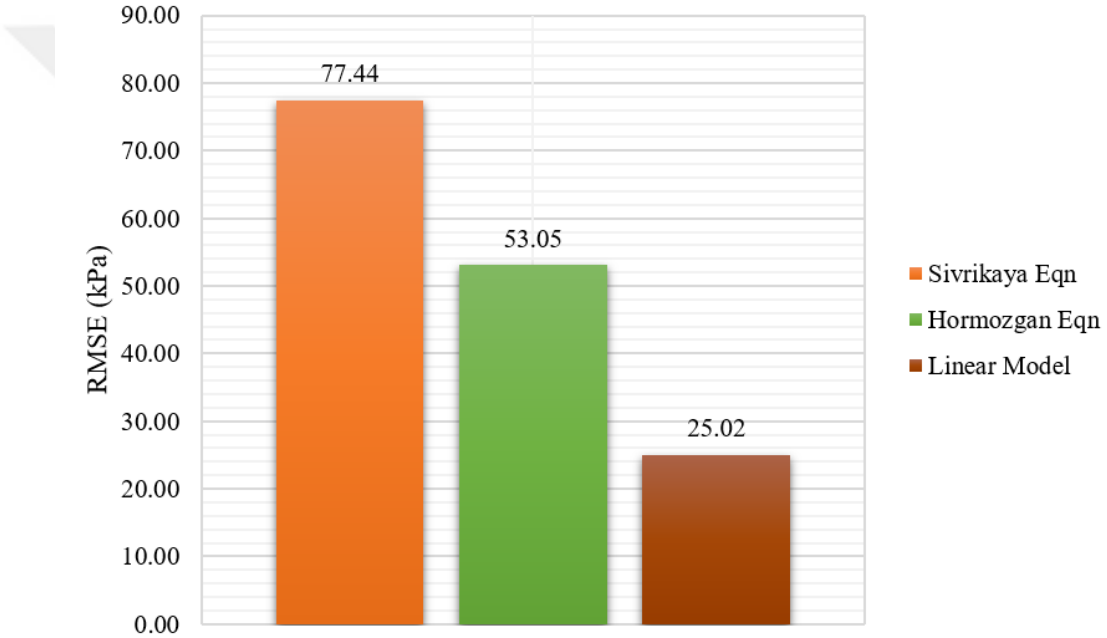


Figure 5.2. RMSE of Linear Model Developed and Existing Correlations.

As evident from Figure 5.2 it is clear that the linear model developed from this thesis outperforms the ones present in the literature. This can be attributed to the different mechanical properties of the soils used to develop these models.

Table 5.2. Comparison of Linear Models to Estimate Undrained Shear Strength.

Model	**RMSE (kPa)**	**R^2**
Linear Model	25.02	0.61
Hormozgan	53.05	0.40
Sivrikaya	77.44	0.52

5.3. Random Forest against Existing Elastic Modulus and Limit Pressure Correlations

Similarly, we would like to see how the Random Forest would compare to the correlations of the elastic modulus and limit pressure. For these comparisons we will the equations proposed by Bozbey and Togrol (Equation 5.3) as well as those proposed by Kayabasi (Equation 5.4).

$$E_m(MPa) = 1.61N_{60}^{0.77} \tag{5.3}$$

$$E_m(MPa) = 0.29N_{60}^{1.4} \tag{5.4}$$

Figure 5.3 shows the plot of the RMSE indices of the existing correlations in the literature and Random Forest developed in this thesis. Random Forest clearly outperforms both equations put forth by the literature.

Furthermore, the same researchers developed Equation 5.5 and 5.6 to predict the limit pressure from in-situ parameter N_{60}.

$$p_L(MPa) = 0.26N_{60}^{0.71} \tag{5.5}$$

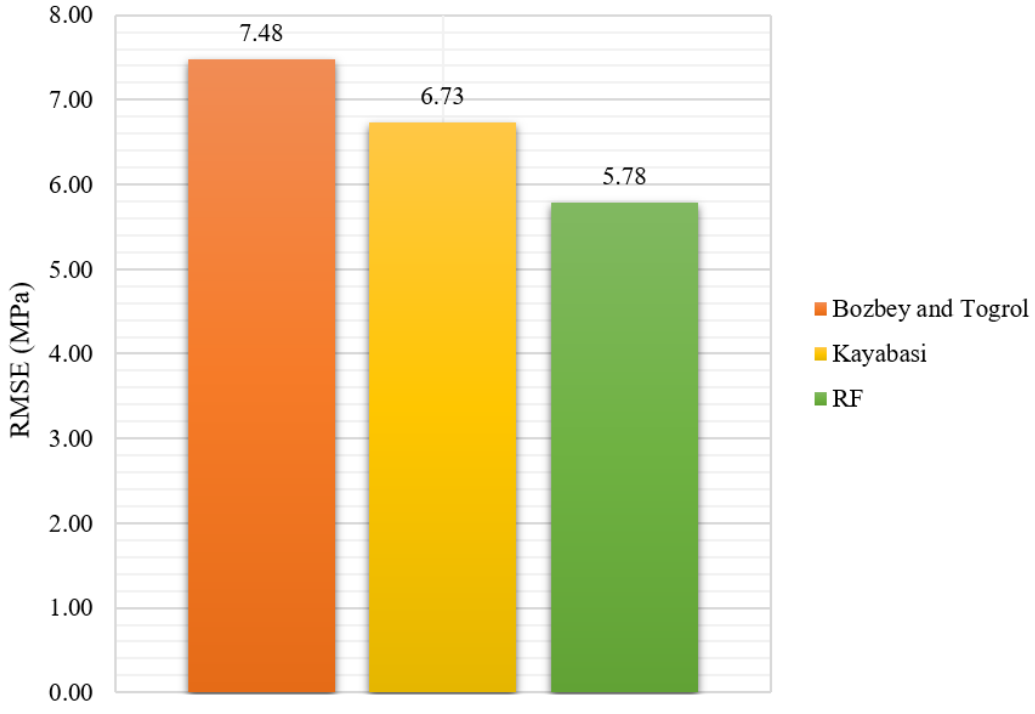


Figure 5.3. RMSE of Random Forest and Existing Correlations of Elastic Modulus.

$$p_L(MPa) = 0.0425 N_{60}^{1.1965} \tag{5.6}$$

An easily interpretable table of both the RMSE of coefficient of determination for models used to determine the elastic modulus is presented in Table 5.3. Results of the RMSE indices in Figure 5.4 clearly show the superiority of Random Forest model over the other correlations that exist in the literature.

Table 5.3. Comparison of Models used to Estimate Elastic Modulus.

Model	**RMSE (MPa)**	**R^2**
Random Forest	5.78	0.82
Bozbey and Togrol	7.48	0.68
Kayabasi	6.73	0.64

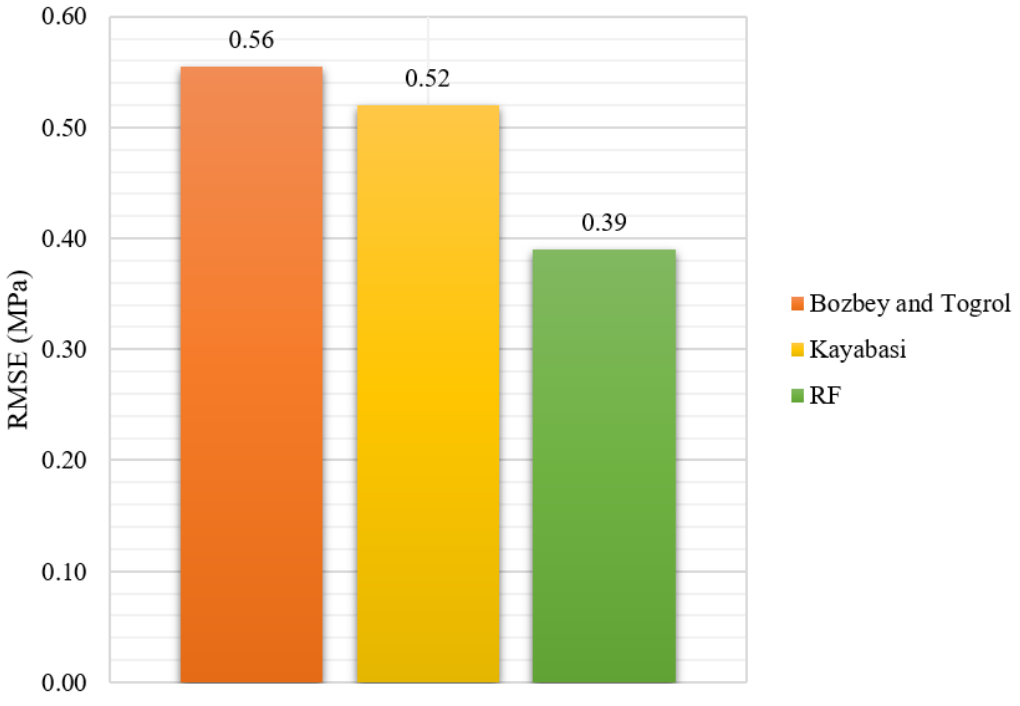


Figure 5.4. RMSE of Random Forest and Existing Correlations of Limit Pressure.

Figure 5.4 and Table 5.4, we see the RMSE and subsequently the coefficient of determination, R^2, of the models compared to the Random Forest. The results indicate that the Random Forest has a better prediction capability compared to the linear models presented in the literature.

Table 5.4. Comparison of Models used to Estimate Limit Pressure.

Model	**RMSE (MPa)**	**R^2**
Random Forest	0.39	0.93
Bozbey and Togrol	0.56	0.83
Kayabasi	0.52	0.80

From the comparison conducted in this chapter, once can easily see the superiority of the machine learning algorithms over the conventionally used equations to predict respective soil parameters. The machine learning algorithms, especially Random Forest out performances the other developed models in this thesis. The Random Forest and Gradient Boosting outperform the conventionally used correlations simply because they have no bias. The linear relationships present in the geotechnical engineering literature assume linear relationships between predictors and their corresponding responses. The results of this thesis show very promising results when machine learning algorithms such as Random Forest and Gradient Boosting are used to predict both undrained shear strength, elastic modulus and limit pressure.

Validation of the built models are explained before was conducted by using an independent test data that was not included in the training phase. The results of the RMSE and R^2 indicate superiority of the built models to the ones that exist in the literature.

6. CONCLUSION

This thesis has aimed to develop machine learning approaches in the field of geotechnical engineering so as to develop more accurate predictive models. Machine learning has gained a lot of popularity recently due to the ever developing field of technology. Data availability has also contributed a lot to this rapid increase in popularity of machine learning approaches. Machine learning has been introduced with all the necessary analyses done on the R programming software.

Three machine learning approaches have been chosen in order to correlate the undrained shear strength, elastic modulus and limit pressure from their respective predictors found in their datasets. All these models have shown acceptable predictive performances. Before embarking on building the model, data preprocessing was performed. This included removal of existing outliers from the sample data and also selecting parameters and checking their correlations. Since the variables present in the sample have been used by previous researchers in determination of undrained shear strength, elastic modulus and limit pressure, these variables were not altered in any manner.

When building the models, a 10 fold cross-validation repeated 5 times was performed. Essentially this means training the model a total of 50 times on the training data set. The purpose of cross-validation is to determine the best parameters for your algorithm that would give the best predictive results. The results of these cross-validations have been clearly presented in the chapters above.

Using an independent set of testing data and RMSE value as a gauge to determine model performance, the built models were then compared among themselves. In predicting the undrained shear strength, the stacked model approach was determined as the best performer while Random Forest outperformed the other approaches in predicting the elastic modulus and limit pressure.

The best performing models again using their RMSE values as gauges to performance were then compared with correlations from different researchers in the literature. Both the stacked model and Random Forest outperformed the correlation equations put forth by the researchers.

From the results of this thesis, we deduce the possibility of using machine learning algorithms in the geotechnical engineering field. The predictive abilities of stacking and Random Forest far outperform that of the conventionally used linear regression. Hence, when accuracy of the response is a critical matter, using these algorithms leads to closer predictions to the actual values. With the presence of essentially one line codes to ran the Random Forest and stacked model algorithms, engineers with little understanding of programming can be able to use them. However, from the results of the models we see that with a properly built linear model, the resulting equation can be used to predict the undrained shear strengths with promising predictive capabilities.

In conclusion, the performance capabilities of these algorithms depends highly on the quality of the data being fed into the model. Data that has not been properly sampled hence containing missing or outliers from the field or the laboratory will result in poor predictive performance of the models. Furthermore, to properly present these algorithms into the field, a larger dataset should be taken into consideration. The more the algorithm learns, the better it is able to predict known and unknown responses.

REFERENCES

1. P.Coduto, D., M. chu Ronald Yeung and W. A. Kitch, *Geotechnical Engineering Principles and Practices*, Pearson Education Inc., Upper Saddle River, New Jersey, 2011.

2. Üzeler, V., *Compressibility of Clays Determined from In-Situ Tests*, MS. Thesis, Middle Eastern Technical University, 2013.

3. O.Sivrikaya, "Comparison of Artificial Neural Networks Models with Correlative Works on Undrained Shear Strength", *Eurasian Soil Science, 2009, No.13 pp.1487-1496*, Vol. 42, 2009.

4. Erzin, Y. and N. Gunes, "The Prediction of Swell Percentage and Swell Pressure by Using Neural Networks", *Mathematical and Computational Applications*, Vol. 16, 2011.

5. Zhou, J., X. Shi, K. Du, X. Qiu, X. Li and H. S. Mitri, "Feasibility of Random-Forest Approach for Prediction of Ground Settlements Induced by the Construction of a Shield-Driven Tunnel", *International Journal of Geomechanics*, Vol. 17, No. 6, p. 04016129, 2016.

6. Kayabasi, A., "Prediction of pressuremeter modulus and limit pressure of clayey soils by simple and non-linear multiple regression techniques: a case study from Mersin, Turkey", *Environmental Earth Sciences*, Vol. 66, No. 8, pp. 2171–2183, 2012.

7. James, G., D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning*, Vol. 112, Springer, 2013.

8. Efron, B. and R. Tibshirani, "Improvements on cross-validation: the 632+ bootstrap method", *Journal of the American Statistical Association*, Vol. 92, No. 438,

pp. 548–560, 1997.

9. Sivrikaya, O. and E. Toğrol, "Determination of undrained strength of fine-grained soils by means of SPT and its application in Turkey", *Engineering geology*, Vol. 86, No. 1, pp. 52–69, 2006.

10. Schmertmann, J. H. and A. Palacios, "Energy dynamics of SPT", *Journal of the Geotechnical Engineering Division*, Vol. 105, No. 8, pp. 909–926, 1979.

11. Robertson, P., R. Campanella and A. Wightman, "Spt-Cpt Correlations", *journal of geotechnical engineering*, Vol. 109, No. 11, pp. 1449–1459, 1983.

12. Skempton, A., "Standard penetration test procedures and the effects in sands of overburden pressure, relative density, particle size, ageing and overconsolidation", *Geotechnique*, Vol. 36, No. 3, pp. 425–447, 1986.

13. Kalantary, F., H. Ardalan and N. Nariman-Zadeh, "An investigation on the S u–N SPT correlation using GMDH type neural networks and genetic algorithms", *Engineering Geology*, Vol. 104, No. 1, pp. 144–155, 2009.

14. Stroud, M., "The standard penetration test in insensitive clays and soft rocks", *Proceedings of the 1st European Symposium on Penetration Testing, Stockholm, Sweden*, Vol. 2, pp. 367–375, 1974.

15. Terzaghi, K., R. B. Peck and G. Mesri, *Soil mechanics in engineering practice*, John Wiley & Sons, 1996.

16. Sivrikaya, O. and E. Toğrol, "Relations between SPT-N and qu", *5th international congress on advances in civil engineering, Istanbul*, pp. 943–952, 2002.

17. Nassaji, F. and B. Kalantari, "SPT capability to estimate undrained shear strength of fine-grained soils of Tehran, Iran", *Electronic Journal of Geotechnical Engineering*, Vol. 16, pp. 1229–1238, 2011.

18. Sanglerat, G., "The penetrometer and soil exploration: Interpretation of penetration diagrams-theory and practice. 1 vols", *Developments in geotechnical engineering (Print), ISSN*, pp. 0165–1250, 1972.

19. Décourt, L., "The standard penetration test, state-of-the-art report", *Proc. 12th ICSMFE, Rio De Janeiro*, Vol. 4, pp. 2405–2416, 1989.

20. Hettiarachchi, H. and T. Brown, "Use of SPT blow counts to estimate shear strength properties of soils: Energy balance approach", *Journal of Geotechnical and Geoenvironmental Engineering*, Vol. 135, No. 6, pp. 830–834, 2009.

21. Clarke, B. G., *Pressuremeters in geotechnical design*, CRC Press, 1994.

22. Menard, L., *Rules for the Calculation of Bearing Capacity and Foundation Settlement Based on Pressure-meter Tests*, Corps of Engineers, US Army Cold Regions Research and Engineering Laboratory, 1972.

23. JE, B., *Foundation Analysis and Design*, The McGraw - Hill Co. Inc, Singapore, 1997.

24. Briaud, J.-L., *The pressuremeter*, AA Balkema, 1992.

25. Baguelin, F., *The pressuremeter and foundation engineering*, Trans Tech publications., 1978.

26. Yagiz, S., E. Akyol and G. Sen, "Relationship between the standard penetration test and the pressuremeter test on sandy silty clays: a case study from Denizli", *Bulletin of engineering geology and the environment*, Vol. 67, No. 3, pp. 405–410, 2008.

27. Chiang, Y. and Y. Ho, "Pressuremeter method for foundation design in Hong Kong", *Proceedings of the Sixth Southeast Asian Conference on Soil Engineering*, pp. 31–42, 1980.

28. Ohya, S., T. Imai and M. Matsubara, "Relationships between N value by SPT and LLT pressuremeter results", *Proceedings 2nd European Symposium on Penetration Testing*, Vol. 1, pp. 125–130, 1982.

29. Bozbey, I. and E. Togrol, "Correlation of standard penetration test and pressuremeter data: a case study from Istanbul, Turkey", *Bulletin of engineering geology and the environment*, Vol. 69, No. 4, pp. 505–515, 2010.

30. Gonin, H., P. Vandangeon and M. Lafeullade, "Correlation study between standard penetration and pressuremeter tests", *Revue Française de Géotechnique*, Vol. 58, pp. 67–78, 1992.

31. Hobbs, N. and J. Dixon, "In-Situ Testing for Bridge Foundations in the Devonian Marl", , 1900.

32. Seber, G. A. and A. J. Lee, *Linear regression analysis*, Vol. 936, John Wiley & Sons, 2012.

33. Liaw, A., M. Wiener *et al.*, "Classification and regression by randomForest", *R news*, Vol. 2, No. 3, pp. 18–22, 2002.

34. Özçelik, K., *Zemin İncelemelerinde Standart Penetrasyon ve Koni Deneyleri*, MS. Thesis, Istanbul Technical University, 2013.

35. Kuhn, M. and K. Johnson, *Applied predictive modeling*, Vol. 810, Springer, 2013.

36. Zhou, J., X. Li and H. S. Mitri, "Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction", *Natural Hazards*, Vol. 79, No. 1, pp. 291–316, 2015.

37. Breiman, L., "Stacked regressions", *Machine learning*, Vol. 24, No. 1, pp. 49–64, 1996.

38. Aladag, C., A. Kayabasi and C. Gokceoglu, "Estimation of pressuremeter modulus

and limit pressure of clayey soils by various artificial neural network models", *Neural computing & applications*, pp. 1–7, 2013.